GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Entity-Centric Text Mining for Historical Documents

Dissertation for the award of the degree

Doctor of Philosophy (PhD)
Division of Mathematics and Natural Sciences

of the Georg-August-Universität Göttingen

within the Programme in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

Maria Coll Ardanuy
from Ivars d'Urgell, Spain

Göttingen, 2017

Member of the Thesis Committee:

Prof. Dr. Caroline Sporleder
Institute of Computer Science
Göttingen Centre for Digital Humanities (GCDH)
Georg-August-Universität Göttingen

Member of the Thesis Committee:

Prof. Dr. Ramin Yahyapour
Institute of Computer Science
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Georg-August-Universität Göttingen

Members of the Examination Board:

Reviewer:
Prof. Dr. Caroline Sporleder (Institute of Computer Science, Georg-August-Universität Göttingen)
Second Reviewer:
Prof. Dr. Ramin Yahyapour (Institute of Computer Science, Georg-August-Universität Göttingen)
Further members of the Examination Board:
Prof. Dr. Ulrich Heid (Institute of Information Science and Language Technology, Universität Hildesheim)
Prof. Dr. Dieter Hogrefe (Institute of Computer Science, Georg-August-Universität Göttingen)
Prof. Dr. Gerhard Lauer (Department of German Philology, Georg-August-Universität Göttingen)
Prof. Dr. Wolfgang May (Institute of Computer Science, Georg-August-Universität Göttingen)

Date of the oral examination: July 7th, 2017

Abstract

Recent years have seen an important increase of digitization projects in the cultural heritage domain. As a result, growing efforts have been directed towards the study of natural language processing technologies that support research in the humanities. This thesis is a contribution to the study and development of new text mining strategies that allow a better exploration of contemporary history collections from an entity-centric perspective. In particular, this thesis focuses on the challenging problems of disambiguating two specific kinds of named entities: toponyms and person names. They are approached as two clearly differentiated tasks, each of which exploiting the inherent characteristics that are associated to each kind of named entity.

Finding the correct referent of a toponym is a challenging task, and this difficulty is even more pronounced in the historical domain, as it is not uncommon that places change their names over time. The method proposed in this thesis to disambiguate toponyms, *GeoSem*, is especially suited to work with collections of historical texts. It is a weakly-supervised model that combines the strengths of both toponym resolution and entity linking approaches by exploiting both geographic and semantic features. In order to do so, the method makes use of a knowledge base built using Wikipedia as a basis and complemented with additional knowledge from GeoNames. The method has been tested on a historical toponym resolution benchmark dataset in English and improved on the state of the art. Furthermore, five datasets of historical news in German and Dutch have been created from scratch and annotated. The method proposed in this thesis performs significantly better on them than two out-of-the-box state-of-the-art entity linking methods when only locations are considered for evaluation.

Person names are likewise highly ambiguous. This thesis introduces a novel method for disambiguating person names from news articles. The method, *SNcomp*, exploits the relation between the ambiguity of a person name and the number of entities referred to by it. Modeled as a clustering problem in which the number of target entities is unknown, the method dynamically adapts its clustering strategy to the most suitable configuration for each person name depending on how common this name is. *SNcomp* has a strong focus on social relations and returns sets of automatically created social networks of disambiguated person entities extracted from the texts. The performance of the method has been tested on three person name disambiguation benchmark datasets in two different languages and is on par with the state of the art reported for one of the datasets, while using less specific resources.

This thesis contributes to the fields of natural language processing and digital humanities. Information about entities and their relations is often crucial for historical research. Both methods introduced in this thesis have been designed and developed with the goal of assisting historians in delving into large collections of unstructured text and exploring them through the locations and the people that are mentioned in them.

Acknowledgements

This dissertation would not have been completed without the support and encouragement of many people. First and foremost, I wish to thank my supervisor, Professor Caroline Sporleder, from whom I have learned much, for her trust, dedicated support and guidance during these years. I would also like to thank Professor Ramin Yahyapour, for accepting to be part of this thesis committee and sharing with me some insightful comments to improve the structure of this dissertation. I would also like to gratefully acknowledge Professors Ulrich Heid, Dieter Hogrefe, Gerhard Lauer, and Wolfgang May, for their willingness to serve on my examination committee.

I am particularly grateful to all the members of the AsymEnc project: Joris van Eijnatten, Jaap Verheul, Toine Pieters, Maarten van den Bos, and Hermione Giffard from Utrech University; and Ulrich Tiedau and Tessa Hauswedell from University College London. It has been a great pleasure to have had the chance to work with you all and to be able to be part of a project that was making tangible progress towards applying text mining techniques to historical reasearch. I especially thank Maarten, who was always open to new ideas, and with whom I worked the most and learned (indirectly and quite unexpectedly) much on European integration and Dutch politics. I consider myself fortunate to have been able to work closely with a historian and see and understand first-hand the challenges this domain poses.

I started this PhD at the University of Trier and continued it at the University of Göttingen. I am very grateful to everyone that made the transition possible and in such a smooth manner. I also want to thank my colleagues in Trier and Göttingen, and in particular to Jürgen Knauth and Andrei Beliankou, who were always ready to help and share their knowledge

with me. I would like to thank Andrei also for helping me prepare and set up the annotation scheme for my toponym disambiguation task. In this respect, I would also like to say a big thank you to the three annotators, Anne, Katharina, and Michael, for persevering in completing a task that may not have been the most exciting one.

During the period of this investigation, I have lived in three different cities. To all my friends in Saarbrücken, Trier, and Göttingen (you know who you are!), an enormous thank you for being an essential part of my life and for being at my side both in the good and the difficult times. There aren't enough words to express how important your friendship has been to me. A very special thank you goes to Evi, for being the best friend one could hope for since my very first days in Germany, and to Pedro, who, even when thousands of kilometers away, has always been there encouraging me all the way; it will be your turn soon.

Finally, to my loving family goes the biggest and most heartfelt thank you for always believing in me and for your unconditional moral and emotional support. I dedicate this thesis to you.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Glossary

List of terms as used in this thesis. The task to which they are related is indicated in square parentheses after their definition: *[TopDis]* for toponym disambiguation and *[PerDis]* for person name disambiguation. The SMALL CAPS indicate a term that is also present and defined in the glossary.

candidate location Also referred to as *candidate*. Any LOCATION that may be referred to by a given TOPONYM. *[TopDis]*

candidate selection The task that aims at selecting all the LOCATIONS that can possibly be referred to by a TOPONYM that has been identified in text. *[TopDis]*

cross-document coreference resolution The task that aims at clustering documents that contain MENTIONS to the same PERSONS. *[PerDis]*

entity A real-world item with a distinct existence. *[PerDis][TopDis]*

entity linking The task that aims at linking MENTIONS to the entries in a knowledge base corresponding to the ENTITIES they refer to, if existing. *[PerDis][TopDis]*

full name A PERSON NAME with at least two tokens (ideally a first and last name, even though this is not necessarily always the case). *[PerDis]*

gazetteer A dictionary of TOPONYMS. *[TopDis]*

geographic reference Also referred to as *geographic footprint*. The unambiguous representation of a LOCATION on the Earth's surface. *[TopDis]*

location Also referred to as *place*. An ENTITY that occupies a spatial area on the Earth's surface. *[TopDis]*

location linking Subtask of ENTITY LINKING. The task that aims at linking TOPONYMS to the entries in a knowledge base corresponding to the LOCATIONS they refer to, if existing. *[TopDis]*

mention The expression of an ENTITY in a text. *[PerDis][TopDis]*

mention name A PERSON NAME that is mentioned in a document and that is not the QUERY NAME. *[PerDis]*

namepart Each of the tokens that form a FULL NAME. *[PerDis]*

non-person mention A MENTION that does not refer to a PERSON. *[PerDis]*

person A human ENTITY. *[PerDis]*

person linking Subtask of ENTITY LINKING. The task that aims at linking PERSON NAMES to the entries in a knowledge base corresponding to the PERSONS they refer to, if existing. *[PerDis]*

person name A MENTION in a text referring to a PERSON. *[PerDis]*

person name disambiguation An umbrella term for the various very similar tasks that aim at distinguishing between different ENTITIES that share the same PERSON NAME. *[PerDis]*

query name In a text, the target PERSON NAME to disambiguate. *[PerDis]*

toponym Also referred to as *place name*. A MENTION in a text referring to a LOCATION. *[TopDis]*

toponym disambiguation The task that aims at finding the most likely CANDIDATE LOCATION to be the referent of a given TOPONYM regardless of its output representation. *[TopDis]*

toponym identification The task that aims at detecting TOPONYMS in a text. *[TopDis]*

toponym resolution The task that aims at disambiguating a TOPONYM by its GEOGRAPHIC REFERENCE. *[TopDis]*

web people search The task that aims at discriminating among different PERSON NAMES in the web domain. *[PerDis]*

Chapter 1

Introduction

The rise of digitization in the cultural heritage domain has opened the way to broaden the boundaries of historical research. As more and more historical textual materials are digitized, new possibilities arise to explore contemporary history in a way that was infeasible until recent years. Natural language processing has traditionally paid relatively little attention to the cultural heritage domain. However, this is fortunately changing steadily. Working with historical documents provides the computational linguist with a series of additional challenges that are often not present when working with modern documents: apart from the obvious and well-identified problem of optical character recognition and non-standard text, historical text mining faces the difficulty of dealing with entities that in many cases no longer exist, and traces them, together with the events that surrounded them, through time.

This thesis introduces two entity-centric text mining methods that are aimed at assisting historians in exploring large collections of digitized historical documents. Entities can act as gates through which to explore and mine historical texts, but the names that are used to refer to them are often ambiguous, therefore hindering the retrieval process. The focus of this thesis is on disambiguating two distinct types of named entities: toponyms and person names.

1.1 Background

The task of disambiguating toponyms and person names has been approached from different perspectives. Some methods, such as most **entity linking** methods, attempt

to link named entities to entries in a knowledge base. Since these kinds of methods do not discriminate between different kinds of entities, they usually treat toponym and person name disambiguation as a single task and attempt to resolve both problems simultaneously. Other methods prefer to treat toponym and person name disambiguation as clearly differentiated and independent tasks. This is the approach preferred in this thesis, as it allows to exploit the inherent characteristics of each kind of entity: locations in the case of toponym disambiguation, and people in the case of person name disambiguation. I provide the motivation, formal definition, and relevant terminology for each of the two tasks in the following subsections.

1.1.1 Toponym disambiguation

The relation between toponyms and locations in the real world is many-to-many: it is often the case that several different locations share the same name, and it is not uncommon that a location is known by more than one possible name. This ambiguity is even more pronounced in historical documents, since toponyms may undergo changes over time, become obsolete or even refer to a location that no longer exists. With the following example, I illustrate the difficulty and importance of the toponym disambiguation task in historical texts:

(1) *In eight days after leaving* ***London*** *one can now be in the* ***Belgian Congo****, and the same applies to travellers from* ***Belgium*** *[...] Motor transport takes them to* ***Stanleyville*** *[...] Passengers who fly north from* ***Capetown*** *can change at* ***Broken Hill*** *to a feeder air service to* ***Elisabethsville****. Here there is a train link to* ***Port Francqui****, at which point connections are established with the Congo airways, which run from* ***Luluabourg*** *to* ***Leopoldville****.*[1]

In the example, toponyms have been marked in bold. In this text, 'London', 'Belgium', and 'Capetown' refer to the entities most commonly referred to by these names, namely, the capital of England, the country in Europe, and (with minor orthographic variation) the city in South Africa. However, there are other locations in the world known by these names, among which a city in Ontario, a village in Illinois, and a locality in California, respectively. At the time of publication of the news article, in 1933, the country currently known as the Democratic Republic of the Congo was a Belgian

[1] **Source:** *The Queenslander*, Brisbane, 9th February 1933.

colony, called the Belgian Congo. After independence, many of the locations mentioned in the text, then part of the Belgian Congo, changed their name. Stanleyville became Kisangani, Elisabethsville became Lubumbashi, Port Francqui became Ilebo, Luluabourg became Kananga, and Leopoldville became Kinshasa. Broken Hill, at the time part of Northern Rhodesia, was renamed Kabwe after Zambian independence.

A good toponym disambiguation system should be able to link toponym mentions to the locations they refer to, overcoming the problem of location renaming and toponym ambiguity. In the same way that a human is expected to (most of the times subconsciously) disambiguate toponyms in texts, a system should also be able to do so when given plain text containing mentions to geographic entities. It should be able to decide that the mention 'Stanleyville' in example 1 refers to the city today known as Kisangani, and not to the Stanleyville community in North Carolina, and that the mention 'Broken Hill' refers to the city today known as Kabwe, and not to the Australian city Broken Hill. Likewise, it should be able to identify that the mention 'London' in example 1 refers to the capital city of England, whereas the 'London' in example 2 refers to the city in Ontario.

(2) *All debts due to and by the firm of ***, in* ***Hamilton*** *and* ***London****, will be received and paid, and the business carried on as heretofore, at these places by the undersigned.*
**** & Co.* ***Toronto****, 31st March 1858.*[1]

1.1.1.1 Terminology

Before proceeding any further, I introduce the terminology associated to the toponym disambiguation task used in this thesis. Different researchers have used different terms for similar concepts and similar tasks. Their use in this thesis is clarified below.

Concepts. A **location** (also referred to as *place*) is an entity that occupies a static spatial area on the Earth's surface. A location might have more or less ambiguous boundaries, and its existence is often more conditioned to the human need of naming it rather than to purely geographic factors. A location is an entity that can be named, and the name to refer to a location is called a **toponym**, a Greek word which literally means 'place name'. Toponyms are often ambiguous: they can sometimes refer to just

[1] **Source:** *The Canada Gazette*, Ottawa, 3rd April 1858.

one location or to more than one location. The potential locations that can be referred to by a toponym are named **candidates** or **candidate locations**. A **gazetteer** is a dictionary of geographic names and must have, at least, two fields of information for each entry: the toponym (a natural language expression that is often ambiguous) and its unambiguous geographic reference (which allows the location to be mapped). The **geographic reference** (also referred to as *geographic footprint*) is the unambiguous representation of a location on the Earth's surface, either represented by points (the latitude and longitude of the centroid of the geographic area of the location) or by a group of points (a set of latitude and longitude coordinates forming a polygon which approximates the shape of the location).

Tasks. Very similar tasks are known by different names, and there is no general consensus among researchers on how the different terms should be used. Many approaches use the term **toponym resolution** to refer to the whole process, from the identification of a toponym in a text to the resolution to its geographic footprint, whereas some approaches use it only to name this last step. Some researchers consider this last step to consist of two separate tasks: disambiguation of the toponym and resolution of the disambiguated entity to its geographic coordinates. The term **entity linking**, on the other hand, refers to the task of linking named entities (among which, toponyms) to the entries (among which, locations) in a knowledge base that refer to them. In this thesis, I use the term **toponym disambiguation** to refer to the task of finding the most likely candidate to be the referent of a given toponym regardless of its output representation, **toponym resolution** to refer to the task of disambiguating a toponym by its geographic reference (e.g. latitude and longitude) and **entity linking** (and if only locations are specifically dealt with, **location linking**) to refer to the task of disambiguating a toponym by linking it to its entry in a knowledge base.

The toponym disambiguation task involves two subtasks: toponym identification and candidate selection. **Toponym identification**[1] is a form of named entity recognition that involves detecting expressions (also called **mentions**) that refer to toponyms. The input is plain text and the output is the set of detected toponyms. It can be

[1]Toponym identification is sometimes also known as geoparsing, geotagging, georecognition, place name identification, place name recognition, place name detection, toponym recognition, or toponym detection.

approached as a part of a **named entity recognition** problem (in which only the *location* class is considered) or as a problem in its own right. **Candidate selection** is the (often not explicit) task of selecting all the possible locations that can possibly refer to a toponym that has been identified in text.

1.1.1.2 Formal definition

The toponym disambiguation task can be formally defined thus: given a document d in which a set of toponyms $T = t_1, t_2, ..., t_n$ has been identified, a set of candidate referents $C_{t_i} = c_{1_{t_i}}, c_{2_{t_i}}, ..., c_{m_{t_i}}$ is found for each toponym. By means of a disambiguation function df, the best candidate is found from the set of candidate entities for each toponym. The task of toponym disambiguation consists of finding the set of entities $E = e_{t_1}, e_{t_2}, ..., e_{t_n}$ that are referred to by the toponyms.

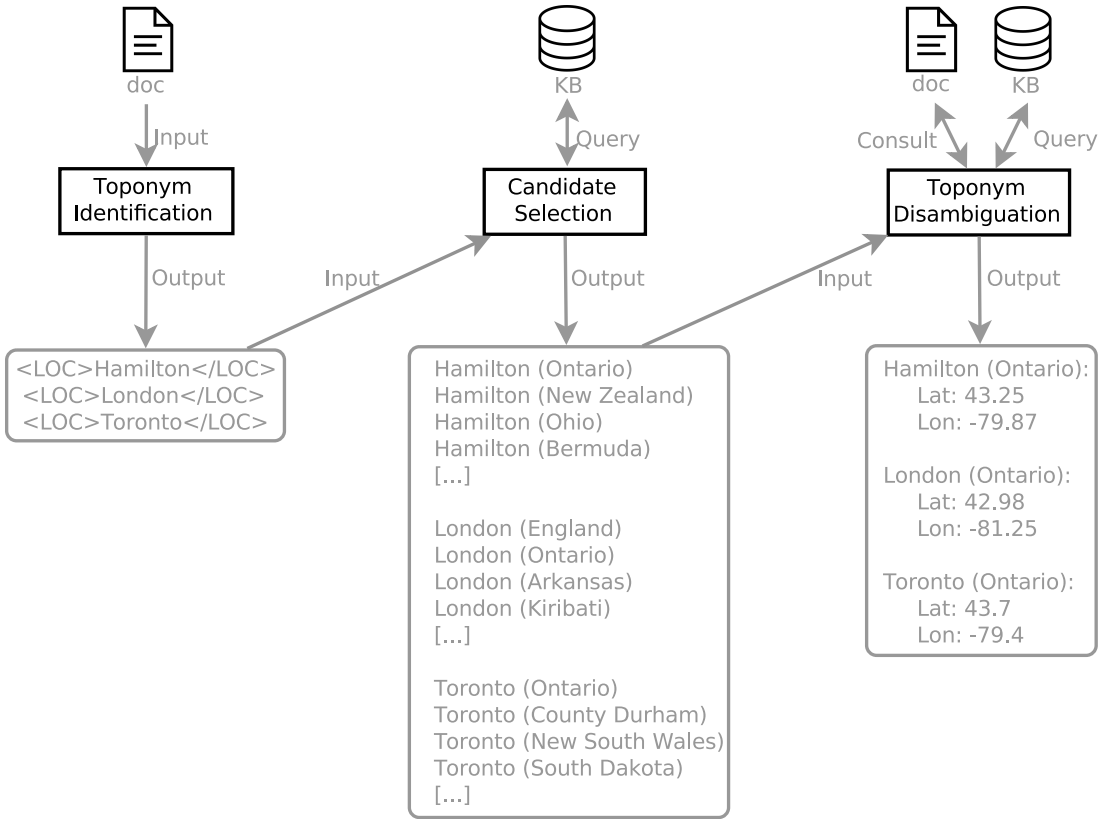


Figure 1.1: Toponym disambiguation workflow.

Figure 1.1 illustrates the steps involved in the task. The input is a plain text

document (i.e. text without any markup) and the final output is the set of toponyms identified in the text, disambiguated to the locations that are being referred to, expressed by means of unambiguous expressions (i.e. 'London, Ontario') or by means of their geographic coordinates (i.e. a latitude of 42.98 and a longitude of -81.25). There are three main steps involved in the process: toponym identification, candidate selection, and the actual toponym disambiguation. **Toponym identification** receives a document in plain text as input (e.g. the text from example 2) and returns the set of toponyms that have been identified (i.e. 'Hamilton', 'London', and 'Toronto'). The list of identified toponyms is then passed as input to the **candidate selection** component. In it, a set of candidate entities is found for each toponym, in accordance with the information present in the consulted knowledge base. Some examples of candidates for the toponym 'Hamilton' are Hamilton (Ontario), Hamilton (New Zealand), Hamilton (Ohio), and Hamilton (Bermuda). Ideally, this step should guarantee that the correct referent is among the selected candidates and that there is an acceptable number of noisy candidates. Once the set of candidates has been generated, the **toponym disambiguation** component is responsible for deciding (by consulting the context of a toponym and the information of the knowledge base) which is the most likely candidate to be the correct referent of the identified toponym in this particular document.

1.1.2 Person name disambiguation

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.[1] With the following examples,[2] I illustrate the importance and difficulty of the task:

(3) 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.

(4) **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

[1] According to the U.S. Census Bureau, only 90,000 different names are shared by up to 100 million people, as stated in Artiles et al. (2009) (12).

[2] **Source:** John Smith Corpus, introduced in Bagga and Baldwin (1998) (14).

and high-tax policies and its commitment to national ownership of industrial assets.

(5) 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.

In the examples, person names have been marked in bold. There is one name, 'John Smith', which occurs in the three texts but which corresponds to a different real-world person in each of them: the CEO of General Motors, the Labour Party leader, and an athletics coach. The goal of person name disambiguation is to find the correct referent for each person name, i.e. the person that is actually meant by the writer of the text. In the literature, there exist two main strategies to approach the task: as an entity linking task or as a cross-document coreference resolution task.

Entity linking resolves the different person name mentions by linking them to their respective entries in a knowledge base. Different kinds of knowledge bases have been used in the past, from encyclopedias like Wikipedia to specific databases created for a given collection. For instance, if the knowledge base is Wikipedia, the 'John Smith' mention of example 3 would be resolved to `http://en.wikipedia.org/wiki/John_F._Smith_Jr.`, the mention of example 4 to `http://en.wikipedia.org/wiki/John_Smith_(Labour_Party_leader)`, and the mention of example 5 to `http://en.wikipedia.org/wiki/John_Smith_(sprinter)`. Besides, the rest of person name mentions would also be linked to their corresponding entries in Wikipedia, if existing.

Cross-document coreference resolution takes a very different approach. Given a query consisting of a person name and given a collection of documents in which this name occurs (e.g. 'John Smith' in the three examples), the task of person name disambiguation is to group these documents according to the different real-world entities (i.e. persons) behind the identical person name. Given the collection where examples 3, 4, and 5 are taken from, a cross-document coreference resolution system would return one cluster for each different entity that answers to the name 'John Smith' in the collection, where each cluster would contain all the documents in which this particular person is being referred to.

Both approaches have clear advantages and disadvantages. Entity linking provides a faster and clearer retrieval of person entities, which are moreover linked to a knowledge base which straightforwardly informs about the entity in particular. However, it is a

classification task in which the potential target entities are limited to the ones present in the knowledge base.[1] This is a major problem when working with historical news articles that very often come from very regional collections, and whose mentioned people might not be recorded in most knowledge bases. For this reason mainly, the approach chosen in this thesis is to treat the problem as a cross-document coreference resolution task.

1.1.2.1 Terminology

Before proceeding any further, for clarity I explain the different terms associated to the person name disambiguation task that are used in this thesis, distinguishing between concepts and tasks.

Concepts. I describe here the concepts used throughout this chapter, illustrated with an example:

(6) 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. The person names in example 6 have been marked in bold. An **entity** (or **person**) is the real-world referent that is referred to by a person name. In the example, '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. I proceed on the largely held 'one sense per discourse' assumption, according to which all occurrences of the same person name within a document are considered to always refer to the same entity. A **mention name** is any person name that is mentioned in a document and that is not the query name (i.e. 'Alexie' in the example). I call a **full name** any person name with at least two tokens (ideally a first and last name, even though this is not necessarily always the case), whereas a **namepart** is each of the tokens that form a full name. In the

[1] Most recent approaches allow marking entities also as NIL if they are not present in the knowledge base. This in practice often means that all entities that are not present (or found) in the knowledge base are classified together, regardless of how different they are.

example, 'John Smith' is the only full name and 'John' and 'Smith' are its nameparts. Finally, by **non-person mention** I mean any named entity expression that does not refer to a person (i.e. 'Wellpinit' and 'Spokane Indian Reservation').

Tasks. **Person name disambiguation** is an umbrella term for the various very similar tasks that are sometimes known by different names. **Cross-document coreference resolution** aims at clustering documents in which the same person (i.e. entity) is mentioned, and often the number of target entities is unknown. **Person linking** (often just a subpart of **entity linking**) aims at linking mentions of person names in texts to their corresponding entries in a knowledge base, if they exist. Finally, if the domain of the documents to cluster or classify are web-pages, the task of discriminating among different person names is also known as **web people search**.

1.1.2.2 Formal definition

The task can be formally defined thus: given a query name qn and a set of documents in which it appears $d_1, d_2, ..., d_j$, person name disambiguation 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 entity $e_1, e_2, ..., e_k$ and each containing the documents in which the mentioned qn refers to the entity in question.

Figure 1.2 shows the main idea of person name disambiguation when treated as a clustering task. The input is a collection of documents mentioning a particular person name, for instance 'John Smith' in examples 3, 4, and 5. The method should ideally output as many clusters of documents as entities being referred to by the query name in the collection, where each cluster should correspond to a different person named with the query name (i.e. the CEO of General Motors, the Labour leader, and the coach, in the examples).

1.2 Research aims

This thesis is a contribution to the study and development of natural language processing technologies for the cultural heritage domain. In particular, its aim is to develop methods for toponym and person name disambiguation that are especially suited for

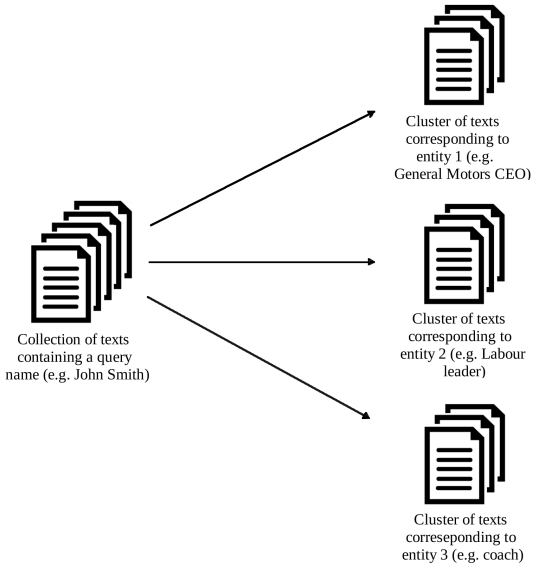


Figure 1.2: Person name disambiguation task.

historical newspaper collections. News articles have been a common source of data for both tasks, but they are usually drawn from current newspapers with an international scope. In the historical domain, regional newspapers are often as interesting to scholars as more generic ones. The disambiguation strategy applied to historical or regional newspapers must significantly differ from that applied to current international newspapers, as the knowledge shared by the expected readership of the collection is assumedly not commonly shared by external not-intended readers. This clearly hinders the disambiguation process, both automatically and human conducted.

In this thesis, I introduce two named entity disambiguation methods — one for toponyms and one for person names — that exploit the intrinsic characteristics of each kind of entity: the social context in the case of person entities and the geographic context in the case of location entities. Both methods are conceived to perform well in historical or regional collections. The toponym disambiguation method takes into consideration the fact that place names may change over time and that locations that existed or had a certain relevance in the past may not exist anymore or not exist in the same manner as were known at the time of writing. The person name disambiguation method is devised in such a way that the presence of unknown persons in the texts (i.e. people who do not appear in knowledge bases or even historical records) does not

have an impact on the performance.

1.3 Significance of the research

From the outset, newspapers and other forms of printed mass media have played a prominent role in the formation of public opinion. For many years, they were the main sources of information. Thus, they are not only records of the important events of a given time (or rather, of the events that were chosen to be reported by the newspaper editor) but they are also valuable indicators of public opinion and debate. In fact, due to their prominent role as an information source, newspapers played a significant role in transmitting opinions, ideologies and values, conditioned on the interests they served, thereby shaping public opinion and stimulating public debate. Even so, printed mass media have not yet enjoyed the popularity among historians that would be expected from such a precious resource.

There are probably two factors that explain why the use of newspapers for historical research has been disadvantaged until recently. First, historical newspapers have not always been easily accessible in the past, with different publications being dispersed and scattered in various libraries and archives. Fortunately, this is changing quickly with the mass digitization of newspapers, gazettes, magazines, pamphlets, and other kinds of materials which ideally puts these resources within easier reach for researchers from all over the world, making it likely that the practice of using them as sources in historical research becomes more common. A second reason why printed mass media have not yet featured prominently as sources of historical research is that — in a humanities context — newspaper archives constitute 'big data' and are therefore difficult to analyze with traditional humanities methods, i.e. close reading. Historians are often not so interested in the detailed content of individual articles but in more general trends of how topics are discussed over time and how this differs from newspaper to newspaper or from country to country. This is a distant reading application that requires dedicated text mining tools.

Entities are many times the starting points through which to explore and dig into historical newspaper collections. The digital humanities community has developed a plethora of typically general purpose text mining tools over the past years, but most of these only support relatively shallow analyses. Needless to say, these can already be

extremely valuable to researchers; a simple keyword search, for example, can already provide a historian with a sense of whether a collection contains material relevant for their research, thus saving many hours of visiting archives and skimming through pages. However, such approaches often fall short of fully supporting the specific needs of historical research. With the growth of digitization of historical documents, there is a pressing need to improve techniques that address information extraction from unstructured data, which make for most of the real-world data with which historians have to deal. High-quality entity mining, though, is at the moment difficult to achieve, in large part due to the high ambiguity which is often associated to named entities.

Resolving person names across documents is an open problem of unquestionable importance in natural language processing. According to Artiles et al. (2005) (13), person names represent 30% of the overall number of queries in the web domain. Person names have an equally significant presence in the news domain, in which people are often at the core of the events reported in articles. This is particularly interesting in historical research, as people are drivers and carriers of change, and newspapers have traditionally been the platform for someone to become a public figure. On the other hand, toponym disambiguation geographically grounds the different texts of a collection. Today, many digital-born data are already geo-localized. The usefulness of representing texts according to their geographic reference (especially in the historical domain) needs no elaboration. Finally, working with historical materials also provides some attractive challenges to the field of natural language processing, the most relevant in an entity-based research being that either entities may have changed over time or the knowledge we have of them may have altered. This forces alternative strategies and approaches to be found in order to deal with them, which results in a better understanding of the nature of entities and named entities.

1.4 Assumptions, scope, and limitations

In the design and development of natural language processing technologies, some assumptions are always necessary. They are many times implicit and presupposed, but they sometimes need to be specified. The main and most relevant assumptions used in developing the toponym and person name disambiguation methods of this thesis are

discussed in the next lines. Besides, I also specify the scope and limitations of the proposed methods.

1.4.1 Assumptions

One of the most common assumptions in disambiguation tasks presupposes one sense per discourse, which considers that a word (or an expression) has the same meaning across all its occurrences in a document. This thesis adopts the one-sense-per-discourse assumption as well.

Wikipedia is the encyclopedic knowledge base that is used as a source of knowledge in the proposed toponym disambiguation method but not in the person name disambiguation method, due to the assumption that Wikipedia has a sufficient geographic coverage to treat toponym disambiguation as a classification task, whereas it does not have sufficient coverage of people. In other words, it is assumed that all locations that are important enough to be mentioned in historical newspapers are likely to be present in Wikipedia, whereas the coverage of people is clearly insufficient for the person name disambiguation task to be treated as a classification problem.

Another assumption of the proposed methods is that treating toponym and person name disambiguation as separate tasks and not as one only task (as in entity linking approaches) allows exploiting specific features for each kind of named entity, and that this results on an improved overall performance.

Finally, two more assumptions were made in the person name disambiguation clustering method: first of all, that the more ambiguous a person name is, the more entities this name can possibly refer to; secondly, that the people mentioned in the same paragraph form each other's social context, and that this social circle around a target entity is a source of evidence for disambiguation.

1.4.2 Scope and limitations

Besides the assumptions made, the methods proposed have a well-defined scope, but also some limitations. These are reviewed in the following lines.

Toponyms A toponym is a mention to a location. Defining location is not easy. Whereas a person is a defined entity with clear delimited borders, this is not the case of a location. In general, it could be said that a location is any entity for which

stable coordinates can be extracted. According to this very general definition, houses, trees, mines, and wrecked ship's remains should all be considered locations. A more restrictive definition considers a location to be any entity for which stable coordinates can be extracted and that can be named (Kripke (1980) (56)). The naming condition is redundant in *fiat* entities such as cities or countries, which are a result of human delimitation (Kavouras and Kokla (2007) (54)), but is necessary in the rest of the cases. To illustrate this with an example, a tree would become a location the moment it is widely known and recognized by a name, as is the case of the Lone Cypress in Pebble Beach, California, in contrast with any other cypresses that can be found in a forest. Similarly, a marsh becomes a location the moment it is named and its extension delimited, such as the Fox Tor Mires in Dartmoor. From a more practical perspective, a location is traditionally defined in toponym resolution systems by its presence in a gazetteer. For example, Amitay et al. (2004) (3) used a gazetteer of only countries and cities of more than 5,000 inhabitants, and therefore the rest of geographical entities were not even considered. In their approach, an entity would be considered a location only if it had an entry in their gazetteer. On the other side of the spectrum, Matsuda et al. (2015) (63) used a gazetteer in which facility entities were also present, such as bus stops, restaurants, and schools. Similarly, in this thesis a location is also defined to be any entity with coordinates that is present in the resource that is used to extract world knowledge, which in this case is based on Wikipedia and complemented with information from the GeoNames gazetteer. This means that subdivisions of populated places (e.g. streets, neighborhoods, etc.) or urban facilities (e.g. bus stops, restaurants, etc.) are most of the times not considered locations. Finally, this thesis does not attempt to resolve compositional geographic mentions, such as '6 km south of Göttingen'.

Person names A person name is a name that refers to a human being. The datasets that I work with are collections of newspaper articles in English, Italian, and Dutch. In most societies, the combination of a given name and a last name (in whichever order) is the most common, but this convention is not universal. In Western societies, the first name usually precedes the last name; in Eastern societies this is the other way round. In Spanish, the custom is that every person has two family names (even though, depending on several factors, people might be known by just one of them or both), and

Javanese allows people to be known by a single token. Be what may, the person naming convention that is associated to a given society is often relatively flexible, thus leaving room for exceptions. In English, for example, most family names consist of one only token, but multiple-token family names are also possible, usually separated by a dash. Even if most family names in Italian only have one token, they sometimes include a particle (e.g. '*del*', '*della*', etc.) preceding the last name, and this is even more usual in Dutch. In these languages, person names are usually expressed as combinations of a first name, a last name (with or without a preceding particle) and, occasionally, one or more middle names. In this thesis, I restrict query names to names with at least two tokens (excluding name particles) and assume them to follow the naming conventions of the three working languages.

Languages The methods proposed in this thesis are largely language-independent, since they can easily be adapted to different languages. To demonstrate this, I tested the toponym disambiguation method on datasets in English, German, and Dutch, and the person name disambiguation method on datasets in English, Italian, and Dutch. The resources used to assist in the disambiguation of toponyms are language-dependent but can automatically be constructed by combining information extracted from Wikipedia (which has versions in multiple languages) and the GeoNames gazetteer. Likewise, the person name disambiguation method can be used in several different languages, given the condition that they follow similar naming conventions as those of the three aforementioned languages. The choice of languages was partly determined by the availability of benchmark data and partly by the demands of the Asymenc[1] research project, within which the current thesis was carried out.

Domain The disambiguation methods presented in this thesis were developed with the aim of performing well on historical newspaper collections. Person name disambiguation in recent years has moved towards the webpages domain, where additional features can be exploited that are not present in the news domain. The method introduced in this thesis for person name disambiguation would most likely perform worse in the webpages domain as it does not exploit any features other than those that can

[1]Humanities in the European Research Area (HERA) research project 'Asymmetrical Encounters: E-Humanity Approaches to Reference Cultures in Europe, 1815–1992': `https://asymenc.wp.hum.uu.nl/`.

be extracted from the unstructured textual context of a mention. Both methods have a general applicability in the news domain and are well suited to perform well in newspaper collections of regional, national, and international scope.

1.5 Outline of the thesis

This thesis is organized into five chapters. This first chapter has been an introduction to the topics covered throughout this thesis. It has provided the research background, defined the aims and significance of the research and has specified its assumptions, scope and limitations. Chapter 2 presents the related work. This thesis undertakes two main tasks: toponym disambiguation and person name disambiguation, which are covered in chapters 3 and 4 respectively. Each follows the same structure: it starts with an overview of the data, describes the resources built and used for the task, presents the disambiguation method, and evaluates the results quantitatively and qualitatively. Finally, chapter 5 concludes this thesis by summarizing its contributions and offering directions for future research.

Chapter 2

Related Work

Until recently, natural language processing has paid relatively little attention to the cultural heritage domain, but this is beginning to change. The first ACL LaTeCH[1] workshop was launched in 2007 as an initiative to fill this gap and has been taking place annually since then. In 2012, the ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities (SIGHUM) was constituted to promote the study of language technologies in the humanities. Over the last years, several new conferences and workshops devoted to this area have appeared. They address the many challenges that arise from working with data from the cultural heritage domain and explore the many possibilities it offers for fostering research in the humanities. This thesis is a contribution to two well-known problems in the field of natural language processing, those of toponym and person name disambiguation, and specifically focuses on the domain of historical texts.

Section 2.1 provides an overview of the literature on toponym disambiguation and section 2.2 reviews the related work of person name disambiguation. Section 2.3 includes an overview of different approaches proposed to access knowledge from historical collections through entity-centric strategies, and section 2.4 concludes this chapter by summarizing its key points.

[1]Language Technology for Cultural Heritage, Social Sciences, and Humanities, of the Association for Computational Linguistics.

2.1 Toponym disambiguation

Historians are often interested in the locations mentioned in digitized collections. However, the surface string forms which are used to refer to them (i.e. their toponyms) can be highly ambiguous and may have changed over time, which makes it especially hard to automatically ground mentions of places in historical texts to their real-world referents. As discussed in the introduction, the problem of mapping toponyms to locations has been approached from two different perspectives in the literature: from a geographical perspective (in which toponyms are grounded by means of their geographical coordinates) and from an entity linking perspective (in which toponyms are linked to the entries in a knowledge base that correspond to the entities that are referred to by them). The first approach corresponds to the **toponym resolution** task and is reviewed in subsection 2.1.1. The latter approach corresponds to the **entity linking** task, which is reviewed in subsection 2.1.2.

2.1.1 Toponym resolution

Buscaldi (2011) (24) groups toponym resolution approaches into three categories: map-based methods, knowledge-based methods, and data-driven or supervised methods. The disambiguation strategy of *map-based* methods relies mostly on geometric information such as the latitude and longitude of the possible candidates and the geographic distance between them. *Knowledge-based* methods exploit external knowledge for properties of the geographic entities. Finally, *data-driven or supervised* methods are based on machine learning techniques and often require hand-annotated training data.

Map-based methods. One method whose strategy relies heavily on the use of geometric features is Smith and Crane (2001) (89). They use rule-based strategies to identify toponyms in a text and then compute a geographic centroid from all the possible interpretations for each toponym, weighted by the number of occurrences. Candidate locations that are more than twice the standard deviation away from the centroid are discarded, and the final disambiguation is based on the distance between each location and the calculated centroid and the rest of toponyms in the text (of which only those unambiguous or already disambiguated are taken into account). A similar method based on Smith and Crane (2001) (89) is Buscaldi and Magnini (2010) (25),

who additionally consider information about the geographical source of the text. The authors found out that most of the toponyms in a local publication (76.2% according to Buscaldi (2011) (24)) are located within 400 kilometers from the place where the newspaper is published. The authors concluded that ambiguous toponyms are spatially autocorrelated.

Knowledge-based methods. Knowledge-based methods are the most common in the literature. Rauch et al. (2003) (82) use a commercial system, MetaCarta, to build a disambiguation method based on the prominence of locations. Disambiguation in their approach is a training process based on how often a toponym refers to a location, and this starting point is only overruled if strong evidence exists, such as a significant population difference between the candidates. Amitay et al. (2004) developed Web-a-where, a rule-based system that geocodes the content of web pages based on the position and cooccurrence of a location in a taxonomy (e.g. *Paris/France/Europe*). The system deals only with countries (and some states and provinces) and cities of more than 5,000 inhabitants. Other methods that rely on the hierarchical structure of locations are Bensalem and Kholladi (2010) (17), who exploit the proximity between toponyms in a hierarchical tree structure of locations, Buscaldi and Rosso (2008) (26), who make use of the conceptual density of the hierarchical paths of the toponyms' candidates from WordNet, and Volz et al. (2007) (96), who build an ontology based on the GNIS and GNS gazetteers and enriched with WordNet relations, and rank the candidates based on weights that are attached to the different classes of the ontology.

The toponym resolution strategy of Lieberman et al. (2010) (60) involves constructing local lexicons of toponyms that are likely to be known to the reader of a certain location. Local newspapers aim at a certain reduced audience for which certain knowledge is presupposed. The authors propose to automatically build lexicons of the locations that might be known to the readers of a certain audience. The authors exemplify it with the following example: Columbus, the capital of Ohio, has in its vicinity places named 'Dublin', 'Africa', 'Alexandria', 'Bremen', 'London', 'Washington', etc. For a reader from Columbus, the first referent of these toponyms might not be their most prominent sense but the places neighboring their city. The authors automatically create lexicons of places with their most probable referent given a certain audience which must comply with the following three characteristics: the local lexicon must be constant

across articles from the news source, the toponyms in it must be close to each other, and the lexicon must contain a sensible number of toponyms, not too few and not too many. The authors also compute a global lexicon of locations that a general audience is likely to know. The toponym resolution is then achieved by combining a number of heuristic rules.

Data-driven methods. Data-driven methods enjoyed little popularity at the beginning due to the lack of annotated data and to the problem of unseen classes. Nevertheless, recent years have witnessed an increase in these kinds of methods. Most approaches that have looked at the document context in order to find disambiguation cues have restricted the textual context of the document to the set of nearby toponyms. It is the case of Overell (2009) (72), who extracts training instances from Wikipedia, using only toponyms as features. In recent years, some methods have appeared that also use non-geographic context words for disambiguation. Roberts et al. (2010) (84) propose a probabilistic model that uses the spatial relationships between locations (collected from the GeoNames gazetteer) and the people and organizations related to these locations (extracted from Wikipedia). The presence of non-geographic entities adds disambiguation power to the task, the authors argue. Qin et al. (2010) (78) represent the different location candidates in a hierarchy tree mined from the Web and base their disambiguation strategy on a score propagation algorithm.

Speriosu and Baldridge (2013) (91) propose a supervised learning method by means of indirect supervision in which non-toponym words are also used for textual context. The authors see the task as a traditional classification task and train models from geotagged Wikipedia articles. A text classifier is learned for each toponym and is used to disambiguate toponyms both in current news articles and historical texts from the American Civil War, in both cases with good results. More recently, DeLozier et al. (2015) (35) developed a method that does not rely on knowledge from gazetteers for the disambiguation of toponyms. Instead, they model the geographic distributions of words over the surface of the Earth: for each word, a geographic profile is computed based on local spatial statistics over a set of language models annotated with coordinates (the authors use GeoWiki, a subset of articles from Wikipedia that contain latitude and longitude). The disambiguator returns the most likely set of coordinates

and the closest referent in a gazetteer. This method significantly outperforms other state-of-the-art toponym resolvers.

2.1.2 Entity linking

In recent years, the task of toponym resolution has somewhat been absorbed by the more general entity linking. Specific work on disambiguation of toponyms and their resolution to geographical reference has become less common in favor of methods that are able to disambiguate different kinds of named entities jointly and link them to entries in a knowledge base. Aware of the increase in popularity of entity linking methods, DeLozier et al. (2015) warn that such approaches are not created specifically to disambiguate toponyms, and therefore do not exploit any of the geographically-specific cues and properties of locations. Having to rely in other forms of knowledge that are not geographically-specific, entity linking approaches exploit features and strategies that are not often explored in toponym resolution.

Han et al. (2011) (50) summarizes entity linking approaches by classifying them into three categories, revised here in order to include the most recent approaches: local compatibility based approaches, simple relational approaches, and collective approaches.

Local compatibility based approaches. First approaches to entity linking emerged together with the rise of popularity of Wikipedia, the free online encyclopedia, which soon became one of the most utilized sources of world knowledge in the field of natural language processing due to its wide coverage, high quality, free availability, and partially-linked structure. First approaches did not take into account interdependence between different mentioned entities: decisions were taken based on the similarity between the textual context of a mention in the document and that of the different Wikipedia entries that constitute the set of candidate entities to be the correct referent. Bunescu and Paşca (2006) (22) train support vector machines to rank the different candidates. For each candidate, two vectors are computed: one based on the cosine similarity between the document context and the text of the Wikipedia article, and one based on word-category correlations. Several other approaches, similar in nature, are Cucerzan (2007) (31), Mihalcea and Csomai (2007) (66), and Fader et al. (2009) (41), who also use an extended Bag-of-Words model. Other more recent prominent methods that enter in this category are Gottipati and Jiang (2011) (47), who link entities to

Wikipedia articles through query expansion, Zhang et al. (2011) (105), who explore acronym expansion and topic modeling, and Dredze et al. (2010) (37), whose system ranks the candidates and learns when not to link if confidence is low.

Simple relational approaches. The problem of local compatibility based approaches is that the interdependence between entities is not taken into account. Medelyan et al. (2008) (64) and Milne and Witten (2008) (64) consider that unambiguous named entities can provide information that may be useful to disambiguate other named entities that have more than one candidate. With the idea that entities mentioned in texts are often interdependent, they rank entities according to their compatibility with entities referring to unambiguous mentions. The main drawback of these kinds of approaches is that, even if some relatedness between the different entities in a text is considered, this is severely limited by the required presence of unambiguous named entities, thus leaving unexploited the information that other entities could also be carrying.

Collective approaches. In an attempt to overcome the above-described disadvantages, Kulkarni et al. (2009) (57) define a score based on the pairwise interdependence between entities. Han et al. (2011) propose a graph-based approach to model global interdependence and use it for disambiguation, where compatibility between the mention and the candidate entities is weighted by context similarity and the relatedness between the different entity nodes is weighted by their coherence. Also graph-based is Hoffart et al. (2011) (51), who propose an interconnected model that combines local and global features and reaches the optimal disambiguation by discarding the least connected candidate in each iteration. Similar approaches are Barrena et al. (2015) (16), who combine the local context of the mention and the Wikipedia hyperlink structure to provide a global coherence measure for the document mentions, Moro et al. (2014) (69), whose disambiguation is not only limited to named entities but also to concepts, and who use a greedy densest subgraph algorithm that selects semantic interpretations with high coherence, and Weissenborn et al. (2015) (98), who similarly to Moro et al. (2014) (69) also base their coherence model on sets of related concepts or named entities that they call semantic signatures.

2.2 Person name disambiguation

Resolving person names across documents is an open problem of considerable importance in natural language processing. As outlined in the introduction, person name disambiguation can be treated either as a classification task in which the possible referents (i.e. person entities a person name can refer to) are known or as a clustering task in which the possible referents are unknown. Entity linking falls into the first category, whereas the approach adopted in this thesis falls into the latter. Therefore, I do not delve into the details of entity linking approaches here, as their perspective of the task is completely different. Person name disambiguation, when envisioned as a clustering task, has the aim of, given a collection of documents, grouping together mentions of the same person entities occurring in them. Unlike traditional coreference resolution, this task does not usually attempt to resolve definite noun phrases and pronouns. Person name disambiguation is very closely related to word sense disambiguation, from which it differs greatly in one key aspect: contrary to word senses, the set of entities referred to by a person name is *a priori* unknown, as there is no available list, census, dictionary, or knowledge base, of all the people in the world.

One of the most relevant earliest works on disambiguating person names is Bagga and Baldwin (1998) (14). The authors provided the first reference set for cross-document coreference resolution (henceforth CDCR), which they modeled as a document clustering problem in which each cluster should ideally correspond to a different entity. To solve the problem, the authors applied the standard vector space model based on cosine similarity. They also proposed a scoring algorithm, named B-Cubed, which assesses a system's accuracy on a per-document basis, on which it builds a global score. Their approach was adapted and extended in several subsequent studies, such as Ravin and Kazi (1999) (83) and Gooi and Allan (2004) (46). The latter used agglomerative clustering as they consider it to perform particularly well for this task. The clustering is made on context vectors of a windows size of 55 words. More recently, Latent Dirichlet Allocation (LDA) and other topic models have been used in order to learn topics per document containing a person name to be disambiguated. This is the case of Song et al. (2007) (90), who learn the distribution of topics with regard to persons and words, and joins documents by means of a hierarchical agglomerative clustering

algorithm, and Kozareva and Ravi (2011) (55), who use LDA to generate the context of each query name.

In 2007, an evaluation campaign, the Web People Search task (WePS), was organized to tackle the problem of name ambiguity on the World Wide Web, and the first large corpus, WePS-1 (Artiles et al. (2007) (11)), was created. The corpus consists of several query names. Each of the query names comes with several web pages, and in each web page there is at least one mention of the query name. This query name can refer always to the same person entity or to different ones throughout the webpages that form the collection. The campaign went through two more editions, and two more corpora were created, each larger than the previous one. Therefore, although the person name disambiguation task started in the news domain, its interest largely moved to the web domain as a consequence of this campaign. Even though often more heterogeneous in form, web pages tend to be more structured than news articles, as these usually consist of plain text. Web pages also provide additional features that, if they are correctly exploited, can be of great assistance to the disambiguation task, such as urls, e-mail addresses, phone numbers, etc. Therefore, methods that perform well in one domain do not necessarily perform well in the other.

Yoshida et al. (2010) (102) distinguishes 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, temporal expressions, and, in the case of webpages, urls, telephones, institutions, etc. Some of the approaches that use biographical facts to assist the disambiguation are Mann and Yarowsky (2003) (62), who use a rich feature space of biographic facts that are obtained by bootstrapping extraction patterns, and Niu et al. (2004) (71), who extract local named entity phrases that identify the correct referent and use them to assist in the disambiguation. Al-Kamha and Embley (2004) (2) use features such as emails, zip codes, and addresses, and Bollegala et al. (2006) (21) use automatically extracted key phrases from different clusters of documents containing the same person name to identify different entities behind them.

In both the news and web domain, one of the most exploited sources of evidence for clustering has been named entities. Some of the many relevant works that have used them in the task are Blume (2005) (20), Chen and Martin (2007) (29), Popescu and Magnini (2007) (77), and Kalashnikov et al. (2007). Artiles et al. (2009) (10) thoroughly study the role of named entities in the person name disambiguation task.

Even though the authors conclude that named entities (and any other kind of strong feature) often increase the precision at the expense of recall, they leave the door open to more sophisticated approaches using named entities, such as in combination with other levels of features, as in Yoshida et al. (2010) (102) or in graph-based approaches. Along these lines, Kalashnikov et al. (2008) work on improving on the quality of most disambiguation approaches by collecting external knowledge of co-occurrences extracted from the Web. The collected information is then used to assist in making clustering decisions. Jiang et al. (2009) (53) use a graph-based approach that use different web-based features to cluster web pages that talk about the same entity. Chen et al. (2012) (28) create a semantic graph for each ambiguous name from all the Wikipedia concepts for the entities that share this person name. This semantic graph captures the topic structure, and the clustering is performed depending on the weight of the learned topics.

In 2011, Bentivogli et al. (2013) (19) proposed an evaluation campaign similar to WePS, the News People Search task (NePS), which returned the focus to the news domain. The aim of this campaign was to evaluate cross-document resolution of person entities in the news domain in a different language than English, in this case Italian. Bentivogli et al. (2008) (18) introduced a large dataset in Italian, the Cross-document Italian People Coreference corpus (CRIPCO). Until then, very few approaches (among which, Popescu (2009) (76)) had warned of the importance of determining the ambiguity of a person name to improve the performance of the disambiguation system. In the creation of the CRIPCO corpus, the authors made sure that names with different degrees of ambiguity were included in the datasets.

The most popular clustering method for person name disambiguation has been pairwise clustering with a fixed similarity threshold. In it, two documents are grouped together if their similarity is higher than a certain similarity threshold, which is fixed. However, 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. In Zanoli et al. (2013), a dynamic threshold similarity is introduced by estimating the ambiguity of each query name. The authors calculated person name ambiguity from a specifically Italian resource, the phone book *Pagine Bianche*. This method, which follows an entity linking approach as person name mentions are linked to a knowledge base that contains person descriptions, is one of the very few existing approaches in which ambiguity of the person name plays a role.

Over the last years, the trend has moved towards using resource-based approaches. The task of person name disambiguation has been in most cases subsumed by the more general entity linking, which does not focus only on person entities, but on other kinds of entities as well. Entity linking (or person linking, if the focus is on persons only) can be regarded as a classification task in which mentions to person names are linked to entries in a knowledge base (and sometimes to the empty class, if they do not have any corresponding entry in it). Bunescu and Paşca (2006) (22) and Cucerzan (2007) (31) are among the earliest and most relevant systems to have exploited the wide coverage of Wikipedia by linking entity mentions to the referring Wikipedia articles. Even though Wikipedia has been the most widely used resource in entity linking methods, other options are possible and have been used, such as knowledge bases like DBpedia (58) and Yago (93). Zanoli et al. (2013) (103) and Dutta and Weikum (2015) (38) are two examples of methods that perform person name disambiguation by linking person name mentions to entries in a knowledge base.

2.3 Entity-centric text mining in the digital humanities

One of the goals of this thesis is to improve on techniques and methods from natural language processing that can strengthen historical research. In the growing field of digital humanities, many voices have expressed a fear of a decline in the role of interpretative close reading and traditional historical research. Through this thesis, I want to show how using information extraction and text mining methods in historical data can be positive for historical research, allowing new and more refined research and to further the scope of inquiry, revealing new perspectives that were not possible until recently.

In this section, I review some of the recent entity-centric literature in the digital humanities field. In subsection 2.3.1, I provide an overview of some approaches that have a geographic perspective and, in subsection 2.3.2, of some approaches with a person-centric perspective. Finally, in subsection 2.3.3, I review some studies that approach the historical collections from a combined person- and location-centric perspective.

2.3.1 Location-centric approaches

Cultures cannot be understood independently of the temporary and spatial framework in which they originated, developed, and blossomed. Culture, understood as the collection of symbols, language, traditions, art and music, values, beliefs, and norms, arise and live in a particular point of time and in a particular place. The research conducted in this thesis was framed in the AsymEnc project, which aimed at tracing the changing influence between cultures in contact through time and space, focusing on the period between 1815 and 1992. The relation between places and cultures is very strong, and it is therefore crucial that historians are able to perform sophisticated queries of locations which go beyond the surface form.

A disambiguated toponym is, most of the times, a toponym for which geographical coordinates can be derived, and which can therefore be placed in a map. In the context of digital humanities, applications are plentiful and indisputable: by geotagging newspaper articles, it is possible to obtain a map of the world as known by the intended readers of the collection; by geotagging military reports, it is possible to trace the hot spots or places of interest during a certain conflict; by geotagging personal correspondence, it is possible to track personal itineraries. In short, by visualizing a collection of historical documents as a map, the historian is given the possibility to explore it from a geographic perspective, at different granularities, going from the overview to the detail.

With the rise of digitization, georeferencing toponyms is becoming more and more common. The number of georeferenced digital libraries is large and growing. The Perseus Project,[1] which originated in the late 1980s, was a major undertaking with the aim to bring the ancient Greek world closer to students and learners, by representing the most relevant Greek works in a dynamic and visual manner. Eventually, the project expanded and today it consists not only of data from the ancient world, but also of data from early modern English literature or from 19th-century history of the United States. Among other functionalities, the Perseus Project has georeferenced over one million toponyms, automatically recognized and resolved by means of simple heuristic rules, a percentage of which were corrected by hand (Smith and Crane (2001) (89)). In the digital humanities, many projects and tools have appeared in the following years that are similar to Perseus, but most of them do not attempt to perform automatic

[1] `http://www.perseus.tufts.edu/hopper/`.

disambiguation of toponyms or, if they do, use very basic methodologies based on proximity, population, or simple heuristics.

2.3.2 Person-centric approaches

To date, not much attention has been paid to person name disambiguation in the historical domain. However, its need is strongly felt in the many attempts to perform person-centric text mining, which is a very common and often necessary approach in historical research. One of the most popular strategies to explore historical collections from a person-centric perspective has been by means of social networks. A social network is a structure that captures the relationships between actors, and several studies have exploited this advantage to answer their research questions. However, they mostly concern pre-modern history, where the source material is much more limited and thus networks can still be constructed by hand. There are very few approaches that automatically create social networks from newspaper collections: most of the work in historical research that relies on social networks creates them either manually or from structured data, thus avoiding one of the greatest challenges in network creation, that of person name disambiguation. The potential of automatic approaches and the importance of performing person name disambiguation is discussed in Stratford and Browne (2015) (92) in the context of analyzing connections and communities in ancient Assyria, and one of the first and few fully-automatic approaches to social network construction for historical research is described in Coll Ardanuy et al. (2015) (8), who highlight the utility of automatically created social networks from newspaper archives in a case study on European integration and provide a very basic strategy to disambiguate named entities.

The utility of social networks for historical research has been repeatedly demonstrated by several studies. Padgett and Ansell (1993) (73) investigated in detail the action and rise of the Medici during the period between 1400 and 1434 by means of social networks; Jackson (2014) (52) used social networks of the medieval Scottish elites from the period between 1093 and 1286 to find hidden relationships; Rochat et al. (2014) (85) focused on the Venetian maritime empire from the end of the 13th century to 1453, and networks of ports and places were used to model maritime connections. More recently, social networks have been used by many scholars to represent historical data from very diverse sources and origins, in order to find relationships that may assist them in their research questions. To mention just two examples, the work

by Zhitomirsky-Geffet et al. (2016) (106) aims at revealing relationships between the different Jewish sages across generations, and Grandjean (2016) (48) connects the different intellectuals from 1919 to 1927 in order to understand how scientific elites were connected and to gain insights about their relations with the rest of the scientific and diplomatic world. Also Moretti et al. (2016) (68) advocate for the use of social networks to explore and visualize modern and contemporary text collections, using as a case study Nixon and Kennedy's speeches of the 1960 presidential campaign, without delving into the problem of person name disambiguation.

Automatically-constructed networks have been used in other humanities areas. In particular, person-centric analyses have gained popularity in quantitative literary analysis, as in Elson et al. (2010) (40), Bamman et al. (2013) (15), and Coll Ardanuy and Sporleder (2014) (5). In them, novels are represented as social networks of characters that typify the skeleton representation of the plot. The self-containing nature of literary works makes for the biggest difference between fiction and real-world data. When a novel ends, its characters cease to exist. When working with data from the news domain, we are not in a microcosmos anymore, and therefore networks are necessarily more spread out, and nodes more disseminated. Works of fiction, therefore, arguably pose less of a challenge for this task than newspaper collections, as the number of distinct entities and thus the expected amount of ambiguity is significantly lower.

The person name disambiguation method that is introduced in chapter 4 was conceived to assist in the creation of social networks from modern and contemporary historical news data. The texts that constitute this domain are often populated by many people that are absent from historical accounts and, therefore, also from knowledge bases or other sources of knowledge. Ter Braake and Fokkens (2015) (94) discuss the problem of biases in historiography and the importance of rescuing long-neglected individuals from the oblivion of history. By refraining from linking entities to a knowledge base (and, therefore, by treating the task as a clustering problem instead of a classification problem), I seek to avoid, to the extent possible, the bias towards favoring entities which are present in it.

2.3.3 Combined approaches

Querying and exploring collections through disambiguated person names and toponyms can be of great assistance to historians. In recent years, some projects have been carried

out in the digital humanities that illustrate the usefulness of exploring collections both through the biographical or social lens and through the geographical lens combined (and obviously subordinated to the temporal dimension). In the field of prosopography (i.e. the collective study of individual biographies), this combined information enables tracing different individuals through time and space, as reported in Buning (2016) (23) and Braake et al. (2014) (95). It is also crucial to be able to discern between different person names and toponyms in the task of historical event extraction. One example is Cybulska and Vossen (2011) (32), who explore historical event extraction from Dutch news articles through a case study, the Srebrenica Massacre in July 1995. Finally, several new digital humanities projects also advocate for the joint use of networks of people and maps to ease exploration of historical documents and assist scholars, as is the case of Škvrňák and Mertel (2016) (97) and Ferreira Lopes et al. (2016) (61), to mention just two.

2.4 Summary

In this chapter, I have provided the related work for the two main parts of which this thesis consists, toponym disambiguation and person name disambiguation. To summarize, toponym disambiguation has been approached from two different perspectives, entity linking and toponym resolution. Both are typically considered as classification tasks in which the potential referents are known. Person name disambiguation has also been approached from two different perspectives: entity linking and cross-document coreference resolution. The first is essentially a classification task, whereas the latter has usually been approached as a clustering task. Entity linking approaches usually do not discriminate between different types of entities: they are usually systems that aim at resolving both person and place names (as well as other kinds of entities) simultaneously. Whereas this has the advantage that it is more general-purpose, its performance may be affected by their not exploiting features that are specific and characteristic of the kind of entity in question. This chapter ends with an overview of the importance of place and person disambiguation in the digital humanities.

Chapter 3

Toponym disambiguation

This chapter introduces a novel method particularly suited for performing toponym disambiguation in historical documents. The proposed method combines the strengths of toponym resolution and entity linking approaches by exploiting both geographic and semantic features. Section 3.1 describes the six datasets that are used to assist in the development of the method and in its evaluation. Section 3.2 gives details of the creation of a new resource that combines geographic and encyclopedic knowledge extracted from GeoNames and Wikipedia respectively, and which is used to select a pool of referent candidates and information about them for each of the toponyms found in a document. The core of the method is presented in section 3.3, and section 3.4 describes the evaluation metrics used, provides the evaluation of the proposed method on the six datasets, and discusses the results.

3.1 Data

This section begins by reviewing existing corpora of historical documents that have been used in the literature to evaluate toponym resolution systems, and proceeds by explaining the process of annotating five new datasets from scratch. I describe how the sampling of documents is performed and outline the annotation schema and guidelines, illustrated by means of some examples. This section concludes by summarizing the different datasets that are used in this thesis in order to develop and evaluate a new toponym disambiguation method.

3.1.1 Brief review of toponym disambiguation historical corpora

Very few corpora exist for developing and evaluating toponym disambiguation in historical texts. In fact, until recently, very few corpora existed for developing and evaluating toponym disambiguation texts at all. Leidner (2008) (59) introduced a dataset with the explicit aim of countering the lack of standard benchmark datasets for the task. The created corpus (**TR-CoNLL**) consists of a subset of 946 documents (REUTERS news articles published in 1996) from the CoNLL 2003 shared task in which all toponyms had been manually identified. Geographic reference of the toponyms was resolved by hand. Speriosu and Baldridge (2013) (91), though, identified several errors in the dataset and corrected some systematic ones. A new version of the annotations was released, but the fact that some idiosyncratic errors still remain, that the corpus is not freely available, and that it can be argued whether it can be considered a historical corpus (is 20 years enough to consider a corpus of news articles historical?) did not make it the best candidate to develop and test the method introduced in this thesis.

Most of the existing historical corpora with geographic annotations were, until last year, either very small or had serious flaws, making them unsuitable (or at least unfavorable) to be used as benchmark datasets against which to compare the performance between the different methods. The most notable example of this is the *Perseus Civil War and 19th Century American Collection* (**CWAR**), from the Perseus Digital Library project[1] (Smith and Crane (2001) (89)). It consists of 341 volumes of biographies, memoirs, and records from the years around the American Civil War. In them, toponyms were automatically identified with a named entity recognizer and coordinates were assigned using simple rules and then post-corrected by hand. However, Speriosu and Baldridge (2013) (91) and DeLozier et al. (2015) (35) detected several issues with the toponym annotations (for instance, states and countries were not identified as place names), making the corpus inappropriate for evaluation.

In 2016, a new corpus was created with the ambition of becoming a benchmark dataset for toponym disambiguation from historical texts. The War of the Rebellion corpus (henceforth referred to as **WOTR**) was introduced in DeLozier et al. (2016), who aimed at creating a large historical corpus for the task of toponym resolution that would overcome the inconsistencies of the corpora available to date. WOTR consists

[1] `http://www.perseus.tufts.edu/hopper/`.

of a selection of documents from the volumes of the *Official Records of the War of the Rebellion*,[1] considered the largest collection of primary sources (mostly military reports and letters) of the American Civil War (1861–1865). A total of 1,644 documents were selected and annotated with two kinds of geographic reference: points (latitude and longitude coordinates) and polygons (sets of latitude and longitude coordinates). The annotators were asked to annotate mentions to locations by either entering their defining pairs of coordinates into a box in a GUI annotation tool, by drawing points (for locations with smaller scope) or polygons (to capture the scope of larger locations, such as states, countries, or bodies of water) directly on a map, or by retrieving the geographic reference from the list of recently annotated locations. However, even if some of the locations were originally annotated with polygons, the authors also provide a simplification of the polygonal shapes to centroids, expressed in latitude and longitude coordinates. The annotators were encouraged to consult various sources in order to obtain the coordinates of the different toponyms, and the most used among them was Wikipedia, even though also other resources such as Google Maps or different websites about the American Civil War were consulted.

3.1.2 Corpus creation and annotation

As mentioned in chapter 2, place name disambiguation can be regarded as a toponym resolution task, but also as an entity linking task in which the only type of entity considered is of the location type. The greatest difference between both manners to approach the task is in the expected output. Whereas toponym resolution aims at returning the geographic reference itself (in the form of points or polygons of latitudes and longitudes), entity linking aims at returning the link to the entry that refers to this location in a database. Advantages and disadvantages exist for both approaches. Evaluation is clearer in entity linking approaches, since the predicted output should coincide perfectly with the reference dataset in order to be considered correct. In toponym resolution, evaluation can be a bit confused, as there is no official standard dictionary of coordinates. A city like Los Angeles, for example, is so extensive that coordinates describing points that are tens of kilometers apart can still be describing the correct position of the city on the Earth's surface. This problem obviously increases in the case of countries, rivers, seas, continents, or oceans. The use of polygons to

[1] `http://ebooks.library.cornell.edu/m/moa//`.

shape larger locations is not devoid of problems either, as it is highly unlikely that the gold standard polygon and the predicted polygon match perfectly. Should therefore any point inside the polygon be considered correct, or only the centroid? As already said, a clear evaluation is one of the advantages of treating the task from an entity linking perspective. However, this has drawbacks as well. In toponym resolution, the geographical reference can be obtained from different resources, therefore maximizing the possibility that a correct geographic reference is found. Entity linking approaches, on the other hand, rely on only one resource to extract coordinates from, namely the resource to which toponyms are linked to location entries, and this can have a negative impact on coverage.

In this section, I describe the creation and annotation of five datasets in two different languages, German and Dutch, to be used in the development and evaluation of toponym disambiguation in historical texts. I opted to treat the task from an entity linking perspective and to link toponyms to the Wikipedia URLs that correspond to their referents. What motivated this decision is the fact that it is relatively straightforward to obtain the geographic reference from the Wikipedia URL, but not viceversa. Therefore, in treating this task as an entity linking task, I was making sure that it could be compared against both entity linking systems and toponym resolution systems. As for the choice of Wikipedia as the resource from which to extract the coordinates, this is motivated by its being the most standard resource used in entity linking tasks as well as its being the most complete freely available multilingual encyclopedic database that exists at the moment. Therefore, annotators were asked to identify all locations mentioned in a text and map them to the URL of the Wikipedia article that refers to them in the language version of the same language in which the text is written. A mention to the French capital, for example, should be mapped to `http://de.wikipedia.org/wiki/Paris` if the text is in German, and to `http://nl.wikipedia.org/wiki/Parijs` if the text is written in Dutch.

3.1.2.1 Corpus sampling

This PhD thesis is framed within the Asymetrical Encounters (AsymEnc) project,[1] which had access to a series of digitized collections of historical newspapers. At the

[1] `http://asymenc.wp.hum.uu.nl/`.

time of creation of the corpora, the available collections were in two languages: Dutch and German.

In **Dutch**, the data came from the Koninklijke Bibliotheek (KB), the National Library in the Netherlands. In 2006, the KB began the Database of Digital Daily newspapers project,[1] which has digitized over one million newspaper pages from national, regional, local, and colonial newspapers from 1618 to 1940. Our selection consisted of OCR processed news articles from three different kinds of newspapers: colonial newspapers with origin in the then Dutch East Indies, colonial newspapers with origin in the Netherlands Antilles,[2] and regional Dutch newspapers. This choice was taken with the consideration that the more regional a collection is, the farther it falls from general world knowledge. Whereas an international news agency such as REUTERS aims at reaching readers from the different corners of the world (and therefore sticks to general world knowledge), regional newspapers are interesting because they aim at an audience that shares certain knowledge about a specific region that others do not. Besides aiming at this specific audience, regional newspapers often have international news sections. Colonial newspapers are particularly interesting cases. These newspapers aimed to reach the Dutch audience that was living in the overseas colonies. They are therefore witnesses of an anomalous (albeit common those days) situation, in which a certain vision of the world coexisted with another one, thousands of kilometers away. These collections are also interesting because they describe a reality that does not exist as such anymore. In 1945, from the Dutch East Indies the Republic of Indonesia was born. Eventually, the islands that were part of the Netherlands Antilles also gained sovereignty. These political changes obviously produced many changes in different levels, one of them being the renaming of many place names.

In **German**, the data came from two distinct origins: Berlin during the years it was the capital of Prussia and St. Vith, a small-sized municipality in the German-speaking region of Belgium. The Prussian collection consists of 2,428 issues from two different newspapers: the *Provinzial-Correspondenz* (issues ranging from 1863 to 1884)

[1] `https://www.kb.nl/en/organisation/research-expertise/digitization-projects-in-the-kb/database-of-digital-daily-newspapers`.

[2] Today five islands with their own government, the Netherlands Antilles were known by this name until 2010. They underwent different changes of status during the 20th Century and were also known as Curaçao and Dependencies and the Territory of Curaçao. For simplification, from now on I will refer to them as the Netherlands Antilles.

and the *Neueste Mittheilungen* (issues ranging from 1882 to 1894). Both have been OCR corrected. The Belgian collection consists of 1248 issues from the *St. Vither Volkszeitung* ranging from 1955 to 1964. This collection's OCR is generally good, but some texts are not well divided due to bad layout recognition. The *St. Vither Volkszeitung* is a very good example of a newspaper that combines very regional news with colonial and international news. The Prussian collection is interesting in that it consists of two national newspapers from a state that does not exist as such anymore. The locations mentioned in these articles that were then part of Prussia are now part of several different countries.

We randomly selected files from the five different sources (Prussia, Belgium, regional Netherlands, the Dutch East Indies and the Netherlands Antilles). Each file was in the form of plain text with no other metadata information than that of the origin of the collection, which was kept as part of the file name. This information was of great assistance to the annotators.

3.1.2.2 Annotation schema

The tool used to annotate the selected files was WebAnno 2.0, a general purpose web-based system described in Yimam et al. (2013) (101) and Yimam et al. (2014) (100) that can be used for a wide range of annotation tasks. Each annotator was given access to the platform and could select which file to annotate, as shown in figure 3.1. Once selected, the annotators could see the selected document, sentence tokenized and ready to be annotated (figure 3.2). Just by selecting the text spanning a toponym (for example, 'Laos'), a window would pop up as in figure 3.3 and then the annotators only needed to fill the `Identifier` box with the Wikipedia URL of the article that refers to the mention in the language in which the text is written, `http://de.wikipedia.org/wiki/Laos`, in this case.

3.1.2.3 Annotation decisions

A toponym is defined here to be any named entity that can be defined according to the pair of static world coordinates of its referent. Parts of populated places (such as buildings, streets, or neighborhoods) are excluded from this definition, and so are adjectival and demonymic forms of place names (such as '**Parisian** university' or '**French** wine-growing regions'). Like most other toponym resolution datasets, I also consider

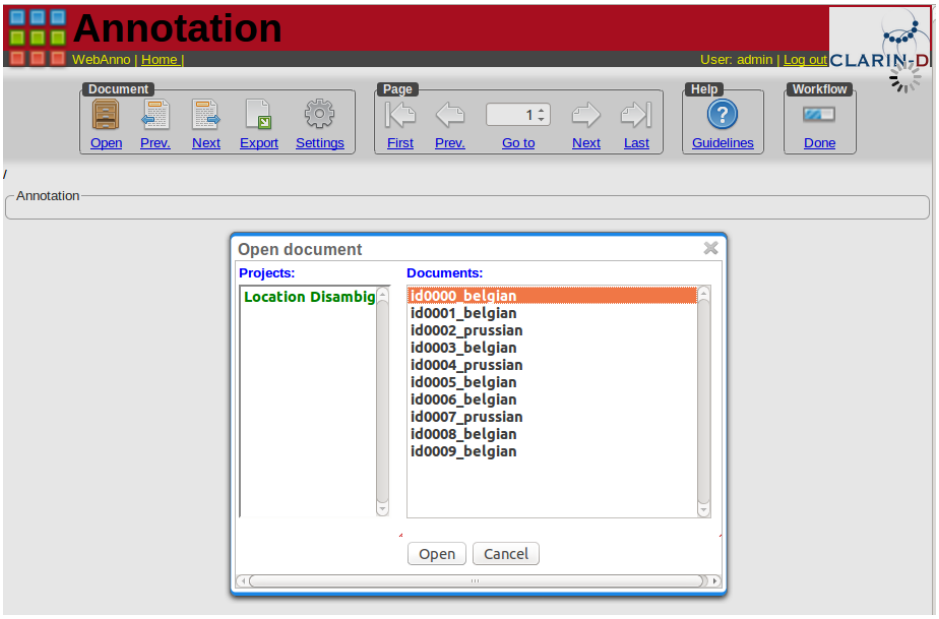


Figure 3.1: Webanno: Selection of a document to annotate.

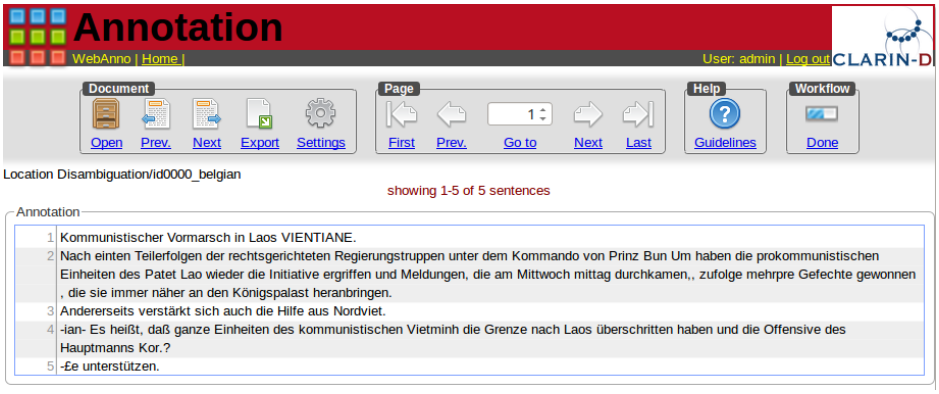


Figure 3.2: Webanno: Selected document ready to be annotated.

as locations place names used as metonyms, as in '**France** signed the deal.' When a place name is embedded in a larger named entity, the annotators were asked not to annotate the place name as a location (e.g. '*The **New York** Times*'). The task of the annotator was to map each location found in the text to the full URL of the Wikipedia article that refers to it, always using the Wikipedia in the same language in which the source text is written. If the entity does not exist in Wikipedia, the mention is tagged as `UNKNOWN`. As is usually the case in any sense disambiguation task, most of the instances are expected to correspond to the most relevant sense. Therefore, it is of the

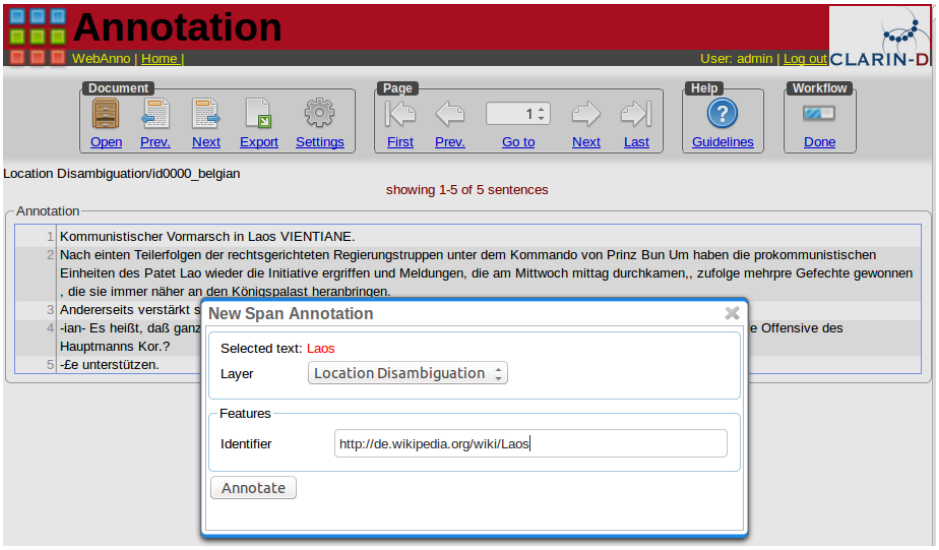


Figure 3.3: Webanno: Annotation window.

utmost importance that unusual cases (such as 'Paris' referring to a city in Kentucky, for instance) are correctly identified and resolved.

Historical texts pose some extra annotation challenges to those inherent to the task of place name disambiguation. In them, there might appear locations that do not exist as such anymore (such as Prussia, a historic German state, or Belchite, a ghost town destroyed during the Spanish Civil War) and locations that have changed their names (such as Königsberg, renamed Kaliningrad after the Second World War). The annotators had the instructions to always annotate with the closest possible referent, not only in terms of geography, but also of time. For example, given a mention 'Königsberg' referring to the city nowadays known as 'Kaliningrad' in a German text, the annotator should provide the Wikipedia URL `http://de.wikipedia.org/wiki/K%C3%B6nigsberg_(Preu%C3%9Fen)`, which refers to the time the city was known as Königsberg, before it became Kaliningrad, and not `http://de.wikipedia.org/wiki/Kaliningrad`. If there is a mention to 'Leningrad' referring to the city now known as Saint Petersburg, the best Wikipedia URL the mention could be annotated with is `http://de.wikipedia.org/wiki/Sankt_Petersburg`, since there is not a distinct entry in the German version of Wikipedia for the city when it was known as 'Leningrad'.

Besides, as mentioned, we have attempted to collect articles from newspapers that are considerably regional. This adds difficulty to the disambiguation task as well as

one additional layer of difficulty to the annotation task. The more global a text is, the more it can be expected from a general reader to know all the entities that are mentioned in it, and therefore the higher the likelihood is that mentions refer to the most common sense for this name. The more regional a text is, the less a general reader is likely to know the referents. To illustrate this with an example, a reader of the *St. Vither Volkszeitung*, a newspaper issued in St. Vith, a German-speaking Belgian municipality, encountering the word 'Born' in a piece of news will know it to refer to the little village which is about seven kilometers away from St. Vith, whereas a reader of any Barcelona-based newspaper will know it to refer to this city's neighborhood, unless otherwise specified.

Annotation examples The following paragraphs are examples of texts in German (containing some OCR errors) annotated with links to the German Wikipedia. Identified toponyms are in boldface:

(7) *Tödlicher Verkehrsunfall.* ***St. Vith****. In* ***Oneux*** *bei* ***Theux*** *stießen am Samstag der Pkw des J. S. aus* ***St. Vith*** *und der Motorradfahrer R. M. aus* ***Oneux*** *zusammen. Letzterer wurde mehrere Meter weit mitgeschleppt. Er zog sich einen Schädelbruch zu, an dessen Folgen er wenig später im Krankenhaus verstarb.*[1]

There are three different toponyms in example 7: St.Vith, Oneux, and Theux. Clearly, the two mentions to St. Vith refer to the Belgian municipality where the newspaper was based. They should be linked to `http://de.wikipedia.org/wiki/Sankt_Vith`. The mention to Theux is also straightforward to annotate, as the only Theux in the German Wikipedia refers to a neighboring municipality to St. Vith. The case of Oneux is not as trivial. There is an article for a location named 'Oneux' in Northern France (in the Somme department), but this is not the place that is mentioned twice in the article. Since there is no article for the Belgian Oneux that the text refers to, the correct label in this case is **UNKNOWN**.

(8) *Königin Elisabeth reist nach dem Kongo.* ***BRÜSSEL****. Am kommienden Freitag wird Königin Elisabeth von Belgien nach* ***Albertville*** *reasen. [...] Die Königin fliegt mit*

[1] **Translation:** Fatal traffic accident. St. Vith. This Saturday the car of J. S. from St. Vith and the motorbiker R. M. collided in Oneux at Theux. The latter was thrown several meters away. He has fractured his skull and died a little later in the hospital.

*ihrer zahlreichen Begleitung zunächst mach **Stan-leyville** und von dort nach **Albertville** wo sie vom Minister der Kolonien, Buisseret begrüßt wird.*[1]

There are three different toponyms in example 8. 'Brüssel' refers to the capital of Belgium. 'Albertville' is the old name for the city nowadays known as 'Kalemie'. Since there is no distinct article for the city before it changed its name, the annotator was expected to link the mention 'Albertville' to the Wikipedia page that refers to Kalemie (`http://de.wikipedia.org/wiki/Kalemie`). Similarly, the annotator was expected to link the string 'Stan-leyville' to the Wikipedia page that refers to Kisangani (`http://de.wikipedia.org/wiki/Kisangani`), the Congolese city known as Stanleyville at the time when the text was written. 'Belgien' has not been annotated as a toponym, as it is considered to be part of a larger named entity ('Königin Elisabeth von Belgien', i.e. Queen Elisabeth of Belgium).

(9) *Von **Athen** aus begibt sich das deutsche Kaiserpaar nach **Konstantinopel**.*[2]

Both toponyms from example 9 are straightforward to annotate, since the titles of the Wikipedia articles that refer to them match the mentions. The mention 'Athen' should be linked to `http://de.wikipedia.org/wiki/Athen` and 'Konstantinopel' to `http://de.wikipedia.org/wiki/Konstantinopel`. The interesting case of this example is to see how 'Konstantinopel' is matched to the Wikipedia page of the city before it became Istanbul. If there was no page on the German Wikipedia on Constantinople, the annotator would have had to annotate the location as referring to Istanbul (as illustrated in example 8 with the cases of Albertville and Stanleyville). The reason for selecting the closest match also in chronological terms is that the geography of the location might have undergone changes through the years (a city might have grown in one direction or had one part destroyed, a region might have annexed other territories from other regions or lost them), and the goal is to always have the best suited pair of coordinates for each mention, when possible. Finally, note how the adjective 'deutsche' in example 9 has not been identified as a location, as specified at the beginning of this section.

[1] **Translation:** Queen Elisabeth travels to the Congo. BRUSSELS. On the coming Friday, Queen Elisabeth of Belgium will travel to Albertville. [...] The Queen flies with her numerous escort to Stanleyville and from there to Albertville, where she will be greeted by the Minister of the Colonies, Buisseret.

[2] **Translation:** From Athens, the German imperial couple went to Constantinople.

3.1.3 Summary of datasets

Five corpora resulted from this annotation process. The **Prussian** corpus consists of 237 news articles in German from the years 1863 to 1894. The **Belgian** corpus consists of 163 news articles in German from the years 1955 to 1964. The **Antilles** corpus consists of 96 news articles in Dutch from the years 1921 to 1995. The **EastIndies** corpus consists of 100 articles in Dutch from the years 1913 to 1946. Finally, the **DRegional** corpus consists of 222 news articles in Dutch from the years 1800 to 1995.

A small selection of 90 documents were annotated by two people to compute their agreement. Over the total number of toponyms that have been identified by at least one of the annotators, the coincidence on the resolved toponyms to locations is of 75%. This percentage indicates the proportion of items that were linked to the exact Wikipedia URL. Some of the differences between the annotators stem from the difficulty of being able to predict the precise entity: populated places and administrative regions that share the same name are often confounded, as are geographic and political entities (e.g. 'Province of Posen' and 'Posen' or 'Ireland' and 'Republic of Ireland'). Another cause of mismatch was that sometimes annotators would link toponyms to redirect pages which eventually link to the same Wikipedia URL that the other annotator provided (e.g. 'Pacific' and 'Pacific Ocean'.)

From all the external corpora available, the **WOTR** corpus is the most adequate to be used in this chapter's experiments. It is a large historical corpus that paid special attention to consistency in the identification and annotation of toponyms. Even though the corpus and its accompanying document were published after we had annotated our data, most of the main principles are the same: metonymic uses of place names are considered locations, demonyms are not, and place names that are embedded in larger named entities are not annotated as locations. The biggest difference between WOTR and the corpora created expressly for this experiment is that the first provides geographical reference instead of links to the Wikipedia articles. Since the links to the Wikipedia articles can easily be converted into coordinates, this is not a major problem. Some of the locations in WOTR were originally annotated with polygons, but the authors of the corpus provide a simplified version of the corpus in which polygons of geographic coordinates have been translated to latitude and longitude centroids. I used this simplified version of the annotations because it is closer to the geographic

	Known	Total	Language	Domain	Country	Years
WOTR	10,380	11,795	English	War letters and reports	United States	1860s
Prussian	1,529	1,745	German	Newspaper articles	Prussia	1863–1894
Belgian	544	560	German	Newspaper articles	Belgium	1955–1964
Antilles	301	335	Dutch	Newspaper articles	Netherlands Antilles	1921–1995
EastIndies	210	262	Dutch	Newspaper articles	Dutch East Indies	1913–1946
DRegional	1,037	1,124	Dutch	Newspaper articles	Netherlands	1800–1995

Table 3.1: Summary of datasets: 'known' refers to the number of identified toponyms for which coordinates are known, 'total' to the total number of identified toponyms, 'language' to the language in which the collection was written, 'domain' to the domain of the collection, 'country' to the country of publication at the time of writing, and 'years' to the years spanning the collection.

reference that can be extracted from Wikipedia. Table 3.1 summarizes the six datasets that have been used for the toponym resolution experiment.

3.2 Resources

Toponym disambiguation is usually treated as a classification task, where a toponym is mapped to the best among the set of candidates that can potentially be referred to by it. Given a mention 'Paris' in a text, for example, the referent may be the capital of France, but other potential referents exist that are known by the same name (e.g. a city in Texas, a town in Tennessee, or a community in Ontario) and should not be ignored, as the possibility exists that they are the true referents of the mention. In order to obtain an adequate pool of candidate locations for each toponym identified in a text, a resource with geographic knowledge is needed. Candidate selection has often been overlooked in many toponym resolution methods, and many papers do not even describe how they obtained the set of candidates. In this section, I start with an overview of the resources exploited in previous toponym disambiguation approaches in order to extract potential candidates and then focus on the advantages and disadvantages of selecting an encyclopedic resource such as Wikipedia over a geographic resource such as GeoNames

as a base from which to build a new resource that combines both kinds of knowledge. I then proceed to describe the decisions taken in its creation and development. I create a resource for each different working language, which I henceforth call $GeoSemKB_{xx}$, where `xx` is the language code of the language in question. The language code is left empty if no specific language is meant. I conclude this section by summarizing its key points.

3.2.1 Brief review of resources

The choice of an appropriate knowledge source is one of the keys to a successful disambiguation of toponyms, since it determines which candidates are extracted for each toponym and limits the ambiguity in the disambiguation task. The choice of a wrong knowledge source may have a great impact on the overall accuracy of the disambiguation task, as it directly influences recall (if the correct referent is not among the candidates) or may affect precision (if the number of candidates is too high). Since the number of candidates that potentially refer to a toponym is directly related to the knowledge source, the decision of using one resource or another can have a profound impact in the final result. After analyzing several prominent entity linking approaches, Hachey et al. (2013) (49) reported that candidate selection accounts for most of the variation between the different examined systems, and by analyzing the output of a toponym resolution system, Quercini et al. (2010) (79) found that candidate selection can have a negative impact on the results both if it is noisy and if it is not sensitive enough.

Some of the first toponym resolution approaches built their own resources from different available sources. Smith and Crane (2001) (89) built a resource from different geographic sources, among which the Getty Thesaurus of Geographic Names,[1] ending up with a gazetteer of above one million place names. Also the Web-a-where system, developed by Amitay et al. (2004) (3), created a gazetteer from different sources and had a total of 40,000 different world locations (with a total of 75,000 different spelling variations). Their gazetteer was possibly heavily biased towards locations in the United States (as they used one source for locations inside the United States and one source for locations outside the United States) and it could be expected to perform poorly in resolving toponyms from local newspapers due to the choice of restricting the gazetteer to only countries and cities of more than 5,000 inhabitants. Also Leidner (2008) (59)

[1] `http://www.getty.edu/research/tools/vocabularies/tgn/`.

built his own gazetteer from different sources: the GNIS gazetteer of the U.S. Geographic Survey,[1] the GNS gazetteer of the National Geospatial-Intelligence Agency,[2] and the CIA World Factbook.[3] The author admits that this gazetteer would not be suited for disambiguating toponyms in historical texts.

A good resource must have high quality, wide coverage, and sufficient depth. In more recent years, several resources have appeared that largely meet these requirements and, combined with their availability, easy usage, and global scope and ambition, have eclipsed other knowledge sources, some of which with more authority. From these, the most used have undoubtedly been GeoNames for toponym resolution and Wikipedia for entity linking. Both are user-edited and maintained, freely accessible, and flexible resources that aim at containing all the knowledge in the world: geographic knowledge the first, encyclopedic knowledge the latter. Despite the undeniable similitude between toponym resolution and entity linking, these two tasks have rarely ever intersected. Each task has its own strategies and techniques and also favorite resources.

With more than ten million locations from different sources and user-contributed and maintained, `GeoNames`[4] has become the most common resource in recent toponym resolution systems (Speriosu (2013) (91), Lieberman et al. (2010) (60), etc.). Each location in GeoNames has several information fields, such as name, latitude and longitude, country, and population, and each location is accompanied by a list of alternate names with which it may also be known (i.e. orthographic variations, historical names, or names in other languages). Most approaches using GeoNames extract candidates for each toponym by just performing shallow surface matching between the mention and the titles of GeoNames entries and their alternate names. Roberts et al. (2010) (84), though, consider that the alternate names provided by GeoNames are not always enough to find the correct candidates. They propose a series of strategies (basic, caseless, safe, moderate, and aggressive) to improve the retrieval of matches. These strategies range from altering capitalization of the toponyms to linking the GeoNames entities to their Wikipedia equivalents and capturing their redirect links and results from the disambiguation pages. Since the ambiguity of a toponym depends exclusively on the choice of gazetteer and candidate selection strategy, it is hardly surprising

[1] `http://geonames.usgs.gov/domestic`.
[2] `http://geonames.nga.mil/gns/html`.
[3] `https://www.cia.gov/library/publications/the-world-factbook/`.
[4] `http://geonames.org`.

that the more aggressive their candidate selection strategy is, the lower the precision reported is and the higher the recall.

The resources used in entity linking approaches differ from the ones commonly used in toponym resolution tasks as they both respond to the different needs of each task. The goal of entity linking is not to return coordinates for each location but to link entity mentions (among which, toponyms) to entries in a knowledge base. `Wikipedia`[1] and Wikipedia-based resources (such as DBpedia (58), YAGO (93) and Babelnet (70)) are indisputably the most common resources in entity linking approaches. These resources have links to geographic resources (among which, GeoNames) from which coordinates can be derived. Candidate extraction from Wikipedia is not as straightforward as from GeoNames and allows for more flexibility. Therefore, there are more decisions to be taken which can affect the final outcome. Since the beginning, most approaches have taken advantage of the same features (page redirects, disambiguation pages, and stripping appositions from page titles), easy to extract due to the linked structure of Wikipedia.

3.2.2 Wikipedia as a resource base

GeoNames is a resource with a global scope which stemmed from several different gazetteers. It is edited and maintained by its users. One of the biggest differences between GeoNames and Wikipedia is that, whereas GeoNames is one single resource, the Wikipedia project in reality consists of 284 different active Wikipedia versions that operate independently, one for every different language for which there is a community. Even though they are closely related, Wikipedia versions in different languages operate independently and therefore are at different stages of development. As of December 2016, the English Wikipedia had 5,304,238 articles, the German Wikipedia 2,007,472 articles, and the Dutch Wikipedia 1,884,991 articles. As in GeoNames, content in Wikipedia is edited and maintained by users, but also generated, and this is one of its main characteristics. This of course has several consequences, one of the most controversial is that of lack of authority. Even if sources are provided, this does not change the fact that it is still very possible to include errors in the articles (both voluntarily and involuntarily). The lack of accuracy in Wikipedia in comparison with other authorative encyclopedias can be debatable (e.g. Giles (2005) (44)), and yet not

[1] `http://www.wikipedia.org/`.

all consequences are as gloomy: the easy and free accessibility and its popularity are also remedies for possible inaccuracies, as it means that more possible editors and correctors exist. Furthermore, the fact that it comes without authority also means that information can easily be more up-to-date.

One of the five pillars of Wikipedia is that it has a neutral point of view.[1] Being user-generated, though, several biases inevitably exist. One of these biases is the geographic systematic bias existing across the different Wikipedia versions, which Overell (2009) (72) quantified. Most language versions of Wikipedia have their focuses shifted towards the locations where the language is spoken. Figure 3.4 shows the location distribution plot in the French and the German versions of Wikipedia. It is manifest that some bias exist towards France and Germany respectively.

Figure 3.4: Heatmaps in the French Wikipedia (left) and in the German Wikipedia (right), source: Overell (2009) (72).

Overell (2009) (72) also concludes that the English version is the most global and therefore the one that has the least geographic bias, which can be partially attributed to the fact that, being English the lingua franca of the Internet, it is widely used also by non-native speakers. To illustrate this bias, he created a series of cartograms. In them, countries shrink or swell depending on whether the country has a higher or lower number of references per inhabitant than average, whereas the color represent the absolute number of references (the darker, the more references). The geographic bias becomes obvious by looking at figure 3.5, which shows the cartograms of references in the Portuguese and Spanish versions of Wikipedia, where the difference in both South America and the Iberian Peninsula is very graphic: in the Portuguese Wikipedia, the size of Brazil and Portugal is comparably bigger than that of their neighboring Spanish-speaking countries, and vice-versa.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Five_pillars`.

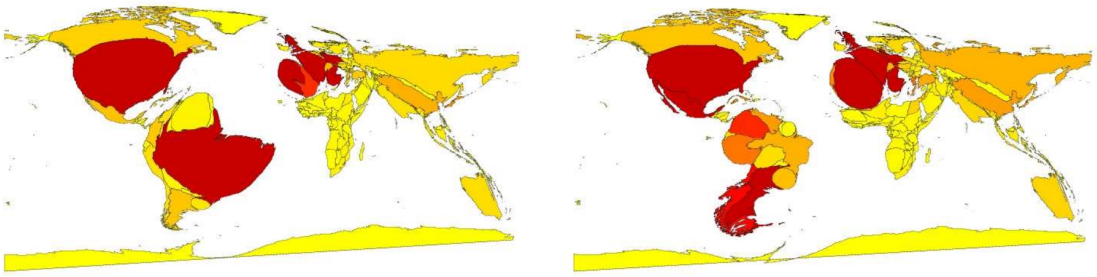


Figure 3.5: Cartograms of references in the Portuguese Wikipedia (left) and the Spanish Wikipedia (right), source: Overell (2009) (72).

The question of whether geographic bias is a negative aspect of Wikipedia could be argued. For many tasks, the answer should probably be a clear yes. However, this bias also reflects the viewpoint of those writing the articles, often native speakers of the language of a particular Wikipedia version. Geographic bias results in countries where the language is widely spoken being overrepresented, as are countries that have had a strong influence or share cultural ties, whereas more unknown regions of the World (e.g. Africa or Central Asia, in figure 3.5) are underrepresented. Even though a biased representation of the world is inaccurate, it could be argued that it is a more faithful representation of the world as seen from the point of view of the majority of the speakers of a certain language. Geographic bias is also a constant in the reporting of news. Mehler et al. (2006) (65) analyzes geographic biases in newspapers, which arise from the premise that people in different places talk about different things. Bias in newspapers is also detected and exploited for toponym disambiguation in Buscaldi and Magnini (2010) (25) and Lieberman et al. (2010) (60).

The decision to use Wikipedia as the main source of knowledge from which to build a resource that will assist in the selection of candidates and disambiguation to the best among them is motivated by the many advantages of it. Despite the awareness that it can be unreliable and biased and has a narrower coverage than GeoNames, Wikipedia is also the most complete and up-to-date online encyclopedia, and the contents of its entries are conveniently structured. The vast majority of entries corresponding to locations (the number of exceptions is minimal) encode geographical coordinates. A definitive advantage of using an encyclopedia in detriment of a gazetteer is that the first comes with a context: the body of the entry, which describes the entity. This is of great assistance both in human as well as machine disambiguation.

Comparing GeoNames and Wikipedia through an example. In GeoNames, the query 'Göttingen' yields several results, among which the following three:

`geonameid:` 3221013	`geonameid:` 2918632	`geonameid:` 6557373
`name:` Landkreis Göttingen	`name:` Göttingen	`name:` Göttingen
`alternatenames:` Goettingen, Landkreis Goettingen, Göttingen, Landkreis Göttingen	`alternatenames:` Choettingen, Getingen, Getynga, Gjottingen, Goettingen, Gotinga, etc.	`alternatenames:` Goettingen, Stadt Goettingen, Göttingen, Stadt Göttingen
`latitude:` 51.5108300	`latitude:` 51.5344300	`latitude:` 51.5129700
`longitude:` 9.9136100	`longitude:` 9.9322800	`longitude:` 9.9535300
`fclass:` A	`fclass:` P	`fclass:` A
`fcode:` ADM3	`fcode:` PPLA3	`fcode:` ADM4
`country:` DE	`country:` DE	`country:` DE
`admin1:` 06	`admin1:` 06	`admin1:` 06
`admin2:` 00	`admin2:` 00	`admin2:` 00
`admin3:` 03152	`admin3:` 03152	`admin3:` 03152
`admin4:`	`admin4:`	`admin4:` 03152012
`population:` 247988	`population:` 122149	`population:` 116650

It may be difficult for a human to grasp the differences between these three entities at first sight. The coordinates inform us that the three points are separated by 2.77, 2.81, and 2.93 kilometers the one from the others. Since it is a convention in GeoNames (as is in most gazetteers) to mark locations regardless of their extension by just one point in the map (sometimes the centroid, sometimes an approximation to the capital city), we could assume that the three locations are probably overlapping. The entity in the left column (`geonameid` *3221013*) refers to the Landkreis Göttingen, the district of Göttingen, which is an administrative division (`fclass` *A*) of third order (`fcode` *ADM3*) and has a population of 247,988. The entity in the middle column (`geonameid` *2918623*) is classified as a populated place (`fclass` *P*) and as a seat of a third-order administrative division (`fcode` *PPLA3*) and refers to the city of Göttingen, capital of the district of Göttingen, and has a population of 122,149. Finally, the entity in the right column (`geonameid` *6557373*) is classified as an administrative division (`fclass` *A*) of fourth order (`fcode` *ADM4*) which is an administrative subdivision of the district of Göttingen, with a population of 116,650, that can also be known as 'Stadt Göttingen'. From the data alone, it cannot be understood which is the relation between the second and third locations. One explanation would be that the third location is an administrative division inside the city of Göttingen with about 5,500 inhabitants less than the number of inhabitants of the city. However, there are in GeoNames no other administrative divisions of the same order that encompass the remaining population.

Another explanation would be that the fourth-order administrative division Göttingen (`geonameid` *6557373*) actually encompasses the populated place Göttingen (`geonameid` *2918623*). Yet, the latter has a smaller population than the should-be-encompassing administrative division.[1]

GeoNames is a traditional gazetteer that contains geographic information. Granularity of geographic information in Wikipedia is coarser and coverage is narrower than in GeoNames (from the three entities in the example, only the district and city of Göttingen have a page in the English Wikipedia), but Wikipedia has the distinct advantage that entities come with a context in natural language: an introduction of the entity and optionally other kinds of information such as the historical background. Figure 3.6 shows the English Wikipedia page for the district of Göttingen.

My preference for Wikipedia as a base from which to build the resource arises from the will of treating toponym disambiguation in an analogous manner to how a human would face the problem. After all, the natural audience of newspapers and other forms of written text are humans, and it is therefore possible that if a person can find more clues for disambiguating a toponym in Wikipedia, so might a text mining system, since a method based solely on geographic information will miss the many clues that context can provide.

3.2.3 Building a resource

I now proceed to describe the process of building a resource based on Wikipedia and complemented with information from GeoNames. Since context words are also exploited in the method proposed in this chapter, it is necessary to build one resource for each different language of each collection, namely English, German, and Dutch. In the following pages I describe how Wikipedia and GeoNames sources were obtained and how locations were extracted from them. For each location, alternate names were found and a series of features were extracted to assist in the disambiguation process.

[1]Note that inaccuracies and inconsistencies (such as repeated or overlapping locations) are not rare in GeoNames, as reported in Ahlers (2013) (1). GeoNames integrates several gazetteers (`http://www.geonames.org/data-sources.html`) and some automatic merging was surely necessary in order to integrate the data. However, the strategy followed to merge the data has not been published.

Figure 3.6: Wikipedia page of the district of Göttingen in English.

3.2.3.1 Obtaining the sources

A resource has been built for each language in a parallel fashion. First, the Wikipedia dumps were downloaded for English,[1] German,[2] and Dutch.[3] To do so, at least the following files are required (where `xx` should be replaced by the language code of the Wikipedia dumps: '`en`' for English, '`de`' for German, and '`nl`' for Dutch):

```
xxwiki-latest-page.sql.gz
xxwiki-latest-pagelinks.sql.gz
xxwiki-latest-pages-articles.xml.bz2
```

[1] `http://dumps.wikimedia.org/enwiki/latest/`, Wikipedia dump from November 2015.
[2] `http://dumps.wikimedia.org/dewiki/latest/`, Wikipedia dump from June 2016.
[3] `http://dumps.wikimedia.org/nlwiki/latest/`, Wikipedia dump from June 2016.

```
xxwiki-latest-redirect.sql.gz
xxwiki-latest-category.sql.gz
xxwiki-latest-categorylinks.sql.gz
```

The JWPL library[1] (Zesch et al. (2008) (104)) was used for processing Wikipedia data. As for GeoNames, its latest version was downloaded.[2]

3.2.3.2 Location extraction

Coordinates are not always easy to find in Wikipedia articles. For most locations, coordinates can be extracted from the `geo_tags` table, which is downloadable from the Wikipedia dumps.[3] Another possibility to link locations to their pair of coordinates is through the DBpedia linking from GeoNames to Wikipedia[4] or from the '`link`' field in GeoNames '`alternatenames`' table, which links the Wikipedia entity to its GeoNames equivalent. Besides, I made use of the interlanguage functionality[5] to transfer coordinates from articles from one Wikipedia version to another, therefore making sure that entities that are locations in one Wikipedia version and that exist in another version are also locations in the latter. Table 3.2 shows the number of articles with content for which coordinates could be extracted.

	Locations
$GeoSemKB_{en}$	972,957
$GeoSemKB_{de}$	392,370
$GeoSemKB_{nl}$	294,122

Table 3.2: Number of articles for which coordinates could be extracted in the English, German, and Dutch resources.

The coverage of GeoNames is much wider than the geographic coverage of any Wikipedia language version. In August 2016, GeoNames had coordinates for 9,084,628 entities, whereas I was able to extract coordinates for 972,957 entities in the largest

[1] `http://dkpro.github.io/dkpro-jwpl/`.

[2] `http://download.geonames.org/export/dump/`, retrieved in August 2016.

[3] `http://dumps.wikimedia.org/xxwiki/latest/enwiki-latest-geo_tags.sql.gz`, where 'xx' should be replaced for the language code of the Wikipedia version in question. It contains information from Wikipedia infoboxes for those entities that encode coordinates (a total of 1,736,396).

[4] `http://wiki.dbpedia.org/Downloads`.

[5] Interlanguage links from version to version of Wikipedia can be downloaded from the Wikipedia dumps: `http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-langlinks.sql.gz`.

Wikipedia version, the English one,[1] more than nine time less than GeoNames. And yet, Wikipedia offers a series of unique advantages that are difficult to dismiss: the contextual information, a system of page redirection and of language linking, semantic indexing, and a system to evaluate subjective relevance of an entity (by number of incoming links) that are characteristics that may well be worth trading off for coverage.

3.2.3.3 Finding alternate names

Locations do not necessarily have a unique name to refer to them. Some locations might accept different spellings (e.g. both 'Marseille' and 'Marseilles' can be used in English to refer to the French city on the shores of the Mediterranean), some locations are known differently in different languages (e.g. Saarbrücken, the city on the French border, is known as Sarrebruck in French), some locations have undergone linguistic variation along the years (e.g. the German financial city of Frankfurt was often spelled 'Frankfort' in English in the past), and some locations are known nowadays with a name completely different from the name it was known by in the past (e.g. the Japanese capital, Tokyo, was known as 'Edo' until 1868). It is important that different naming options are associated to the locations they refer to, even more so when dealing with historical documents, as place names have obviously undergone more changes.

For each location, the list of name variations has been derived both from Wikipedia and GeoNames. Wikipedia's policy is that every article has a unique unambiguous title. If more than one entity shares the same name, disambiguators are added to the title as in the examples below:

- `Trois-Rivières`
- `Trois-Rivières,_Martinique`
- `Trois-Rivières,_Guadeloupe`

Convention dictates that the most common entity does not have a disambiguator element in the title (as is the case of the first Trois-Rivières in the example, referring to a city in Quebec). Less-common entities have an apposition that is a good disambiguation clue (Martinique and Guadeloupe in the example). Both the full title and the title

[1] Note that coordinates provided by the Geographical coordinates WikiProject, which can be downloaded from the dumps, has over 1.7 million geocoded entries, but many of them do not correspond to article pages with content.

where the apposition has been stripped (i.e. 'Trois-Rivières' in the three examples) are kept as alternate names.

Redirect pages automatically send Wikipedia users from an empty page whose title coincides with the users' query (e.g. 'NYC') to a page with content (e.g. 'New York City') and informs them with a '*Redirected from*' warning, in order to facilitate their navigation. This characteristic is highly valuable for disambiguation, since also different spellings and historical place names are often also encoded. For example, if 'Leningrad' is queried, the user will be transferred to the `Saint_Petersburg` page. Titles of all pages that are redirected to a location are kept as alternate names (i.e. 'NYC' is stored as an alternate name for `New_York_City` and 'Leningrad' is stored as an alternate name for `Saint_Petersburg`).

The data for this thesis was provided by the European project AsymEnc, which investigated the transnational contact in Europe during the years between 1815 and 1992. This was a period of time that witnessed a constant movement of borders which, in combination with dramatic changes of political systems and status of countries, led to many renamings of places. For this reason, I also store as alternate name for each location the name in different languages, restricted to the six most widely spoken languages in Europe: German, Italian, English, French, Spanish, and Polish. Different language versions of Wikipedia have different stages of development. For this reason, it is possible that a redirect page in one language is not a redirect page in another. Aware that by performing this step it is easy to introduce also some noise, I transfer the redirect pages from the three Wikipedia versions with which I am operating (English, German, and Dutch) from the one to the other.

Finally, for each location in Wikipedia for which there is a link to a location in the GeoNames database, I store all the alternate names from GeoNames as alternate names of the location. Table 3.3 displays the number of locations in the built resources in each language and the total number of alternate names. There are 2.9 names in average for each location in English, 3.7 in German, and also 3.7 in Dutch.

3.2.3.4 Extraction of geographic features

For each entry in the resource (i.e. for each Wikipedia article for which coordinates could be retrieved), the following features are also extracted, when possible, describing its geography and associated characteristics:

	Locations	Names
$GeoSemKB_{en}$	972,957	2,862,779
$GeoSemKB_{de}$	392,370	1,414,455
$GeoSemKB_{nl}$	294,122	1,098,499

Table 3.3: Number of locations and the names they may be known by in the English, German, and Dutch resources.

- **Latitude and longitude:** geographic coordinates that specify the distance of a location north or south of the Equator and east or west of the Greenwich meridian, represented as one point on the Earth's surface and measured in decimal degrees.
- **Population:** number of inhabitants of a populated place or administrative unit. This information is extracted from the Wikipedia version of the language in question, if existing and found. Otherwise, it is extracted from any of the other Wikipedia versions or, in the last instance, from GeoNames.
- **Number of inlinks:** number of incoming links to the Wikipedia article, i.e. links from other Wikipedia articles that point to the location in question. For example, if the location is the city of Göttingen, this feature stores the number of articles in Wikipedia that have a link to the article on the city of Göttingen.
- **Country:** country inside which the location is situated, if any, expressed as its ISO code. This information is again extracted from the Wikipedia version of the language in question, if existing. Otherwise, it is extracted from any of the other Wikipedia versions or, in the last instance, from GeoNames.

3.2.3.5 Extraction of context features

Additionally, for each location a series of semantic features are extracted, complementing the information about the location in natural language. These include the title of the Wikipedia entry of the location[1] and, if there is an apposition in the title, both the stripped title and the apposition (e.g. in 'Trois-Rivières,_Martinique', both 'Trois-Rivières' and 'Martinique' should be stored as semantic features). Also the name of the country and the region of the location are included as semantic features. Finally, I also extract context words from parts of the body of the Wikipedia article that have

[1]The last part of the Wikipedia URL.

history content. In order to do so, some text processing is required, as is described below.

Obtaining context words The body of any Wikipedia article is edited in the Wikipedia Markup Syntax, which facilitates its visualization when displaying it in a browser. The markup contains some information that is relevant in terms of content (such as section splitting, lists, and internal links) but also some metadata that are irrelevant for the research here conducted (such as pronunciation aids, images, or references and citing sources). Figure 3.7 shows an example of wikitext from the Wikipedia article of Göttingen.

```
'''Göttingen''' ({{IPA-de|ˈgœtɪŋən}} {{Audio|De-Göttingen-pronunciation.ogg|
<small>listen</small>}}; [[Low German]]: ''Chöttingen'') is a [[college
town|university city]] in [[Lower Saxony]], [[Germany]]. It is the [[Capital
(political)|capital]] of the district of [[Göttingen (district)|Göttingen]].
The [[River Leine]] runs through the town. In 2011 the population was
116,052.

==General information==
The origins of Göttingen lay in a village called ''Gutingi, ''first mentioned
in a document in 953 AD. The city was founded northwest of this village,
between 1150 and 1200 AD, and adopted its name. In [[Middle Ages|medieval]]
times the city was a member of the [[Hanseatic League]] and hence a wealthy
town.

[[File:Göttingen Gänseliesel März06.jpg|thumb|left|Landmark [[Gänseliesel]]
fountain at the main market.]]
Today, Göttingen is famous for its old university (''Georgia Augusta'', or
[[University of Göttingen|"Georg-August-Universität"]]), which was founded in
1734 (first classes in 1737) and became the most visited university of
Europe. In 1837, seven professors protested against the absolute sovereignty
```

Figure 3.7: Fragment of the source text of the Wikipedia entry for Göttingen.

In order to remove unnecessary metadata, some preprocessing is wanted. Some steps to clean the wikitext from irrelevant information are common in all languages, such as removal of different xml tags and other markup elements. In order to reduce noise, I remove birth and death dates of some people with a very naive strategy. They are usually introduced the first time a person is mentioned and, assuming the person has an entry in Wikipedia, they are usually expressed in this form: `[[person name]] (birthyear-deathyear)`, as in `[[Albert Einstein]] (1879-1955)`.

In any Wikipedia entry, the body of an article ideally consists of two parts: the lead section and the content. The lead section consists of an introduction to the entity in question and comes before the table of content, if existing. If this is the case, the content of the entity follows the table of contents and is distributed into different sections. The context information provided by some sections is more useful than that provided by others. The titles of all sections and subsections are marked by two or more surrounding equal signs ('='), as in `==History==` or `===Modern history===`. The names of the sections are not standardized and are therefore obviously different from one language to the other. I removed several full sections that I considered irrelevant for the task since they rarely contain context words of interest for the disambiguation strategy and could possibly introduce some noise. They are the following, in each language:

- **English:** Notes, See also, References, External links, Citations, Further reading, Bibliography, Sources, Footnotes, Gallery, Twin cities, Other sources, Web resources, Etymology, Twin towns, Sister cities.
- **German:** Literatur, Quellen, Einzelnachweise, Siehe auch, Bilder, Partnerschaft, Fußnoten, Namensherkunft, Etymologie, Städtepartnerschaften, Galerie, Weblinks, Allgemeine Quellen, Literatur und Quellen, Name, Filmografie, Weiterführende Informationen, Anmerkungen, Bildergalerie, Namensgebung.
- **Dutch:** Stedenbanden, Zie ook, Externe link, Noot, Referenties, Citaten, Literatuurlijst, Bronnen, Afbeeldingen, Naam, Foto's, Noten, Literatuur, Galerij, Naamsgeschiedenis, Bronvermelding en referenties, De naam, Etymologie.

My aim is to extract historical contexts for the locations, specifically from late modern and contemporary history, spanning the 19th and 20th centuries. In extracting words related to these periods from the Wikipedia entry corresponding to a location, two scenarios are possible:

- **The article does not have a 'History' section:** If the article does not have an explicit 'History' section, all sentences from the remaining sections that contain a number that is between 1800 and 1999 are set aside. For each sentence, each mention of a number between 1800 and 1999 is converted into a decade (e.g. if there is the mention '1834' in a sentence, this number is converted into 1830),

and the sentence is stored as part of this decade. If more than one decade is found in a sentence, this sentence is assigned to each decade that is present in the sentence. If there is a time period in a sentence, the sentence is stored for each decade the period spans. For example, the sentence '*The beginning of the Napoleonic Wars (1804–15) is often chosen as a convenient point in time with which to date the end of the Enlightenment*' would be assigned to the decades 1800 and 1810, as would if the period was expressed as 'between 1804 and 1815' and in similar choices of word rephrasings.

- **The article has a 'History' section:** I approach articles that have an explicit 'History' section ('Geschichte' in German, 'Geschiedenis' in Dutch) in a different manner. In this case, only this section and its subsections are considered, whereas the rest of the text of the article is disregarded. Every sentence is processed one after the other. The first time a number between 1800 and 1999 appears in a paragraph, the decade to which it belongs is stored. If there is no date specified in the following sentences, the stored decade is assigned to them, and so forth until the next sentence in which a year is specified. This decision rests on the ground that the history section often follows a chronological order and coherence.

For each Wikipedia page referring to a location (i.e. for each Wikipedia page for which coordinates could be extracted), a set of tuples decade–sentence are obtained. For each decade, sentences are tokenized and function and frequent words and punctuation signs are removed. The rest of the words are considered context words.[1] Below are some examples of words extracted for some decades for two locations, Göttingen and Curaçao, according to $GeoSemKB_{en}$:

Göttingen

`1800:` westphalia, electorate, prussia, brunswick, hanover, lüneburg, napoleon, etc.

`1860:` province, prussia, voted, cause, war, kingdom, declaring, neutral, hannover, etc.

`1930:` war, housed, prevented, albert, kristallnacht, process, etc.

`1940:` heavily, caused, comparatively, kassel, bombing, impact, timbered, etc.

[1] Free links can be customized in Wikipedia articles, the text displayed on the browser might be different than the title of the article that the link points to. For example, the string `[[Intercity-Express|ICE]]` produces the text 'ICE' but links to the `Intercity-Express` page. In these cases, both elements before and after the vertical bar are taken as context words.

1970: dissolved, district, enlarged, orm, münden, hannoversch, incorporating, etc.

Curaçao

1810: napoleonic, island, dutch, colony, wars, dependencies, incorporated, stable, etc.

1860: cane, owner, exchange, caribbean, harvest, sugar, plantation, master, slave, etc.

1910: government, maracaibo, traffic, panama, increase, attracted, future, dutch, etc.

1960: discontent, process, protest, uprising, islanders, rioting, antagonisms, social, etc.

1980: downturn, venezuela, independence, stagnation, atmospheric, tourism, creole, etc.

Besides, for each entity also the context words from the lead section (i.e. the introduction of the article) are stored as context features, limited to three paragraphs when it is longer than this. Some of the 'intro' context words for Göttingen are 'town', 'lower', 'saxony', 'university', 'capital', 'leine', and 'germany'; and for Curaçao are 'uninhabited', 'papiamentu', 'dissolution', 'sea', 'antilles', and 'caribbean'.

3.2.4 Summary of resources

In this section I have described the design of a resource that can assist in a fast and robust manner in the task of disambiguating toponyms from historical documents. The resource has been built using Wikipedia as a base and has been fed with additional knowledge from GeoNames. It does not only have geographic knowledge, but also contains semantic information. For this reason, there needs to be one resource for each different working language. I have therefore created three different resources: $GeoSemKB_{en}$ for English, $GeoSemKB_{de}$ for German, and $GeoSemKB_{nl}$ for Dutch. The English resource consists of 972,957 locations and a total of 2,862,779 alternate names, the German resource consists of 392,370 locations and a total of 1,414,455 alternate names, and the Dutch consists of 294,122 locations and a total of 1,098,499 alternate names. Each location in the resource has a series of features that can be classified into two types: features that describe the geography and associated characteristics of the location (latitude and longitude, population, number of Wikipedia inlinks, and country it belongs to) and context features extracted from the body of the Wikipedia article that refers to it. The context features are extracted from the introduction of the article and from the historical sections and fragments. The resources described in this section are key to the success of the toponym disambiguation method that is introduced in the next section.

3.3 Disambiguating toponyms

I present in this section the core of the proposed method. Three distinct subtasks can be distinguished: **toponym identification** takes plain text as input and recognizes mentions that refer to locations, **candidate selection** generates sets of potential locations that may be the referent of each identified toponym, and **toponym disambiguation** per se selects the best candidate and provides its unique identifier in the resource and its latitude and longitude. They are described in detail in the next subsections, and the main ideas of the method are summarized at the end of this section.

3.3.1 Toponym identification

I perform toponym identification through an existing and widely-used out-of-the-box named entity recognizer, Stanford Named Entity Recognizer[1] (SNER from now on) (Finkel et al. (2005) (43)), trained on modern day newspaper data.[2] On the recognizer's output, I applied several post-processing steps.

The heuristics I use can be classified into two types. The first type uses context clues that provide strong evidence that the mention corresponds to a location. For example, in English, the presence of certain words such as 'cape', 'port', 'pass', 'town', 'lake', 'river', or 'valley' in the immediate vicinity of a capitalized expression is usually a clear indicator of it being a location, as are endings in 'burg' and 'ville', or two-token expressions starting with 'San', 'Saint', and derivative forms. The second type of heuristics uses context clues that provide strong evidence that the mention does not correspond to a location, but to a person. It is common practice that the first time a person is mentioned in a text that has a wide audience the person is referred to either with a full name (first and last name) or with a title. Therefore, in these kinds of texts, the first mention of a person is rarely a one-token expression. If the first token of a multi-token named entity expression is a title (such as 'Mr.', 'Professor', 'Countess', etc.) or if the context of an expression is a clear indicator that the named entity is a person (such as 'the 28-year-old' before a capitalized expression), the mention is

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

[2] For English, I used a classifier that comes by default with the software, for German I used the classifier based on Faruqui and Padó (2010) (42) (`http://nlp.stanford.edu/software/stanford-german-2015-10-14-models.jar`) and for Dutch the training data from the CoNLL-2002 shared task (`http://www.cnts.ua.ac.be/conll2002/ner`).

considered a person. Since this seems to provide stronger evidence than the first type of heuristics, it overrules it. Stanford NER tags entities sequentially and therefore different occurrences of the same mention in the same document might obtain different labels. Similarly to most disambiguation methods, I proceed on the one-sense-per-discourse assumption, according to which a mention refers always to the same entity throughout the document. Therefore, for consistence, a mention should likewise be always classified into the same named entity class throughout a document. If evidence is clear about an occurrence of a named entity expression being of the person class, I consider all full- or partial-matches of these expressions that can be found in a document to also be classified as a person. Both types of heuristics are combined and, finally, a last post-processing step to identify place names is implicitly performed by the selection of candidates: only if at least one candidate location is found for a mention, can this mention be considered a place name.

3.3.2 Candidate selection

The goal of candidate selection is to find the set of potential referents for each of the toponyms identified through the previous step. Ideally, the correct referent should be found among the set of extracted candidates. Subsection 3.2.3.3 describes the strategy to obtain alternate names for each location, which is a first and necessary step to perform candidate selection.

As an example, table 3.4 shows the result of identifying alternate names of the Indonesian city Surabaya (once the largest city in Dutch East Indies), according to the resource built for Dutch. Its alternate names were obtained from different sources: the main title 'Soerabaja', the redirect pages 'Soerabaya' and 'Surabaja', and the page name retrieved through interlanguage linking (in this case, German) 'Surabaya' were all retrieved from the resource in the same language as that of the source text (in this case, Dutch). The redirect pages 'Sourabaya', 'Soreabaia', and 'Kutisari' were retrieved from the English Wikipedia. Finally, all alternate names from GeoNames written in Latin characters, such as 'Sourampagia', 'Surabaia', and 'Surabajo' were also transferred to the $GeoSemKB_{nl}$ resource.

The possibility that a location is known by a name that has had to be derived from Wikipedia versions in other languages seems small. However, as mentioned, there have been many movements of borders in the period spanning the different collections,

WikiID	Alternate name	Source	Type
91643	Soerabaja	Dutch Wikipedia	mainttl
91643	Soerabaya	Dutch Wikipedia	redirect
91643	Surabaja	Dutch Wikipedia	redirect
91643	Surabaya	Dutch Wikipedia	translation:de
91643	Sourabaya	English Wikipedia	redirect
91643	Soerabaia	English Wikipedia	redirect
91643	Kutisari	English Wikipedia	redirect
91643	Sourampagia	GeoNames	alternatename
91643	Surabaia	GeoNames	alternatename
91643	Surabajo	GeoNames	alternatename
91643	...	...	...

Table 3.4: Some alternate names of the city of Surabaya in Indonesia according to the Dutch resource.

and places have changed names often depending on their belonging to one country or another. Table 3.5 shows the alternate names of the town of Bielice, in the Lower Silesia Province of Poland, according to the English resource, $GeoSemKB_{en}$. In this case, the English Wikipedia could only contribute with the main title (with and without apposition), whereas the German Wikipedia also supplied the name 'Bielendorf', as the town was known until 1945. Similar cases of location renaming abound all over Europe.

WikiID	Alternate name	Source	Type
91643	Bielice, Lower Silesian Voivodeship	English Wikipedia	mainttl
91643	Bielice	English Wikipedia	strippedttl
91643	Bielendorf	German Wikipedia	redirect

Table 3.5: Alternate names of the town of Bielice, in the Lower Silesia Province in Poland, according to the English resource.

And yet, the confidence that a name truly refers to an entity is not as strong in these cases, as a lot of noise can be introduced in this manner. Therefore, I propose three levels of confidence into which each alternate name falls:

- **Strong confidence** for exact matches[1] between the identified toponym and a

[1]Yet, some small variations are allowed, such as presence of absence of a dash, 'ß' matching 'ss', 'ä' matching 'ae' or 'St.' matching 'Saint'.

main title or an alternate name extracted from the Wikipedia version of the same language as the collection.

- **Medium confidence:** for exact matches[1] between the identified toponym and an alternate name extracted from GeoNames or a Wikipedia version in a different language than that of the collection.

- **Weak confidence:** for some partial matches between the identified toponym and any alternate name, such as when cardinal points are removed (e.g. 'Südchina' matching 'China') or when other kinds of political, administrative, and geographic words such as 'city', 'river', or 'province' are either removed or added.

At this stage, for each document I have a set of identified toponyms and, for each, the set of possible candidate entities that may refer to each of them. In the next section, I describe the approach to decide which of the candidates is the most likely to refer to each toponym in the text.

3.3.3 Toponym disambiguation

This is a weakly-supervised method which involves some feature combination and parameter tuning. The features used for the disambiguation are described in the following sections. I have considered two different kinds of features, local and global, which measure the compatibility between the mention and an entity and the compatibility between entities. They exploit both geographic and semantic knowledge.

3.3.3.1 Local disambiguation features: mention to mention compatibility

Local features measure the compatibility of a candidate location to be the referent of a toponym without regard to its compatibility with co-occurring toponyms. I compute three different local features for each candidate: **(1)** historical context similarity, **(2)** geographic closeness to the base location, and **(3)** relevance of the candidate. They are explained in detail in the following paragraphs.

[1] Ibid.

Historical context similarity Context similarity measures the similarity between a mention of a toponym and a candidate entity according to context information, and is common in entity linking approaches. Historical context similarity is a modification of it, and estimates the likelihood between the context of a toponym in the text and the historical context of the candidate. The intuition behind the usefulness of this feature is the following: some locations are known for something in particular or for some historical event. The following example, taken from the Belgian collection,[1] is explanatory of this:

(10) *A new series of nuclear weapons trials will be conducted at a high altitude over the island of Johnston in the Pacific in four days. The first explosion, which is supposed to be a little less than a megaton, will take place at a distance of ten kilometers, the two following in the ionosphere, i.e. at an altitude of about one hundred kilometers. The first of these two bombs will have an explosive force of a megaton, while the explosive force of the second one will be slightly below a megaton.*

Particularly interesting is the case of the toponym 'Johnston', for which seven candidates exist in $GeoSemKB_{en}$: Johnston Atoll (an island in the Pacific ocean), which in this case is the correct referent, and six more: a city in Iowa, a town in Rhode Island, a town in South Carolina, a community in Pembrokeshire, a community in Pennsylvania, and a suburb of an Australian city. The context from the Wikipedia article on the Johnston Atoll leaves little room for doubt that this is the most suited referent in this example. These are some examples of context words taken from Wikipedia fragments associated with the 1950s and the 1960s: 'nuclear', 'weapons', 'trials', 'conducted', 'altitude', 'pacific', 'explosion', 'megaton', 'explosive', 'force', etc. During the years of the Belgian collection (1955–1964), the Johnston Atoll was used a base for testing nuclear and biological weapons, and this newspaper article is representative of that.

Example 11, from the War of the Rebellion corpus, is also illustrative of the usefulness of this feature. The historical context for the correct referent of Fort Monroe contains the following words: 'Butler', 'major', 'general', 'war', 'headquarters', 'army', '1864', and 'Richmond'. Fort Monroe, Virginia, was the site of an important episode during the American Civil War.

[1]Translated here into English for illustrative purposes. In the experiments, the original language of the data was always respected.

(11) *BENJ. F. BUTLER, Major-General, Commanding. SECRETARY OF WAR. HEADQUARTERS EIGHTEENTH ARMY CORPS, Fort Monroe February 12, 1864. GENERAL: I have the honor to forward to you with commendation the report of Brigadier-General Wistar of his brilliantly and ably executed movement upon Richmond which failed only from one of those fortuitous circumstances against which no foresight can provide and no execution can overcome.*

I compute similarity between the context of the toponym and that of the candidate based on the bag of words model, where a toponym t is represented as a vector T of its contextual content words in the text[1] and candidate location c is represented as a vector C of the words from the historical contexts (words from the historical fragments that coincide in time with the period when the text was written) extracted from the Wikipedia article as explained in subsection 3.2.3.5:

$$CS(t,c) = \frac{T \cdot C}{|T||C|} \tag{3.1}$$

This is then normalized over the set K of all candidate locations that can refer to t:

$$CSnorm(c) = \frac{CS(t,c)}{\sum_{k \in K} CS(t,k)} \tag{3.2}$$

Closeness to base location This feature requires a piece of information that is most of the times available in historical collections: the base location of the publication. In this thesis, base location is described as the location (approximated to either the centroid or the capital city, if a larger location such as a province or country is given) of the expected audience, which often coincides with the origin of the publication. This is not the first approach that takes into consideration the source of the text to be disambiguated: Buscaldi and Magnini (2011) (24) worked with very local collections and realized that adding information about the base location was of great assistance for the disambiguation task, as most of the locations were found to be in the vicinity of the source of the texts. Also in Lieberman et al. (2010) (60) is information about the origin of the source included in the method, as it is stated that different audiences from different places have different shared knowledge.

I assigned a base location to each of the toponym resolution corpora. The Belgian corpus consists of articles from the *St. Vither Volkszeitung*, which is based in the Belgian

[1]The window size is set to 50 words, as in Pedersen et al. (2005) (74).

municipality of St. Vith and is aimed at a local audience (there are international news, but also very local ones). Having no information about which newspapers constituted the three Dutch corpora (`Antilles`, `EastIndies`, and `DRegional`), the locations of the (then) capitals of the Netherlands Antilles, the Dutch East Indies and the Netherlands are taken as base locations. The `Prussian` corpus contains two newspapers based in Berlin, then the capital of Prussia. Finally, the `WOTR` corpus has documents from all over the United States, so I pick the location that is the closest to the geographic center of the United States: Lebanon, Kansas. These are the following Wikipedia URLs for the base locations of the different corpora:

- **Prussian corpus:** Berlin, `http://de.wikipedia.org/wiki/Berlin`
- **Belgian corpus:** St. Vith, `http://de.wikipedia.org/wiki/Sankt_Vith`
- **Antilles corpus:** Amsterdam, `http://nl.wikipedia.org/wiki/Amsterdam`
- **EastIndies corpus:** Jakarta, `http://nl.wikipedia.org/wiki/Jakarta`
- **DRegional corpus:** Willemstad, `http://nl.wikipedia.org/wiki/Willemstad_(Curaçao)`
- **WOTR corpus:** Lebanon, `http://en.wikipedia.org/wiki/Lebanon,_Kansas`

The distance between each candidate and the base location is calculated with the great-circle distance algorithm, which measures the shortest distance between two points on a sphere measured along the surface. Even if the Earth is not a perfect sphere, the great circle distance is often used to calculate the distance 'as the crow flies' between two points on Earth. I calculate the great-circle distance d between two locations a and b with the Haversine formula[1] (see equation 3.3).

$$d(a,b) = 2 \cdot r \cdot arcsin\sqrt{sin^2\left(\frac{\Delta lat}{2}\right) + cos(lat_1) \cdot cos(lat_2) \cdot sin^2\left(\frac{\Delta lon}{2}\right)} \tag{3.3}$$

where r is the radius of the Earth (6,371.2 kilometers), lat_1 and lat_2 are the latitudes on Earth of the two comparing locations, lon_1 and lon_2 are the longitudes on Earth of the two comparing locations, Δlat is the absolute difference between the latitudes of both points, and Δlon is the absolute difference between the longitudes of both points. Latitude and longitude must be expressed in radians. Thus, for each candidate, its distance with regard to the base location of the collection is calculated following equation 3.3. I define two distance degrees:

[1] Attributed to Mendoza y Ríos (1805) (34) by Cajori (1928) (27), the Haversine formula was widely used in navigation and is still nowadays widely used for computing distances between two points on Earth.

- **near** if both the base location and the candidate belong to the same country or the distance between them is smaller than the radius of the country area where the base location is situated, approximated as a circle:[1] given the area of the country where the base location is situated, the radius is calculated as $\sqrt{\frac{Area}{\pi}}$. The radius is the distance in kilometers from the base location inside which a candidate location is considered to be near. The reason why not only candidate locations that fall inside the same country as the base location are considered *near* is two-fold: on the one hand, whereas coordinates exist for all locations, not all locations have information about the country they belong to; on the other hand, they need to take into consideration bordering areas between countries (e.g. a person from the German city of Saarbrücken is likely to have a good knowledge of locations in the neighboring French region of Lorraine). The closeness feature of candidates that fall inside the *near* distance degree is 1.0.

- **far** if the distance between the candidate and the base location fall outside the afore-mentioned circle and both locations are not in the same country. The closeness of candidate locations that are considered as *far* is calculated as the distance in kilometers to the base location, reverse normalized against the sum of distances from all the candidates of the toponym:

$$closeness(c) = 1 - \frac{d(c, baseloc)}{\sum_{k \in K} d(k, baseloc)} \tag{3.4}$$

  where c is the candidate and *baseloc* the base location, and K is the set of all candidate locations of the toponym.

Relevance Toponym resolution systems have traditionally used **population** as a measure to determine the relevance of a location. It has been used either as part of the disambiguation procedure (Rauch et al. (2003) (82), i.a.) or as a strong baseline (DeLozier et al. (2015) (35), i.a.). Entity linking methods usually measure relevance of an entity by its **number of inlinks** (also known as incoming links or backlinks), which indicates how many articles mention this entity. In this thesis, the latter is preferred to population for a series of reasons, discussed below.

[1] This is of course a very loose approximation (more in some countries, like Chile, than others, but serves as a normalization against the true dimensions of a country to define a 'close-by' distance).

Population as a measure for relevance is faulty: some locations are unquestionably relevant but have no population associated with them, as is the case of oceans, seas, the Everest, or Antarctica, to mention just a few. Given the ambiguous toponym 'Pacific', for example, population indicates that the most relevant candidate to be the referent of such mention is a city in Missouri named 'Pacific' that has a population of 7,002 inhabitants, since the Pacific Ocean has a population of 0. It could be argued that in general — or, at least if we were looking just at populated places — population is a better choice because it is an objective measure, whereas the number of inlinks taken from the Wikipedia articles is a measure that naturally imbibes the systemic bias of Wikipedia. However, this bias (in this case of a geographic nature) is not necessarily a negative aspect, since it may be seen as to reflect the world as viewed by one particular community. To exemplify this with an example, the toponym 'Waterloo' has a different meaning for the citizens of Canada than for the citizens of Belgium. Apart from the Battle of Waterloo (which is the most relevant sense for 'Waterloo' in both the English and Dutch versions of Wikipedia), there are at least two prominent locations that are known as 'Waterloo': one in Ontario, Canada, and the other one in Belgium (after which the battle was named). The English Wikipedia considers the Canadian Waterloo about five times as relevant as the Belgian Waterloo, whereas the Dutch Wikipedia considers the Belgian Waterloo about 7.5 times as relevant as the Canadian location. Relevance from Wikipedia is also a measure that favors places that have had an important role in history (again, seen from the perspective of a certain Wikipedian community). This has already been shown in the case of the Battle of Waterloo. Another example concerns the toponym 'Toledo'. With a population of 84,000 inhabitants, the Spanish city of Toledo is smaller than its homonyms in the Philippines, Brazil and the United States. However, it is the city in Spain the one that has the most inlinks in the English Wikipedia (and also in the German and the Dutch versions), probably due to its administrative and historical prominence in the past, as it once was the de facto capital of the Visigothic Kingdom and capital of Spain until the 16th Century. Having population as a measure for relevance is especially debatable when dealing with historical texts, as none of these bigger cities named 'Toledo' existed prior to 1833.

Relevance of a candidate location c is calculated as the number of inlinks of this location r_c, normalized over the sum of the number of inlinks of the rest of the candidates

K of the toponym:

$$relevance(c) = \frac{r_c}{\sum_{k \in K} r_k} \tag{3.5}$$

Even though measuring relevance by the number of Wikipedia incoming links is not a perfect measure (on the one hand, as mentioned, it is susceptible to be affected by the biases that surround Wikipedia, some of them unwanted; on the other hand, the relevance of locations is also changing throughout the years, and a location that is nowadays popular may not have necessarily been relevant in the past), I am certain that the advantages of using this relevance measure outweigh its disadvantages. This manner of measuring relevance is a very common measure in information retrieval and, in the same way as the most common sense in word sense disambiguation, provides a very strong baseline.

3.3.3.2 Global disambiguation features: entity-to-entity compatibility

Global features take into account the interdependence between entities to measure the compatibility of a candidate. I propose two different global features to compute for each candidate: **(1)** semantic relatedness and **(2)** overall geographic closeness.

Semantic relatedness This feature measures the semantic relatedness between different candidates, regardless of the context of the document. This is to collectively favor locations that were historically related at the time when the text was written. I illustrate this with the following example:

(12) *Marshal Zhukov had reached the middle **Oder** at a number of places between a point 15 miles east of **Kustrin** and **Glogau**. He was within 9 miles of **Frankfurt** and had also advanced the north-west flank of his salient. **Posen** continued to offer stubborn resistance.*[1]

The toponyms in this text refer to the river Oder and the cities Kostrzyn nad Odrą, Głogów, Frankfurt an der Oder, and Poznań, all of them locations that happened to be in the way of the Red Army as it advanced towards Berlin. Their semantic contexts,

[1] **Source:** *The National Archives*, catalogue reference: CAB/65/49/14, feb 1945.

restricted to the 1940s decade, are therefore more similar than those of the rest of the candidates[1] that were not part of the same events.

The semantic relatedness between any two locations a and b is calculated following the measure proposed by Milne and Witten (2008) (67):

$$sr(a,b) = 1 - \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))} \tag{3.6}$$

where A and B are the sets of context words for candidates a and b respectively. As described in subsection 3.2.3.5, the context words of a Wikipedia article are the title of the entry (and, if there is an apposition in the title, both the stripped title and the apposition), the name of the country and the region, and its set of historical words matching the period of the collection. W refers to all the context words in the resource in question. Once this measure has been applied to all pairs of entities, it is possible to calculate the semantic relatedness between a candidate entity of a toponym and the rest of candidate entities of the rest of the toponyms in the document in the following manner:

$$OSR(c) = \frac{1}{|E|} \sum_{e \in E} sr(c,e) \tag{3.7}$$

where c is a potential candidate of toponym t_c, and E is the set of potential candidates of all toponyms other than t_c.[2] This is then normalized and smoothed with Laplace correction:

$$OSRnorm(c) = \frac{OSR(c)+1}{\sum_{e \in E}(OSR(e)+1)} \tag{3.8}$$

Overall geographic closeness This feature measures how close a location is to the rest of candidates of the rest of toponyms and favors locations that minimize this distance. The same example 12 is also illustrative of its usefulness. The mention 'Frankfurt', for example, is geographically closer to the correct disambiguations of 'Kustrin' (now Kostrzyn nad Odrą), Glogau (now Głogów), and Posen (Poznań). This overall

[1] The English resource $GeoSemKB_{en}$ provides 32 different candidates for 'Frankfurt', 2 for 'Oder', 2 for 'Glogau', 2 for 'Kustrin', and 6 for 'Posen'.

[2] If, for example, t_c is the mention 'Frankfurt' and the current candidate c is Frankfurt an der Oder, E would be all the candidates that are potential referents of the toponyms 'Oder', 'Kustrin', 'Glogau', and 'Posen'.

distance is significantly smaller than if the larger and more commonly referred city Frankfurt am Main is taken as candidate instead of Frankfurt an der Oder. However, we do not know which are the correct disambiguations of the other toponyms in the text, and therefore I compute the closeness between a candidate normalized over the set of all the candidates of the rest of the toponyms in the text. The overall geographic closeness between an entity c and a set of entities E is calculated in this manner:

$$OGC(c) = \frac{1}{|E|} \sum_{e \in E} d(c, e) \tag{3.9}$$

where c is a potential candidate of toponym t_c, E is the set of potential candidates of all toponyms other than t_c,[1] and $d(c, e)$ is the geographic distance calculated with the great-circle distance according to equation 3.3. This is reverse normalized and smoothed with Laplace correction:

$$OGCnorm(c) = 1 - \frac{OGC(c) + 1}{\sum_{e \in E}(OGC(e) + 1)} \tag{3.10}$$

3.3.3.3 Feature and parameter combination

The local features are combined and weighted together with the penalization parameter $conf$, which determines the confidence of the candidate being a correct retrieval given a toponym:

$$locComp(c) = \frac{\alpha \cdot CSnorm(c) + \beta \cdot closeness(c) + \gamma \cdot relevance(c)}{\alpha + \beta + \gamma} \cdot conf \tag{3.11}$$

The local compatibility of a candidate given a toponym is then combined with the global features and the most likely candidate is taken as the referent of the toponym, as described in equation 3.12:

$$bestcand = max \left(\frac{\delta \cdot locComp(c) + \epsilon \cdot OSRnorm(c) + \zeta \cdot OGCnorm(c)}{\delta + \epsilon + \zeta} \right) \tag{3.12}$$

A development set is used to tune the different parameters.

[1] Ibid.

3.3.4 Summary of the disambiguation method

In this section, I have described a novel end-to-end toponym disambiguation method, which first identifies the toponyms in a document, finds candidate locations that may be referred to by each toponym, and selects the best referent. The method described is weakly-supervised and particularly suited for the historical domain. The disambiguation step combines the use of geographic and semantic features, which can be classified into two categories: local and global. The local disambiguation features are historical context similarity, closeness to the base location, and relevance, and the global disambiguation features are semantic relatedness and overall geographic closeness. This section ends with the description of how the features are combined with the different parameters that weight the importance of each feature in each dataset.

3.4 Experimental results

As already mentioned, the task of toponym disambiguation has been approached from two different yet closely related tasks: toponym resolution and entity linking. One of the main differences between them is in the format of the returned output. Toponym resolution traditionally returns the geographic coordinates of the disambiguated location, whereas entity linking approaches return the link to the corresponding entry in a knowledge base. Evaluation is clearer in the latter case, as an item is considered correct only if the linking matches the gold standard, and incorrect if it does not. The evaluation of toponym resolution systems based on predicted coordinates is often fuzzy, as there is no standardized dictionary or mapping from locations to their defining coordinates. Therefore, it is possible that a toponym has been correctly disambiguated but is considered incorrect if its latitude and longitude do not coincide with the gold standard coordinates. This problem is most severe in the case of extensive locations such as provinces, countries, rivers, or seas, as different resources do not always agree on their centroid point.[1]

In this section, I start by reviewing several evaluation metrics that have been used in the toponym resolution and entity linking literature and that I also use to test the method proposed in the previous section. I compare its performance against that of the

[1]For instance, Wikipedia provides coordinates for the northern Canadian territory of Nunavut that lie 315 kilometers away from the coordinates provided for this location in GeoNames.

different baselines and state-of-the-art methods, and provide and discuss the evaluation results.

3.4.1 Evaluation metrics

I introduce in subsection 3.4.1.1 the metrics used for evaluating my method from an entity linking perspective and in subsection 3.4.1.2 the metrics for evaluating the method from a toponym resolution perspective. In both cases, as in Leidner (2008) (59), identified toponyms that were tagged as 'unknown' by the annotator have been disregarded. In the WOTR dataset, 'unknown' locations are instances for which the annotators failed to provide coordinates because they were unsure of which place was being referred to in the text, either because of lack of candidates or because of lack of clues that could help them select the correct referent. In any other case, the annotator should have been able to extract coordinates for the toponym, either by retrieving them from a knowledge base or a gazetteer (such as Wikipedia or GeoNames, etc.) or by placing them directly on a map. In the rest of the datasets, 'unknown' locations are, besides the cases just mentioned, also those for which there exists no entry in Wikipedia.

3.4.1.1 Entity linking evaluation metrics

I use the traditional information retrieval measurements, **precision**, **recall**, and **f-score**, to evaluate the method from an entity linking perspective. Entity linking evaluation is very precise: an instance is correct only if the link of the predicted location in a knowledge base matches the gold standard. Since coordinates are extracted afterwards from this same resource, there is no room for evaluation inaccuracy. In these cases, a true positive is a toponym that has been correctly identified and resolved. If a toponym has been incorrectly identified, it is considered a false positive; whereas a missing toponym counts as a false negative. Finally, if a toponym has been correctly identified but incorrectly resolved, it counts both a false negative and a false positive. Precision and recall are then calculated as follows:

$$Precision = \frac{truepositives}{truepositives + falsepositives} \tag{3.13}$$

$$Recall = \frac{truepositives}{truepositives + falsenegatives} \tag{3.14}$$

And F-score is then calculated as the harmonic mean between the two:

$$Fscore = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.15}$$

3.4.1.2 Toponym resolution evaluation metrics

Several evaluation metrics have been proposed to date to overcome the problem of inexactness in toponym resolution evaluation. If **precision**, **recall**, and **f-score** are to be used, they need to be approximated. I use the same approximation as proposed in Speriosu and Baldridge (2013) (91), who define a true positive to be 'the prediction of a correctly identified toponym's location to be as close as possible to its gold label, given the gazetteer used'. False negatives and false positives are defined as in the section above, and precision, recall, and f-score are calculated following the equations 3.13, 3.14, and 3.15 respectively.

Another metric that has often been used in the literature is that of **accuracy**, which can be defined as the fraction of toponyms that have been correctly resolved given a gold standard. Leidner (2008) (59) provides the formula $Acc = T_C/T_N$, where T_C corresponds to the number of correctly resolved toponyms, and T_N to the total number of resolvable toponyms provided by the gold standard. As mentioned, there is no such thing as a standardized dictionary of coordinates for locations: each gazetteer provides its own set of latitude and longitude coordinates for each location. It is therefore possible that the coordinates produced by the system do not match the gold standard coordinates and yet describe the correct location, if different sources are used to extract the geographic reference. This problem is not as grave in small locations (e.g. villages) as it is in larger locations (e.g. regions, countries, continents). Cheng et al. (2010) (30) were the first to introduce **accuracy@k**, an accuracy metric that allows for some fuzziness in the accepted pair of coordinates, where k is the number of kilometers inside which error is allowed. Recent approaches have set k to 161 (roughly 100 miles), thought to be an approximation to what can be considered a 'large metropolitan area of correct', as described in Speriosu (2013) (91). Besides, I also provide results for accuracy@k when k is set to 16 (roughly 10 miles).

Finally, the **mean error distance** and **median error distance** calculate the mean and median error distance in kilometers between the predicted coordinates of a toponym and the gold standard coordinates. They have been used by Eisenstein et

al. (2010) (39), Wing and Baldridge (2011) (99), Roller et al. (2012) (86), DeLozier et al. (2015) (35) and the first of these two metrics is also provided for the WOTR corpus in DeLozier et al. (2016) (36).

3.4.2 Baselines

I provide three baselines common to both location linking- and toponym resolution-based evaluations:

1. The ***Random* baseline** picks randomly one out of the possible candidates extracted for each toponym. Therefore, if a toponym has one only candidate, the correct candidate will obviously be selected. The reason behind using this baseline is to understand the nature of the corpus and of the resource used to extract the candidates. This baseline is averaged over three trials, as in Speriosu and Baldridge (2013) (91) and DeLozier et al. (2016) (36).

2. The ***MaxPopulation*** returns the candidate that has the largest population. This is a common baseline in toponym resolution.

3. The ***MaxInlinks*** returns the candidate that has the most inlinks in the Wikipedia version of the same language as the collection. This is a common baseline in entity linking approaches. It is considered to be analogous to the most common sense in other sense disambiguation tasks and it has proved to be a very strong baseline.

The three baselines select one candidate from the pool of candidate locations that are potential referents of a given toponym. In order to obtain this pool, I used the same candidate selection procedure that I proposed for my method, even though it is evident that the performance of the three baselines can change greatly depending on this step.

3.4.2.1 Entity linking comparing methods

I provide results for two out-of-the-box state-of-the-art entity linking systems: ***DBpedia Spotlight***, introduced in Daiber et al. (2013) (33) and ***Babelfy***, described in Moro et al. (2014) (69). I opted for these two systems because they are also competitive for languages other than English, among which German and Dutch, and because

they can be used off the shelf, which reduces the risk that errors are made in the re-implementation of the method.[1] DBpedia Spotlight and Babelfy do not only resolve locations: they both return also different kinds of resolved entities (such as people and organizations) and the latter returns also some non-entities. To avoid these off-target entities that are identified by these methods to unfairly count as false positives, during the evaluation of these methods I only consider as identified toponyms those mentions for which coordinates could be extracted from the predicted entity.

3.4.2.2 Toponym resolution comparing method

Finally, I also report results for a state-of-the-art toponym resolution system on the WOTR dataset, ***TopoClusterGaz***, described in DeLozier et al. (2015) (35), who achieved state-of-the-art performance in two standard toponym resolution datasets. This paper is the only one for which results on the WOTR dataset are provided.

3.4.3 Experimental settings

I provide results on the six datasets described in section 3.1 and summarized in table 3.1: *Belgian*, *Prussian*, *Antilles*, *EastIndies*, *DRegional*, and *WOTR*. There are three datasets in Dutch, two in German, and one in English. Two metadata fields are required: the decades the collection spans and the base location (see table 3.6), which are two pieces of information that are usually known in historical collections. The base location is given as the Wikipedia URL of the entity referring to it in the same Wikipedia version of the language of the collection. Coordinates are then derived from it via the *GeoSemKB* resources.

Collection	Decades	WikiURL of base location
Prussian	1860s–1890s	`http://de.wikipedia.org/wiki/Berlin`
Belgian	1950s–1960s	`http://de.wikipedia.org/wiki/Sankt_Vith`
Antilles	1920s–1990s	`http://nl.wikipedia.org/wiki/Amsterdam`
EastIndies	1910s–1940s	`http://nl.wikipedia.org/wiki/Jakarta`
DRegional	1800s–1990s	`http://nl.wikipedia.org/wiki/Willemstad_(Curaçao)`
WOTR	1860s	`http://en.wikipedia.org/wiki/Lebanon,_Kansas`

Table 3.6: Summary of metadata for each dataset.

[1] Entity linking systems are typically difficult to re-implement, as there are many factors and parameters that may alter greatly the overall result. Besides, results are often difficult to replicate due to fact that they usually rely on resources that are constantly being updated.

This is a weakly-supervised method which requires combining features and tuning parameters. Each corpus has been split into two sets: a **development set** and a **test set**. Unlike traditional supervised classification, not a large amount of data is needed to learn feature combinations and parameter tuning, and therefore I have split the data in a way that 40% is used for development and the remaining 60% is set apart for evaluation. Table 3.7 shows the number of documents in the development and test set of each corpus.

Collection	Development set	Evaluation set	Total
Prussian	65	98	163
Belgian	94	143	237
Antilles	38	58	96
EastIndies	40	60	100
DRegional	88	134	222
WOTR	215	322	537

Table 3.7: Number of total documents for each corpus, divided into development and evaluation set.

The documents were randomly classified into each set. Note that I use the same test set of the WOTR corpus used to evaluate *TopoClusterGaz* in DeLozier et al. (2016) (36). They had a different distribution of the data: 80% was used for training and 20% for testing. In order to have comparable datasets, I randomly selected the 40% of the evaluation set size from the training set, which results in a development set of 215 documents.

The optimal parameter settings are determined through experimentation with the development set. Table 3.8 shows the optimal parameter combination for each dataset. The penalization parameter $conf$ was set to 1.0 for strong confidence cases, 0.66 for medium confidence cases, and 0.33 for weak confidence cases.

3.4.4 Results and discussion

My method is exposed to the two types of evaluation: toponym resolution evaluation (in subsection 3.4.4.1) is conducted on the WOTR dataset and location linking evaluation (in subsection 3.4.4.2) is performed on the rest of the datasets. The reason behind using a different type of evaluation for the English dataset than for the rest of the German and Dutch datasets is not linguistically-motivated, but stems from the nature of the

	α	β	γ	δ	ϵ	ζ
Belgian	0.25	0.75	1.00	1.00	0.25	0.25
Prussian	0.25	0.75	1.00	1.00	0.75	0.50
Antilles	0.50	0.75	0.75	1.00	0.50	0.50
EastIndies	0.50	0.75	1.00	1.00	0.50	0.50
DRegional	0.50	0.25	0.75	1.00	0.25	0.50
WOTR	0.25	1.00	0.25	1.00	0.50	0.50

Table 3.8: Optimal parameter weights.

datasets. The German and Dutch datasets were annotated with Wikipedia URLs and can be compared against other entity linking systems that also provide Wikipedia URLs for their predicted cases, whereas the English dataset was annotated with geographic coordinates. The results are analyzed and discussed in subsection 3.4.4.3.

3.4.4.1 Toponym resolution evaluation

Table 3.9 shows the results of applying several different methods on the WOTR dataset, the only one considered that does not provide links to Wikipedia URLs, but to geographic coordinates. The metrics used to test the performance of the different approaches are approximate precision, recall, and f-score (where a true positive is the closest possible candidate to the gold standard), accuracy at 161 kilometers and at 16 kilometers (i.e. roughly 100 and 10 miles respectively), and mean and median error distance. To be fully comparable to the results reproduced by DeLozier et al. (2016) (36), the first three metrics test the whole method's procedure, from plain text to the predicted locations, whereas the remaining four metrics were calculated on texts where toponyms had been manually recognized. Results are provided for seven different methods: apart from my own method, which I henceforth refer to as *GeoSem*, I provide results for the three afore-mentioned baselines (*Random*, *MaxPopulation*, and *MaxInlinks*), the toponym resolution state-of-the-art comparing method *TopoClusterGaz*, and the two baselines provided by DeLozier et al. (2016) (36), authors of the WOTR corpus: a random baseline (*RandomTCG*) and a baseline based on selecting the candidate with the most population (*MaxPopTCG*).

My method, *GeoSem*, substantially improves on the state-of-the-art system provided for comparison (*TopoClusterGaz*) in approximate precision, recall, and f-score, its mean error distance is 23 kilometers smaller, and it also beats the different baselines. Accuracy@161 (according to which locations that fall inside an error distance of

	$\sim Precision$	$\sim Recall$	$\sim F_1S$	$Acc@161$	$Acc@16$	$Mean$	$Median$
Random	0.25	0.26	0.26	0.33	0.26	1711	700
MaxPopulation	0.51	0.52	0.51	0.62	0.44	913	15
MaxInlinks	0.52	0.54	0.53	0.65	0.49	723	10
RandomTCG	0.15	0.06	0.09	0.22	—	2216	—
MaxPopTCG	0.42	0.18	0.25	0.63	—	1483	—
TopoClusterGaz	0.38	0.31	0.34	**0.72**	—	468	—
GeoSem	**0.56**	**0.58**	**0.57**	0.68	**0.52**	**445**	**7**

Table 3.9: Toponym resolution evaluation on the WOTR dataset.

161 kilometers from the gold standard are considered correct predictions) is the only metric for which my method produces lower results than *TopoClusterGaz*. Even though DeLozier et al. (2016) do not provide results in terms of accuracy at 16 kilometers and in terms of median error distance, I provide them nonetheless, as they may be helpful to understand the behavior of the methods and be useful to future research.

It is interesting to note the difference between the two random methods and the two population methods. *Random* takes a random item from the set of candidates retrieved with the knowledge base created for the *GeoSem* method, which is described in section 3.2, whereas *RandomTCG* randomly selects one from the set of candidates extracted for the *TopoClusterGaz* method, based on GeoNames alternate names. Both baselines provided by DeLozier et al. (2016) (36) perform significantly worse in terms of most of the metrics (and especially recall) than their equivalent baselines provided here. Some speculations can be ventured on what may cause this difference. The random baseline is a good indicator of the difficulty of the disambiguation step, but also of the size of the used gazetteer or resource. One reason why their random baseline's performance is lower has to do with the fact that GeoNames has a much wider geographic knowledge than Wikipedia, and therefore the average in number of candidates per toponym is significantly higher than in the $GeoSemKB_{en}$ resource. However, the fact that both the *RandomTCG* and *MaxPopTCG* baselines report a much lower recall than precision and high accuracy indicates an inferior ability to recognize toponyms in plain text. Finally, it is important to note that the relevance measure based on number of inlinks from Wikipedia produces better results than a relevance measure that uses population as a criterion.

3.4.4.2 Location linking evaluation

In the location linking evaluation (as in the more general entity linking evaluation), the items to compare are entries in a knowledge base, and not coordinates, which are often assigned in a somewhat arbitrary fashion for larger geographical areas. I report results on the two German datasets, *Belgian* and *German*, in tables 3.10 and 3.11 respectively, and results on the three Dutch datasets, *Antilles*, *East Indies*, and *DRegional*, in tables 3.12, 3.13, and 3.14 respectively. The metrics used to compare the different methods are precision, recall, and f-score (where a true prediction is an exact match with the gold standard), accuracy at 161 and at 16 kilometers, and the mean and median error distances. Contrary to the accuracies and error distances reported in table 3.9 for the WOTR dataset, here they are provided as end-to-end results, from the plain text to the predicted locations. I compare the performance of my method (*GeoSem*) with the three baselines (*Random*, *MaxPopulation*, and *MaxRelevance*) and the two state-of-the-art systems (*DBpedia Spotlight* and *Babelfy*) described above. Both for *DBpedia Spotlight* and *GeoSem*, the parameters were optimized for each collection with the development set.

	Precision	*Recall*	F_1S	*Acc*@161	*Acc*@16	*Mean*	*Median*
Random	0.48	0.39	0.43	0.52	0.49	1011	0
MaxPopulation	0.63	0.51	0.56	0.60	0.57	533	0
MaxInlinks	**0.76**	0.62	**0.68**	0.66	0.63	49	0
DBpedia Spotlight	0.63	**0.63**	0.63	0.65	0.63	133	0
Babelfy	0.62	0.55	0.58	0.61	0.59	32	0
GeoSem	0.75	0.62	**0.68**	**0.67**	**0.64**	**23**	0

Table 3.10: Location linking evaluation on the Belgian dataset.

	Precision	*Recall*	F_1S	*Acc*@161	*Acc*@16	*Mean*	*Median*
Random	0.37	0.39	0.38	0.52	0.47	1286	0.4
MaxPopulation	0.64	0.66	0.65	0.75	0.73	269	0
MaxInlinks	**0.72**	**0.75**	**0.73**	**0.78**	0.77	54	0
DBpedia Spotlight	0.53	0.56	0.55	0.64	0.63	94	0
Babelfy	0.64	0.62	0.63	0.67	0.67	**23**	0
GeoSem	0.71	**0.75**	**0.73**	**0.78**	**0.78**	65	0

Table 3.11: Location linking evaluation on the Prussian dataset.

In general, the *GeoSem* clearly outperforms the two state-of-the-art systems, whose lower performance may be explained by the fact that they attempt to resolve not only

locations, but all kinds of entities. Because of that, they do not exploit any geographic-specific features in the disambiguation step. This is possibly one of the reasons for their lower performance, another one being a consequence of it: both systems had the added difficulty of having to choose among a more diverse pool of candidates, not all of which locations.[1] This could easily explain why a baseline based only on relevance of the toponym candidates (*MaxInlinks*) performs much better than them.[2]

Baselines based on the most common sense have proved to be very strong in sense disambiguation tasks. Both entity linking and toponym resolution are disambiguation tasks, and it is therefore not strange that the performance based on the most common sense (i.e. the most relevant sense) is high. It is interesting to note that the relevance baseline based on the number of inlinks extracted from Wikipedia returns significantly better results throughout the different collections than the one based on population, which confirms the intuition that it is a much stronger indicator of relevance than population. On the whole, *MaxInlinks* and *GeoSem* present similar performances on the two German datasets, with the accuracies being slightly higher for the latter, but precision being better for the first. The *MaxInlinks* baseline has the best results in terms of precision, recall, and f-score for the three datasets in Dutch, whereas my method is competitive in terms of the two accuracy measures.

	Precision	*Recall*	F_1S	*Acc*@161	*Acc*@16	*Mean*	*Median*
Random	0.44	0.41	0.42	0.46	0.45	2282	0
MaxPopulation	0.62	0.59	0.60	0.67	0.65	317	0
MaxInlinks	**0.72**	**0.68**	**0.70**	0.70	0.68	**113**	0
DBpedia Spotlight	0.41	0.56	0.48	0.70	0.67	145	0
Babelfy	0.44	0.55	0.49	0.57	0.57	267	0
GeoSem	0.66	**0.68**	0.67	**0.71**	**0.70**	169	0

Table 3.12: Location linking evaluation on the Antilles dataset.

[1]However, as already mentioned, only elements for which coordinates could be extracted were considered for evaluation. For instance, a linking from a mention to a person entity does not count as a false positive even if the mention does not appear in the gold standard.

[2]Take the mention 'Churchill', for instance. If only locations are considered, the system needs to choose among the different locations that may be known as 'Churchill' (20 according to the $GeoSemKB_{en}$ resource, 36 according to GeoNames). If other kinds of entities are also considered, the pool of candidates becomes larger to include people and other kinds of entities.

	Precision	*Recall*	F_1S	*Acc*@161	*Acc*@16	*Mean*	*Median*
Random	0.40	0.38	0.39	0.56	0.52	1383	0
MaxPopulation	0.55	0.52	0.54	0.65	0.62	402	0
MaxInlinks	**0.70**	**0.67**	**0.69**	**0.71**	0.69	110	0
DBpedia Spotlight	0.51	0.60	0.55	0.66	0.65	**44**	0
Babelfy	0.52	0.55	0.54	0.62	0.62	506	0
GeoSem	0.67	**0.67**	0.67	**0.71**	**0.70**	157	0

Table 3.13: Location linking evaluation on the EastIndies dataset.

	Precision	*Recall*	F_1S	*Acc*@161	*Acc*@16	*Mean*	*Median*
Random	0.42	0.32	0.36	0.48	0.44	1149	0
MaxPopulation	0.55	0.43	0.48	0.57	0.55	291	0
MaxInlinks	**0.72**	**0.56**	**0.63**	0.59	0.59	45	0
DBpedia Spotlight	0.51	0.54	0.53	**0.68**	**0.66**	41	0
Babelfy	0.52	0.54	0.53	0.59	0.59	67	0
GeoSem	0.69	**0.56**	0.62	0.59	0.59	**17**	0

Table 3.14: Location linking evaluation on the DRegional dataset.

3.4.4.3 Discussion

The results reported above hint that the resource from which candidates and their features are extracted play an important role in the overall performance. After analyzing several prominent entity linking approaches, Hachey et al. (2013) (49) reported that candidate selection accounts for most of the variation between the different examined systems. I have already discussed why I preferred Wikipedia over GeoNames as the base upon which to build the resources that would later be used to assist in the disambiguation of toponyms, even though the geographic coverage of the latter is much wider. In the following lines, I show how this difference does not necessarily imply a negative impact when using the smaller resource. In order to do so, I analyze the performance of both resources on the test set of the WOTR corpus, which is the only one that has been annotated with coordinates taken from different sources and is therefore the one that can assess more objectively the coverage of both resources to compare: GeoNames and my Wikipedia-based resource for English, $GeoSemKB_{en}$. I extracted candidates for each toponym in the WOTR corpus both with $GeoSemKB_{en}$ and GeoNames. In the first case, I have explained in detail how alternate names were found in subsection 3.2.3.3 and how these were used to select candidates for each toponym in subsection 3.3.2. In the case of GeoNames, I have used the same strategy to select candidates, but I extract potential location names from the GeoNames 'alternatename' table.

	Accuracy@161	**Accuracy@16**
Oracle GeoSemKB$_{en}$	0.81	0.64
Oracle GeoNames	0.75	0.61

Table 3.15: Performance of the oracle methods based on *GeoSemKB$_{en}$* and GeoNames.

The random baseline based on GeoNames has a 19% chance to select the best candidate randomly, whereas *GeoSemKB$_{en}$* has a 33% chance, in both cases averaged over three trials. This means that the number of candidates for each toponym is, on average, substantially higher in GeoNames. This is hardly surprising, as it has about ten times the number of locations Wikipedia has. In order to determine to what degree this difference of candidates per toponym affect the overall performance, I provide accuracy at 161 and at 16 kilometers of an **oracle method**, which selects the candidate that is the closest (in terms of geographic distance) from the gold standard. Interestingly enough, even if the geographic coverage of Geonames is much wider than that of Wikipedia, the English resource based on Wikipedia (and complemented with knowledge from GeoNames) achieves better oracle results than GeoNames, as shown in table 3.15.

An analysis of incorrect or missing GeoNames predictions revealed that they mainly stem from two different sources: **(1)** it is a location that no longer exists and GeoNames does not have an entry for it (such as Camp Wright, Fort Union, La Libertad, Camp Lincoln, or Texana; all old settlements, posts, or towns that do not exist as such anymore), or **(2)** the correct location could not be retrieved from the toponym (e.g. 'Mississippi' is not an alternate name for the Mississippi River in GeoNames). The first type of error source had a smaller impact in *GeoSemKB$_{en}$*, and the second was almost not present. Finally, there was another source of considered incorrect predictions, which in this case affected both resources, and which is due to the inherently fuzzy evaluation of toponym resolution by coordinates: some locations were actually correctly identified, but their coordinates fell more than 161 kilometers away from the gold standard. This is the case of the United States in the *GeoSemKB$_{en}$* resource (whose coordinates point at 174 kilometers away from the gold standard) and Alabama in GeoNames (which is 212 kilometers away from the gold standard), to mention just two examples.

The selection of candidates possibly affects the performance of location linking approaches, as the German and Dutch resources $GeoSemKB_{de}$ and $GeoSemKB_{nl}$ enlarge the list of alternate names of a location with names extracted from GeoNames and other Wikipedia language versions, whereas the candidate extractors from DBpedia Spotlight and Babelfy seemingly do not. This had more impact in some collections than in others. An 11% of the toponyms of the *EastIndies* datasets could not be possibly matched to the correct location only by using the Dutch version of Wikipedia when using the traditional mechanisms to obtain alternate names, such as title stripping or page redirection,[1] whereas in the case of the *Belgian* collection it was only a 5% of the toponyms that could not be possibly retrieved by only using the German version of Wikipedia.

The *GeoSem* method does not require a very large amount of data to be annotated, and yet it would be desirable, for the sake of usability, that no annotated data were needed at all. Unfortunately, as is usually the case, each collection has different characteristics that render it unique, and a method's configuration that is optimal for one collection very often does not perform quite as well in a different one. Some relation might be expected to exist between the characteristics of a corpus and the optimal parameter values. The **WOTR** collection, for example, is very US-centric, as it mentions very few locations from beyond the borders of the United States. It mostly consists of military reports and letters that describe military moves and events that took place during the American Civil War and which often did not occur in relevant or well-known places (often, the reports follow the movements of an army through a territory and mention the different locations that are found on its way). It is therefore understandable that the optimal local distance parameter β is 1.0, as it favors places inside the same country. It is likewise hardly surprising that the relevance parameter γ is lower than for the rest of the datasets, as relevance is a less decisive factor to disambiguate toponyms in military reports than it is in newspapers. The relevance parameter is high in the rest of the datasets, which is not strange, considering that all of them (even the most regional ones, like the Belgian collection) cover international news, for which often a common shared knowledge is expected from the reader. The **Prussian** collection stands out for having locations in its news articles that are more

[1]From them, a 50% could be extracted from redirect links on the English Wikipedia, a 6% from redirect links on the German Wikipedia, and a 39% from alternate names transferred from GeoNames.

semantically related than in the rest of the collections, which might be due to the fact that the newspapers it consists of were very close to the Prussian government and were therefore very Prussian-centric. On the other hand, it is also one of the most ancient collections, and most of the then Prussian locations share a similar historical background. The **DRegional** has the lowest local distance parameter β, which is possibly due to the fact that this collection consists of several very regional newspapers for which we did not know their real base location. I provided an approximate base location to the corpus: the capital of the Netherlands, Amsterdam. However, very regional newspapers aim at a very specific readership, and it is therefore not surprising that a feature that favors any location in the Netherlands does not provide much assistance to disambiguate toponyms in very regional news articles. The **Belgian** collection is the most modern one. It is a regional newspaper with base in the Belgian municipality of St. Vith, but which also has a wide coverage of international news. These two characteristics might explain the low geographic and semantic relatedness among the different locations in the articles. Finally, the two colonial collections, **Antilles** and **EastIndies**, have a combination of high local context similarity and a high local distance parameter β. Colonies often have locations named after places of the colonizing country. This may explain why the distance parameter is high but not the highest possible (as most locations mentioned are in the colonies, but locations in the Netherlands are greatly penalized due to the large distance from the base location) and why the local context similarity is higher than in other datasets (as context often plays an important role in specifying whether a toponym refers to the location in the colony or to its namesake in the colonizing country).

Further analysis on the importance of relevance for disambiguating toponyms showed that this measure is often determining if relevance is highly unequal among the different candidates, but not so much otherwise. The more similar the relevance between the candidates is, the more the method needs to rely on other features. This is clearly seen through an example: given the toponym 'Berlin', the likelihood that it refers to the most relevant candidate (the capital of Germany, which has 7,156 inlinks according to $GeoSemKB_{en}$) is much higher than that it refers to the second most relevant candidate (the city now known as Kitchener, in Ontario, which has 354 inlinks); whereas given the toponym 'Springfield', the likelihood that it refers to the most relevant candidate (the capital city of Illinois, which has 782 inlinks) is similar to the likelihood that it

refers to the second most relevant one (a city in Massachusetts, which has 764 inlinks). This can be another reason why relevance in the WOTR dataset plays a lesser role than in the other datasets, as for the 76% of the toponyms in the WOTR corpus the candidate with the largest number of inlinks has more inlinks than the sum of inlinks of the rest of the candidates, whereas in other collections this number ranges from 84% to 97%.

Finally, the confidence that a given toponym may refer to a location is not always the same, as discussed in subsection 3.3.2. Because of that, I had defined three different levels of confidence: weak, medium, and strong, whose effect to the overall performance of the method is regulated by means of the $conf$ parameter with a view to minimize a possible negative impact. In some of the datasets, the confidence parameter proved to have an impact on the final results: without taking it into account, the method's performance dropped by four points in f-score when tested on the Prussian dataset, two points on the Antilles and the WOTR datasets, and one point on the Belgian dataset.

3.5 Summary

In this chapter, I have introduced *GeoSem*, a weakly-supervised toponym disambiguation method that combines the strengths of toponym resolution and entity linking systems. I have analyzed its performance on several datasets: one standard benchmark dataset for toponym resolution in English (*WOTR*) and five datasets that were created from scratch in German (*Prussian* and *Belgian*) and Dutch (*Antilles*, *EastIndies*, and *DRegional*), and which were annotated with links to Wikipedia. The good performance of the method on the six datasets in three different languages shows that the combined use of geographic and semantic knowledge is promising. The *GeoSem* method has been compared to three state-of-the-art toponym resolution and entity linking systems, and improved on their performance.

Chapter 4

Person name disambiguation

This chapter introduces a novel method for disambiguating person names from news articles. Given the assumption that a person name always refers to the same entity in a document, person name disambiguation amounts to document clustering. The method 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, this method dynamically adapts its clustering strategy to the most suitable configuration for each name depending on how common this name is. It is a partially-supervised approach that returns as output a set of social networks, one for each disambiguated entity. Section 4.1 describes the datasets that will be used for development and experimentation. Section 4.2 proposes a strategy to calculate person name ambiguity. The disambiguation method is explained in detail in section 4.3 and its performance is assessed and compared against other methods in section 4.4.

4.1 Data

To the best of my knowledge, there are no existing or available datasets for person name disambiguation in the historical news domain. In this section, I review existing corpora of contemporary news articles that have been annotated with person entities. In particular, I describe three benchmark datasets: the CRIPCO Corpus, the John Smith Corpus, and the NYTAC Pseudo-name Corpus. I also present the Banning–Schillebeeckx Corpus, a historical dataset for person name disambiguation created specifically for experimenting and assessing the usefulness of the method in historical research. Even if it

did not go through a process of annotation of person entities, it nevertheless proved useful for evaluating qualitatively the performance of the method in the historical domain. I conclude this section by summarizing the four different datasets.

4.1.1 CRIPCO Corpus

The Cross-document Italian People Coreference Corpus (CRIPCO), introduced in Bentivogli et al. (2008) (18), was created to provide a benchmark dataset to the task of person name disambiguation in a language other than English, in particular Italian. The corpus comes with a development and test set, of 105 and 103 query names respectively. For each query name, several documents containing this name are classified into different sets, each corresponding to a different entity. The query name 'Anthony Hopkins', for example, has eighty-nine documents (i.e. news articles) mentioning it, all into one only subset, as they all refer to the world-renowned actor; whereas the query name 'Andrea Bianchi' has a total of 127 documents mentioning this name, classified into eighteen different subsets, each referring to a different person named thus. There are a total of 43,328 documents in the corpus, 22,574 in the development set and 20,754 in the test set. There is an average of 3.22 entities per query name.

4.1.2 John Smith Corpus

The John Smith Corpus was introduced in Bagga and Baldwin (1998) (14) as the first reference set for cross-document coreference resolution. It consists of only one query name, 'John Smith', the most common name of the English language. It consists of 197 different news articles from the New York Times, each of them containing at least one mention of 'John Smith' corresponding to one of 35 possible different entities. The documents are not equally distributed among the different entities: there are 24 entities that are only mentioned in one document, and there is one entity that is mentioned in 88 documents.

4.1.3 NYTAC Pseudo-name Corpus

The NYTAC[1] Pseudo-name Corpus is an artificial corpus introduced in Rao et al. (2010) (81), created by conflating dissimilar person names together. With a total of 19,360

[1]New York Times Annotated Corpus (Sandhaus (2008) (87)).

news articles from the New York Times, this dataset consists of 99 pairs of conflated person names (i.e. 198 entities), matching in gender. Half of them are topically similar, such as Robert Redford and Clint Eastwood (actors), Plácido Domingo and Luciano Pavarotti (opera singers), or Viswanathan Anand and Garry Kasparov (world chess champions), whereas the remaining entities are arbitrarily conflated. The corpus consists of a total of 19,360 documents in which all mentions of the query names have been rewritten for concealment.

4.1.4 Banning–Schillebeeckx Corpus

Due to the lack of annotated data from the historical news domain, I created, with the help of one of the historians of the AsymEnc project, a dataset that would assist us to assess the impact of the approach introduced in this chapter in the social sciences. The dataset was not annotated with the person entities behind the person names, and was used just to provide a qualitative analysis of the method. In the creation of the dataset, we focused on two actors that played an important role in the religious transformations of the postwar years in the Netherlands: Willem Banning and Edward Schillebeeckx.

The 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, some heuristics were applied to disregard those articles in which the query name was at least once preceded by any capitalized word not matching with their first and middle names, their initials, or with any title. We restricted the data to the years of interest, namely between 1930 and 1970 in the case of Willem Banning, and 1950 and 1990 in the case of Edward Schillebeeckx. The resulting dataset consisted of 3,267 news articles for Banning and 2,172 news articles for Schillebeeckx. 'Banning' is a much more common last name than 'Schillebeeckx', which is probably the reason behind the difference in the number of articles between the two. Whereas all mentions of 'Schillebeeckx' in the collection seem to refer to the Edward Schillebeeckx in which we were interested, a quick search at the beginning of the experiment revealed that there were several different people with the name 'Banning', among which at least a shopkeeper, a swimming champion, a man on trial, and an amateur fisherman.

Corpus	Resolved	Language	Nature	DevSet	docs	querynames	entities
CRIPCO	Yes	Italian	Natural	Yes	20,754	103	354
JohnSmith	Yes	English	Natural	No	197	1	35
NYTAC	Yes	English	Artificial	No	19,360	99	198
BanSchillb	No	Dutch	Natural	No	5,439	2	unknown

Table 4.1: Summary of datasets: 'resolved' refers to whether there is a gold standard, 'language' refers to the language of the collection, 'nature' to whether it is a natural or artificial corpus, 'devset' to whether it comes with a development set, and 'docs', 'querynames' and 'entities' refer to the number of documents, query names, and entities of the test set.

4.1.5 Summary of datasets

In this section, I have reviewed three of the existing person name disambiguation datasets: the CRIPCO Corpus, the John Smith Corpus, and the NYTAC Pseudo-name Corpus. Given the lack of data from the historical domain, I also describe the creation of a dataset of historical news articles that will be used to assess the performance of the method from a qualitative point of view: the Banning–Schillebeeckx Corpus. The characteristics of the four datasets are summarized in table 4.1.

4.2 Ambiguity of person names

The method proposed in this chapter regards person name disambiguation as a document clustering task. Documents containing a query name are clustered into an unknown number of clusters, each corresponding to a different entity that is known by the same referring name (i.e. the query name). The method relies on the assumption that the number of clusters is directly related to the ambiguity of a person name. The more ambiguous a name is (e.g. 'John'), the more likely it is that it refers to several different people and that, therefore, it yields more clusters. Conversely, the less ambiguous it is (e.g. 'Edward Cornelis Florentius Alfonsus Schillebeeckx'), the more likely it is that it refers to one only person. 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. In this section, I present a method to calculate ambiguity from person name mentions in text. This calculation consists of two steps: (1) building name lists for all languages

under investigation (subsection 4.2.1), and (2) computing the ambiguity of the person names (subsection 4.2.2).

4.2.1 Building name inventories for three languages

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. In Zanoli et al. (2013) (103), an Italian-specific resource, the phonebook *Pagine Bianche*,[1] is used for their experiments on an Italian corpus. 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, and this person is usually the male head. To overcome this problem, I choose a different approach, consisting of collecting person names from a large corpus of text. In this manner, even though I cannot guarantee that the gender bias is not there anymore, at least it is not as explicit and conscious.

The datasets that I work with are collections of newspaper articles in English, Italian, and Dutch. Below, I detail the specificities of the resources built for each of these languages:

- **Italian resource:** I downloaded the unannotated Italian corpus PAISÀ,[2] a collection of Italian texts from the Internet that amounts to 250 million tokens (1.5GB) and used the named entity recognizer TextPro (Pianta et al. (2008) (75)) to identify person names in the Italian texts. The extracted list of 718,568 person names is not a census of the Italian population, but a list of people mentioned in webpages or blogs.

- **English resource:** I used the `Persondata` information from the DBpedia project (only available for English and German so far), which was built by collecting all the Wikipedia articles about people. At the moment of download, the English `Persondata` database had 7,889,574 entries.

- **Dutch resource:** Since DBpedia does not provide a `Persondata` table for Dutch, I downloaded the list of people from the English version of Wikipedia and selected

[1] `http://www.paginebianche.it/`
[2] `http://www.corpusitaliano.it/`

the names of people who were born in the Netherlands. The most occurring Dutch name was given the maximum occurrence value of the list and the rest of names (i.e. names of people not born in the Netherlands) were normalized against them.

This resulted in a list of person names for each language. Each name was split into their different tokens[1] and used for building three lists: a list of first tokens, a list of last tokens, and a list of middle tokens (which I will henceforth call, respectively, list of first names, last names, and middle names, for simplification, since it roughly coincides). A list of person name stopwords was created to avoid tokens such as '*Junior*', '*Sr.*' and '*III*' or particles such as '*van*' and '*der*' to be considered as person name tokens.

4.2.2 Person name ambiguity calculation

Given a newspaper article, this step assigns a numerical ambiguity to each person name that has been identified in the text. I propose an ambiguity scale that spans from 0 to 1, in which very ambiguous names occupy the highest range and very non-ambiguous names take the lowest range. Two assumptions are made in this step: that the more common a person name is, the more ambiguous it is; and that the more tokens a person name has, the less ambiguous it is. Formally, I distinguish three types of person names that can be encountered in plain text: single-token names, two-token names, and multiple token-names:

- **Single-token person names:** Person names that consist of only one token (such as 'John' or 'Smith') are among the most ambiguous person names that can be found. Their ambiguity is calculated by multiplying their relative frequency in either the list of first names or the list of last names (I select the one that provides a higher frequency) by 0.2 and adding 0.8. These two steps assure that the calculated ambiguity of single-token person names falls inside the upper end (*[0.8–1.0]*) of the spectrum.

- **Two-token person names:** Person names that consist of two tokens (such as 'John Smith' are the most likely to occur in a text, as they often correspond to a first and last name combination. Yet, their ambiguity is in many cases obviously

[1]Since precision is more important than recall in this step, I consider only names consisting of at least two tokens, since single tokens are more likely to be misidentified or misclassified by the recognizer.

smaller, since the person referred to by them is specified by using two tokens (e.g. 'John Smith') and not only one (e.g. 'John'). In the three languages I use for the experiments (Italian, Dutch, and English), first names are less diverse than last names. In English, in order to cover around 96% of the total of first names, 5,000 first names are enough; whereas to cover the same number of last names, 70,000 last names are needed, according to Popescu (2009) (76). Given that the diversity of first and last names is not necessarily the same in all languages, I calculate the weighted average[1] between the relative frequency of the first token (taken from the list of first names) and the relative frequency of the second and last token (taken from the list of last names). The weight of the most common two-token name ('Giovanni Rossi' for Italian, 'John Smith' for English, 'Jan (de) Vries' for Dutch) is taken as the maximum value against which the relative frequency of any other two-token name combination is calculated. The resulting number is then multiplied by 0.6 and added to 0.2, to guarantee that the ambiguity of two-token person names falls in the middle part (*[0.2–0.6]*) of the spectrum.

- **Multiple-token person names:** Person names that consist of three or more tokens (such as 'Edward John Smith') are among the least ambiguous person names that can be found, since more tokens than in the previous cases are used to specify the person they refer to. Their ambiguity is calculated in a similar manner than in the case of two-token names, but distributing the weight of the first and middle names equally.[2] The resulting number is multiplied by 0.2, to guarantee that the ambiguity of a multiple-token person name always falls in the lower end (*[0.0–0.2]*) of the spectrum.

The different ambiguity ranges and their relation to the three formally defined classes of person names are summarized in figure 4.1. Table 4.2 shows examples of English person names that fall into each range.

[1]The weighted average in each language is based on the fraction of the sum of occurrences of the ten most common first names by the sum of occurrences of the ten most common last names.

[2]This would be addressed differently if I was dealing with other languages such as Spanish or Portuguese, in which names are usually composed of at least two family names.

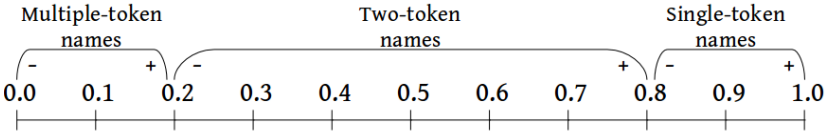


Figure 4.1: Person name ambiguity spectrum.

Ambiguity range	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 4.2: Examples of English names that fall into each ambiguity range.

4.2.3 Summary

In this section, I have given details on the creation of language-dependent lists of person names that can be then used to assess the ambiguity of person names in each different working language. Based on them, a calculation metric is described that finds an ambiguity value for each person name, which falls into one of ten ambiguity ranges. Ambiguity of person names is a crucial aspect of the method that is introduced in the next section.

4.3 Disambiguating person names

Given the assumption that a person name always refers to the same entity in a given document, person name clustering amounts to document clustering in which each document mentioning a certain person name (i.e. query name) should be grouped together with the other texts in the collection that contain the same person name mention and that actually refer to the same entity (i.e. person). In order to cluster documents, a similarity measure is needed. The core idea is that two documents should be clustered

together if and only if they are similar enough, i.e. if there exists enough evidence that they belong together. As already hinted in the previous section, 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 should be required to ensure that only documents that refer to the same entity are clustered together.

In section 4.2, I have described how I assess person name ambiguity. In this section I 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. This model relies heavily on the social dimension of news, so I model document similarity based on social network similarity. Thus, for each query name, I represent documents as social networks in which the nodes are the people mentioned in them. In order to determine network similarity I take two types of information into account: the amount of node overlap (for which I learn a threshold from a small manually labeled data set) and the ambiguity of the overlapping nodes (for which I 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 I additionally make use of further features in those cases where networks do not provide sufficient evidence: BoW representations of the content, the dominant topic according to a topic modeling algorithm, and the overlap of other named entity types (in particular, toponyms and organization names). These additional features are added to model the document's content.

In this section, I propose a method for disambiguating person names that simultaneously produces a visualization of the relations between the different persons in the collection of news articles in the form of social networks of the mentioned people. In this manner, a researcher interested in a certain historical actor is able, not only to retrieve all the documents in which this person is mentioned, but also to visualize all the related people according to that particular newspaper collection. Thus, the proposed disambiguation method automatically extracts networks of the people mentioned in news stories, weighted according to their significance in the news and distributed according to their co-occurrence in the text.

4.3.1 Building social networks from documents

The first step of the method is to represent each document containing the query name as a social network of the people mentioned in it. A network consists of two main components: nodes and edges. In a social network, the nodes are the actors and the edges represent the relations between them.

4.3.1.1 Obtaining the nodes

A social network is a structure that captures the relations between a set of actors. Therefore, the first step for the creation of a social network must necessarily be the extraction of person names from plain text. I used a named entity recognizer in order to identify person names: Stanford Named Entity Resolution[1] (Finkel et al. (2005) (43)) for English and Dutch,[2] and TextPro[3] (Pianta et al. (2008) (75)) for Italian. In order to enhance the performance of the named entity recognition, I have applied some basic heuristic filtering steps. On many occasions, newswire text introduces a person name by his or her description, as in '63-year-old Frank Donoghue', 'captain Ben Shaw', or '21-year-old pianist Theo'. Such linguistic cues are very reliable, since it can be expected that most of the times a capitalized word following an age, a title, or a profession will be a person name. I created a set of basic language-specific rules to capture age, titles, and professions.

To capture age, regular expressions suffice: in English and Dutch, every sequence of capitalized words (including initials and middle particles such as '*del*', '*van der*', or '*v.d.*') following the expression '*XX-year-old*' and '*XX-jarige*' (where `XX` is a number, expressed numerically or alphabetically, and where the dash is optional) respectively is considered a person name. In Italian, I consider a person name every sequence of capitalized words preceding a comma-delimited apposition of the form '*(di) XX anni*' (where `XX` is again a number, expressed numerically or alphabetically, and where preposition '*di*' is optional). To capture the title or the profession, I relied on lists of professions and titles. Finally, newspapers tend to personalize institutions or organizations such as political entities, which is the reason why, unlike in other domains

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

[2] For English, I used a classifier that comes by default with the software, for Dutch I used the training data from the CoNLL-2002 shared task (`http://www.cnts.ua.ac.be/conll2002/ner`).

[3] `http://textpro.fbk.eu/`.

such as literature, verbs of utterance are not reliable clues to identify human names unequivocally.

The list of person names appearing in a news article is not necessarily its set of nodes. In example 13, person names are marked in bold:

(13) *Mr.* ***Parker Cramer*** *and Mr.* ***Bert Hassell*** *left Rockford, Illinois, in August,1929, to fly to Sweden, via Greenland, but had to make a forced landing in Greenland, where they were rescued by a scientific expedition under Professor* ***Hobbs****, after they had been given up as lost. They were taken to Denmark.*
Mr. ***Cramer*** *was one of the pilots with Sir* ***Hubert Wilkins*** *on his second expedition to the Antarctic in 1928–29.*[1]

In the example, 'Parker Cramer' and 'Cramer' clearly correspond to one entity and should therefore occupy only one node in a social network. It is common in newspaper articles that the first mention of a person in an article is by using his or her full name or the form that is most widely known by the readers and that, at the same time, results less ambiguous; and that further mentions use a shorter form (surname, title plus surname, or first name if proximity wants to be shown). This is a recommendation common in several press style books, among which *The Associated Press Stylebook* (45) or the style book of the Italian weekly newspaper *Internazionale*.[2]

Within-document co-reference of person names is resolved by naive surface string matching: given the assumption that two identical surface forms in the same article will refer to the same person, a reduced form will likewise always refer to an encompassing string, unless there is a contradiction of title. In this case, each surface form is supposed to be a different entity only when the titles clearly indicate two different genders (such as 'Mr. de Muis' and 'Mrs. de Muis). In any other case, all identical surface forms or reduced forms matching longer strings will be considered to correspond to one only entity, and the longer name will be stored. In Dutch, it is very common to introduce people in news articles by the initials of their given names, so the matching strategy also takes initials into account, which can be initials of first names or of particles (e.g. '*v.d*' matching '*van der*').

Naming conventions vary greatly in different cultures of the world. It is a convention that many Western societies name people by a given name followed by a family name

[1] ***Source:*** *New Zealand Herald, Volume LXVIII, Issue 20949, August 12th, 1931.*

[2] `http://www.internazionale.it/dalla_redazione/2015/06/25/libro-stile-internazionale`.

(commonly known as the *Western order*), whereas Eastern societies often prefer the last name preceding the given name (commonly known as the *Eastern order*). There are several exceptions for both cases, and in some societies both name orders may co-exist. It was not uncommon in Italian, especially some years ago rather than today, that names would be presented sometimes in the Eastern order due to the influence of bureaucratic use. Example 14 shows a case from the CRIPCO corpus[1] in which the family name ('*Grigolli*') comes before the given name ('*Giorgio*'), and example 15 shows the inverse case, in which the given name '*Giorgio*' comes before the family name '*Grigolli*', as is usually the rule in Italian:

(14) *In riferimento a quanto apparso il 10 settembre sul quotidiano "L'Adige" a firma del sig.* ***Grigolli Giorgio*** *ben vengano i ricercatori storici sui fatti di Malga Zonta, purché siano documentati e non politicizzati.*[2]

(15) *Per* ***Giorgio Grigolli*** *"sessant'anni dopo, il proposito c'è. Occorre scrivere intera (non riscrivere) la storia di Malga Zonta. Era tempo, si può prenderne atto, convintamente."*[3]

I take this possible name inversion into account when performing within-document coreference resolution, and store the name combination that is more likely to be in the Western order according to the name lists created in subsection 4.2.1 (i.e. the combination whose first token is more frequent in the list of first names and whose last token is more frequent in the list of last names).

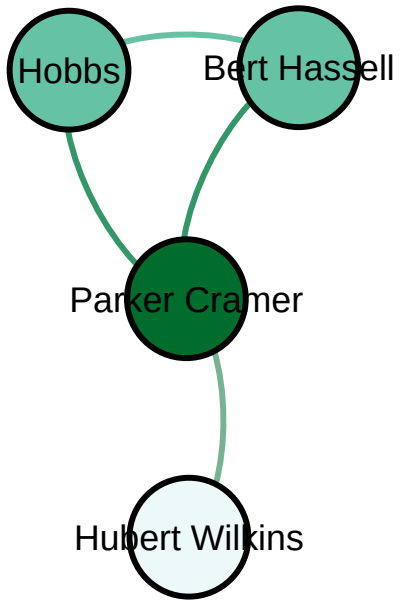
4.3.1.2 Linking the nodes

Once all the nodes corresponding to the different entities in an article have been identified, I define the kind of relation that is drawn between them in the following manner: each pair of nodes is linked in the network if they co-occur in the same paragraph in the news article. The resulting network is undirected and based on the co-occurrence of nodes. It is also weighted, as the more two entities co-occur throughout an article, the

[1] `http://hlt-nlp.fbk.eu/technologies/cripco`.

[2] **Translation:** With reference to what appeared on the newspaper *L'Adige* on September 10th, signed by Grigolli Giorgio: historical researchers on the Malga Zonta events are welcome, but these have to be documented and not advertised.

[3] **Translation:** Quoting Giorgio Grigolli "sixty years after, the intention is there. The whole history of Malga Zonta needs to be written (not re-written). It was time, you can take note of it, decisively."



Hobbs–Bert Hassell

`Content words:` left, rockford, illinois, august, 1929, fly, sweden, via, greenland, forced, landing, rescued, scientific, expedition, given, lost, taken, denmark.

`Named entities:` rockford, illinois, sweden, greenland, denmark.

Hobbs–Parker Cramer

`Content words:` left, rockford, illinois, august, 1929, fly, sweden, via, greenland, forced, landing, rescued, scientific, expedition, given, lost, taken, denmark.

`Named entities:` rockford, illinois, sweden, greenland, denmark.

Bert Hassell–Parker Cramer

`Content words:` left, rockford, illinois, august, 1929, fly, sweden, via, greenland, forced, landing, rescued, scientific, expedition, given, lost, taken, denmark.

`Named entities:` rockford, illinois, sweden, greenland, denmark.

Parker Cramer–Hubert Wilkins

`Content words:` pilots, second, expedition, antarctic, 1928, 29.

`Named entities:` antarctic.

Figure 4.2: Social network representation of the news article from example 13. The darker the node, the higher its degree. On the right, the list of words for each edge in the network.

stronger the relation between them will be. In an edge attribute, I keep the list of files in which both nodes of the edge co-occur. Besides, for each edge in the network I store the list of all the content words in the paragraph where both nodes appear (stopwords removed) and the list of named entities expressions that are not person names. Figure 4.2 represents the text from example 13 as a social network, in which each different person mentioned in it is given a distinct node, and these are linked if their mentions in the article occur in the same paragraph.

4.3.2 Similarity of documents represented as social networks

The core idea behind assessing the similarity of documents as social networks is that the social circle of people may be a useful indicator of who they are: it is their social context. This similarity measure assists in the decision of whether two documents mentioning the same person name should be clustered together and their network representations

should therefore be merged. A very naive strategy could be to join together all the documents and merge their networks if they share at least one person name (apart from the query name, which is common in all the documents). This would mean that nobody knows two different people who share the same name. This is a naive and obviously dangerous strategy to follow, as person names in plain text may be introduced as unspecific as by just the first or the last name (e.g. two networks would also be joined if a very ambiguous name, such as 'John', is a node in both of them).

This strategy assumes that a node that is shared by two networks (henceforth *overlapping node*) corresponds to the same entity. This would mean that the same person name occurring in two different documents corresponds to the same person. This is of course not necessarily the case. Besides, as already seen, mention names can range from single tokens to multiple tokens, and they can be extremely 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 in great measure on the 'quality' of these overlaps. An overlapping person name that provides greater evidence that we are dealing with one only entity (i.e. a name that has a low ambiguity) is considered of higher quality than an overlapping name that provides little evidence that it corresponds to one only entity (i.e. a name that has a high ambiguity).

In section 4.2.2, I have described the process of assigning a numerical value that assesses the ambiguity of each person name that can be encountered in a text. I distinguish between three degrees of ambiguity: low, medium, and high. **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. Finally, **medium ambiguity** consists of the names that fall into the middle spectrum (see figure 4.3).

Each name in each document (and its respective node in the social network representation of the document) is assigned a numerical ambiguity value and given an ambiguity degree. Table 4.3 provides the resulting ambiguity values for the names from example 13.

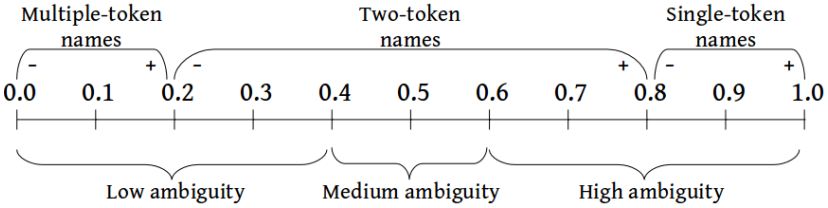


Figure 4.3: Person name ambiguity spectrum with ambiguity degrees.

Person name	Numerical ambiguity value	Ambiguity degree
Hobbs	80.06	High
Bert Hassell	20.36	Low
Parker Cramer	20.58	Low
Hubert Wilkins	21.16	Low

Table 4.3: Numerical ambiguity values and ambiguity degree of the names from example 13.

4.3.2.1 Learning clustering probabilities

In order to understand how reliable it is to cluster documents together when their social network representations share a certain number of nodes independently from the quality of these nodes, I decided to learn clustering probabilities from the development set from the CRIPCO corpus, consisting of 105 different query names. I computed the ambiguity for each of these names according to the method described in subsection 4.2.2 and classified them into ten different ranges (i.e. `[0.0-0.1]`, `[0.1-0.2]`, `[0.2-0.3]`, etc.). For each query name in each ambiguity range, I learned the probability that two documents belong to the same entity if their social network representations had no nodes in common (apart from the query name), one node in common, two nodes in common, or three nodes in common. It was observed that, in all ambiguity ranges, whenever two networks shared four or more nodes, there was a near-unity probability that their respective documents belonged together.

The clustering probabilities learned from the development set revealed the importance of assessing the ambiguity of the target person name to disambiguate. Whereas two documents mentioning a name of the lowest ambiguity range had a very high probability (up to 0.96) of referring to the same entity even when no nodes (other than the one corresponding to the query name) existed in common in their network representations, a highly ambiguous name had a much lower probability (0.13) to refer to the

same entity in the case of no nodes in common. The data showed that the larger the number of overlapping nodes between both networks is, the higher the probability is that two documents containing the same query name refer to the same person.

4.3.2.2 Penalizing lower quality overlaps

The learned probabilities do not take into account the quality of the person names in common. Returning to example 13, supposing 'Hobbs' is the query name, it is evident that two social networks whose names in common have a low ambiguity (such as 'Parker Cramer', 'Bert Hassell', and 'Hubert Wilkins') have a higher probability of referring to the same person (i.e. the very same person behind the name 'Hobbs') than two social networks whose names in common have a high ambiguity (such as 'John', 'Mary', and 'William'). A penalty function is defined to lower the learned probabilities when applied to networks with overlapping nodes of lower quality. The probability of two networks being clustered together is lowered according to the decreasing probability function (dp) in equation 4.1:

$$dp = \frac{Pr(n[i]) - Pr(n[i-1])}{i+1} \tag{4.1}$$

where i is the number of overlapping nodes (excluding the overlapping node corresponding to the query name) between two documents, n is the set of networks sharing a certain number i of nodes, and $Pr(n[i])$ is the probability that two networks belong together if they have i nodes in common. This is then applied to the *penalty* function as shown in equation 4.2:

$$penalty = Pr(n[i]) + ooq \cdot dp \tag{4.2}$$

The *penalty* function adds to the probability that two networks belong together (given an i number of nodes in common) the decreased probability dp multiplied by the overall quality of the nodes in common ooq, which is the sum of the quality of the nodes that are shared by the two networks. The quality of the nodes is inversely proportional to the ambiguity of the person names of the nodes. A high-ambiguity person name has a low quality, whereas a low-ambiguity person name has a high quality. A numeric value is given to each node provided its quality: a node with a low-ambiguity person name ($\downarrow$) is represented by a 0 value, a node with a medium-ambiguity person name ($\rightarrow$) is

represented by -1, and a node with a high-ambiguity person name ($\uparrow$) is represented by -2. If there is more than one mention node in common in the two networks, the numeric values behind the ambiguity degree of each node are added up. Table 4.4 shows how probabilities are recalculated, according to the penalty function in equation 4.2, depending on the number of nodes in common and the ambiguity (i.e. inverse quality) of the nodes' person names. The idea behind the *penalty* function is that the probability that two documents that mention the same query name actually refer to the same entity is lower the more ambiguous the overlapping nodes in their social network representations are, and vice versa.

4.3.3 Other similarity metrics

The evidence social network similarity can provide is limited: the overlapping of nodes is not always a sufficient source of information to decide whether two networks belong together or not, especially so in cases of small overlap or no overlap at all. As discussed in Artiles et al. (2009) (10), approaches that focus on named entities tend to achieve high precision at the cost of recall. A similarity measure between documents based solely on the node overlaps of their social networks representations 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, and it should not be inferred that the documents involved should not be clustered together just because of lack of evidence.

In order to address this problem, I additionally make use of other kinds of features: for each ambiguity range, I learn probabilities that two documents are clustered together in terms of a bag-of-words vector representation of simple word counts, of a bag-of-words vector representation with tf-idf weightings, and of the number of common non-person mentions. Finally, I applied LDA to our datasets using collapsed Gibbs sampling, as implemented in the `scikit-learn` Python library,[1] to produce a lower dimensional representation of the datasets. Each document was assigned the most relevant latent topic.

[1] `http://pypi.python.org/pypi/lda`.

ONE OVERLAPPING NODE

$dp = \frac{Pr(n[1]) - Pr(n[0])}{2}$

Node ambiguity	Recalculated probability
$\uparrow$	$Pr(n[1]) - 2 \cdot dp = Pr(n[0])$
$\rightarrow$	$Pr(n[1]) - 1 \cdot dp$
$\downarrow$	$Pr(n[1]) - 0 \cdot dp = Pr(n[1])$

TWO OVERLAPPING NODES

$dp = \frac{Pr(n[2]) - Pr(n[1])}{3}$

Node ambiguity	Probability recalculated
$\uparrow\uparrow$	$Pr(n[2]) - 4 \cdot dp$
$\uparrow\rightarrow$	$Pr(n[2]) - 3 \cdot dp = Pr(n[1])$
$\rightarrow\rightarrow$	$Pr(n[2]) - 2 \cdot dp$
$\uparrow\downarrow$	$Pr(n[2]) - 2 \cdot dp$
$\rightarrow\downarrow$	$Pr(n[2]) - 1 \cdot dp$
$\downarrow\downarrow$	$Pr(n[2]) - 0 \cdot dp = Pr(n[2])$

THREE OVERLAPPING NODES

$dp = \frac{Pr(n[3]) - Pr(n[2])}{4}$

Node ambiguity	Probability recalculated
$\uparrow\uparrow\uparrow$	$Pr(n[3]) - 6 \cdot dp$
$\uparrow\uparrow\rightarrow$	$Pr(n[3]) - 5 \cdot dp$
$\uparrow\uparrow\downarrow$	$Pr(n[3]) - 4 \cdot dp = Pr(n[2])$
$\uparrow\rightarrow\rightarrow$	$Pr(n[3]) - 4 \cdot dp = Pr(n[2])$
$\uparrow\rightarrow\downarrow$	$Pr(n[3]) - 3 \cdot dp$
$\rightarrow\rightarrow\rightarrow$	$Pr(n[3]) - 3 \cdot dp$
$\downarrow\rightarrow\rightarrow$	$Pr(n[3]) - 2 \cdot dp$
$\downarrow\downarrow\uparrow$	$Pr(n[3]) - 2 \cdot dp$
$\downarrow\downarrow\rightarrow$	$Pr(n[3]) - 1 \cdot dp$
$\downarrow\downarrow\downarrow$	$Pr(n[3]) - 0 \cdot dp = Pr(n[3])$

Table 4.4: Recalculation of probabilities. The left column shows the combination of nodes according to their ambiguity degree. Each arrow represents one node: $\uparrow$ a high-ambiguous name, $\rightarrow$ a medium-ambiguous name, and $\downarrow$ a low-ambiguous name. In the right column, the probability of two networks being clustered together based on the number of nodes they share is recalculated according to the quality of their nodes.

4.3.4 Clustering strategy

Given a query name and a set of documents that contain a mention of this query name, the ideal output is a set of clusters of documents, each cluster corresponding to a different entity bearing the same query name. In parallel, the clustering algorithm returns one social network per cluster (i.e. ideally, per entity), which is the result of merging the different social network representations of the documents of each cluster by their common nodes. The clustering is performed taking the following assumption into consideration: that the more ambiguous a person name is, the more entities it can potentially refer to.

The first step is to turn all the documents containing a given query name into their social network representations, as described in subsection 4.3.1. I initiate the clustering by taking the social network that has the highest number of nodes (i.e. the network corresponding to the document that has the most person names), and sort the remaining networks by decreasing number of nodes overlapping with the largest network. The similarity between the document corresponding to the largest network and the document with the largest number of overlaps is computed. If there is evidence enough that these two documents mentioning the same query name actually refer to the same person, I cluster both documents together and merge their social network representations. The resulting network is now the largest, and I re-rank the remaining networks by sorting them again by decreasing number of nodes overlapping with the updated largest network. Again, similarity between the updated largest network and the network with the most overlapping networks is computed, and the documents are clustered and the networks merged if enough evidence exists. This process is repeated until no partially-overlapped network is found. In this case, I repeat the whole procedure of finding the largest remaining network and finding its fully- or partially-overlapping networks. This process continues until all the documents/networks have been considered. This is a greedy algorithm, and it is therefore of prime importance that two documents are only clustered together if there exists enough evidence that this should be the case.

I 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. To determine which is the amount of evidence needed to join two documents and merge their networks, first of all 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. Since a less ambiguous name tends to correspond to fewer entities than a more ambiguous one, non-ambiguous names allow documents with a low similarity to be clustered together, whereas ambiguous names are less permeable and require high-document similarity for them to be clustered together. Decision-making is learned from observation of the CRIPCO development set. Given a low-ambiguity query name, if any of the extracted features is true, this evidence is considered 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. An overlap of five entities in a high-ambiguity query name (be them person names, locations, or organizations) proves in most of the cases to provide 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 have to be true. Finally, medium-ambiguity names have a middle stance between low-ambiguity and high-ambiguity names when it comes to permeability.

4.3.5 Summary

In this section, I have introduced a new person name disambiguation method. The method, which has a strong focus on the social dimension of news articles, returns clusters of documents containing mentions to the same person represented as one social network. In subsection 4.3.1, I describe how single documents can be represented as social networks. In order to decide whether two networks corresponding to two documents belong together, I look foremost at the similarity between their social network representations. This measure is explained in subsection 4.3.2. Other features also considered have been outlined in subsection 4.3.3, and the clustering strategy has been described in subsection 4.3.4.

4.4 Experimental results

In this section, I evaluate the performance of the method that has been proposed in this chapter, which I henceforth call *SNcomp*. I introduce the baselines against which I test the method and describe the system's settings. In subsection 4.4.3, I provide

results on three datasets: the CRIPCO corpus, the NYTAC pseudo-name corpus, and the John Smith corpus. In subsection 4.4.4, I provide a qualitative analysis in order to assess the contribution of the method to the social sciences.

4.4.1 Baselines

I compare the **SNcomp** method with two baselines: (1) **SNsimple** is the base case, the most naive representation of the method, in which two documents are grouped together whenever their network representations share at least one node other than the node corresponding to the query name; and (2) **TopicModel** clusters together the documents by their most relevant topic as returned when applying LDA using collapsed Gibbs sampling.[1] Besides, I also provide for the CRIPCO dataset the results from Zanoli et al. (2013) (103), which is the state of the art on this dataset. On the NYTAC pseudo-name corpus, I provide the results from Rao et al. (2010) (81), who are the state-of-the-art for this dataset and also presented results on the John Smith corpus. From Rao et al. (2010) (81), I provide the results of the two system configurations that work the best both for the NYTAC pseudo-name corpus and for the John Smith corpus. The method introduced in this chapter is the only one so far that has been evaluated for the three datasets simultaneously.

4.4.2 Settings

My method does not require a big amount of training data, but just a representative selection spreading over the ambiguity spectrum 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, as using the whole dataset makes no significant difference in the performance. I randomly selected the query names, making sure I would, when possible, have a query name for each of the ten ambiguity ranges.[2] The CRIPCO training dataset does not have a query name

[1] As implemented in the scikit-learn Python library: `http://pypi.python.org/pypi/lda`.

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

for all of the ambiguity ranges: I 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*]). I made sure that if a query name from the testing dataset falls into one of these ranges, it would take the probability of its immediately precedent less ambiguous range. Besides the learning of probabilities, the mentioned fifteen instances from the development set have been used to find the optimal combination of features. The learned probabilities and strategy have been applied directly, without further learning nor tuning, to the other two datasets.

4.4.3 Quantitative analysis

Table 4.5 shows the results on the three datasets. The evaluation metrics used to compare the performance of the different systems with regard to the gold standard are extended B-Cubed precision, recall, and their harmonic mean, F_1Score, as introduced in Artiles et al. (2009) (12) for the WePS-2 task. I used the official scorer provided for the task.

	cripco				**nytac_sel**			**johnsmith**		
Approach	P	R	F_1S		P	R	F_1S	P	R	F_1S
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	0.62
Rao et al. 2010 *[2]*	–	–	–		0.82	0.24	0.36	0.85	0.59	0.70
SNcomp	0.87	0.95	0.91		0.63	0.75	0.68	0.79	0.60	0.68

Table 4.5: Evaluation results.

While the results of the main method introduced in this chapter, *SNcomp*, are slightly lower than Zanoli et al. (2013), this difference is not statistically significant (Wilcoxon test, $p = 0.054$), an yet this result is obtained without needing to rely on expensive resources, such as a knowledge base. A further advantage of the *SNcomp* method is that it is easily adaptable to other datasets from the news domain. The only work that at the time of the experiment had reported results for the NYTAC pseudo-name corpus is by the creators of the dataset themselves, i.e. Rao et al (2010) (81), who also report results for the John Smith corpus. The NYTAC dataset was

artificially created by conflating pairs of names, and therefore some of the assumptions of my method 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, as is the case of former tennis players Chris Evert and Lindsay Davenport or of opera singers Plácido Domingo and Luciano Pavarotti, in which cases both names appear very often mentioned in the same contexts. Therefore, their social networks have much less predictive power than in natural data, where it can be assumed that two people with the exact same name have a low probability to share a big portion of their social networks. That could explain the lower precision reported for this dataset, and yet the overall results obtained are comparable to those from the best of the two models from by Rao et al. (2010) (81).

The results reported for the John Smith corpus improve upon recent models, such as Singh et al. (2011) (88), who obtained 0.664, but is far from the most recent approach at the time of writing (Rahimian et al. (2014) (80)), who reported around 0.80. This might be well due to the fact that there was only one query name in the CRIPCO development set that had high ambiguity, which was, still, far from being as ambiguous as 'John Smith'. My method works overall better than any of the two methods from Rao et al. (2010) (81) when I 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. The F_1Score for the John Smith corpus drops to 0.37 if the name 'John Smith' is considered a low-ambiguity name, and to 0.52 if it is considered a medium-ambiguity name. The F_1Score for the CRIPCO dataset drops to 0.77 when the ambiguity range of the query names of this dataset is randomly assigned.

4.4.4 Qualitative analysis

The method returns a set of networks, one for each disambiguated entity. Figure 4.4 shows a fragment of the resulting social network for Donald Regan (Ronald Reagan's Secretary of the Treasury and later Chief of Staff) created from the NYTAC Pseudo-name Corpus. As mentioned, each edge is a container of information (context words

and entities weigthed with tfidf) relevant to the two nodes it links. In the example network, the edge Donald Regan and Ronald Reagan contains among others the following terms: 'astrology', 'treasury', 'abdominal', 'surgery', 'williamsburg', and 'iran'. Small knowledge of the characters will suffice to understand the relevance of these terms in regards to the relation between these historical actors. This information is encoded for each pair of nodes that can be found in the network. Mikhail Gorbachev and Margaret Thatcher are two of the many minor nodes that populate the social network of Donald Regan. They appear in only one document, from March 1987. Some relevant terms in common between these two politicians (and implicitly also Donald Regan) are: 'diplomacy', 'agreement', 'East-West', 'Middle East', 'summit' or 'Reykjavik', in reference to the 1986 Reykjavík Summit.

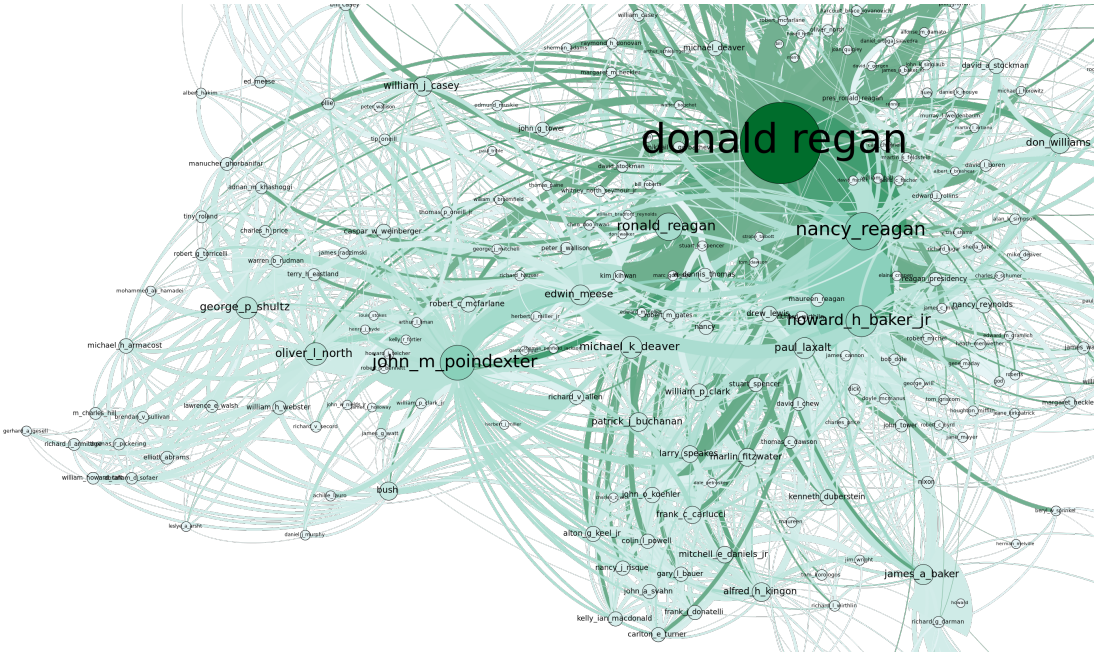


Figure 4.4: Fragment of the resulting social network for Donald Regan from the NYTACps corpus.

To assess the impact of this approach in the social sciences, I introduce here a case study that analyzes its performance and proves its contribution. The case study was conducted with the assistance and close collaboration of one of the historians of the

AsymEnc project, and resulted in a publication: Coll Ardanuy and van den Bos (2016) (7). Due to the lack of annotated data, we could only provide a qualitative analysis. As a use case, we focused 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. I have already described in subsection 4.1.4 how the corpus was created.

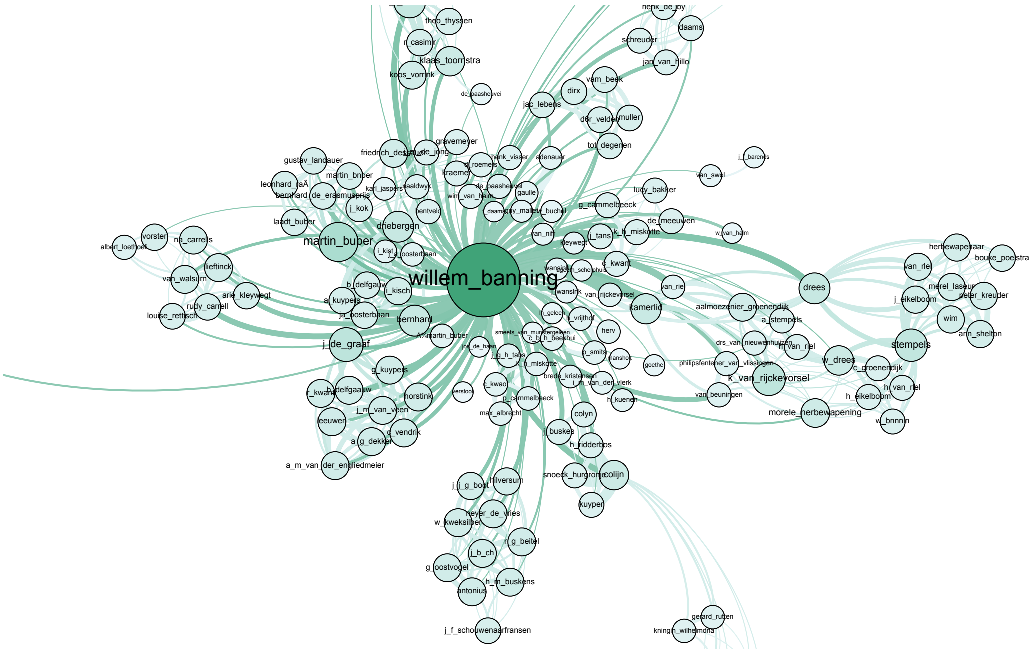


Figure 4.5: Fragment of the resulting social network for Willem Banning for the year 1963.

The amount of noise that can be found in the networks created from digitized historical newspapers is clearly higher than in the standard benchmark datasets, mostly due to incorrect character recognition. As a result, the named entity recognizer, trained on modern Dutch,[1] performs worse, even if the final networks do not suffer much from this, since noisy nodes are pushed to the periphery of the networks. The historian in the team was able to find only expected names in the center of the networks, and

[1]Training data from CoNLL-2002: `http://www.cnts.ua.ac.be/conll2002/ner/`.

exceptions were few. 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. All this was information that could have been expected in the networks, but also unexpected results were obtained: the 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.

In order to ease the task of going back to the original sources to the historian, each edge also stores the list of documents in which both nodes are present. Allusions to the controversial relation between Ronald Reagan and astrology or to the death from abdominal cancer of his Secretary of the Treasury Donald Regan in a hospital near Williamsburg could have been expected to be found in the networks. The presence of expected information in the networks is interesting and necessary because it proves the validity of the approach. Even more interesting is, though, the presence of unexpected results in the network, since they are potential hypotheses that may challenge the dominant narratives of history. 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 the method, whether there is some truth in the information yielded by the networks.

4.5 Summary

In this chapter, I have introduced *SNcomp*, a novel method for person name disambiguation which explores the relationship between name ambiguity and the amount of different entities that can be referred to by the same name. It is a partially supervised approach and has proved to be competitive in different languages and throughout very different collections without need to retrain it. Its performance is on par with the

state-of-the-art reported for the CRIPCO dataset, while using less specific resources. 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, as demonstrated in the qualitative analysis performed on the Banning–Schillebeeckx dataset.

Chapter 5

Conclusions

With the recent explosion of digitization projects in the cultural heritage domain, an increasing effort is being devoted to address the problem of information extraction from unstructured text. Several studies have hinted that deep text mining may offer great possibilities to broaden the boundaries of historical research by allowing a more effective and accurate exploration of digital texts. However, to date, most digital humanities tools and environments provide just basic search functionalities based on keywords, which often disregard the problem of word ambiguity, even though there is no doubt that it is a question of crucial importance in order to obtain a meaningful extraction of information from any collection of unstructured data. In this thesis, I have focused on two kinds of ambiguities: toponym ambiguity and person name ambiguity.

Toponym disambiguation and person name disambiguation allow humanities researchers to perform an exploration of a collection based on entities (locations and persons) and not surface word forms. This is particularly useful for historians, who often use entities as starting points through which to explore and dig into collections of historical documents, and it is certainly useful for other researchers from the humanities and social sciences too, such as literary scholars (e.g. to trace characters' movements and interactions throughout a literary work) or politics scholars (e.g. to investigate patterns and hidden structures of power and the places where these manifested themselves), to mention just two. This exploration can be assisted by a visualization of the data in the form of maps and social networks, respectively, through which the scholar obtains a location- or person-centered overview of a collection or of a meaningful selection of it, restricted, for example and in the case of press collections, in terms of dates

or of ideology of the data that form the collection, or in terms of topic by means of query restriction.

The idea behind representing a given collection as a set of social networks or as a map of mentioned locations is to allow scholars to have a bird's eye view of the scenario of a certain period of time from a social and geographic perspective. These representations of the data need not be empty structures: they can be containers of information stored for each entity. The resulting structures are likely to reproduce some expected results: expected people appearing in prominent positions in the network, expected relations between actors, expected relevant terms between one actor and another, expected locations being mentioned more often than others, etc.), which underlines the reliability of the approach; and, more interestingly, some unexpected results that may raise new questions. I conclude this thesis by reviewing its main contributions and laying the ground for future research.

5.1 Contributions

This thesis has introduced two methods that tackle the problem of named entity disambiguation: a method for toponym disambiguation and a method for person name disambiguation. They are treated as clearly differentiated tasks in order to be able to exploit the properties that are inherent to each kind of entity (locations for toponym disambiguation, and people for person name disambiguation), while additionally making use of features that are standardly used by methods that perform toponym and person name disambiguation simultaneously.

5.1.1 Toponym disambiguation

The method proposed for toponym disambiguation is weakly-supervised. Unlike most previous approaches to disambiguate toponyms, it uses both semantic and geographic features, thus combining their strength. The proposed features are mostly based on features used in either geographically-based toponym resolution systems or semantically-based entity linking systems. To the best of my knowledge, this is the first method that takes the historical component into account: the local semantic feature of the method selects the candidate whose historical context (i.e. the sentences in the Wikipedia article that refer to the same period the collection spans) has the highest similarity to the

context of the toponym in the news article, whereas the global semantic feature selects the candidate whose historical context is overall most similar when considering the rest of candidates of the rest of the toponyms in the text.

I also proposed a new strategy to build a resource that assists in a fast and robust manner in the task of disambiguating toponyms from historical documents. The resources (I use the plural, as a different one is required for each working language) are based on the Wikipedia version of the language in question and complemented with geographical information from GeoNames as well as from the Wikipedia versions in the other languages. This knowledge is exploited to extract all locations and find alternate names for them. Besides, for each location geographic information is stored (latitude and longitude, population, number of Wikipedia inlinks, and country it belongs to, if any) as well as semantic knowledge (the context words, mostly extracted from the historical contexts of each Wikipedia article matching the period of the collection). The creation of such a resource allows a fast and robust retrieval of the knowledge needed to disambiguate toponyms.

In order to assess the performance of the method, five datasets of historical newspapers were created from scratch and annotated, which added to an existing corpus of historical military reports. The five new datasets were annotated by students following clear and principled guidelines in an entity linking manner: mentions of toponyms in texts were asked to be linked to Wikipedia articles in the Wikipedia version in the same language as that of the text where the mention is found. Unlike the already existing dataset (WOTR), the annotation was performed from an entity linking perspective for reasons of practicality: coordinates are straightforward to extract from Wikipedia links, but not viceversa. Therefore, by annotating toponyms with their corresponding Wikipedia URL, I was making sure that they could be used to compare the performance of the method both against state-of-the-art entity linking methods and toponym resolution methods. The five new datasets and the already existing one are in three different languages (English, German, and Dutch) and consist of historical news articles. The collections, though belonging to similar or the same genres, have inherent differences: one collection has a regional scope, two of them have a national scope, and two of them have a national scope but are based in colonies. As mentioned, this is a weakly-supervised method which needs tuning several parameters. The evaluation results for

each dataset hint that different characteristics might be determinant for choosing the optimal parameters and thus obtaining the best performance of the method.

The proposed method performed well when compared both with entity linking and toponym resolution state of the art methods. For the evaluation, a wide range of metrics was used, from perfect Wikipedia URL matches to accuracy at 16 and 161 kilometers, and mean and median error distances.

5.1.2 Person name disambiguation

The proposed method for person name disambiguation explores the relationship between a name ambiguity and the amount of different entities that can be referred to by the same name. A strategy to quantify person name ambiguity is presented, which is crucial for the good performance of the method. There exist very few approaches that take into account ambiguity of a person name to determine how many different entities it can refer to. In the method proposed in this thesis, name ambiguity works as a sort of valve that controls which documents mentioning the same query name are similar enough to be considered as corresponding to the same entity. Person names that are less ambiguous are more permeable: less evidence is needed for two documents mentioning the same person name to be considered as corresponding to the same entity. Conversely, person names that are more ambiguous are more impermeable: more evidence is needed for two documents mentioning the same person name to be considered as corresponding to the same entity.

Evidence is represented as the number of features that need to be true. These features include standard ones in these kinds of tasks such as cosine similarity, but also a novel one is introduced: social network similarity, which is based on the number of nodes in common between the social network representations of two documents. A penalty function is defined to lower the probabilities of two documents belonging together given a certain number of person names in common whenever these are ambiguous. This measure is introduced to penalize those instances in which the person names in common in two documents are not clear indicators that they correspond to the same entity. To the best of my knowledge, this is a novel idea that has never been exploited in previous person name disambiguation approaches.

The method has proved to be competitive in different languages (has been quantitatively tested in English and Italian) and throughout different collections without

need to retrain it. It does not require specific resources such as a knowledge base of people information or other kinds of expensive resources. It is easily portable and can easily be adapted to different datasets and several different languages without the need of learning new parameters. Finally, 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.

5.1.3 Entity-centric text mining in the digital humanities

This thesis also makes a tangible contribution to the field of digital humanities. I have already motivated the need for an improvement of text mining strategies from plain text given the large amounts of new historical materials that are made available to historians on a daily basis. In this thesis, two new methods have been proposed that can offer great assistance to mine historical collections from an entity-centric perspective.

The toponym disambiguation method has been specifically designed to work optimally with historical texts. It receives as input plain texts and returns all mentioned toponyms identified and linked to their respective entry in Wikipedia and, more relevantly maybe for scholars, to their geographic coordinates. This allows a map-based visualization of the collections, therefore allowing the historians to explore a collection by the locations mentioned in it and in which they are interested. Even though it is not the first method that attempts at resolving the toponyms of historical texts, it is, to my knowledge, currently the only one that has been conceived with this particular aim and which combines the use of geographic and semantic features to boost its performance. The good performance of the method on six historical collections and in three different languages shows that the approach is promising.

Finally, the person name disambiguation method has been designed to perform well in historical newspaper collections. The method returns automatically constructed social networks of disambiguated person entities from news articles. To the best of my knowledge, it is the first method that automatically builds social networks of disambiguated persons from large collections of unstructured historical data. The advantages of obtaining a person-centric data visualization of a given collection are numerous, as they allow a deep exploration of the original sources from a social perspective. The historian can have an overview of the collection easily and then explore it at ease: the

different nodes, their centrality in the collection, their relations, the information contained in news articles where both are mentioned, etc. A case study was conducted with the assistance of a historian to assess the value of such an approach for historical research. The outcome was very positive.

5.1.4 Publications

Much of the work presented in this thesis has been published (or will be published in the near future) in peer-reviewed international conferences and workshops in the fields of language technology and digital humanities. The references to the papers are the following:

- Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder. Laboratories of community: how digital humanities can further new European integration history. In *Social Informatics – SocInfo 2014 International Workshops, Revised Selected Papers, Lecture Notes in Computer Science (LNCS)*, pages 284–293, Barcelona, Catalonia, Spain, 2015. Springer.

- Mariona Coll Ardanuy and Maarten van den Bos. Entity centric historical text mining: A people-centric approach to modern Dutch religious history. In *Proceedings of DHBenelux 2016*, University of Luxembourg, Luxembourg, 2016.

- Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder. You shall know people by the company they keep: Person name disambiguation for social network construction. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '16, pages 63–73, Berlin, Germany, 2016.

- Mariona Coll Ardanuy, Jrgen Knauth, Andrei Beliankou, Maarten van den Bos, and Caroline Sporleder. Person-centric mining of historical newspapers collections. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries 2016*, TPDL '16, pages 320–331, Hannover, Germany, 2016.

- Mariona Coll Ardanuy and Caroline Sporleder. Weakly-supervised toponym disambiguation in historical documents using semantic and geographic

features. In *Proceedings of Digital Access to Textual Cultural Heritage*, DATeCH '17, 2017 *(forthcoming)*.

5.2 Future work

Naturally, despite the contributions, there remain some open questions and possible future research directions related to the task of disambiguating toponyms and person names from historical texts. I summarize the possible further steps for each of the tasks in the following paragraphs.

Toponym disambiguation I envision several possible directions for future work in the proposed research on toponym disambiguation:

- I would like to study the role of two more features in the overall disambiguation method: the type of location (e.g. city, river, ocean, etc.) and the hierarchy of the location (e.g. 'Paris/France/Europe'). The first can be extracted easily from GeoNames, where the types are a closed set and are always in English, but it is not as straightforward in the case of Wikipedia as, even if the location page has an infobox, many more location types are allowed and change from language to language.[1] If the Wikipedia page does not have an infobox and its GeoNames corresponding entry is not found, the type of location would have to be extracted either from the categories of the page or from the introductory paragraph of the body of the article. Regarding the hierarchy of locations, this is a characteristic that can be found in several gazetteers and that some approaches have exploited. It would be interesting to see to what degree its use could contribute to the good performance of the method proposed in this thesis.

- A location has been defined to be an entity with coordinates that appears in Wikipedia. This has proved not to be a too restricting approach. However, it would be interesting to explore whether a finer granularity could be achieved with the assistance of external resources. Names of streets, neighborhoods, and stations (to mention just some of the location types that have been in most cases

[1] For example, the infobox of the city of Liège in Belgium describes it as a 'Belgium Municipality' whereas the infobox of the city of Lubumbashi in the Democratic Republic of the Congo describes it as a 'Settlement'.

disregarded by this approach) are extremely ambiguous. However, because of their high ambiguity, they tend to have clear disambiguation clues in the near context.

- Whereas a shallow analysis of the different historical collections and their potential relation to the weighting of the different parameters has been performed, it would be interesting to explore whether patterns exist of feature combinations and parameter weightings which can be inferred from the intrinsic characteristics of a given collection. To be able to do so, though, more annotated collections would probably be needed and they should be analyzed by experts in the field. Similarly, closeness is a subjective concept: what *near* is depends on each person's perspective of space and, from a collective perspective, on the readership scope of the text. The Prussian collection had a nationwide scope, whereas the Belgian collection had a very local scope. The distinction between *near* and *far* in the manner is presented in this thesis corresponds to the will of finding a measure of closeness that accommodates most corpora. It would be interesting to further explore to what degree this distinction can affect the outcome of this method when applied to the different collections.

- Finally, it would be interesting, in order to improve the visualization, to define locations by means of the polygons that shape them. If hierarchy was one of the characteristics of the knowledge base, polygons could be approximated from all the locations that are inside another one. Working with polygons would allow dealing with compositional geographic descriptions such as '30 kilometers south of London', where the polygon should occupy the extension of the correct London referent and an approximated area of 30 kilometers to the South.

Person name disambiguation There are several directions for future work:

- In the proposed approach, social similarity does not exploit the weight of the edges. I would like to study whether it may have an impact in the performance of the method, i.e. whether nodes that are closer in the text to the query name are more informative that nodes that are farther from it.

- That person names can be highly ambiguous has already been discussed and stressed throughout this thesis. However, person names have another characteristic, that they can be referred to through different alternative lexicalizations. In the method, I have only dealt with names that largely follow the most conventional form in English, Dutch, and Italian, but alternative cases should be also considered in future research. I worked on the assumption that most person names are often introduced by a first and last name, at least once in the article (often the first time the person is mentioned). However, this is not necessarily always like this, especially so in the case of public figures, who are well-known by the readers and therefore require less introduction. This brings us to the next question: whether person names of public figures should be dealt in a special manner, as they are very likely to appear in many more documents and therefore possibly in many more social networks. In the future, it would be interesting to explore whether the fact that a person name is borne by a well-known figure can produce an impact on the performance of the method in the same way ambiguity does.

- Finally, it would also be interesting to explore other languages that do not follow the same naming conventions of English, Dutch, and Italian, such as Spanish, Icelandic, Russian, or languages that follow the Eastern order of placing the last name before the given name. It would be also interesting to explore whether ambiguity in person names is more common in some languages than others, and to observe how this could impact on the final result.

I will end this section with a future step envisaged for both methods. Both the toponym and place name disambiguation methods that have been proposed in this thesis are conceived to assist historians who want to dive into a large collection of unstructured text. They allow the visualization and exploration of collections from the perspective of the people and locations mentioned in them, through their representations as sets of social networks and maps. As a future step, the information that is contained in the networks and the map, which is at the moment not stored anywhere and is therefore lost after the historian has finished exploring the collection, could be used to feed a knowledge base. The toponym and person name disambiguation methods are at the moment clearly independent works. It would be interesting in the future to combine

them for the purpose of linking in this manner the people and locations of any digitized collection of historical texts.

Bibliography

[1] Dirk Ahlers. Assessment of the accuracy of geonames gazetteer data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 74–81, 2013. 49

[2] Reema Al-Kamha and David W. Embley. Grouping search-engine returned citations for person-name queries. In *Proceedings of the 6th ACM International Workshop on Web Information and Data Management*, WIDM '04, pages 96–103, 2004. 24

[3] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280, 2004. 14, 43

[4] Mariona Coll Ardanuy, Jürgen Knauth, Andrei Beliankou, Maarten van den Bos, and Caroline Sporleder. Person-centric mining of historical newspapers collections. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries 2016*, TPDL '16, pages 320–331, Hannover, Germany, 2016.

[5] Mariona Coll Ardanuy and Caroline Sporleder. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature at EACL 2014*, CLfL '14, pages 31–39, Gothenburg, Sweden, 2014. 29

[6] Mariona Coll Ardanuy and Caroline Sporleder. Weakly-supervised toponym disambiguation in historical documents using semantic and geographic

features. In *Proceedings of Digital Access to Textual Cultural Heritage*, DATeCH '17, 2017 *(forthcoming)*.

[7] Mariona Coll Ardanuy and Maarten van den Bos. Entity centric historical text mining: A people-centric approach to modern Dutch religious history. In *Proceedings of DHBenelux 2016*, University of Luxembourg, Luxembourg, 2016. 111

[8] Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder. Laboratories of community: how digital humanities can further new European integration history. In *Social Informatics – SocInfo 2014 International Workshops, Revised Selected Papers*, Lecture Notes in Computer Science (LNCS), pages 284–293, Barcelona, Catalonia, Spain, 2015. Springer. 28

[9] Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder. You shall know people by the company they keep: Person name disambiguation for social network construction. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '16, pages 63–73, Berlin, Germany, 2016.

[10] Javier Artiles, Enrique Amigó, and Julio Gonzalo. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 534–542, Singapore, 2009. Association for Computational Linguistics. 24, 103

[11] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 64–69, Prague, Czech Republic, 2007. 24

[12] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. WePS-2 evaluation campaign: Overview of the web people search clustering task. In *Proceedings of the 2nd Web People Search Evaluation Workshop at the WWW Conference*, WEPS '09, 2009. 6, 108

[13] Javier Artiles, Julio Gonzalo, and Felisa Verdejo. A testbed for people searching strategies in the WWW. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 569–570, Salvador, Brazil, 2005. 12

[14] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics*, COLING '98, pages 79–85, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics. 6, 23, 88

[15] David Bamman, Brendan O'Connor, and Noah A. Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 352–361. The Association for Computer Linguistics, 2013. 29

[16] Ander Barrena, Aitor Soroa, and Eneko Agirre. Combining mention context and hyperlinks from wikipedia for named entity disambiguation. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 101–105, Denver, Colorado, 2015. Association for Computational Linguistics. 22

[17] Imene Bensalem and Mohamed-Khireddine Kholladi. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, **6**(6):653–659, 2010. 19

[18] Luisa Bentivogli, Alessandro Marchetti, and Emanuele Pianta. Creating a gold standard for person cross-document coreference resolution in italian news. In *Proceedings of the LREC Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, LREC '08, pages 19–26, Marrakech, Morocco, 2008. 25, 88

[19] Luisa Bentivogli, Alessandro Marchetti, and Emanuele Pianta. The news people search task at EVALITA 2011: Evaluating cross-document coreference resolution of named person entities in Italian news. In *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, Revised Selected Papers*, EVALITA '12, pages 126–134. Springer, 2013. 25

[20] Matthias Blume. Automatic entity disambiguation: benefits to NER, relation extraction, link analysis, and inference. In *Proceedings of the 1st International Conference on Intelligence Analysis*, pages 2–4, 2005. 24

[21] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 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?*, CLIIR '06, pages 17–24, Sydney, Australia, 2006. Association for Computational Linguistics. 24

[22] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 9–16, Trento, Italy, 2006. 21, 26

[23] Robin Buning. Experimental prosopographical network analysis and visualization with 'Early Modern Letters Online'. In *DHBenelux 2016: Conference Abstracts (Poster)*, 2016. 30

[24] Davide Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, **3**(2):16–19, 2011. 18, 19, 64

[25] Davide Buscaldi and Bernardo Magnini. Grounding toponyms in an Italian local news corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 1–5, Zurich, Switzerland, 2010. 18, 47

[26] Davide Buscaldi and Paulo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, **22**(3):301–313, 2008. 19

[27] Florian Cajori. *A History of Mathematical Notations: Two volumes bound as one: II Notations mainly in higher mathematics*, chapter Trigonometry, page 171. The Open Court Publishing Company, 1928. 65

[28] Liwei Chen, Yansong Feng, Lei Zhou, and Dongyan Zhao. Explore person specific evidence in web person name disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning, EMNLP-CoNLL '12, pages 832–842. Association for Computational Linguistics, 2012. 25

[29] Ying Chen and James Martin. Towards robust unsupervised personal name disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP–CoNLL '07, pages 190–198, Prague, Czech Republic, 2007. Association for Computational Linguistics. 24

[30] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, Toronto, Canada, 2010. ACM. 73

[31] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP–CoNLL '07, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 21, 26

[32] Agata Cybulska and Piek Vossen. Historical event extraction from text. In *Proceedings of the 5th ACL–HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 39–43, Portland, Oregon, 2011. 30

[33] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124. ACM, 2013. 74

[34] José de Mendoza y Ríos. *A complete collection of tables for navigation and nautical astronomy: With simple, concise, and accurate methods, for all the calculations useful at sea; particularly for deducing the longitude from lunar distances, and the latitude from two altitudes of the sun and the interval of time between the observations.* Printed by T. Bensley, sold by R. Faulder (etc.), 1805. 65

[35] Grant DeLozier, Jason Baldridge, and Loretta London. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2382–2388, 2015. 20, 32, 66, 74, 75

[36] Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. Creating a novel geolocation corpus from historical texts. In Association for Computational Linguistics, editor, *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, pages 188–198, 2016. 74, 76, 77, 78

[37] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285, 2010. 22

[38] Sourav Dutta and Gerhard Weikum. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. In *Transactions of the Association for Computational Linguistics (TACL)*, **3**, pages 15–28, 2015. 26

[39] Jacob Eisenstein, Brendan OConnor, Noah A. Smith, and Eric P. Xing. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768, 2010. 74

[40] David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, pages 138–147, 2010. 29

[41] Anthony Fader, Stephen Soderland, and Oren Etzioni. Scaling wikipedia-based named entity disambiguation to arbitrary web text. In *Proceedings of the Wikipedia and AI Workshop*, 2009. 21

[42] Manaal Faruqui and Sebastian Padó. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of Kon-*

ferenz zur Verarbeitung natürlicher Sprache, KONVENS '10, pages 129–133, 2010. 59

[43] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL 05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. 59, 96

[44] Jim Giles. Internet encyclopaedias go head to head. *Nature*, **438**(900–901), 2005. 45

[45] Norm Goldstein, editor. *Associated Press Stylebook and Briefing on Media Law: With Internet Guide and Glossary*. The Associated Press, 2003. 97

[46] Chung Heong Gooi and James Allan. Cross-document coreference on a large scale corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL–HLT '2013, pages 9–16, 2004. 23

[47] Swapna Gottipati and Jing Jiang. Linking entities to a knowledge base with query expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 804–813, 2011. 21

[48] Martin Grandjean. Archives distant reading: Mapping the activity of the league of nations intellectual cooperation. In *Digital Humanities 2016: Conference Abstracts*, pages 531–534, Jagiellonian University & Pedagogical University, Kraków, 2016. 29

[49] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with wikipedia. *Artificial Intelligence*, **194**:130–150, 2013. 43, 81

[50] Xianpei Han, Le Sun, and Jun Zhao. Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 765–774, New York, NY, USA, 2011. ACM. 21

[51] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 22

[52] Cornell Jackson. Using social network analysis to reveal unseen relationships in medieval scotland. In *Digital Humanities Conference*, Lausanne, 2014. 28

[53] Lili Jiang, Jianyong Wang, Ning An, Shengyuan Wang, Jian Zhan, and Lian Li. Grape: A graph-based framework for disambiguating people appearances in web search. In *Proceedings of IEEE International Conference on Data Mining*, pages 199–208, 2009. 25

[54] Marinos Kavouras and Margarita Kokla. *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*, chapter Ontologies, pages 71–72. CRC Press, 2007. 14

[55] Zornitsa Kozareva and Sujith Ravi. Unsupervised name ambiguity resolution using a generative model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 105–112, Edinburgh, Scotland, 2011. Association for Computational Linguistics. 24

[56] Saul A. Kripke. *Naming and Necessity*. Harvard University Press, 1980. 14

[57] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 457–466, New York, NY, USA, 2009. ACM. 22

[58] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014. 26, 45

[59] Jochen L. Leidner. *Toponym resolution in text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* PhD thesis, University of Edinburgh, Universal Press, Boca Raton, FL, USA, 2008. 32, 43, 72, 73

[60] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the International Conference on Data Engineering*, pages 201–212, 2010. 19, 44, 47, 64

[61] Patricia Ferreira Lopes, Francisco Pinto Puerto, Antonio Jimenez Mavillard, and Juan Luis Suarez. Seeing andalucia's late gothic heritage through gis and graphs. In *Digital Humanities 2016: Conference Abstracts*, pages 501–504, Jagiellonian University & Pedagogical University, Kraków, 2016. 30

[62] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL '03, pages 33–40, Edmonton, Canada, 2003. 24

[63] Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. Annotating geographical entities on microblog text. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 85–94, Denver, Colorado, 2015. Association for Computational Linguistics. 14

[64] Olena Medelyan, Ian H. Witten, and David Milne. Topic indexing with Wikipedia. In *Proceedings of the AAAI Wikipedia and AI workshop*, 2008. 22

[65] Andrew Mehler, Yunfan Bao, Xin Li, Yue Wang, and Steven Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, **12**(5):765–772, 2006. 47

[66] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM. 21

[67] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM. 69

[68] Giovanni Moretti, Sara Tonelli, and Stefano Menini. Building large persons' network to explore digital corpora. In *Digital Humanities 2016: Conference Abstracts*, pages 625–629, Jagiellonian University & Pedagogical University, Kraków, 2016. 29

[69] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association of Computational Linguistics*, **2**:231–244, 2014. 22, 74

[70] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**:217–250, 2012. 45

[71] Cheng Niu, Wei Li, and Rohini K. Srihari. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 598–605, Barcelona, Catalonia, Spain, 2004. Association for Computational Linguistics. 24

[72] Simon E. Overell. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. PhD thesis, Imperial College London, Department of Computing, 2009. vii, 20, 46, 47

[73] John F. Padgett and Christopher K. Ansell. Robust action and the rise of the medici, 1400–1434. *American Journal of Sociology*, pages 1259–1319, 1993. 28

[74] Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. Name discrimination by clustering similar contexts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 226–237, 2005. 64

[75] Emanuele Pianta, Christian Girardi, and Roberto Zanoli. The TextPro tool suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pages 2603–2607, Marrakech, Morocco, 2008. European Language Resources Association. 91, 96

[76] Octavian Popescu. Person cross document coreference with name perplexity estimates. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 997–1006, Singapore, Singapore, 2009. Association for Computational Linguistics. 25, 93

[77] Octavian Popescu and Bernardo Magnini. IRST-BP: Web people search using name entities. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 195–198. Association for Computational Linguistics, 2007. 24

[78] Teng Qin, Rong Xiao, Lei Fang, Xing Xie, and Lei Zhang. An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 53–60, 2010. 20

[79] Gianluca Quercini, Hanan Samet, Jagan Sankaranarayanan, and Michael D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, 2010. 43

[80] Fatemeh Rahimian, Sarunas Girdzijauskas, and Seif Haridi. Parallel community detection for cross-document coreference. In *Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, WI-IAT '14, pages 46–53. IEEE Computer Society, 2014. 109

[81] Delip Rao, Paul McNamee, and Mark Dredze. Streaming cross document entity coreference resolution. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1050–1058, Beijing, China, 2010. Association for Computational Linguistics. 88, 107, 108, 109

[82] Erik Rauch, Michael Bukatin, and Kenneth Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, **1** of *HLT-NAACL-GEOREF '03*, pages 50–54. Association for Computational Linguistics, 2003. 19, 66

[83] Yael Ravin and Zunaid Kazi. Is Hillary Rodham Clinton the President? disambiguating name accross documents. In *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp '99, pages 9–16, College Park, Maryland, 1999. Association for Computational Linguistics. 23

[84] Kirk Roberts, Cosmin Adrian Bejan, and Sanda Harabagiu. Toponym disambiguation using events. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*, FLAIRS '10, pages 271–276. Association for the Advancement of Artificial Intelligence Press, 2010. 20, 44

[85] Yannick Rochat, Melanie Fournier, Andrea Mazzei, and Frédéric Kaplan. A network analysis approach of the Venetian Incanto system. In *Digital Humanities Conference*, Lausanne, 2014. 28

[86] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, , and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP–CoNLL 12, pages 1500–1510. Association for Computational Linguistics, 2012. 74

[87] Evan Sandhaus. The New York Times Annotated Corpus overview. *Linguistic Data Consortium*, 2008. 88

[88] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Large-scale cross-document coreference using distributed inference and hierarchical methods. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 793–803, Portland, Oregon, 2011. Association for Computational Linguistics. 109

[89] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 127–136. Springer Berlin Heidelberg, 2001. 18, 27, 32, 43

[90] Yang Song, Jian Huang, Isaac G. Councill, Jia Li, and C. Lee Giles. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 342–351, Vancouver, Canada, 2007. ACM. 23

[91] Michael Speriosu and Jason Baldridge. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, **1** of *ACL '13*, pages 1466–1476, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 20, 32, 44, 73, 74

[92] Edward Stratford and Jeremy Browne. Linkedin circa 2000 bce: Towards a network model of puu-kens commercial relationships in old assyria. In *Digital Humanities Conference*, Sydney, 2015. 28

[93] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th international World Wide Web conference*, WWW '07, pages 697–706, Banff, Canada, 2007. ACM. 26, 45

[94] Serge ter Braake and Antske Fokkens. How to make it in History. working towards a methodology of canon research with digital methods. In *Proceedings of the First Conference on Biographical Data in a Digital World*, pages 85–93, Amsterdam, The Netherlands, 2015. 29

[95] Serge ter Braake, Antske Fokkens, and Fred van Lieburg. Mining ministers (1572–1815). using semi-structured data for historical research. In *HistoInformatics2014 – the 2nd International Workshop on Computational History. Social Informatics. Volume 8852 of the series Lecture Notes in Computer Science*, pages 279–283, 2014. 30

[96] Raphael Volz, Joachim Kleb, and Wolfgang Mueller. Towards ontology-based disambiguation of geographical identifiers. In *Proceedings of the Workshop Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web*, I3 '07, 2007. 19

[97] Jan Škvrňák and Adam Mertel. Linking graph with map for the purpose of historical research. In *Digital Humanities 2016: Conference Abstracts*, pages 887–888, Jagiellonian University & Pedagogical University, Kraków, 2016. 30

[98] Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-objective optimization for the joint disambiguation of nouns and named entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL–IJCNLP, pages 596–605, Beijing, China, 2015. Association for Computational Linguistics. 22

[99] Benjamin P. Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT 11, pages 955–964, Portland, Oregon, 2011. Association for Computational Linguistics. 74

[100] Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. Automatic annotation suggestions and custom annotation layers in webanno. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 91–96, Baltimore, Maryland, 2014. Association for Computational Linguistics. 36

[101] Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '13, pages 1–6, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 36

[102] Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Person name disambiguation by bootstrapping. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 10–17, Geneva, Switzerland, 2010. ACM. 24, 25

[103] Roberto Zanoli, Francesco Corcoglioniti, and Christian Girardi. Exploiting background knowledge for clustering person names. In *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, Revised Selected Papers*, EVALITA '11, pages 135–145, 2013. 26, 91, 107

[104] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pages 1646–1652, Marrakech, Morocco, 2008. European Language Resources Association. 51

[105] Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. Entity linking with effective acronym expansion, instance selection and topic modeling. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, pages 1909–1914, Barcelona, Catalonia, Spain, 2011. Association for the Advancement of Artificial Intelligence Press. 22

[106] Maayan Zhitomirsky-Geffet, Gila Prebor, and Amichai Feigenboim. Towards a cross-generation social network for jewish sages. In *Digital Humanities 2016: Conference Abstracts*, pages 916–917, Jagiellonian University & Pedagogical University, Kraków, 2016. 29