Context-based Toponym Disambiguation regarding Historical Gazetteer

Bram Schmidt[1]

Supervisors: Victor de Boer[2], Marieke van Erp[3], Rombert Stapel[4]

[1] Master Information Sciences, Vrije Universiteit Amsterdam, S2651459
[2] Vrije Universiteit Amsterdam, Department of Computer Science
[3] KNAW - Humanities Cluster, Digital Humanities Lab
[4] International Institute of Social History

Abstract

Place names are very ambiguous and may change over time. This makes it hard to link mentions of places to their corresponding modern entity and coordinates, especially in a historical context. We focus on historical Toponym Disambiguation approach of entity linking based on identified context toponyms. We analyze the additional descriptions that come with toponym entries of the historical database of the American Gazetteer. These texts contain fundamental information about major places in its vicinity. By identifying and exploiting these tags, we aim to estimate the most likely position for the historical entry and accordingly link it to its corresponding contemporary counterpart.

Therefore, in this case study we examined the toponym recognition performance of state-of-the-art Named Entity Recognition (NER) tools spaCy and Stanza concerning historical texts and we tested two new heuristics to facilitate efficient entity linking to the geographical database of GeoNames. We tested our method against a subset of manually annotated records of the gazetteer.

Results show that both NER tools do function insufficiently in their task to automatically identify relevant toponyms out of the free text of a historical lemma. However, exploiting correctly identified context toponyms by calculating the minimal distance among them proves to be successful as long as the input is of sufficient value. We developed an optimized combined algorithm based on the initial results that achieved a substantially improved recall score. Future search should focus on consulting external knowledge bases to account for outdated place names and implementing data-driven methods to also process non-geographical entities in the context descriptions.

1 Introduction

From carvings in clay tablets from around 1000 B.C. to modern-day GPS systems, people have been drawing and using maps for a long time [15]. They have become essential in deciding upon the fastest route from A to B or finding the nearest supermarket. Considering the data shown on maps is spread (to their geographical location) and thus unsorted, they could be supplemented by gazetteers. These are resources, mostly in the form of a dictionary or directory, that consist of a list of geographical names, accompanied by physical features and historical information [15]. Originally, they took the form of reference books that documented the spelling of places, or more formally called toponyms. Additional context or descriptions can be included, but the emphasis was on the naming. Since place names can change over time or whole places disappear, the index of place names for any region is subject to many changes. That is why modern gazetteers usually provide a good overview of contemporary topography but give inadequate insight into historical transformations [12]. This notion constitutes the underlying foundation of the OpenGazAm project, which provides the context for this Master project.[5]

Whereas most existing digital gazetteers focus on modern place names, we focus on historical data, which makes research more complicated. Namely, when a place and both its name and location are identified, the historical element can be linked to the modern geographic entity but the validity of the place name is still temporally bounded. As such, names are not stable and without checking carefully for timeliness, toponyms can easily be mixed up [6]. A crucial step herein is reserved for the linking of place names from newly added data to that of existing digital gazetteers, a process known as reconciliation. The presence of many places having the same name or, especially in this historical context, the development of place names or even the disappearances of whole places does complicate this process and makes it very ambiguous.

To address these problems, we inspect the additional descriptions that come with the entries of the historical database. These descriptions, in the form of free text, contain fundamental information about administrative regions the place belongs to and mention neighbouring places. As such, these details clarify the relative locations of place records and therefore we assume that the process of disambiguation could be enhanced by taking this crucial information into account. That is to say, if we focus on any mentioned adjacent site and we state that the specific entry must be in its immediate vicinity, we could perform calculations based on the possible coordination pairs to establish the most likely position for the historical entry.

[5] https://clariah.nl/projecten/research-pilots/opengazam

Therefore we performed a **case study that focuses on the disambiguation of historical toponyms**. We examine the functioning of state-of-the-art Named Entity Recognition tools *spaCy*[6] and *Stanza*[7] to the task of identifying descriptive geographical information. Additionally, we assess two newly developed heuristics that process this selected data to calculate and determine which specific modern geographical entity is referred to. We apply these methods to the digitized historical dataset of the American Gazetteer. By facilitating and analyzing toponym recognition and reconciliation with their modern GeoNames referent, we provide **insight in the operation and limitations of applied methods and promote the amplification of digitally available historical gazetteers**. In this way, we improve access to spatio-temporal knowledge and broaden the possibilities for research on global history. The following main research question was formulated:

How can the process of toponym disambiguation with respect to historical works be improved by interpreting free text of place record lemmas?

In order to help answering the main research question, we defined these subquestions:

- *How can relevant spatial information be identified from the free text of a place record lemma?*
- *To what extent can processed relevant spatial information facilitate successful toponym reconciliation?*

This thesis document is structured as follows. The context is summarized in section 2. Related work is covered in section 3 and the datasets used are described in section 4. In section 5, we describe the first step of recognizing toponyms, followed by their candidate selection (section 6) and the final step of applying context-based heuristics to disambiguate them in section 7. The results of both the recognition and reconciliation are presented and discussed in section 8. Conclusions and future work are outlined in section 9.

[6] https://spacy.io/

[7] https://stanfordnlp.github.io/stanza/ner.html

2 Context

The main resource used for our research project is the American Gazetteer. We use the geographical data that this dictionary contains as a test set for our developed heuristics. The original book was written and compiled by Jedidiah Morse in 1797 and contains an alphabetical index of 6852 toponyms, i.e. names of geographic entities (states, counties, cities, mountains, rivers, etc.). These places are all located on the American continent and the West-Indies islands [16]. Since the names of these places are often ambiguously interpreted and their locations are only partly established, more accurate annotations are required.

By enhancing this disambiguation process and promoting the delivery of an annotated historical dataset, we contribute to the *World-Historical Gazetteer (WHG) project.*[8] Researchers and developers from different institutions cooperate in this joint project. The aim of the project is to create an online gazetteer of historical places for the period after 1500 CE. Specific objectives comprise the creation of a set of standards for worldwide place documentation, a consistently formatted places database, the facilitation of incorporation of additional data and introducing a user interface to apply the gazetteer functionality [14].

A crucial step is reserved for the linking of place names from newly added data to that of existing digital gazetteers. For this, the WHG focuses on the geographical database of GeoNames [9] and DBpedia [10] and already provides services for the Getty Thesaurus of Geographic Names (TGN) [11] and the place authority resources of WikiData [12]. With regard to our research project, we selected GeoNames as modern gazetteer for querying place names. We prefer to use GeoNames because the larger part of the manual references in the available dataset concerns annotations by GeoNames ID code. It is open to wisdom of the crowds. This means that it allows volunteered data, thus everyone can contribute to it [7]. A disadvantage of (only) using GeoNames as annotation provider is that we are limited to the data present in their web service. However, it is one of the largest gazetteers with a size of over 25 million geographical names and contains a great variety of entities, along with relevant metadata such as coordinates and alternative names [1].

GeoNames provides an API, which we used for querying the database and obtaining the necessary IDs and coordinates. In section 6.1 the query approach will be further explained.

[8] http://whgazetteer.org/
[9] https://www.geonames.org/
[10] https://wiki.dbpedia.org/
[11] https://www.getty.edu/research/tools/vocabularies/tgn/
[12] https://www.wikidata.org/wiki/

3 Related work

This section will cover the most relevant methods for toponym disambiguation. The main task of all methods is to automatically label geographic mentions in plain text, similar to how a human would. For this, multiple approaches have been developed, that are mostly based on assigning scores to possible referents and selecting the most suitable option [9]. The methods to be discussed can be used for dual purposes. The general format is the selection of the right toponym referent among a set of candidates. However, the form of their output differs. Toponym resolution aims to find the geographic coordinates of mentioned localities, whereas entity linking connects these locations to their referents in a knowledge base by their identifier. [4]. Since we strive to annotate historical toponyms with their modern GeoNames ID, our reconciliation process is a matter of entity linking. After all, it is insufficient to only find the correct coordinates, as these could neither be processed by the WHG nor be properly evaluated.

3.1 Toponym Disambiguation methods

Toponym disambiguation approaches are grouped into three categories: map-based methods, knowledge-based methods and data-driven methods.

Map-based methods. Map-based methods use explicit representations of places on a map and are very sensitive to context [9]. The general approach gathers all toponyms occurring in the same document, paragraph or sentence, considers their possible locations and selects the most likely location based on a varying geographical calculation. As such, they do not need any information other than the coordinates of the mentioned places [8]. For instance, Smith and Crane (2001) developed a rule-based strategy and their main disambiguation was based on the distance between each possible referent location and the calculated geometric centre of all other mentioned unambiguous toponyms in the text [19]. Buscaldi and Magnini (2010) extended this approach and also took the geographical source of the text into consideration [10]. In our case, the specific context could either be formed by the entry toponyms of the document, the toponyms mentioned in every lemma or a combination of both. Since the document comprises entities spread over and around the American continent, the geographical variation of the whole document is quite high and therefore considering all these entities would not make sense. The lemma itself, however, is more precisely defined and thus provides a good starting point for applying map-based disambiguation methods.

Knowledge-based methods. Knowledge-based methods exploit external knowledge sources for specific properties and are most commonly used [3]. These methods resemble the more generic Word Sense Disambiguation and mostly use a non-geographical approach to geographic references that is based on either external facts or the correlation between the sense of a given word and its context. [8] The used knowledge sources could be population data or ontologies, for example. Rauch et al. (2003) observed the prominence of locations, i.e. conditioned disambiguation, based on how often locations are referred to by toponyms [18]. Also, the position of locations in taxonomies and hierarchical tree structures have been studied [2] [5]. For our research, using external knowledge might advance selection of and reconciliation to the right candidate. Hence we decided to investigate the effect of including relevance as ordering principle.

Data-driven methods. Data-driven methods are based on machine learning techniques and rely on annotated data. Because of the lack of geographically tagged data and the fact that unseen toponyms are hard to classify they are not frequently used in relation to toponym disambiguation. In contrast to knowledge-based models, these methods could also exploit non-geographical content [8]. For example, a mentioned person or organisation could be based at a specific place and therefore be an important clue. Since the amount of annotated data is growing, the popularity of these methods in increasing [3]. These methods are less relevant for our research, as most descriptions only comprise geographical entities and we strive to develop an overarching, consistently applicable method. Nonetheless, for further research it would be interesting to apply these methods to the specific entries that contain extensive historical anecdotes.

3.2 Evaluation

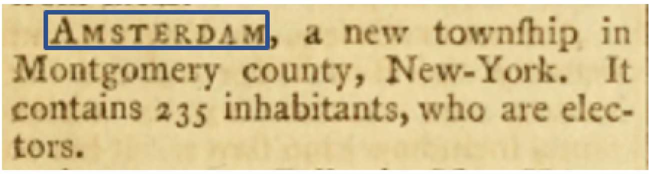
For a toponym disambiguation method to be successful, it should extract unique geographic identifiers to the specific locations that are specified within a text. Hereby it must overcome the problem of location ambiguity and renaming [4]. To evaluate these methods, a gold standard or reference corpus of the data source is required. In this dataset, all geographic names are manually annotated. These annotations can then be compared to the automated system's output to measure the accuracy of the method [13]. In recent toponym disambiguation method research, the primary measures are precision and recall. Precision is calculated as the number of correctly disambiguated toponyms divided by the number of disambiguated toponyms. Recall is calculated as the number of correctly disambiguated toponyms divided by the number of toponyms in the test collection. The harmonic mean between these measures, and as such the measure of overall accuracy, is termed the F-measure or F1-score [9].

4 Data

In this section, we describe the framework of the main dataset, the preprocessing steps taken and the selection of the gold standard.

4.1 Dataset Framework

The primary dataset used for this study is the digitized version of the historical work of the American Gazetteer. Every record consists of a toponym, the label under which the entity is identified, accompanied by a lemma that specifies the type of entity, the administrative region(s) it belongs to and its demographics, as shown in Fig. 1. The mentioned entity types range from regional areas (counties, states) to more concentrated localities (towns, townships, villages).



AMSTERDAM, a new townſhip in Montgomery county, New-York. It contains 235 inhabitants, who are electors.

Fig. 1. Example of a place record within the American Gazetteer

Context Structure In addition to the mentioned basal metadata, the majority of lemmas within the dataset contain supplementary context descriptions and clues regarding the location of the entry toponym. This condition forms the basis of our disambiguation approach. The context description in the form of free text takes many forms and varies in length, composition and usefulness. Next to the corresponding county and state that are mentioned at the beginning of every lemma, which already provide constructive bounding clues, additional location cues are often provided further down the description. As illustrated in Table 1, the formulation of these cues ranges on a scale of specificity, from vague directions with respect to adjacent places to relative positions and distances or even concrete coordinates. However, not every mentioned toponym within the free text may be considered relevant with regard to the location of the entry toponym. Occasionally toponyms are called in context because of alternative associations, as represented in Table 2. This distinction complicates the recognition of relevant tags and as such the facilitation of disambiguation within our research.

Type of description	Textual example
Administrative region	*“..in Washington co. New-York..”, “..is the capital of..”*
Nesting	*“..situated in/on..”, “..lies on the coast of..”*
Linear demarcation	*“..lies between ... and..”, “..lies on the road from ... to..”*
Vague direction	*“..above..”, “..opposite to..”,“..situated near..”*
Specific direction	*“..bounded N. by”, “..on the W. side of..”*
Specific distance	*“..a mile and an half from..”*
Specific direction and distance	*“..4 miles S. S. E. of..”*
Provided coordinates	*“..lies in S. lat. 3. 56. W. long. 32. 43..”*

Table 1. Examples of geographically relevant toponym-mentioning descriptions

Type of description	Textual example
Former placename	*“..first called Morristown..”*
Similarity	*“..these are similar to those in Tennessee and Virginia..”*
Comparison	*“..50 feet higher than the fall of Niagara..”*
Dependency	*“..belonging to Spain..”*
Historical	*“..where the settlers of New-England first landed..”*
Trade-related	*“..produce better fruit than in Portugal..”,* *“..it exceeds Port au Prince in the value of its productions”*

Table 2. Examples of geographically irrelevant toponym-mentioning descriptions

4.2 Preprocessing

To establish a workable dataset, some preprocessing steps were taken. The pages of the original source text of the historical work were already scanned and converted into machine-encoded text using Optical Character Recognition (OCR). This resulted in a digital text file containing all place records. Additionally, the text file had been converted into an annotated spreadsheet. Nonetheless, the dataset that we had at our disposal still involved spelling mistakes within its records. To improve reliability, we manually corrected noticed errors in the lemma descriptions.

4.3 Gold Standard

Part of the dataset was manually checked and annotated with the appropriate location ID. Likewise, the correct coordinates for these entries are established. Entries for which no applicable referent could be found were filtered out. This subset of 197 manually annotated records serves as the Gold Standard for our research project.

5 Toponym Recognition

In this section, we will initiate the approach of our research project. Fig. 2 represents an overview of the applied disambiguation pipeline. As explained in section 4.2 and shown in the overview, the historical document of the American Gazetteer was digitized and pre-processed. Subsequently, we analyzed the lemmas and identified the mentioned neighbouring geographical entities out of the free text. In the overview, this step is the first of the pipeline. It has been marked in red and relates to sub-research question 1. The steps taken within this process are further explained in the following subsections.

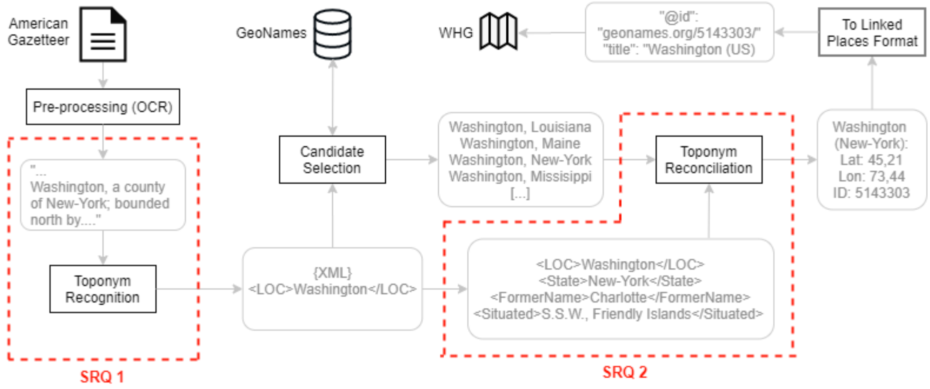
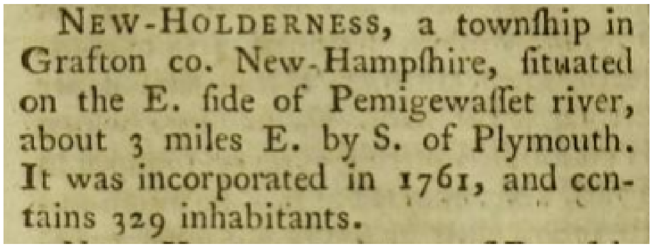

American Gazetteer
GeoNames
WHG
"@id": "geonames.org/5143303/" "title": "Washington (US)
To Linked Places Format
Pre-processing (OCR)
"... Washington, a county of New-York; bounded north by...."
Toponym Recognition
Candidate Selection
Washington, Louisiana
Washington, Maine
Washington, New-York
Washington, Missisippi
[...]
Toponym Reconciliation
Washington (New-York): Lat: 45,21 Lon: 73,44 ID: 5143303
{XML} <LOC>Washington</LOC>
<LOC>Washington</LOC>
<State>New-York</State>
<FormerName>Charlotte</FormerName>
<Situated>S.S.W., Friendly Islands</Situated>
SRQ 1
SRQ 2

Fig. 2. Overview of the disambiguation approach, represented as a processing pipeline.

5.1 Context Toponym Identification

As mentioned in section 4, all entries of our dataset feature a rich description, stating the administrative regions they belong to and their neighbouring geographical entities. Although the scope of the description varies per entry, the county and state in question are almost always mentioned. These are essential tags that need to be detected and processed to facilitate disambiguation. Since the original gazetteer is a 200-year-old document with a rather ambiguous font, small recognition errors remained after digitization. Besides, to limit space, the descriptions contain abbreviations, as seen in Fig. 3. Both facts complicate the process to adequately identify relevant toponyms. Hence we revised the free text of the dataset in advance by filtering out any detected OCR errors and eliminating the following abbreviations:

- 'N.' → North, 'S.' → South, 'E.' → East, 'W.' → West
- 'co.' or 'C.' → county, 'R.' → River, 'I.' → Island



NEW-HOLDERNESS, a townſhip in Grafton co. New-Hampſhire, ſituated on the E. ſide of Pemigewaſſet river, about 3 miles E. by S. of Plymouth. It was incorporated in 1761, and ccntains 329 inhabitants.

Fig. 3. As seen in this fragment, some phrases are printed unclear ('fitwated' → 'situated') and indications of counties and directions are abbreviated

To recognize relevant toponyms out of the revised free text, different methods are available. The most obvious is the use of Named Entity Recognition (NER) software. Such tools recognize the separate entities and label them with their according type. We used both *spaCy* and *Stanza* for this purpose. After the NER processor has labelled all entities of the free text, the relevant types must be selected. With regard to our research project, we are only interested in geographical entities, which means that among other things mentioned persons or nationalities are filtered out. Although we decided to exclude this information from our research project, it may become suitable for further research as any mentioned unique non-geographical entity could potentially narrow results as well. For this project, we decided to solely include geopolitical entities (countries, cities, states) and other location tags (mountain ranges, bodies of water).

Although the performance of both these NER tools has been proved to be certainly sufficient with F1 scores of around 85% (85.4 and 88.8, respectively) in previous work [17], the tokenization and identification of geographical entities are never completely flawless. Therefore we also annotated the gold standard part of the dataset with manually identified context toponyms. For these 197 entries, we have extracted and listed the correct context toponyms from the free text ourselves, without acquiring the corresponding GeoNames IDs or coordinates, since this is part of the disambiguation phase. This manual toponym identification allows us to evaluate the operation of spaCy and Stanza as a contribution to the overall disambiguation approach.

5.2 Context Toponym Selection

After having identified the relevant types of entities out of the free text, the relevant specific entities must be selected. Indeed, not every mentioned site is mentioned in the context of adjacency. As seen in table 2, in some cases, places are mentioned because of similarity, dependency or historical association. By manually reading and checking lemmas we concluded that the most relevant information is often mentioned in the first sentences of the free text. In the case of towns, townships, post towns or other single point entities generally the corresponding county in question is mentioned first and the state is mentioned second. When the entry concerns a county, the corresponding state is mentioned first. With regard to other, more extensive locations (rivers, lakes, capes, islands etc.) there is no fixed listing pattern to be discovered. After the administrative regions and neighbouring towns occasionally follows a varying historical description of the demography or trading position of the site. Taking this into consideration, we decided *to only extract and process the first five identified context toponyms for every entry.* If fewer toponyms were identified we took all of them into account. This not only prevents mismatching by use of irrelevant tags but also prevents the computing capacity from being overloaded by taking too many possible locations into account.

6 Candidate Selection

After having bounded the collection of context toponyms for every lemma, we pursue to use the coordinates of these tags to predict the location of the entry toponym based on assumed adjacency. Therefore we compiled a selection of possible GeoNames entities to which every identified context toponym corresponds. Likewise, it is necessary to compile such a selection of candidates for the entry toponym itself. After all, the most apparent entry toponym candidate must be determined based on this selection. In the overview of the approach (Fig. 2), this is the second step of the pipeline. In this section, the steps and considerations are further explained.

6.1 Query filtering

Before determining the right entity to which the entry toponym refers, we need to make a selection of candidates for both the entry toponym and the context toponyms to elect among. As a starting point, we perform a GeoNames query using the fully written out toponym name of the lemma as search term. To increase reconciliation chances we decided to *implement a fuzzy search with factor 0.8* (Levenshtein distance 1), as some place names are spelt slightly different nowadays (e.g. Fannet → Fannett, Falsington → Fallsington, Followfield → Fallowfield). This takes into account minor spelling changes, but at the same time does not allow overly deviated results. To narrow down the number of irrelevant results we only *select entities that lay between longitude -135 West and -20 East.* Hereby we filter out any results outside of the geographic scope of the database (roughly the large Pacific/Atlantic area around the American Continent). The returned query results are sorted by relevance, in descending order as determined by the GeoNames API. To maximize chances of finding the right entry candidate among the returned GeoNames records we select *the first 1000 results for the entry toponym*, which is the maximum that GeoNames allows. If fewer records are known, we include all records. In the case of context toponyms, often more eminent, distinctive sites are mentioned and that is why we decided to take *the first five results for the context toponyms* and their according coordinates for further processing. This limit both prevents irrelevant entities from negatively impacting results and prevents overloading the computing capacity because of accounting for an unnecessarily large number of results. An overview of the exact query settings is provided in Table 3.

Intention	Filter
Only include relevant coordinates	Demarcate area -35W:-20E
Only include relevant entity types	Include Feature Class A, H, P, S, T, L, U
Accounting for spelling changes	Fuzzy = 0.8
Sort results	Order by relevance
Delimit search records	Entry toponym: n = 1000 (max), Context toponym: n = 5

Table 3. Overview of the (initial) GeoNames API query settings

7 Toponym Reconciliation

After the collection of candidates for both the entry itself and the geographical entities in its free text, it is time for the final step of toponym reconciliation. In the overview of the approach (Fig. 2) this is the third step of the pipeline, it has been marked in red and relates to research question 2 and 3. In this section, the applied disambiguation heuristics will be explained in more detail.

7.1 Baseline

Random candidate First, we created a random baseline algorithm as a simple look-up approach. This baseline was implemented by querying the concerning place name and coupling the place record to one of the results returned by the API query. The settings as described in Table 3 are taken into account to guarantee an equal starting point. From the fixed list of returned records, a candidate is picked by applying a randomly generated index number. This means that the initial reconciliation to the chosen candidate is solely based on similar naming, without any interference based on relevance or position.

First candidate Secondly, we used a more validated approach by building upon the GeoNames' sorting algorithm that is based on relevance. Hereby, instead of random picking, the first result out of the returned query list is selected. We assume that this candidate would give higher chances of an accidental correct match since according to GeoNames this entry is considered the most relevant, and as such the most plausible result to the specific query. Especially in the case of unique place names (e.g. 'Santa Fe d'Antiochia') or toponyms that contain explicitly mentioned entity types (e.g. 'Fort Edward', 'Flint river') the search process is already profoundly optimized, which increases the chance of efficiently encountering the correct result and thus enhances the chance of the correct result appearing at the top as well.

7.2 Heuristics

As stated in section 5 and 6, we assume that there is a high likelihood that a toponym lies in the vicinity of the places mentioned in its description. Accordingly, after having established the candidates for both the entry toponym and its context toponyms, we suggest exploiting the selected entities in a complementary way to determine the most likely toponym candidate for every entry. We used different heuristics for this determination and will explain them in more detail.

Minimal distances The Minimal Distances heuristic relies on the foundation that any toponym is generally positioned on a relatively small *'as the crow flies'* distance from the mentioned sites in its context. Therefore, this heuristic measures and constantly compares the sum of the distances from the entry toponym to the separate context toponyms. For every entry toponym candidate, the closest context toponym candidates are determined by calculating the minimal possible distance. These distances are determined by using the Haversine formula, which takes into account the curvature of the earth and herewith measures the great-circle distance between two points [11]. The determined smallest distances are then all added up. By comparing these total numbers for every entry toponym candidate the most likely candidate is selected by opting for the smallest total distance.

Smallest Polygon The Smallest Polygon heuristic suggests that the toponym location could be established by calculating the minimum area that spans the entry toponym and the selected toponyms out of its context. To achieve this, the candidate positions of the selected context toponyms could, together with the candidate positions of the entry toponym itself, become potential vertices of a polygon to be constructed. For each of the combinations of the coordinates of both the entry toponym candidate and the selected context toponym candidates, the system creates a polygon that covers these points and by continual selection, it encounters the smallest possible area. After having determined this shape, it verifies which of the vertices is formed by the entry toponym candidate (the other vertices are formed by context toponyms). Consequently, the toponym candidate contributing to the final shape is selected as the most likely option for reconciliation.

8 Results and Discussion

In this section, we present and discuss our results. These results are measured by applying both the recognition tools and disambiguation heuristics to the 197 records that were manually annotated as a gold standard.

8.1 Recognition Evaluation

The free-text of the gold standard entries was processed both manually and automatically in order to test the operation of the named entity recognition tools. By comparing the outcome for both tools to the manually processed tags we have evaluated their accuracy. As shown in Table 4 both tools performed rather moderately at this task, with a small margin in favour of Stanza. By and large, both processors are able to tag about half of the toponyms correctly. We will highlight some examples to illustrate the complexity and deficiencies in functionality.

Processor	Precision	Recall	F1-score
spaCy	0.59	0.49	0.54
Stanza	0.65	0.52	0.58

Table 4. Precision, recall and F1-score with regard to the context toponym recognition task of both tested NER-tools

The entry fragments of Fairfield (1) and Falmouth (2) contain fairly ordinary context descriptions, mentioning both its corresponding county and state and other adjacent sites. As indicated in Table 5 and 6, Stanza is able to correctly distinguish most of the mentioned context toponyms, while spaCy ignores some tags and also incidentally marks wind directions as relevant tags.

Fairfield, a plantation in ***Lincoln county*** *district of* ***Maine****, on the South East bank of* ***Kennebeck River*** *South of* ***Canaan****, and opposite* ***Hancock*** *; about 17 miles from* ***Pittstown****, and 7 from* ***Fort Halifax****. It contains 492 inhabitants, and is 225 miles North East of* ***Boston****.* (1)

Falmouth, a township in ***Hants county Nova-Scotia*** *; situated on the South East side of the* ***Basin of Minas****, oppofite* ***Windsor****, 28 miles North West of* ***Halifax****.* (2)

Processor	Identified tags
Stanza	[’Lincoln county’, ’Maine’, ’Kennebeck River’, ’Canaan’, ’Hancock’, ’Pittstown’, ’Fort Halifax’, ’Boston’]
spaCy	[’Lincoln county’, ’Maine’, ’Canaan’, ’Pittstown’, ’Boston’]
Manual (GS)	[’Lincoln county’, ’Maine’, ’Kennebeck River’, ’Canaan’, ’Hancock’, ’Pittstown’, ’Fort Halifax’, ’Boston’]

Table 5. Toponym recognition results for free text of Fairfield (1).

Processor	Identified tags
Stanza	[’Hants county’, ’Nova-Scotia’, ’the Basin of Minas’, ’Windsor’]
spaCy	[’Hants’, ’south east’, ’Basin’, ’Minas’, ’Halifax’]
Manual (GS)	[’Hants county’, ’Nova-Scotia’, ’Basin of Minas’, ’Windsor’, ’Halifax’]

Table 6. Toponym recognition results for free text of Falmouth (2).
Black = correct, red = incorrect, orange = partly correct

When the context is described more extensively, as is the case with Fishing Bay (3), it becomes clear that both tools fall short (Table 7). Automatically recognizing and naming mentioned rivers, creeks and islands when they are not expressly described as such, seems too ambitious. To underline the importance of correct identification and geocoding of context toponyms for determining the entry location, we have mapped these as shown in Fig 4.

Fishing Bay, in ***Maryland****, lies on the East side of* ***Chesapeak bay****,*
partly in ***Dorchester*** *and* ***Somerset counties****. t receives several*
rivers *from each county, the chief of which are* ***Wicomico Nanticoke****;* (3)
also ***Transquaking*** *and* ***Blackwater creeks****. The entrance into this*
large bay lies between ***Goldshorough*** *and* ***Devil’s islands****.*

Processor	Identified tags
Stanza	['Maryland', 'Chesapeak', 'Dorchester', 'Somerset', 'Transquaking', 'Goldshorough', 'Devil']
spaCy	['Maryland', 'Chesapeak bay', 'Goldshorough']
Manual (GS)	['Maryland', 'Chesapeak bay', 'Dorchester county', 'Somerset county', 'Wicomico river', 'Nanticoke river', 'Transquaking creek', 'Blackwater creek', 'Goldshorough island', 'Devil's island']

Table 7. Toponym recognition results for free text of Fishing bay (3).
Black = correct, orange = partly correct, green = clarifying addition

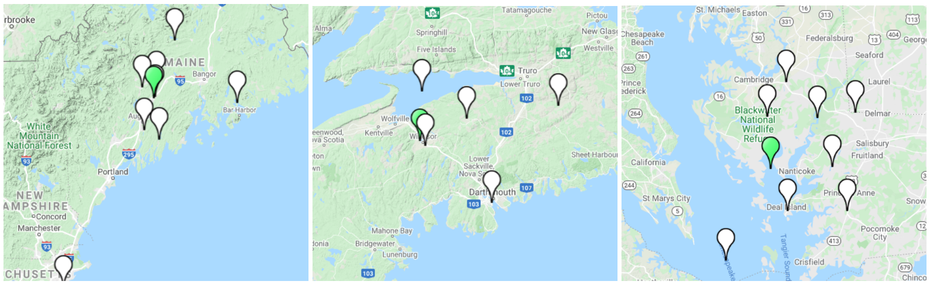


Fig. 4. The mapped context toponyms (white) and entry toponym (green) of Fairfield (1), Falmouth (2) and Fishing bay (3)

8.2 Reconciliation Evaluation

Now that we have examined to what extent the context toponyms can be identified, we investigated in what way they could be applied to reconciliate toponyms. The disambiguation heuristics have been applied to the manually identified context toponyms and to the context toponyms identified by spaCy and Stanza. Since we want to independently assess the functioning of the developed heuristics, we mostly **focused on the manually annotated context toponym selection**. We analyzed the applied disambiguation heuristics both quantitatively and qualitatively to get a complete overview of their functionality in the context of our project. The used heuristics are evaluated with respect to the random baseline score, that is computed as a comparison guideline. Additionally, we have scrutinized the formed connections and performed an error analysis to eventually find the most optimal way to extract, process and geocode geographical data from historical documents.

Evaluation metrics As explained in section 3.2, the main measures we used to evaluate disambiguation functionality are precision (number of correct matches/number of matches) and recall (number of correct matches/number of entries). Additionally, we calculated the coverage for every heuristic, defined as the percentage of matched toponyms (irrespective of correctness) relative to the total number of entries in the test collection. Lastly, we computed the F1-score as the harmonic mean between precision and recall. We compared all these scores to the gold standard and investigated the distribution of mean distance errors. This mean distance between the found coordinates and the factual location (of the gold standard) was calculated using the Haversine formula. By comparing the established Gold Standard IDs to the returned GeoNames query lists we found that the correct candidate was present in 113 of the 197 queries. This determines the upper bound and leads to a maximum achievable recall score of 57.4%.

Heuristic	Correct	Correct	Correct	Mean	No result
	@10km	@100km	@1000km	distance off	found
Baseline (Random)	15.7	19.3	46.2	1523.3 km	21
First Candidate	24.4	30.0	50.7	1359.0 km	21
spaCy + Min. Dist.	37.6	46.2	61.9	603.6 km	46
Stanza + Min. Dist.	43.1	51.3	70.6	648.1 km	28
Manual + Min. Dist.	47.8	59.9	74.1	612.7 km	22
spaCy + Polygon	24.9	30.5	51.8	951.0 km	46
Stanza + Polygon	24.9	28.4	55.8	1122.2 km	28
Manual + Polygon	23.4	27.9	58.9	1109.4 km	22

Table 8. Coverage of the heuristics within 10, 100 and 1000 km of the gold standard coordinates, applied to the manually annotated toponyms. The table shows the percentage of entries for which the error distance is within the specified range, compared to the random baseline, along with the mean distance off and the number of cases for which no coordinates are found. These results are based on the Gold Standard subset (n = 197) and the query settings as outlined in Table 3.

Upon analysis of the coverage scores (Table 8), calculated as the percentage of entries for which the offset falls within a specified distance range, we conclude that both context-based heuristics decrease the mean error distance of the results. At a limit of 1000 km, the coverage score for both has increased compared to the random baseline and the first candidate heuristic. However, this only implies that the number of matches, thus linked entities, has increased. The correctness of these matches is evaluated by calculating the precision and recall score and weighing their harmonic mean value, the F1-score.

As seen in Table 9, the First Candidate heuristic gets a recall score of 0.23. This indicates that, without any manipulation, 23 per cent of the entries in our dataset obtain a correct match simply by returning the first result that comes up in the GeoNames query. As such, a large share of the entry toponyms evidently refer to the most prominent geographical entity which bears their name. It makes sense to think that this is specifically the case with places that only yield one or a few results, but in that case the random baseline should deliver a similar outcome. After analyzing the correct results it is confirmed that the good score is not just the result of the uniqueness or rarity of specific place names, as the average number of records returned for these correctly coupled entries is 18. Apart from any interference of steering context toponyms, this is reasonably a slightly promising output. Likewise, this means that both context-related disambiguating heuristics have to cross a relatively high entry threshold to deliver improved results.

Heuristic	Precision	Recall	F1-score
Baseline (Random)	0.66	0.13	0.21
First Candidate	0.76	0.23	0.35
spaCy + Min. Dist.	0.71	0.33	0.45
Stanza + Min. Dist.	0.70	0.36	0.48
Manual + Min. Dist.	0.67	0.40	0.50
spaCy + Polygon	0.70	0.21	0.33
Stanza + Polygon	0.77	0.22	0.34
Manual + Polygon	0.67	0.19	0.29

Table 9. Precision, recall and F1-score of the different heuristics at 100 km from the coordinates in the gold standard. These results are based on the query settings as outlined in Table 3.

The evaluation results of the context-based heuristics indicate that this threshold is particularly a major challenge for the smallest polygon algorithm. Upon analysis, we concluded that the cause for this lies in the specific calculations that are being performed. The polygon heuristic performs overarching calculations to measure the total affected area. Here, all included toponym candidates, projected as virtual vertices, are treated equally. This means that the mutual distance distributions between the context toponyms themselves are weighed for a considerable share, although these specific proportions are not certainly relevant.

From the obtained results we can derive that the Minimal Distances heuristic performs substantially better than the Smallest Polygon heuristic. With a recall score of 0.40 when using manually annotated toponyms, the performance has increased compared to the first candidate heuristic, which is promising, but still far from a satisfying and sufficient outcome. Whereas the polygon heuristic measures areas and thereby compares distances between all the toponym candidates, the distances heuristic continuously compares the minimal distance between the individual context toponym candidates and the separate entry candidates and thus puts more emphasis on the critical role of the entry toponym. We assume this is the underlying cause that generates improved results.

Since the Minimal Distances heuristic has yielded the most impressive results, from now on we will focus on the analysis of the operation of this heuristic concerning manually annotated toponyms to accomplish further advancements.

8.3 Error analysis: observations and trade-offs

By thoroughly examining how the impact of context toponyms influences the final selection of the heuristic, we present minor adjustments to optimize the results. In this subsection, we discuss the most remarkable observations and their consequences. The effect of adjusting these settings is represented in Table 10.

Outdated place names After analyzing and comparing the names of the selected entries to the Gold Standard, we found out that 35 places of the 197 entries are now known under a completely different name. With this, small name changes that could be bridged by fuzzy search (Levenshtein distance 1) have not been taken into account. This implies that no matter how well the heuristic functions, for these specific place records the right candidate could not be established without consulting external knowledge sources.

Query limiting One of the variables we took into consideration was the number of weighed entry candidates. Here we mean in particular the number of returned GeoNames records found by the entry toponym query that we need to take into account. As said, the first candidate heuristic performed fairly well, which implies that the correct match is commonly located in the upper part of the results. On the other hand, many entities in the dataset are labelled by very common names, among which are also small, unknown towns. These are logically placed lower in the returned record list, so the search field must be broadened to include them. Taking both views into consideration leads to a trade-off between increased opportunities for relevant results and including as many records as possible.

Inclusion of feature types Out of the 197 correctly annotated entries, 34 are of GeoNames feature type A (mostly counties and townships), 36 of type H (mostly rivers), 91 of type P (towns) and 23 of type T (capes and islands). Lastly, there are 11 entries of type S (forts), 1 of type L (a park) and 1 of type U (a shoal). For the sake of completeness, all these types were initially included in the query. While it is apparent to include types A, H, P and T, we assumed that the other types produced more noise than they improve results. After excluding these feature types from the query, results for all heuristics considerably improved (Table 10). The lack of the ability to set the correct link for a few single entries is hereby accepted. Especially candidates of type S appear to have negatively influenced the results. For example, it often happens that the ‘county jail’ entity or the ‘county courthouse’ is preferred above the (correct) county entity itself by a slightly more accurate location. Another example is the ‘State park’, for which the heuristic gives precedence over the state itself.

Effect of Fuzzy search To increase reconciliation chances, we implemented a fuzzy search with factor 0.8 (Levenshtein distance 1). This has yielded partly the desired result because it facilitated the linking of some entries that contained small spelling changes. However, at the same time, entries could be linked to the wrong entity, because their context toponyms are incorrectly coupled to fuzzy-induced resembled context entities that steer them to the wrong location.

Similarity In contrast to enabling a more flexible search by using fuzzy search, some entries do require a stricter approach. This is the case for entries that are coupled to an entity that goes by a slightly different but analogous name. For example, the 'Fairlee' entry is location-wise linked to the West Fairlee town with a small lead, while second-to-best and correct option Fairlee town would be the best strict match language-wise. Another example is 'Falls' a township entry that gets coupled by the Minimal distances heuristic to the village of 'Enosburg Falls', while its correct match is the record of 'Township of Falls'. To solve the problem of the first example, GeoNames does have the query option to only include options with an exact match of strings. However, in the second case, a more strict equal search will not trigger a proper match either. Therefore, a way should be found to include more elaborate name descriptions but no irrelevant places that only contain part of the place name.

Heuristic	Default	n=10	n=100	-S,L,U	F=0.6	F=1.0	Equal
Baseline	0.13	0.16	0.17	0.17	0.09	0.09	0.17
First Cand.	0.23	0.28	0.28	0.29	0.27	0.27	0.28
Min. Distances	0.40	0.43	0.49	0.51	0.42	0.18	0.49
Sm. Polygon	0.19	0.17	0.22	0.27	0.13	0.14	0.23

Table 10. Recall scores of the different heuristics for all further tested additional settings: a query limit of 10 & 100 entry candidates, excluding incidental feature types (S,L,U), a stricter (0.6) and a disabled (1.0) fuzzy search and implementation of equal search. These results are acquired by exploiting the manually annotated context toponyms.

As we analyse the results of the adjusted query settings, we conclude that some modifications improved results compared to the default query. Restricting the query limit, excluding incidental feature types and implementing equal search enhanced recall scores for all heuristics. This information offers new opportunities for further research on finding the most optimized heuristic.

8.4 Combining First Candidate + Minimal Distances

While manually checking, in particular the incorrectly matched output of the Minimal distances heuristic and the comparative intermediate steps taken to get there, we observed a recurring pattern. During run-time, with a majority of cases, the correct entry candidate is temporarily selected and therefore considered the most prominent candidate for a large part of the process. This correct candidate often largely reduces the total distance compared to its previously weighted competitors, but, unfortunately, gets eliminated later on as a result of a rather small decrease in total distance by means of a less prominent entry candidate further down the query list. This situation requires a way to only consider large, significant distance decreases while ignoring minor ones, hereby combining the distance-weighing heuristic with the increased probability of more relevant, high positioned results.

To account for this, we built in a discriminating distance variable as a threshold that only approves substantial distance differences in the comparison method. The minimal distances heuristic further maintains its functionality. As outlined in Table 11 we strive to find the optimal distance difference by comparing the influence of several thresholds on the outcome. Up to a discriminating distance of 100 km, all measures increase. Here, the recall score approaches the upper bound of 57.4%. This is a promising observation, as this means that the heuristic works well in itself, but future research should focus on how the range of results, and as such the availability of the correct results among them, can be extended.

Measure	MD (DD=0)	DD=1	DD=10	DD=100	DD=1000
Coverage	0.60	0.73	0.73	0.75	0.55
Precision	0.67	0.67	0.73	0.75	0.80
Recall	0.40	0.49	0.53	0.57	0.46
F1-score	0.50	0.56	0.61	0.65	0.59

Table 11. Evaluation scores of the newly composed heuristic with a varying discriminating distance at 100 km from the coordinates in the gold standard, compared to the initial Minimal Distances heuristic. These results are acquired by exploiting the manually annotated context toponyms.

9 Conclusion and Future Work

The aim of this study was to explore the process of toponym disambiguation regarding historical works. We intended to find possible improvements by interpreting the free text of place record lemmas. The results of the evaluation of the disambiguation methods show that the context toponyms can certainly contribute to accurate entity linking, although there are still many drawbacks in the process. With regard to the sub-questions, the following can be stated:

- *Q1: How can relevant spatial information be identified from the free text of a place record lemma?*
  As illustrated in Table 2, not all toponym-mentioning descriptions are necessarily relevant to the process of toponym disambiguation. Although a large portion is called because of adjacency, other places are mentioned because of similarity or historical associations. For humans, a fair distinction can be made between them based on the content of the text. For software, it already proves difficult to recognize toponyms, let alone qualify them within a specific substantive target group. However, upon checking the lemmas, we concluded that the less relevant toponyms are often mentioned later in the description. By only processing the commencing context toponyms, we account for this relevance issue without having to assess them on their content. By using Named Entity Recognition software about half of the toponyms could be extracted from the historical text and correctly classified by their type. Certainly in the case of vaguely defined or not explicitly stated entities, it turns out to be difficult to automatically tag context toponyms as such. By incorporating machine learning and advanced tuning of the tools to historical and geographical texts, this output can probably be increased.

- *Q2: To what extent can processed relevant spatial information facilitate successful toponym reconciliation?*
  After the list of relevant context toponyms for every lemma was established and their possible candidates were selected, we investigated several ways to accurately process this information. Our two developed heuristics were based on calculating the smallest total distance between the separate context toponyms and the entry toponym and calculating the minimum spanning area covering the entry toponym and selected context toponyms. The first method turned out to be substantially more effective in its task to correctly reconciliate entry toponyms to their corresponding GeoNames ID. By further implementing the proved emphasis on relevance in the functionality of this heuristic, we managed to increase recall scores reaching up to the upper bound that was defined by the maximum capacity of the API query. This suggests that contextual spatial information definitely allows for enrichment of the toponym reconciliation process and enhances acknowledged map-based methods.

The obtained results from our study revealed that the functionality of the newly developed and refined discriminating distances heuristic is evident, though to really excel in its applicability and overall performance the conditions should be formed in an optimal way. Both the automated toponym recognition process and the formation of a complete selection of candidates have proven to be inadequate. These limitations at the beginning of the disambiguation pipeline induce unfavourable conditions for the concluding heuristics. Moreover, we could state that this case study helped to gain more insight into the crucial considerations of both automated recognition and calculation-based reconciliation with respect to historical works.

To further progress towards the goal of efficient historical entity linking, some steps have to be taken in all aspects of the disambiguation pipeline. Future research could focus on creating knowledge bases that include all former names of places. By combining this external database to our proposed heuristic by knowledge-based methods the historical relations could be examined. Furthermore, the remaining fragments of the context descriptions that we consciously disregarded, definitely contain more distinctive information. Obviously, the coordinates, directions and distances that were provided for part of the entries could demarcate the search area. As indicated in Table 2, the other clues are mostly non-geographical, but by implementing data-driven methods these contents can also be of decisive value. Lastly, it might be interesting to account for the fact that an entry is either a larger, common place or a smaller, unknown site and adjust the number of query results to this. After all, setting more accurate query limits increases the chances of finding and selecting the right entity.

In conclusion, we have taken another step towards efficient historical toponym disambiguation, but there are still many opportunities for improvement.

References

1. Ahlers, D.: Assessment of the accuracy of geonames gazetteer data. In: Proceedings of the 7th workshop on geographic information retrieval. pp. 74–81 (2013)
2. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 273–280 (2004)
3. Ardanuy, M.C.: Entity-centric Text Mining for Historical Documents. Ph.D. thesis, Georg-August-Universität Göttingen (2017)
4. Ardanuy, M.C., Sporleder, C.: Toponym disambiguation in historical documents using semantic and geographic features. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage. pp. 175–180 (2017)
5. Bensalem, I., Kholladi, M.K.: Toponym disambiguation by arborescent relationships. Journal of Computer Science **6**(6), 653 (2010)
6. Bol, P.K.: On an infrastructure for historical spatial analysis (2012)
7. Bol, P.K.: On the cyberinfrastructure for gis-enabled historiography: Space–time integration in geography and giscience. Annals of the Association of American Geographers **103**(5), 1087–1092 (2013)
8. Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Science **22**(3), 301–313 (2008), https://vu.on.worldcat.org/oclc/281985459
9. Buscaldi, D.: Approaches to disambiguating toponyms. SIGSPATIAL Special **3**(2), 16–19 (2011), https://vu.on.worldcat.org/oclc/4806399669
10. Buscaldi, D., Magnini, B.: Grounding toponyms in an italian local news corpus. In: Proceedings of the 6th workshop on geographic information retrieval. pp. 1–5 (2010)
11. Chopde, N.R., Nichat, M.K.: Landmark based shortest path detection by using a* and haversine formula. International Journal of Innovative Research in Computer and Communication Engineering **1**(2), 298–302 (2013)
12. Goodchild, M.F., Hill, L.L.: Introduction to digital gazetteer research. International Journal of Geographical Information Science **22**(10), 1039–1044 (2008)
13. Leidner, J.L., Lieberman, M.D.: Detecting geographical references in the form of place names and associated spatial natural language. SIGSPATIAL Special **3**(2), 5–11 (2011), https://vu.on.worldcat.org/oclc/4806399667
14. Manning, P., Mostern, R.: World-historical gazetteer (2015)
15. Mitchell, S.: Where in the world? an online guide to gazetteers, atlases and other map resources. Internet reference services quarterly **8**(1-2), 183–194 (2003)
16. Morse, J.: The American Gazetteer. Thomas & Andrews (1797)
17. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
18. Rauch, E., Bukatin, M., Baker, K.: A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references. pp. 50–54 (2003)
19. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries. p. 127–136. ECDL '01, Springer-Verlag, Berlin, Heidelberg (2001)