So Far Away and Yet so Close: Augmenting Toponym Disambiguation and Similarity with Text-Based Networks

Andreas Spitz Johanna Geiß Michael Gertz

Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
{spitz, geiss, gertz}@informatik.uni-heidelberg.de

ABSTRACT

Place similarity has a central role in geographic information retrieval and geographic information systems, where spatial proximity is frequently just a poor substitute for semantic relatedness. For applications such as toponym disambiguation, alternative measures are thus required to answer the non-trivial question of place similarity in a given context. In this paper, we discuss a novel approach to the construction of a network of locations from unstructured text data. By deriving similarity scores based on the textual distance of toponyms, we obtain a kind of relatedness that encodes the importance of the co-occurrences of place mentions. Based on the text of the English Wikipedia, we construct and provide such a network of place similarities, including entity linking to Wikidata as an augmentation of the contained information. In an analysis of centrality, we explore the networks capability of capturing the similarity between places. An evaluation of the network for the task of toponym disambiguation on the AIDA CoNLL-YAGO dataset reveals a performance that is in line with state-of-the-art methods.

CCS Concepts

•**Computing methodologies** → **Natural language processing;** •**Information systems** → *Information retrieval;*

Keywords

Location network, location similarity, toponym disambiguation, toponym extraction, centrality, Wikipedia, Wikidata

1. INTRODUCTION

The processes of learning, understanding, and deriving knowledge are often described by the idiom of "connecting the dots" in the English language, an expression that reflects the way in which we as humans store and process information in our brains. As a result, artificial representations of knowledge frequently take a similar form, such as the web of information in knowledge bases. This, of course, poses the question of how such connections can be made or – more formally – how the dots (or nodes) that make up information can be connected to form a network. Often, this information is not explicitly given but can be derived or extracted from unstructured sources. In the case of places, spatial proximity is a natural candidate for establishing connections that are frequently reflected in reality. For example, a mention of the Eiffel Tower brings to mind the city of Paris where the tower is located. Here, spatial proximity directly corresponds to semantic relatedness between the two place mentions. In other cases, however, spatial proximity is misleading. Consider, for example, the Canary Islands, which are located on the African continental plate in the Atlantic Ocean. Based on proximity, they are related to the countries of Morocco and Mauritania, to which they are closest. A more thorough investigation reveals, however, that they are an autonomous community of Spain and thus belong to the European Union politically and culturally. Depending on the intended application, such a cultural or political relation can be preferable. Thus, a network that is to be used in support of information processing and retrieval should capture these semantic relations. Here, we describe an approach to the creation of such a network from place mentions (so called *toponyms*) in unstructured text and provide a network of places derived from the text of the English Wikipedia with links to the knowledge base Wikidata.[1]

The connectedness of places in such a network enables us to explore the notion of centrality, i.e., the importance of places with regard to each other, which implies asymmetric relations between places that can be captured in directed networks. By using such a network, we then consider the value of the context of places for the disambiguation and resolution of toponyms, i.e., the linking of a toponym to a unique entity with geo-coordinates. While the context of toponyms has been utilized in the area of toponym resolution before, much of this work is focused on the distribution of words in the context of toponyms [3, 7] or, alternatively, on linguistic features of the surrounding text [8] (for a recent overview of methods, see [12]). As a result, the context is considered primarily in a language-dependent manner. In contrast to this, we show that an approach that is based only on the relations between places themselves is equally viable. By using inherently language-independent networks between places instead of the language-specific context of

This is the author's version of the work. It is posted here for your personal use, not for redistribution. The definitive version is published in:

GeoRich'16, June 26-July 01 2016, San Francisco, CA, USA

© 2016 ACM. ISBN 978-1-4503-4309-1/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2948649.2948651

[1]The network is available for download at our website: `http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

previous approaches, we take a step towards tackling the open question of language independent methods for entity disambiguation [16]. Furthermore, by linking the network to a language-agnostic knowledge base such as Wikidata, we provide a novel approach to the problems that arise due to alternative names and spellings of locations.

The remainder of this paper is structured as follows. In Section 2, we discuss related approaches. We present the models for generating location networks from textual sources in Section 3 and apply them to the English Wikipedia in Section 4. In Section 5, we evaluate the network's performance for the task of toponym resolution and give a summary and outlook in Section 6.

2. RELATED WORK

Numerous approaches exist for toponym resolution, i.e., the task of disambiguating toponyms and linking them to locations. Therefore, the following list is by no means comprehensive, but serves as an overview of methods that are either central to the task or most related to our approach.

Most existing work builds directly on textual features. DeLozier et al. use a world-wide geographic distribution of words and include additional statistics from Wikipedia for toponym resolution [3]. Based on similarity features, Santos et al. employ machine learning for the task of toponym disambiguation [12]. Liu et al. present an approach that measures the relatedness of toponyms by using provinces of China as a fixed set of locations, for which they analyze textual co-occurrence patterns [9]. Speriosu and Baldridge extract document-level geotags for text-driven toponym resolution [14]. For the analysis of streaming news data on small localities but for large domains of locations, Lieberman and Samet propose an approach for the resolution of toponyms based on context features [8]. Beyond toponyms, there are numerous approaches to the disambiguation of named entities in general (for an introduction, see [7]).

In addition to these approaches, a number of previous studies also focus on the use of networks for disambiguation. Volz et al. introduce a topology-based approach to the disambiguation of toponyms [17]. Quercini and Samet study the spatial relatedness in Wikipedia, explore graph-based similarity measures to determine related concepts for each location, and build so-called local lexicons for occurring locations as a resource for toponym disambiguation [11]. Alencar et al. use Wikipedia links to generate a semantic network for the geographical classification of documents [10].

Many of the above approaches use Wikipedia, Wikipedia links or knowledge bases derived from the former to extract similarities of place mentions. However, they do not include text-based distances between the mentions of places within each document as a measure of the inherent similarity between places. We extract a comprehensive network of text-based place similarities first and then use disambiguation as just one of the possible applications of such a network. While naturally occurring spatial networks have been a focus of network analysis for quite some time (for an overview, see [1]), to our knowledge, they do not yet include spatial networks that are constructed from textual proximity. In this paper, we therefore extend previous work on the extraction of implicit networks of temporal expressions [15], persons [5], and locations [6] from textual distance in document collections and show how the obtained network of locations can be used in geographic information retrieval.

3. MODEL

Based on the premise that entities share a relation if they are mentioned in the same context, we construct a network of locations as a graph whose nodes represent the locations while edges indicate relations between them. In contrast to approaches that build knowledge bases from structured data, we create this network from unstructured text. We consider two toponyms to share a context if they occur in the same document. This approach entails two challenges, namely the number of induced edges and the textual distances between toponyms that entail a diminishing strength of connection with increasing distance in the text. To resolve these issues, we employ an edge aggregation technique and a weighting scheme for edges as described in the following.

3.1 Undirected Co-occurrence Network

Following the above intuition, we construct a network from the co-occurrence of place mentions in a collection of documents, which we assume to be tagged for place mentions, i.e., the toponyms are identified. First, we construct a bipartite graph $B = (V \cup O, E_B)$ to model the co-occurrences, where the set of nodes V corresponds to the set of locations and O denotes the set of documents. For two nodes $v \in V$ and $o \in O$, the bipartite edge set E_B then contains an edge (v, o) iff document o contains a toponym that corresponds to location v. To obtain a network of locations, we project B onto V, i.e., we include an edge between two locations if they occur in the same document. This induces parallel edges, since a given pair of locations may co-occur in more than one document. A document with k distinct place mentions induces $\binom{k}{2}$ edges between locations, which makes the projection rather dense. In the following, we describe a scheme for aggregating and weighting the resulting edges.

Let $M = (V, E_M)$ denote the multi-graph over the set of all locations V that we obtain by projecting B. Two locations $v, w \in V$ are connected by an edge $e = (v, w, i) \in E_M$ if there exists an instance i where toponyms of v and w co-occur in a document. Note that co-occurrence instances do not necessarily correspond to documents, since a document may contain multiple instances of a given toponym. As a result, E_M is a multiset that contains a distinct edge for each co-occurrence of v and w. Given an instance of a co-occurrence i between locations v and w, we define the distance $d(v, w, i)$ between them as the number of sentences that separate their occurrences in the document that corresponds to instance i. If they occur in the same sentence, we set $d(v, w, i) = 0$. This provides us with a kind of dissimilarity between the two toponyms that increases with their distance in the text. To create weights that indicate similarity for edges in M based on this dissimilarity, we introduce edge weights that decay exponentially, i.e. $\varphi : E_M \to \mathbb{R}$, as

$$\varphi(e = (v, w, i)) := \exp\left(-\frac{d(v, w, i)}{2}\right) \tag{1}$$

To aggregate multiple parallel edges into a single edge, we then use a cosine similarity of adjacency vectors of nodes in the weighted node-edge incidence matrix of M. Computations on the full node-edge incidence matrix would be quite complex due to the enormous number of edges. We note, however, that we can use the sparseness of the matrix to reduce the complexity to the local neighbourhood of two nodes, since it corresponds to a weighted cosine similarity of neighbourhoods for the two incident nodes. There-

fore, let $N_v := \{x \in V | (v, x, \cdot) \in E_M\}$ denote the set of nodes that are adjacent to a given node v. Furthermore, let $N_{vw} := N_v \cap N_w$ denote the shared neighbourhood, i.e., the set of nodes that are adjacent to both v and w. $E_v := \{(v, x, \cdot) \in E_M | x \in N_v\}$ then denotes the set of edges that are incident to nodes in a given neighbourhood N_v. We denote edges in the shared neighbourhood of two nodes as $E_{vw} := \{(v, x, \cdot) \in E_v | x \in N_{vw}\} \cup \{(w, x, \cdot) \in E_w | x \in N_{vw}\}$. Based on these neighbourhoods, we obtain a *distance cosine* similarity of node-edge incidence vectors from the exponentially decaying similarity measure as

$$dicos(v, w) := \frac{\sum_{e \in E_{vw}} \varphi(e)^2}{\sqrt{\sum_{e \in E_v} \varphi(e)^2} \sqrt{\sum_{e \in E_w} \varphi(e)^2}} \quad (2)$$

If the shared neighbourhood of two nodes is empty, we set $dicos(v, w) := 0$. To obtain the desired simple graph $G = (V, E)$, we regard G as the complete graph over nodes V and define a weight function $\omega : E \to \mathbb{R}$ as $\omega(e) := dicos(v, w)$. Edges with weight zero are not considered to obtain a sparse representation. The resulting graph is still dense, but an edge threshold can be applied if a sparser graph is required.

3.2 Directed Co-occurrence Network

The graph that results from the above model is inherently undirected since there is no direction in the co-occurrence of place mentions. While it is possible to consider the order of sentences that contain the toponyms, this is not intuitive due to the large possible distances between sentences that span across paragraphs. However, a meaningful notion of direction arises when we consider the number of co-occurrences of two locations in relation to their individual number of co-occurrences overall. For example, Paris is more important for the Eiffel Tower than the tower is for Paris since the city has many other points of interest, while the tower is located in only one city. Thus, more influential locations tend to be more important for their less influential neighbours than the other way around. This is reflected in the number of toponym co-occurrences, where a much larger fraction of the overall textual co-occurrences of a point of interest will be with the enclosing city, while the co-occurrences of the city with the point of interest constitute only a smaller fraction of its occurrences in the text. Based on this intuition, we derive a directed network by normalizing the weights of outgoing edges of a node with the sum of all incident edges, i.e., we construct a set of directed edges A with weights $\omega : A \to \mathbb{R}$ such that

$$\omega(v \to w) := \frac{\omega(v, w)}{\sum_{x \in V} \omega(v, x)} \quad (3)$$

The resulting edge weights are distinct for reciprocal edges and encode the importance of one place for another. The procedure for creating both the directed and undirected networks is summarized in Algorithm 1, which has an average worst-case complexity of $\mathcal{O}(\langle d \rangle^2 \cdot |O|)$, where $\langle d \rangle$ is the average number of toponyms per document.

3.3 Linking Locations

A network G or D as described above provides a representation of implicit textual relations between the locations V. However, if external knowledge is available for locations, further augmentation of the network is possible. On the one hand, nodes in V can be linked to entities in a knowledge base to make additional node attributes available, such as the type of place (e.g., city, country, point of interest, etc.), or the population of a city or country. On the other hand, additional relations from such a knowledge base can induce a new set of edges. For example, it is possible to include geographical hierarchies in the form of a new set of labelled edges A_H, such that the resulting graph $H = (V, A_H)$ describes relationships of nodes within the hierarchy. In the case of Wikipedia as a data source, the steps of toponym extraction and toponym resolution can be combined into a single process by using Wikidata, thus circumventing the problems that typically arise during entity linking.

Algorithm 1 Creation of the location networks for a collection of documents that have been tagged for locations.

Input: Documents O, locations V

Initialize $V \leftarrow \emptyset$, $E_M \leftarrow \emptyset$, $E \leftarrow \emptyset$, $A \leftarrow \emptyset$ and $i = 0$
for $o \in O$ **do**
  $V_o \leftarrow \{v \in V | v \in o\}$
  while $V_o \neq \emptyset$ **do**
    Take $v \in V_o$ and set $V_o \leftarrow V_o \setminus \{v\}$
    for $w \in V_o$ **do**
      $E_M \leftarrow E_M \cup \{(v, w, i)\}$
      $\varphi(v, w, i) \leftarrow \exp(-\frac{1}{2} d(v, w, i))$
      $i \leftarrow i + 1$
for $v \in V$ **do**
  for $w \in N_v$ where $w > v$ **do**
    $E \leftarrow E \cup \{(v, w)\}$
    $\omega(v, w) \leftarrow dicos(v, w)$
for $v \in V$ **do**
  $s \leftarrow \sum_{x \in V} \omega(v, x)$
  for $w \in N_v$ **do**
    $A \leftarrow A \cup \{(v \to w)\}$
    $\omega(v \to w) \leftarrow \frac{1}{s} \omega(v, w)$

Output: $G = (V, E, \omega)$ and $D = (V, A, \omega)$

4. NETWORK EXPLORATION

Based on the model presented above, we now describe the construction of such a location network from the text of the English Wikipedia dump of June 2, 2015. We augment it with information from the Wikidata dump of June 1, 2015.

4.1 Network Construction and Properties

To construct the location network, we use the English Wikipedia as a comprehensive source of text. In line with our model, we include only unstructured text, i.e., we do not consider info boxes, references or pages of lists. For the extraction of toponyms, we make use of Wikipedia links (WL), which are links between Wikipedia pages that are already tagged in Wikipedia. Here, the surface text of a Wikipedia link is considered to be a toponym if the link's target corresponds to a Wikipedia page that is connected to a Wikidata entity that either has geocoordinates or is classified as a location. For all such links, we assign the Wikidata ID to the location. Thus, toponym resolution is a trivial matter for the construction of the network itself. While this approach has been used before, note that we do not generate a network between Wikipedia pages based on their connection through links but between locations based on the co-occurrences of toponyms in the text of any Wikipedia page, and only use the WLs to identify toponyms. A possible

city	c_{deg}	c_{indeg}	c^H_{deg}	c^H_{indeg}
Paris	63,150	89.87	8,064	7.56
New York City	79,398	71.74	9,294	12.12
Chicago	54,217	51.84	8,074	7.70
Los Angeles	49,961	51.47	7,276	7.76
Washington, D.C.	62,858	51.05	8,138	8.65
Boston	45,895	50.43	6,121	6.08
Philadelphia	51,237	45.19	6,372	5.03
Vienna	35,724	44.55	4,827	7.44
Moscow	29,026	43.77	4,644	19.47
San Francisco	43,759	40.87	6,029	4.76
Rome	43,500	40.31	5,825	6.10
Baltimore	31,490	38.88	4,582	2.31
Toronto	34,716	37.95	5,273	4.19
Berlin	40,451	37.65	5,750	10.57
Madrid	24,753	35.94	3,381	4.22
Stockholm	22,386	35.64	3,512	5.79
Buenos Aires	20,920	35.40	3,214	13.12
Portland	22,515	34.66	3,858	2.51
Worcester	8,458	34.36	1,631	1.94
Lyon	15,852	33.42	2,141	2.90

Table 1: The twenty highest ranked cities according to c_{indeg}, along with all four centrality scores.

problem of this approach arises from the Wikipedia guidelines, which state that entities should only be linked to their Wikipedia sites the first time they appear on a page. This could negatively influence the recall of our approach since we may overlook subsequent location mentions. However, as we found in a previous experiment for building a social network from Wikipedia in which we employed a string search to locate subsequent mentions of identified person mentions, the effect on the resulting network is negligible [5].

An added benefit of using WLs is the direct integration of Wikidata as a knowledge base. Wikidata links its entries to Wikipedia pages, which means that matching a toponym to a Wikidata entry is a simple matter of matching the link targets. As a result, we effectively resolve the toponyms and obtain additional information such as the population for countries or cities and the type of location (e.g., city, country, continent, point of interest or geographic feature). Furthermore, we can extract a hierarchy of places from Wikidata to augment the network with an alternative set of edges that encodes the relations *city* < *country* < *continent*.

By using this data from a total of 4.8M content pages on Wikipedia that contain 20.6M Wikipedia links to locations, we are able to identify 723,779 locations and 178.9M aggregated edges in the undirected graph. The clustering coefficient $cc = 0.56$ of the (undirected) graph is high in comparison to its global density of $\delta = 6.8 \cdot 10^{-4}$, which indicates a good local density [6]. If we limit the set of locations to those for which we have hierarchy information, the network contains 96,444 locations, the majority of which are cities, and 58.5M edges with non-zero weights.

4.2 Network Centrality

When presented with a network, one of the first questions that comes to mind concerns the position of nodes in relation to each other, i.e., the question of centrality. Many measures of centrality have been suggested in the literature and not all of them can be applied in every setting. Here, we use the degree centrality as a basic measure, i.e., the number of adjacent edges of a node, which corresponds to the number of distinct co-occurrences of a toponym. We denote with $c_{deg}(v)$ the degree centrality of node v in the entire network and with c^H_{deg} the centrality that is obtained by only counting edges between nodes for which we have hierarchy information. As a more involved measure, we also include the in-degree centrality c_{indeg}. Formally, it measures the strength of incoming edges in the directed network as

$$c_{indeg}(v) := \sum_{w \in V} \omega(w \to v) \tag{4}$$

which corresponds to the first iteration of a PageRank algorithm. The hierarchical in-degree centrality c^H_{indeg} is defined analogously on edges between locations for which we have hierarchy information. In Table 1, we show the 20 top-ranked cities by in-degree centrality. The degree centrality and in-degree centrality capture different notions of importance as the example of Worchester shows, which is much less connected than the other top-ranked cities. However, the city is very central based on its historical importance for connected places. To provide a more intuitive visualization of the local cohesion between places in the network, we show the strongest connections ($\omega > 0.07$) between the top-ranked cities by in-degree centrality in Figure 1. Despite some inaccuracies (e.g., London is missing due to inconsistencies in Wikidata labels), we find that the network reflects local structures well, but is not limited to geographical proximity as it also includes a number of intercontinental connections.

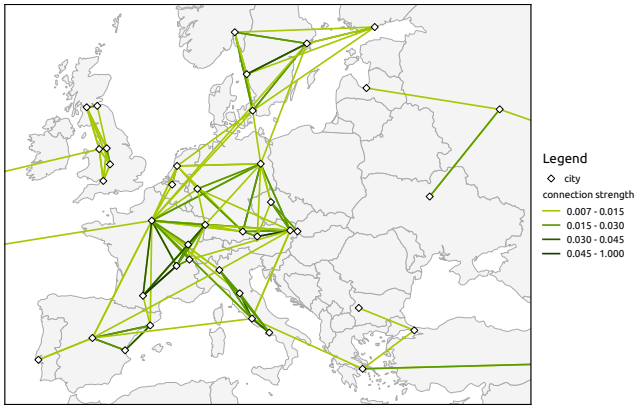

Figure 1: European map section of the geographically embedded subnetwork with $\omega > 0.007$ between the worldwide 100 highest ranked cities by c_{indeg}.

Finally, we consider the usefulness of the centrality scores for classification tasks within the geographical hierarchy. Intuitively, countries should be more central than cities, according to the definition we used for the centrality scores, since they are more likely to be better connected. There are, of course, exceptions where large metropolises are justifiably more important than very small countries. We classify locations in the network into countries and cities by their centrality and use the hierarchy data from Wikidata for evaluation. Since the data is imbalanced with many more cities than countries, we provide a curve of precision vs. recall for the different centrality scores in Figure 2. The in-degree centralities perform better for lower recall values, while the basic degree centrality performs best for higher recall. Limiting the selection to nodes with hierarchical data improves the result for the in-degree centrality, but has a strong negative effect for the degree centrality. While the results are expectedly not perfect, the correlation of centrality to a notion of importance of the locations is visible.

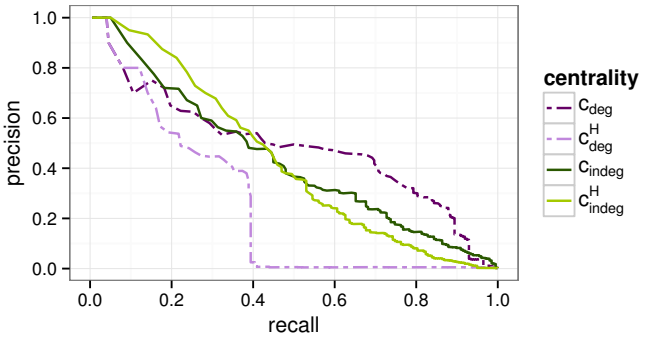

Figure 2: Precision vs. recall curves of the classification of countries by using the different centrality indices as a ranking.

5. TOPONYM DISAMBIGUATION

Disambiguation as a task stands to profit immensely from the availability of a network that represents the context of entities, as the context alone is often sufficient to disambiguate a term. Consider the toponym `Heidelberg`, for example. Worldwide, at least 11 settlements are referred to by that name, two of which even have a university. When given a text about Heidelberg University, it is therefore not directly clear to which place it refers. If the text contains mentions of additional places, however, this situation changes. For example, if the river Neckar is mentioned, the text likely refers to Heidelberg in Germany. In the following, we therefore present an approach that uses the location network extracted from Wikipedia and Wikidata as a tool for toponym disambiguation. To this end, we use uniquely identified points of reference in the text, the so-called *seeds*, in an extension of a similar approach to the disambiguation of person mentions based on a social network [4].

5.1 Toponym Extraction and Entity Look-up

The preparatory step for toponym disambiguation requires the tagging of toponyms in the target text, a task for which any named entity recognizer or gazetteer can be applied. Our aim is then to map these mentions to a reference list of unique locations. The first step is therefore a matching of the extracted toponyms to candidate locations in the location network. Since the set of nodes directly corresponds to Wikidata entries, we consider them to be sufficiently comprehensive as a repository of locations, i.e., as a gazetteer. To match the toponyms in the text to nodes in the network, we use a set of string matching algorithms (for a more thorough description, see [2]). As a result of this step, it is of course possible to obtain more than a single candidate, which leads to three possible outcomes for each toponym: *(i)* no match is found, *(ii)* exactly one match is found (unambiguous mention), or *(iii)* more than one match is found (ambiguous mention). In the first case, there is no match for the toponym in the location network and it can therefore not be linked to any location. In the second case, the location mention is unambiguous, and the task of linking it to a node in the network is straightforward. In the following, we also refer to these uniquely identified toponyms as *seeds*. In the third case, we find more than one matching location and obtain a list of possible candidate locations $L_t \subseteq V$ for toponym t. We select the best candidate among them by using the location network as described in the following.

5.2 Network Disambiguation

To determine the best candidate in L_t, we use the neighbourhoods in the location network to compute a ranking of candidates by their relation to other location mentions in the documents. Then, we select the highest ranked candidate and link it to the toponym. Key to this approach are the unambiguous location mentions that are identified in the string matching phase (see *(ii)* above), which we refer to as the set of seeds S. For these, we know that they are accurate links between toponyms in the document and locations in the network. Thus, S does not depend on the specific toponym t that we are trying to disambiguate, but rather on the document that contains t. We therefore use these seeds as points of reference for linking t to the appropriate location in network. To do this, we compute the strength of the tie between each candidate $l \in L_t$ to the seed locations as the sum of edge weights between them. Intuitively, we try to find the location for which the neighbourhood is the best fit. Formally, we compute the distance to all seeds as

$$\varrho(l, S) := \sum_{s \in S} \omega(l, s) \tag{5}$$

Based on ϱ, we rank all candidates $l \in L_t$ and select the candidate for which $\varrho(l, S)$ is maximal, meaning that we find the candidate for which the fit with the surrounding points of reference in the network is highest.

5.3 Evaluation

To provide a benchmark of our approach that allows a comparison to other methods, we evaluate it on a standard data set, namely the AIDA CoNLL-YAGO data [7], of which we use the *test-b* set. Since our method does not require training, the training set is used to adjust the entity look-up methods to maximize the number of mentions for which candidates are found in the location list. Of the 4,485 mentions of different types of entities that are annotated with a Yago2 entity and a Wikipedia link in the original data set, we select the subset of 1,493 location mentions (excluding demonyms). We use the Wikipedia link information to map the mentions to Wikidata IDs for the ground truth. As baselines, we include two heuristics. For B_{DIST}, the candidate with the shortest distance to the seed locations is selected. The candidate with the lowest Wikidata IDs is used for the baseline B_{MIN}, based on the assumption that more important locations were added to Wikidata sooner and thus have a lower ID. To provide a direct system comparison, we installed AIDA [7] locally and evaluated it on the described subset. We use two metrics for evaluation, namely precision, which is the ratio between correctly linked toponyms and all linked toponyms, and the mean distance in kilometres between the selected and the correct location. To calculate great circle distances, we use the haversine formula [13].

For a total of 251 toponyms, the evaluation framework as described above is unable to select a target location for linking, due to the following three reasons: *(i)* no candidate is found in the location list for 23 toponyms, *(ii)* for 166 mentions it is impossible to establish a set of seeds (51 documents contain no seed locations) and *(iii)* in 62 cases there is no edge between any possible candidate and any seed location. These issues can be reduced by using a larger location list, different string matching algorithms or a filtering of candidates. We observe that our approach returns no result, rather than a wrong result, in these cases. For

	P in %			mean dist in km		
	all	seeds	ambig.	all	seeds	ambig.
WLND	**85.7**	86.0	**85.6**	**327.5**	522.9	**179.1**
AIDA	84.9	86.0	83.2	*120.4*	*87.7*	*142.3*
B_{DIST}	81.6	86.0	78.5	683.1	522.9	800.8
B_{MIN}	81.4	86.0	78.8	650.9	522.9	745.0

Table 2: Precision (P) and mean distance in km values for toponym disambiguation based on our Wikipedia Location Network (WLND), AIDA and the two baselines (B_{DIST} and B_{MIN}).

1,065 of the remaining 1,242 mentions, our approach links the toponym to the correct place, which results in a precision of 85.7%. In Table 2, we show the precision of all approaches along with the baselines. We omit recall, as it would evaluate the extraction instead of the disambiguation of toponyms. In lieu with our approach, we distinguish between toponyms for which we find a single entry in the location list (seeds) and ambiguous toponyms. For the seeds, the precision of our Wikipedia Location Network Disambiguation method (WLND) is identical to the baselines, since the baseline methods only vary in the selection of ambiguous candidates. AIDA does not distinguish between the two subsets, but for completeness we include the results in relation to our approach. However, the focus of our evaluation lies on the ambiguous toponyms (labelled *ambig.* in Table 2). Here, our approach outperforms both the baselines and the AIDA system. With a higher precision in identifying the seeds, the quality of our disambiguation approach increases further. For the mean error distance, our method outperforms the two baselines, but fares worse than AIDA. Here, we note that no geo-coordinate information is available for a third of locations that are incorrectly linked by AIDA, which results in a misleadingly low mean error distance.

Overall, we find that disambiguation based on the location network achieves top precision and performs comparably to state-of-the-art systems. In contrast to existing disambiguation tools, our approach is based only on co-occurrences of toponyms in Wikipedia pages. Local or linguistic phenomena in the text of the test data such as coherence among entities are not taken into account. As a result, the location network disambiguation is a novel approach that can be further improved through combination with existing tools, for example by substituting their recommendations as seeds in the network instead of relying on string matching.

6. CONCLUSION

In this paper, we presented a novel approach to modelling the relations between places by extracting locations from a collection of documents and computing a similarity that is based on their distances in the text. Unlike traditional knowledge bases, the presented approach has the advantage of being feasible for a construction directly from unstructured text, while still retaining the capability as a powerful tool for supporting natural language processing tasks such as disambiguation or summarization. We created a location network from the English Wikipedia and linked it to Wikidata as a knowledge base, based on which we showed the intuitive nature of relations in the network.

Currently, we are working on including more entity types to increase the scope of the network. Furthermore, we are extending the network by including the semantics of relations between toponyms as a typing of edges on the basis of semantic embedding and term co-occurrences. Based on these extensions, we are working towards a combination of further entity networks extracted from unstructured text towards a multi-partite network between different classes of entities. Such a network will serve to further improve the performance of natural language processing, which can be supported by similarities that are only implicitly contained in document collections.

7. REFERENCES

[1] M. Barthélemy. Spatial Networks. *Physics Reports*, 499(1):1–101, 2011.

[2] P. Christen. A Comparison of Personal Name Matching: Techniques and Practical Issues. In *ICDMW*, 2006.

[3] G. DeLozier, J. Baldridge, and L. London. Gazetteer-independent Toponym Resolution Using Geographic Word Profiles. In *AAAI*, 2015.

[4] J. Geiß and M. Gertz. With a Little Help from my Neighbors: Person Name Linking Using the Wikipedia Social Network. In *WWW Companion*, 2016.

[5] J. Geiß, A. Spitz, and M. Gertz. Beyond Friendships and Followers: The Wikipedia Social Network. In *ASONAM*, 2015.

[6] J. Geiß, A. Spitz, J. Strötgen, and M. Gertz. The Wikipedia Location Network - Overcoming Borders and Oceans. In *GIR*, 2015.

[7] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *EMNLP*, 2011.

[8] M. D. Lieberman and H. Samet. Adaptive Context Features for Toponym Resolution in Streaming News. In *SIGIR*, 2012.

[9] Y. Liu, F. Wang, C. Kang, Y. Gao, and Y. Lu. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *T. GIS*, 18(1), 2014.

[10] R. Odon de Alencar, C. A. Davis Jr, and M. A. Gonçalves. Geographical Classification of Documents Using Evidence from Wikipedia. In *GIR*, 2010.

[11] G. Quercini and H. Samet. Uncovering the Spatial Relatedness in Wikipedia. In *SIGSPATIAL*, 2014.

[12] J. Santos, I. Anastácio, and B. Martins. Using Machine Learning Methods for Disambiguating Place References in Textual Documents. *GeoJournal*, 80(3):375–392, 2015.

[13] R. W. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68(2):158, 1984.

[14] M. Speriosu and J. Baldridge. Text-Driven Toponym Resolution using Indirect Supervision. In *ACL*, 2013.

[15] A. Spitz, J. Strötgen, T. Bögel, and M. Gertz. Terms in Time and Times in Context: A Graph-based Term-Time Ranking Model. In *TempWeb*, 2015.

[16] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL at HLT-NAACL*, 2003.

[17] R. Volz, J. Kleb, and W. Mueller. Towards Ontology-based Disambiguation of Geographical Identifiers. In *I3*, 2007.