GeoJournal (2015) 80:375–392
DOI 10.1007/s10708-014-9553-y

Using machine learning methods for disambiguating place references in textual documents

João Santos · Ivo Anastácio · Bruno Martins

Published online: 7 May 2014

Abstract This paper presents a machine learning method for disambiguating place references in text. Solving this task can have important applications in the digital humanities and computational social sciences, by supporting the geospatial analysis of large document collections. We combine multiple features that capture the similarity between candidate disambiguations, the place references, and the context where the place references occur, in order to rank and choose from a set of candidate disambiguations, obtained from a knowledge base containing geospatial coordinates and textual descriptions for different places from all around the world. The proposed method was evaluated through English corpora used in previous work in this area, and also with a subset of the English Wikipedia. Experimental results demonstrate that the proposed method is indeed effective, showing that out-of-the-box learning algorithms and relatively simple features can obtain a high accuracy in this task.

Keywords Place reference disambiguation · Geographic text mining and retrieval · Entity linking in text · Learning to rank

J. Santos (✉) · I. Anastácio · B. Martins
Instituto Superior Técnico, Lisbon, Portugal
e-mail: joaosantos.010@gmail.com

Introduction

Given the large amounts of textual data that are currently available and published daily on different types of Web platforms, research on information extraction methods to automatically extract structured information from these sources is getting increasingly important. Moreover, we have that geographic information is pervasive over these textual contents, since most documents contain references to particular locations. An important text mining problem is therefore related to resolving the place names that are referenced in texts, an activity that can be generally divided into two separate sub-tasks, namely (1) place reference identification, and (2) place reference disambiguation. The first sub-task is deeply related to the problem of named entity recognition (NER), which has been thoroughly studied in the natural language processing (NLP) community (Nadeau and Sekine 2007). The second sub-task involves re-expressing the recognized references into a standard format which precisely describes their location on the surface of the Earth (e.g., assigning place references to unique identifiers such as geospatial coordinates). This sub-task is, in turn, deeply related to the problem of named entity disambiguation (Ji and Grishman 2011), which has also been receiving substantial attention. Our work specifically addresses the later sub-problem, i.e. place reference disambiguation over textual documents.

Consider the following sentences, each from a different document, and consider that the word *Georgia* has been recognized as a place reference:

1. ***Georgia***, *on the year of 1788, was the fourth state to ratify the constitution of the United States of America.*
2. *Films set in* ***Georgia*** *include Gone with the Wind and Driving Miss Daisy.*
3. *Joseph Stalin was born in* ***Georgia*** *on 1879.*

By analyzing the context surrounding each reference to the place named *Georgia*, a disambiguation system should assign the same identifier to the first two references, as they both refer to the US state named Georgia, while the place reference in the third document is referring to the the Eastern European country with the same name, thus corresponding to a different identifier. Although this example considered place references sharing the exact same characters, references that are misspelled or that can be referenced by multiple equivalent names (e.g., *New York City*, *NYC* and *Big Apple*) should also be assigned to the same unique identifier (e.g., the same entry in a well-established geographic gazetteer, or the same geospatial coordinates).

Possible applications for place reference resolution include (1) enriching contents in digital libraries with links to geographic gazetteers, (2) producing map-based visualizations that support the exploration of textual collections, and (3) supporting more advanced geographic information retrieval applications. Place reference resolution can, for instance, be particularly useful for humanities scholars (Brown et al. 2012; Smith and Crane 2001) and for social scientists analyzing media coverage (Mehler et al. 2006), enabling them to study geographic patterns emerging from place references occurring over large document repositories (Adams and McKenzie 2013).

Our work addresses the disambiguation of place references, initially identified through a standard NER model, with a method that relies learned models in order to rank and choose, from a set of candidate locations described in a knowledge base built from Wikipedia, the most likely disambiguation for each place reference made in the text. Results from an extensive set of experiments demonstrate that the proposed method is indeed effective, showing that out-of-the-box machine learning algorithms from the current state-of-the-art (e.g., the LambdaMART learning to rank algorithm (Burges 2010), together with relatively simple and computationally inexpensive features, can obtain a high accuracy in this particular disambiguation task.

The following section presents related work, while section ("Disambiguating place references over textual contents") presents the proposed approach. Section "Experimental validation" presents experimental results. Finally, section "Conclusions and future work" summarizes our conclusions, and presents possible directions for future work.

Related work

This section presents previous work related to the research that is reported in this paper, starting with previous studies that addressed the general problem of named entity disambiguation in text. Section "Place reference disambiguation" describes previous work that specifically focused on the disambiguation of place references. Finally, section "Geocoding the entire contents of textual documents" describes previous research on the related problem of geocoding the contents of entire textual documents.

General approaches for named entity disambiguation

Many previous works have modeled the general problem of named entity disambiguation as the task of assigning each entity mention to the corresponding Wikipedia entry. This representation makes it possible to address named entity disambiguation as a ranking problem, where each entity mention should be assigned to its most similar Wikipedia page. The previous works by Bunescu and Pasca (2006) and by Cucerzan (2007) are two of the earliest and most notorious proposals following this methodology. For instance Bunescu and Pasca developed a linear similarity function which considered contextual and categorical features, with weights optimized using supervised learning. The authors also addressed the problem of finding the correct referent for entity mentions not included in Wikipedia (i.e., the NIL entities referenced in the documents). The proposed solution involved defining a similarity threshold below which no assignment was performed by the system.

The named entity disambiguation method proposed by Cucerzan (2007) also relies on Wikipedia as an external knowledge repository. However, contrary to most other approaches, Cucerzan considered the remaining entity references, made in the same document, as the context for each entity reference being disambiguated, instead of the surrounding words. His approach uses the traditional Vector Space Model, comparing document vectors with vectors for the candidate referents. The document vector contains the categories of all possible referents for all entity references found in the text, as well as the number of occurrences of each reference. The referents have binary feature vectors with all the categories and entity references found in the corresponding Wikipedia entry. Interestingly, the similarity measure used by the author does not normalize the feature values, thus privileging important entities, which tend to have longer descriptions, more mentions, and more categories. Also, the author argues that the errors originating from the usage of the *one sense per discourse* principle (Gale et al. 1992) are non-negligible, and he proposed to determine a reference's context in an iterative fashion. Whenever more than one entity scores higher than a predefined threshold, the considered context is shrunk to the level of a paragraph, and possibly to the level of an individual sentence.

Mihalcea and Csomai (2007) proposed the Wikify! system for performing word sense disambiguation based on Wikipedia articles. Their approach involves four main steps, namely one for selecting the candidate referents, two disambiguation modules that independently determine the most probable referent, and a fourth step that checks if the disambiguation modules agree. If there is no agreement for the most probable reference, then no referent is assigned (i.e., we have a NIL entry). On what regards the disambiguation modules, we have that one of them measures the contextual overlap between the reference and candidate referents, while the other leverages on the manually assigned links, existing inside Wikipedia articles, to train a supervised learning approach based on a Naïve Bayes model. The feature vectors include not only the word terms, but also their parts of speech categories (e.g., noun, verb, etc.).

Given the success of learning to rank approaches in document retrieval, Zheng et al. (2010) evaluated learning to rank methods in the named entity disambiguation task. The considered ranking methods included representative algorithms of point-wise (e.g., SVM regression), pair-wise (i.e., Ranking Perceptron), and list-wise (i.e., ListNet) learning to rank approaches. The authors used approximately twenty features to represent candidates, divided into three groups, namely (a) surface features, which measure the name similarity between the reference and the candidate referents, (b) context features, that measure the context similarities, and (c) special features, which represent an entity's geographical and categorical aspects. Experiments showed that the list-wise approach was the most successful. In order to address the cases where references had no correct referent in the knowledge base, the authors supply the top ranked referent to a binary classifier which, using a set of features very similar to the ones used for ranking, decides whether that referent is correct or not.

Place reference disambiguation

Several previous works have also addressed disambiguation tasks focusing on specific types of entities, including place references. Similarly to the general case of named entity disambiguation, the main challenges are related to ambiguity in natural language. For instance Amitay et al. (2004) characterized place reference ambiguity problems according to two types, namely geo/non-geo (e.g., *Turkey*, the country or the bird) and geo/geo (e.g., *Paris* in Texas or in France) problems.

In the context of his PhD thesis, Leidner (2007) surveyed approaches for handling place references in text. He concluded that most methods rely on gazetteer matching for performing the identification, together with NLP heuristics such as default senses (e.g., each disambiguation should be made to the most important referent, estimated with basis on population counts), or geographic heuristics such as the spatial minimality (e.g., disambiguations should minimize the bounding polygon that contains all candidate referents) for performing the disambiguation. Some of the geospatial features used in our system are based on those surveyed by Leidner.

Martins et al. (2010) experimented with the usage of hidden Markov models for the recognition of place references in textual documents, together with a disambiguation model based on SVM regression that leveraged on features also inspired by the heuristics surveyed by Leidner. The regression model captured

correlations between features describing the candidate disambiguations, and the geospatial distance between these candidates and the correct interpretation for each place reference. Initial experiments showed that the SVM regression method could achieve a performance of approximately 0.6 in terms of the F_1 metric, in the task of assigning place references to a correct gazetteer identifier.

Mani et al. (2008) proposed the SpatialML scheme for annotating place references in text, together with a test collection annotated in this format. These authors have also reported experimental results with a statistical ranking model for place reference disambiguation, although without presenting much details about the considered approach. Specifically, the authors report a result of 0.93 in terms of the F_1 measure for the disambiguation of the recognized references.

Lieberman et al. proposed an heuristic method for the resolution of place references in textual documents, focusing on mentions to small and highly ambiguous locations (Lieberman et al. 2010; Lieberman and Samet 2011). The proposed method relies on local lexicons built automatically from regional news documents, involving three main steps, namely (1) inferring local lexicons, (2) performing toponym recognition, and (3) performing toponym resolution. The inference of local lexicons is made by recognizing place names in news articles from local sources, through a simple fuzzy geotagging process which returns a set of possible interpretations for ambiguous toponyms. Toponym recognition is made through a hybrid method that focuses on achieving a high recall, and that uses parts-of-speech tags for identifying proper nouns, together with lexicons and a previously-trained NER system. Finally, toponym resolution is made through a pipeline of heuristics, capturing place prominence and geographic coherence in the interpretations. A particularly interesting contribution from this work is the LGL dataset, containing a collection of news documents from local sources, which can be used to assess the accuracy of disambiguation systems in the case of highly ambiguous place references. In subsequent work, Lieberman and Samet (2012), also proposed to address the place reference disambiguation problem through a binary classification approach relying on a large set of features, by training a Random Forest classifier that decides, for each candidate disambiguation, if it is correct or not. Besides place prominence, the considered features reflect aspects such as the geospatial proximity between toponyms mentioned in a given textual context, and sibling relationships between disambiguations in a geographic hierarchy, for toponyms mentioned in a given textual context (i.e., an adaptive window of textual terms surrounding the place reference).

Speriosu and Baldridge (2013) noted that most previous works that addressed the place reference disambiguation task have neglected textual contexts not corresponding to toponyms, although spatially relevant words like *downtown* or *beach*, that are not explicit toponyms, can be strong cues for disambiguation, given the spatial heterogeneity associated with their distributions. Previously, the connection between non-spatial words and locations had been successfully exploited in data-driven approaches to the problem of document geolocation, estimating the most likely geospatial coordinates for a given textual document (Roller et al. 2012). Therefore, Speriosu and Baldridge proposed to learn resolvers that use all words in local or global document contexts, using similar methods. Essentially, the authors attempt to learn text classifiers for disambiguating toponyms with basis on contextual information obtained from the surrounding text (e.g., classifiers that essentially correspond to a set of language models learned from the textual contents associated with specific regions), training these models using geotagged Wikipedia articles. The authors performed experiments with three different corpora, namely with the collection that was also used in the work of Leidner (2007), the Perseus Civil War and nineteenth century American collection of books written about and during the American Civil War (Smith and Crane 2001), and a dataset containing over one million articles from the English Wikipedia. The obtained results showed that the proposed approach, based on text classifiers, is more accurate than algorithms based solely on spatial proximity or metadata features.

Geocoding the entire contents of textual documents

Several previous studies have addressed a related problem to the one that is considered in this paper, by focusing on the assignment of an encompassing geographic context to the entire contents of a given textual document. For instance the seminal work by

Ding et al. (2000) introduced heuristic techniques for automatically computing the geographic scope of Web pages, based on their textual contents (i.e., based on place names mentioned in the pages, disambiguated through simple heuristics), as well as on the geographic distribution of their hyperlinks.

Authors like Roller et al. (2012) or Dias et al. (2012) have also investigated the automatic geocoding of entire documents, describing supervised methods based on language modeling that, using the raw document text as evidence together with binned representations of the Earth's surface, classify individual documents as belonging to particular regions (i.e., to the bins from the representation of the Earth), and afterwards assign geospatial coordinates of latitude and longitude with basis on these results. The authors concluded that the task of identifying a single location for an entire document provides a convenient way of evaluating approaches for connecting textual documents with locations, although we can have many documents that refer to multiple locations. Nonetheless, these types of approaches can be used in the development of features that aid in the disambiguation of individual place references.

Disambiguating place references over textual contents

The disambiguation of place references can be seen as a particular case of the more general problem of named entity disambiguation. Previous works have, for instance, modeled named entity disambiguation as a ranking task, where the named entity reference is the equivalent to a query, and where the assigned referent should be the highest ranked candidate. In the 2011–2013 editions of the entity linking task of the text analysis conference (TAC–KBP), we participated with a learning-based system similar to the prototype used in the experiments reported here. In these joint evaluation efforts, we obtained accuracy results that were above the median scores of all participants (Anastácio et al. 2011; Santos et al. 2013). Different participants at the yearly TAC–KBP challenge have noted that place references are the most elusive entity type for disambiguation (i.e., better results are generally achieved for person and organization names, as can be seen in the overview paper by Ji and Grishman (2011).

The system used in this study relies on a standard NER tool for initially identifying the place references in a given text (i.e., the English model distributed with Stanford NER[1] version 3.2.0, a toolset that has been described by Finkel et al. (2005)), afterwards addressing the disambiguation sub-task through the following steps:

1. **Query expansion**: Places may be referenced by several alternative names, some perhaps less ambiguous. Given a reference, we apply expansion techniques that try to identify other names, in the source document, that reference the same entity. We considered two simple mechanisms, namely one that finds alternative names by looking for a textual pattern that corresponds to a set of capital words followed by the alternative name inside parentheses (i.e., finding expressions like *United States (US)*), or vice-versa, and another that looks for longer entity mentions in the source text (i.e., *New York* is an expansion for *NY*). Each query is thus expanded with the set of possible alternative names.
2. **Candidate generation**: This step searches the Knowledge Base (KB) for entries that might correspond to the query, based on string similarity. We used a subset of the pages in the English Wikipedia[2] as the KB, containing (1) all the pages with geospatial coordinates in their info-boxes, and (2) all pages categorized in DBPedia[3] as corresponding to either persons, organizations and locations. Although we are just interested in the disambiguation of place references, being able to disambiguate other types of entities occurring in the same documents is also useful for some of the ranking features that are considered latter. Some of Wikipedia's hyperlink structure is also used to obtain alternative names (e.g., disambiguation pages, redirects, anchors, etc.). We return the top 50 most likely entries in the KB (i.e., those whose name(s) are more similar to the entity reference), according to an *n*-gram retrieval model supported by a Lucene[4] index.
3. **Candidate ranking**: This step sorts the retrieved candidates according to the likelihood of being the

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

[2] http://dumps.wikimedia.org/index.html

[3] http://dbpedia.org/index.html

[4] http://lucene.apache.org/index.html

correct referent, using the LambdaMART learning to rank algorithm (Burges 2010) as implemented in the RankLib[5] library. The ranking model was trained to optimize accuracy (i.e., the precision at the first position of the ranked list) over sets of disambiguation examples that were automatically gathered from Wikipedia (i.e., we used hypertext anchors from links towards entities in the Knowledge Base, namely entities associated with geospatial coordinates, occurring in Wikipedia documents different from those in the Knowledge Base). The ranking model leverages on a rich set of features for representing each candidate, including (1) candidate authority features such as the PageRank of the candidate in Wikipedia's hyperlink graph, (2) textual similarity features such as the cosine similarity between bag-of-word representations for the query's source text and the candidate's textual description, (3) topical similarity features based on Latent Dirichlet Allocation (LDA) probabilistic topic models that compute the topical similarity between the query and the candidate's description, (4) name similarity features such as Jaccard's similarity coefficient computed between the candidate's name and the query, (5) entity overlap features such as the number of named entities shared by the query and the candidate's textual description, (6) document-level features such as the number of shared entities between the most likely candidates for each reference in the document, and (7) geographic features. In total, we considered a set of 58 different ranking features.

4. **Candidate validation**: This step decides whether the top ranked referent is an error, resulting from the fact that the correct referent is not given in the knowledge base, through a Random Forest classifier that reuses the features from the ranking model, and that also considers some additional features for representing the top ranked referent (e.g., the candidate ranking score, or the results from well known outlier detection tests, that try to see if the top ranked candidate is significantly different from the others). The validation model uses a total of 64 features (i.e., the 58 ranking features plus 6 validation-only features).

[5] http://people.cs.umass.edu/~vdang/ranklib.html

A thorough description of the named entity disambiguation method is also given in the separate publications that describe our participation in the TAC–KBP evaluation campaign. Due to length restrictions, we do not provide here a detailed description of all the considered features for candidate ranking and validation. However, Table 1 gives a brief overview on the considered features, according to the groups that were listed in the description of the candidate ranking step. Given their particular importance to this study, we describe more thoroughly the set of geographic features that were considered, which essentially try to capture aspects related to place prominence (i.e., important places should be preferred as disambiguations) and geographic coherence in the disambiguations (e.g., textual documents tend to mention places related among themselves). These features, inspired by previous works in the area such as those of Leidner (2007) or of Lieberman and Samet (2012), are as follows:

- **Candidate count**. The number of times that the candidate appears also as a disambiguation candidate for other place references in the same document, or in a window of 50 tokens surrounding the reference (i.e., two separate features, considering two different textual sources as the context).
- **Population count**. This feature takes the value of a particular attribute that is commonly associated with the entries in the knowledge base, corresponding to the number of inhabitants of a given candidate place. A total of 256,497 knowledge base entries have this information available.
- **Geospatial area**. This feature also takes the value of a particular attribute that is commonly associated with places described in the knowledge base, corresponding to the area of the region in squared kilometers. A total of 117,951 knowledge base entries have this information available.
- **Common geo. entities**. The number of place references that are shared by both the query's source text and the candidate's textual description that is taken from Wikipedia (i.e., our knowledge base was built from Wikipedia and, as such, we have textual descriptions associated to the entries in the knowledge base).
- **Jaccard similarity between geo. entities**. The Jaccard similarity coefficient, computed between

Table 1 Overview on the considered ranking and validation features

Feature group	Individual features
Authority	PageRank score of the candidate, over Wikipedia's link graph
	Length of the textual description given in Wikipedia for the candidate
	Number of alternative names associated with the candidate
	Ranking order of the candidate, in the list of all candidates, according to the above scores (i.e. PageRank, description length, and alternative names)
Textual similarity	Cosine similarity between TF-IDF representations for the query document and for the candidate's textual description in Wikipedia
	Ranking order of the candidate according to the cosine similarity
	Cosine similarity between TF-IDF representations, but only considering a window of 50 tokens surrounding all occurrences of the query entity
	Cosine similarity between TF-IDF representations, but only considering the first 150 tokens from the candidate's textual description
	Cosine similarity between TF-IDF representations for the candidate's textual description and for the query entity
	Occurrence of the query entity in the candidate's textual description
	Occurrence of the candidate's name(s) in the query document
Topical similarity	Similarity between topic-based representations for the query document and for the candidate's description, obtained from an LDA topic model, according to the cosine metric and to the Kullback-Leibler divergence
	Match between the most probable LDA topic for the query document and according to the cosine metric and to the Kullback-Leibler divergence
	The probabilities for the most likely LDA topics, for both the query according to the cosine metric and to the Kullback-Leibler divergence
Name similarity	Exact match between the query entity and one of the candidate's names
	Containment between the query entity and one of the candidate's names
	Query entity begins, or ends, with one of the candidate's names
	One of the candidate's names begins, or ends, with the query entity
	Maximum number of common words between query and candidates
	Maximum similarity between the query entity and one of the candidate's names, according to the Levenshtein and Jaro-Winkler character-leve metrics, and the Jaccard, Soft-Jaccard, and Soft-TF-IDF metrics
Entity overlap	Number of common entity names in the query document and the candidate
	The type of the query entity (i.e., person, place, organization, or unknown)
	The type of the candidate (i.e., person, place, organization, or unknown)
	Match between the type of the query entity, and that of the candidate
	Jaccard similarity between the set of entity names in the query document, and the set of entity names occurring in the candidate's description
Document-level	Number of links connecting the candidate to the best-ranked candidates of other entities in the same document, according to Wikipedia's hyperlinks
	Contextual PageRank score, over a graph where nodes are the candidates for all entities in the document, together with their neighbors in
	Wikipedia's graph, and where edges are the existing hyperlinks
	Ranking order of the candidate, according to the contextual PageRank
Geographical	Population and geospatial area of the candidate, as given in Wikipedia …(*this particular group of features is detailed in the paper*)

Table 1 continued

Feature group	Individual features
Validation	Ranking score for the candidate, given by the ranking model
	Mean ranking score of all candidates, and the standard deviation
	Difference between the candidate's ranking score and the mean score
	Standard deviations separating the candidate's ranking score and the mean
	Dixon's Q test for seeing if the candidate's score is an outlier value

the set of place references occurring in the query document, and the set of place references from the candidate's textual description.

- **Missed geo. egtities**. The number of place references in the source text that are not mentioned in the candidate's textual description from Wikipedia.
- **Geospatial distance**. Taking inspiration on the previous work by Dias et al. (2012), we used an efficient similarity search method based on min-hash and locality-sensitive hashing (Broder 1997) to assign geospatial coordinates of latitude and longitude to the entire contents of the query document, afterwards measuring the geospatial distance between the coordinates of the document and those of the candidate, using the geodetic formulae from Vincenty (1975).
- **Geospatial containment**. We again used a similarity search method based on min-hash, but this time to assign the entire contents of the query document to a geospatial region defined over the surface of the Earth, afterwards seeing if the candidate's coordinates are contained within this geospatial region.
- **Average and minimum distance**. The mean, and the minimum, geospatial distance between the candidate disambiguation, and the best candidate disambiguations for other place references in the same document, computed through Vincenty's formulae. The best candidates correspond to those having the highest textual name similarity (i.e., the first in the candidate lists retrieved from Lucene).
- **Distance to closest reference**. The geospatial distance between the candidate disambiguation, and the best candidate for the place reference that appears closer in the same query document. The best candidate is again that which has the highest textual name similarity. This distance feature takes the value of zero if the document contains a single place reference in its text.
- **Area of the geometric hull**. The area of the convex hull, and of the concave hull, obtained from the geospatial coordinates of the candidate disambiguation, and from the coord inates of the best candidates for other place references made in the same document. Best candidates are again those with the highest textual name similarity, i.e. the first candidates retrieved by Lucene.

Notice that the geospatial distance and geospatial containment features, from the previous enumeration, rely on the assignment of a global *geographic context* to the query document (i.e., the entire contents of the query document should be assigned to geospatial coordinates of latitude and longitude, or to a geospatial region, prior to the computation of these features). To efficiently address these particular problems, we relied on a simple method based on interpolating from the coordinates of the most similar geo-referenced entries in the knowledge base. Specifically, we extract all character 7-g occurring in the query document, and we search for the 5 most similar geo-referenced knowledge base entries, in terms of having many character 7-g in common. Efficient nearest neighbor search is implemented through a simple locality-sensitive hashing (LSH) method that leverages min-hash signatures (Broder 1997) to compress the sets of 7-grams associated with each document, preserving the expected Jaccard similarity coefficient between pairs sof documents. Figure 1 illustrates the main steps involved in the assignment of a global geographic context to each query document.

When indexing the knowledge base entries, we start by generating min-hash signatures, with 300 integer values, from the character 7-grams associated with the textual contents of each entry. The signatures are then

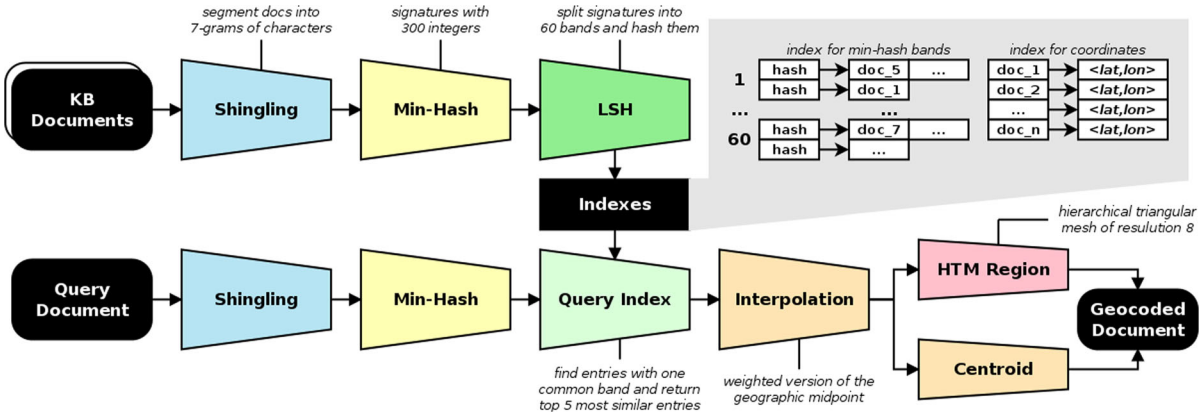

Fig. 1 Geocoding a query document through a locality-sensitive hashing procedure

split into 60 bands of 10 integers each, and we hash the bands into a data structure that associates particular values, for the min-hash bands, to the corresponding knowledge base entries. When geocoding a given query document, we start by also computing a min-hash signature from the textual contents. Knowledge base entries with at least one identical band are considered as candidates, and their Jaccard similarity coefficient, towards the query document, is then estimated using the complete min-hashes. The candidates are sorted according to their similarity, and the geospatial coordinates from the top-5 most similar entries are interpolated, in order to estimate the geospatial coordinates of the query document. The interpolation step is based on finding the geographic midpoint from the available coordinates (Jenness 2008), using the values from the Jaccard similarity coefficient as weights for the coordinates of the top-5 most similar documents.

From the geospatial coordinates discovered through the interpolation procedure, we also infer the geospatial region that is associated with the query document. This is made through the application of the hierarchical triangular mesh procedure (Dutton 1996), which builds a multi-level recursive decomposition of the Earth's surface into triangular regions with roughly the same area. We assign the query document to the triangular region that contains the corresponding geospatial coordinates, considering a resolution of 8 for the multi-level triangular decomposition (i.e., the Earth's sphere is initially divided into 8 spherical triangles, and each triangle is then recursively subdivided 7 times into 4 separate spherical triangles, finally resulting in regions of approximately 1000 km^2 each).

Due to length restrictions, we do not detail in this paper the particular issue of document geocoding through LSH and min-hash, particularly in the experiments reported in section (Experimental validation). However, the parameters involved in this procedure (i.e., the size of the character *n*-grams, the size of the min-hash signatures, the number of LSH bands, the number of nearest neighbors, and the resolution for the triangular decomposition) where tuned through a particular set of experiments. Many other parameters involved in the computation of some of the considered features were also tuned through specific experiments (e.g., the number of topics in the LDA model that was used in the computation of the topical similarity features was tuned by minimizing perplexity on a held-out set of documents (Blei et al. 2003).

Experimental validation

We experimentally compared two different configurations of the proposed place reference disambiguation approach, namely one configuration corresponding to a standard named entity disambiguation setting (i.e., a system similar to the ones that have participated in previous TAC–KBP named entity disambiguation competitions), and another introducing the usage of the geographic features described in the previous section. We used documents from a recent dump of the

English Wikipedia (i.e., the dump dated from October the 1st 2013), both as the textual sources containing place references to be disambiguated, and as the Knowledge Base (KB) entries supporting the disambiguation. Besides Wikipedia, we also measured results over two previously available collections for place reference disambiguation studies, namely the local-global lexicon (LGL) dataset introduced by Lieberman et al. (2010), and the SpatialML dataset available from the Linguistic Data Consortium (Mani et al. 2008).

In brief, SpatialML is an annotation scheme that considers a `PLACE` tag for annotating place references in text. There are nine different attributes defined for the `PLACE` tag, with the values of these attributes expressing the semantics of the annotated place references. A `LATLONG` attribute contains the geospatial coordinates corresponding to the place, encoded in a textual format (e.g., numeric values for degrees, minutes, and seconds, referring to latitude and longitude coordinates in the WGS-84 datum). SpatialML distinguishes between named places (e.g., names for cities, points of interest, etc.) and place nominals (e.g., *the village*, *the city*, etc.), and our study focused on the named places containing associations to geospatial coordinates, ignoring the nominals. A document collection annotated according to the SpatialML scheme has been made available by Mani et al. (2008), containing a total of 428 documents with 210,065 words.

As for the LGL dataset, it focuses on articles from a variety of small and geographically-distributed newspapers, thus being better-suited for evaluating place reference disambiguation on a local level and at a thinner granularity. The dataset contains a total of 588 articles with 213,446 words.

Table 2 presents characterization statistics for the considered datasets (i.e., for the SpatialML and LGL datasets, as well as for the entity disambiguation dataset built from Wikipedia itself, by using hypertext anchors from links in the Wikipedia documents as the query entities to be disambiguated). In the experiments, about 80 % of the documents from the entity disambiguation dataset that was built from Wikipedia were used for training the disambiguation system, whereas the remaining 20 % of the documents were used for model testing. In Table 2, we specifically present the total number of place references available in each collection, and also the number of place references of the types *country* and *city*. The remaining entities are referred to as *other* (e.g., states, continents, lakes, counties, etc.). As for the knowledge base supporting the disambiguation, it has 1,265,307 entries obtained by filtering the full set of English Wikipedia pages and keeping those that contain geospatial coordinates in their *info-boxes*, or that correspond to entities described in DBPedia as either locations, persons, or organizations, even though these last entries may lack an association to coordinates. Notice that in the case of the Wikipedia dataset, Table 2 only considers the non-NIL queries (i.e., in Table 2 we only present the number of disambiguation queries, from the Wikipedia documents, that correspond to non-NIL knowledge-based entries containing associations to geospatial coordinates, although these same documents may also contain references corresponding to NILs). When considering the NILs, we have that the Wikipedia dataset contains a total of 68,833 training references, and a total of 17,186 references for testing.

Table 2 Number of geo-referenced place references in the considered evaluation datasets

Dataset	Country references	City references	Other references	Total references
SpatialML	2,354	1,392	697	4,443
LGL	785	2,186	1,491	4,462
Wiki (test)	—	—	—	12,446
Wiki (train)	—	—	—	49,813

We mainly used the geospatial distance between the coordinates returned as the disambiguation, and the correct geospatial coordinates, as the evaluation metric. We also used accuracy (i.e., the precision at the first ranking position, obtained from the ratio between the number of queries disambiguated to the correct KB entry, over the total number of queries) and the Mean Reciprocal Rank (i.e., the average of the multiplicative inverses of the ranking positions for the correct disambiguation entries) evaluation metrics, as well as the number of candidate misses (i.e., the number of times the correct candidate was not even chosen in the candidate generation step). When measuring accuracy and the MRR in the SpatialML and LGL datasets, we considered that a distance smaller than 5, 50 or 250 km corresponds to a correct

disambiguation. This was required since, in these two datasets, the place references were not originally disambiguated to the corresponding Wikipedia entries, only having associations to the corresponding geospatial coordinates. Although measuring accuracy and the MRR with basis on thresholds over the geospatial distance can have some problems (e.g., a place reference can be disambiguated to a location that is close by, but that is outside its real borders), we still believe that it can be useful to analyze the results in terms of these metrics.

Table 3 therefore presents the results obtained in terms of the average and median distances, in kilometers, between the geospatial coordinates assigned by the algorithm and the correct coordinates as given in the annotations of the different datasets. The table also presents the number of place references that were geocoded with the proposed method (i.e., that were assigned to a knowledge base entry containing an association to the corresponding geospatial coordinates of latitude and longitude).

Notice again that the distance-based evaluation approach has some limitations, particularly for the case of the SpatialML and LGL datasets where we can only assess the correctness of our results by comparing geospatial coordinates (i.e., references in these datasets were not originally disambiguated to Wikipedia entries). Given that distances can only be computed for the place references that were resolved to geospatial coordinates (i.e., for those that are disambiguated to knowledge base entries associated with coordinates), we have that different strategies may result in a different number of disambiguations being used in the computation of averages (i.e., a particular configuration of the system can produce more NIL results, or it may return more disambiguations to knowledge base entries without coordinates, and we will not directly account with these results in the computation of average distances for the configuration under evaluation). To address these limitations, Table 3 presents results for different experimental settings on what concerns the measurement of distances, namely (1) a **regular** setting where distances were only measured for those candidates to which the system assigned a non-NIL disambiguation having geospatial coordinates in Wikipedia, (2) a **maximum distance** setting where we penalize all disambiguations made to NILs or to Wikipedia pages having no coordinates, by assigning them with a distance value of 20,038 km (i.e., half of the length of the equatorial circumference of the Earth), (3) a setting where we only used the results from the **ranking** module, ignoring NIL classifications and measuring distances towards the coordinates of the top-ranked candidate, and (4) a setting in which we used the **min-hash** procedure to assign geospatial coordinates to the non-NIL disambiguations that did not originally have coordinates in Wikipedia.

Table 4 presents results in terms of the average accuracy (i.e., P@1) and MRR evaluation metrics, and there we can see that the accuracy across the different datasets and for both configurations remains approximately similar and reasonably high. The results in Table 4 were measured using the regular strategy from Table 2. We can calculate exact accuracy and MRR values on experiments with the Wikipedia dataset, given that in this case we have the correct Wikipedia disambiguation associated with each reference, therefore not needing to measure the results with basis on thresholds over the geospatial distance. However, for facilitating comparisons across the datasets, we still present the results achieved with distance based metrics for Wikipedia. As for the remaining datasets, the P@1 and MRR metrics were measured using the disambiguations made according to the regular strategy, seeing if candidates had a distance to the correct disambiguation bellow a given threshold. In Table 4, we also present the number of non-NIL references where a correct disambiguation has been made, and the number of references in which our system failed to retrieve an appropriate disambiguation candidate (i.e., the candidate misses).

The results from Tables 3 and 4 show that the system's performance benefits from the introduction of the geographic features in some situations, although one can also observe that the improvements are not significant. The geo-specific features seem to have a limited impact over a strong baseline system, which uses an extensive set of features based on textual similarity. It is important to notice that the relatively high values for the average and median geospatial distances are often due to cases such as large countries, whose centroid geospatial coordinates appear differently in Wikipedia than in the original annotations given in the SpatialML and LGL datasets. One aspect that we verified, through a detailed analysis of the results, is that the system tends to assign more NILs when using the set of geographic features. We can see

Table 3 The obtained results with the four different distance-based evaluation methodologies

	Without geographic features			With geographic features		
	Wikipedia	LGL	SpatialML	Wikipedia	LGL	SpatialML
Regular						
Geocoded	11,865	3,127	3,549	11,777	3,167	3,559
Avg. (km)	23.962	763.137	**136.103**	**21.739**	**742.040**	139.615
Med. (km)	**0.000**	**2.435**	**27.820**	**0.000**	2.790	28.706
Max. dist.						
References	12,446	4,462	4,443	12,446	4,462	4,443
Avg. (km)	**958.228**	6,529.897	4,140.572	1,097.631	**6,342.136**	**4,098.593**
Med. (km)	**0.000**	92.734	**54.776**	**0.000**	**79.896**	**54.776**
Rank. only						
Geocoded	12,361	3442	4,148	12,439	4,270	4,379
Avg. (km)	40.558	783.639	**174.072**	**40.383**	**735.475**	231.445
Med. (km)	**0.000**	**3.672**	**54.282**	**0.000**	15.484	**54.282**
Min-hash						
Geocoded	11,898	3,555	3,758	11,777	3,167	3,560
Avg. (km)	33.980	906.836	285.790	**21.739**	**742.040**	**140.871**
Med. (km)	**0.000**	**7.965**	**41.439**	**0.000**	2.790	54.473

this by comparing the *regular* and the *ranking only* tests, where we can see that the number of disambiguated references rises when ignoring the NILs in the geographic models. Moreover, using geographic features, the system usually performs the disambiguation to a KB entry containing coordinates, and this is why in the *min-hash* test there is almost no difference in comparison to the *regular* test (i.e., all the disambiguations assigned in the geographic test already contained coordinates).

In Table 5, we present the results obtained with 4 different baseline methods that, instead of using a learning to rank method, choose the correct disambiguation with basis on one of the features that is considered by the system. These baselines are (1) choosing the candidate with the largest population, (2) choosing the candidate with the largest area, (3) choosing the candidate with the highest PageRank over the Wikipedia hyperlink graph, and (iv) choosing the candidate with the highest textual similarity towards the query document. In the case of baselines (1) and (2), and given that there are many KB entries for which we do not know the corresponding area and/or population, we use the textual similarity feature as a second candidate ranking criterium (i.e., if we do not know the population/area of the candidate disambiguations, then we choose the candidate having the highest textual similarity).

In Table 5, the average and median distances are measured using the regular evaluation setting that was considered for Table 3, while the results in terms of accuracy and of MRR were measured with the threshold value that corresponds to a distance bellow or equal to 5 km. The results for the baselines in Table 5 are generally lower than those obtained with the learned ranking model, thus showing that combining multiple features is indeed beneficial for place reference disambiguation.

Regarding comparisons with the current state-of-the-art, the authors of the SpatialML dataset reported a result of 0.93 in terms of the F_1 measure for the disambiguation of the recognized geographical expressions (Mani et al. 2008). On what regards the LGL dataset, the most recent study reported on a disambiguation quality of approximately 0.95 in terms of both the precision and F_1 measures (Lieberman and Samet 2012). However, since these authors did not use Wikipedia entries as the knowledge base supporting place reference disambiguation in their studies, a direct comparison with these previous results cannot be made.

Table 6 shows the place names with the highest average errors in terms of the geospatial distance,

Table 4 Results obtained with and without the proposed set of geospatial features

	Without geographic features			With geographic features		
	Wikipedia	LGL	SpatialML	Wikipedia	LGL	SpatialML
Exact						
Total	12,446	–	–	12,446	–	–
Correct	12,027	–	–	**12,074**	–	–
Misses	5	–	–	5	–	–
P@1	0.966	–	–	**0.971**	–	–
MRR	0.980	–	–	**0.983**	–	–
≤5 km						
Geocoded	11,865	3,127	3,549	11,777	3,167	3,559
Correct	**11,741**	**1,734**	1,499	11,660	1,720	1,497
Misses	3	773	1,829	4	831	1,836
P@1	0.989	**0.554**	**0.420**	**0.990**	0.543	0.417
MRR	0.994	**0.620**	**0.441**	**0.995**	0.610	**0.441**
≤50 km						
Geocoded	11,865	3,127	3,549	11,777	3,167	3,559
Correct	**11,766**	2,028	1,942	11,687	**2,067**	**1,943**
Misses	2	371	1,207	3	392	1,212
P@1	0.991	0.648	**0.544**	**0.992**	**0.653**	0.541
MRR	0.995	0.718	0.564	**0.996**	**0.723**	**0.570**
≤250 km						
Geocoded	11,865	3,127	3,549	11,777	3,167	3,559
Correct	**11,787**	2,409	**3,166**	11,705	**2,456**	3,141
Misses	2	92	107	3	94	118
P@1	0.993	0.770	**0.887**	**0.994**	**0.775**	0.876
MRR	0.996	0.840	**0.912**	**0.997**	**0.849**	0.909

collected from the LGL (on the right) and the SpatialML (on the left) datasets, and also the place references that occur more frequently in each dataset. The errors in terms of distance were measured for the case of models using the set of geographic features and the regular methodology from Table 3.

After a careful analysis of each case, we can see that for LGL, which is a dataset containing mostly references to small places, the most significant errors are related to the decision of disambiguating to the most popular entry in the knowledge base. For instance, the references *Jordan*, *Malta*, *Belgrade*, or *Paris* (which occur frequently, and also have a high average distance in the obtained results) were disambiguated to the most popular entries that share these names (i.e., Jordan in the middle east, Paris in France, etc.), although these cases referred to small towns in the United States. We also have cases like *Georgia*, that have a considerably high average distance associated to them, despite being correctly disambiguated. This happens because the centroid geospatial coordinates appear differently in the dataset and in Wikipedia. As for the SpatialML dataset, we can also see errors resulting from choosing the most popular entry (e.g., *Aberdeen*, a reference to a city in the United States that the system resolved to the city in Scotland). Other errors were not so clear to interpret, but many appear to come from wrongly geocoding the support document, through the method based on min-hash and LSH.

In a particular experiment, we retrieved the closest KB entry for each of the place references present in either the SpatialML or the LGL datasets, and we then measured the average and the standard deviations for these shortest distances between the KB entries and

Table 5 Results with the four baseline methods that were considered

	Population			Geospatial Area		
	Wikipedia	LGL	SpatialML	Wikipedia	LGL	SpatialML
Num. references	12,446	4,462	4,443	12,446	4,462	4,443
Geocoded	10,096	2,968	3,846	10,159	3,314	3,810
Precision@1	0.473	0.219	0.270	0.415	0.163	0.220
MRR	0.666	0.386	0.344	0.598	0.301	0.307
Average dist. (km)	1,694.805	2,126.289	663.798	1,888.250	1,781.425	893.229
Median dist. (km)	17.464	393.154	151.894	99.174	538.262	227.998
	PageRank			Textual Similarity		
	Wikipedia	LGL	SpatialML	Wikipedia	LGL	SpatialML
Num. references	12,446	4,462	4,443	12,446	4,462	4,443
Geocoded	10,027	2,969	3,926	9,617	2,999	2,599
Precision@1	0.550	0.281	0.288	0.789	0.473	0.424
MRR	0.743	0.420	0.354	0.972	0.602	0.525
Average dist. (km)	1,428.010	1675.110	671.035	290.061	756.242	303.354
Median dist. (km)	0.0	290.066	151.894	0.0	11.439	19.197

the ground truth. We respectively obtained results of 11.79 ± 30.33 km for the SpatialML dataset, and of 6.53 ± 22.08 km for the LGL dataset. These values can be seen as lower bounds on the distance errors that can be obtained by a disambiguation system that uses our particular knowledge base.

In a separate set of experiments, we also attempted to quantify the impact that a variable such as the size of the query document has on the quality of the results. The chart in Fig. 2 shows the distribution of the error values that were obtained for documents of different sizes, in terms of the distance towards the correct disambiguations, for the SpatialML and LGL datasets and using the regular methodology and the full set of features. The results show that there are no important differences when analyzing different documents of different sizes. The proposed method seems to be equally able to disambiguate place references in documents of different sizes.

Figure 3, on the other hand, shows the distribution of the error values, in terms of geospatial distances, for different types of places being referenced, in the case of the Wikipedia dataset (i.e., for the 10 place types that appear more frequently). The per-type analysis from Fig. 3 was made through the usage of a mapping from Wikipedia categories, for the individual pages associated with the different categories, into the categories from the OpenCyc ontology, using the methodology described by Pohl (2010). The OpenCyc category system offers a better organization of concepts than that of Wikipedia, facilitating the analysis of our results on a per-type basis. We should notice that the entries that were correctly disambiguated (i.e., place references having a disambiguation error of 0 km, given that in test dataset built from Wikipedia we the place references assigned to the exact same geospatial coordinates that appear in the KB) are not represented in Fig. 3 (i.e., we only show the cases corresponding to errors).

Through the analysis of the results from Fig. 3, we can see that the different types of places appear to have a distinct distribution in terms of the errors that are produced. For instance places associated with the category *Port Cities* seem to be particularly hard to disambiguate, whereas place types such as *Railway Stations* generally present small errors, in terms of geospatial distance.

In Table 7, we illustrate the obtained results for the case of four short example documents, containing references to some of the place references that were shown in Table 6. The phrases in bold correspond to the place references that were recognized by the

Table 6 References with the highest distances towards the correct result, or with the highest occurrence frequencies, respectively in the LGL and SpatialML datasets

	Occurrences	Average error (km)
LGL		
Jordan	11	10108.346
Clare	3	9913.766
Petersburg	1	8543.474
Malta	2	8401.731
Belgrade	3	8166.633
US	83	0.000
Georgia	62	422.304
Paris	55	7135.410
Texas	52	79.896
Israel	51	54.282
SpatialML		
Baden	1	6663.568
Bristol	1	6262.896
Loudoun	2	5538.943
Aberdeen	1	5518.309
Westwood	3	4117.233
Iraq	483	54.776
Baghdad	268	3.715
Washington	096	37.619
Israel	91	54.282
Iran	90	112.775

Stanford named entity recognition system. The geospatial coordinates in a regular typesetting correspond to the cases where the place reference was correctly disambiguated into the corresponding Wikipedia page, whereas coordinates in bold and typeset in italics indicate an error in the disambiguation.

Conclusions and future work

Despite the recent advances in the general area of named entity disambiguation, few previous works have specifically focused on place references, leaving several open questions (e.g., to what extent can the entity disambiguation systems from the current state-of-the-art in the area of information extraction, be effectively used for place reference disambiguation). In this work, we report on an extensive set of experiments with an adapted version of a state-of-the-art named entity disambiguation system, evaluating its performance on the specific task of place reference disambiguation, using previously existing datasets such as the SpatialML dataset described by Mani et al. (2008), or the LGL corpus. Our experimental results demonstrate that the proposed system is indeed effective, showing that out-of-the-box learning algorithms and relatively simple features can obtain a high accuracy. Our results also showed that the introduction of geo-specific features seems to have a limited impact over a strong baseline, which was essentially based on textual similarity.

Place reference disambiguation is a particularly interesting problem from the perspective of computational approaches for natural language processing, and the task can be of use to a wide range of studies in the digital humanities and the computational social sciences. We argue that linking place references, occurring in

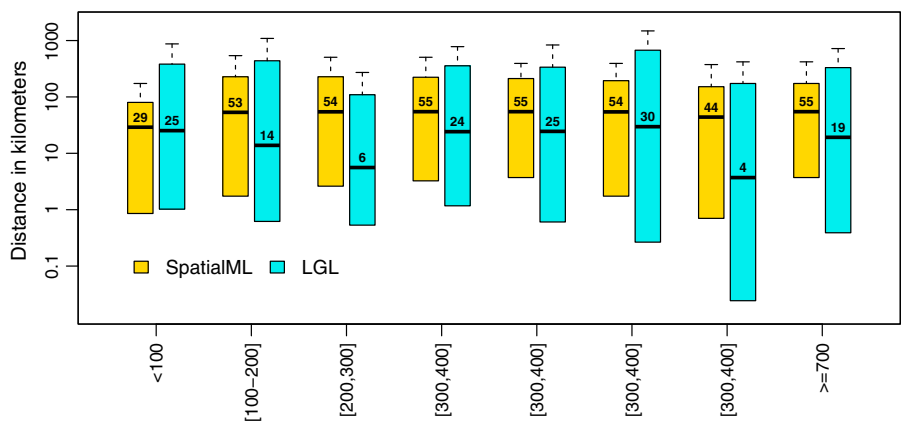

Fig. 2 Geospatial distances when disambiguating references in documents of different sizes

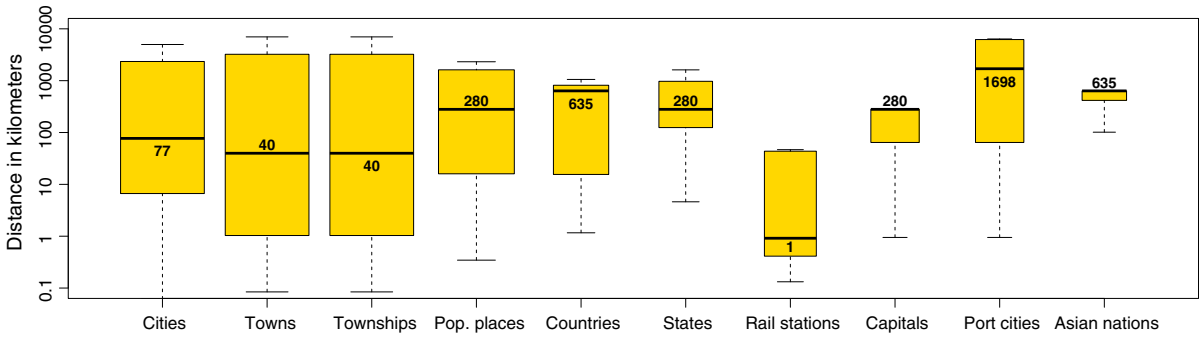

Fig. 3 Geospatial distances for different types of place references being referenced

Table 7 Disambiguations that are produced in the case of four example documents

Example document	Place reference	Latitude	Longitude
From 1922 to 1936, **Georgia** was part of the Transcaucasian Socialist Federative Soviet Republic, which united the Soviet Socialist Republics of **Georgia**, **Armenia** and **Azerbaijan**	Georgia	*41°43′N*	*44°47′E*
	Armenia	40°11′N	44°31′E
	Azerbaijan	40°25′N	49°50′E
Paris has actually been named the best small town in the state of **Texas**, in a 1998 book by Kevin Heubusch	Paris	***48°51′N***	***2°21′E***
	Texas	31°00′N	100°00′W
Bristol has the fifth highest per-capita GDP of any city in the state, after larger cities like **London** or **Glasgow**, and third highest GDP per capita of any English city	Bristol	51°27′N	2°35′W
	London	51°30′N	7°39′W
	Glasgow	55°51′N	4°15′W
Boeing Employees' Credit Union remains based in the **Seattle** area, though it is now open to all residents of **Washington**	Seattle	47°36′N	122°19′W
	Washington	***38°53′N***	***77°12′W***

different types of textual documents, to unambiguous identifiers such as geospatial coordinates is a fundamental geographical tool for scholarly research in these areas. By solving this problem with a high accuracy, we can support a wide range of studies involving the access and analysis of information encoded over large document collections, through geospatial constraints (Brown et al. 2012; Adams and McKenzie 2013).

Despite the interesting results reported here, there are also many other ideas for improving the place reference disambiguation system, and for improving the validation methodology that was considered in this paper. We plan, for instance, to evaluate the specific contribution that different groups of features have on the disambiguation performance, besides just singling-out the set of geographic features.

Specifically regarding the geographic features, and noticing that efficiently geocoding textual documents is essential to the computation of some of the considered features, we would like to experiment with alternative document geocoding methods. Adams and Janowicz (2012) have for instance proposed a method for geocoding textual documents with basis on their contents, leveraging a combination of the latent dirichlet allocation (LDA) topic model and the kernel density estimation (KDE) technique. In brief, the authors discover the parameters an LDA topic model with basis on a large document collection and, for each of the K topics resulting from the LDA model, the authors estimate a density surface encoding the incidence of textual contents, related to that particular topic, over different geographical regions. The KDE technique is used to estimate the topic-specific density surfaces, with basis on the coordinates of geo-referenced textual documents whose majority of contents are generated by the particular topic. To geocode the

textual contents of a previously unseen document, the authors start by using the LDA model to discover the document's representation as a mixture of K topics. They then compute a surface encoding the likelihood of having the document associated with specific geographic regions, with basis on a weighted average of the surfaces associated with the K different topics. The authors finally geocode the document by finding the geospatial region, or the specific point, corresponding to the highest likelihood. In our place reference resolution system, given that an LDA topic model is already used in the computation of some of the considered features (e.g., we use the Kullback-Leibler divergence between topical representations for the source document and for the candidate entry, as one of our features), we could perhaps use a similar method to geocode the textual documents, instead of relying on a simple heuristic method based on interpolating from the k nearest neighbors in terms of common n-grams, i.e. from the most similar documents that are already themselves associated with geospatial coordinates.

Also regarding the geo-specific features that are used in our system, it would be interesting to experiment with the introduction of additional features capturing the fact that the source documents can either have a very broad geographical context (i.e., documents that discuss only very large geographical areas, corresponding to entire countries or continents), or they can relate to small geographic regions (e.g., documents corresponding to local news). Through features indicating if the query document is indeed discussing a small/local or a broad geographic region, we could perhaps train our models to better weight the individual contribution of the remaining geographical features. Our experiments with the LGL dataset have, for instance, shown that the disambiguation errors were often due to the assignment of popular places (e.g., *Paris*, the French capital) to small towns having the same name, and features capturing place prominence, such as the area or the number of inhabitants, should perhaps have a lower weight in the context of documents having a small/local geographic scope. For future work, to capture the fact that a source document can have either a local or a global geographic scope, we would like to experiment with the training of a binary classification model leveraging on Wikipedia contents (e.g., textual contents associated to either very large or small geographic areas), which could then be used to provide a feature to our disambiguation models, capturing the geographic context (i.e., local versus global) of the source documents.

The information retrieval community has also recently started to look at the problem of relational learning to rank, explicitly considering cases in which there exists a relationship between the objects to be ranked (Qin et al. 2008). For future work, and noticing that entities referenced in the same context (e.g., in the same document or in documents from a same collection) should be similar to one another, we would particularly like to experiment with relational learning methods in order to explore document- or collection-level disambiguation directly at the level of the learning algorithm, going beyond the document-level features that were already considered here.

References

Adams, B., & Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of the international AAAI conference on weblogs and social media*.

Adams, B., & McKenzie. (2013). Inferring thematic places from spatially referenced natural language descriptions. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, Springer.

Amitay, E., Har'El, N., Sivan, R., & Soffer A. (2004). Web-a-where: Geotagging web content. In *Proceedings of the ACM SIGIR conference on information retrieval*.

Anastácio, I., Calado, P., & Martins B. (2011). Supervised learning for linking named entities to wikipedia pages. In *Proceedings of the text analysis conference*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(1), 993–1022.

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings of the conference on compression and complexity of sequences*.

Brown, T., Baldridge, J., Esteva, M., & Xu, W. (2012). The substantial words are in the ground and sea: Computationally linking text and geography. In *Texas studies in literature and language: Linguistics and literary studies: Computation and convergence*.

Bunescu, R., & Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European conference of the association for computational linguistics*.

Burges, C. J. C. (2010). *From RankNet to LambdaRank to LambdaMART: An overview*. Microsoft research technical report.

Cucerzan, S.-P. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning*.

Dias, D., Anastácio, I., & Martins, B. (2012). Geocoding textual documents through hierarchical classifiers based on language models. *Linguamática, Revista para o Processamento Automático das Línguas Ibéricas, 4*(2), 13–25.

Ding, J., Gravano., & Shivakumar, N. (2000). Computing geographical scopes of web resources. In *Proceedings of the International Conference on Very Large Data Bases,* Cairo, Egypt.

Dutton, G. (1996). Encoding and handling geospatial data with hierarchical triangular meshes. In *Advances in GIS research II*.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Annual Meeting on Association for Computational Linguistics,* Michigan, USA.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the MLT workshop on speech and natural language*.

Jenness, J. (2008). Calculating areas and centroids on the sphere. In *Proceedings of the annual ESRI international user conference*.

Ji, H., & Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the annual meeting of the association for computational linguistics*.

Leidner, J. (2007). *Toponym resolution: A comparison and taxonomy of heuristics and methods*. PhD thesis, University of Edinburgh.

Lieberman, M., & Samet, H. (2011). Multifaceted toponym recognition for streaming news. In *Proceedings of the ACM SIGIR conference on information retrieval*.

Lieberman, M., & Samet, H. (2012). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the ACM SIGIR conference on information retrieval*.

Lieberman, M., Samet, H., & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the IEEE international conference on data engineering*.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., & Wellner B. (2008). SpatialML annotation scheme, corpora, and tools. In *Proceedings of the international conference on language resources and evaluation*.

Martins, B., Anastácio, I., & Calado, P. (2010). A machine learning approach for resolving place references in text. In *Procedings of the AGILE international conference on geographic information science*.

Mehler, A., Bao, Y., Li, X., Wang, Y., & Skiena, S. (2006). Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics, 12*(5).

Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the ACM conference on conference on information and knowledge management*.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes, 1*(30), 3–26.

Pohl, A. (2010). Classifying the wikipedia articles into the opencyc taxonomy. In *Proceedings of the ISWC workshop on the web of linked entities*.

Qin, T., Liu, T.-Y., Zhang, X.-D., Wang, D.-S., Xiong, W.-Y., & Li, H. (2008). Learning to rank relational objects and its application to web search. In *Proceeding of the international conference on world wide web*.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning*.

Santos, J., Anastácio, I., & Martins, B. (2013). The entity linking system from dmir at the 2013 tac-kbp entity linking tasks. In *Proceedings of the text analysis conference*.

Smith, D. A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *Proceedings of the European conference on digital libraries*.

Speriosu, M., & Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the annual metting of the association for computational linguistics*.

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review, XXIII*(176), 88–93.

Zheng, Z., Li, F., Huang, M., & Zhu, X. (2010). Learning to link entities with knowledge base. In *Proceedings of the conference of the North American chapter of the association for computational linguistics*.