JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 17 (2018), pp. 31–62
doi:10.5311/JOSIS.2018.17.417

RESEARCH ARTICLE

Georeferencing places from collective human descriptions using place graphs

Hao Chen, Maria Vasardani, and Stephan Winter

Dept. of Infrastructure Engineering, The University of Melbourne, Australia

Received: February 2, 2018; returned: May 21, 2018; revised: September 4, 2018; accepted: November 19, 2018.

Abstract: Place descriptions in everyday communication or in online text provide a rich source of spatial knowledge about places. Such descriptions typically consist of references to places and spatial relationships between them. An important step to utilize such knowledge in information systems is georeferencing the referred places. Beside place name disambiguation, another challenge is that a significant proportion of place references in such descriptions are not official place names indexed by gazetteers, thus cannot be resolved easily. This paper presents a novel approach for georeferencing places from collective descriptions using place graphs, regardless of whether they are referred to by gazetteered names or not. The approach leverages spatial relation models for approximate locating and matching. Different models are proposed and evaluated using several metrics.

Keywords: place description, place graph, georeferencing, qualitative spatial relation, spatial language

1 Introduction

Place descriptions are a way of encoding and transmitting spatial knowledge about places between individuals [51, 52]. They are conveyed in either verbal or written form in everyday communication. The web provides a plethora of place descriptions such as news articles, social media texts, trip guides, and tourism articles [24]. Place descriptions typically provide a qualitative reference system for describing geographic locations, and consist essentially of references to places and their qualitative spatial relationships, e.g., "The *courtyard* is on the *campus*, beside the *clock tower*," describing the location of the courtyard in relation to two other places.

In order to enable computers to digest the place information communicated in such descriptions, the references to places have to be located in space through georeferencing. The place information can then be used to enrich geographic information systems and to facilitate a wide range of applications such as geographic information retrieval [22, 40, 47], smoothing human-computer interaction [9, 41, 55] by providing an interface with the capacity of locating place references given in vernacular contexts, as well as for general place search and analysis purposes.

In this research, we regard place descriptions as textual documents. Georeferencing place from text is not a new problem. With the increasing volume of unstructured text documents published online, and the growing need for place related information, extensive studies have focused on identifying and locating place names from text, e.g., [29,31], thanks to the rapid development of text mining and natural language processing techniques.

However, methods proposed in these previous studies are typically based on identifying and disambiguating gazetteered, i.e., officially indexed, place names, and ignore references to places that are not. Everyday place descriptions, on the other hand, often include vernacular and thus potentially non-gazetteered place references, such as synonyms or place types (e.g., "the large square"), references to places that are from too fine-grained environments to be captured by gazetteers (e.g., "the dean's office"), and references to vague places or vernaculars that exist only in limited contexts (e.g., "*the BBQ area* near the tree in front of our department"). In communications, such places are typically located by providing spatial relations to some landmarks. In the previous example, the "campus" is a functional reference instead of an official name, and "courtyard" and the "clock tower" are referring to non-gazetteered places, but all are used as landmarks.

Non-gazetteered places, including fine-grained places, often have higher ambiguities to resolve than coarse-grained ones [3]. The current approaches, mostly designed for larger geographic features such as populated places (e.g., cities or countries) or natural geographic features (e.g., rivers or mountains), use heuristics for ambiguity resolution based on the sizes of the features (e.g., population) or their hierarchical containment relationships. Such approaches are not applicable for features in everyday communication, which are significantly more numerous and more similar to each other. Thus, this research is facing a more complex problem than the one addressed previously.

In order to address both the reference and the ambiguity problem, this research will take advantage of *place graphs*, representing the extracted spatial knowledge from place descriptions. In these place graphs references to places are embedded in a neighbourhood, or spatial context. The hypothesis of this research is that integrating place graphs into the georeferencing process allows to address these two problems, regardless of whether they are referred by gazetteered names or not. The approach below first identifies and disambiguates anchor place names that are gazetteered and easier to resolve. Next, it derives approximate location representations for the remaining place references based on their spatial relationships to the anchors. The derived approximate locations are used for best-matching to gazetteer entries, as well as for locating these places on a map even if they are not gazetteered. Thus the contributions of this paper are:

1. an approach that leverages spatial relations to georeference places from place descriptions, even if they are referred by references that are not officially indexed names.
2. the use of both formal models for spatial relations, as well as contextualized probabilistic-based ones that are contextualized and were trained in a reverse-

engineering manner. We also propose a method to integrate different search spaces to compute the approximate locations of places.
3. a matching procedure to link unrecognizable place references to gazetteer entries based on string and semantic similarities and relation satisfaction.
4. a test of our approach on several datasets collected from different sources and with different sizes, granularity, and place density, evaluating the approach using various metrics.

The remainder of the paper is structured as follows. In Section 2 a review of related work is given. Section 3 clarifies the input and the georeferencing approach. Section 4 shows implementation and experiment results on test datasets. Section 5 presents a discussion on the case study and the obtained results. Section 6 concludes this paper.

2 Related work

People talk about space by referring to places [56]. Place-based research is an increasingly popular field in GIScience as an alternative and complement to research with surveying based data, and its importance has been widely acknowledged (e.g., see [16,17,55]). In this section, related works about georeferencing place from text, modelling qualitative spatial relations, and place graphs will be introduced.

2.1 Georeferencing place from text

In order to locate place names on a map with precise coordinates, gazetteers are often used. A gazetteer typically contains three core components: place name, feature type, and footprint [20] and is often regarded as a geospatial dictionary of geographic names. A place name is what people usually use when they search for this place, and is typically an official name gazetteered by an authority. Some gazetteers may also store alternative names. A place type is a category from a feature type thesaurus for classifying places according to their semantics. A footprint represents the location of a place, typically by a single coordinate tuple for the center of the place, and sometimes by a polygon or a polyline instead.

The task for resolving the locations of places referred to in text is often called *toponym resolution* [29], and it is a core task for building geographic information retrieval and document geotagging systems [32, 40]. It comprises two tasks, namely place name *recognition*, and *disambiguation* (disambiguation because place names are rarely unique; e.g., geonames.org lists 14 populated places "Melbourne" worldwide). Toponym recognition is typically done by gazetteer matching, thus non-gazetteered place references are ignored. In this research, place references were instead extracted using a parser that is able to capture references identified by spatial relationships [23, 34], regardless of whether they are gazetteered or not. Many of these place references are even not proper names, but descriptions. For the disambiguation task, extensive methods have been proposed that can be classified as map-, knowledge-, or machine learning-based, and are often used in conjunction with various heuristics. The selection of an approach is highly task and source dependent [3]. Most of these existing approaches cannot be applied directly for the task of this research, as they are designed for coarse-granularity places such as natural geographic features and populated places. Fine-grained places may not have sufficient differentiators

for such heuristics, e.g., popularity, prominence, and hierarchical containment relationships [1, 4], and are often more ambiguous to resolve.

Some other studies attempt to overcome the limitations. Palacio et al. developed an approach for disambiguating fine-grained toponyms based on Euclidean distance and topographic similarity to anchor toponyms from the discourse [39]. However, the approach still requires the toponyms to be disambiguated being captured by some databases such as a gazetteer, in order to retrieve ambiguous candidate entries with type and location information. Moncla et al. discussed the possibility of leveraging natural language spatial relationships to approximately locating non-gazetteered toponyms; however, their actual implementation only relies on the convex hull and circumscribed circle of anchors [38]. Spitz et al. leverage the network for toponym disambiguation based on computing a ranking of candidates by their co-occurrence with other toponyms mentioned in the documents [49]. However, the network must be georeferenced beforehand, which means it cannot be used to resolve new toponyms that did not appear in the network. Finally, some studies focus on developing gazetteer independent approaches. For example, language models have been used for georeferencing toponyms and documents [46, 54]. These methods typically discretize the earth into cells, and train language models to associate documents with these cells. Then, similarity scores are computed to decide which cells are best corresponding to a given test documents. DeLozier and Baldridge also developed a gazetteer independent approach that calculates the likelihood of seeing a word at a certain location, and find points of strongest overlap for a toponym and context words [11]. However, the mean and median distance errors (as well as distance uncertainties caused by the size of the cells) of these gazetteer independent methods are generally at the scale from hundreds to thousands of kilometers, which are much less useful in locating fine-grained places.

2.2 Modeling qualitative spatial relations

People use qualitative spatial relations often when describing places from memory, based on their cognitive image of the environment. Qualitative spatial relations have been extensively studied in Artificial Intelligence (e.g., see [33]), where they are formalized in logical or algebraic calculi. In English, such qualitative spatial relationships are often expressed by prepositions thus can be identified and extracted from texts.

Some studies derive uncertainty fields for spatial relations (but not always qualitative) in locative expressions referring to some known locations (e.g., "10km east of Berkeley") based on probabilistic models [35]. However, their parameters require manual configuration by the user when the models are being used. Fu et al. assign different search radii for relations such as *near* or *north* based on the size of the places for spatial querying, the distance parameters are again empirically adjusted [14]. Other studies attempt to quantify qualitative spatial relations using data-driven methods. Delboni et al. focus on determining semantic equivalence of distance relations for query expansion purposes [10]. Hall et al. quantify spatial relations in terms of distance and orientation [19]. Skoumas et al. derive probabilistic models for spatial relations and choose only major metropolises as their case study [48]. Derungs and Purves use web n-grams to model vague spatial relation concepts, and also have a strong focus on prominent places [12]. The way of deriving some of the spatial relation models in this research is essentially not so different compared to [48] and [12]; however we also aim at generalizing the models to make them scalable to different places for georeferencing purposes as well as to make them contextualized.

Other studies suggest bringing in context for interpreting qualitative spatial relations. However, most of them only remain at conceptual level, and the knowledge required in these models may not be extracted from text using existing techniques. For example, Cai proposed a framework to contextually model geospatial data considering tasks and transportation tools [6]. Wallgrün et al. propose identifying key context features affecting human usage of spatial relation expressions, in order to produce contextualized models for answering spatial queries [53]. Despite the fact that no implementation is yet provided, the goal is similar to the current research. Yao and Thill studied context explicitly by investigating how contextual features could affect the interpretation of proximity measures such as *near* and *not so far* [58]. However, most of such contextual features, e.g., familiarity with the area, financial and time budget, network connectivity, and personal characteristics, may not be extracted from text either. In this research we focus on contextual features that are derivable from place descriptions.

Despite the extensive studies on modeling qualitative spatial relations and the expected efficacy of these models in georeferencing place from text, spatial relation models are hardly leveraged for this purpose. In a recent review of current approaches for geocoding textual documents, spatial relationships other than hierarchical containment are not discussed [37]. In the approach below, we propose several models for various qualitative spatial relation families, and test their performance on georeferencing places.

2.3 Place graphs

The input of this research are preprocessed place graphs instead of raw place descriptions. Vasardani et al. regard place descriptions as place references and spatial relationships embedded in locative expressions, which can be extracted using a parser and modeled by *triplets* [51]. For example, the description "The *courtyard* is on the *campus*, beside the *clocktower*" can be modelled in the form of triplets of a *locatum L*, the reference to a place that is to be located, a *relatum R*, the reference to a place that is already located, and a spatial relationship *r* between the two: <*L*: courtyard, *r*: on, *R*: campus> and <*L*: courtyard, *r*: beside, *R*: clocktower>. A place graph $G = (V, E)$ can then be constructed from such triplets [25, 51], with the directions of the edges starting from the locata and ending with the relata. A parser being able to extract triplets from place descriptions have been developed [23,34]. Each triplet is stored as two nodes, one each for locatum and relatum, and an edge in between for the spatial relationship. The two example triplets can be used to create a simple place graph, as shown in Figure 1.

Figure 1: A simple place graph representing the spatial references "the courtyard is on the campus" and "the courtyard is beside the clocktower."

When a place graph is constructed from collective place descriptions, the challenge is to identify nodes referring to identical places. For example, if the two triplets are from two different description discourses, the graph in Figure 1 cannot be created unless the two *courtyard* references are detected to be referring to the same place. Kim et al. developed a comprehensive approach to merge different place graphs by identifying place identity

with several measurements [27]. Sufficiently similar places are stored as a single node, i.e., each node has a unique identifier and potentially multiple place references. Note that this task is different from toponym disambiguation, as node merging does not require the place references to be gazetteered, and the process will not link places to locations either.

Compared to the object and field based models for places, a place graph additionally captures the *network* dimension [28] of places by their co-occurrence and spatial relationships in descriptions. Place graphs have been used already for several tasks including creating plausible sketch maps [25] and identifying landmarks [26].

3 Methodology

This section first clarifies the input of the georeferencing approach below, the three core subtasks, and the workflow of this research. It will then guide through each subtask.

3.1 Overview

Each node in a place graph has a unique identifier and at least one, but potentially multiple place references. Between places (as conceptualized in the real world) and place references are *n:m* relationships, i.e., a place may be referred to by one or more different place references, while the same place reference may be used to describe different places in different contexts. For example, two references "Flinders Street Railway Station" (gazetteered reference) and "the train station" (non-gazetteered reference) come from conversational contexts where they refer to the same, gazetteered place (Flinders Street Railway Station). In a different context, the reference "the train station" may refer to another train station.

Figure 2 shows a sample place graph, which consists of six places represented by nodes (labeled *a, b, c, d, e, f*) as well as seven spatial relationships represented by labeled edges. A list of place references from the original place descriptions for each node is shown in the solid line rectangles. Each dashed line rectangle shows the gazetteered name(s) for these places ("-" for non-gazetteered places). The ground truth names are only shown here for demonstration purposes, and are not available from the input place graph for the below georeferencing process.

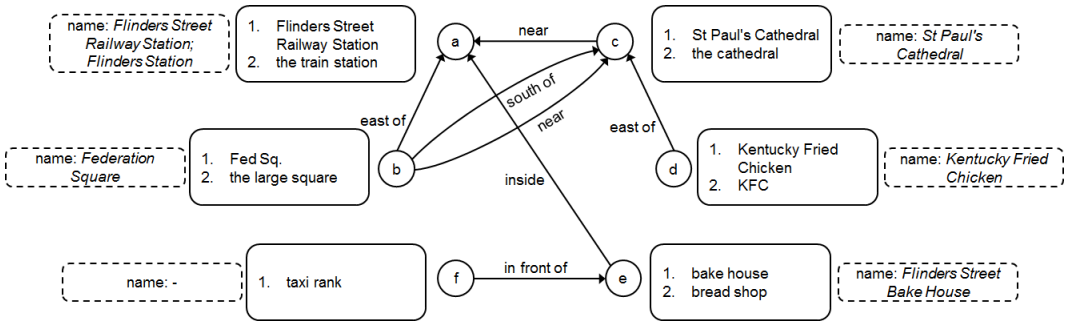

Figure 2: A sample place graph with six nodes and seven edges, each node is stored with one or more place references merged from collective place descriptions.

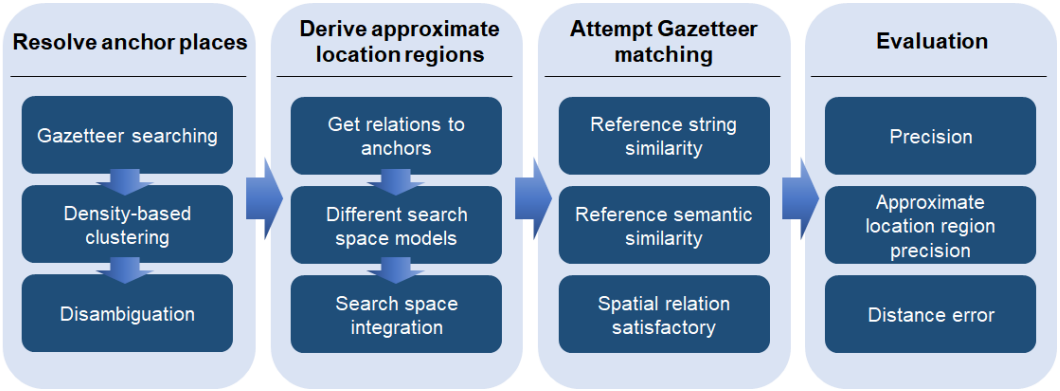

Figure 3: The workflow of this research, with the first three phases corresponding to the three major subtasks of the georeferencing approach.

Note the difference between a *non-gazetteered reference* and a *non-gazetteered place*: a non-gazetteered reference may either be a synonym referring to a gazetteered place, or a reference referring to a non-gazetteered place, while a non-gazetteered place does not have any corresponding gazetteer entries. Thus, three situations can be distinguished for nodes in a place graph (taking the sample graph as example, judged by the ground truth information):

1. A gazetteered place with at least one gazetteered reference and possibly other non-gazetteered references (synonyms) (nodes a, c, d);
2. A gazetteered place with no gazetteered references (nodes b, e); and
3. A non-gazetteered place (node f).

Such a place graph is the input of the georeferencing approach below. Additionally, a separate index file is kept, storing the information of which references are from the same original descriptions (by a description identifier). Thus, it is possible to trace back reference co-occurrence by discourse. The task of this research is to georeference every node from an input place graph. The solution provided in this research can be divided into three main phases, as shown in Figure 3.

The first phase attempts to identify and disambiguate some anchor places in an input place graph, i.e., nodes from Situation 1 (gazetteered reference). This is done by searching all place references in a gazetteer and disambiguation through a density based clustering method. In the second phase, the approximate location regions of the remaining places (Situations 2–3) are derived based on their spatial relationships to the resolved anchors. Finally, the derived regions will be used for matching with gazetteered entries within them.

3.2 Resolving anchor places

In the first phase, all place references in an input place graph are looked up using a gazetteer. If a place has at least one associated place reference that can be found in the gazetteer (i.e., Situation 1), it is regarded as an anchor place.

The next step is disambiguation, i.e., assigning each anchor place with one entry from its ambiguous candidates. We choose to use a map-based approach as it does not require additional knowledge of places. Place descriptions often contain place names of fine-grained

features, while knowledge-based disambiguation approaches developed in the literature typically focus on larger geographic features. Such approaches quickly fail when dealing with the fine-grained places. Even disambiguation approaches based on machine-learning techniques are difficult to be applied for fine-grained places due to the lack of good quality training data.

Map-based disambiguation typically relies on clustering, with the locations of all ambiguous entries for all places as the input point cloud. Among the algorithms used in the literature (e.g., [5, 18, 38]), DBSCAN seems most suitable for the task of this research when compared to other simplified heuristics. However, DBSCAN has a major disadvantage as it requires manual input parameters, particularly *Eps*—the distance threshold to put points into one cluster. This parameter was empirically adjusted in [38], which requires a-priori knowledge of the context of the collected data. While in many large-scale contexts this is given, in our context of everyday communication—collecting descriptions of potentially various conversational contexts—no a-priori value can be assumed.

Therefore, for this step we use an algorithm presented in [7]. The algorithm does not require manual input parameter values, and it is able to identify clusters with significantly large point densities that are likely to be corresponding to spatial contexts. The function shown in Equation 1 is firstly defined, which computes point density according to different distance thresholds d. Δd is for discretizing the function and can easily be set to a small distance such as $100m$. Then, the value of d where the point density first falls at two standard deviations plus the mean density after the global maximum value is selected as the cluster distance for deriving clusters.

$$K(d) = \frac{1}{\pi d^2 - \pi(d - \Delta d)^2} \times \frac{1}{n} \sum_{i=1}^{n} count(p \in region(p_i, (d - \Delta d, d])) \quad (1)$$

Figure 4 illustrates the disambiguated anchor places from the sample graph through clustering. The performance of the algorithm along with its robustness and sensitivity has been evaluated [7]. The result confirms its superiority over DBSCAN as well as other competitive clustering algorithms for the task of disambiguating place names from place descriptions, in terms of disambiguation precision and distance error. The algorithm is robust with input from various conversational contexts, this is shown by having relatively low variance in disambiguation precision and distance error. It is also able to derive clusters that are well-matched to the actual spatial contexts for these inputs.

We also introduce an additional process for further disambiguation, as it is possible that a cluster may contain more than one entry for an anchor place. In this case, these places are temporarily removed from the anchor place list, and will be georeferenced together with the remaining places in the next phase, where spatial relationships will be used for further disambiguation.

3.3 Deriving approximate location region representations

This section first introduces *search spaces* for spatial relations, either as formally defined ones or as contextualized probabilistic-based ones trained from data. *Approximate location regions* will then be introduced for integrating different search spaces in order to approximately locate the remaining places from the previous phase.

Figure 4: Disambiguated anchor places *a*, *c*, *d* from the sample graph. The dashed circle indicates approximately the spatial context of the sample graph, where the identified cluster (the three entries shown on the map) are within (*source*: Google Map, 2015).

3.3.1 Formal search spaces

The semantics of qualitative spatial relationships from four relationship families are considered, as shown in Table 1. In order to be able to use these families, a mapping of spatial prepositions (in the place graph) to the formal relationships has to be applied. We used a classification schema in a look up table, as also implemented in the literature [25]. For example, the natural language preposition *close* will be mapped to the qualitative distance relation *near*, and *W* and *western* will be mapped to the cardinal direction relation *west*.

Table 1: Spatial relationships considered for search space modeling in this research

Spatial relationship family	Spatial relationships
Cardinal direction	north, south, east, west, northeast, southeast, northwest, southwest
Qualitative distance	near
Relative direction	in front of, behind, left of, right of
Topological	inside, covered by, overlap, meet, disjoint, cover, contain, equal

A search space is defined for each relation from Table 1 to represent the constrained location of a locatum that satisfies the spatial relation to an already georeferenced anchor place (relatum). The search spaces below are defined for anchor place geometries of points, polylines, or polygons, although in most gazetteers places are represented by points.

Cardinal direction relations. The search spaces for cardinal direction relations are defined based on Frank's half-plane models for point type relatum [13], as shown in Figure 5 (a, b, c). The model can be extended to support polygon-based relata (Figure 5 (d)).However, in this research we chose to use the centroid of any polygon to derive the half-planes. The reason is that a cardinal direction preposition might express an internal relation [36], e.g., "in the north (in the northern part) of the city," which is misinterpreted by a polygon model.

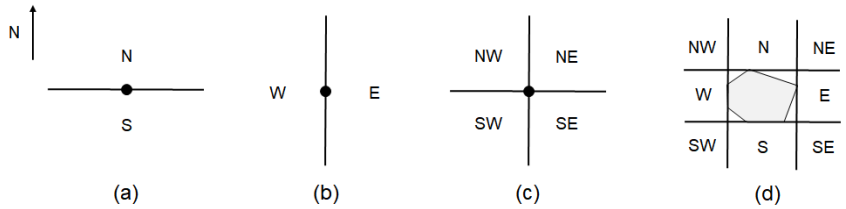

Figure 5: Search spaces for cardinal direction relations based on the centroid of the relatum (a, b, c), and an alternative model for non-point relata (d) that is not applied in this research.

Qualitative distance relations. The search spaces for qualitative distance relations are defined by buffer regions as shown in Figure 6 (a, b, c) for different relatum geometry types. Similar to the ones proposed in [35] Figure 6 (d). Buffer regions are a generally accepted model for quantifying qualitative distances in applications such as in local search applications or geographic information retrieval engines. The buffer distances, which are highly context dependent, are defined here empirically, and then adapted to the semantic context considering the size of the relata as well as to the size of the *spatial context* (which will be introduced in Section 3.3.3), as shown in Equation 2. d stands for the buffer distance, α is a constant, and β, γ are two coefficients that make d positively correlated with the area of the relatum, as well as the area of the spatial context.

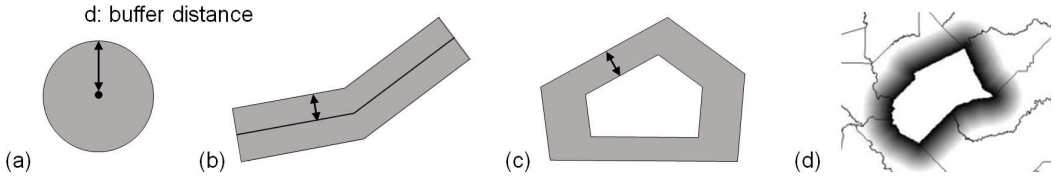

Figure 6: Search spaces for the qualitative distance relations in this research (a, b, c), and a comparison to the model by Liu et al. (d).

$$d = \alpha + \beta * \text{getArea}(relatum) + \gamma * \text{getArea}(spatialContext) \quad (2)$$

Relative direction relations. Search spaces for relative direction relations are defined based on orientation reference frames used by people, and can be either deictic, intrinsic or extrinsic [43]. Assuming that the reference frame used is known (and stored in the place

graph), the search spaces could be defined as shown in Figure 7. The arrow in the figure shows the direction of "in front of" in a known reference frame.

However, current natural language parsers are unable to infer the reference frames of prepositions automatically from place descriptions. If spatial reference frame knowledge is unavailable, the search spaces is set to be the same as *near*, as a fallback approach.

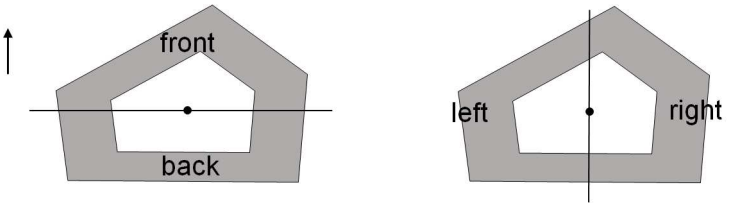

Figure 7: Search spaces for relative directions given a reference frame.

Topological relations. If the relatum is polygon-based, the search spaces for different topological relations are defined as shown in Figure 8, otherwise no search spaces will be enforced. These search spaces are used as initial filters. In the later best-matching stage, topological relations will further validated through geometry computation for excluding unsatisfactory gazetteer entries.

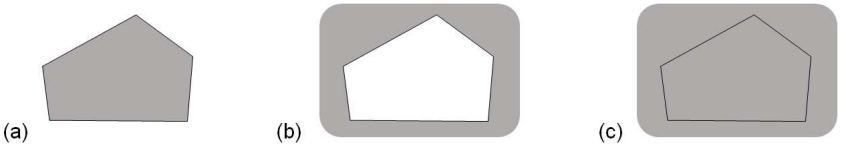

Figure 8: Search spaces for *covered by, equal, inside* (a), *disjoint, meet* (b), and the other three topological relations *overlap, cover, contain* (c).

3.3.2 Contextualized probabilistic search space models

As refinements to the formal search space models, we propose contextualized probabilistic search space models. Search spaces in this section are derived from training data and are contextualized by four factors. The values of the factors are only dependent on the input information introduced previously, and thus are automatically obtainable.

Granularity of the relatum. The semantics of a triplet's relatum can affect the interpretation of the triplet's spatial relation. For instance, "near a restaurant", "near an airport", and "near Melbourne" should be interpreted differently for defining search spaces. This factor has also been used in the formal models for qualitative distance relations introduced above. Here we group the semantic types of relata into five categories based on spatial granularity, inspired by the classification in Richter et al. [44]: *finer than building-, building-, street-, district-, or city-level* and *beyond city*. The underlying assumption is that places from the same spatial granularity level generally have similar search spaces for spatial relations.

Granularity of the locatum. Similarly, the semantics of the locatum can also affect search spaces. For example, "the building is near the CBD" and "the suburb is near the CBD" differ in their search spaces for the locatum. The theory of contrast sets by Winter and Freksa [57] offers an explanation for this. The contrast sets of the two locata are other buildings and other suburbs in the current conversational contexts, respectively. Therefore, the search spaces for suburbs will be larger than for buildings. The same five categories of place granularity are used for this factor.

Prominence of the relatum. Landmarks are cognitively salient spatial objects in terms of prominence and distinctiveness [45] and are often used to locate other, less prominent places. Thus, landmarks should ideally have larger search spaces considering their influences compared to less prominent places from the same granularity. The degree of prominence of a relatum can be measured by the frequency of references to this place in all collected place descriptions. Also, prominence can be discretised using a two-valued logic: *prominent* and *not prominent*.

Granularity of the spatial discourse. This factor is similar to the *scale effect* identified by Yao and Thill [58], introduced here as the granularity of the spatial discourse. For example, the relation *near* in the description "near Eiffel Tower" can be interpreted differently in different discourses. For example, a place description could completely be located to a limited area near the tower, or could cover the whole city of Paris. For a triplet, the granularity of the spatial discourse can be obtained by first collecting all places from the same description (which has been indexed) and selecting the coarsest granularity category (from the named five) among these places, or, if all places are of the same category, one level up. Consider the example "Richmond is near the CBD" (both are from *district-level*). In this case, it makes sense to limit the spatial discourse to city-level, since neither a suburb nor a city's centre can be larger than the city that contains both.

A combination of contextual factor values is called a *contextual criteria set* (CCS), e.g., {granularity of the relatum: *building*, granularity of the locatum: *building*, prominence of the relatum: *prominent*, granularity of the discourse: *district*}. For each of the four contextual factors, an additional value *undetermined* is defined in case a value cannot be determined. A spatial relation will have one search space derived for each possible CCS. Using this method, search spaces for relations even not included in the formal ones discussed above (e.g., *at*) can be derived.

For a given triplet with relatum as an anchor place, the following approaches are used to associate it to one of the defined CCSs. The granularity of the relatum will be determined by mapping the stored place type of the relatum in the gazetteer (which is typically from a taxonomy) to one of the six granularity categories by a dictionary. The granularity of the locatum will be determined similarly, through identifying place type keywords from the stored place references of the locatum. For example, if keywords such as "building", "park", or "city" occur, the locatum will be assigned with a granularity level accordingly. If dictionary matching fails, e.g., the granularity of the reference "the place" cannot be determined, the value of locatum granularity will be *undetermined*. Since the granularities of all places (as locata or relata embedded in triplets) can be determined, the granularity of the spatial discourse can be derived as well based on the rule defined above. For the prominence of the relatum, we used node in-degree as the number of references made to

particular places, similar to the approach proposed in [26]. To translate absolute measures of prominence (in-degrees) into relative measures of significance (*prominent* and *not prominent*), the median in-degree value is used as threshold.

In the remaining part of this section, we propose three different models for representing search spaces.

Density surface model. The first model is based on Kernel Density Estimation (KDE), a non-parametric method to estimate the probability density function of some observation data. It provides a tool to visually represent vague concepts, and each region on the generated density surface represents the relative likelihood of a new observation within it.

Figure 9 (a) shows an example search space generated through KDE for *near* given a CCS: {granularity of the relatum: *building*, granularity of the locatum: *building*, prominence of the relatum: *not prominent*, granularity of the discourse: *street*}. Assuming a set of training triplets that satisfy the CCS, for each triplet, the location of the locatum is regarded as a position vector relatively to the location of the relatum and mapped on a 2D plane. The result is a point cloud as the input of KDE. The generated density surface provides an intuitive representation of the search space of *near* of the CCS.

Regression model. The regression model aims at smoothing the density surface generated through KDE and avoid overfitting. For this purpose a Gaussian Process Regressor is used. A Gaussian Process is a generic supervised learning method designed to solve regression and probabilistic classification problems. The prediction interpolates the observations and is Gaussian probabilistic, and thus allows for deriving meaningful ALRs. Another reason for applying GPR is that, individuals use and understand spatial relation phrases differently, and thus results in multi-component distributions aggregated for search spaces. Figure 9 (b) shows the result after regression using the same data.

Tessellation-based model. As the second model, a hexagonal partition of the space is defined. Examples (with different cell sizes) using the same training data are shown in Figure 9 (c, d). The model generalizes certain details and has a reduced computational complexity compared to the KDE model when used in the next phase. Choosing different cell sizes can affect the generated search spaces, as can be seen from the two subfigures. Generally, a smaller cell size will more likely result in more dynamics (less smoothing).

3.3.3 Approximate location region

An *approximate location region* (ALR) is a derived region that represents the approximate location of a place based on all known spatial relationships to some anchors, and is computed by integrating all the search spaces of these relationships, as well as the *spatial context* by intersection. The spatial context of a place graph is defined, in this search, as a buffer region based on the minimum bounding box of the locations of all the anchor places, with a buffer distance equal to the cluster distance. A default spatial context is acting as a fallback approach to locate places if they do not have any available spatial relationship knowledge to anchor places. An example of a spatial context is illustrated by the dashed circular region shown in Figure 4.

For the sample input place graph, integrating different search spaces as formal models for deriving ALRs for nodes *b* and *e* is illustrated in Figure 10 (a). Place *b* from the sample

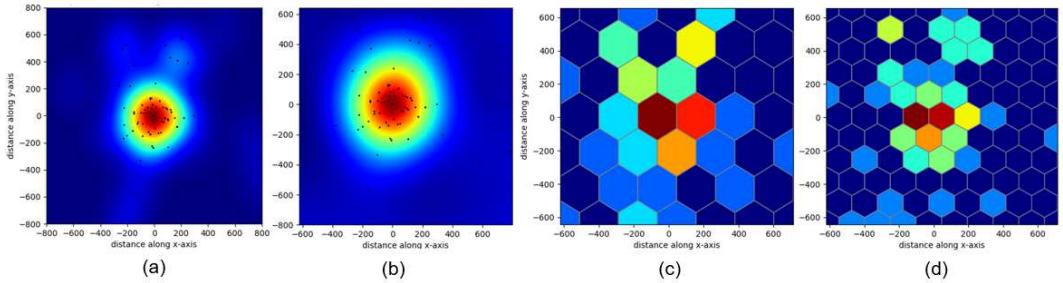

Figure 9: Example of the density surface generated by KDE (a), the density surface generated by regression (b), and the hexagon representation generated by tessellation (c, d) for *near* trained for a specified CCS, based on relative locatum locations (distance [meters]).

graph has no gazetteered references; however, three relationships are available, i.e., east of *a*, south of *c*, and near *c*. Knowing that *a* and *c* are already georeferenced in the previous phase, the location of *b* can be constrained by the shaded region representing the ALR where *b* is most likely within. Thus, gazetteer entries outside the region will not be considered for matching in the best-matching phase (following).

Integrating search spaces as probabilistic models, i.e., KDE-, tessellation- and regression-based, leads to slightly different ALRs. Given n search spaces generated by the KDE or regression models, Equation 3 presents a product operation for integration. In the equation, $s(x, y)$ stands for the value of a search space at location (x, y), and $p(x, y)$ stands for the value of the derived ALR at location (x, y). The value of $p(x, y)$ represents the relative likelihood of a place to occur at that location. For the hexagon tessellation model, $s(x, y)$ is instead computed by the number of observation points within the cell divided by the total number of points in the input point cloud. Figure 10 (b) illustrates the integration process for two search spaces generated by KDE (the blue and green contour lines) into an ALR density surface. In order to use the new search spaces in the later georeferencing process, the values in such an ALR are normalized between 0 and 1. If a crisp boundary is required (e.g., for visualization), a threshold value for membership can be selected.

$$p(x, y) = \prod_{i=1}^{n} s_i(x, y) \tag{3}$$

3.4 Gazetteer best-matching

In the last (third) phase, ALRs are used for attempting gazetteer entry matching. This is done by first collecting all gazetteer entries within the ALR of each of the remaining place, and then choosing the one that is most likely to be the actual entry the place reference addresses. Three measurements (each between 0 and 1) are considered for best matching.

Reference string similarity. Reference string similarity measures how well a place name from a candidate gazetteer entry matches a place reference in string. Existing algorithms are many, and a comprehensive comparison for toponym matching is provided in [42]. The

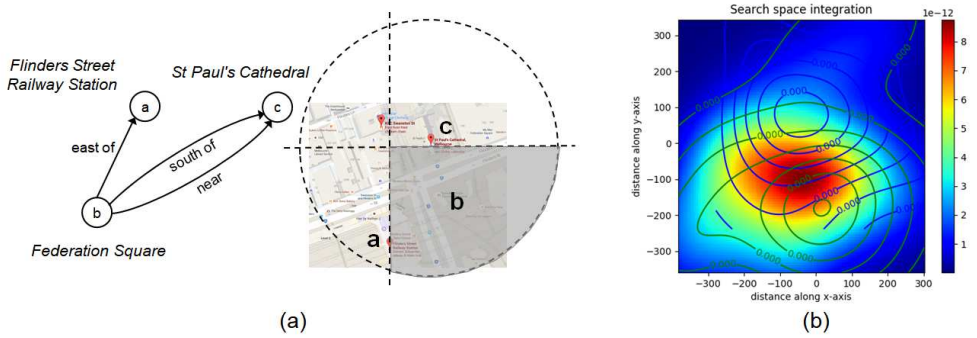

Figure 10: An example of deriving the ALR (the shaded region) for Place *b* through integrating three search spaces (*Source*: Google Maps, 2015) (a) and deriving an ALR by integrating two probabilistic search spaces into a new density surface (b) (distance [meters]).

selection of the algorithm is dataset dependent, and in this research we use the Damerau-Levenshtein algorithm [8,30]. It is a commonly used algorithm for matching tasks such as gazetteer conflation or points of interest matching in the literature. It is expected to perform well on our dataset since many place references from the dataset are short, incomplete, and vernacular.

Semantic similarity. Semantic similarity measurements have been extensively studied in communities such as information retrieval. In this research we use the Jiang-Conrath distance [21] over WordNet synsets as lexicons for measuring semantic similarity word-wise, e.g., "woods" and "forest", or "department" and "section". It is a common algorithm and similar implementations for other tasks already exist (e.g., [2]). Abbreviations (e.g., "bldg" vs. "building") are considered as having 1.0 semantic similarity. Additionally, we consider place type keywords associated with gazetteer entries as well to assist matching, if available. Taking the gazetteer of OpenStreetMap[1] for example, tagging information is stored with most entries, e.g., {name: *Peter Hall Building*; type: *building*; organization: *unimelb*; department: *Mathematics*}. The highest word-wise semantic similarity value will be returned.

Spatial relation satisfaction. Spatial relation satisfaction is for measuring how well a gazetteer entry at a certain location satisfies the given spatial relationships. For formal search space models, this is computed considering orientation, distance, and topology. For example, if two entries obtained for the place reference *the large square* for node *b* in the sample input place graph have the same name, they can only be ranked by their closeness to the anchor place *St Paul's Cathedral* given the spatial relationship *near*.

Methods for computation are shown in Figure 11. The shaded regions indicate search spaces. Nearness satisfaction is measured by the distance between the locations of the entry and the relatum, and must be between 0 (furthest) and 1 (closest). Orientation satisfaction is measured by the angle between the displacement vector starting from the relatum to the

[1]https://www.openstreetmap.org/

entry location, as well as the direction specified by the relation (1 for $0°$ and 0 for $90°$). Topology satisfaction is measured by computing the topological relation between the two places, and can be either 0 (not satisfied) or 1 (satisfied). If an entry does not satisfy a given topological relation constrain, the entry will be excluded immediately. Topological relation computation can be implemented using existing libraries with models such as DE-9IM [50].

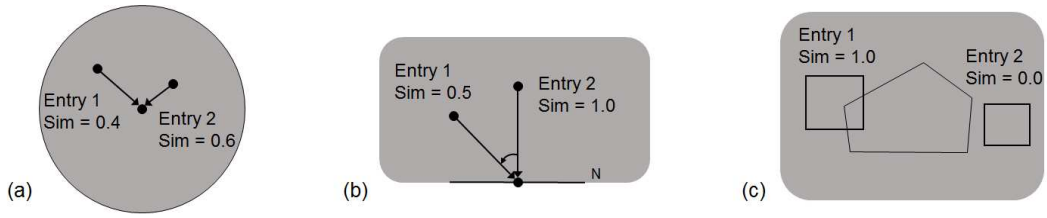

Figure 11: Illustration for spatial relation satisfaction for *near* with relatum in the middle (a), *north of* with relatum at the bottom (b), and *overlap* with relatum in the middle (c).

For contextualized probabilistic-based search spaces, an ALR represents the likelihood of a locatum being at different locations. Thus, the value for spatial relation satisfaction for a given candidate entry is simply the value of the ALR at the location of the entry. The values have been normalized and, thus, must be between 0 and 1.

Overall scoring. For each candidate gazetteer entry, the overall score is calculated by Equation. 4 in a weighted multi-attribute manner. Different weights will be tested in the implementation stage. Table 2 shows an example of calculating the overall scores for three candidate entries for node *b*. For each of the two place references *Fed Sq.* and *large square* stored with node *b*, values of the three measurements for the three candidate entries (*Ian Potter Centre*, *Federation Square*, and *Kirra Galleries*) are calculated. After overall scoring, the highlighted cell in the last column of the table, i.e., *Federation Square* with the highest score 0.7, will be used for georeferencing node *b*.

$$OverallScore = W_1 * StringSim + W_2 * SemanticSim + W_3 * SpatialSat \quad (4)$$

Place	Place reference	Candidate entry	Overall score
node *b*	Fed Sq.	Ian Potter Centre	0.37
		Federation Square	**0.70**
		Kirra Galleries	0.22
	large square	Ian Potter Centre	0.43
		Federation Square	0.63
		Kirra Galleries	0.27

Table 2: Example of best-matching for node *b* based on computed overall scores.

At the end of this phase, a score threshold is necessary to decide whether the matching process was able to find a gazetteer entry. Different threshold values will be tested in the implementation stage. A non-gazetteered place, such as node *f* from the sample graph, will then be georeferenced only by its ALR. With such a representation, the location of the

place can further be described using *anchoring theory* [15]. Thus, the place can be regarded as anchored to a location just by stating what is known with certainty and leaving the rest for further reasoning. Here, the place can be described as *anchored in* its derived ALR.

4 Implementation and experiments

The approach has been implemented using Python. The Neo4j graph database[2] is used for storing place graphs and for querying spatial relationships. This section first describes the input datasets and preprocessing procedure, and then experimental results from the three phases will be given.

4.1 Data overview and preprocessing

Two input place graphs are used for experiments. The first one contains descriptions submitted by graduate students about the University of Melbourne campus. It has richer spatial relationships among places and more focused spatial coverage compared to the second dataset. The second graph was harvested from web texts for places around and inside the area of the Greater Melbourne, Australia [24]. The sources for the Melbourne dataset include: WikiMapia as a collaborative mapping platform with user generated place descriptions; Wikipedia articles with descriptions of places; business sites or official sites with descriptions related to locations such as of companies, shops, and restaurants; and blogs with descriptions focused on individual interests such as tourism. The types of geographic features in the datasets vary from fine-grained local points of interest to large geographic features such as nature reserves. Descriptions from certain sources are more likely to include certain types of places. For instance, business sites typically focus on urban contexts, while tourism articles may be from either urban contexts or natural environment contexts. It is also observed that places from urban contexts are noticeably finer in granularity and more frequent than places as natural geographic features from the datasets. Two example descriptions from the two datasets are shown below respectively, with place references highlighted:

> "... If you go into the **Old Quad**, you will reach **a square courtyard** and at the back of **the courtyard**. You can either turn left to go to the **Arts Faculty Building**, or turn right into the **John Medley Building** and **Wilson Hall** [...] If you continue walk along **the road** on the right side where you're facing **Union House**, you can see the **Beaurepaire** and **Swimming Pool**. There will also be a **sport track** and the **University Oval** behind it ..."

> "... **St Margaret's School** is an independent, non-denominational day school with a co-educational primary school to Year 4 and for girls from Year 5 to Year 12. **The school** is located in **Berwick**, a suburb of **Melbourne** [...] In 2006, St Margaret's School Council announced its decision to establish **a brother school for St Margaret's**. **This school** opened in 2009 named **Berwick Grammar School**, currently catering for boys in Years 5 to 12 ... "

[2] https://neo4j.com/

The two input place graphs were preconstructed from triplets extracted by the parser described in [34] and have been used in previous research [25, 26]. Before the construction of these graphs, all triplets had been cleaned manually, since the parser also extracted noises such as person names, temporal expressions, and metaphors. Triplets with such noises were stripped off beforehand, as place reference extraction is outside the scope of this research.

In order to create ground truth for both training the contextualized probabilistic search spaces as well as for evaluation, the locations and granularities of all places from the two datasets are manually grounded, and anchor places are linked to their corresponding gazetteer entries. We also refer to the original descriptions for assistance. The two graphs consist of more than 5000 triplets, and proportions of different places from the three situations regarding gazetteer entries are shown in Table 3. Part of the campus graph is shown in Figure 12 as an example, illustrating the input to the following experiments.

Table 3: Proportions of places from the three situations in the two input place graphs.

Place graph	Total nodes	Anchor place	Gazetteered	Non-gazetteered
Campus	256	56 (21.9%)	143 (55.9%)	47 (18.4%)
Melbourne	3520	974 (27.8%)	1684 (47.8%)	862 (24.5%)

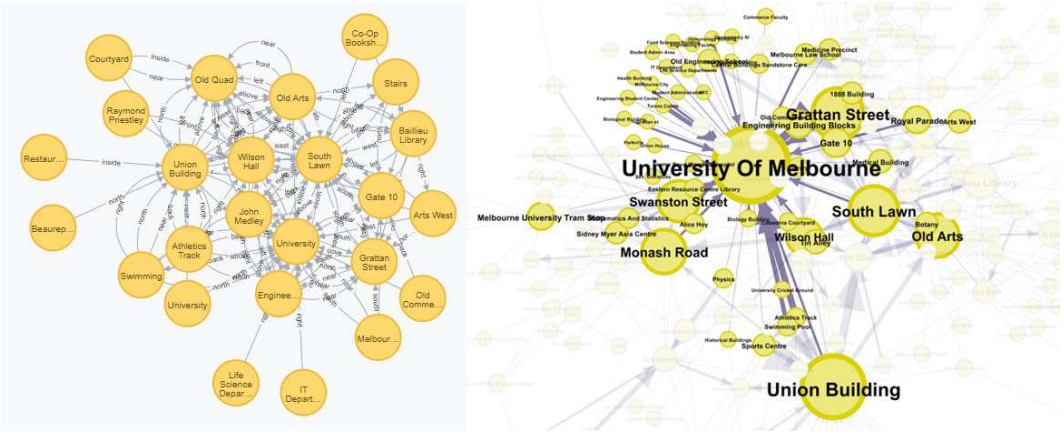

Figure 12: Part of a place graph constructed from University of Melbourne campus descriptions (left); and a different visualization with node size corresponding to node degree, and edge size corresponding to number of relationships between the linked nodes (right).

4.2 Resolving anchor places

Three gazetteers are used in conjunction for retrieving entries, aiming for completeness: OpenStreetMap Nominatim geocoder [3], GoogleV3 geocoder [4], and GeoNames [5]. The num-

[3] https://nominatim.openstreetmap.org/
[4] https://developers.google.com/maps/documentation/geocoding/intro
[5] http://www.geonames.org/

bers of ambiguous gazetteer entries retrieved for anchor places from the two input place graphs are shown in Figure 13, representing the ambiguity of references to these places. For example, the name *St Margaret's School* in the description example from the last section has a total of 11 corresponding entries from our used gazetteers.

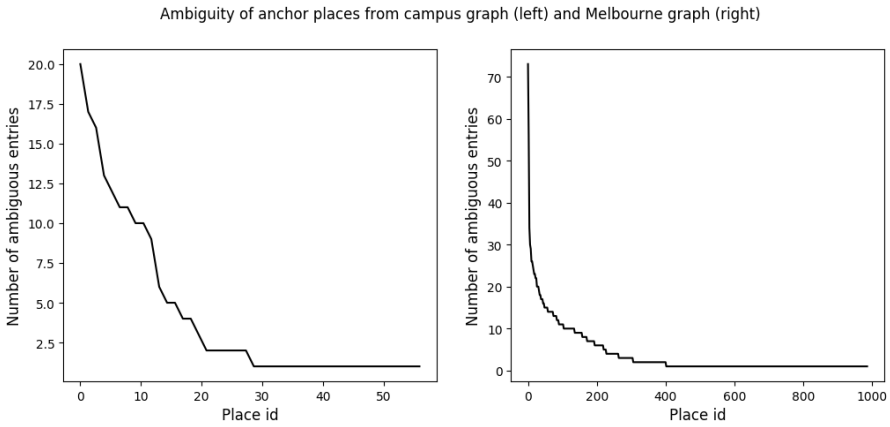

Figure 13: Numbers of ambiguous entries of anchor places from the two input place graphs.

Figure 14 illustrates the procedure from initial input point clouds to the disambiguated anchor place locations after clustering. Precisions of disambiguation are given in Table 4.

Table 4: Precisions of anchor place disambiguation.

Place graph	Campus	Melbourne
Precision	96.4% (54 out of 56)	91.9% (895 out of 974)

4.3 ALR derivation and best-matching

The input to this phase are the remaining unresolved places from the first task. This section first clarifies how contextualized probabilistic search spaces are trained, and then results will be given.

We use 10-fold cross validation for training and evaluating contextualized probabilistic-based search spaces. Specifically, triplets from the input place graphs are divided into 10-folds, and the search spaces for triplets in each fold are trained using ones from the rest based on the annotated granularity and location information. In the testing stage, the georeferencing procedure does not require further manual intervention. WordNet is used for creating dictionaries for determining the granularities of place from the test dataset.

We also tested the robustness of the training approach by reducing the amount of training data and compare the results, in order to evaluate the sensitiveness of the approach regarding the amount of training data. A comparative example is presented in Figure 15. The figure on the left side is same as Figure 9 (a), while the figure on the right side shows the search space derived based on removing 80% of random training points. The result

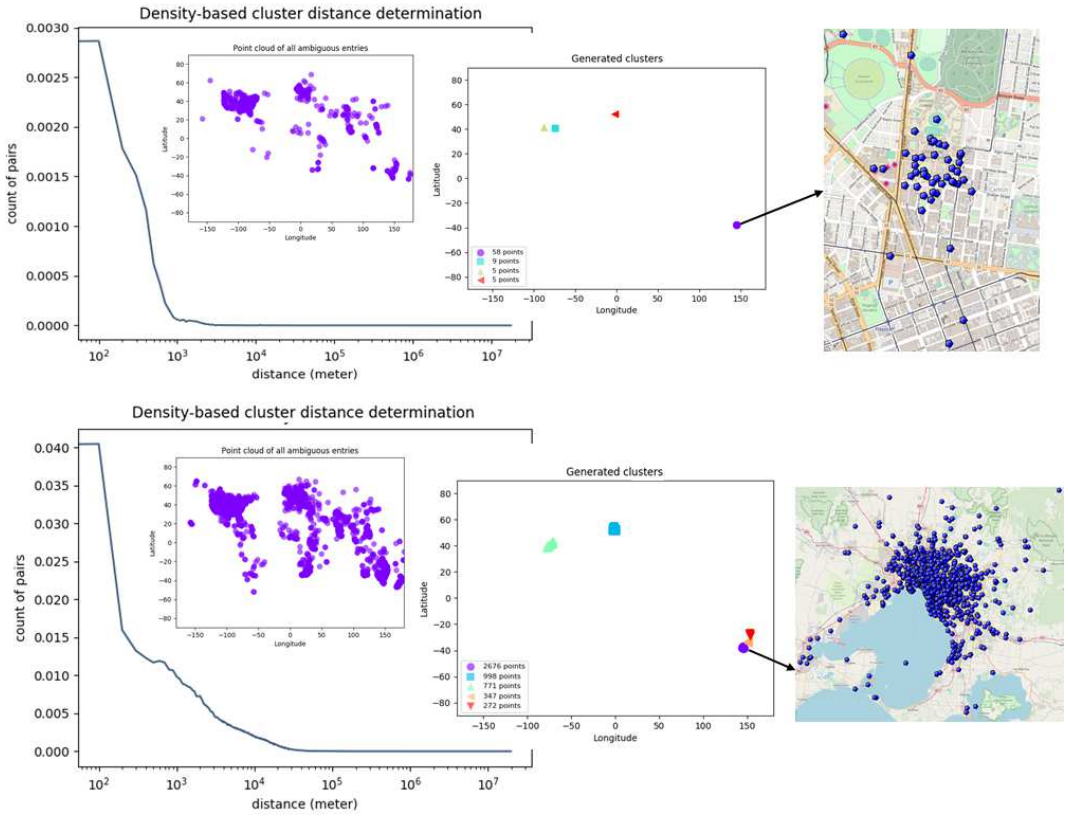

Figure 14: Deriving cluster distance for input point clouds from the campus graph (top) and the Melbourne graph (bottom); generating clusters for disambiguation (middle) from point clouds, and disambiguated anchor places forming spatial contexts (right).

indicates that even with a largely reduced amount of training data, meaningful and similar search spaces can be derived. Note that similarity is defined here by comparing to search spaces generated for other CCSs in terms of distributions by distance (c.f., Figure 16). Therefore, it is safe to say the training approach is reasonably robust and does not rely on large amounts of training data.

Figure 16 gives examples of several trained search spaces. As shown in the three figures at the top, the area of the search spaces grows as the granularity of the relata becomes coarser, with other contextual factor values preserved. The search space for *any relative direction* is derived using training triplets with any of the four relative direction relationships (*in front of*, *left*, *right*, *back*); since, for this particular case, we are more interested to explore the metric distance details through the generated search space, as there is currently no technique to automatically infer the reference frames and directions used in place descriptions.

The results of georeferencing using both formal and contextualized probabilistic-based search spaces are shown in Table 5. Three evaluation metrics are employed:

1. Precision: the percentage of places correctly linked to their gazetteer entries.

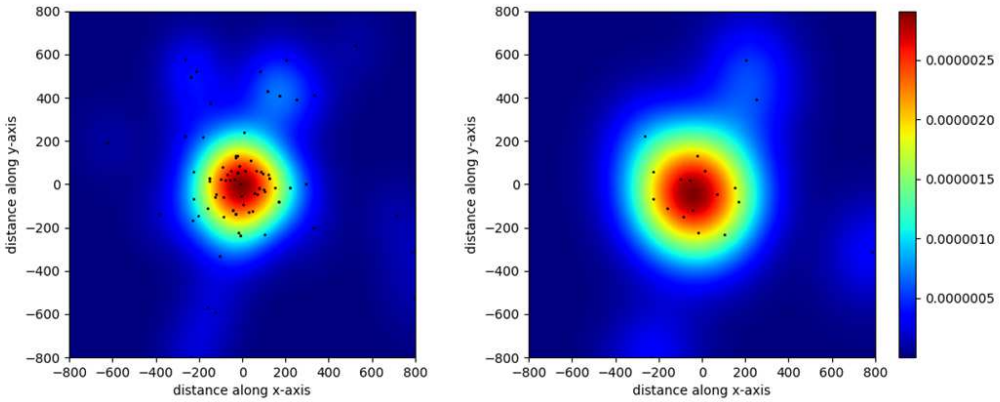

Figure 15: Comparison of KDE-based search space generated by removing 80% of random input training points (right) with the search space generated with full training points (left) for testing the robustness of the training approach (distance [m])

2. ALR precision: the percentage of places with their corresponding gazetteer entries located within their derived ALRs.
3. Mean and median distance error: the mean and median distances for all distances between the gazetteer entries matched and the ground truth ones.

Table 5: Georeferencing performance by search space models for the two input graphs

Graph	Evaluation metrics	Formal	KDE	Tessellation	Regression
Campus	Matching precision	36.3%	39.5%	39%	40.5%
	ALR precision	84.2%	93.2%	85.3%	94.7%
	Mean distance error	195m	153m	171m	144m
	Median distance error	96m	72m	88m	69m
Melbourne	Matching precision	29.4%	32.5%	30.1%	34%
	ALR precision	73.8%	92.6%	85.5%	97.5%
	Mean distance error	7203m	5863m	6752m	5217m
	Median distance error	2410m	1675m	2250m	1469m

Figure 17 (a) shows matching precision according to overall score, i.e., the precision of places matched with score equal or greater than a threshold. Figure 17 (b) and (c) shows the distance errors for individual places from the two input graphs using different models.

In order to further understand how each of the three similarity measurement influences the best-matching result, a comparison of matching precisions when applying different weights of Equation. 4 is provide in Table 6. The experiment is based on a grid search of weights with 0.1 as change interval (except for the equal weighted function as shown in the fist row). The previous results shown in Table 5 are based the best-performing weights.

Finally, we classified non-gazetteered places by testing different thresholds of best-matched scores, and the resulting precisions and recalls are shown in Figure 17 (d). Fig-

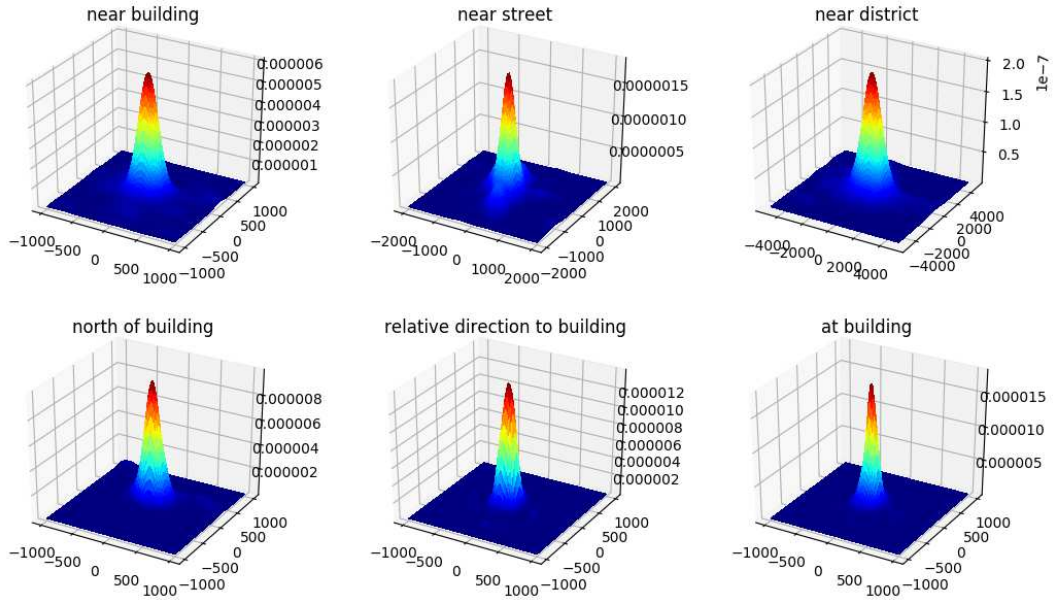

Figure 16: Search space examples for triplets with building-level locata that have certain spatial relationships to relata from different levels, with prominence and spatial discourse granularity undetermined, generated by KDE model (distance [meters]).

Table 6: Matching precisions with different weights of the overall measurement function when applying the regression model

Overall measurement function	W_1	W_2	W_3	Matching precision
Equal weights for three similarities	0.33	0.33	0.33	25.3%
Only string similarity	1.0	0.0	0.0	28.1%
Only semantic similarity	0.0	1.0	0.0	12.3%
Only spatial similarity	0.0	0.0	1.0	2.2%
String and semantic similarity	0.5	0.5	0.0	22.5%
String and spatial similarity	0.5	0.0	0.5	10.9%
Semantic and spatial similarity	0.0	0.5	0.5	6.4%
...	...	...	...	...
Best performing weights	0.5	0.3	0.2	**34.4**%

ure 18 provides an example of visualizing the approximate location of a non-gazetteered place *swimming pool* on the map by its ALR, given two relationships <swimming pool, near, University Oval> and <swimming pool, right of, Tin Alley> (the two relata are anchor places). The search spaces have been given crisp boundaries for visualization purposes using different thresholds.

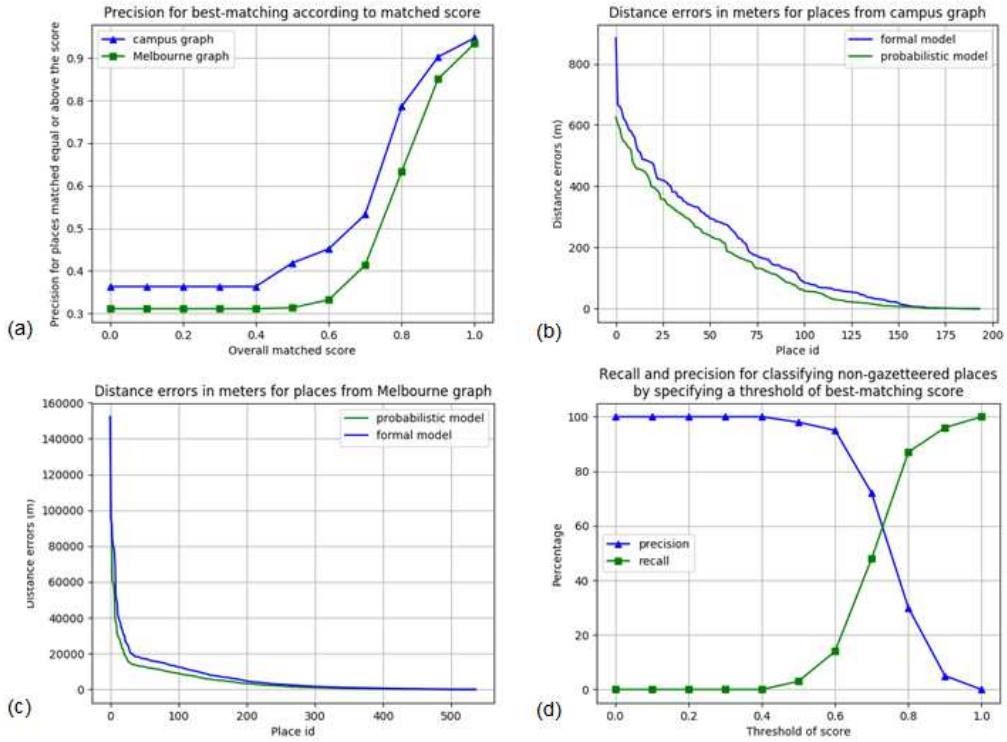

Figure 17: Precision by best-matching scores (a); distance errors by place for the two graphs (b, c) for both formal and regression-based model (distance [meters]); precision and recall trade off for identifying non-gazetteered places by thresholding (d).

5 Discussion and evaluation

Given the ambiguities of anchor places shown in Figure 13, the result from Table 4 indicates that the clustering-based approach is feasible for disambiguation. Failures are due to three reasons. First, some place references are classified as anchor places but are actually not. For example, *Gate 4* is referring to an non-gazetteered place in the University of Melbourne campus but was identified as an anchor place, since there is a gazetteer entry with the same name and it was captured by clustering. Second, references to different places may be merged incorrectly by the graph merging approach developed in [27], causing incorrect georeferencing of some references. For example, two buildings with similar references are both described to be near the same landmarks in the campus datasets, which are distinct places but are merged to the same node. In our dataset this only affects gazetteer matching (step three) of few places. Third, some anchor places are still ambiguous after the additional disambiguation process, as no sufficient spatial relations are available for further disambiguation.

Georeferencing places without gazetteered references is the main focus and motivation of this research, and the task has been largely ignored in the literature. The formal models are borrowed from fields including Artificial Intelligence and geographic information

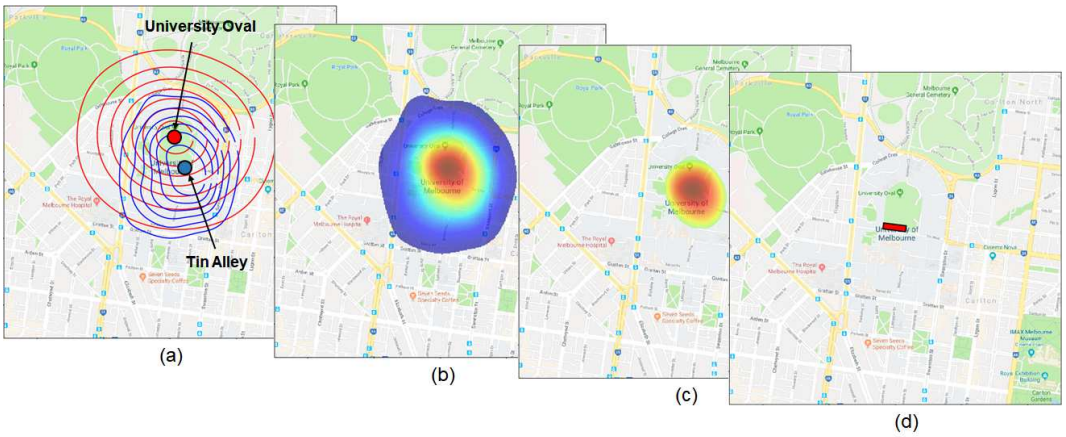

Figure 18: Example of representing the location of *swimming pool* on map, given two spatial relationships to two anchor places. Search spaces of the two relations as contours (a); crisp ALR with 0.95 as threshold (b); crisp ALR with 0.5 as threshold (c); ground truth location of the place (d) (*Source*: Google Maps, Jan 2018).

retrieval. The contextualized probabilistic based models are novel as they are generalized, contextualized, and scalable, instead of being specific to certain relata as in previous studies, such as a density field representing places near London [12, 19, 48]. We also leverage these models for the new problem (i.e., georeferencing) at hand. The approach works on a place graph constructed from a single, short description as well, e.g., "*the place* is in the University of Melbourne, near the ERC library"; although for experiments we use collective, merged descriptions in order to georeference more places by providing a more complete knowledge network.

Some examples of trained contextualized probabilistic search spaces given some CCSs are shown in Figure 16. The training process relies on manually annotated data, while the corresponding georeferencing process is purely automatic. Some interesting observations can be made during the training stage. In the sample corpora, people tend to use certain relations under specific contexts. For instance, prepositions expressing relative directions are frequently used between building-level places, but rarely for places with granularities equal to or above district level. Prepositions expressing cardinal directions, on the other hand, are used more flexibly. Prepositions expressing qualitative distance relationships, such as *near* and *at*, are generally less frequently used when referring to places with granularities larger than street level. Topological relationships are typically used to describe relationships between places of different granularity levels (mostly *inside*). We also notice that the granularity of the relatum is the most influential contextual factor on the shapes of search spaces in most of the contexts, followed by granularity of the locatum, spatial discourse granularity, and prominence. Some factors are more influential in certain contexts, e.g., the search spaces for *near* with prominent building-level places as relata are significantly larger than with less prominent ones.

The georeferencing results for places without gazetteered references are shown in Table 5. The regression model performs best for all of the four evaluation metrics for both place graphs. The results given by the KDE and the tessellation model still perform better

than the formal model. A trade off exists between (matching) precision and ALR precision. This is because larger ALRs tend to result in higher ALR precision, but at the same time are less restrictive, thus may reduce matching precision. Therefore, we also use mean and median distance errors to provide additional information about to what spatial resolution the derived ALRs are limiting the location of the places to be georeferenced. In summary, the ALR precisions show that most ALRs derived capture the location of the places to be georeferenced, while the distance errors are constraining enough considering the spatial resolution of the spatial contexts shown in Figure 14.

When comparing the formal models and contextualized probabilistic models, the increases in ALR precisions are not simply because the areas of the new search spaces are larger. In fact, search space areas for places that are finer in spatial granularity (e.g., building- and street-level) have generally decreased; yet most of them can still capture the ground truth locations of places to be georeferenced. Search spaces for places from coarser granularities, on the other hand, have generally increased and become able to capture more ground truth locations. Thus, contextualized probabilistic based search spaces are more flexible to accommodate different contexts compared to the formal ones. Additionally, they provide likelihood distribution information which is useful for location visualization, particularly for non-gazetteered places. An example is given in Figure 18.

Additional experiment results are provided in Figure 17. Figure 17 (a) shows the georeferencing precision by best-matching according to matched score. Place references matched with similarities over 0.9 are generally around 90%. Overall, places matched with higher overall similarities are more likely to be correctly georeferenced. Figure 17 (b) and (c) plots the distance error for each individual place, to assist the interpretation of the previous provided mean and median distance errors. There are some (relatively) small proportions of places with significantly larger distance errors than the other places, due to either incorrectly georeferenced anchor places (error propagation) or the lack of spatial relationship knowledge. For example, if a place has no spatial relationship available to any anchor places, its ALR will be determined loosely by the whole spatial context, which could be significantly larger compared to other ALRs constrained by spatial relationships. In addition, an input place graph may include spatial relationships that are not true, either due to the imperfection of the parser, or mistakes by the descriptors. Still, such places can be located with a reasonable distance error, when comparing several kilometers as shown in Figure 17 to the whole area of the Melbourne dataset (Melbourne has a diameter of 120 km).

For the best-matching process, there are two reasons for failures. One is that some derived ALRs are not capturing the true locations of the corresponding places. The other is because some place references are too vernacular and different from their gazetteered names thus are challenging to be linked matched. Different weights of Equation. 4 for overall similarity measurement have been tested, as shown in Table 6. The result shows that string similarity generally plays the most important role in the matching process, while spatial similarity is least important. A likely reason is that the obtained gazetteer entries for each place to be matched have already been filtered by spatial relationship search spaces, thus string and semantic similarity are more effective for further ranking these entries than spatial similarity.

6 Conclusions

Place descriptions occur in everyday communication as a way of conveying spatial information about place, and the web provides a plethora of place descriptions as texts in various forms. An important step to utilize the contained knowledge of these descriptions in an information system is to decode the spatial references, as well as to identify the places referred to, including their location. This research develops an approach for georeferencing references to places regardless of whether the references are gazetteered or not. The approach is also able to locate places that do not exist in gazetteers, based on spatial context as well as relations to landmarks. The approach has been implemented and tested with datasets collected from different sources and with different sizes, place densities, and spatial granularities.

This research starts from a place graph constructed from collective place descriptions as input, which is regarded as a knowledge base model for place references and their spatial relationships. The proposed georeferencing approach consists of three main stages. First, some anchor places are identified and disambiguated for an input graph based on gazetteer matching and a density-based clustering method. Next, for each of the remaining places, their spatial relationships to the anchor places are used for deriving representations of their approximate locations. Finally, gazetteer entries that satisfy the relational constrains for each place are used for best-matching considering string, semantic and relational similarity satisfaction, and the top ranking entry will be used for georeferencing the place. Even if the matching fails, i.e., the place is non-gazetteered, the previously derived approximate location representation can be used to visually locate the place on a map.

We used both formal spatial relationship models as well as several contextualized probabilistic based ones trained from data, for various prepositions referring to spatial relationships from four different families: cardinal directions, relative directions, qualitative distances, and topology. The contextualized probabilistic based models are determined by four factors as variables and generalized with sound scalability. We use multiple metrics to evaluate these models, and the result shows that the contextualized probabilistic models are able to accommodate flexible contexts compared to the formal models. The method performs reasonably well in terms of precisions and distance errors, considering the spatial resolutions of the graph coverages as well as the novelty of the problem compared to relevant research discussed in the literature.

We interpreted the obtained results and discussed the major observations, failure cases as well as their reasons. The major limitation of this research is the relatively small training dataset for deriving the contextualized probabilistic search spaces, particularly under certain context criteria. Although the presented models are designed to be generalized enough and only require a small number of training samples, a richer training dataset is still expected to further increase the georeferencing performance. Also, in this research we only consider contextual factors that are automatically obtainable from an input place graph, while in the literature there are other factors identified that can affect the interpretation of spatial relationships, such as traveling mode and familiarity with the environment. In the future, the contextualized probabilistic search space models proposed in this work can be further refined by these factors. In addition, there is currently no link between search spaces for different contexts, as they are trained using different data, even though the contexts may be similar. This results in difficulties in interpreting how the contextual factors affect search spaces of spatial relations.

This research presents a feasible approach to georeference places in a place graph that is constructed from collective place descriptions. The outcome has potential benefits to other research fields including geographic information retrieval which heavily relies on techniques that are able to automatically georeference places from text documents. Another application area is the automated location of callers to emergency authorities during accidents or a crisis, which can quickly fail when facing vernacular place descriptions with non-gazetteered place references and qualitative spatial relationships. The standard available geographic information systems (such as national address files) used in such situations are possibly not detailed enough for localization with regard to vernacular or granularity. Furthermore, the presented approach is able to enrich authoritative datasets, such as digital gazetteers and address databases, with people's local geographic knowledge. Finally, research about place knowledge using place graphs could help with better understanding human descriptions as input to spatial services, and thus support smoother human-computer interaction. The developed approach in this work can better capture the vagueness of locations, and can be used to build interfaces to better communicate the location of unknown locata by providing visual aids.

Acknowledgement

The support by the Australian Research Council grant DP170100109 is acknowledged.

References

[1] ADELFIO, M. D., AND SAMET, H. Structured toponym resolution using combined hierarchical place categories. In *Proc. of the 7th Workshop on Geographic Information Retrieval - GIR'13* (2013), ACM, pp. 49–56. doi:10.1145/2533888.2533931.

[2] BALLATORE, A., BERTOLOTTO, M., AND WILSON, D. C. Grounding linked open data in WordNet: The case of the OSM semantic network. In *Web and Wireless Geographical Information Systems - W2GIS'13* (2013), S. H. L. Liang, X. Wang, and C. Claramunt, Eds., vol. 7820 of *Lecture Notes in Computer Science*, Springer, pp. 1–15. doi:10.1007/978-3-642-37087-8_1.

[3] BUSCALDI, D. Approaches to disambiguating toponyms. *SIGSPATIAL Special 3*, 2 (2011), 16–19. doi:10.1145/2047296.2047300.

[4] BUSCALDI, D., AND ROSSO, P. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science 22*, 3 (2008), 301–313. doi:10.1080/13658810701626251.

[5] BUSCALDI, D., AND ROSSO, P. Map-based vs. knowledge-based toponym disambiguation. In *Proc. of the 2nd International Workshop on Geographic Information Retrieval - GIR '08* (2008), C. Jones and R. Purves, Eds., ACM, p. 19. doi:10.1145/1460007.1460011.

[6] CAI, G. Contextualization of geospatial database semantics for human-GIS interaction. *GeoInformatica 11*, 2 (2007), 217–237. doi:10.1007/s10707-006-0001-0.

[7] CHEN, H., VASARDANI, M., AND WINTER, S. Disambiguating fine-grained place names from descriptions by clustering. *arXiv preprint* (2018).

[8] DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM 7*, 3 (1964), 171–176. doi:10.1145/363958.363994.

[9] DAVIES, C., HOLT, I., GREEN, J., HARDING, J., AND DIAMOND, L. User Needs and the Implications for Modelling Place. *Proc. of the International Workshop on Computational Models of Place 8*, 3 (2008), 1–14.

[10] DELBONI, T. M., BORGES, K. A. V., LAENDER, A. H. F., AND DAVIS, C. A. Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS 11*, 3 (2007), 377–397. doi:10.1111/j.1467-9671.2007.01051.x.

[11] DELOZIER, G., BALDRIDGE, J., AND LONDON, L. Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proc. of the 29th AAAI Conference on Artificial Intelligence* (2015), AAAI Press, pp. 2382–2388. doi:10.1016/j.celrep.2016.09.063.

[12] DERUNGS, C., AND PURVES, R. S. Mining nearness relations from an n-grams Web corpus in geographical space. *Spatial Cognition and Computation 16*, 4 (2016), 301–322. doi:10.1080/13875868.2016.1246553.

[13] FRANK, A. U. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing 3*, 4 (1992), 343–371. doi:10.1016/1045-926X(92)90007-9.

[14] FU, G., JONES, C. B., AND ABDELMOTY, A. I. Ontology-based spatial query expansion in information retrieval. In *Proc. of the OTM Confederated International Conferences* (2005), R. Meersman and Z. Tari, Eds., vol. 3761, Springer, pp. 1466–1482. doi:10.1007/11575801_33.

[15] GALTON, A., AND HOOD, J. Anchoring: A New Approach to Handling Indeterminate Location in GIS. In *Proc. of the International Conference of Spatial Information Theory - COSIT'05* (2005), A. Cohn and D. Mark, Eds., vol. 3693 of *Lecture Notes in Computer Science*, Springer, pp. 1–13. doi:10.1007/11556114_1.

[16] GOODCHILD, M. F. Citizens as sensors: The world of volunteered geography. *GeoJournal 69*, 4 (2007), 211–221. doi:10.1007/s10708-007-9111-y.

[17] GOODCHILD, M. F. Formalizing Place in Geographic Information Systems. In *Communities, Neighborhoods, and Health* (2011), Springer, pp. 21–33. doi:10.1007/978-1-4419-7482-2_2.

[18] HABIB, M. B., AND VAN KEULEN, M. Improving Toponym Disambiguation by Iteratively Enhancing Certainty of Extraction. In *Proc. of the 4th International Conference on Knowledge Discovery and Information Retrieval - KDIR 2012* (Barcelona, Spain, 2012), A. L. N. Fred and J. Filipe, Eds., SciTePress, pp. 399–410.

[19] HALL, M. M., SMART, P. D., AND JONES, C. B. Interpreting spatial language in image captions. *Cognitive Processing 12*, 1 (2011), 67–94. doi:10.1007/s10339-010-0385-5.

[20] HILL, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Research and Advanced Technology for Digital Libraries* (2000), J. L. Borbinha and T. Baker, Eds., vol. 1923 of *Lecture Notes in Computer Science*, Springer, pp. 280–290. doi:10.1007/3-540-45268-0_26.

[21] JIANG, J. J., AND CONRATH, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proc. of the 10th International Conference on Research in Computational Linguistics - ROCLING'97* (1997). doi:10.1.1.269.3598.

[22] JONES, C. B., AND PURVES, R. S. Geographical information retrieval. *International Journal of Geographical Information Science 22*, 3 (2007), 219–228. doi:10.1080/13658810701626343.

[23] KHAN, A., VASARDANI, M., AND WINTER, S. Extracting Spatial Information From Place Descriptions. In *Proc. of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP'13* (2013), S. Scheider, B. Adams, K. Janowicz, M. Vasardani, and S. Winter, Eds., pp. 62–69. doi:10.1145/2534848.2534857.

[24] KIM, J., VASARDANI, M., AND WINTER, S. Harvesting large corpora for generating place graphs. In *International Workshop on Cognitive Engineering for Spatial Information Processes* (2015), vol. 12.

[25] KIM, J., VASARDANI, M., AND WINTER, S. From descriptions to depictions: A dynamic sketch map drawing strategy. *Spatial Cognition and Computation 16*, 1 (2016), 29–53. doi:10.1080/13875868.2015.1084509.

[26] KIM, J., VASARDANI, M., AND WINTER, S. Landmark Extraction from Web-Harvested Place Descriptions. *Künstliche Intelligenz 31*, 2 (2017), 151–159. doi:10.1007/s13218-016-0467-3.

[27] KIM, J., VASARDANI, M., AND WINTER, S. Similarity matching for integrating spatial information extracted from place descriptions. *International Journal of Geographical Information Science 31*, 1 (2017), 56–80. doi:10.1080/13658816.2016.1188930.

[28] KUHN, W. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science 26*, 12 (2012), 2267–2276. doi:10.1080/13658816.2012.722637.

[29] LEIDNER, J. L. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers, 2007. doi:10.1145/1328964.1328989.

[30] LEVENSHTEIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady 10*, 8 (1966), 707–710.

[31] LIEBERMAN, M. D., AND SAMET, H. Adaptive context features for toponym resolution in streaming news. In *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'12* (2012), p. 731. doi:10.1145/2348283.2348381.

[32] LIEBERMAN, M. D., SAMET, H., AND SANKARANARAYANAN, J. STEWARD: architecture of a spatio-textual search engine. In *Proc. of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems - GIS'07* (2007), H. Samet, C. Shahabi, and M. Schneider, Eds., ACM, pp. 186–193. doi:10.1145/1341012.1341045.

[33] LIGOZAT, G., AND RENZ, J. What is a qualitative calculus? A general framework. In *PRICAI: Trends in Artificial Intelligence* (2004), C. Zhang, H. W. Guesgen, and W.-K. Yeap, Eds., Springer, pp. 53–64. doi:10.1007/978-3-540-28633-2_8.

[34] LIU, F., VASARDANI, M., AND BALDWIN, T. Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. In *Proc. of the 4th International Workshop on Location and the Web - LocWeb'14* (2014), D. Ahlers, E. Wilde, and B. Martins, Eds., pp. 9–16. doi:10.1145/2663713.2664426.

[35] LIU, Y., GUO, Q. H., WIECZOREK, J., AND GOODCHILD, M. F. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science 23*, 11 (2009), 1471–1501. doi:10.1080/13658810802247114.

[36] LIU, Y., WANG, X. I. M., JIN, X., AND WU, L. On internal cardinal direction relations. In *Proc. of the International Conference on Spatial Information Theory - COSIT'05* (2005), A. G. Cohn and D. M. Mark, Eds., vol. 3693 of *Lecture Notes in Computer Science*, Springer, pp. 283–299. doi:10.1007/11556114_18.

[37] MELO, F., AND MARTINS, B. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS 21*, 1 (2017), 3–38. doi:10.1111/tgis.12212.

[38] MONCLA, L., RENTERIA-AGUALIMPIA, W., NOGUERAS-ISO, J., AND GAIO, M. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proc. of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2014), Y. Huang, M. Schneider, M. Gertz, J. Krumm, and J. Sankaranarayanan, Eds., pp. 183–192. doi:10.1145/2666310.2666386.

[39] PALACIO, D., DERUNGS, C., AND PURVES, R. Development and evaluation of a geographic information retrieval system using fine grained toponyms. *Journal of Spatial Information Science 2015*, 11 (2015), 1–29. doi:10.5311/JOSIS.2015.11.193.

[40] PURVES, R. S., CLOUGH, P., JONES, C. B., ARAMPATZIS, A., BUCHER, B., FINCH, D., FU, G., JOHO, H., SYED, A. K., VAID, S., AND OTHERS. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science 21*, 7 (2007), 717–745. doi:10.1080/13658810601169840.

[41] RAUBAL, M. Cognitive engineering for geographic information science. *Geography Compass 3*, 3 (2009), 1087–1104. doi:10.1111/j.1749-8198.2009.00224.x.

[42] RECCHIA, G., AND LOUWERSE, M. M. A Comparison of String Similarity Measures for Toponym Matching. In *Proc. of The 1st ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP'13* (2013), S. Scheider, B. Adams, K. Janowicz, M. Vasardani, and S. Winter, Eds., ACM, pp. 54–61. doi:10.1145/2534848.2534850.

[43] RETZ-SCHMIDT, G. Various views on spatial prepositions. *AI Magazine 9*, 2 (1988), 95.

[44] RICHTER, D., WINTER, S., RICHTER, K.-F., AND STIRLING, L. How people describe their place: Identifying predominant types of place descriptions. In *Proc. of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD'12* (2012), pp. 30–37. doi:10.1145/2442952.2442959.

[45] RICHTER, K.-F., AND WINTER, S. *Landmarks: GIScience for Intelligent Services*. Springer Science & Business, 2014.

[46] ROLLER, S., SPERIOSU, M., RALLAPALLI, S., WING, B., AND BALDRIDGE, J. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2012), J. Tsujii, J. Henderson, and M. Pasca, Eds., Association for Computational Linguistics, pp. 1500—-1510. doi:10.1.1.261.2865.

[47] SILVA, M. J., MARTINS, B., CHAVES, M., AFONSO, A. P., AND CARDOSO, N. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems 30*, 4 (2006), 378–399. doi:10.1016/j.compenvurbsys.2005.08.003.

[48] SKOUMAS, G., PFOSER, D., KYRILLIDIS, A., AND SELLIS, T. Location Estimation Using Crowdsourced Spatial Relations. *ACM Transactions on Spatial Algorithms and Systems 2*, 2 (2016), 1–23. doi:10.1145/2894745.

[49] SPITZ, A., GEISS, J., AND GERTZ, M. So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks. In *Proc. of the 3rd International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich'16* (2016), A. Züfle, B. Adams, and D. Wu:, Eds., ACM, pp. 1–6. doi:10.1145/2948649.2948651.

[50] STROBL, C. Dimensionally Extended Nine-Intersection Model (DE-9IM). In *Encyclopedia of GIS*, S. Shekhar, H. Xiong, and X. Zhou, Eds. Springer, 2017, pp. 470–476. doi:10.1007/978-3-319-17885-1_298.

[51] VASARDANI, M., TIMPF, S., WINTER, S., AND TOMKO, M. From descriptions to depictions: A conceptual framework. In *Proc. of the 11th International Conference of Spatial Information Theory - COSIT'13* (2013), T. Tenbrink, J. G. Stell, A. Galton, and Z. Wood, Eds., vol. 8116 LNCS of *Lecture Notes in Computer Science*, Springer, pp. 299–319. doi:10.1007/978-3-319-01790-7-17.

[52] VASARDANI, M., WINTER, S., AND RICHTER, K. F. Locating place names from place descriptions. *International Journal of Geographical Information Science 27*, 12 (2013), 2509–2532. doi:10.1080/13658816.2013.785550.

[53] WALLGRÜN, J. O., KLIPPEL, A., AND KARIMZADEH, M. Towards Contextualized Models of Spatial Relations. In *Proc. of the 9th Workshop on Geographic Information Retrieval - GIR '15* (2015), R. S. Purves and C. B. Jones, Eds., ACM, pp. 3–4. doi:10.1145/2837689.2837692.

[54] WING, B., AND BALDRIDGE, J. Hierarchical Discriminative Classification for Text-Based Geolocation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing - EMNLP'14* (2014), A. Moschitti, B. Pang, and W. Daelemans, Eds., pp. 336–348. doi:10.1177/0022427810393016.

[55] WINTER, S., BALDWIN, T., RENZ, J., TOMKO, M., AND KUHN, W. Place knowledge as a trans-disciplinary research challenge for Geographic Information Science. In *Proc. of the UCGIS Symposium* (2016), J. Mennis, Ed.

[56] WINTER, S., BENNETT, R., TRUELOVE, M., RAJABIFARD, A., DUCKHAM, M., KEALY, A., AND LEACH, J. Spatially Enabling 'Place' Information. In *Spatially Enabling Society: Research, Emerging Trends and Critical Assessment* (2010), A. Rajabifard, Ed., GSDI Association, pp. 55–68.

[57] WINTER, S., AND FREKSA, C. Approaching the notion of place by contrast. *Journal of Spatial Information Science 5*, 5 (2012), 31–50. doi:10.5311/JOSIS.2012.5.90.

[58] YAO, X., AND THILL, J. C. How far is too far? - A statistical approach to context-contingent proximity modeling. *Transactions in GIS 9*, 2 (2005), 157–178. doi:10.1111/j.1467-9671.2005.00211.x.