Deriving the Geographic Footprint of Cognitive Regions

Heidelinde Hobel, Paolo Fogliaroni and Andrew U. Frank

Abstract The characterization of *place* and its representation in current Geographic Information System (GIS) has become a prominent research topic. This paper concentrates on places that are cognitive regions, and presents a computational framework to derive the geographic footprint of these regions. The main idea is to use Natural Language Processing (NLP) tools to identify unique geographic features from User Generated Content (UGC) sources consisting of textual descriptions of places. These features are used to detect on a map an initial area that the descriptions refer to. A semantic representation of this area is extracted from a GIS and passed over to a Machine Learning (ML) algorithm that locates other areas according to semantic similarity. As a case study, we employ the proposed framework to derive the geographic footprint of the *historic center of Vienna* and validate the results by comparing the derived region against a historical map of the city.

Keywords Geographic information retrieval · Cognitive regions · User generated content · Natural language processing · Machine learning · Semantic similarity

H. Hobel (✉)
Doctoral College Environmental Informatics, Vienna University of Technology, SBA Research, Vienna, Austria
e-mail: hobel@geoinfo.tuwien.ac.at

P. Fogliaroni · A.U. Frank
Department for Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria
e-mail: fogliaroni@geoinfo.tuwien.ac.at

A.U. Frank
e-mail: frank@geoinfo.tuwien.ac.at

T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,
Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-319-33783-8_5

1 Introduction

The characterization of *place* and its representation within Geographic Information System (GISs) are becoming prominent research topics in the field of geographic information science (Gao et al. 2013; Goodchild 2011; Scheider and Janowicz 2014; Kuhn 2001). The notion of place is strictly related to people's conceptualization of space, and may correspond to different things, e.g. points of interest, geographic regions (Montello 2003), or settings (Schatzki 1991) (i.e., aggregations of spatial features).

A place is generally regarded as a region of space that is homogeneous with respect to certain criteria. We adopt the taxonomy for geographic regions proposed in Montello (2003) and focus on the category of so–called *cognitive regions*: conceptual regions derived by people as they experience the world. The geographic interpretation of cognitive regions may (and usually does) differ slightly among several individuals, as shown, for example, with the cognitive regions of downtown Santa Barbara (Montello et al. 2003), and Southern and Northern California and Alberta (Montello 2014).

In this paper we propose a novel approach to derive the geographic footprint (i.e. the location and extension) of a cognitive region from User Generated Content (UGC) sources containing textual descriptions of places. We argue that this type of UGC is a valuable knowledge base to derive an approximated geographic footprint of a cognitive region from, as it contains the conceptualizations that several people have of that region. In particular, we focus on a special type of cognitive regions: those that are conceptualized as homogeneous areas in terms of the activities they allow to be performed. To derive the geographic footprint of these regions we propose a novel framework that employs Natural Language Processing (NLP) tools to extract from textual descriptions of a place a set of named geographic features. These are used to detect on a map an initial area that the descriptions refer to, and to retrieve the activities one can perform in it. These activities provide a simplified semantic representation of the cognitive region of interest that is passed over to a Machine Learning (ML) algorithm to extend the initial area by locating other areas offering similar opportunities.

To the best of our knowledge, this is the first computational approach that exploits NLP and ML techniques based on the categorical attributes to derive an approximation of the geographic footprint of cognitive regions. As a case study we used the suggested framework to derive the geographic footprint of the cognitive region *historic center of Vienna*. Indeed, while the *historic center of Vienna* is clearly a concept that is widely referred to by people and has even a dedicated entry in Tripadvisor,[1] it is generally not retrievable from current GISs, at least at the time of writing this paper. As a preliminary evaluation we compared the derived area with a historic map of Vienna dating back to 1850, and we found that the two mostly coincide. Meanwhile,

[1] http://www.tripadvisor.at/.

we are designing a questionnaire to assess the quality of the derived region as done in Montello et al. (2003). A pilot investigation showed that the region we derived matches very well with the conceptualizations of the subjects interviewed so far. We plan to publish the final results of this more detailed evaluation in a further paper.

The remainder of this paper is structured as follows. In Sect. 2, we review related work in the fields of place representation, Geographic Information Retrieval and Natural Language Processing. We present the framework in Sect. 3 and discuss the implementation and the results for the case study in Sect. 4. Section 5 concludes the paper, also discussing limitations of the presented approach and sketching future work.

2 Related Work

In this section, we discuss related work in the fields of Place and Vague Regions and Geographic Information Retrieval and Natural Language Processing.

2.1 Place and Activities

The notion of place plays a relevant role in everyday life (Winter et al. 2009; Winter and Truelove 2013). In the field of geographic information science, different research directions have emerged which investigate the representation of places (Goodchild 2011), classify and categorize various forms of places (Schatzki 1991), and model places according to their relations to activities and affordance theory (Jordan et al. 1998). According to Schatzki (1991): "[. . .] places are defined by reference to human activity". Such a statement is supported by further research (Alazzawi et al. 2012; Montello et al. 2003; Rösler and Liebig 2013; Scheider and Janowicz 2014) implying that place semantics are closely related to activities.

In Schatzki (1991) it is argued that places organize into settings, local areas, and regions. This general notion of a hierarchical structuring of space is relatively undisputed and supported by findings of other researchers (Couclelis and Gale 1986; Freundschuh and Egenhofer 1997; Montello 1993; Richter et al. 2013). More specifically, Schatzki (1991) distinguishes two types of settings: those demarcated by barriers (e.g. apartment building), and those identified by bundles of activities that occur in them (e.g. playing in a park, shopping at a mall). Recently, the idea of equipping next-generation geographic search engines and recommendation systems with models that view places as aggregated entities has been receiving increasing attention (Ballatore 2014; Hobel et al. 2015).

2.2 Geographic Information Retrieval and Natural Language Processing

Geographic Information Retrieval (GIR) is a specialization of traditional information retrieval supported by geographic knowledge bases that enables the retrieval of geographic information and geotagged objects. The respective tools enable the identification and disambiguation of place names, the mapping of place names onto spatial features and vice versa, and the derivation of place semantics. Regarding the latter, the literature is mainly focused on the identification and classification of places (Tversky and Hemenway 1983; Smith and Mark 2001) and on the automatic generation of ontologies (Popescu et al. 2008).

To enhance the capabilities of the next generation of geographic search engines, different approaches are currently being pursued to facilitate the retrieval of georelated content. Applications range from the conceptualization of space into a metric space algebra (Adams and Raubal 2009), to the contextualization of unstructured text (Adams et al. 2015; Adams and McKenzie 2012) to relate concepts to places, to the development of content-rich knowledge bases and vocabularies (Ballatore 2015), and to semantic similarity measures for geographic terms (Ballatore et al. 2013).

Interesting approaches of automatically mapping spatial content is pursued in different fields. Jones et al. (2008) focused on modeling vague regions by statistical density surfaces and mining place descriptions in natural language to infer the approximate region. Grothe and Schaab (2009) exploited freely available georeferenced photographs to derive the geographic footprint of imprecise regions by using Kernel Density Estimation and Support Vector Machines. Cunha and Martins (2014) derived imprecise regions by exploiting machine learning for interpolating from a set of point locations. Lüschner and Weibel (2013) concentrated on using characteristics of topographical data to delineate regions.

The current focus on similarity measures for geographic terms (Ballatore 2015; Ballatore et al. 2014) is further proof that there is an interest in the disambiguation of places and place descriptions. One of the goals is to prepare shared and universally accepted vocabularies to facilitate the interpretation and the resolution of spatial requests. For instance, if the task is to search for a place where *one can get something to eat*, there are more possible matches than just restaurants. Coffee shops, pubs, or even supermarkets may also fulfill the requirements of the request.

The availability of mature Natural Language Processing (NLP) tools (Manning et al. 2014) allows for advanced processing of textual spatial descriptions (Chang et al. 2015, 2014; Chang 2014; Coyne and Sproat 2001) where tokenization and part-of-speech taggers are used to automatically break text into meaningful symbols—a selection of Part-of-Speech Tags (POST) is shown in Table 1. Two recent interesting approaches are presented in Alazzawi et al. (2012) and McKenzie et al. (2013). The former builds upon current state-of-the-art NLP to extract spatial activities from unstructured text; the latter presents a model to derive user similarity from spatial topics they discuss on social media.

Table 1 A selection of part-of-speech tags (POST) (Santorini 1990)

POST		POST	
Tag	Definition	Tag	Definition
CC	Coordinating conjunction	DT	Determiner
IN	Preposition or subordinating conjunction	WRB	Adverb
JJ	Adjective	WP	Pronoun
NN	Noun, singular or mass	TO	To
NNP	Proper noun, singular	VB	Verb

3 Deriving the Geographic Footprint of Cognitive Regions

In the following, we outline a processing workflow (see Fig. 1) to derive the geographic footprint of a given cognitive region from textual descriptions of that region. Details of the single steps involved are given in further sections.

The proposed approach relies on two types of data sources (depicted as white databases in the figure): (i) a User Generated Content (UGC) database containing

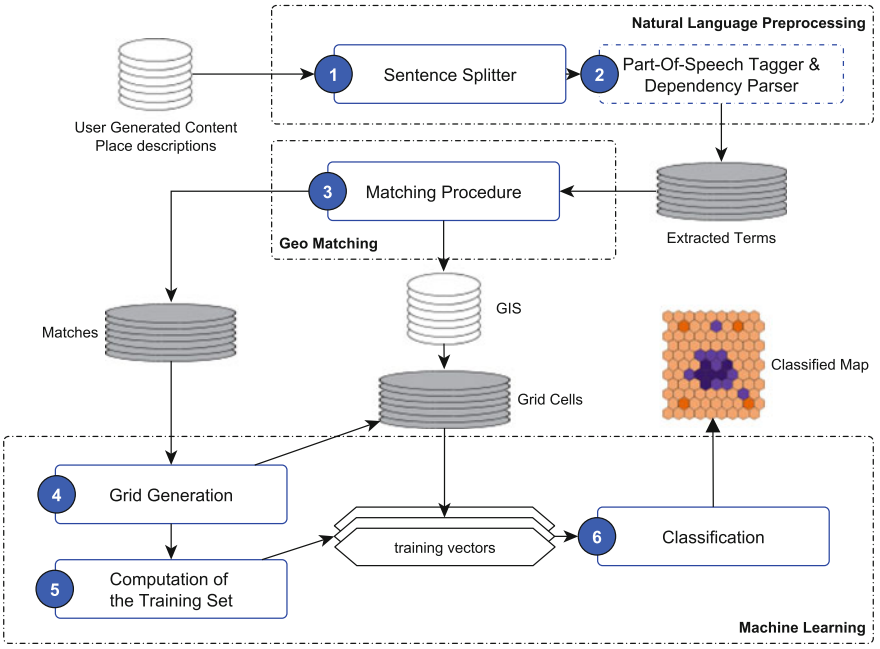

Fig. 1 Schematic illustration of the proposed workflow to derive the geographic footprint of cognitive regions

textual descriptions of a given cognitive region, and (ii) a Geographic Information System (GIS).

The workflow consists of three main stages labeled in Fig. 1 as *Natural Language Preprocessing*, *Geo Matching*, and *Machine Learning*, respectively. First, the textual descriptions undergo a natural language processing phase in order to extract from them a set of nouns referring to geographic features. In the next step, this set is compared to the geonames available in the spatial database in order to assign each a location on the map. Finally, a grid of regular cells is superimposed onto the map and the cells containing at least one of the geographic features mentioned in the textual descriptions are selected. These, together with a different set of cells selected randomly from the grid as counterexamples, are used as training samples for a machine learning algorithm that categorizes all other cells according to the activities they allow. As a result, each cell is associated to either of the two training sets, unless too little information is known about it—in which case it is marked as "unclassified".

3.1 Natural Language Preprocessing

The natural language preprocessing stage relies on the Stanford CoreNLP Natural Language Processing Toolkit (Manning et al. 2014). More specifically, it relies on three of the tools it provides: the *sentence splitter*, the *part-of-speech tagger*, and the *dependency parser*.

The sentence splitter tokenizes each UGC description into sentences (step 1 in Fig. 1) that are passed over to the part-of-speech tagger and the dependency parser (step 2 in Fig. 1). The tagger classifies every word in a sentence according to its syntactical class, e.g. noun (NN), verb (VB), adjective (JJ) (see Table 1 for a more complete list of syntactical classes and tags). The parser generates a so–called dependency tree whose nodes denote the syntactical class of each word in a sentence, with edges representing the hierarchical structure of grammatical relations between the words. For example, given the sentence "*The Karntner Strasse*[2] *is a bit touristy, but generally the area is where one could spend most of one's time in Vienna.*", the part-of-speech tagger and the dependency parser produce the tree shown in Fig. 2. Note that each term is also lemmatized, i.e. it is transformed into its base form.

Given a dependency tree, it is easy to extract from it the set $\mathcal{N}$ of common and proper nouns—tagged NN and NNP, respectively. Possibly, this set contains any reference to geographic features contained in the textual description that we are interested in locating on the map. Since the name of a geographic feature may be a compound noun (e.g. Kärntner Straße, St. Stephen's Cathedral), we need to further process the set of nouns before trying to match them with geonames available in the geographic database.

[2]The correct spelling in German language is Kärntner Straße. This comment has been retrieved from the web and is purposely reported in its original, wrongly spelled, form.

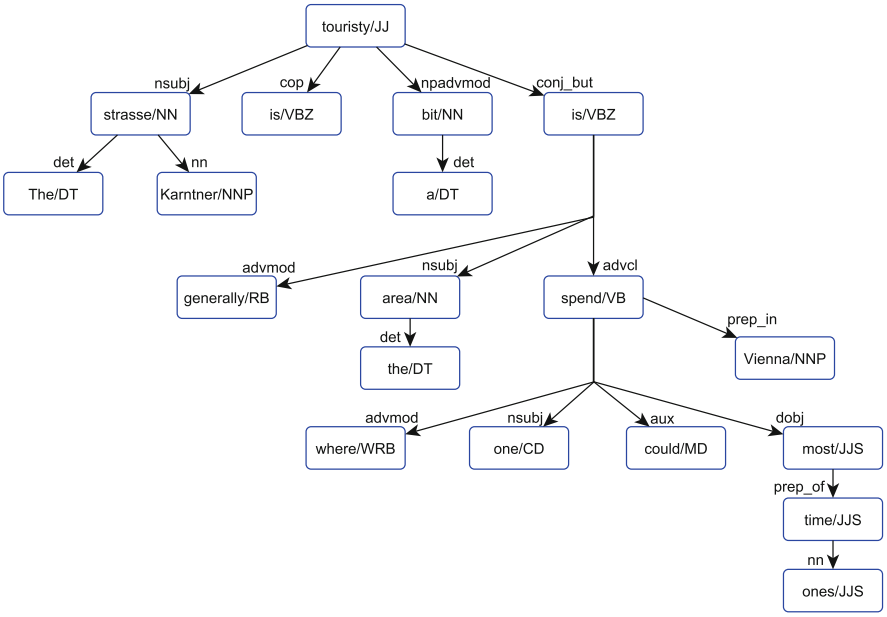

Fig. 2 The dependency tree generated by the Stanford's part-of-speech tagger and dependency parser (Manning et al. 2014) for the sentence "*The Karntner Strasse is a bit touristy, but generally the area is where one could spend most of one's time in Vienna*"

Algorithm 1 Finding candidate compound geonames

Input
$\mathcal{N} = \{nouns\ in\ UGC\ descriptions\}$,
$x = maximum\ number\ of\ words\ making\ up\ a\ compound\ geoname$
Output $\mathscr{C} = \{\mathscr{C}_n, n \in \mathcal{N}\} = \{candidate\ geonames\ for\ each\ noun\ n\ in\ \mathcal{N}\}$
1: **procedure** COMPOUNDGEONAMES
2: $\mathscr{C} \leftarrow \emptyset$
3: **for all** $n \in \mathcal{N}$ **do**
4: $\mathscr{C}_n \leftarrow \emptyset$
5: $\mathscr{D} \leftarrow \{n\} \cup RetrieveDependencies(n, x)$
6: **for all** $d \in 2^{\mathscr{D}}$ **do**
7: $\mathscr{C}_n \leftarrow \mathscr{C}_n \cup \{PermutationsOf(d)\}$
8: $\mathscr{C} \leftarrow \mathscr{C} \cup \mathscr{C}_n$

We propose the procedure reported in Algorithm 1 that, given the set of nouns $\mathcal{N}$, produces a set $\mathscr{C}$ consisting of simple and compound nouns that we refer to as *candidate geonames*. For each noun $n \in \mathcal{N}$ we access again the dependency tree to retrieve other nouns that, together with n, might make up a compound noun. This is done through the function *RetrieveDependencies(n, x)* (line 5) which, starting from the node corresponding to n, traverses the tree upwards (towards the root) and downwards (towards the leaves) and retrieves up to $x \in \mathbb{N}$ other nouns in both directions. These nouns, together with n, are stored in the set $\mathscr{D}$. The final set $\mathscr{C}_n$ of candidate

compound nouns associated to n consists of all possible permutations of the power-set of $\mathscr{D}$ (lines 6–7). The complete set of candidate geonames consists of the union (line 8) of all such sets of candidate geonames: $\mathscr{C} = \bigcup_{n \in \mathscr{N}} \mathscr{C}_n$.

In our example, from the dependency tree in Fig. 2 we derive:

$$\mathscr{N} = \{Strasse,\ Karntner,\ bit,\ area,\ time,\ Vienna\}$$

And for the noun $n = Karntner$ we have:

$$\mathscr{D}_{Karntner} = \{Karntner,\ Strasse\}$$
$$\mathscr{C}_{Karntner} = \{\emptyset, Karntner,\ Strasse,\ Karntner\ Strasse,\ Strasse\ Karntner\}$$

Note that in this case the number x of dependencies to be retrieved does not influence the sets of candidate compound names, as far as $x > 0$.

3.2 Geographic Matching

This stage does not rely on any external tool. The objective is trying to match every candidate geoname obtained in the previous stage against a unique feature in the geographic database according to name comparison (step 3 in Fig. 1). The result is a set $\mathscr{G}$ that, for each noun $n \in \mathscr{N}$, contains at most one geographic feature from the database: the one whose name best matches the candidate geonames for n (i.e., in $\mathscr{C}_n$). This implies that we also discard nouns referring to categorical features (e.g. street, square), as our final goal is to pinpoint an initial area on the map that the textual descriptions refer to.

Algorithm 2 Geographic matching

Input
$\mathscr{N} = \{nouns\ in\ UGC\ descriptions\}$,
$\mathscr{C} = \{\mathscr{C}_n, n \in \mathscr{N}\} = \{candidate\ geonames\ for\ each\ noun\ n\ in\ \mathscr{N}\}$,
$\varepsilon = threshold$
Output $\mathscr{G} = \{matched\ geonames\}$

1: **procedure** GEOMATCHING
2: $\mathscr{G} \leftarrow \emptyset$
3: **for all** $n \in \mathscr{N}$ **do**
4: $\mathscr{P} \leftarrow patternMatch(n)$
5: $(\underline{p}, \underline{d}) \leftarrow (nil, +\infty)$
6: **for all** $(c, p) \in \mathscr{C}_n \times \mathscr{P}$ **do**
7: $d \leftarrow Levenshtein(c, p.name)$
8: **if** $d \leq \varepsilon \cdot WordsIn(c) \wedge d < \underline{d}$ **then**
9: $(\underline{p}, \underline{d}) \leftarrow (p, d)$
10: **if** $\underline{p} \neq nil \wedge$ IsUnique$(\underline{p})$ **then**
11: $\mathscr{G} \leftarrow \mathscr{G} \cup \{\underline{p}\}$

We propose the procedure reported in Algorithm 2 that works as follows. For each noun $n \in \mathcal{N}$ we retrieve (line 4) from the geographic database a set $\mathcal{P}$ of features whose names pattern-match (i.e. via regex expression) against n. In defining the regex expressions, particular attention must be given to encode case-insensitivity and special characters (e.g. vowel mutations) to deal with spelling issues occurring when people write place names in a non-native language (e.g. the German word Straße vs. Strasse). Of all the retrieved features $\mathcal{P}$ we are only interested in selecting (lines 5–9) one whose name best matches against the set $\mathcal{C}_n$ of candidate geonames associated to n. For each candidate geoname $c \in \mathcal{C}_n$ and for each feature $p \in \mathcal{P}$ we compute the Weighted Levenshtein distance[3] between c and the name of p (line 7). The Weighted Levenshtein distance is a reasonable choice in case of UGC as it allows to cope with incompleteness and irregularities typical of UGC. To find possible matches (line 9) we enforce (line 8) the distance not to be bigger than a given threshold ε. Since a candidate geoname might be a compound name, we multiply ε by the number of words making up the candidate geoname. The best matching, then, is the one with the smallest Levenshtein distance. At the end of the loop the variable p is either empty or it contains a geographic feature. In the first case no match has been found. Otherwise we must make sure that the feature is unique in the geographic database (line 10). This might not be the case for features like e.g. shops or restaurants that have several branches in the same city.

Let us resume the example sentence introduced in Sect. 3.1 and whose dependency tree is shown in Fig. 2. Assume that for the noun $n = Karntner$ and for the case-insensitive regex expression "k(alaelä)rntner" the function *patternMatch* (line 4) returns only one feature named 'Kärntner Straße'. The following table then shows the resulting Levenshtein distance for each candidate geoname in $\mathcal{C}_n$ and the threshold (assumed to be $\varepsilon = 3$) multiplied by the number of words in each noun:

$c \in \mathcal{C}_n$	Levenshtein distance	$\varepsilon \cdot WordsIn(c)$
Ø	15	0
'Karntner'	8	3
'Strasse'	11	3
'Karntner Strasse'	**3**	**6**
'Strasse Karntner'	12	6

It is easy to see that there is only one entry in this table whose distance is admissible and is minimum: the entry 'Karntner Strasse'.

[3]The Levenshtein distance is a string metric that measures similarity by the minimal number of required editing steps to transform one string into the other string.

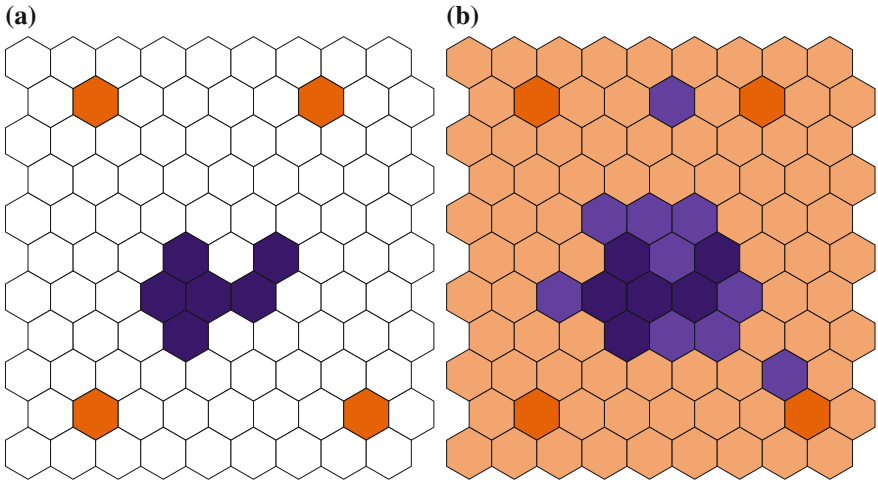

Fig. 3 A schematic representation of the classification process—given training vectors for the cognitive region of interest (*purple cells* in (**a**)) and for the counter-examples (*orange cells* in (**a**)), the machine learning classifier associates each other cell to one of the two classes (*light purple* and *light orange* in (**b**)). **a** Initial configuration. **b** Classified area

3.3 Machine Learning

This stage relies on a machine learning model called Multinomial Naïve Bayes: a probabilistic approach mainly used for text classification that learns from a given set of pre-classified samples (called training vectors) how to classify other, unclassified feature vectors according to their similarity with the given training vectors.

We adapt Multinomial Naïve Bayes to classify geographic areas as either being part of the cognitive region of interest (class 1) or not being part of it (class 2). The training vectors are obtained by tessellating the map with a regular grid (step 4 in Fig. 1) and retrieving the cells $\mathscr{I}_1$ containing at least one of the geographic features $\mathscr{G}$ derived in the previous stage. Such cells are the training vectors for the first class. The training vectors $\mathscr{I}_2$ for the second class consist of the same number of randomly selected cells that do not contain any of the geographic features in $\mathscr{G}$.

We adopt a bag-of-words model[4] to obtain a simplified 'semantical' representation of the training cells by extracting certain categorical attributes from all the geographic features contained in each such cell (step 5 in Fig. 1). Let $\mathscr{T} := \{t_i : i = 1, \ldots, n\}$ be a vocabulary containing all categorical attributes of interest from the whole map. Then, each cell is represented by a vector $(x_1, \ldots, x_n)$, where x_i is the

[4]The bag-of-words model is typically used for text classification. A text is represented as the bag (multiset) of its words and the frequency of occurrence of each word is used as a feature vector for training a classifier.

frequency of the categorical attribute t_i in this cell. Since our focus is on cognitive regions conceptualized as homogeneous areas in terms of the activities they allow to be performed, we only (so far manually based on an educated guess of the most typical activities for a cognitive region such as historic city centers[5]) select categorical attributes proper of geographic features that offer a service (e.g. bars, shops, restaurants, banks, ...).

Given the two training sets $\mathscr{I}_1$ and $\mathscr{I}_2$ as described above, the machine learning procedure is capable of classifying all the remaining cells (step 6 in Fig. 1) as graphically exemplified in Fig. 3.

4 Evaluation

This section describes an implementation of the processing workflow described in Sect. 3 and the results we obtained for the case study of the cognitive region *historic center of Vienna*.

As data sources (see Fig. 1) we selected two well-known UGC and Volunteered Geographic Information (VGI) projects: TripAdvisor[6] and OSM.[7] By means of a customized crawler we retrieved English textual descriptions of the *historic center of Vienna* from a dedicated comment page on TripAdvisor. For the geographic database we used the OSM extract of Vienna as provided by Mapzen Metro Extracts.[8] OSM provides spatial data in the form of *points* (e.g. a park bench), *ways* (e.g. streets and buildings), and *relations* (e.g. spatial entities consisting of several parts). Semantic information such as name and categorical attributes are defined as 'tags', which are key-value pairs. For example, OSM contains an entry for the "Hofburg Imperial Palace" that includes the name of the feature in several languages and is described by the following tags (among others): *(building, yes)*, *(historic, castle)*, *(castle_type, palace)*, *(tourism, attraction)*. The spatial dataset (see Fig. 4) was stored in a dedicated database where the geometry of ways and relations was simplified by their centroid.

Finally, for the implementation of the machine learning stage (see Sect. 3.3) we resorted to a hexagonal grid with uniform cells with an edge-length of 0.0025°,[9] and we used the MatLab implementation of the Multinomial Naïve Bayes[10] classifier.

[5] We are working on an extension to select activities from textual descriptions.

[6] http://www.tripadvisor.com/.

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

[8] https://mapzen.com/.

[9] The cell size can be shrinked or enlarged to obtain finer-grained or coarser results, respectively.

[10] http://de.mathworks.com/help/stats/naive-bayes-classification.html.

Fig. 4 Visualization of the OpenStreetMap (OSM) dataset of Vienna as used in our experiments—the whole dataset consists of 290,586 nodes, 368,112 ways, and 810,145 relations, for a total of 1,468,843 features

4.1 Experimental Results

We ran our workflow implementation on two experimental scenarios. Both scenarios use the same data sources with the following difference: in the first scenario (see Fig. 5), the training vectors have been kept in their integrity. In the second scenario (see Fig. 6), we manually removed outlier cells from the training vector associated to the cognitive region—i.e., those cells that fall far away from the actual city center (compare the distribution of dark purple cells in Figs. 5 and 6).

For the pictorial representations of the results we adopted the following color scheme: dark purple cells represent training vectors for the cognitive region *historic center of Vienna* as extracted from the textual descriptions; dark orange cells represent training vectors for the counter-example. Light purple and light orange cells show the areas classified as *historic center of Vienna* and counter-example, respectively. White cells denote areas that have not been classified because of insufficient semantic information.

Since counter-examples are selected randomly from the grid we decided to perform several runs for each scenario. Figure 5 shows the results obtained for five runs on the first scenario. Figure 6 shows similar results for the second scenario, where outlier cells were removed from the training vector of the cognitive region. The results for the two scenarios mostly coincide, and the cells classified as similar to the cognitive region *historic center of Vienna* form a region approximately corresponding to the central district of the city and its immediate surroundings. Interestingly,

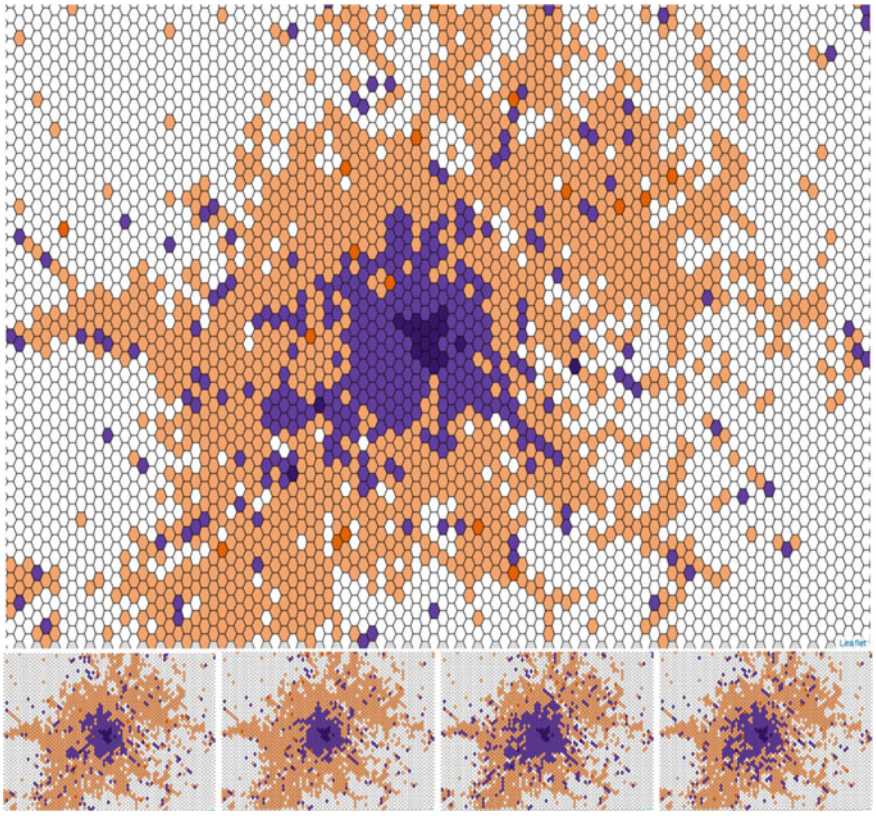

Fig. 5 Visualization of classification results for the first scenario (several runs)—*dark purple cells* represent training vectors for the cognitive region *historic center of Vienna*; *light purple cells* are classified as *historic center of Vienna*; *dark orange cells* represent training vectors for the counter-example; *light orange cells* are classified as counter-example; *white cells* are unclassified

the cells that were manually removed in the second scenario are associated to the class corresponding to the cognitive region anyway.

To mitigate the effects of using randomly selected counter-examples, we performed ten runs for each scenario and intersected the results to obtain 'robust' results: only cells classified as *historic center of Vienna* that occur in the result of each run form the robust results, as shown in Fig. 7.

4.2 Preliminary Evaluation

A sound evaluation of the results would require investigating how the derived cognitive regions fit to human conceptualization. To that end we are currently in the process of designing a questionnaire similar to that used in Montello et al. (2003).

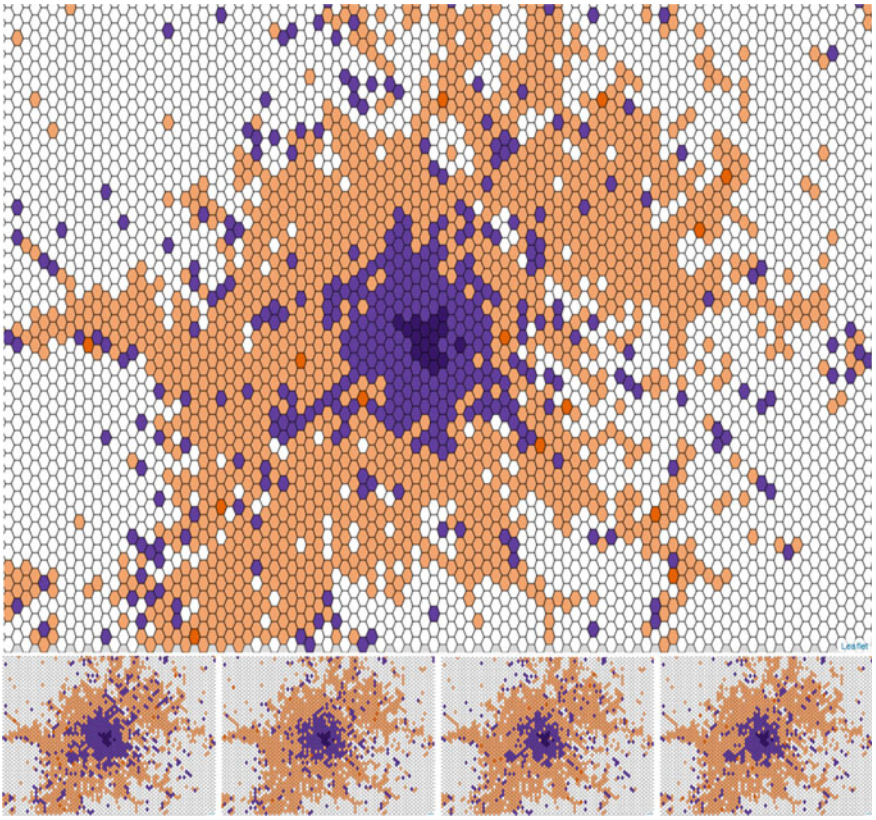

Fig. 6 Visualization of classification results for the second scenario (several runs)—*dark purple cells* represent training vectors for the cognitive region *historic center of Vienna*; *light purple cells* are classified as *historic center of Vienna*; *dark orange cells* represent training vectors for the counter-example; *light orange cells* are classified as counter-example; *white cells* are unclassified

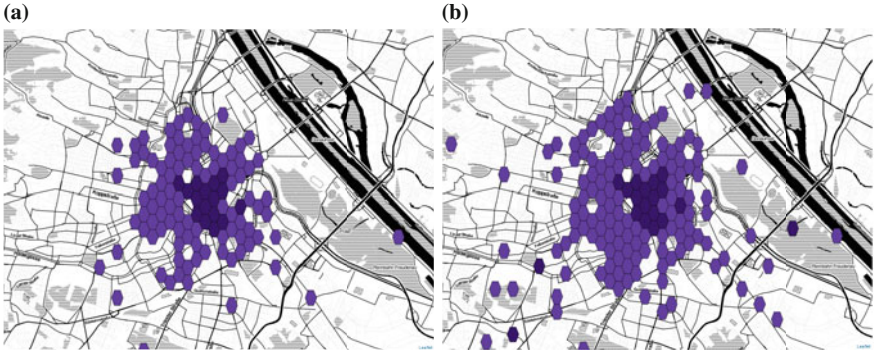

Fig. 7 Visualization of robust results. **a** Scenario 1. **b** Scenario 2

A pilot study with a small user group already showed that the footprint derived for the cognitive region *historic center of Vienna* fits well to human conceptualization. We plan to publish the final results of this study in a future publication.

Meanwhile, we present here a preliminary qualitative evaluation of the outcomes by comparing the obtained robust results with a historical map of the city of Vienna that dates back to 1850. For that, we geographically overlaid the derived regions with the map, as shown in Fig. 8 for the first scenario. It is easy to see how the shape and extent of the derived region nicely fit with the city boundaries of 1850: The outer boundary of the main part of the classified area coincides with a physical separation which is now a major street of the city, while the few outlier cells correspond to historical sites that are not reported in the historical map (e.g. the Schönbrunn Palace).

In summary, the approach, which relied solely on a knowledge base derived from VGI and crowdsourced information sources, shows promising results.

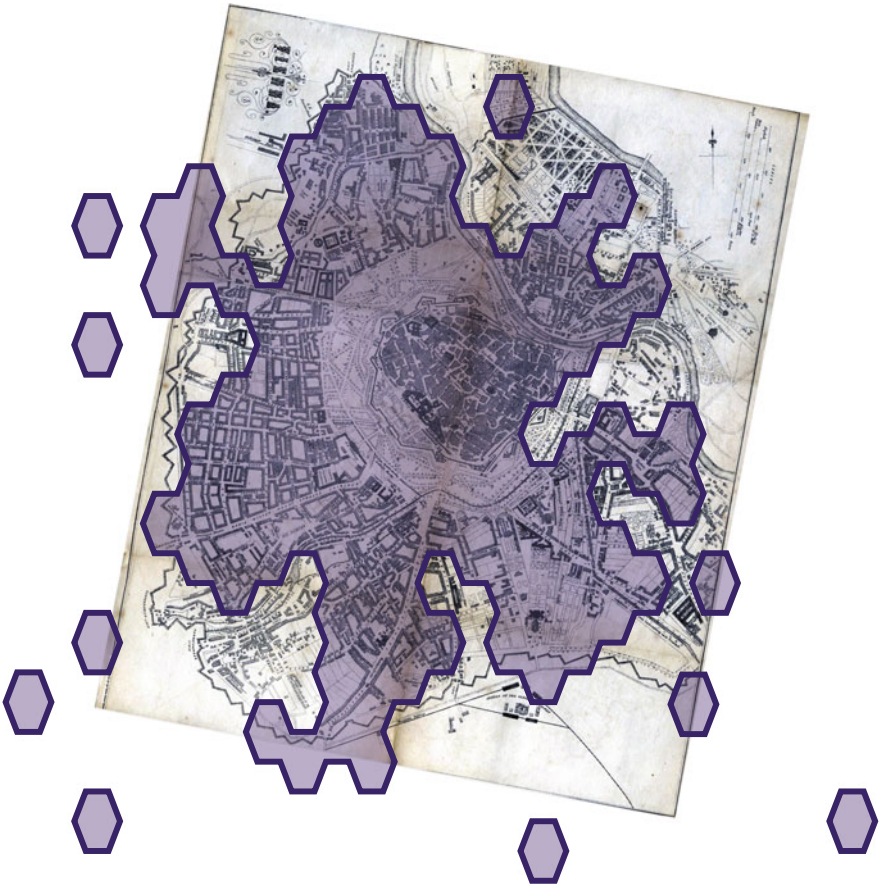

Fig. 8 Approximate overlay of the robust result (Scenario 1) over a historic representation of Vienna (map retrieved from http://www.valentina.net)

5 Conclusion and Future Work

We presented a novel automated approach to derive the geometric extent of "cognitive regions" by utilizing solely crowd-sourced geographic information as the fundamental knowledge bases. Based on Natural Language Preprocessing and a combinatorial place matching procedure tailored to identify unique geonames, the conceptualization of regions perceived as a whole based on the activities they allow is translated into a machine-processable form.

The proposed approach builds upon a representation of semantic attributes of geographic features, and allows for the automated clustering of cities into "cognitive regions". We pointed out that the classification problem can be efficiently solved by utilizing the Multinomial Naïve Bayes model as classifier. For that, a bi-classification approach was discussed that operates on initial seeding cells identified by the combinatorial place matching procedure. Counter-examples are derived using a Monte Carlo approach.

While the method presented in this paper reveals promising results, it presents a number of limitations that we plan to overcome in future work.

First, our approach relies on uniquely identifiable places to derive the initial cells for the machine learning model, while non-unique features (like shops with several branches) are completely discarded and not used to create the training vectors. Different approaches can be devised to also exploit such non-unique features, according to wether they are mentioned in a comment together with uniquely identifiable features or not. In the first case, one approach would be to use the dependency tree to locate the syntactically closest unique feature mentioned in the text. This could be used as a reference point to locate on the map the spatially closest feature matching the non-unique reference. In the second case, a solution would be to run a two-step geomatching. In the first step only uniquely identifiable features are used (as done in the current approach) to generate a starting set of training cells. In the second step this initial set is recursively extended by disambiguating non-unique features according to their vicinity to the training cells.

It could be argued that the random selection of counter-examples for the machine learning model can be improved by applying sophisticated methods. For example, one could derive an ontology of cognitive regions and select counter-examples from those that are semantically furthest away from the region of interest.

We adopted a bag-of-word model as a semantical approximation of the training cells. This is a rather coarse semantical representation, as it only accounts for the frequency of categorical attributes in a given cell. An improvement would be to resort to a model that also takes into consideration the ontological relations among the attributes as well as their spatial distribution and configuration.

Finally, we are working on a further extension of this approach that also exploits verbs and other syntactical classes to derive the activities that can be carried out at a given place. This extension may allow for a variety of more advanced applications such as the enrichment and/or validation of semantic attributes in geographic

databases, as well as enabling natural language interfaces for Geographic Information Retrieval (GIR) systems.

Acknowledgments We acknowledge the work of © OpenStreetMap contributors (http://www.openstreetmap.org/copyright), and Leaflet (http://leafletjs.com). This research was partially funded by the Vienna University of Technology through the Doctoral College Environmental Informatics.

References

Adams B, McKenzie G (2012) Frankenplace: an application for similarity-based place search. ICWSM

Adams B, McKenzie G, Gahegan M (2015) Frankenplace: interactive thematic mapping for ad hoc exploratory search. In: Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 12–22

Adams B, Raubal M (2009) A metric conceptual space algebra. In: Spatial information theory. Springer, pp 51–68

Alazzawi AN, Abdelmoty AI, Jones CB (2012) What can i do there? Towards the automatic discovery of place-related services and activities. Int J Geogr Inf Sci 26(2):345–364

Ballatore A (2014) The search for places as emergent aggregates

Ballatore A, Bertolotto M, Wilson DC (2015) A structural-lexical measure of semantic similarity for geo-knowledge graphs. ISPRS Int J GeoInf 4(2):471–492

Ballatore A, Wilson DC, Bertolotto M (2013) Computing the semantic similarity of geographic terms using volunteered lexical definitions. Int J Geogr Inf Sci 27(10):2099–2118

Ballatore A, Bertolotto M, Wilson DC (2014) An evaluative baseline for geo-semantic relatedness and similarity. Geoinformatica 18(4):747–767

Chang AX, Savva M, Manning CD (2014) Interactive learning of spatial knowledge for text to 3d scene generation. Sponsor: Idibon, p 14

Chang AX, Savva M, Manning CD (2014) Learning spatial knowledge for text to 3d scene generation. EMNLP

Chang A, Monroe W, Savva M, Potts C, Manning CD (2015) Text to 3d scene generation with rich lexical grounding. arXiv:1505.06289

Couclelis H, Gale N (1986) Space and spaces. Geografiska annaler. Series B. Hum Geogr 68(1):1–12

Coyne B, Sproat R (2001) Wordseye: an automatic text-to-scene conversion system. In: Proceedings of the 28th annual conference on computer graphics and interactive techniques. ACM, pp 487–496

Cunha E, Martins B (2014) Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. Int J Geogr Inf Sci 28(11):2220–2241

Freundschuh SM, Egenhofer MJ (1997) Human conceptions of spaces: implications for gis. Trans GIS 2(4):361–375

Gao S, Janowicz K, McKenzie G, Li L (2013) Towards platial joins and buffers in place-based gis. In: Proceedings of the 1st ACM SIGSPATIAL international workshop on computational models of place (COMP'2013), pp 1–8

Goodchild MF (2011) Formalizing place in geographic information systems. In: Communities, neighborhoods, and health. Springer, pp 21–33

Grothe C, Schaab J (2009) Automated footprint generation from geotags with kernel density estimation and support vector machines. Spat Cogn Comput 9(3):195–211

Hobel H, Abdalla A, Fogliaroni P, Frank AU (2015) A semantic region growing algorithm: extraction of urban settings. In: AGILE 2015. Springer, pp 19–33

Jones CB, Purves RS, Clough PD, Joho H (2008) Modelling vague places with knowledge from the web. Int J Geogr Inf Sci 22(10):1045–1065
Jordan T, Raubal M, Gartrell B, Egenhofer M (1998) An affordance-based model of place in gis. In: 8th international symposium on spatial data handling, SDH, vol 98, pp 98–109
Kuhn W (2001) Ontologies in support of activities in geographical space. Int J Geogr Inf Sci 15(7):613–631
LÃscher P, Weibel R (2013) Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. Comput Environ Urban Syst 37:18–34
Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
McKenzie G, Adams B, Janowicz K (2013) A thematic approach to user similarity built on geosocial check-ins. In: Geographic information science at the heart of Europe. Springer, pp 39–53
Montello DR (1993) Scale and multiple psychologies of space. In: Spatial information theory a theoretical basis for GIS. Springer, pp 312–321
Montello DR (2003) Regions in geography: process and content. In: Foundations of geographic information science, pp 173–189
Montello DR, Goodchild MF, Gottsegen J, Fohl P (2003) Where's downtown? Behavioral methods for determining referents of vague spatial queries. Spat Cogn Comput 3(2–3):185–204
Montello DR, Friedman A, Phillips DW (2014) Vague cognitive regions in geography and geographic information science. Int J Geogr Inf Sci 28(9):1802–1820
Popescu A, Grefenstette G, Mo ëllic PA (2008) Gazetiki: automatic creation of a geographical gazetteer. In: Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries. ACM, pp 85–93
Richter D, Vasardani M, Stirlng L, Richter K-F, Winter S (2013) Zooming in-zooming out hierarchies in place descriptions. In: Krisp JM (ed) Progress in location-based services. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 339–355
Rösler R, Liebig T (2013) Using data from location based social networks for urban activity clustering. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) Geographic information science at the heart of Europe, Lecture notes in geoinformation and cartography. Springer, pp 55–72
Santorini B (1990) Part-of-speech tagging guidelines for the penn treebank project (3rd revision)
Schatzki TR (1991) Spatial ontology and explanation. Ann Assoc Am Geogr 81(4):650–670
Scheider S, Janowicz K (2014) Place reference systems: a constructive activity model of reference to places. Appl Ontol 9(2):97–127
Smith B, Mark DM (2001) Geographical categories: an ontological investigation. Int J Geogr Inf Sci 15(7):591–612
Tversky B, Hemenway K (1983) Categories of environmental scenes. Cogn Psychol 15(1):121–149
Winter S, Truelove M (2013) Talking about place where it matters. In: Raubal M, Mark DM, Frank AU (eds) Cognitive and linguistic aspects of geographic space. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 121–139
Winter S, Kuhn W, Krüger A (2009) Guest editorial: does place have a place in geographic information science? Spat Cogn Comput 9(3):171–173