Taylor & Francis
Taylor & Francis Group

Computing the semantic similarity of geographic terms using volunteered lexical definitions

Andrea Ballatore[a]*, David C. Wilson[b] and Michela Bertolotto[a]

[a]*School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland;* [b]*Department of Software and Information Systems, University of North Carolina, University City Bouleward, Charlotte, NC, USA*

(*Received 16 November 2012; accepted 24 March 2013*)

Volunteered geographic information (VGI) is generated by heterogenous 'information communities' that co-operate to produce reusable units of geographic knowledge. A consensual lexicon is a key factor to enable this open production model. Lexical definitions help demarcate the boundaries of terms, forming a thin semantic ground on which knowledge can travel. In VGI, lexical definitions often appear to be inconsistent, circular, noisy and highly idiosyncratic. Computing the semantic similarity of these 'volunteered lexical definitions' has a wide range of applications in GIScience, including information retrieval, data mining and information integration. This article describes a knowledge-based approach to quantify the semantic similarity of lexical definitions. Grounded in the recursive intuition that similar terms are described using similar terms, the approach relies on paraphrase-detection techniques and the lexical database WordNet. The cognitive plausibility of the approach is evaluated in the context of the OpenStreetMap (OSM) Semantic Network, obtaining high correlation with human judgements. Guidelines are provided for the practical usage of the approach.

Keywords: lexical definitions; semantic similarity; volunteered geographic information; crowdsourcing; geo-semantics; WordNet; OpenStreetMap

1. Introduction

Geographic knowledge is a crucial asset in human activities. To share geographic information across a community, it is necessary to extract concepts from the chaotic repository of implicit knowledge that lies in human minds, which Sowa (2006) aptly called the 'knowledge soup'. The first step in gathering and sharing geographic knowledge consists of describing terms in natural language, focussing on semantic aspects relevant to the target community. In this sense, lexical definitions constitute the core of any geographic lexicon, storing informational wealth about the properties and the meaning of each term.

Over the past decades, a variety of models have been developed to represent and share geographic information. Vector and raster formats, the entity-relationship model and, to a lesser extent, the object-oriented paradigm have been adopted in geographic information systems (GISs) (Goodchild et al. 2007). Semantic Web technologies attempt to provide a platform to share geographic knowledge using logic formalisms and ontologies (Ashish and Sheth 2011). Despite these attempts, geographic knowledge is highly fragmented

*Corresponding author. Email: andrea.ballatore@ucd.ie

across different 'information communities', i.e. groups of individuals whose interactions depend on the successful interpretation of shared information.

Lexical definitions play a pivotal role across these diverse formats and models. The descriptive definitions found in language dictionaries have the purpose of accounting for accepted usages of a term within a linguistic community. By contrast, lexical definitions in a GIS lexicon or in a geo-ontology are prescriptive, i.e. indicate the intended usage of a term within a relevant scope, such as an organisation, an information system or a specific data set. For example, the definition of the term 'field' in a land registration GIS may specify that it should be used to label discrete objects delimiting a currently active agricultural field, whilst in a scientific GIS, it may be used to represent sampled intensities of the Earth's magnetic field. Lexical definitions are an effective tool to delimit the semantics of terms in a lexicon with normative effects, narrowing the semantic gap between the different conceptualisations that human agents construct in their minds.

Volunteered geographic information (VGI) is a key development in the processes of production and consumption of spatial knowledge (Goodchild 2007). In the context of VGI crowdsourcing projects, the role of lexical definitions is of paramount importance in order to allow the collaborative production of information by diverse and heterogenous human semantic agents. In online projects such as OpenStreetMap (OSM) and Ushuaidi,[1] contributors collect and share large amounts of geographic information, relying on Web 2.0 tools. Unlike professional information communities commonly found in private companies and public institutions, amateur online communities do not share similar formal training and do not have the possibility of creating a solid common semantic ground through years of intense inter-personal communication and collaborative work.

In VGI, short lexical definitions are often the only resource utilised to decide what terms describe a new piece of geographic knowledge to be shared. For example, the semantic model underlying OSM is an open tagging system with a recommended set of terms called 'tags'. Users who want to add features to the OSM vector data set can use editors such as Java OpenStreetMap (JOSM) and Potlatch, which offer a catalogue of core terms.[2] However, especially for users unfamiliar with the editing process and for creating rare or ambiguous features, lexical definitions are necessary to dispel doubts and constrain the terms' interpretation to a shared meaning. The usage of a new, undocumented term, e.g. *site* = *carpark* instead of the documented *amenity* = *parking*, can result in semantic ambiguity and possibly uninterpretable data.

Semantic similarity is a specific type of semantic relatedness, typically involving hyponym–hypernym relations between terms. For example, 'river' and 'stream' are semantically similar, while 'river' and 'boat' are dissimilar but semantically related. In the context of ever-growing online geographic information, computing the similarity of terms is a valuable endeavour in order to increase co-operation within and across information communities, providing support for a wide variety of tasks in information retrieval and integration (Janowicz et al. 2011, Kuhn 2013).

To the best of our knowledge, no geo-semantic similarity measure focussed on lexical definitions has been devised for VGI, which normally lacks advanced ontology engineering and design. To fill this lacuna, this article presents a technique to compute the semantic similarity of geographic terms, based on their lexical definitions, drawing from natural language processing techniques. The lexical database WordNet is utilised as a source of term-to-term semantic similarity and is then combined into a definition-to-definition similarity.

The remainder of this article is organised as follows: Section 2 reviews related work in the areas of VGI, OSM and semantic similarity measures. The salient aspects of

volunteered lexical definitions are subsequently discussed in Section 3, with particular focus on OSM. Section 4 outlines our approach to computing the semantic similarity of geographic terms and Section 5 presents an empirical evaluation. Finally, Section 6 draws conclusions from this work and suggests directions for future research.

2. Related work

Our approach to computing the semantic similarity of lexical definitions of geographic terms is informed by several research areas, including geo-semantics and lexical similarity measures. Several semantic similarity measures for geospatial data have been devised at the intersection between cognitive science, psychology, ontology engineering and geographic information retrieval (GIR) (Janowicz et al. 2011). A detailed survey of geo-similarity has been conducted by Schwering (2008), including geometric, featural, network, alignment and transformational models. Lexical measures based on natural language processing are not included in the survey.

Notably, Rodríguez and Egenhofer (2004) have extended Tversky's set-theoretical ratio model in their Matching-Distance Similarity Measure (MDSM). Schwering and Raubal (2005) proposed a technique to include spatial relations in the computation of semantic similarity. Furthermore, Janowicz et al. (2007) developed Sim-DL, a similarity measure for geographic terms based on description logic (DL), a family of formal languages for the Semantic Web. As such approaches rely on rich, formal definitions of geographic terms, they are unfit to compare volunteered lexical definitions. In our previous work, we have explored techniques suitable for VGI, such as graph-based measures of semantic similarity on the OSM Semantic Network (Ballatore et al. 2012a). We have enriched the OSM semantic model with Semantic Web resources (Ballatore and Bertolotto 2011), and we have outlined a GIR system based on the semantic similarity of map viewports (Ballatore et al. 2012b).

Keßler (2007) surveyed the idea of context in existing geo-similarity measures. All approaches share the idea of adjusting the weights of the dimensions that are considered in the similarity measurement, to reflect a specificity of the context. The context is modelled by restricting the computation to a subset of terms contained in the knowledge base. In this article, we focus on general semantic similarity of definitions and the role of context in the comparison between lexical definitions is left implicit.

2.1. Term-to-term semantic similarity

A semantic similarity measure quantifies the association between two given terms ($sim(t_a, t_b) \in \Re$). Approaches to computing the semantic similarity of terms (as opposed to larger semantic entities) can be classified in two main families: *knowledge-based* and *corpus-based*. Knowledge-based techniques rely on representational artifacts, such as semantic networks, taxonomies or full-fledged ontologies. Under a structuralist assumption, most of these techniques observe the relationships that link the terms, assuming, for example, that the taxonomical distance is inversely proportional to the semantic similarity (Rada et al. 1989).

The lexical database WordNet, because of its focus on semantic relationships and dense connectivity, has been successfully used as a support tool to compute similarity (Fellbaum 1998). A variety of measures have been devised and tested to compute generic lexical similarity on WordNet (e.g. Jiang and Conrath 1997, Hirst and St-Onge 1998, Leacock

Table 1. WordNet-based similarity measures.

Name	Reference	Description	SPath	Gloss	InfoC
path	Rada et al. (1989)	Edge count	√		
lch	Leacock and Chodorow (1998)	Edge count scaled by depth	√		
res	Resnik (1995)	Information content of *lcs*	√		√
jcn	Jiang and Conrath (1997)	Information content of *lcs* and terms	√		√
lin	Lin (1998)	Ratio of information content of *lcs* and terms	√		√
wup	Wu and Palmer (1994)	Edge count between *lcs* and terms	√		
hso	Hirst and St-Onge (1998)	Paths in lexical chains	√		
lesk	Banerjee and Pedersen (2002)	Extended gloss overlap		√	
vector	Patwardhan and Pedersen (2006)	Second order co-occurrence vectors		√	
vectorp	Patwardhan and Pedersen (2006)	Pairwise second order co-occurrence vectors		√	

Notes: SPath: the measure uses the shortest path in the Word-Net taxonomy; Gloss: the measure exploits lexical definitions (glosses); InfoC: the measure uses the information content of terms.

and Chodorow 1998). The characteristics of 10 popular WordNet-based measures are summarised in Table 1. This table shows the core tenets of each measure, i.e. whether they rely on network topology, information content or on glosses, that is, lexical definitions. As such, measures are complementary, and they can be combined to overcome their limitations (Ballatore et al. 2012c).

Unlike knowledge-based approaches, corpus-based techniques do not need explicit relationships between terms and compute semantic similarity of two terms based on their co-occurrence in a large corpus of text documents (e.g. Turney 2001). An underlying assumption of these techniques was famously put forward by Harris (1954) as the 'distributional hypothesis', which states that words that occur in the same contexts tend to have similar meanings. Going beyond simple co-occurrence, latent semantic analysis (LSA) has become a prominent approach to extract a similarity model from a text corpus (Landauer et al. 2007). Instead of only looking at the total number of co-occurrences, LSA considers detailed patterns of co-occurrence in individual sentences.

2.2. Text-to-text semantic similarity

A family of semantic similarity measures focus on the similarity of segments of text, instead of isolated terms. A classic area of research, whose core preoccupation is such type of similarity, is the detection of plagiarism in academic tests and publications (Potthast et al. 2010). Anti-plagiarism techniques rely on the distributional hypothesis to detect suspiciously close citation patterns and similarities in writing styles across different text documents. More recently, the problem of paraphrase detection has become an active research area (Corley and Mihalcea 2005). For example, the sentence 'The Iraqi Foreign Minister warned of disastrous consequences if Turkey launched an invasion of Iraq' should be classified as a paraphrase of 'Iraq has warned that a Turkish incursion would have disastrous results' (Fernando and Stevenson 2008, p. 2). The challenge lies in the detection of sentences that convey roughly the same meaning through a different lexicon.

Mihalcea et al. (2006) have developed a hybrid approach to text similarity, combining WordNet-based and corpus-based techniques. In terms of precision, the knowledge-based measures consistently outperform the corpus-based ones, as well as the combined approach. Tackling similar issues, Islam and Inkpen (2008) combined corpus-based term similarity with string similarity to compute the similarity of short texts. Furthermore, random walks on graphs represent an emergent alternative approach to network-based similarity (Ramage et al. 2009). Because of their conceptual simplicity and effectiveness, knowledge-based approaches by Corley and Mihalcea (2005) and Fernando and Stevenson (2008) were included in our technique.

A fundamental trade-off lies between knowledge-based and corpus-based approaches. Knowledge-based approaches rely on expert-authored knowledge bases, resulting in more precise similarity models, but at the cost of limiting the coverage to the set of terms included in the knowledge base. By contrast, corpus-based approaches tend to cover a wider set of terms, but at the cost of lower precision in the similarity model. Given the restricted domain covered by the lexical definitions of geographic terms, we adopt a knowledge-based approach, favouring precision over coverage.

3. Lexical definitions

Lexical definitions play a fundamental role as a tool to create a shared semantic ground in an information community. In this section, we discuss lexical definitions of geographic terms, and we focus specifically on the context of VGI. Following a well-established semiotic tradition, we refer to 'terms' as symbols pointing to 'concepts' (also known as 'thoughts'), complex psychological entities utilised by agents to produce and interpret geographic information about real-world entities (Kuhn 2005). Such external entities are called 'objects' or 'referents'. The relationship between symbols and concepts is not static and absolute but dynamic and rooted in social agreements among semantic agents.

Because of their ubiquity, lexical definitions have been at the centre of centuries-long debates in philosophy and epistemology. A classical tenet is that a 'definition signifies the necessary and sufficient conditions for the application of the term being defined' (Kavouras and Kokla 2008, p. III). The term being defined is called the *definiendum*, whilst the lexical definition is the *definiens*. As terms are intrinsically vague, lexicons suffer from semantic ambiguity. Even basic geographic terms such as 'mountain' and 'lake' refer to objects with fuzzy boundaries, whose definitions can vary considerably across different information communities (Varzi 2001).

Lexical definitions rely implicitly on broad lexical semantic theories. Lexical semantic theories are divided into *constructive* and *differential* (Miller et al. 1990). In constructive theories, a lexical definition lex_a has to provide the agent h with enough information to accurately construct the concept a. However, in practice, the requirements for a constructive lexical definition are rarely met, as lexical definitions tend to be heavily underspecified, i.e. relying on agents' prior knowledge of the domain. Differential theories, on the other hand, envisage lexical definitions to help the agent distinguish between semantically similar terms.

3.1. Geographic lexical definitions

A typical lexical definition in a geographic lexicon is *intensional*, describing properties of the concept, and *precising*, narrowing the usage of the target term to the desired scope. Furthermore, lexical definitions are *prescriptive*, showing the intended usage of the term

in a given informational context. In some cases, a lexical definition can be *stipulative*, i.e. defines a novel meaning for a term conventionally bearing a different meaning. In the tradition of Linnaean taxonomies, lex_a may include *genus* (the family) and *differentia* (the salient differences to other members of the same family). This is often the case when very similar concepts have to be distinguished and the existing lexicon does not clearly differentiate them.

The role of lexical definitions in the interaction between semantic agents can be expressed formally from a set-theoretical perspective. Given a term t_a in a geographic lexicon L (e.g. 'river'), a lexical definition lex_a provides information about t_a (e.g. 'A large natural stream of water flowing in a channel'). The lexicon L enables information sharing within an information community of human agents H. The term t_a symbolises the concept a_h, held in the human agent's mind according to their prior experience and knowledge of rivers. The lexical definition lex_a is a function that enables the construction of concept a_{h_1} for agent h_1 : $lex_a(t_a, h_1) \rightarrow a_{h_1}$ The success of communication between agents h_1 and h_2 depends on the overlap between the concepts generated by the two agents $a_{h_1} \cap a_{h_2}$. On one extreme, the identity $a_{h_1} \equiv a_{h_2}$ would correspond to perfect communication. On the other, if $a_{h_1} \cap a_{h_2} = \emptyset$, communication between h_1 and h_2 fails.

A definition lex_a is successful to the degree to which the constructed concepts a_h converge towards a shared concept a for all agents $h \in H$. If lex_a contains too much information or information not salient to the agents, it is overspecified. For example, in a given geographic information community it might be irrelevant to specify the typical biological population of different kinds of 'river'. By contrast, lex_a is underspecified if the constructed concepts a_h do not converge to a. Naturally, in real information communities, it is unlikely that all agents $h \in H$ share precisely the concept a. Some convergence towards a is sufficient to enable semantic co-operation and information production in H, limiting the semantic failures to a non-critical degree.

As Kavouras et al. (2005) pointed out, lexical definitions of geographic terms differ to non-geographic terms in a number of respects. Semantic properties capture the *location*, the *purpose* of man-made entities and the *cause* that brought an entity into existence. Similarly, semantic relations indicate the spatial *adjacency* to, the *surroundedness* by and *association* with other geographic entities. While lexical definitions in expert-authored geo-lexicons, such as those by the British Ordnance Survey,[3] show a high degree of consistency, VGI differs in a number of ways.

3.2. Volunteered lexical definitions

In VGI projects, a large number of human agents co-operate to generate reusable geographic information (Goodchild 2007). The negotiation of a shared geographic lexicon is therefore crucial to the success of any productive effort. In the context of geo-semantics, it is important to assess how volunteered lexical definitions differ from traditional ones. Unlike traditional expert-centred production processes, the ontological ground of volunteered endeavours is thin, dynamic and often ontologically unsound. Rather than relying on conceptual models carefully engineered and validated by experts, these projects are based on open tagging mechanisms and folksonomies (Gruber 2007).

From a semantic perspective, VGI faces a tension between the desiderata of ontology engineering (correctness, minimality, expressivity, precision, etc.) and the accessibility to non-expert contributors. The contributors' drive to share pieces of information can be easily hampered by complex insertion and validation procedures. For example, a contributor can quickly share the location of her favourite restaurant, simply by tagging a point with the

label 'restaurant', receiving instant gratification. Tapping this drive is essential to guarantee the initial survival and eventual flourishing of crowdsourcing projects (Shirky 2010).

To discuss lexical definitions in VGI, the collaborative project OSM constitutes a paradigmatic case. The OSM contributors form a large and heterogenous information community that co-operate to produce a general-purpose world map. The shared semantic ground is provided through the OSM wiki website, which hosts crowd-sourced lexical definitions of OSM terms. This semantic model is a semi-structured folk-sonomy, i.e. a lexicon of terms defined by contributors in a negotiation process. Terms, called 'tags' in the OSM terminology, are structured in key/value pairs, where key is a general category and the value is more specific (e.g. *amenity* = *university*). OSM contributors define, update and refer to these definitions to create and utilise the OSM vector data set. The OSM Semantic Network, a machine-readable semantic artefact, was extracted from the wiki website (Ballatore *et al*. 2012a).[4] The average length of lexical definitions in the OSM Semantic Network is 44 terms, with a standard deviation of 33. Only a small number of definitions are substantially larger.

From an ontological viewpoint, OSM adopts a fully object-based perspective, without representing field-based views (Goodchild *et al*. 2007). The terms are used to provide meta-data for discrete objects, encoded as points, polylines and polygons. The lexicon is dynamic and the choice of terms is left loosely regulated, following the slogan 'any tags you like'.[5] Table 2 shows a sample of lexical definitions of OSM terms. The semantics of OSM terms is heavily underspecified: it includes no explicit taxonomy and does not encode a partonomy. For example, the term *room* = * is described as part of a building, but this relation is not defined formally. The conceptualisation does not distinguish between attributes and relations, keeping all terms at the same representational level (e.g. *width* = * is not modelled as an attribute of another entity). Terms are generally vastly underspecified, encouraging the creative generation of new terms.

Depending on the terms, definitions can be very detailed to the point of overspecification (e.g. *leisure* = *garden*), or circular, when the term is deemed to be self-explanatory (e.g. *building* = *university*). Some definitions adopt a differential perspective, indicating

Table 2. Examples of lexical definitions of OSM terms, extracted from the OSM Semantic Network on 25 August 2012.

Term (Key = Value)	Lexical definition
leisure = *garden*	Place where flowers and other plants are grown in a decorative and structured manner or for scientific purposes. . . . A garden can have aesthetic, functional and recreational uses. Not to be confused with *shop* = *garden-centre*. Meant to tag the land area itself.
waterway = *stream*	A naturally-forming waterway that is too thin to be classed as a river. Maybe you can just jump over it. It may not even be permanently filled with water. An active, able-bodied person should be able to jump over it if trees along it are not too thick.
building = *university*	A university building. Completed with the tag *amenity* = *university*.
room = *	Room marks a room inside a building. Man made. The room Key is used to mark nodes or areas as a room, typically inside a building. Part of Proposed features/indoor.
natural = *wetland*	An area subject to inundation by water or where waterlogged ground may be present. Examples include swamps, tropical mangroves, tidal, marshland and bogs. Wetlands are also typically found at the fringes of rivers, lakes or along the coastline.

relevant differences from semantically related terms (e.g. *leisure* = *garden* is 'Not to be confused with *shop* = *garden-centre*'). Some definitions contain OSM-specific details (e.g. 'Part of Proposed features' and 'Approved') and prescriptive and stipulative details, indicating to the human agent how the term should be used in the context of OSM, focussing on counter-intuitive aspects of the lexicon. The main issue with the OSM lexical definitions is the frequent underspecification and the high variability of length, semantic content and quality across the terms. To meet the challenges of this uncertain semantic territory, we investigate a bag-of-words approach to compute the semantic similarity of OSM terms, outlined in the next section.

4. A semantic similarity measure for lexical definitions

This section outlines an approach to compute semantic similarity of lexical definitions in a geographic lexicon, based on WordNet as the source of semantic similarity and paraphrase-detection techniques. The purpose of this measure is to quantify the semantic similarity s of two given geographic terms t_a and t_b as a real number, based on their lexical definitions lex_a and lex_b. The real number $s(t_a, t_b)$ is not meaningful in isolation, but conveys useful information when compared with other values of other pairs of terms.

The psychological intuition behind our approach to geo-semantic similarity is that similar terms are described using similar terms. The infinite regression that would ensue is avoided by using WordNet to compute the similarity scores of definitional terms. The approach computes the semantic similarity of two terms $s(t_a, t_b)$ based on four input parameters $\{POS, C, sim_t, sim_v\}$: a part-of-speech (POS) filter, which consists of a set of valid POS (e.g. nouns, verbs and adjectives); a text corpus C; two similarity functions sim_t and sim_v. The four steps of the similarity algorithm are the following:

(1) Given two terms t_a and t_b, lemmatise and POS-tag their definitional terms $\{t_{a1}. . .t_{an}\}$ and $\{t_{bl}. . .t_{bm}\}$.
(2) Construct semantic vectors $\vec{a}$ and $\vec{b}$, including definitional terms based on the POS filter, and retrieving weights w_t from corpus C.
(3) Construct similarity matrices M_{ab} and M_{ba}. Each cell of these matrices contains a term-to-term similarity score sim_t (t_{ai}, t_{bj}), relying on WordNet as a knowledge base.
(4) Compute similarity score $s(t_a, t_b)$ from the similarity matrices using vector-to-vector similarity sim_v, based on paraphrase-detection techniques.

The architecture of the approach and its resources are schematised in Figure 1. For illustrative purposes, we consider lexical definitions from the OSM Semantic Network. The

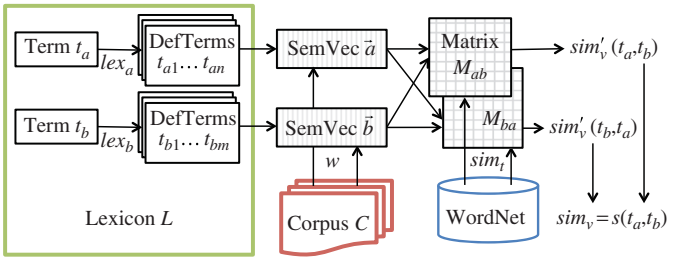

Figure 1. Approach to semantic similarity of lexical definitions *lex* for terms t_a and t_b.

remainder of this section describes in detail the four steps to compute the semantic similarity of geographic terms.

4.1. Extraction of definitional terms

Given a geographic term t_a, the lexical definition lex_a can be seen as a function associating t_a with a set of definitional terms $\{t_{a1} \ldots t_{an}\}$, i.e. words contributing to the overall meaning of the term t_a. From a linguistic perspective, the term t_a is the *definiendum*, whilst the definitional terms are part of the *definiens*. Definitional terms can either be single words (e.g. 'lake') or compound words (e.g. 'swimming pool') and are described by a part-of-speech tag ($t = word/POS$), following the format of the Penn Treebank (Marcus et al. 1993). Among all the Penn tags, we restrict our method to nouns (NN), verbs (VB) and adjectives (JJ). To facilitate their usage, the definitional terms must be lemmatised (e.g. 'rivers' to 'river', 'ate' to 'eat'). For example, the OSM lexical definition of 'garden' as a 'place where flowers and other plants are grown', reported in Table 2 contains the definitional terms *place*/*NN*, *flower*/*NN*, *plant*/*NN* and *grow*/*VB*.

4.2. Construction of semantic vectors

As discussed in the previous section, the definitional terms $\{t_{a1} \ldots t_{an}\}$ contribute to conveying the meaning of t_a. A geographic term t_a can therefore be represented as a semantic vector, whose components represent the definitional terms. This representation is often called bag-of-words, as it does not take into consideration the syntactical structure of text, but only the presence of terms, regardless of their order. To represent the geographic term t_a in a semantic vector space, we define a multidimensional vector $\vec{a}$, whose scalars are non-negative weights and components are the definitional terms $\{t_{a1} \ldots t_{an}\}$:

$$\vec{a} = w_{a1} \cdot t_{a1} + ... + w_{an} \cdot t_{an} \qquad \forall w_{ai} : w_{ai} \in \mathbb{R}_{\geq 0} \qquad \sum_{i=1}^{n} w_{ai} = 1 \tag{1}$$

The weights w capture the fact that, in a lexical definition, terms can contribute to the overall meaning of the *definiendum* to different degrees. For example, in the definition of OSM term *building* = *university* in Table 2, the terms *university*/*NN* and *building*/*NN* carry more meaning than *completed*/*JJ* or *tag*/*NN*. Thus, an appropriate weighting scheme has to be used to compute the weight w_t of a term t in the corpus.

Hence, a corpus of text documents C is utilised to compute the weights of terms on a statistical basis. The corpus C is a set of text documents, each containing a set of terms. Formally, let the document $d_i = \{t_1 \ldots t_n\}$ and $C = \{d_1 \ldots d_m\}$. A term can be non-unique ($\exists(t_i, t_j) : t_i = t_j$). A popular approach to computing the weight of the terms in a corpus is the term frequency–inverse document frequency (TF-IDF). IDF scores represent how common a term t is in the whole corpus, while TF is the number of occurrences of t in document d. A relatively infrequent term in a corpus is expected to have a higher weight than a very frequent one. The result of this step consists of the semantic vectors $\vec{a}$ and $\vec{b}$.

4.3. Construction of similarity matrices

The similarity $s(t_a, t_b)$ could, in principle, be computed as a simple vector similarity $\vec{a}$ and $\vec{b}$, e.g. as the inverse of the distance between the vectors, such as the Euclidean, cosine and Chebyshev distances. However, these techniques are effective only when there is an overlap

between the components of the two vectors. When there is little or no overlap in the vectors, the vast majority of similarity scores are 0. For this reason, the term-to-term similarity function sim_t is necessary to capture a continuous semantic distance between terms in $\vec{a}$ and $\vec{b}$. We define a normalised, term-to-term semantic similarity measure $sim_t(t_i, t_j) \in [0,1]$.

In principle, any term-to-term semantic similarity measure might be adopted as sim_t. However, in order to compute semantic similarity exploiting relationships between terms, knowledge-based approaches are generally deemed to have higher precision than corpus-based techniques (Mihalcea et al. 2006, Agirre et al. 2009). In Table 1 (Section 2.1), we described 10 WordNet similarity techniques, which are utilised in this phase. These measures consider semantic relationships such as synonyms (e.g. 'buy' and 'purchase'), hyponyms and hypernyms (e.g. 'oak' and 'tree'). As the evaluation will confirm, the choice of sim_t impacts on the cognitive plausibility of the measure. In this sense, sim_v is a function $f(sim_t)$. Two similarity matrices, M_{ab} and M_{ba}, have to be constructed with sim_t:

$$\begin{aligned} |M_{ab}| &= |\vec{a}| \times |\vec{b}| \quad \forall i,j : M_{i,j} = sim_t(t_{ai}, t_{bj}) \\ |M_{ba}| &= |\vec{b}| \times |\vec{a}| \quad \forall i,j : M_{j,i} = sim_t(t_{bj}, t_{ai}) \end{aligned} \tag{2}$$

If sim_t is symmetric ($sim_t(t_i, t_j) = sim_t(t_j, t_i)$), $M_{ba} = M_{ab}^T$. The matrices M_{ab} and M_{ba} are used by the vector-to-vector measures sim_v.

4.4. Computation of overall similarity

Having constructed the semantic vectors $\vec{a}$ and $\vec{b}$ and the matrices M_{ab} and M_{ba}, it is possible to compute the vector-to-vector similarity sim_v. As the comparison of two semantic vectors is in many ways analogous to paraphrase detection, in this step, we consider techniques originally developed in this area of research. An asymmetric similarity measure of semantic vectors $sim'_v(\vec{a}, \vec{b})$ can be formalised as follows:

$$\begin{aligned} &sim'_v(\vec{a}, \vec{b}) = \sum_{i=1}^{|\vec{a}|} w_{ai} \cdot \hat{s}(t_{ai}, \vec{b}, M_{ab}), \quad sim'_v(\vec{b}, \vec{a}) = \sum_{i=1}^{|\vec{b}|} w_{bi} \cdot \hat{s}(t_{bi}, \vec{a}, M_{ba}) \\ &sim'_v(\vec{a}, \vec{b}) \neq sim'_v(\vec{b}, \vec{a}), \; sim'_v(\vec{a}, \vec{b}) \in [0, 1] \end{aligned} \tag{3}$$

where function $\hat{s}$ returns a similarity score between a definitional term, a semantic vector, based on a similarity matrix. Two alternative functions can be adopted as $\hat{s} : \hat{s}_{com}$ or $\hat{s}_{fes}$. Based on the approach presented by Corley and Mihalcea (2005), abbreviated as *com*, the function $\hat{s}$ corresponds to the maximum semantic similarity score between a term t_i and a vector $\vec{v}$, based on the similarity scores in matrix M:

$$\hat{s}_{com}(t_i, \vec{v}, M) = M_{i,k} \quad \forall t_j \in \vec{v} : M_{i,j} \leq M_{i,k} \tag{4}$$

More recently, Fernando and Stevenson (2008) have developed an alternative approach, which we abbreviate as *fes*. Instead of considering only the terms with maximum similarity to the vector being analysed, $\hat{s}$ is the sum of all similarities in M, where $t_j \in \vec{v}$:

$$\hat{s}_{fes}(t_i, \vec{v}, M) = \sum_{j=0}^{|\vec{v}|} M_{i,j} \tag{5}$$

A symmetric measure $sim_v \in [0, 1]$ can be easily obtained from sim'_v as:

$$sim_v(a, b) = \frac{sim'_v(a, b) + sim'_v(b, a)}{2}, sim_v(a, b) = sim_v(b, a) \quad (6)$$

This knowledge-based approach relying on semantic vectors enables the computation of the semantic similarity of lexical definitions. In order to give precise indications of its practical applicability, it is necessary to analyse the spatio-temporal computational complexity of the approach. First, the construction of semantic vectors can be considered to have a linear complexity $O(|\vec{a}| + |\vec{b}|) = O(n)$, where n is the number of terms present in the two lexical definitions. Subsequently, the computational complexity of the term-to-term similarity scores depends on the selected sim_t. WordNet sim_t measures vary from shortest-path algorithms (*path*, *wup*, *res*, etc.), to very complex lexical chains (*hso*).

As a general estimate of the complexity for sim_t, we consider Dijkstra's classic shortest-path algorithm $O(|\vec{a}| \cdot |\vec{b}| \cdot (|W_E| + |W_V| log |W_V|)) = O(n^3)$, where W_E and W_V are the edges and vertices of WordNet. The vector-to-vector measures construct two similarity matrices of size $|\vec{a}| \times |\vec{b}|$. The overall upper bound of the lexical approach complexity is cubic: $O(n + n^3 + n^2) \leq O(n^3)$. Considering the limited size of geographic lexicons (for example, the OSM Semantic Network contains about 6500 terms) and applying the appropriate pre-computations, the cubic complexity does not constitute an obstacle for the practical usage of the technique.

5. Evaluation

This section presents an empirical evaluation of our approach to compute the semantic similarity of volunteered lexical definitions. The hypotheses that this evaluation purports to validate are the following:

(1) The volunteered lexical definitions of geographic terms allow the computation of a more plausible semantic similarity measure than the terms in isolation.
(2) Our bag-of-words, knowledge-based approach can reach a high cognitive plausibility, handling the high variability of volunteered lexical definitions.

To evaluate similarity measures, two complementary approaches have been widely utilised: *cognitive plausibility* and *task-based evaluation*. The approach based on cognitive plausibility aims at quantifying the effectiveness of a similarity measure through psychological tests, directly comparing human similarity judgements on sets of pairs with machine-generated scores. The plausibility is therefore proportional to the technique's ability to mimic human behaviour. Task-based evaluation, by contrast, applies the similarity measure to a specific task, for example, in the area of natural language processing. The performance of the similarity measure is inferred from how satisfactorily the task is carried out, using appropriate information retrieval metrics such as precision and recall.

In this evaluation, we adopted the cognitive plausibility approach, utilising a human-generated data set as ground truth. Formally, given the human-generated rankings for term pairs R_{H} and the corresponding machine-generated rankings R_{M}, the cognitive plausibility can be computed as Spearman's $\rho(R_{\mathrm{H}}, R_{\mathrm{M}})$, in range $[-1, 1]$. This evaluation permits the detailed observation of the impact of the algorithm parameters, providing empirical insights.

5.1. Ground truth

A set of human rankings of geographic term pairs, the MDSM evaluation data set, was utilised as ground truth. This data set was originally collected by Rodríguez and Egenhofer (2004) from 72 human subjects and was utilised to evaluate the MDSM, their semantic similarity measure. Because these terms were defined with short lexical definitions without focussing on ontology-specific details, they are suitable to study the cognitive plausibility of our approach. The original data set contains contextual and non-contextual similarity judgements for 108 geographic term pairs, grouped in 10 questions, with 10 or 11 pairs in each question. The term pairs cover 33 terms, ranging from large natural entities (e.g. 'mountain' and 'forest') to man-made features (e.g. 'bridge' and 'house'). The subjects were asked to rank the pairs from the most to the least similar (e.g. ⟨athletic field, ball park⟩ → . . . → ⟨athletic field, library⟩).

In order to compare the human rankings with the rankings generated by our technique, the terms were manually mapped onto the corresponding terms in OSM and WordNet, based on their lexical definitions. For example, the term 'tennis court' was matched to OSM tag *sport = tennis* and WordNet synset *tennis court#n#1*. While 29 terms have a satisfactory equivalent in OSM, four terms ('terminal', 'transportation', 'lagoon' and 'desert') were discarded because they did not have a precise matching term in OSM. The four questions that include contextual judgements, outside the scope of this study, were excluded. The resulting data set consisted of the rankings of 62 term pairs, grouped in six questions, and covering 29 geographic terms.[6] In order to summarise the cognitive plausibility of the algorithms evaluated in the experiments that follow, the results across six questions had to be aggregated. For this purpose, we utilised the Hunter–Schmidt meta-analytical method by obtaining a tie-corrected Spearman's p as a weighted average across the six questions (Field 2001).

5.2. Preliminary experiments

To validate the first hypothesis, we ran three preliminary experiments. First, the semantic similarity of the 62 term pairs in the MDSM evaluation data set was computed using the 10 WordNet-based similarity measures outlined in Section 2.1. These measures quantify the semantic similarity of two input terms by observing different aspects of the WordNet taxonomy. The tie-corrected Spearman's ρ was obtained between each of the 10 WordNet-based rankings and the human rankings. The cognitive plausibility of this approach turned out to be poor, with ρ falling in the interval [.24, .5]. Moreover, the statistical significance of the correlation test was largely insufficient for several WordNet measures ($P > .1$). The inadequacy of this approach validates the hypothesis, supporting the necessity of including the lexical definitions in the similarity computation.

The second experiment focussed on the similarity measure based on set-theoretical overlap of terms. A term is similar to another term to the degree to which they share the same terms in their lexical definition, seen as a bag-of-words, using Tversky or related set-theoretical approaches, such as the Dice coefficient (Tversky 1977). The lexical similarity of the pairs contained in the MDSM evaluation data set was computed using the Text::similarity tool on the OSM definitions.[7] As the lexical definitions share a very limited number of terms (mostly generic terms such as 'refer' or 'tag'), these techniques incur a limited-information problem, showing no correlation with the human scores ($\rho \in [-.1, .1]$ for all the set-theoretical measures).

To support the claim that existing corpus-based measures do not handle the similarity of volunteered lexical definitions, we ran a third experiment focussing on latent semantic analysis (LSA), a state-of-art semantic similarity measure (Landauer et al. 2007).

We selected two typical LSA similarity analyses, term-based and document-based, using the Touchstone Applied Science Associates (TASA) data set as a general-purpose English corpus.[8] In the case of the term-based comparison, the similarity of term definitions is computed in the vector space of terms, extracted from the corpus, whilst the document-based case analyses the texts in the document vector space.

Hence, we computed the similarity scores of the term pairs in the MDSM evaluation data set using the two LSA analyses. For example, the similarity of pair <lake,mountain> was based exclusively on the terms 'lake' and 'mountain'. This approach obtained a cognitively plausibility of $\rho = .54$ for the term-based analysis and of .56 for the vector-based analysis (with $P < .05$ in both cases). Subsequently, the same analyses were run on the lexical definitions, with more plausible results ($\rho = .61$ and .64, with $P < .01$). In this case, the term 'lake' was compared with 'mountain' based on their full OSM lexical definitions. Therefore, the cognitive plausibility of LSA fell in the interval [.54, .64]. The experiment described in the following section shows that our knowledge-based approach obtains considerably better results.

5.3. Experimental set-up

In order to explore in detail the performance of our approach to lexical similarity of geographic terms, we included several options for the four parameters $\{POS, C, sim_t, sim_v\}$: three POS combinations, 10 WordNet-based term-to-term similarity measures, four text corpora and two vector-to-vector measures (Table 3). Further details of the corpora C are summarised in Table 4. Each of the 240 combinations of parameters $\{POS, C, sim_t, sim_v\}$ returns a unique set of similarity scores for the term pairs, which can be compared with the human-generated ground truth.

The lexical definitions of the 29 terms were extracted from the OSM Semantic Network, and the definitions were lemmatised with Stanford CoreNLP and POS tagged with the Stanford log-linear part-of-speech tagger.[9] Of all the tagged terms, only nouns, verbs and adjectives were selected (*NN*, *VB* and *JJ,* respectively). As a result, 1406 terms

Table 3. Experiment set-up: resources included for each of the four input parameters.

Symbol	Number	Description
POS	3	POS filters (*NN*, *VB*, *NN VB*). Adjectives (*JJ*) were initially included, but most sim_t measures are designed to handle only nouns and verbs, making a direct comparison difficult.
C	4	We collected three corpora: the OSM wiki website, a set of random news stories from the newspaper *Irish Independent* and Wikipedia. The OSM wiki website corpus is strongly skewed towards geographic and OSM-related terms, whilst the *Irish Independent* and Wikipedia represent examples of primarily non-geographic corpora (see Table 4). The *Null* corpus corresponds to constant weights, i.e. a constant $w > 0$.
sim_t	10	The term-to-term similarity function sim_t is utilised to construct the similarity matrices needed to compute the similarity. Table 1 in Section 2.1 shows the 10 WordNet-based semantic similarity measures included in the experiment: *hso*, *jcn*, *lch*, *lesk*, *lin*, *path*, *res*, *vector, vectorp* and *wup*.
sim_v	2	Two vector-to-vector similarity measures, originally developed to detect paraphrases, were included: *com* (Corley and Mihalcea 2005) and *fes* (Fernando and Stevenson 2008)
Total	240	$\lvert POS\rvert \cdot \lvert C\rvert \cdot \lvert sim_t\rvert \cdot \lvert sim_v\rvert$

Table 4. Details of text corpora C, used to weight terms in semantic vectors.

Corpus name	Extracted on	Doc Number	Term Number	Description
OSM wiki website	13 October 2011	6.4×10^3	2.5×10^6	Wiki website of OSM.
Irish Independent	28 May 2011	4.9×10^3	2.2×10^6	News stories.
Wikipedia	11 November 2011	3.8×10^6	2.5×10^9	English Wikipedia

were selected, of which 789 were nouns, 236 were adjectives and 381 were verbs. Because of the complexity of computing their similarity in WordNet, adjectives were finally discarded.

The similarity matrices M_{ab} and M_{ba} were constructed with the tool WordNet::Similarity (Pedersen et al. 2004). This pre-computation offers the possibility of empirically observing the temporal complexity of each measure. Bearing in mind the variability induced by the implementation of the measures, it is possible to notice that shortest path-based measures (*path*, *res*, *jcn*, *lin* and *wup*) are remarkably faster to compute than both the gloss-based ones (*lesk*, *vector* and *vectorp*) and the lexical-chains of *hso*.

5.4. Experimental results

For each of the 240 combinations $\{POS,\ C,\ sim_t,\ sim_v\}$, we computed the tie-corrected Spearman's correlation coefficient ρ, which expresses the correlation between the rankings computed by our similarity approach applied to the OSM Semantic Network and the MDSM evaluation data set. The correlation ρ is computed between the human and the machine-generated rankings over the 62 term pairs.[10] This correlation captures the cognitive plausibility of our computational approach, using the MDSM evaluation data set as ground truth, where $\rho = 1$ corresponds to perfect correlation, $\rho = 0$ corresponds to no correlation and $\rho = -1$ corresponds to perfect inverse correlation. For the cases including only verbs ($POS = \text{VB}$), the correlations with the human data set were very weak ($\rho \in [-.1,.4]$) and obtained insufficient statistical significance ($P > .1$). For this reason, we excluded these cases from the analysis, and the resulting correlations are not reported. By contrast, all the other parameter combinations obtained high statistical significance ($P < .001$).

To measure the performance of a particular parameter (e.g. $C = Null$ or $sim_t = lin$), all the cases where that parameter is used are treated as a set of ρ. By looking at the distribution of ρ within these sets, it is possible to notice that the distributions are often skewed to the left, i.e. towards lower values. For this reason, we consider the median to be a more robust descriptor of central tendency than the mean. Table 5 reports the detailed results of the cognitive plausibility of our approach to lexical similarity. For example, when the POS filter is set to *NN*, the distribution of ρ results in a $\tilde{\rho} = .68$, with a maximum value of .81. The distributions of ρ are displayed in Figure 2. This allows for an intuitive understanding of the impact of each parameter option on the ρ distribution.

POS filter. The selection of certain parts of speech (POS) have a noticeable impact on the results. As is reasonable to expect, given the descriptive nature of the OSM Semantic Network definitions, nouns (*NN*) carry most of the term semantics, with $\tilde{\rho} = .68$ Verbs (*VB*) showed a very weak correlation with the human data set and, because of the low statistical significance of their correlation tests, were not included in the analysis. The combination of nouns and verbs (*NN VB*) performs marginally better than nouns by themselves

Table 5. Results summary of cognitive plausibility. The central tendency of each parameter is summarised by the median $\tilde{\rho}$, its lower and upper quartiles and its maximum value.

Param name	Param value	Median $\tilde{\rho}$	25%–75% quartiles		Max ρ
sim_v	*com*	.7*	.64	.75	.81
	fes	.66	.56	.73	.79
POS	*NN VB*	.7*	.63	.75	.81
	NN	.68	.56	.73	.81
	VB	–	–	–	–
C	*Null*	.7*	.62	.76	.81
	Wikipedia	.7	.6	.73	.8
	Irish Indep	.7	.6	.74	.79
	OSM Wiki	.67	.59	.74	.81
sim_t	*jcn*	.75*	.73	.75	.77
	lch	.74	.69	.76	.81
	wup	.74	.66	.78	.8
	res	.72	.69	.77	.81
	hso	.71	.7	.73	.76
	path	.7	.66	.76	.77
	lin	.67	.58	.75	.81
	vector	.62	.58	.66	.68
	vectorp	.56	.55	.64	.69
	lesk	.52	.45	.55	.6
All	–	.69	.6	.74	.81

Note: *Indicates best performance.

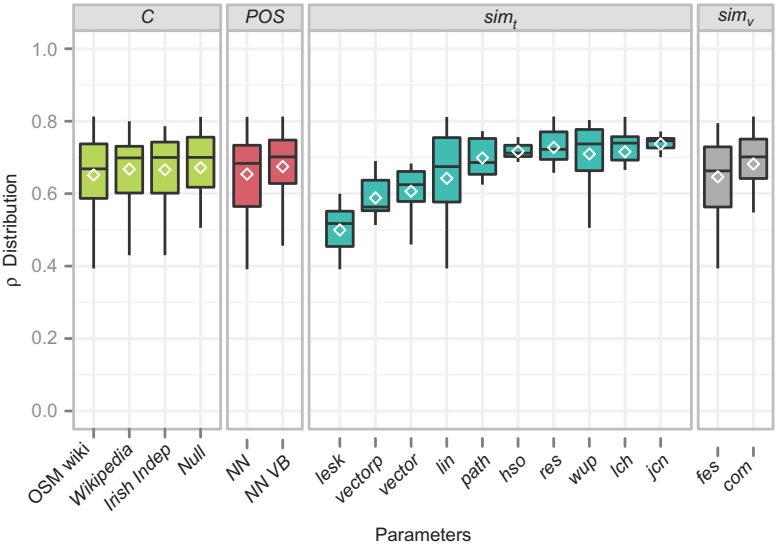

Figure 2. Cognitive plausibility of the similarity measure against the MDSM evaluation data set. The box plot shows the distribution of ρ, sorted by the median in ascending order. For each value of the parameter, the box plot shows the smallest ρ, the 25% quartile, the median, the 75% quartile and largest ρ. The white diamond represents the mean.

($\tilde{\rho} = .7$). The inclusion of verbs also slightly reduces the dispersion, and the best results are obtained when both nouns and verbs are included. When compared with a paired *t*-test, the correlations of cases with *NN VB* and *NN* are significantly different ($P < .01$).

Corpus *C*. The corpus *C* determines the vector weights. Figure 2 shows ρ grouped by the four corpora included in this study. The *Null*, *Irish Independent* and Wikipedia corpora have very close medians (.7), with the quartiles of the distribution of *Null* being higher. The OSM wiki website shows slightly lower cognitive plausibility ($\tilde{\rho} = .67$). Therefore, it is clear that the corpus-based weights do not improve the cognitive plausibility of the approach. This is consistent with the findings of Fernando and Stevenson (2008), who showed that in the area of paraphrase detection such weighting schemes slightly *worsen* the performance. The reasons for this counter-intuitive result can be various. The OSM wiki website and *Irish Independent* corpora bear an obvious bias (i.e. a geographic lexicon and news stories), and Wikipedia obtains similar results, indicating that the corpus size has no impact on the cognitive plausibility. The weighting scheme (classic TF–IDF) could be sub-optimal compared to other schemes, such as BM25 (Jonesh et al. 2000).

Vector-to-vector similarity measure sim_v. The vector similarity measure, sim_v, has a stronger impact than the corpus on the approach performance. Corley and Mihalcea's *com* approach shows high cognitive plausibility ($\tilde{\rho} = .7$), while Fernando and Stevenson's *fes* tends to obtain lower results ($\tilde{\rho} = .66$). When compared with a paired *t*-test, the results of the two approaches are significantly different ($P < .01$). While in paraphrase detection *fes* performs slightly better than *com*, in the context of semantic similarity the opposite is true (Fernando and Stevenson 2008).

Term-to-term similarity measure sim_t. The term-to-term measure, sim_t, displays higher variability than the other three parameters, and its analysis requires more caution. The measure *jcn* has the highest mean (.75) and reaches .77 as max ρ. Measures *lch* and *wup* have a comparable median (.74) and obtain a higher max value ($\approx .8$). However, as it is possible to notice in Figure 2, *wup* has a wider distribution, with a minimum substantially lower than *jcn* and *lch* (.51). Another measure that falls in the top performing group is *res*, with a lower median (.72) but relatively high quartiles and maximum (.69, .77 and .81, respectively). Thus, the three measures that provide the most promising results are *jcn* (Jiang and Conrath 1997), *lch* (Leacock and Chodorow 1998) and *res* (Resnik 1995). All the other measures have either considerably lower medians, maximum values or have very wide distributions that show low stability in the results.

Fixing the other parameters to optimal values ($POS = NN\ VB$, $C = Null$, $sim_v = com$), the cognitive plausibility of the three top performing sim_t is $\rho = .77 \pm .23$ for *jcn*, $.76 \pm .17$ for *lch*, and $.77 \pm .1$ for *res*, where the confidence intervals were computed with the Hunter–Schmidt meta-analytical method, with $P < .05$. These three measures maintain a high performance even with suboptimal POS (*NN*), respectively, $.76 \pm .18$ (*jcn*), $.81 \pm .09$ (*lch*) and $.7 \pm .08$ (*res*). The stability of these results is further confirmed when selecting a suboptimal corpus *C*, such as the OSM wiki website. Even in this case, the correlation with the human data set falls in the interval [.71, .81], indicating high cognitive plausibility.

6. Conclusions

In this article, we tackled the challenge of devising a computational model for the semantic similarity of terms in a volunteered geographic lexicon. Based on the recursive intuition that similar terms tend to be defined using similar terms, the proposed approach to compute the semantic similarity of volunteered lexical definitions combines existing WordNet and paraphrase-detection techniques. The following conclusions can be drawn:

- Volunteered lexical definitions of geographic terms show high variability in length and low consistency, as generally observed in crowdsourcing projects. As shown in Section 3, such definitions are often circular, underspecified and overspecified. This makes automatic semantic analyses challenging, making existing techniques unsatisfactory. Simple measures based on term overlap between lexical definitions are inadequate ($\rho \in [-.1,.1]$.). WordNet similarity measures applied directly to geographic terms obtain low cognitive plausibility ($\rho \in [.24, .5]$). Corpus-based technique LSA falls in interval $\rho \in [.54, .64]$.
- Our approach quantifies the semantic similarity of geographic terms by comparing their volunteered lexical definitions, using paraphrase-detection techniques and WordNet as a knowledge base. The method was evaluated against human-generated similarity judgements, obtaining high cognitive plausibility ($\rho \in [.71, .81]$). Even without expert- authored, sophisticated conceptual models, our approach provides a highly-plausible measure of semantic similarity for volunteered geographic terms, purely relying on volunteered lexical definitions.
- From a pragmatic viewpoint, the following guidelines to select input parameters $\{POS, C, sim_t, sim_v\}$ can be adopted. Nouns should be included in the POS filter; verbs slightly improve the results. The TF–IDF weighting scheme does not seem to improve the results, making $C = Null$ preferable to other corpora because it is computationally less expensive. Term-to-term similarity measures sim_t either by Jiang and Conrath (1997), Leacock and Chodorow (1998) or Resnik (1995) obtain good results. As vector-to-vector similarity function sim_v, the paraphrase-detection technique by Corley and Mihalcea (2005) outperforms the technique by Fernando and Stevenson (2008). As these guidelines refer to general similarity judgements on geographic terms, particular similarity-based tasks might require a different parameter tuning.

The approach we have presented in this article can be applied to a number of complex semantic-centred tasks, within the same lexicon or between different lexicons. In the context of data mining, it can be used to cluster similar terms together and identify patterns in the spatial distribution of terms. The comparison of terms based on their lexical definitions can also support information integration and alignment between different VGI lexicons, increasing semantic interoperability across diverse geographic communities and the emergent constellation of open knowledge bases (Ballatore et al. 2013). Semantically similar terms are needed to perform query expansion and relaxation in GIR, supporting the users in the querying process (Cardoso and Silva 2007).

Possible extensions to this work might focus on the syntactical aspects, which are ignored in the bag-of-words model (Kavouras et al. 2005). Although the approach presented in this article is geared towards the context of VGI, it could be applied to other geographic information repositories. Furthermore, the validation of the approach on larger human-generated data sets may reveal insights on the strengths and weaknesses of the approach. A deeper understanding of similarity measures would facilitate the usage of the ever-increasing informational wealth extracted by human agents from the implicit geographic 'knowledge soup' residing in their minds.

Notes

1. http://openstreetmap.org, http://ushahidi.com (acc. March 19, 2013)
2. http://wiki.openstreetmap.org/wiki/Editing (acc. March 19, 2013)

3. http://www.ordnancesurvey.co.uk/oswebsite/ontology (acc. March 19, 2013)
4. http://wiki.openstreetmap.org/wiki/OSM_Semantic_Network (acc. March 19, 2013)
5. http://wiki.openstreetmap.org/wiki/Any_tags_you_like (acc. March 19, 2013)
6. The data set is available online at http://github.com/ucd-spatial/Datasets
7. http://text-similarity.sourceforge.net (acc. March 19, 2013)
8. See implementation at http://lsa.colorado.edu
9. http://nlp.stanford.edu/software/corenlp.shtml (acc. 19 March 2013)
10. The 62 pairs in the MDSM evaluation data set are grouped in six questions. To compute ρ, we utilise the Hunter–Schmidt meta-analytical method across the six questions (Field 2001).

References

Agirre, E., *et al*, 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. *In*: NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 2009 Boulder, Colorado, USA. Stroudsburg: Association for Computational Linguistics, 19–27.

Ashish, N. and Sheth, A., eds., 2011. *Geospatial semantics and the semantic web: foundations*, *algorithms*, *and applications*. Vol. 12. New York: Springer.

Ballatore, A. and Bertolotto, M., 2011. Semantically enriching VGI in support of implicit feedback analysis. *In*: K. Tanaka, P. Fröhlich, and K-S. Kim, eds. *Proceedings of the web and wireless geographical information systems international symposium* (*W2GIS 2011*). March 3–4 2011, Kyoto, Japan, Vol. 6574 of *LNCS*. Berlin: Springer, 78–93.

Ballatore, A., Bertolotto, M., and Wilson, D., 2012a. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, Avaliable from: http://link.springer.com/article/10.1007%2Fs10115-012-0571-0 [Accessed 24 March 2013].

Ballatore, A., Wilson, D., and Bertolotto, M., 2012b. A holistic semantic similarity measure for viewports in interactive maps. *In*: S. D. Martino, A. Peron, and T. Tezuka, eds. *Proceedings of the web and wireless geographical information systems international symposium* (*W2GIS 2012*) April 12–13, 2012, Naples, Italy, Vol. 7236 of *LNCS*. Berlin: Springer, 151–166.

Ballatore, A., Wilson, D., and Bertolotto, M., 2012c. The similarity jury: combining expert judgements on geographic concepts. *In*: S. Castano, P. Vassiliadis, L. V. Lakshmanan and M. L. Lee, eds. *Advances in conceptual modeling*. *ER 2012 Workshops* (*SeCoGIS*) April 12–13, 2012, Naples, Italy, Vol. 7518 of *LNCS*. Berlin: Springer, 231–240.

Ballatore, A., Wilson, D., and Bertolotto, M., 2013. A survey of volunteered open geo-knowledge bases in the semantic web. *In*: G. Pasi, G. Bordogna and L. Jain, eds. *Quality issues in the management of web information*, Vol. 50 of *Intelligent Systems Reference Library*. Berlin: Springer, 93–120.

Banerjee, S. and Pedersen, T., 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *In:* A. Gelbukh, ed. *Computational linguistics and intelligent text processing*, February 17–23 2002, Mexico City, Mexico, Vol. 2276 of *LNCS*. Springer, 117–171.

Cardoso, N. and Silva, M., 2007. Query expansion through geographical feature types. *In*: R. Purves and C. Jones, eds. *Proceedings of the 4th ACM workshop on geographical information retrieval*, November 6–10, 2007, Lisbon, Portugal, New York: ACM, 55–60.

Corley, C. and Mihalcea, R., 2005. Measuring the semantic similarity of texts. *In*: B. Dolan and I. Dagan, eds. *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, June 30 2005 Ann Arbor, MI: Association for Computational Linguistics, 13–18.

Fellbaum, C, ed., 1998. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Fernando, S. and Stevenson, M., 2008. A semantic similarity approach to paraphrase detection. *In*: *Proceedings of computational linguistics UK (CLUK 2008*), *11th Annual Research Colloquium*, March 2008, Oxford, UK. UK Special Interest Group for Computational Linguistics, 1–7.

Field, A., 2001. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, 6 (2), 161–180.

Goodchild, M., 2007. Citizens as sensors: The world of volunteered geography. *Geo-Journal*, 69 (4), 211–221.

Goodchild, M., Yuan, M., and Cova, T., 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21 (3), 239–260.

Gruber, T., 2007. Ontology of folksonomy: a mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, 3 (1), 1–11.

Harris, Z., 1954. Distributional structure. *Word*, 10 (2–3), 146–162.

Hirst, G. and St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *In*: C. Fellbaum, ed. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press, 305–332.

Islam, A. and Inkpen, D., 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2 (10), 1–25.

Janowicz, K., *et al*., 2007. Algorithm, implementation and application of the SIM-DL similarity server. *In*: *GeoSpatial semantics: second international conference*, *GeoS 2007*, Vol. 4853 of *LNCS*. Mexico City: Springer, 128–145.

Janowicz, K., Raubal, M., and Kuhn, W., 2011. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2 (1), 29–57.

Jiang, J. and Conrath, D., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *In*: *Proceedings of international conference on research in computational linguistics*, *ROCLING X*, August 22–14 1997, Taiwan. Vol. 1, Taiwan: Association for Computational Linguistics and Chinese Language, 19–33.

Jonesh, K., Walker, S., and Robertson, S., 2000. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36 (6), 779–808.

Kavouras, M. and Kokla, M., 2008. *Theories of geographic concepts: ontological approaches to semantic integration*. Boca Raton, FL: CRC Press.

Kavouras, M., Kokla, M., and Tomai, E., 2005. Comparing categories among geographic ontologies. *Computers & Geosciences*, 31 (2), 145–154.

Keßler, C, 2007. Similarity measurement in context. *In*: B. Kokinov, D. C. Richardson, T. R. Roth-Berghofer, and L. Vieu, eds. *Proceedings of the 6th international and interdisciplinary conference on modeling and using context*, August 20–24 2007, Roskilde, Denmark, Vol. 4635 of *LNCS*. Berlin: Springer, 277–290.

Kuhn, W., 2005. Geospatial Semantics: Why, of What, and How?. *In*: *Journal of data semantics III. Special issue on semantic-based geographical information systems*, Vol. 3534 of *LNCS*. Berlin: Springer, 1–24.

Kuhn, W., 2013. Cognitive and linguistic ideas and geographic information semantics. *In*: LNGC *cognitive and linguistic aspects of geographic space*. Berlin: Springer, 159–174.

Landauer, T., *et al*., 2007. *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Leacock, C. and Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. *In*: C. Fellbaum, ed. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press, 265–283.

Lin, D., 1998. An information-theoretic definition of similarity. *In*: *Proceedings of the 15th international conference on machine learning*, 1998, Madison, WI. Vol. 1. San Francisco, CA: Morgan Kaufmann, 296–304.

Marcus, M., Marcinkiewicz, M., and Santorini, B., 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*, 19 (2), 313–330.

Mihalcea, R., Corley, C, and Strapparava, C, 2006. Corpus-based and knowledge-based measures of text semantic similarity. *In*: A. Cohn, ed. *Proceedings of the twenty-first national conference on artificial intelligence*, July 16–20, Boston, Massachusetts. Palo Alto: AAAI Press, Vol. 21, 775–780.

Miller, C, *et al.*, 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3 (4), 235–244.

Patwardhan, S. and Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *In*: *Proceedings of the EACL 2006 workshop making sense of sense – bringing computational linguistics and psycholinguistics together*, April 3–7 2006, Trento Italy, Vol. 1501, The European Chapter of the ACL, 1–8.

Pedersen, T., Patwardhan, S., and Michelizzi, J., 2004. WordNet::Similarity: measuring the relatedness of concepts. *In*: *Proceedings of human language technologies: the 2004 annual conference of the North American chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, 38–41.

Potthast, M., *et al*., 2010. An evaluation framework for plagiarism detection. *In*: *Proceedings of the 23rd international conference on computational linguistics: posters*, 997–1005.

Rada, R., *et al*., 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems*, *Man and Cybernetics*, 19 (1), 17–30.

Ramage, D., Rafferty, A., and Manning, C, 2009. Random walks for text semantic similarity. *In*: *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, 23–31.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. *In*: C. S. Mellish, ed. *Proceedings of the 14th international joint conference on artificial intelligence*, *IJCAI'95* August 20–25 1995, Montreal, Vol. 1. Morgan Kaufmann, 448–453.

Rodríguez, M. and Egenhofer, M., 2004. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18 (3), 229–256.

Schwering, A., 2008. Approaches to semantic similarity measurement for geo-spatial data: a survey. *Transactions in GIS*, 12 (1), 5–29.

Schwering, A. and Raubal, M., 2005. Spatial relations for semantic similarity measurement. *In*: *Perspectives in conceptual modeling*, Vol. 3770 of *LNCS*. Springer, 259–269.

Shirky, C., 2010. *Cognitive surplus: creativity and generosity in a connected age*. London: Penguin.

Sowa, J., 2006. The challenge of knowledge soup. *In*: J. Ramadas and S. Chunawala eds. *Research trends in science, technology and mathematics education*. Mumbai, India: Homi Bhabha Centre, 55–90.

Turney, P., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *In*: *Proceedings of the 12th European conference on machine learning*, *ECML'01*, Vol. 2167 of *LNAI*. Springer, 491–502.

Tversky, A., 1977. Features of similarity. *Psychological review*, 84 (4), 327–352.

Varzi, A., 2001. Vagueness in geography. *Philosophy & Geography*, 4 (1), 49–65.

Wu, Z. and Palmer, M., 1994. Verbs semantics and lexical selection. *In*: *Proceedings of the 32nd annual meeting of the association for computational linguistics*, *ACL-94*, 133–138.