Towards Domain-Specific Semantic Relatedness: A Case Study from Geography

Shilad Sen
Macalester College
ssen@macalester.edu

Isaac Johnson
University of Minnesota
joh12041@umn.edu

Rebecca Harper
Wilamette College
rcharper@willamette.edu

Huy Mai
Brandeis University
huymai@brandeis.edu

Samuel Horlbeck Olsen
Macalester College
shorlbec@macalester.edu

Benjamin Mathers
Macalester College
bmathers@macalester.edu

Laura Souza Vonessen
University of Arizona
lvonessen@email.arizona.edu

Matthew Wright
University of Minnesota
mlwright84@gmail.com

Brent Hecht
University of Minnesota
bhecht@cs.umn.edu

Abstract

Semantic relatedness (SR) measures form the algorithmic foundation of intelligent technologies in domains ranging from artificial intelligence to human-computer interaction. Although SR has been researched for decades, this work has focused on developing *general SR* measures rooted in graph and text mining algorithms that perform reasonably well for many different types of concepts. This paper introduces *domain-specific SR*, which augments general SR by identifying, capturing, and synthesizing domain-specific relationships between concepts. Using the domain of geography as a case study, we show that domain-specific SR — and even geography-specific signals alone (e.g. distance, containment) without sophisticated graph or text mining algorithms — significantly outperform the SR state-of-the-art for geographic concepts. In addition to substantially improving SR measures for geospatial technologies, an area that is rapidly increasing in importance, this work also unlocks an important new direction for SR research: SR measures that incorporate domain-specific customizations to increase accuracy.

1 Introduction

Semantic relatedness (SR) measures help computers understand the strength of relationships between concepts. Due to the broad importance of this task, SR measures have become critical to research and applications in a wide range of domains (e.g. natural language processing, information retrieval, human-computer interaction, spatial computing, bioinformatics, artificial intelligence). SR measures have been studied for decades, with dozens of approaches to SR published (e.g. [Rubenstein and Goodenough, 1965; Gabrilovich and Markovitch, 2007; Strube and Ponzetto, 2006; Halawi *et al*., 2011]). These approaches, however, all draw from the same general family of natural language processing and information retrieval techniques. Specifically, they apply either graph or text mining algorithms to a large repository of general world knowledge (usually WordNet or Wikipedia).

This paper extends these *general* SR approaches by introducing and exploring the notion of *domain-specific SR*, in which domain-specific methods are used in concert with traditional approaches to assess the relatedness of within-domain concepts. More specifically, we ask an important, but unstudied, question: if we focus on a single domain, can we utilize domain-specific techniques (along with traditional SR approaches) to improve SR estimation?

We address this question through a case study in the domain of geography, an area of increasing importance to computer science, as well as to SR-based systems specifically. For instance, 13-15% of search queries contain place names and one-third of all queries have some geographic component [Jones and Purves, 2008; Parsons, 2012; Sterling, 2012]. Thus, search technologies like the Google Knowledge Graph need SR to be highly accurate for geographic queries. Geographic SR also supports the widely used task of geographic named entity disambiguation (NED), also called toponym resolution [Leidner, 2004; Overell and Rüger, 2008; Moncla *et al.*, 2014]. Many NED-reliant systems use SR to, e.g., distinguish mentions of "London" (England) from those of "London" (Ontario, Canada).

In the experiments reported below, we find evidence that *domain-specific approaches to SR can be remarkably effective*. Specifically, we show that a domain-specific *geography-enhanced SR* measure (GESR) that intelligently extends general SR with geography-specific signals (e.g. distance, containment) significantly outperforms the state-of-the-art in general SR for within-domain SR assessment (Spearman's correlation of 0.810 vs. 0.656). We also show that a geography-only SR measure (GOSR) that completely eschews the complex techniques in the SR literature and only uses straightforward geography-specific signals also surpasses the SR state-of-the-art, although by a smaller margin than the hybrid GESR approach.

Our work points to a future SR that unifies both general SR and the wide array of domain-specific semantic relatedness-like approaches that exist in a vast array of domains (e.g. audio signal processing, protein sequence comparison, co-visitation/co-purchasing patterns). Our results suggest that these domain-specific SR approaches may actually outperform the sophisticated graph algorithms, text mining techniques, and other technical approaches that have been developed in the extensive (general) SR literature. Our results also suggest that, when combined together, general SR and these domain-specific approaches can exceed the capability of either approach alone.

Our findings are enabled by a novel gold standard dataset of relatedness estimates for pairs of places (i.e. geographic concepts) that we collected for this paper. These estimates aggregate 23,941 relatedness judgements from 913 people in nine countries. The 754 distinct place pairs were robustly selected to vary in their geographic *class* (country, city, museum, stadium, etc), spatial distance, estimated SR, and familiarity level. This is the first dataset of its kind, and we are releasing it along with a reference implementation of GESR to advance both geography-specific SR research and domain-specific SR research more generally.[1]

To support extending our approach to new domains, in our geographic case study we outline a four-step process for developing a domain-specific SR algorithm. First, we *align* domain-specific datasets with Wikipedia articles. Second, we develop a domain-specific *gold standard* to evaluate SR performance. Third, we extract domain-specific *signals* that correlate with SR and are rooted in domain theories and techniques (e.g. theories and techniques from geography and geographic information science). Finally, we *combine* these signals using machine learning approaches (to develop our GOSR and GESR approaches in our case). As we describe each of the four steps, we highlight key generalizable insights that will assist future SR researchers.

To summarize, this paper makes the following contributions:

1. We introduce a geography-enhanced SR measure (GESR) that significantly outperforms the state-of-the-art for geographic concept pairs by intelligently combining geography-specific signals (e.g. distance, containment) with traditional general SR approaches.
2. We show that a geography-only SR measure (GOSR) also outperforms the current state-of-the-art, but by less of a margin than GESR. GOSR uses only straightforward geography-specific signals, yet is able to more accurately predict SR for geographic concepts than the many complex SR approaches that have been proposed in the extensive SR research.
3. We introduce the notion of *domain-specific SR* more generally, in which domain-specific signals (e.g. distance and containment for geography) are used to assess the relatedness of two concepts in a domain, either in combination with traditional (general) SR or alone. We outline a four-step process future domain-specific SR researchers can follow consisting of dataset alignment, gold-standard development, signal extraction, and signal combination.
4. We release the first gold standard SR dataset consisting of geographic concept pairs (i.e. place pairs). This dataset consists of relatedness assessments from 913 people in nine countries for 754 concept pairs.

[1] https://github.com/shilad/geo-sr

2 Related Work

Semantic relatedness algorithms output a single number (usually between 0 and 1) that summarizes the number and strength of relationships between two concepts [Hecht and Gergle, 2010]. Several extensive meta-reviews of the SR literature have been published [Budanitsky and Hirst, 2006; Zesch and Gureyvch, 2009; Zhang *et al*. 2013], and these provide a detailed overview of the various approaches that have been used to calculate SR.

As noted above, all major existing SR approaches address the problem of *general SR*. That is, they attempt to estimate the number and strength between any two concepts in any domain (rather than a single domain), and they usually do this by applying open-domain graph or text mining algorithms to a large repository of general knowledge. Historically this repository was often WordNet, but in the past decade, Wikipedia has become the repository of choice [Zesch and Gurevych 2010]. While approaches other than text mining and graph algorithms have been proposed - most notably information theoretic approaches (e.g. [Resnick 1995; Pirró and Seco 2008]) operating on ontologies - these approaches also only consider general SR.

Researchers have supported SR applications in specific domains by *applying general SR algorithms to domain-specific sources of world knowledge*. For example, in the biomedical domain, Pederson et al. [2007] adapted WordNet-based SR algorithms to a medical ontology, and Liu et al. [2012] propose an SR algorithm that analyzes word co-occurrence vectors for a pair of terms and their ontological relatives in biomedical knowledge bases. While these approaches are focused on a single domain, they are applications of *general SR*; no domain-specific methods or metrics are utilized. A truly domain-specific biomedical approach might, for example, incorporate numerical measures of the similarity between two drugs' proteins (e.g. DNA sequence alignment) or a geometric comparison of their three-dimensional structures.

The field of geographic information science (GIScience) has also applied general SR algorithms to domain-specific knowledge bases. However, for historical reasons,[2] this work has framed SR as a comparison of *classes* of geographic entities instead of *instances* of those classes. For example, rather than considering *Mississippi River* and *Lake Superior*, this literature has focused on *lake* and *river.* As a

[2] The reasons for this approach are rooted in historically important geographic use cases, such as geographic information retrieval [Jones, Alani, and Tudhope 2001] and the alignment of geographic ontologies [Uitermark *et al.*, 1999].

result, despite the disciplinary origins of these approaches, all existing GIScience SR algorithms eschew explicitly geographic signals and instead rely on the same high-level techniques leveraged in computer science, namely graph algorithms applied to knowledge repositories (see Ballatore et al. [2014] for an overview of this literature). Our work incorporates this class-based approach as one of many signals of relatedness between two geographic entities, but extends this approach with *distance metrics* and *containment relationships* specific to the geographic domain. As a side contribution, we also show that the best Wikipedia-based general SR algorithms from computer science significantly outperform the best existing SR algorithms from GIScience for the traditional class-based GIScience SR problem.

Several SR researchers have analyzed the relationship between estimated SR and distance in Wikipedia (this correlation is predicted by Waldo Tobler's well-known "First Law of Geography" [1970], described below). Hecht and Moxley [2009] found that Wikipedia articles about places at closer distance classes were more likely to link to each other. Li et al. [2014] showed that this result held with more robust general SR measures. Quercini and Samet [2014] similarly propose a general SR approach called *SynRank*, and find that given a spatial target Wikipedia article *A*, the most related other spatial articles *B* frequently fall within a 500km radius of *A*. All three works show that general SR estimates between spatial articles correlate with the spatial distance between articles. Our work furthers this line of research by modeling *human perceptions* of geographic SR using both general SR algorithms and domain-specific geographic signals. Notably, this approach was not possible in prior research because no geographic SR dataset with place pairs existed.

3 Survey Methodology

We used a web-based survey to develop a gold-standard dataset that captures human perceptions of the semantic relatedness between different geographic concepts. The survey gathered *SR assessments* from subjects; a single SR assessment is a relatedness rating between 0 and 4 (inclusive) by a subject for a place pair (e.g. a rating of 4 for (*Great Britain*, *United Kingdom*)). Figure 1 shows the SR rating screen from the survey. As noted above, the use of

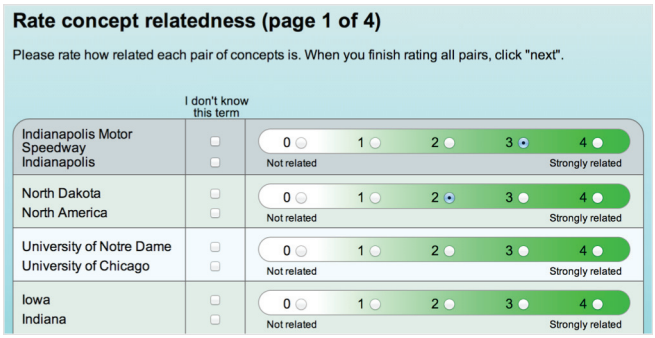

Figure 1: The rating interface workers on Mechanical Turk used to assess the relatedness of concepts.

SR assessments from humans is the most common way in which SR algorithms are evaluated [Budanitsky and Hirst, 2006; Zesch and Gurevych, 2010]. In our data collection process, we followed best practices in the SR literature (e.g. [Pedersen *et al.,* 2007; Radinsky *et al.,* 2011; Halawi *et al*., 2012; Sen *et al.,* 2015]), but adapted them for the task of collecting geographic concepts rather than general concepts.

After consenting to the study, subjects entered basic demographic information (gender, education level) and listed all locations where they had lived for at least one month. These locations were used to estimate place familiarity levels. Next, subjects provided 37 SR assessments. Subjects could indicate that they "did not know a place" instead of providing an SR rating.

Subject recruitment: To encourage a wide variety of geographic perspectives, we surveyed crowdworkers on Mechanical Turk ("Turkers") who live in the nine countries that account for over 90% of Mechanical Turk workers [Ipeirotis, 2010]: United States, Pakistan, India, France, Australia, Spain, Canada, the United Kingdom, and Brazil. To accommodate differences in time zones, we released the study in ten equally spaced intervals throughout the day. All workers had a 98% approval rating and history of at least 1,000 tasks. Following suggested practice (e.g. [Caverlee, 2011]), we made sure to compensate crowdworkers in excess of the active US minimum wage.

Selecting concepts: As candidate concepts, we considered the 3,000 most-viewed geotagged articles in the English Wikipedia. To remove daily variation, we aggregated page views for the 25th day of each of the first five months of 2014. We used this set of concepts because it strikes a balance between diversity of popularity and the likelihood of a subject to recognize a concept. A substantial portion of these concepts were major corporations whose geotags reflected their headquarters. We removed these concepts as they have an ambiguous geospatial interpretation, which left us with 1,985 final candidate concepts. To support the algorithms that follow, we incorporated geographic point representations of places from the Wikidata project, a language-neutral human-editable database of 54M facts about 16.7M concepts (typically Wikipedia articles) [Vrandečić and Krötzsch, 2014]. Polygon representations of countries and first-order administrative districts (e.g. states) come from the NaturalEarth project [Kelso and Patterson, 2009].

Subjects assessed 37 pairs of concepts, with all concepts drawn from the 1,985 candidates. For each subject, we selected a random sample of concept pairs stratified along three dimensions: estimated SR (high, medium, low), spatial distance (within 100km, 500km, beyond) and geographic class (country, state, city, landmark, natural, and other).

To reduce the effects of person-level variation, SR datasets commonly collect responses from five to twenty subjects for each concept pair and average them [Halawi *et al*., 2012]. While this is not difficult for *general knowledge* concept pairs such as (*television, radio*), it is much more challenging to identify ten subjects who are familiar with geography-specific concepts such as (*South India*,

Tiruchirappalli). Thus, a major challenge of our survey was overcoming large individual differences in geographic familiarity while still maintaining a reasonable level of diversity in concepts pairs.

To address this problem, we developed a simple model of geographic familiarity through a pilot survey based on the distance between a concept and the closest location where a person had lived to the concept. In addition, once a specific "new" concept pair was presented for the first time, the survey sought to find additional respondents who were familiar with a concept pair. Thus, the survey introduced more new concept pairs to early respondents (e.g. the first respondent, by definition, rated all "new" concept pairs) and fewer new concept pairs for late respondents (e.g. any "new" concept pairs asked of the last respondent could, by definition, only receive one response). As other domain-specific SR studies may run into similar sparsity issues, this adaptation could prove useful in other domains as well.

Following best practices for Mechanical Turk [Sen *et al*., 2015], the survey also included three *validation concept pairs* that attempted to identify subjects who were not completing the survey in good faith. The concept pairs consisted of one pair that was assumed to be very related (*United Kingdom*, *Great Britain*) and two pairs that were assumed to be very unrelated (*Florida*, *Hong Kong* and *Bermuda Triangle*, *Minnesota*). Respondents needed to answer all three correctly to be included in our dataset.

Basic statistics: Out of the 1,000 survey respondents who provided 36,802 SR assessments, 913 subjects fully completed the survey with correct ratings for the validation concept pairs. Respondents indicated that they were not familiar with 19.7% of responses. Although 1,124 distinct concept pairs were rated by at least one subject, consistent with the existing SR literature's practice of aggregating responses from multiple subjects [Halawi *et al*., 2012], we consider the 754 concept pairs that had at least 10 known responses.[3] To summarize, the rest of this paper analyzes 913 subjects' 23,941 individual responses indicated as familiar. These relatedness assessments covered 754 distinct concept pairs and averaged 1.75 on the survey's [0-4] scale.

Domain-specific SR framework: Above we outlined the first two steps in the framework we proposed for domain-specific SR development. The first step *aligned* domain-specific datasets to Wikipedia articles. Wikidata (and similar projects like DBpedia) offers rich layers of structured facts about Wikipedia articles that can prove valuable to the alignment process. Wikidata exposes third-party identifiers (e.g. geographic FIPS codes, PubChem identifiers for chemicals, ISBNs for books) and other structured information (e.g. containing country for geographic articles, scientific classifications for animals) that can help match datasets to Wikipedia articles. In addition, we used general SR itself to match text in external database records (e.g. NaturalEarth names) to likely articles. Using these two approaches, we created a simple set of rules that performed the alignment with high precision.[4] This approach can be generally applicable in other domains.

This section also described the second step in the framework, which involved collecting a *gold standard* from crowdworkers to train and evaluate domain-specific SR metrics. In addition to following established SR best practices [Pedersen *et al.,* 2007; Radinsky *et al.,* 2011; Halawi *et al*., 2012; Sen *et al.,* 2015], we selected a stratified sample that focused on highly related concept pairs crucial to many real-world applications. We also collected information about each respondent's level of domain expertise by asking their level of *familiarity* with each concept, as it has been shown to substantially affect SR judgements [Sen *et al.,* 2015]. Finally, we used Wikipedia page views statistics to identify a set of candidate concepts that struck a balance between being reasonably well known and diverse. By following these practices, domain-specific SR researchers can develop robust gold standards that meet the needs of real-world applications.

4 Signals for Geospatial SR

This section describes all relatedness signals used in this paper. We first introduce the geography-specific signals we use in the GOSR (geography-only SR) and GESR (hybrid geography-enhanced SR) models and we highlight the theoretical motivation for each signal. We next discuss our implementation of state-of-the-art general SR, as well as our implementation of the type of general SR studied in the geography community.

Containment (*contain*): Previous research has suggested that human geographic perceptions of relatedness incorporate containment relationships [Janowicz *et al.,* 2015]. As such, we encoded geospatial containment relationships among the three most prominent classes of spatial entities in our data: countries, states, and points of interest (POIs). Table 1 shows all possible containment relationships (c = contains; dc = does not contain). We define the relationship such that a class cannot contain itself (e.g. a point cannot contain a point) and, as such, relationships along the diagonal are only "dc". The lower right cells are empty because concept A and concept B can be swapped if the scale of B is larger than A (e.g. B is a country and A is a point). The nine possible relationships in Table 1 are encoded using nine binary dummy variables.

Great-arc distance (*arc*): Tobler's First Law of Geography (TFL) states that "everything is related to

		Concept *B*		
		POI	**state**	**country**
Concept *A*	**country**	c / dc	c / dc	dc
	state	c / dc	dc	
	POI	dc		

Table 1: The 9 containment classes used by the containment metric. Concepts A either contains concept B (c) or does not contain concept B (dc) .

[3] Other thresholds for minimum number of respondents (e.g. 5 and 20) did not change our results meaningfully.

[4] We manually verified all NaturalEarth polygon alignments.

everything else, but near things are more related than distant things" [Tobler, 1970]. TFL predicts that spatial entities that are closer together on the surface of the Earth will be more related. Thus, we include as a signal the great-arc distance between the point representations of two places (i.e. the distance between them in meters).

Ordinal distance (*ordinal*): While TFL predicts that closer places are in general more related, the spatial heterogenity of human populations (i.e. the variations in population density across the surface of the Earth) is thought to serve as a moderating factor in this relationship (e.g. [Li *et al.,* 2014]). Consider two concepts *A* and *B* that are 100 km apart. One might expect *A* and *B* to be more related for a concept *A* located in an extremely rural location (e.g. Patagonia) compared to an *A* located in a highly urban area (e.g. Buenes Aires). We model these relationships through a non-parametric *ordinal* distance metric. Given concepts *A* and *B*, the ordinal distance from *A* to *B* is the rank of *B* in *A*'s distance-ordered neighbor list. For example, if *B* is the 10th nearest neighbor to *A*, *ordinal*(A, B) = 10.

Countries between (*countries*): A large body of social science literature has provided strong evidence of the reduction of cultural, religious, and economic ties that occurs due to geopolitical boundaries such as country and province borders (i.e. 'border effects') (e.g. [Fellman *et al.*, 2007; Singh and Marx, 2012]). This suggests that, in aggregate, two concepts that lie in the same country will be more related than two concepts that do not. We operationalize this signal by determining the number of countries that separate two geographic concepts. If concepts *A* and *B* lie in the same country, *countries*(*A*, *B*) = 0, and if *A* and *B* lie in neighboring countries, *countries*(*A*, *B*) = 1. In other cases, the shortest country path between two concepts is used.

States between (*states*): Same as countries, but for first-order general administrative districts (e.g. states).

General SR (*general-SR*): To capture the state-of-the-art in general SR, we created a machine-learning ensemble that combines several well-known general SR approaches. Each approach draws upon Wikipedia as a knowledge base, mapping a concept pair *(A, B)* to two articles *(C, D)*, and then examining the relationship between *C* and *D*'s category structure, link graphs, and text. We used the implementations of each SR algorithm provided by Sen et al.'s WikiBrain toolkit[5] [2014] combined in a linear ensemble to achieve state-of-the-art performance.[6] The SR ensemble includes: (1) Milne and Witten's [2008] SR algorithm, which calculates the overlap in the links to and from *A* and *B*, (2) Strube and Ponzetto's *WikiRelate* [2006], which measures the shortest distance between *A* and *B* in Wikipedia's category graph, and (3) Gabrilovich and Markovitch's *Explanatory Semantic Analysis* [2007], which assesses the relationship between articles that mention *A* and *B* using a text mining-style approach.

Geographic class-level general SR (*class-SR*): We include an improved approach to the traditional class-based general SR methods (Section 2) developed in GIScience. As noted above, existing GIScience approaches estimate *SR(Mississippi River, Lake Superior)* using geographic classes: *general-SR(river, lake)*. While GIScience has primarily applied ontology-based general SR algorithms, we applied the general SR ensemble described above, and found it yielded better results (Spearman's correlation of 0.813 vs 0.737) on the most recent class-level Geographic SR dataset [Ballatore *et al.* 2013, 2014]. To our knowledge, this is the first attempt to apply modern Wikipedia-based SR algorithms to geographic class-based SR. To incorporate this feature into our analysis, we assigned classes to all 754 concepts using a procedure similar to Ballatore *et al*. [2014].[7]

Additional implementation details: We log transformed the four distance metrics (arc, ordinal, countries, states) because they exhibited right-skewed distributions. All features and metrics exhibited 100% coverage for the 754 concept pairs except for countries-between (96.8% coverage) and states-between (94.4%) due to the nature of continents and the oceans that surround them. For missing data, we impute the maximum values for each distance metric.

Domain-specific SR framework: The above section described the third step of domain specific SR development, which *extracted* domain-specific *signals*. Our case study introduced new signals that drew upon geographic theory. While some domains may take this theory-driven approach, many domains offer existing similarity or association metrics. For example, DNA similarity could enhance SR for protein sequences in the bioinformatics domain just as we used containment and distance in the geography domain. The same could apply to existing metrics that compare beats per minute, instrumentation, or chordal structure in the domain of music analysis, and metrics that compare the visual elements of movies in the film domain, and so on.

5 Results

5.1 Individual features and metrics

To evaluate each individual signal, we adopted the SR community's standard practice of calculating Spearman's rank correlation over all 754 concept pairs between the output of general SR, domain specific signals, and the average human assessment for the pair. Table 2 shows the correlation matrix between the gold-standard (human), the

[5] http://wikibrainapi.org

[6] The linear ensemble achieved a Spearman's correlation of 0.76 on WordSim353 [Finkelstein *et al.*, 2001]. This slightly lower correlation compared to some other published results is the result of our use of cross-validation in evaluation (not commonly used in SR), which yields lower, but more robust results.

[7] As in [Ballatore *et al.*, 2013], we assigned geographic classes to each place from OpenStreetMap's Semantic Network. OSM's search tool Nominatim was used to provide an initial value for the class. We manually verified the results and adjusted the classes when no OpenStreetMap feature existed, an incorrect feature was returned by the initial search, or the initial class was not listed in the OSM Semantic Network.

	human	general-SR	class-SR	arc	ordinal	countries	states
human	1.000	0.656	0.052	-0.543	-0.643	-0.212	-0.608
general-SR	0.656	1.000	0.373	-0.356	-0.456	-0.205	-0.426
class-SR	0.052	0.373	1.000	0.145	-0.126	-0.051	-0.105
arc	-0.534	-0.356	0.145	1.000	0.860	0.326	0.738
ordinal	-0.643	-0.456	0.126	0.860	1.000	0.242	0.733
countries	-0.212	-0.205	-0.051	0.326	0.242	1.000	0.411
states	-0.608	-0.426	0.105	0.738	0.733	0.411	1.000

Table 2: Correlation between signals of geographic relatedness.

two existing SR approaches (general-SR and class-SR) and the four distance metrics (arc, ordinal, countries, states).

The correlations between human SR assessments and each approach (row 1) provide important insights into the predictive power of each signal. While general SR exhibits the strongest correlation with human judgements (σ_s =0.656), both states (σ_s =-0.608) and ordinal (σ_s =-0.643) nearly match the widely established approach. Interestingly, arc, the mostly widely-used measure of geographic distance, shows lower correlation with human judgements than the density-sensitive ordinal and the boundary-aware states. This finding provides the strongest evidence thus far that custom domain-specific SR algorithms that incorporate relatively simple domain-specific relationships can substantially enhance SR performance.

Surprisingly, class-SR, the traditional GIScience approach to SR, shows the *lowest* correlation with human judgements of all the approaches we consider (σ_s =0.052). While this does suggest that the standard GIScience approach has limited applicability outside of the class-based SR problem, this finding deserves additional study. For example, class-based SR may be more effective for rare (or "tail") concepts that have little general knowledge describing them.

5.2 Geography-only and Geography-enhanced SR

Table 3 lists the accuracy of machine learning ensembles combining the six signals defined earlier. We used gradient-boosted trees [Ganjisaffar, Caruana, and Lopes 2011] as implemented in the scikit-learn machine learning library with seven-fold cross-validation. However, we note that a much simpler multiple linear regression also showed similar results with small (0.01 to 0.02) decreases in correlations compared to gradient-boosted trees. The paired non-parametric approach to SR significance testing described in Sen et al. [2015] finds that differences in σ_s >= 0.02 between all pairs of ensembles are significant at $p < 0.05$ and differences of σ_s >= 0.04 are significant at $p < 0.01$.

	σ_s	general SR	class SR	Domain-Specific SR					
				ordinal	contain	states	arc	countries	
1	0.656	X							**SR**
2	0.693			X	X	X	X		**GOSR**
3	0.748		X	X	X	X	X	X	**N/A**
4	0.743	X		X					**GESR**
5	0.772	X		X	X				
6	0.788	X		X	X	X			
7	0.801	X		X	X	X	X	X	
8	**0.810**	X	X	X	X	X	X	X	

Table 3: Accuracy of different SR, GESR, and GOSR algorithms.

General SR, with a correlation of 0.656 (row 1), is significantly outperformed by both geography-only SR (GOSR, rows 2 and 3, σ_s=0.693 or 0.748), and geography-enhanced SR (GESR, row 8, σ_s=0.810). Row 2 shows that GOSR – which incorporates none of the complex approaches used in modern SR and only uses geographic signals like distance and containment – outperforms state-of-the-art SR (σ_s=0.693 v. 0.656). Including class-SR in the ensemble raises GOSR's performance to 0.748. These findings point to the power of domain-specific SR approaches alone, showing that a combination of domain-specific signals can achieve a correlation that is substantially better than sophisticated general SR baselines.

Rows 4 through 8 show the stepwise results for building GESR using a forward search that iteratively adds the strongest feature. The single addition of the ordinal distance metric to general SR increased the correlation from general SR's 0.656 to 0.743. Interestingly, the addition of the ordinal distance metric was significantly more powerful than the addition of the arc distance metric, which only boosted correlation to 0.723 (not shown in table). Further signal additions showed incremental improvements, yielding an accuracy of σ_s = 0.810 for the full GESR model.

Figure 2 provides a higher-level view of GESR by showing a simplified descriptive multiple regression model that achieves an accuracy of σ_s = 0.77. The terms on the left show regression coefficients for the numeric geographic signals in the model (states, ordinal, and general SR), while the table shows values for the nine levels of the categorical containment variable introduced in Table 1. For example, consider the relationship between concept *A=Japan* and two possible concepts B={*Kyoto*, *Buenos Aires*} (upper left cell in table). The containment offset for (*Japan*, *Kyoto*) will be +1.68 (upper diagonal), while the containment offset for (*Japan*, *Buenos Aires*) will be -0.27. The effects of POI containment shows a strong positive SR signal at both the state level (+0.77) and country level (+1.68). Overall, we see that many predictions from geographic theory (boundary effects, distance, and containment) hold true in this context.

5.3 Descriptive Comparison of SR Algorithms

To provide deeper insights into the differences between general and domain-specific SR, we analyzed the concept pairs that exhibited the biggest differences between general SR and GESR. Table 4 shows 10 exemplars from the 20

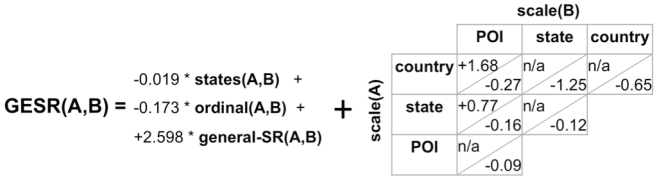

Figure 2: Simplified descriptive multiple-regression GESR model. The lefthand terms show coefficients for three distance metrics, while the righthand table shows containment coefficents.

	Entity 1	Entity 2	Human	GESR	SR
1.	Ajanta Caves	India	85.6	**94.0**	4.3
2.	Epcot	Florida	94.3	**83.4**	26.2
3.	Juilliard School	Queens	42.8	**49.6**	1.7
4.	Bill Gates's house	Mount Rainier	31.8	**42.7**	4.5
5.	Louisiana	Oklahoma	28.7	**37.1**	73.9
6.	Oman	Sri Lanka	0.5	**35.9**	72.8
7.	Leicester	Southampton	37.2	**51.2**	89.0
8.	Ohio	Rock and Roll Hall of Fame	82.3	**78.7**	22.4
9.	Oberlin College	Ohio	73.1	**83.2**	46.1
10.	Oberlin College	Rock and Roll Hall of Fame	9.3	48.1	**11.2**

Table 4: Concept pairs that show the greatest difference between domain-specific (GESR) and general SR estimates.

pairs with largest disagreements. We normalized numerical values by converting them to within-column percentiles. GESR outperforms general-SR for 18 of the 20 pairs — often substantially so. Inspecting the list, it appears that the domain-specific methods we have developed can avoid significant errors experienced by the general approach.

Several pairs clearly indicate GESR's ability to understand containment relationships at the country and state level (rows 1, 2, 8, 9). General SR also appears to overestimate the relatedness of entities in the same class. For example, it predicts both Louisiana v Oklahoma (both U.S. states), and Oman v Sri Lanka (both countries) to be related. Domain-specific SR incorporates the geographic distance between these entities, deeming them (like humans) to be mostly unrelated. Interestingly, this result suggests that general SR may overestimate the importance of the class-based relationship targeted by traditional Geographic approaches to (general) SR.

Rows 8, 9, and 10, which compare the state Ohio, the Rock and Roll Hall of Fame (located in Cleveland, Ohio) and Oberlin College (located in Oberlin, Ohio), provide insight into the strengths and weaknesses of both algorithms. Humans (and GESR) seem to identify the containment relationship between Ohio and the two landmarks as important, while general SR does not. However, despite the spatial proximity of Oberlin College and the Rock and Roll Hall of Fame (approximately 50 km), humans deem them to be largely unrelated, confusing our domain-specific algorithm. This example may point to differences in performance for places that are considered *primarily geographic* (e.g. states, countries, cities, etc.) and those that have other dominant characteristics (e.g. museums, educational institutions, athletic stadiums, etc). While GESR performs better for concepts that are perceived as "primarily geographic", general SR's focus on non-geographic features may allow it to better understand other places. This data point suggests that one area for improvement lies in algorithms that capture and use the perceived level of "domain specificity" for a concept.

5.4 Domain-Specific SR Framework

The above section described the fourth and final step of our domain-specific SR development process, which used machine learning techniques to *combine* domain-specific and general SR signals into a final SR metric using machine learning best practices. We specifically compared general SR, domain-only SR (GOSR), and domain-enhanced SR (GESR). We also identified a minimal set of signals that approached maximum performance. These points of comparison will assist application developers seeking to understand the costs and benefits associated with different implementation choices for domain-specific SR.

6 Discussion and Conclusion

This paper introduces domain-specific SR and, through a case study in geography, shows that it can significantly outperform state-of-the-art general approaches to SR. Indeed, the new domain-specific geographic signals we introduce significantly outperform state-of-the-art general SR algorithms *by themselves*, without the use of any of the traditional SR techniques (e.g. graph and text mining algorithms), although best performance is achieved when traditional (general) SR and domains-specific SR are combined. To support future research in geographic SR and domain-specific SR more broadly, we have released a new evaluation dataset[8] that contains SR judgements from 917 participants in nine countries about 754 distinct concept pairs — the first dataset of its kind.

While we focused on the domain of geography, our domain-specific approach offers promise for many other domains. The four steps we proposed and examined (dataset alignment, gold-standard development, signal extraction, signal combination) should support future domain-specific SR development. A critical next step in domain-specific SR involves re-running our experiments in a number of new domains. This work will be important for understanding the variation in improvement one can obtain with domain-specific SR. Our geographic analyses serve as a proof-of-concept case study, albeit one that has applied value through its consideration of a domain that is significant for SR-based systems.

One major potential benefit of domain-specific SR not considered here is that domain-specific signals are not dependent on a general knowledge repository's (e.g. Wikipedia's) coverage of a given content area. As such, an SR measure that uses only domain-specific signals like GOSR will be able to calculate SR equally accurately for well-known concepts (i.e. places) and very obscure concepts (i.e. places), which is not true of state-of-the-art general SR. This makes domain-specific SR potentially preferable for applications that frequently consider "long tail" concepts. In addition, state-of-the-art general SR is subject to *biases* in the content coverage of their underlying knowledge repositories (e.g. along cultural and gender lines [Hecht and Gergle, 2010; Lam *et al.* 2011]), something that may be more avoidable in domain-specific approaches like GOSR. Future work should examine these phenomena more closely.

[8] https://github.com/shilad/geo-sr

Our work also has implications specifically for the many GIScience applications that model relationships between objects. We show that the best Wikipedia-based general SR algorithms outperform the best GIScience general SR algorithms on the *class-based* SR problem historically studied in GIScience. More critically, we move beyond this class-based approach to introduce (and provide an evaluation dataset for) the problem of geographic concept SR. This problem, following the computer science SR literature, focuses on class instances rather than classes.

7 Acknowledgements

We would like to thank Toby Li for sharing his expertise on geographic SR signals, Andrew Beveridge for his programatic support of the MAXIMA research experience for undergraduates (REU) program, and the crowdworkers on Mechanical Turk that provided SR ratings. This research was generously supported by the Institute for Mathematics and its Applications, the National Science Foundation (IIS-0964697, DMS 0931945, IIS-1421655, DMS-1156701), a University of Minnesota College of Science and Engineering Graduate Fellowship, science education award 52007550 to Macalester College from the Howard Hughes Medical Institute, and an Individual Engagement Grant from the Wikimedia Foundation.

References

[Ballatore *et al*., 2013a] Andrea Ballatore, Michela Bertolotto, and David C. Wilson. 2013. The semantic similarity ensemble. *Journal of Spatial Information Science* , 7 (December 2013).

[Ballatore *et al.*, 2013b] Andrea Ballatore, David C. Wilson, and Michela Bertolotto. 2013. A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. *Quality Issues in the Management of Web Information*. Springer, Berlin Heidelberg, 93–120.

[Budanitsky and Hirst, 2006] Alexander Budanitsky and Graeme Hirst. 2006. "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness." Computational Linguistics 32 (1): 13–47.

[Caverlee 2011]. "Exploitation in Human Computation Systems." *Handbook of Human Computation*. Ed. Pietro Michelucci. Springer New York, 2013. 837–845. *link.springer.com*.

[Fellmann *et al.*, 2007] Jermoe D. Fellmann, Arthur Getis, and Judith Getis. 2007. Human Geography. 9th ed. McGraw-Hill.

[Lev Finklestein *et al.*, 2001] Lev Finkelstein et al. 2001. Placing Search in Context: The Concept Revisited. In WWW '01. ACM, 406–414.

[Gabrilovich and Markovitch, 2007] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI* (2007).

[Ganjisaffar et al., 2011] Yasser Ganjisaffar, Rich Caruana, and Cristina Videira Lopes. 2011. Bagging Gradient-boosted Trees for High Precision, Low Variance Ranking Models. In *SIGIR* (2011).

[Halawi *et al.,* 2012] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale Learning of Word Relatedness with Constraints. In *KDD* (2012).

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (November 2009), 10–18.

[Hecht and Gergle, 2010] Brent Hecht and Darren Gergle. 2010. "The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context." In *CHI* (2010).

[Hecht and Moxley, 2009] Brent Hecht and Emily Moxley. 2009. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. In *Spatial Information Theory*. Springer, Berlin Heidelberg, 88–105.

[Ipeirotis, 2010] Panagiotis G. Ipeirotis. 2010. *Demographics of Mechanical Turk*, Rochester, NY: Social Science Research Network.

[Janowicz *et al.,* 2015] Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. 2015. The semantics of similarity in geographic information retrieval. *Jour. of Spatial Information Science* 0, 2 (Aug 2015), 29–57.

[Jones *et al.*, 2001] Christopher B. Jones, Harith Alani, and Douglas Tudhope. 2001. Geographical Information Retrieval with Ontologies of Place. In Daniel R. Montello, ed. *COSIT 2011*.

[Keslo and Patterson, 2009] Nathaniel Vaughn Kelso and Tom Patterson. 2009. Natural Earth Vector. *Cartographic Perspectives*, 64 (2009).

[Lam, S.K., A. Uduwage, Z. Dong, S. Sen, D.R. Musicant, Loren Terveen, and John Riedl. 2011. "WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance." In *WikiSym '11* 1–10.

[Leidner, 2004] Leidner, Jochen L. 2004. "Toponym Resolution in Text: 'Which Sheffield Is It?.'" *SIGIR* (2004).

[Li *et al.*, 2014] Li, Toby Jia-Jun, Shilad Sen, and Brent Hecht. 2014. "Leveraging Advances in Natural Language Processing to Better Understand Tobler's First Law of Geography." *SIGSPATIAL* (2014).

[Liu *et al.*, 2012] Ying Liu, Bridget T. McInnes, Ted Pedersen, Genevieve Melton-Meaux, and Serguei Pakhomov. 2012a. Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet. IHI (2012).

[Milne and Witten, 2008] David Milne and Ian H. Witten. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *AAAI (*2008*)*.

[Moncla *et al.*, 2014] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. "Geocoding for Texts with Fine-Grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus." In ACM SIGSPATIAL '14. Dallas, TX.

[Overell and Rüger, 2008] Simon Overell and Stefan Rüger. 2008. "Using Co-occurrence Models for Placename Disambiguation." International Journal of Geographical Information Science 22 (3): 265–87.

[Parsons, 2012] Ed Parsons. 2012. "Using Location Data for Business: Past, Present, and Future." https://www.youtube.com/watch?v=ucYiMBfyNfo.

[Pirro and Seco, 2008] Giuseppe Pirró and Nuno Seco. 2008. "Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content." In On the Move to Meaningful Internet Systems: OTM 2008. Springer .

[Pedersen *et al*., 2007] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* 40, 3 (June 2007), 288–299.

[Quercini and Samet, 2014] Giancula Quercini and Hanan Samet. 2014. "Uncovering the Spatial Relatedness in Wikipedia." In ACM SIGSPATIAL '14. Dallas, TX.

[Radinsky *et al*., 2011] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. "A Word at a Time: Computing Word Relatedness Using Temporal Semantic Analysis." In WWW '11: 20th International Conference on World Wide Web, 337–46. Hyberabad, India.

[Resnik, 1995] Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI* (1995).

[Rubenstein and Goodenough, 1965] Herbert Rubenstein and John B. Goodenough. 1965. "Contextual Correlates of Synonymy." Commun. ACM 8 (10): 627–33. doi:10.1145/365628.365657.

[Sen *et al*., 2014] Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. WikiBrain: Democratizing Computation on Wikipedia. OpenSym (2014).

[Sen *et al*., 2015] Shilad Sen, Matthew Lesicko, Margaret Giesel, Rebecca Gold, Benjamin Hillman, Samuel Naden, Jesse Russell, Zixaio Wang, and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. *CSCW* (2015).

[Singh and Marx, 2012] Jasjit Singh and Matt Marx. 2012. *Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity*, Rochester, NY: Social Science Research Network.

[Strube and Ponzetto, 2006] Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI '06*. AAAI Press, 1419–1424.

[Tobler, 1970] Waldo Tobler. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography 46: 234–40.

[Uitermark *et al*., 1999] Harry T. Uitermark, Peter J.M. van Oosterom, Nicolaas J.I. Mars, and Martien Molenaar. 1999. Ontology-Based Geographic Data Set Integration. *Spatio-Temporal Database Management*. Springer, 60–78.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." Communications of the ACM 57 (10): 78–85.

[Zhang *et al*., 2013] Ziqi Zhang, Anna Lisa Gentile, and Fabio Ciravegna. 2013. "Recent Advances in Methods of Lexical Semantic Relatedness – a Survey." Natural Language Engineering 19 (04): 411–79. doi:10.1017/S1351324912000125.

[Zesch and Gurevych, 2009] Torsten Zesch and Iryna Gurevych. 2009. "Wisdom of Crowds versus Wisdom of Linguists – Measuring the Semantic Relatedness of Words." Natural Language Engineering 16 (1): 25–59