The semantic similarity ensemble*

Andrea Ballatore¹, Michela Bertolotto¹, and David C. Wilson²

¹School of Computer Science and Informatics, University College Dublin, Ireland
²Department of Software and Information Systems, University of North Carolina, USA

Received: December 3, 2012; returned: June 20, 2012; revised: August 16, 2013; accepted: September 18, 2013.

Abstract: Computational measures of semantic similarity between geographic terms provide valuable support across geographic information retrieval, data mining, and information integration. To date, a wide variety of approaches to geo-semantic similarity have been devised. A judgment of similarity is not intrinsically right or wrong, but obtains a certain degree of cognitive plausibility, depending on how closely it mimics human behavior. Thus selecting the most appropriate measure for a specific task is a significant challenge. To address this issue, we make an analogy between computational similarity measures and soliciting domain expert opinions, which incorporate a subjective set of beliefs, perceptions, hypotheses, and epistemic biases. Following this analogy, we define the semantic similarity ensemble (SSE) as a composition of different similarity measures, acting as a panel of experts having to reach a decision on the semantic similarity of a set of geographic terms. The approach is evaluated in comparison to human judgments, and results indicate that an SSE performs better than the average of its parts. Although the best member tends to outperform the ensemble, all ensembles outperform the average performance of each ensemble’s member. Hence, in contexts where the best measure is unknown, the ensemble provides a more cognitively plausible approach.

Keywords: semantic similarity ensemble, SSE, lexical similarity, semantic similarity, ensemble modeling, geo-semantics, expert disagreement, WordNet

1 Introduction

The importance of semantic similarity in geographical information science (GIScience) is widely acknowledged [21]. As diverse information communities generate increasingly

*This article extends work presented at the 6th International Workshop on Semantics and Conceptual Issues in Geographical Information Systems (SeCoGIS 2012) [7]
large and complex geo-datasets, semantics play an essential role to constrain the meaning of the terms being defined. The automatic assessment of the semantic similarity of terms, such as river and stream, enables practical applications in data mining, geographic information retrieval, and information integration.

Research in natural language processing and computational linguistics has produced a wide variety of approaches, classifiable as knowledge-based (structural similarity is computed in expert-authored ontologies), corpus-based (similarity is extracted from statistical patterns in large text corpora), or hybrid (combining knowledge and corpus-based approaches) [29, 32]. Several similarity techniques have been tailored specifically to geographic information [35]. In general, a judgment on semantic similarity is not simply right or wrong, but rather shows a certain degree of cognitive plausibility, i.e., a correlation with human behavior. Hence, selecting the most appropriate measure for a specific task is non-trivial, and represents in itself a challenge.

From this perspective, a semantic similarity measure bears resemblance with a human expert being summoned to give her opinion on a complex semantic problem. In domains such as medicine and economic policy, critical choices have been made in uncertain, complex scenarios. However, disagreement among experts occurs very often, and equally credible and trustworthy experts can hold divergent opinions about a given problem [25]. To overcome decisional deadlocks, an effective solution consists of combining diverse opinions into a representative average. Instead of identifying a supposedly “best” expert in a domain, an opinion is gathered from a panel of experts, extracting a representative average from their diverging opinions [10]. Similarly, complex computational problems in machine learning are often tackled with ensemble methods, which achieve higher accuracy by combining heterogeneous models, regressors, or classifiers [34]. This idea was first explored in our previous work under the analogy of the similarity jury [7].

Rather than developing a new measure for geo-semantic similarity, we explore the idea of combining existing measures into a semantic similarity ensemble (SSE). In order to gain insight about the merits and limitations of the SSE, we conducted a large empirical evaluation, selecting ten WordNet-based similarity measures as a case study. The ten measures were combined into all of the possible 1,012 ensembles, exploring the entire combinatorial space. To measure the cognitive plausibility of each measure and ensemble, a set of 50 geographic term pairs including 97 unique terms, selected from OpenStreetMap and ranked by 203 human subjects, was adopted as ground truth. The results of this evaluation confirm that, in absence of knowledge about the performance of the similarity measures, the ensemble approach tends to provide more cognitively plausible results than any individual measure.

The remainder of this paper is organized as follows. Section 2 reviews relevant related work in the areas of geo-semantic similarity and ensemble methods. Section 3 describes the WordNet-based similarity measures selected as a case study. The SSE is defined in Section 4, while Section 5 presents and discusses the empirical evaluation. Finally, Section 6 draws conclusions about the SSE, and indicates directions for future work.

2 Related work

The ability to assess similarity between stimuli is considered a central characteristic of human psychology. Hence, it should not come as a surprise that semantic similarity is
widely studied in psychology, cognitive science, and natural language processing. Over
the past ten years, a scientific literature on semantic similarity has emerged in the context of
GIScience [4, 6, 17]. Schwering [35] surveyed and classified semantic similarity techniques
for geographic terms, including network-based, set-theoretical, and geometric approaches.
Notably, Rodríguez and Egenhofer [33] have developed the matching-distance similarity
measure (MDSM) by extending Tversky’s set-theoretical similarity for geographic terms. In
the area of the semantic web, SIM-DL is a semantic similarity measure for spatial terms ex-
pressed in description logic (DL) [16]. As these measures are tailored to specific formalisms
and data, we selected WordNet-based measures as a more generic case study (see Section
3).

A key element in this article is the combination of different semantic similarity mea-
sures, relying on the analogy between computable measures and domain experts. The idea
of combining divergent opinions is not new. Indeed, expert disagreement is not an excep-
tional state of affairs, but rather the norm in human activities characterized by uncertainty,
complexity, and trade-offs between multiple criteria [25]. As Mumpower and Stewart [26]
put it, the “character and fallibilities of the human judgment process itself lead to persistent
disagreements even among competent, honest, and disinterested experts” (p. 191). From
a psychological perspective, in cases of high uncertainty and risk (e.g., choosing medical
treatments and long term investments), decision makers consult multiple experts, and try
to obtain a representative average of divergent expert judgments [10]. In the context of
risk analysis, mathematical and behavioral models have been devised to elicit judgments
from experts, suggesting that simple mathematical methods such as the average perform
quite well [11]. The underlying intuition has been controversially labeled as “wisdom of
crowds,” and can account for the success of some crowdsourcing applications [37].

In complex domains such as econometrics, genetics, and meteorology, ensemble methods
aggregate different models of the same phenomenon, trying to overcome the limitations
of each model. In the context of machine learning, a wide variety of ensemble methods
have been devised and evaluated [34]. Such methods aim at generating a single classifier
from a set of classifiers applied to the same problem, maximizing its overall accuracy and
robustness [27]. Similarly, clustering ensembles obtain a single partitioning of a set of ob-
jects by aggregating several partitionings returned by different clustering techniques [36].
In computational biology, ensemble approaches are currently being used to compute the
similarity of proteins [19].

Forecasting complex phenomena can also benefit from ensemble methods. Armstrong
[2] pointed out that “combining forecasts is especially useful when you are uncertain about
the situation, uncertain about which method is most accurate, and when you want to avoid
large errors” (p. 417). Notably, a study of the Blue Chip Economic Indicators survey indi-
cates that forecasts issued by a panel of seventy economists tended to outperform all the
seventy individual forecasts [9]. To date, we are not aware of studies that explore system-
atically the possibility of combining semantic similarity measures through an ensemble
method. The next section describes in detail the similarity measures that we selected as a
case study.
3 WordNet similarity measures

In this study, we selected WordNet-based semantic similarity measures as a case study for our ensemble technique, the semantic similarity ensemble (SSE). In the context of natural language processing, WordNet [13] is a well-known knowledge base for the computation of semantic similarity. Numerous knowledge-based approaches exploit its deep taxonomic structure for nouns and verbs [8,22,32,38]. From a geo-semantic viewpoint, WordNet terms have been mapped to OpenStreetMap [5]. Table 1 summarizes the salient characteristics of ten popular WordNet-based measures. In order to compute the similarity scores, each measure adopts a different strategy. Seven measures rely on the shortest path between terms in the noun/verb taxonomy, assuming that the number of edges is inversely proportional to the similarity of terms. This approach is limited by the variability in the path lengths in the different semantic areas of WordNet, determined by arbitrary choices and biases of the knowledge base’s owners. Paths in dense, well-developed parts of the taxonomy tend to be longer than those in shallow, sparse areas, making the direct comparison of term pairs from different areas problematic. Missing edges between terms make the score drop to 0.

To overcome these limitations, three measures include the information content of the two terms and that of the least-common subsumer, i.e., the more specific term that is an ancestor to both target terms (e.g., [32]). Hence, at the same path length, terms with a very specific subsumer (“building”) are considered to be more similar than terms with a generic subsumer (“thing”). Although this approach mitigates the issues of the shortest paths, a new issue lies in the extraction of the information content from a text corpus. Text corpora tend to be biased towards specific semantic fields, underestimating the specificity of terms contained in those fields, resulting in skewed similarity scores. An alternative approach that do not rely on taxonomy paths consists of comparing the term glosses, i.e., the lexical definition of terms. Definitions can be compared in terms of word overlap (terms that are defined with the same words tend to be similar), or with co-occurrence patterns in a text corpus (terms that are defined with co-occurring words tend to be similar) [28]. The results of this approach are sensitive to noise in the definitions (e.g., very frequent or rare words that skew the scores), and to the arbitrary nature of definitions, which can be under- or over-specified.

Empirical research suggests that the performance of these measures largely depends on the specific ground-truth dataset utilized in the evaluation [24]. Therefore, these measures constitute a striking example of alternative models of the same phenomenon, none of which can be considered to be uncontroversially better than the others. Each measure is sensitive to specific biases in the knowledge base, and tends to reflect these biases in the similarity scores. For this reason, we consider these measures to be a suitable case study for the ensemble approach, formally defined in the next section.

4 The semantic similarity ensemble (SSE)

A computable measure of semantic similarity can be seen as a human domain expert summoned to rank pairs of terms, according to her subjective set of beliefs, perceptions, hypotheses, and epistemic biases. When the performance of an expert can be compared against a gold standard, it is a reasonable policy to trust the expert showing the best per-
performance. Unfortunately, such gold standards are difficult to construct and validate, and the choice of most appropriate expert remains highly problematic in many contexts. To overcome this issue, we propose the semantic similarity ensemble (SSE), a technique to combine different semantic similarity measures on the same set of terms. This ensemble of measures can be intuitively seen as a jury or a panel of human experts deliberating on a complex case [7]. Formally, the similarity function \( sim \) quantifies the semantic similarity of a pair of geographic terms \( t_a \) and \( t_b \): \( sim(t_a, t_b) \in [0, 1] \). Set \( P \) contains all term pairs whose similarity needs to be assessed, while set \( M \) contains a set of selected semantic similarity measures from which the ensembles will be formed:

\[
P = \{ \langle t_{a1}t_{b1} \rangle, \langle t_{a2}t_{b2} \rangle \ldots \langle t_{an}t_{bn} \rangle \}
\]
\[
M = \{ sim_1, sim_2 \ldots sim_m \}
\]

A measure \( sim \) from \( M \) applied to \( P \) maps the set of pairs to a set of scores \( S_{sc} \), which can then be converted into rankings \( S_{rk} \) from the most similar (e.g., stream and river) to the least similar (e.g., stream and restaurant):

\[
sim(P) \rightarrow S_{sc} = \{ s_1, s_2 \ldots s_n \} \quad s \in \mathbb{R}_{\geq 0}
\]
\[
rank(S_{sc}) \rightarrow S_{rk} = \{ r_1, r_2 \ldots r_n \}
\]

For example, a measure \( sim \in M \) applied to a set of three pairs \( P \) might return \( S_{sc} = \{ .45, .13, .91 \} \), corresponding to rankings \( S_{rk} = \{ 2, 3, 1 \} \). The rankings \( S_{rk}(P) \) can be used to assess the cognitive plausibility of \( sim \) against a human-generated rankings \( H(P) \). The cognitive plausibility of \( sim \) can be estimated with the Spearman’s correlation \( \rho \in [-1, 1] \).
between $S_{rk}(P)$ and $H_{rk}(P)$. If $\rho$ is close to 1 or -1, $sim$ is highly plausible, while if $\rho$ is close to 0, $sim$ shows no correlation with human behavior.

In this context, a SSE is defined as a set $E$ of unique semantic similarity measures:

$$E = \{sim_1, sim_2 \ldots sim_k\}, \ \forall j \in \{1, 2 \ldots k\} : sim_j \in M$$

$$\forall i \in \{1, 2 \ldots |M| - 1\} : sim_i \neq sim_{i+1}, \ \ k \leq m, |E| \leq |M|$$

For example, considering the ten measures in Table 1, ensemble $E_a$ has two members $\{jcn, lesk\}$, while ensemble $E_b$ has three members $\{jcn, res, wup\}$.

Several techniques have been discussed to aggregate rankings, using either unsupervised or supervised methods. Clemen and Winkler [11] stated that simple mathematical methods, such as the average, tend to perform quite well to combine expert judgments in risk assessment. Hence, we define two aggregation approaches $A$ to compute the rankings of ensemble $E$:

(1) mean of the similarity scores: $A_s = rank(mean(S_{sc1}, S_{sc2} \ldots S_{scn}))$; and
(2) mean of the similarity rankings: $A_r = rank(mean(S_{rk1}, S_{rk2} \ldots S_{rkn}))$.

The first approach, $A_s$, combines directly the similarity scores, while the second approach flattens the scores into equidistant rankings. Rankings contain less information than scores: for example, scores $\{.01, .02, .98, .99\}$ and $\{.51, .52, .53, .54\}$ have very different distributions, but result in the same rankings $\{1, 2, 3, 4\}$. For this reason, in some cases, $A_s \neq A_r$.

If two measures on five term pairs generate the scores $S_{sc1} = \{.9, .9, .38, .44, .31\}$ and $S_{sc2} = \{.28, .47, .14, .61, .36\}$, the resulting $A_s$ is $\{4.5, 1.3, 2\}$, whilst $A_r$ is $\{3, 5, 1, 4, 2\}$.

A given similarity measure has a cognitive plausibility, i.e., the ability to approximate human judgment. A traditional approach to quantify the cognitive plausibility of a measure consists of comparing rankings against a human-generated ground truth [14]. The ranked similarity scores are compared with the rankings or ratings returned by human subjects on the same set of term pairs. Following this approach, we define $\rho_{sim}$ as the correlation of an individual measure $sim$ (i.e., an ensemble of size one) with human-generated rankings $H_{rk}$, while $\rho_E$ is the correlation of the judgment obtained from an ensemble $E$.

When knowledge of $\rho_{sim}$ is available for the current task, the optimal $sim \in M$ can be simply the $sim$ having highest $\rho_{sim}$. However, in real settings this knowledge is often absent, or incomplete, or unreliable. The same semantic similarity measure can obtain considerably different degrees of cognitive plausibility based on the specific dataset in consideration. In such contexts of limited information, the SSE offers a viable alternative to an arbitrary selection of a $sim$ from $M$. The empirical evidence discussed in the next section supports this claim.

5 Evaluation

This section discusses an empirical evaluation conducted on the SSE in real settings. The purpose of this evaluation is to assess the performance of the SSE in detail, highlighting strengths and weaknesses. Ten semantic similarity measures are tested on a set of pairs of geographic terms utilized in OpenStreetMap. A preliminary evaluation of an analogous technique on a small scale was conducted in [7]. Ensembles of cardinalities 2, 3, and 4 were generated from eight similarity measures, for a total of 154 ensembles. The evaluation described below is conducted on a larger scale, adopting a larger set of geographic
terms, ranked by 203 human subjects as ground truth. To obtain a complete picture of ensemble’s performance, the entire combinatorial space is considered, for a total of 1,012 unique ensembles. The remainder of this section outlines the evaluation criteria by which the performance of the SSE is assessed (Section 5.1), the human-generated ground truth (Section 5.2), the experiment set-up (Section 5.3), and the empirical results obtained, including a comparison with the preliminary evaluation (Section 5.4).

5.1 Evaluation criteria

The performance of an ensemble $E$ is measured on its cognitive plausibility $\rho_E$, with respect to the plausibility of its individual members $\rho_{sim}$. Intuitively, an ensemble succeeds when it provides rankings that are more cognitively plausible than those of its members. Four criteria are formally defined in this evaluation:

- **Total success.** The plausibility of the ensemble is strictly greater than all of its members: $\forall sim \in E : \rho_E > \rho_{sim}$.
- **Partial success.** The plausibility of the ensemble is strictly greater than a member: $\exists sim \in E : \rho_E > \rho_{sim}$.
- **Success over mean.** The plausibility of the ensemble is strictly greater than the mean plausibility of its members: $\rho_E > \text{mean}(\rho_{sim_1}, \rho_{sim_2}, \ldots, \rho_{sim_n})$.
- **Success over median.** The plausibility of the ensemble is strictly greater than the median plausibility of its members: $\rho_E > \text{median}(\rho_{sim_1}, \rho_{sim_2}, \ldots, \rho_{sim_n})$.

5.2 Ground truth

In order to assess the cognitive plausibility of the similarity measures and the ensembles, a human-generated ground truth has to be selected. In the preliminary evaluation described, a human-generated set of similarity rankings was extracted from an existing dataset [7]. That dataset contains similarity rankings of 50 term pairs, containing 29 geographic terms, originally collected by Rodríguez and Egenhofer [33], and is available online.¹ In order to provide a thorough assessment of the SSE in the present article, a new and larger human-generated dataset was adopted as ground truth.

As part of a wider study on geo-semantic similarity, we selected 50 pairs of geographic terms commonly used in OpenStreetMap, including 97 man-made and natural features. The terms were subsequently mapped to the corresponding terms in WordNet, as exemplified in Table 2. A Web-based survey was subsequently prepared on the set of 50 term pairs, asking human subjects to rate the pairs’ similarity on a five-point Likert scale, from very dissimilar to very similar. In order to be understandable by any native speaker of English, regardless of knowledge of the geographic domain, the survey only included common and non-technical terms, aiming to collect a generic set of geo-semantic judgments. The survey was disseminated online through mailing lists, and obtained valid responses from 203 human subjects. The subjects’ ratings for each pair were normalized on a [0, 1] interval and averaged, obtaining human-generated similarity scores $H_{sc}$, then ranked as $H_{rk}$. Table 3 outlines a sample of term pairs, with the similarity score and ranking assigned by the 203 human subjects. This dataset was utilized as ground truth in the experiment outlined in the next section.

¹See http://github.com/ucd-spatial/Datasets
Table 2: Sample of the 97 terms extracted from OpenStreetMap and mapped to WordNet.

<table>
<thead>
<tr>
<th>Term</th>
<th>OpenStreetMap tag</th>
<th>WordNet synset</th>
</tr>
</thead>
<tbody>
<tr>
<td>bay</td>
<td>natural=bay</td>
<td>bay#n#1</td>
</tr>
<tr>
<td>canal</td>
<td>waterway=canal</td>
<td>canal#n#3</td>
</tr>
<tr>
<td>city</td>
<td>place=city</td>
<td>city#n#1</td>
</tr>
<tr>
<td>post box</td>
<td>amenity=post box</td>
<td>postbox#n#1</td>
</tr>
<tr>
<td>floodplain</td>
<td>natural=floodplain</td>
<td>floodplain#n#1</td>
</tr>
<tr>
<td>historic castle</td>
<td>historic=castle</td>
<td>castle#n#2</td>
</tr>
<tr>
<td>motel</td>
<td>tourism=motel</td>
<td>motel#n#1</td>
</tr>
<tr>
<td>supermarket</td>
<td>shop=supermarket</td>
<td>supermarket#n#1</td>
</tr>
</tbody>
</table>

Table 3: Human-generated similarity scores ($H_{sc}$) and rankings ($H_{rk}$) on 50 term pairs.

<table>
<thead>
<tr>
<th>Term A</th>
<th>Term B</th>
<th>$H_{sc}$</th>
<th>$H_{rk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>motel</td>
<td>hotel</td>
<td>.90</td>
<td>1</td>
</tr>
<tr>
<td>public transport station</td>
<td>railway platform</td>
<td>.81</td>
<td>2</td>
</tr>
<tr>
<td>stadium</td>
<td>athletics track</td>
<td>.76</td>
<td>3</td>
</tr>
<tr>
<td>theatre</td>
<td>cinema</td>
<td>.87</td>
<td>4</td>
</tr>
<tr>
<td>art shop</td>
<td>art gallery</td>
<td>.75</td>
<td>5</td>
</tr>
<tr>
<td>water ski facility</td>
<td>office furniture shop</td>
<td>.05</td>
<td>46</td>
</tr>
<tr>
<td>greengrocer</td>
<td>aqueduct</td>
<td>.03</td>
<td>47</td>
</tr>
<tr>
<td>interior decoration shop</td>
<td>tomb</td>
<td>.05</td>
<td>48</td>
</tr>
<tr>
<td>political boundary</td>
<td>women’s clothes shop</td>
<td>.02</td>
<td>49</td>
</tr>
<tr>
<td>nursing home</td>
<td>continent</td>
<td>.02</td>
<td>50</td>
</tr>
</tbody>
</table>

5.3 Experiment setup

To explore the performance of an SSE versus individual measures, we selected a set of ten WordNet-based similarity measures as a case study. Table 4 summarizes the resources involved in this experiment. The ten similarity measures were not applied directly to the term pairs, but they were applied to the their lexical definitions, using a paraphrase-detection technique [3]. In order to explore the space of all the possible ensembles, we considered the entire range of ensemble sizes $|E| \in \{2, 3 \ldots 10\}$ for $M$. The entire power set of $M$ was computed. Increasing the ensemble cardinality from 2 to 10, respectively 45, 120, 210, 252, 210, 120, 45, 9, 1 ensembles were generated, for a total 1,012 ensembles. The experiment was carried out through the following steps:

1. Compute $S_{sc}$ and $S_{rk}$ for each of the ten measures on the 50 term pairs from OpenStreetMap.
2. Generate 1,012 ensembles, combining the measures on either similarity scores ($E_s$) or rankings ($E_r$).
3. For each of the ten measures, compute the cognitive plausibility $\rho_{sim}$ against human-generated rankings $H_{rk}$.
4. For each of the 1,012 ensembles, compute the cognitive plausibility $\rho_{E}$ against $H_{rk}$.
5. Compute the four evaluation criteria (total success, partial success, success over mean, success over median) for each measure and ensemble.

2The WordNet:Similarity tool [29] was used to compute the similarity scores.

www.josis.org
10 similarity measures \( \text{sim} \in \mathcal{M} \): \{jcn, lch, hso, lesk, lin, path, res, vector, vectorp, wup\} (see Table 1)

9 ensemble cardinalities \(|\mathcal{E}|\): \{2, 3, 4, 5, 6, 7, 8, 9, 10\}

Number of unique ensembles \(\mathcal{E}\): \{45, 120, 210, 252, 210, 120, 45, 9, 1\}; Total: 1,012

2 types of ensembles \(\mathcal{E}\): ensemble of scores \(\mathcal{E}_s\) and ensemble of rankings \(\mathcal{E}_r\)

Ground truth: 50 term pairs ranked by 203 human subjects by semantic similarity

4 evaluation criteria:
(a) total success; (b) partial success;
(c) success over mean; (d) success over median

Table 4: Experiment setup.

5.4 Experiment results

The experiment was carried out on two types of ensemble, once with \(\mathcal{A}_s\) (mean of scores), and once with \(\mathcal{A}_r\) (mean of rankings). These two approaches obtained very close results, with a slightly better performance for \(\mathcal{A}_r\), with each evaluation criterion always within a 5% distance from \(\mathcal{A}_s\). To avoid repetition, only cases with \(\mathcal{A}_r\) are included in the discussion. All the cognitive plausibility correlations obtained statistically significant results at \(p < .01\).

The experiment results are summarized in Table 5, showing the cognitive plausibility of each measure, and the four evaluation criteria across all the ensemble cardinalities. For example, the ensembles of cardinality 2 containing measure \(\text{wup}\) obtains partial success in 86.1% of the cases. The cognitive plausibility of the ten measures are in the range \(\rho \in [.562, .737]\), where \(\text{vector}\) is the best measure, and \(\text{lin}\) the worst. Whilst both total and partial success change considerably and are fully reported, the success over mean and median obtain homogeneous results and only the means are included in the table. The general trends followed by the evaluation criteria are depicted in Figure 1.

**Total success.** The total success for the 1,012 ensembles falls in the interval \([0, 55.6]\) percent, with a mean of 9.7%. On average, small cardinalities (2 and 3) obtain the best total success rate (≈ 25%). As the cardinality increases, the total success decreases rapidly, dropping below 10% with cardinality greater than 4. This makes sense intuitively, as the larger the ensemble, the less likely the ensemble can outperform every single member. The total success varies across the different measures too, falling in the interval \([3.4, 15.9]\). No statistically significant correlation exists between a measure’s cognitive plausibility and its rate of total success. In other words, ensembles containing the best measures do not necessarily have better or worse total success rate. Although ensembles do not tend to outperform all of their members, the plausibility of an ensemble is never lower than that of all of its members, \(\exists \text{sim} \in \mathcal{E} : \rho_{\mathcal{E}} > \rho_{\text{sim}}\).

**Partial success.** Partial success rate is considerably greater than that of total success. Over the entire space of ensembles, the partial success rate varies widely between 0% and 100%, with a global average of ≈ 70%. The ensembles’ cardinality has no clear impact on the mean partial success rate, which remains in the interval \([0.2, 71.7]\) both with small and large ensembles. Unlike total success, partial success rate is affected by each measure’s cognitive plausibility \(\rho_{\text{sim}}\). The top measures in \(\mathcal{M}\) (\(\text{vector}\), \(\text{lch}\), and \(\text{path}\)) obtain low partial success (< 20%), whereas ensembles consistently outperform the bottom measures (100%).
Table 5: Overall results of the experiment, including cognitive plausibility $\rho_{\text{sim}}$, and the four evaluation criteria. Succ. mean: success over mean; Succ. med.: success over median.

Figure 1: The four evaluation criteria w.r.t. ensemble cardinality.
The average partial success rates bear strong inverse correlation with the measures’ plausibility, i.e., $\rho = -0.87$ ($p < .05$). Ensembles tend to outperform the worst measures, and tend to be outperformed by the top measures. The total and partial success of each measure is displayed in Figure 2. We note that the three top measures do not benefit from being aggregated within the ensemble, whereas all the others do. While in this experiment a ground truth is given, in many real-world settings the best measures are unknown, and therefore the SSE constitutes a viable alternative to the arbitrary selection of a measure. In particular, ensembles of cardinality 3 obtain optimal results over other cardinalities.

**Success over mean and median.** Unlike total and partial success, the success of ensembles over the mean and median of their members’ plausibilities is consistent. All 1,012 ensembles obtain higher plausibility than the mean of their members’ plausibilities (100%). Similarly, 98% of the ensembles are more plausible than the median of their members’ plausibilities. Hence, an ensemble is more than the mean (or the median) of its parts. In order to quantify more precisely the advantage of the ensembles over the mean of their members’ plausibilities, we computed the difference between the ensemble’s plausibility $\rho_E$ and the mean (or median) of all the $\rho_{\text{sim}}$, where $\text{sim} \in E$. On average, the ensembles’ plausibility is .042 higher than the mean of their members (+4.2%), and .046 over the median (+4.6%). Figure 3 depicts the advantage of the ensemble in terms of cognitive plausibility over mean and median, with respect with the cardinality of the ensemble. The advantage is directly proportional to the ensemble’s size, i.e., the larger the ensemble, the larger the improvement over mean and median.

In other words, by combining the rankings, the ensemble reduces the weight of individual bias, converging towards a shared judgment. Such shared judgment is not necessarily the best fit in absolute terms, but tends to be more reliable than most individual judgments.
Comparison with preliminary experiment. To further assess the SSE, the empirical evidence described above can be compared with the preliminary evaluation we conducted in [7], discussing their commonalities and differences. That evaluation included only eight of the ten WordNet-based similarity measures, on ensembles of cardinality 2, 3, and 4, called similarity juries. These measures and ensembles were compared against an existing similarity dataset, originally collected by Rodríguez and Egenhofer [33]. The salient characteristics of the two evaluations are summarized in Table 6. The comparison of the two evaluations reveals that the same general trends are observable across the board. The total success of the current evaluation appears to be lower than in the preliminary evaluation, and this is because the current evaluation includes larger ensembles, which tend to have lower total success than the small ensembles of cardinality smaller than 5. On average, the partial success rates are very similar in both evaluations (≈ 70%). The success over mean is very high in both evaluations, consistently falling between from 93% to 100%.

Although the mean plausibility of the measures is consistent across the two evaluations, the relative performances of the individual measures vary widely. Notably, the measure jcn is the most plausible measure in the preliminary evaluation, while being the second-last in the current evaluation. Similarly, vector is the top measure in the current evaluation, and ranks among the worst in the preliminary evaluation. By contrast, leh, wup, and lesk maintain almost the same relative position in terms of cognitive plausibility. The two sets of plausibilities do not show any statistically significant correlation (Spearman’s $\rho \approx .1$). Although the measures fall within a similar range in both evaluations, it is difficult to identify measures that are always optimal or inadequate. These results confirm the difficulty of identifying optimal semantic similarity measures, suggesting that the SSE offers a way to proceed in a context of limited and uncertain information.

Figure 3: Improvement in cognitive plausibility of the ensembles over the mean and median of their members’ plausibility.

www.josis.org
Table 6: Comparison between the preliminary evaluation in [7] and the evaluation in this article.

6 Conclusions

In this paper we have outlined, formalized, and evaluated the semantic similarity ensemble (SSE), a combination technique for semantic similarity measures. In the SSE, a computational measure of semantic similarity is seen as a human expert giving a judgment on the similarity of two given pairs. Like human experts, similarity measures often disagree, and it is often difficult to identify unequivocally the best measure for a given context. The ensemble approach is inspired by findings in risk management, machine learning, biology, and econometrics, which indicate that analysts that aggregate expert opinions from different experts tend to outperform analyses from single experts [2, 11, 34]. Based on empirical results collected on WordNet-based similarity measures in the context of geographic terms, the following conclusions can be drawn:

- An ensemble $E$, whose members are semantic similarity measures, is generally less cognitively plausible than the best of its members, i.e., $\max(\rho_{\text{sim}}) > \rho_E$. In $\approx 9\%$ of cases, the ensemble obtains total success, i.e., it outperforms the most plausible measure. The larger the ensemble, the less frequently the ensemble outperforms its best member.
- On average, similarity ensembles $E$ tend to be more cognitively plausible than any of their individual measures $\text{sim}$ in isolation (mean of partial success ratio $\approx 70\%$). In our evaluation, ensembles with 3 members are the most successful.
- The SSE confirms what Cooke and Goossens [12] pointed out in the context of risk assessment: “a group of experts tends to perform better than the average solitary expert, but the best individual in the group often outperforms the group as a whole” (p. 644).
- In the vast majority of cases ($\geq 98\%$), the cognitive plausibility of an SSE is higher than the mean and median of its members’ plausibilities. An ensemble is more plausible than the mean (or median) of its parts. These results are overall consistent with a preliminary evaluation [7].
Individual similarity measures obtain widely different cognitive plausibility on different ground truths and contexts. In a context of limited information in which the optimal measure is unknown, we believe that the SSE should be favored over any individual similarity measure.

Several issues should be considered for future work. This study focused exclusively on ten WordNet-based similarity measures and, to gather more empirical evidence, the ensemble approach should be extended to different similarity measures. Moreover, to aggregate the similarity scores, we have adopted two simple ensemble methods (the mean of scores and the mean of rankings). More sophisticated ensemble techniques based on machine learning could be explored to increase the ensemble’s performance [31]. Furthermore, the empirical evidence presented in this paper was limited to the geographic context. General-purpose semantic similarity datasets, such as that devised by Agirre et al. [1], could be used to further evaluate the ensemble across various semantic domains.

The evaluation utilized in this study is based on ranking comparison, which allows to quantify the cognitive plausibility of semantic similarity measure directly. Although this approach is the most popular in the literature, it has several drawbacks, as extensively discussed by Ferrara and Tasso [14]. Alternatively, task-based evaluations could be used to assess the cognitive plausibility of measures indirectly by observing their ability to support a specific task. Suitable tasks in geographic information retrieval and natural language processing, such as geographic query expansion, could be devised and deployed to evaluate the SSE further. In this study, similarity is modeled as a continuous score, but it can also be represented as a set of discrete classes. More importantly, the evaluation discussed in this article focuses on acontextual judgments of similarity of geographic terms. Context, however, has been identified as a crucial component of similarity [20], and the SSE should extended to capture specific facets of the observed terms. The effectiveness of the ensemble should be assessed when observing either the affordances, the size or the physical structure of geospatial entities.

The importance of semantic similarity measures in information retrieval, natural language processing, and data mining can hardly be underestimated [17, 21]. In this article, we have shown that a scientific contribution can be given not only by devising new similarity measures, but also by studying the combination of existing measures. The SSE provides a general approach to obtain more cognitively plausible results in settings where the ground truth is unstable and shifting.

Acknowledgments

The research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References


www.josis.org


www.josis.org


