Journal of Intelligent Information Systems manuscript No.
(will be inserted by the editor)

Semantic Similarity Aggregators for Very Short Textual Expressions: A Case Study on Landmarks and Points of Interest

Jorge Martinez-Gil

Received: date / Accepted: date

Abstract Semantic similarity measurement aims to automatically compute the degree of similarity between two textual expressions that use different representations for naming the same concepts. However, very short textual expressions cannot always follow the syntax of a written language and, in general, do not provide enough information to support proper analysis. This means that in some fields, such as the processing of landmarks and points of interest, results are not entirely satisfactory. In order to overcome this situation, we explore the idea of aggregating existing methods by means of two novel aggregation operators aiming to model an appropriate interaction between the similarity measures. As a result, we have been able to improve the results of existing techniques when solving the GeReSiD and the SDTS, two of the most popular benchmark datasets for dealing with geographical information.

Keywords Knowledge Engineering · Data Integration · Semantic Similarity Measurement

1 Introduction

The computation of semantic similarity has traditionally been considered an important method in many areas of computer research since methods of this kind are of vital importance for successfully addressing a number of complex problems [19]. Automatically determining a similarity score for a pair of text expressions based on their real meaning is a problem that attracts many researchers from many scientific fields. In our particular case, the automatic assessment of the semantic resemblance of landmarks and points of interest

Software Competence Center Hagenberg GmbH
Softwarepark 21
4232, Hagenberg, Austria
Tel: +43 7236 3343 838
E-mail: jorge.martinez-gil@scch.at

(POIs), such as *lake* and *loch*, facilitates a number of practical applications in geographic information retrieval [21], or schema and ontology integration [16]. Therefore, automatic learning to identify the most appropriate semantic similarity measures (ssm) can be considered as a key research challenge [38].

In recent times, the vast amount of geospatial information available on the Web has made such as functionality as searching, and recommending landmarks and POIs has become a major challenge for researchers from this field. As this huge amount of information is produced to be consumed by people, the methods and tools working with information of this kind should positively correlate with human judgments [14].

In our particular case, it is important to note that semantic similarity measurement has direct implication in many problems concerning urban systems or natural environments. For example, location-based services offer landmarks and POIs by analyzing previously visited places [23]. Semantic similarity is also applied to perform data integration between different geospatial sources [6]. In this way, the more resources of this kind become available, the higher is the need for appropriate methods and tools being able to deal with them [15]. In fact, many applications dealing with geographical information require the support of some kind of semantic similarity measurement [9].

For this reason, and despite the fact that it is possible to find an important number of works concerning semantic similarity in the literature; if we discard all those approaches just addressing simple words, and also those works addressing sentence or paragraph similarity, none or little attention have been paid to achieve optimal solutions for facing the problem of dealing with very short textual expressions. In consequence, it is necessary to design robust measures to face scenarios like this, whereby the coverage of landmarks and POIs is one of the most representative use cases. Our working hypothesis is that by using some novel aggregation operators (a.k.a. aggregators), it is possible to leverage the results of existing methods and achieve interesting insights towards this goal. In fact, the major highlights of this work can be summarized as follows:

- The proposal of a novel operator that is able to face the non-stochastic uncertainty of natural language using an aggregation approach built following the principles of fuzzy logic.
- The proposal of another aggregation operator based on artificial neural mechanisms for the aggregation of ssm from the geospatial field. This operator uses a gradient descent approach that handles the subjectivity and imprecision associated with natural language in a proper manner.
- The evaluation of these two novel aggregators using the GeReSiD dataset [5]. This benchmark dataset includes human judgments about 50 pairs of terms and covers a unique set of landmarks and POIs.
- The evaluation of the two novel aggregators using a well-known dataset generated from the Spatial Data Transfer Standard (SDTS) [36], which is intended to easily share geographical data on a wide range of different computer systems.

The remainder of this work is organized in the following way: Section 2 reports the state-of-the-art on geospatial similarity measurement with a special focus on the context of landmarks and POIs. Section 3 presents our novel operators for the aggregation of simple ssm. Section 4 reports the empirical evaluation of our aggregation operators and the analysis of the results that we have achieved from their evaluation. Finally, we outline the conclusions and future lines of research.

2 State-of-the-art

In the past, many researchers from a wide range of research fields have proposed a vast amount of new ssm: n-grams, wordnet, semantic analysis, etc. [7]. This is mainly due to its key role in many application-oriented disciplines from the information technology fields [20]. One of the key aspects is that these computational techniques are usually exploited for handling textual representations which enable effective representation of linguistic items at multiple levels, from word senses to full-text .

In this context, it is very important to distinguish between semantic similarity and relatedness; while two expressions that are similar have to be always related, the opposite is not true. For example, *filling station* and *fuel* are two expressions that are highly related, but that are far from having the same meaning. In this work, and mainly due to our focus on aspects associated to data integration, we deal with the semantic similarity of landmarks and POIs, but there are also some works that have addressed the problem of relatedness [3] since it has direct implications in a number of disciplines such as geospatial query expansion and recommendation.

In addition, when focusing on our case study, it is possible to observe that there are intrinsic issues associated with the geospatial information that makes the problem different from the rest [12]; one of these issues is related to the fact that dealing with semantic similarity is not a homogeneous problem, and it usually involves any of these three clearly differentiated cases:

1. The first case is related to the assessment of semantic similarity for simple words, such as those collected in the Miller & Charles dataset [30] (automobile-car, gem-jewel, and so on). In this context, there are works proving that by using some methods (e.g. information content, path calculation, etc.) over a dictionary such as WordNet being able to automatically assess the semantic similarity between two terms with very good results [35].
2. The second case is related to the assessment of the semantic similarity between sentences or paragraphs, i.e. entities that have a full structure. In cases like these, there are also a number of solutions that try to search for similar words in the two sentences [2]. If the methods are able to identify a high degree of common features between them, then it is quite possible that both sentences and paragraphs are semantically similar. For example, the Word Mover's Distance (WMD) computes the distance between two

textual expressions as the cumulative sum of them minimum distance that each term in the source textual expression has to move to the closest one in the target textual expression [22].

3. Finally, there is another case involving very short textual expressions for which it is not usually possible to use dictionaries (since the meaning of words where together might alter the overall meaning of the complete expression) nor search for common features. For example, such case as *public transport station* and *railway platform* implies that there is no dictionary, no sentence structure nor sufficient overlapping words that can give clues when it comes to solving the problem. As a result, methods for simple words or sentences cannot properly tackle the problem.

However, there are similarity measures relying on distributional semantics, i.e. those measures that extend the representation of each term with the more likely words to appear with it. These kinds of measures have proven to be able to successfully handle short texts [40]. These methods can handle word combinations that do not appear in dictionaries, but can be found recurrently in large text corpora. Our hypothesis is that we could use them when facing the problem of using ssm for dealing with very short textual expressions, in special with landmarks and POIs. To do that, it is possible to choose among three major families of methods that are able to exploit distributional semantics:

- *MCS methods* This family of methods implements a greedy strategy that tries to identify the largest substructure the two expression pairs to be compared have in common [37].
- *LSA methods* are methods to extract and represent the meaning of words through statistical calculations over large corpora. The rationale behind this idea is that those contexts in which a particular word appears can provide a series of reciprocal restrictions, which largely resemble the meaning of the words [18].
- *UMBC methods* provide two prevailing approaches to compute semantic similarity, based on either using of a thesaurus (e.g., WordNet) or statistics from a large text corpus [11].

Concerning the aggregation of different information sources leading to take shared decisions, it is possible to find a great corpus of technical literature mainly due to the importance of dealing with pools of heterogeneous sources that should find a way to provide single results [34][42]. In fact, most of these works belong to different categories: multi-expert in neural-fuzzy networks, ensemble learning, and deep learning among others.

The work that we present here clearly belongs to the first category since we wish to avoid the risks of using a single similarity measure in exploitation settings. Relevant works in this context are Hsu and Chen present a method for aggregating individual fuzzy opinions into a group fuzzy consensus opinion is presented. To do that, authors firstly define the index of consensus of each expert to the other experts using a similarity measure. Then, they aggregate the experts using the index of consensus and the importance of each expert

[13]. Kuncheva presents a method for classifier fusion for continuous-valued individual classifier outputs that mitigate the risk of using just one classifier [17], and Medina et al. proposed to extend a similarity function for appropriately processing vectors to calculate the membership grades of the input measures using a fuzzy neural network [29]. However, our approach is the first attempt to use aggregators for improving the assessment of semantic similarity in the context of landmarks and POIs.

3 Semantic similarity in the geospatial field

A ssm can be defined as a function that maps the likeness of two textual representations into real value in the range [0, 1]. In this way, the value 0 informs about not overlapping features between the two textual representations to be compared, and the value 1 means that all distinguishing features are common [25]. This is mainly due to the fact semantic similarity judgment is not always right or wrong, but it obtains a certain degree of plausibility, depending on how it reflects the human way of thinking [4].

Current approaches on semantic similarity measurement are usually following an approach based on the aggregation of similarity scores retrieved from a number of different ssm, i.e. aggregation methods try to accurately aggregate different viewpoints to come to a final decision [10]. As consequence, some authors have proposed some similarity aggregation techniques that have achieved good results in the past [8]. The rationale behind this approach is very intuitive; if there are some ssm not being able to perform reasonably well for a particular comparison, their effects on the overall performance can be blurred by other ssm being able to provide better results [27]. Therefore, we can define aggregators (shorter form for aggregation operators) as mathematical functions being able to reduce a set of numeric inputs into a meaningful number reflecting a proper interaction between those inputs.

The great advantage of aggregators is that they can work with methods that might differ on the characteristics and the data they can work with. In this way the most widely used configurations are those covering different a wide range of features and background resources ranging from dictionaries to large document collections. In fact, a number of different approaches aiming to aggregate the results of distant methods have been proposed. It is possible to see that many of them have succeeded by overcoming the traditional problems from simple ssm [27]. In the next subsections, we explain how we have designed our fuzzy and neural aggregators, and how these aggregators are able to overcome the limitations of traditional methods.

3.1 Designing a fuzzy aggregator

The great advantage of fuzzy operators is that it allows assigning relevance to sets of ssm and not only to individual measures as it happens for other measures such as the weighted mean. Our fuzzy operator is inspired on the idea of ssm have to be aggregated by not considering dissident scores in case of a common consensus can be achieved or trying to impose a compromise in case of a common recommendation from the ssm to be aggregated is not possible. Operators like this have shown a good performance when solving cases from other application domains [26].

An operator like this is usually implemented by the design of three different stages: fuzzification, reasoning, and defuzzification.

For the process of fuzzification is necessary to define the membership functions, i.e. the way that the real values will be transformed into a function that can be used to reason with. In this case, we have chosen to define the membership function as traditional trapezoids in the form:

$$m(h; x_1, x_2, x_3, x_4) = max\left(min\left(\frac{h - x_1}{x_2 - x_1}, 1, \frac{x_4 - h}{x_4 - x_3}\right), 0\right)$$

The fuzzy reasoning is going to be implemented using rules. These rules take a number of variables as input and produce a unique output that it is going to be a real number between 0 and 1. This output is just the consensus from the different input ssm (if any) or a compromise between two or more choices stating different values. In this case, the final value needs to be calculated by the fusion of their semantic similarity score [27]. Such an operator follows these properties:

1. *Idempotence:* $f(x, x, ..., x) = x$
2. *Neutral element:* $f^n(x_1, ..., e, ..., x_{n-1}) = f^{n-1}(x_1, ..., x_{n-1})$
3. *Associativity:* $f(x_1, x_2, x_3) = f(f(x_1, x_2), x_3) = f(x_1, f(x_2, x_3))$
4. *Symmetry:* $f(x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(n)}) = f(x_1, x_2, ..., x_n)$
5. *Non Absorbent Element:* $f(x_1, ..., a, ..., x_n) \neq a$
6. *Pareto compensation:* $\min_{i=1}^{n}(x_i) \leq f(x_1, x_2, ..., x_n) \leq \max_{i=1}^{n}(x_i)$

Where (1) guarantees determinism, (2) monotone extendibility, (3) and (4) independence of the implementation, (5) avoid the possibility of veto, and (6) an output dependent of the input.

In order to set up the parameters for the fuzzy operator, it is necessary to study many aspects such the overlapping between the membership functions, thresholds, etc. in order to decide whether a pair of landmarks or POIs could be considered semantically equivalent or not [28]. This configuration can be set up in an initial parametric study. A study of this kind consists of optimizing the configuration in order to increase the chances of reaching our goal, which in this case consist of replicating the behavior of experts for all the expression pairs to be compared. In this particular scenario, we want to set up the operator so that the comparisons might be performed with a minimum amount of errors. It could be also optimized for maximizing the number of correct predictions, but this might penalize the overall performance, and therefore, it is not so interesting for our purposes.

Concerning the inverse process, i.e. getting the real number that represents the result from the aggregation process, we need a method that can generate a score for representing the resulting fuzzy set. To do that, we have chosen the Center of Gravity (CoG) method since it does not give any preference to any membership functions. This method can be expressed as follows:

$$CoG = \frac{\sum_{x=a}^{b} \mu_A(\chi) x}{\sum_{x=a}^{b} \mu_A(\chi)}$$

In this case, the CoG is the center of the area of the fuzzy set (centroid), and uses the value at which this occurs as the final aggregated score.

3.2 Designing a neural aggregator

Research on neural networks was initially inspired by the purpose of modeling the brain and other kinds of neural systems, but as collateral effect, researchers realized that neural networks can also be successfully applied in the completion of many computational tasks. Neural network operators are based on the working mode of cells called neurons. In this way, a neuron is a kind of biological artifact that has several inputs that can be activated by external processes. Depending on the degree of activation, the neuron produces its own results and sends them to the outputs. In addition, given output paths could be weighted higher than other ones in order to achieve a specific goal [39].

Therefore, an artificial neural network is a set of connected artificial neurons which has an input layer, a number of hidden layers and an output layer. The information goes in one direction only, i.e. from the input layer to the output layer. During this process, the information goes through the hidden layer. Each node in a layer is connected to every node from the next layer. Weights between nodes store what researchers call the knowledge. This means that after training solved cases from the experts, the network should be able to solve future cases without human supervision using the previously learned knowledge.

In this context, our neural aggregation operator can be used to smartly combine different ssm into one. In this way, the result of the soft aggregation operator might consider (to some variable extent) all the individual values. Therefore, we have created a neural operator, which given a number of ssm as an input, could be able to learn how to aggregate them in order to produce a very accurate score on the likeness of different landmarks and POIs expressed in a textual way. To do so, we model our operator as follows:

$$y_i = f(x_i, w) = w^T x_i$$

Our task is then to find the weights that provide the best fit to past solved cases. In order to calculate our fit, we compute the sum squared error of our network's predictions over our dataset in this way [31]:

$$E(\mathbf{w}) = \sum_i (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 = \sum_i (\hat{y}_i - y_i)^2$$

Therefore, in order to find the line of best fit, we aim to minimize the value $E(w)$. As a result, we get the best combination of simple ssm that is able to solve benchmark datasets concerning geographic information with a minimum error. Then these values are updated using the solved cases.

In order to guide the search within this huge solution space, it is necessary to design a technique based on gradient descent so that we can minimize the distance from the existing function to the ideal one. To do that, it is necessary to compute the derivative of the gradient with respect to the weight associated to each ssm in this way:

$$\begin{aligned}\frac{\partial}{\partial w_{j\to k}} E(\mathbf{w}) &= \frac{\partial}{\partial w_{j\to k}} \sum_i \left(f(\mathbf{x}_i, \mathbf{w}) - y_i\right)^2 \\ &= \sum_i \frac{\partial}{\partial w_{j\to k}} \left(f(\mathbf{x}_i, \mathbf{w}) - y_i\right)^2 \\ &= \sum_i 2\left(f(\mathbf{x}_i, \mathbf{w}) - y_i\right) \frac{\partial}{\partial w_{j\to k}} f(\mathbf{x}_i, \mathbf{w})\end{aligned}$$

It is necessary to save all the obtained information in a vector, so the the output from our aggregation operator is given by:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \left(\frac{\partial E(\mathbf{w})}{\partial w_1}, \frac{\partial E(\mathbf{w})}{\partial w_2}, ..., \frac{\partial E(\mathbf{w})}{\partial w_n}\right)$$

$$= \left(\sum_i 2x_i^{(1)} f(\mathbf{x}_i, \mathbf{w}), \sum_i 2x_i^{(2)} f(\mathbf{x}_i, \mathbf{w}), ..., \sum_i 2x_i^{(n)} f(\mathbf{x}_i, \mathbf{w})\right)$$

Since this operator is not deterministic and has a cold start based on random values, it is not possible to guarantee the same mathematical properties that our fuzzy operator follows. Therefore, results in this context are always presented as an average of several executions.

4 Results

The rationale behind the evaluation is to show how the two new approaches are able to improve the results from current methods. In this context, it is necessary to remark that the most usual way of comparison between the scores from novel solutions and human judgments is usually expressed as a correlation between the two vectors of ratings [33]. This means that the goal is to obtain the degree of likeness between novel results and human judgments. The reason to follow this way to measure the likeness between geographic terms is to compare the degree of correlation between an artificial and a natural solution using the Pearson correlation coefficient [1]. Each of both solutions contains all the similarity scores associated with each particular case from the benchmark dataset. The final result will be between the values -1 (human ratings and results from the proposed solution present an opposite correlation) to 1 (human ratings and results from the proposed solution present a perfect correlation). Obviously, our challenge here is to get a result of 1 what means that our approach can replicate the behavior of the experts who initially solved the benchmark dataset.

To assess the quality of our aggregators, we have used the two existing benchmark datasets from this field. We summarize here the experiments that we have performed and the results that we have obtained. The rest of this section is organized in the following way: in subsection 4.1 we solve the GeReSiD benchmark dataset what is considered as the standard benchmark dataset for working with landmarks and POIs, and we analyze the results that we have obtained from these experiments. In subsection 4.2, we extend our evaluation by performing additional experiments over the so-called SDTS benchmark dataset, and we include an analysis of the results that we have obtained. Finally, in subsection 4.3, we discuss the major insights that can be extracted from the overall evaluation process.

4.1 Solving GeReSiD

This GeReSiD benchmark dataset [5] has been designed to include a pool of 97 terms that have been grouped in 50 pairs. Human judgments of similarity were collected on the different 50 pairs. Table 2 shows us these 50 pairs. These pairs range from those that are not similar at all (nursing home & continent) to other ones that are almost identical (motel & hotel) according to human judgment.

expr1	**expr2**	**human**	**expr1**	**expr2**	**human**
nursing home	continent	0.0169	speed bump	car park	0.3893
political boundary	women's clothes shop	0.0208	sea	island	0.3914
greengrocer	aqueduct	0.0310	managed forest	significant tree	0.3992
interior decoration shop	tomb	0.0504	swimming pool	water reservoir	0.4174
water ski facility	office furniture shop	0.0517	industrial land use	landfill	0.4385
community center	stream	0.0579	mountain hut	hilltop	0.4897
city suburb	antiques furniture shop	0.0717	barracks	shooting range	0.5145
vending machine	gate	0.0806	church	historic ruins	0.5348
fashion shop	swimming spot	0.0847	glacier	body of water	0.5574
beauty parlor	fire station	0.0943	canal	dock	0.5943
football pitch	corporate office	0.1086	police station	prison	0.6107
panoramic viewpoint	race track	0.1240	tower	lighthouse	0.6168
bed and breakfast	school building	0.1393	administrative office	town hall	0.6209
shelter	agricultural field	0.1488	historic castle	city walls	0.6446
ambulance station	city	0.1542	restaurant	beverages shop	0.6496
arts center	bureau de change	0.1612	historic battlefield	monument	0.6680
supermarket	surveillance camera	0.2042	art shop	art gallery	0.7480
post box	town	0.2097	bay	body of water	0.7623
school	toy shop	0.2172	stadium	athletics track	0.7643
canoe spot	hunting shop	0.2354	tram way	subway	0.7643
office building	academic bookstore	0.2686	floodplain	wetland	0.7686
car store	cycling facility	0.2727	basketball court	volleyball facility	0.7807
heritage item	valley	0.2896	public transport station	railway platform	0.8115
city	railway station	0.3279	theater	cinema	0.8730
picnic site	stream	0.3689	motel	hotel	0.9037

Table 1 GeReSiD benchmark dataset contains geographic terminology including 97 unique terms which have been grouped in 50 unique pairs

For achieving our results, we propose to use three different families of semantic similarity measures that we reviewed in Section 2: a) Maximum Common Substructure (MCS) method [37], b) Latent Semantic Analysis (LSA) methods [18] and c) University of Maryland Baltimore County (UMBC) methods [11].

- Concerning the greedy strategy followed by the *MCS method*, we have chosen the algorithm implemented by Rus et al [37].
- Concerning *LSA*, we have chosen the algorithm implemented by Landauer et al [18]. We use two alternatives corpora:
  - **LSA1**. We will use LSA with Touchstone Applied Science Associates (TASA) corpus. This corpus has near 60,000 samples from 6,000 textbooks and other pieces of literature. It corresponds approximately to the total amount of text that an average college student in USA has experienced in its life.
  - **LSA2**. We will also use LSA with the 1st-year-College Corpus. This corpus near 12 million terms that belongs to the category of general readings up to 1st year college in the USA.
- Concerning *UMBC methods*, we propose to use the three alternative methods presented by Han et al. [11]:
  - Firstly, we will use **UMBC1** method with the WebBase corpus. This corpus is a dataset containing a collection of English paragraphs with over three billion words processed from the Stanford WebBase project.
  - Secondly, we will use the **UMBC2** method with the English Gigaword Corpus which is a comprehensive archive of news-wire text data that has been acquired over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania.
  - Finally, we will try **UMBC STS**. This method uses a lexical similarity feature that combined LSA word similarity and WordNet knowledge.

Table 2 presents a summary of the results achieved by the existing methods for measuring semantic similarity concerning landmarks and POIs. First column shows us the approach we have used. Score is the fitness each approach has achieved when solving the GeReSiD benchmark dataset [5]. Finally, the p-value represents the mathematical probability to find the current result if the correlation were in fact zero. If this probability is below than $5.0 \cdot 10^{-2}$, it is usually assumed that the correlation is statistically significant.

Our proposed operators have been able to improve the results from existing approaches when dealing with the GeReSiD dataset. In fact, our strategies have outperformed all cutting-edge ssm. The reason for that is, unlike existing techniques, our aggregation strategies are more appropriate for dealing with subjectivity and imprecision that human language brings when labeling landmarks and points of interest.

Method	Score	p-value
MCS	0.21	$7.1 \cdot 10^{-2}$
LSA1	0.38	$3.2 \cdot 10^{-3}$
UMBC2	0.43	$9.1 \cdot 10^{-4}$
UMBC1	0.49	$1.5 \cdot 10^{-4}$
LSA2	0.54	$2.5 \cdot 10^{-6}$
UMBC STS	0.63	$4.7 \cdot 10^{-7}$
Fuzzy operator	**0.67**	$5.0 \cdot 10^{-8}$
Neural operator	**0.67**	$5.0 \cdot 10^{-8}$

Table 2 Summary of the results achieved by the existing methods for measuring semantic similarity on the GeReSiD benchmark dataset. Results can be considered statistically significant from a p-value of $5.0 \cdot 10^{-2}$

4.1.1 Analysis of the results

Now, we are going to analyze the results from the experiments that we have performed for the GeReSiD benchmark dataset. In this way, the figures we can see below are the graphical representations allowing the visual inspection of the results from the experiments that we have performed. GeReSiD bars represent the 50 textual pairs representing landmarks and points of interest that we have shown in Table 1. Meanwhile, the red bars represent the results achieved by each of the methods that we are describing. Finally, the correlation coefficient and its associated explanation are expressed in the figure captions.

In Figure 1, we can see that the MCS method has obtained a poor result (score of 0.21) when solving the GeReSiD benchmark dataset. As we can see, the reason is that the method generates a lot of false positives for text expressions that are far from being semantically similar, and it is not particularly good when recognizing cases of semantic similarity, i.e. it generates a lot of false negatives.

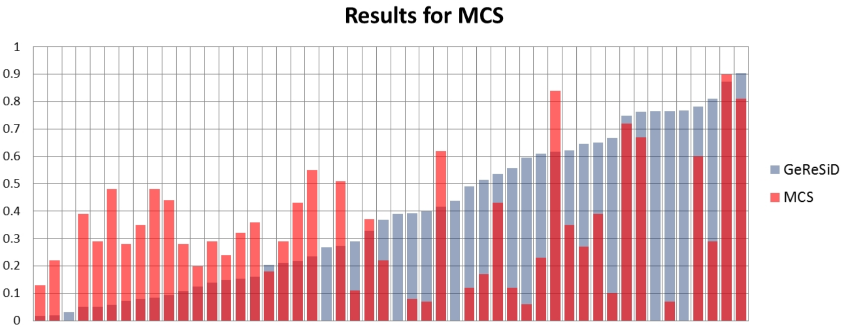

Fig. 1 *MCS method (score of 0.21) does not seem to be very suitable for solving the benchmark dataset. The reason is that the method generates a lot of false positives for text expressions that are far from being semantically similar*

Figure 2 presents the results for the LSA1 method. Despite the fact this method has been able to generate less false positives than its predecessor; a poor score of 0.38 has been achieved in the evaluation. The reason is that it is largely fails to recognize pairs that are semantically similar.

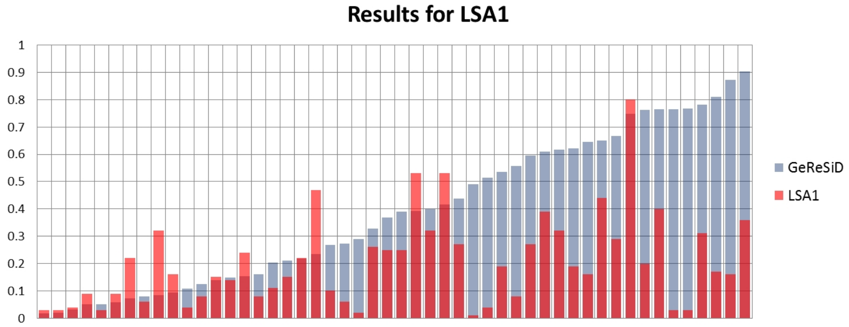

Fig. 2 *LSA1 (score of 0.38) is able to generate less false positives than its predecessor, but it still fails to recognize pairs that are semantically similar*

Figure 3 shows us the results for the method UMBC2. This method is able to produce less false positives than the previous methods. However, it still generates a serious false positive when determining the semantic similarity of the pair: beauty parlour/fire station. Additionally, the results are poor when recognizing equivalent pairs, i.e. it generates many false negatives. These mistakes largely contribute to the achievement of a score of 0.43.

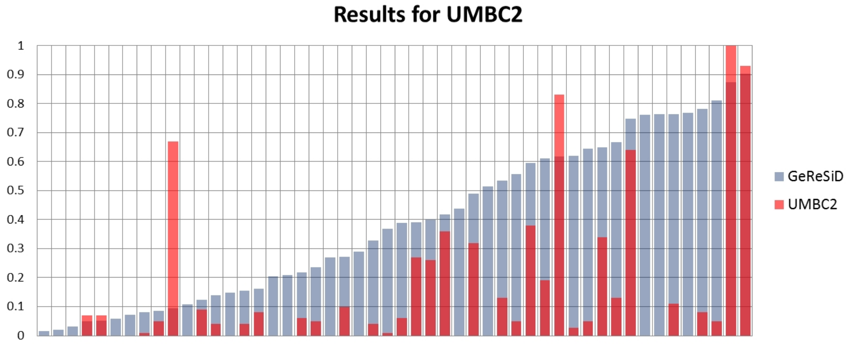

Fig. 3 *UMBC2 (score of 0.43) can produce less false positives than the previous methods. However, it creates a very serious false positive when determining the semantic similarity of the pair: beauty parlour/fire station. Moreover, the large amount of false negatives adds even more penalization*

In Figure 4, we can see that the UMBC1 method. Results are slightly better than in previous experiments. However, the method still presents a very serious false positive when determining the semantic similarity of the pair: beauty parlour/fire station. Additionally, it generates false negatives for the comparisons of: bay/body of water, stadium/athletics track, and floodplain/wetland. As a result, the method has achieved a score of 0.49.

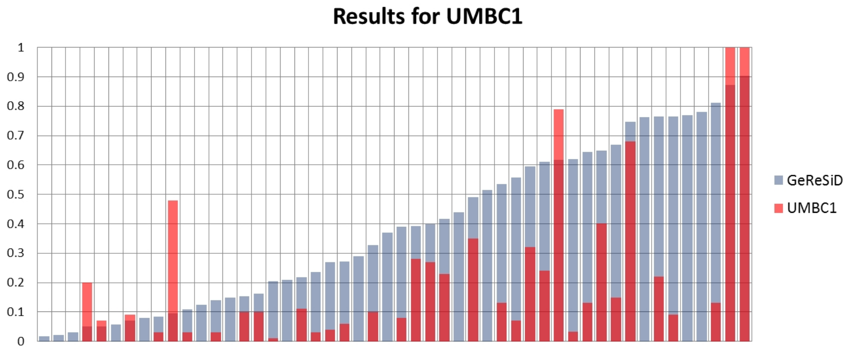

Fig. 4 *UMBC1 (score of 0.49) still presents a very serious false positive when determining the semantic similarity of the pair: beauty parlour/fire station. Additionally, it generates serious false negatives for: bay/body of water, stadium/athletics track and floodplain/wetland*

Figure 5 shows us the results for the method LSA2. This method reaches a decent score of 0.54. The reason for that is that it presents good results in the sense that it is not generating any false positive, but it still needs to be more decisive in stating that two pairs are semantically equivalent.

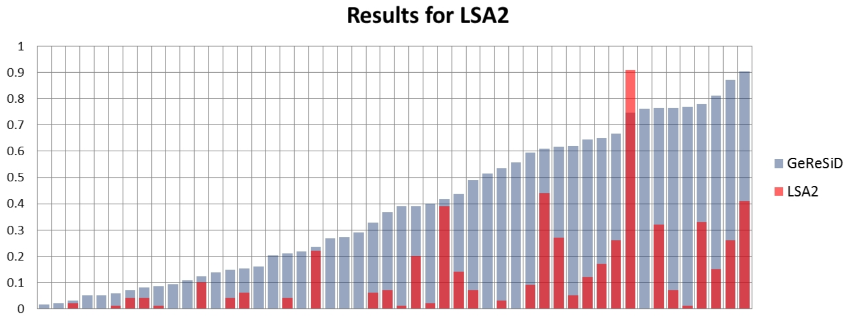

Fig. 5 *LSA2 (score of 0.54) presents good results. In fact, it does not generate any serious false positive, but it still needs to be more decisive in stating that two pairs are semantically equivalent*

Figure 6 shows us the results for the method UMBC STS. This method got very decent results (score of 0.63), mainly due to the fact of generating a very low number of false positives, and being quite successful when correctly determining semantic similarity. This method was the best before our approach was proposed.

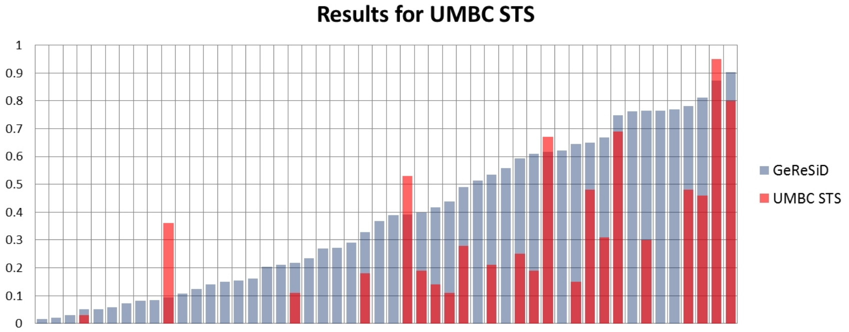

Fig. 6 *UMBC STS (score of 0.63) got decent results, mainly due to the fact of generating a very low number of false positives, and being successfully when determining semantic similarity. This method was the best until now*

In Figure 7, we can see that our fuzzy operator (score of 0.67) presents a very good behavior. The method has successfully worked with intermediate cases of semantic similarity. There is still, however, some place for improvement. For instance, the method has not been able to hit on complex cases such as bay/body of water or floodplain/wetland.

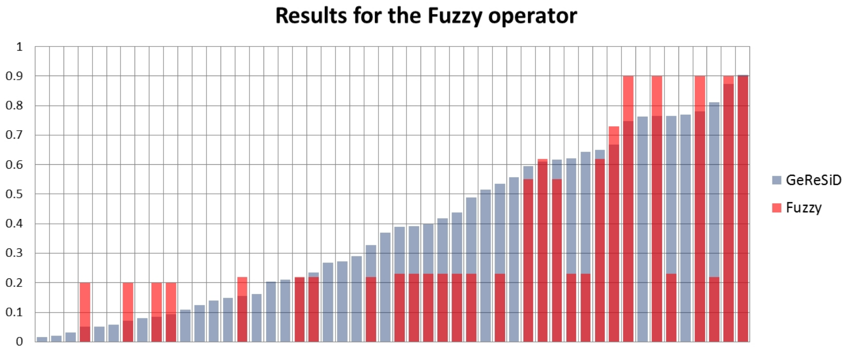

Fig. 7 *The fuzzy operator (score of 0.67) beats the existing state-of-the-art methods. This method can successfully work with intermediate cases of semantic similarity*

Figure 8 shows us that our neural operator (score of 0.67) presents a very good behavior too. Results are in line with those from the fuzzy operator. However, the distribution of the hits is slightly different. The great advantage of the neural operator is that it can replicate the upward pattern of similarity. However, we have prevented over fitting by performing a 10 cross-fold validation, what means that it is strongly penalizing for the aggregation operator to adhere strictly to the data, but reflects a more behavior closer to what one could expect in a real setting.

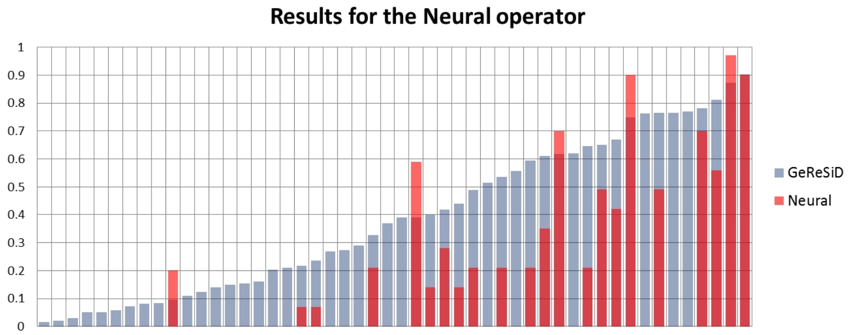

Fig. 8 *The neural operator (score of 0.67) presents a very good behavior. The reason is that it is not generating any false positive and it is doing a great job recognizing semantically similar pairs*

4.2 Solving SDTS

In order to complement the evaluation of our aggregators, now we are going to use another benchmark dataset that is also well known and relevant in the geographical information domain. This benchmark dataset was created by Rodriguez and Egenhofer using the Spatial Data Transfer Standard (SDTS) [36] as a base. The rationale behind it is to verify the behavior of different ssm when working with geographical nomenclature to be shared by a wide range of different computer platforms and systems. In contrast with GeReSiD, this dataset contains just a limited number of geographic terms whereby all are compared against all. Table 3 shows us the 21 pairs representing the ground truth for this benchmark.

expr1	expr2	human	expr1	expr2	human
Athletic field	Ballpark	0.83	Building	Road	0.10
Athletic field	Building	0.17	Building	Sports Arena	0.48
Athletic field	Road	0.12	Building	Stadium	0.30
Athletic field	Sports Arena	0.49	Building	Theater	0.44
Athletic field	Stadium	0.70	Road	Sports Arena	0.10
Athletic field	Theater	0.16	Road	Stadium	0.14
Ballpark	Building	0.16	Road	Theater	0.10
Ballpark	Road	0.10	Sports Arena	Stadium	0.78
Ballpark	Sports Arena	0.49	Sports Arena	Theater	0.58
Ballpark	Stadium	0.74	Stadium	Theater	0.38
Ballpark	Theater	0.14			

Table 3 STDS benchmark dataset contains geographic terminology on 7 unique terms which have been grouped in 21 unique pairs

As in the previous experiment, we are testing here the correlation between the artificial and the natural solutions using the Pearson correlation coefficient [1]. The methods used and their associated configuration remains the same: MCS from [37], LSA1, and LSA2 from [18], with the exception of the UMBC family of methods for which (as of February 2019) no data are available for this benchmark dataset. As a consequence, we have decided to include the classical Lin [24] and Pilehvar [32] approaches in order to have new references for later comparison.

In this context, Table 4 presents a summary of the results achieved by the existing methods for measuring semantic similarity concerning landmarks and POIs using SDTS. As in the previous experiment, the first column shows us the approach we have used. Score column represents the fitness that each semantic similarity measure has achieved when solving SDTS, and the last column gives us information about the statistical significance of the result obtained what means that just in those cases where the p-value is below 5.0, the associated results can be considered significant.

Method	**Score**	**p-value**
Lin	0.13	$2.8 \cdot 10^{-1}$
MCS	0.30	$9.3 \cdot 10^{-2}$
LSA1	0.58	$3.0 \cdot 10^{-3}$
Pilehvar	0.77	$2.2 \cdot 10^{-5}$
LSA2	0.81	$4.3 \cdot 10^{-6}$
Neural operator	**0.83**	$1.6 \cdot 10^{-6}$
Fuzzy operator	**0.86**	$2.9 \cdot 10^{-7}$

Table 4 Summary of the results achieved by the existing methods for measuring semantic similarity on the SDTS benchmark dataset. Results can be considered statistically significant from a p-value of $5.0 \cdot 10^{-2}$

4.2.1 Analysis of the results

Once again, we provide figures to facilitate the visual inspection of the results concerning the SDTS benchmark dataset. The figures we can see below are the graphical representations of the results from the experiments that we have performed. SDTS bars represent the 21 textual pairs representing the geographical terms that we have shown in Table 3. Meanwhile, the red bars represent the individual results achieved by each of the methods considered within this work. Finally, the correlation coefficient achieved is explained in each corresponding caption.

In Figure 9, we can see that the Lin method (score of 0.13) presents a very poor behavior. The reason is that the method is able to work with a limited number of cases of semantic similarity only. In addition, the results are not very accurate for those cases, so this means that the overall value for the correlation is really low. Therefore, it would not be advisable to use it by itself in a real setting.

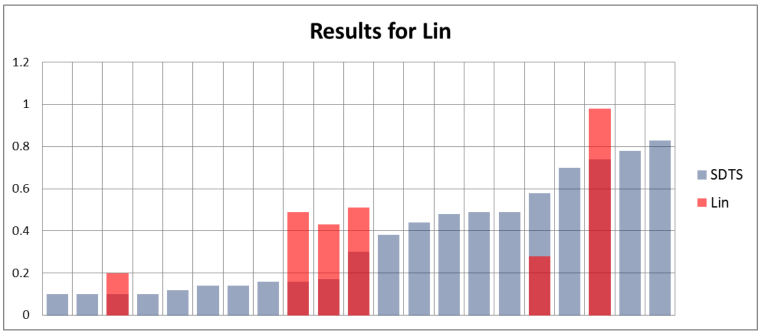

Fig. 9 *Lin method (score of 0.13) does not seem to be very suitable for solving the benchmark dataset. The reason is that the method can work with a limited number of cases of semantic similarity only*

In Figure 10, we can see that the MCS method (score of 0.30) presents a result distribution which is in line with the one obtained in the previous experiment. The reason is similar: it generates a lot of false positives for very short textual expressions that are far from being semantically similar, and it does not particularly provide good results when recognizing cases of positive semantic similarity. Therefore, it would only make sense to use it within an aggregation strategy.

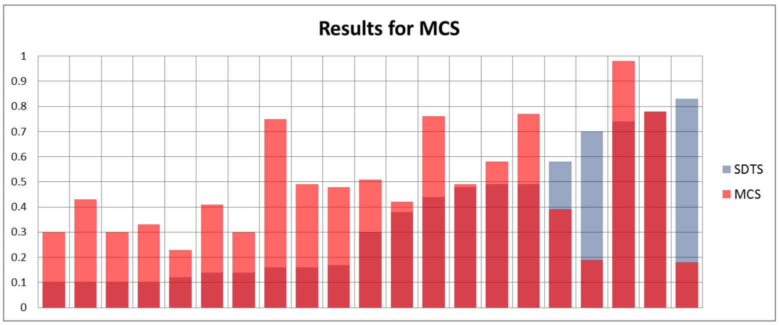

Fig. 10 *MCS method (score of 0.30) does not seem to be very suitable for solving the benchmark dataset. The reason is that the method generates a lot of false positives for text expressions that are far from being semantically similar*

In Figure 11, we can see that the results for the LSA1 method (score of 0.58). Despite the fact that this method is able to work with a limited number of cases of semantic similarity only, it does a good work of not generating any false positives, and it is able to produce some hits on a limited number of occasions. As a result, it is able to achieve a decent global correlation score. However, it might be a little risky to make it operate alone in a real environment.

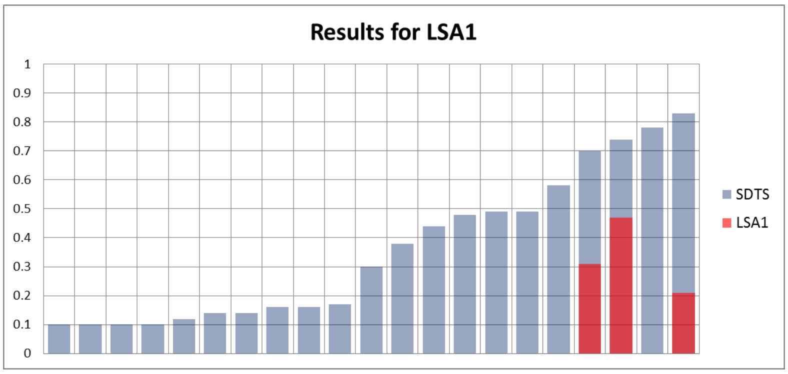

Fig. 11 *LSA1 method (score of 0.58) achieves a decent score when solving the SDTS benchmark dataset. The reason is that the method does not generate any false positives, and it is able to produce some hits on a limited number of occasions*

In Figure 12, we can see that the results for the Pilehvar method (score of 0.77). This method does very good work since the result distribution is very balanced. Even so, it cannot achieve a higher correlation with human judgment as it does not exactly match the value of positive cases. However, the method could work on its own with certain guarantees that it will work well.

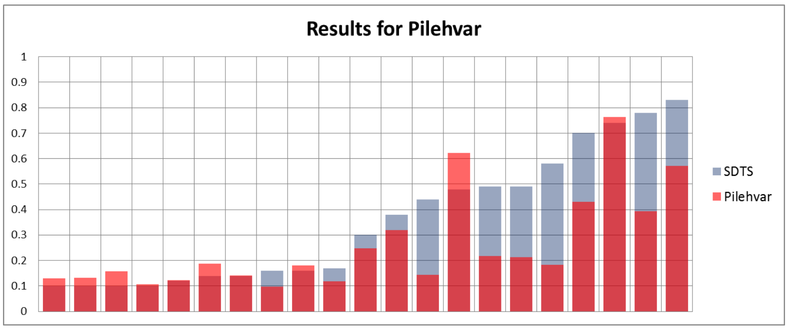

Fig. 12 *Pilehvar method (score of 0.77) seem to be suitable for solving the SDTS benchmark dataset. The reason is that this similarity measure generates a result distribution that is quite balanced*

In Figure 13, we can see that the results for LSA 2 (score of 0.81) which presents a really good behavior. Just like it happened in the previous experiment, the method has successfully worked fine in most cases. As a consequence, we can affirm that this ssm is the best of those operating alone. There is still, however, some place for improvement.

In Figure 14, we can see that our neural aggregator (score of 0.83) presents a very good behavior. The method has successfully identified most of the cases of semantic similarity. As in the analog experiment carried out before, we have prevented overfitting by performing a 10 cross-fold validation. There is still,

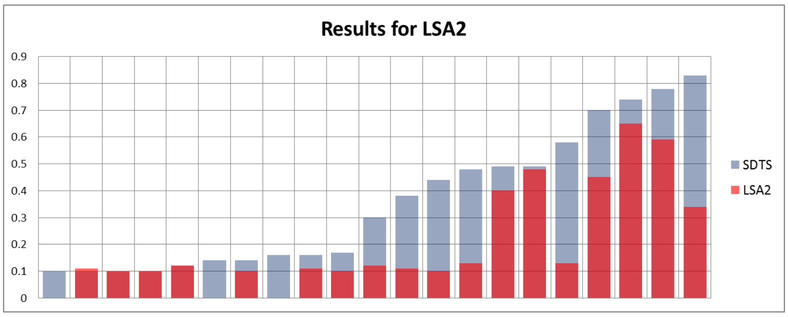

Fig. 13 *LSA2 method (score of 0.81) seem to be very suitable for solving the SDTS benchmark dataset. The reason is that the method has successfully worked fine in most cases. This ssm has been the one that has yielded the best results operating alone*

however, some place for improvement. This is mainly due to the fact that the limited number of instances to proper calibration does not allow it to achieve the best results, i.e. when working with larger datasets, much better results could be expected.

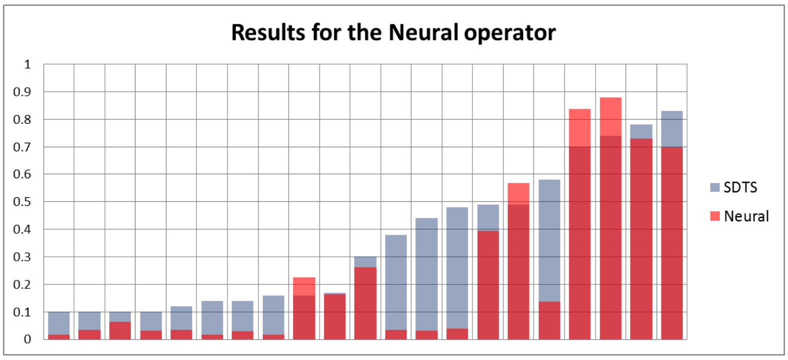

Fig. 14 *The neural aggregator (score of 0.83) presents a very good behavior and it is able to beat the rest of ssm. In fact, this aggregator does not generate any false positive and it is doing fine when recognizing semantically similar pairs*

In Figure 15, we can see that our fuzzy aggregator (score of 0.86) presents the best results. This also happened in the previous experiment. It is clear that this aggregator is greatly benefited here by the fact that the fuzzy aggregation strategy requires no training and can operate on any dataset with great reliability. In scenarios with a greater number of training instances, it is supposed to go hand-in-hand with the neural operator.

From the experiments, it is clear that our proposed aggregation operators have been able to beat the existing methods when solving cases of semantic similarity for landmarks and POIs for both benchmark datasets. In this context, we have achieved an improvement over existing methods.

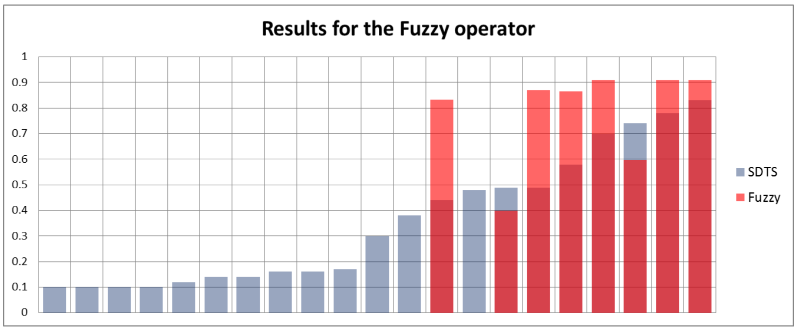

Fig. 15 *The fuzzy operator (score of 0.86) is able to beat the existing state-of-the-art methods. This method makes almost no mistakes, and as result, it is able to achieve a high correlation coefficient*

5 Conclusions

In this work, we have presented two novel operators for the semantic similarity measurement of very short textual expressions, with a special interest in the coverage of landmarks and POIs. These two operators are based on the fuzzy and neural aggregation of semantic similarity measures respectively. The rationale behind these two novel operators is the aggregation of existing similarity measures so that in the case some particular measures cannot perform well for a particular scenario; their effects may be blurred by other measures that perform much better, but without falling into the shortsighted strategy from the existing averaging solutions.

The experiments reported in this work show that our two novel operators are able to improve the results from existing methods when solving the GeReSiD and the STD benchmark datasets. These two datasets are the most well-known benchmark dataset for landmarks and POIs. This means that these two new approaches could be considered as a new improvement, and therefore, we can conclude that it seems that fuzzy and neural aggregation operators are appropriate for handling the vagueness induced from the subjectivity of the natural language when labeling landmarks and POIs. As a result, we think that these novel operators can be exploited by the research community in order to implement innovative software solutions dealing with semantic similarity ranging from systems to help people discover interesting locations in the context of urban systems or natural environments to solutions providing accurate personalized recommendation of landmarks or the creation of social relationships among users operating under areas of similar geographical typology.

Concerning future research, it is important to keep working for improving the overall quality of the existing methods so that it could be possible to recognize even more complex cases of semantic similarity. In this sense, it seems appropriate to address aspects related to ensemble learning [41] to reduce uncertainty in scenarios in which it is not possible to know very well what the right result is. In situations like this, relying on several inputs is usually considered as fairer resolution, since none of each individual measures prevail over the rest. It is also worth considering spatial and temporal aspects since meaning is not something static, but has multiple shades and is able to evolve over time. These are important research questions which have attracted none or little attention in the literature, and therefore, additional research efforts are needed to shed light on our understanding of semantic similarity and its implications.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improve this work. The research reported in this paper has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria under the frame of the COMET Center SCCH [FFG: 844597].

References

1. P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *JASIST*, 54(6):550–560, 2003.
2. S. Amir, A. Tanasescu, and D. A. Zighed. Sentence similarity based on semantic kernels for intelligent text retrieval. *J. Intell. Inf. Syst.*, 48(3):675–689, 2017.
3. M. B. Aouicha and M. A. H. Taieb. G2WS: gloss-based wordnet and wiktionary semantic similarity measure. In *12th IEEE/ACS International Conference of Computer Systems and Applications, AICCSA 2015, Marrakech, Morocco, November 17-20, 2015*, pages 1–7, 2015.
4. A. Ballatore, M. Bertolotto, and D. C. Wilson. Geographic knowledge extraction and semantic similarity in openstreetmap. *Knowl. Inf. Syst.*, 37(1):61–81, 2013.
5. A. Ballatore, M. Bertolotto, and D. C. Wilson. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, 18(4):747–767, 2014.
6. A. Ballatore, D. C. Wilson, and M. Bertolotto. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *International Journal of Geographical Information Science*, 27(10):2099–2118, 2013.
7. D. Buscaldi, J. L. Roux, J. J. G. Flores, and A. Popescu. LIPN-CORE: semantic text similarity using n-grams, wordnet, syntactic analysis, ESA and information retrieval based features. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA.*, pages 162–168, 2013.
8. J. M. Chaves-González and J. Martinez-Gil. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. *Knowl.-Based Syst.*, 37:62–69, 2013.
9. C. Feng and D. M. Flewelling. Assessment of semantic similarity between land use/land cover classification systems. *Computers, Environment and Urban Systems*, 28(3):229–246, 2004.
10. M. Grabisch, J. Marichal, R. Mesiar, and E. Pap. Aggregation functions: Construction methods, conjunctive, disjunctive and mixed classes. *Inf. Sci.*, 181(1):23–43, 2011.
11. L. Han, T. Finin, P. McNamee, A. Joshi, and Y. Yesha. Improving word similarity by augmenting PMI with estimates of word polysemy. *IEEE Trans. Knowl. Data Eng.*, 25(6):1307–1322, 2013.

12. H. Hobel, P. Fogliaroni, and A. U. Frank. Deriving the geographic footprint of cognitive regions. In *Geospatial Data in a Changing World - Selected Papers of the 19th AGILE Conference on Geographic Information Science, Helsinki, Finland, 14-17 June 2016*, pages 67–84, 2016.
13. H. Hsu and C. Chen. Aggregation of fuzzy opinions under group decision making. *Fuzzy Sets and Systems*, 79(3):279–285, 1996.
14. K. Janowicz, M. Raubal, and W. Kuhn. The semantics of similarity in geographic information retrieval. *J. Spatial Information Science*, 2(1):29–57, 2011.
15. K. Janowicz, M. Raubal, A. Schwering, and W. Kuhn. Semantic similarity measurement and geospatial applications. *Trans. GIS*, 12(6):651–659, 2008.
16. Q. Ji, P. Haase, and G. Qi. Combination of similarity measures in ontology matching using the OWA operator. In *Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice*, pages 281–295. 2011.
17. L. Kuncheva. Using measures of similarity and inclusion for multiple classifier fusion by decision templates. *Fuzzy Sets and Systems*, 122(3):401–407, 2001.
18. T. K. Landauer and J. Psotka. Simulating text understanding for educational applications with latent semantic analysis: Introduction to LSA. *Interactive Learning Environments*, 8(2):73–86, 2000.
19. J. J. Lastra-Díaz and A. García-Serrano. A new family of information content models with an experimental survey on wordnet. *Knowl.-Based Syst.*, 89:509–526, 2015.
20. J. J. Lastra-Díaz and A. García-Serrano. A novel family of ic-based similarity measures with a detailed experimental survey on wordnet. *Eng. Appl. of AI*, 46:140–153, 2015.
21. X. Li, G. Cong, X. Li, T. N. Pham, and S. Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 433–442, 2015.
22. Y. Li, D. McLean, Z. Bandar, J. O'Shea, and K. A. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.*, 18(8):1138–1150, 2006.
23. K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera. Personalized tour recommendation based on user interests and points of interest visit durations. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1778–1784, 2015.
24. D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pages 296–304, 1998.
25. J. Martinez-Gil. An overview of textual semantic similarity measures based on web intelligence. *Artif. Intell. Rev.*, 42(4):935–943, 2014.
26. J. Martinez-Gil. Accurate semantic similarity measurement of biomedical nomenclature by means of fuzzy logic. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24(2):291–306, 2016.
27. J. Martinez-Gil. Coto: A novel approach for fuzzy aggregation of semantic similarity measures. *Cognitive Systems Research*, 40:8–17, 2016.
28. J. Martinez-Gil and J. M. Chaves-Gonzalez. Automatic design of semantic similarity controllers based on fuzzy logics. *Expert Systems with Applications*, 131:45–59, 2019.
29. J. A. Medina-Hernández, F. Gomez-Castañeda, and J. A. Moreno-Cadenas. An evolving fuzzy neural network based on the mapping of similarities. *IEEE Trans. Fuzzy Systems*, 17(6):1379–1396, 2009.
30. G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
31. M. T. Musavi, K. Kalantri, W. Ahmed, and K. H. Chan. A minimum error neural network (MNN). *Neural Networks*, 6(3):397–407, 1993.
32. M. T. Pilehvar and R. Navigli. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artif. Intell.*, 228:95–128, 2015.
33. G. Pirrò. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.*, 68(11):1289–1308, 2009.
34. N. Ranjbar, F. Mashhadirajab, M. Shamsfard, R. H. pour, and A. V. pour. Mahtab at semeval-2017 task 2: Combination of corpus-based and knowledge-based methods to

measure semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 256–260, 2017.

35. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 448–453, 1995.
36. M. A. Rodríguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Trans. Knowl. Data Eng.*, 15(2):442–456, 2003.
37. V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu. SEMILAR: the semantic similarity toolkit. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 163–168, 2013.
38. M. Rybinski and J. F. Aldana-Montes. Domesa: a novel approach for extending domain-oriented lexical relatedness calculations with domain-specific semantics. *J. Intell. Inf. Syst.*, 49(3):315–331, 2017.
39. R. Setiono. Generating linear regression rules from neural networks using local least squares approximation. In *Connectionist Models of Neurons, Learning Processes and Artificial Intelligence, 6th International Work-Conference on Artificial and Natural Neural Networks, IWANN 2001 Granada, Spain, June 13-15, 2001, Proceedings, Part I*, pages 277–284, 2001.
40. P. D. Turney. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *TACL*, 1:353–366, 2013.
41. G. I. Webb and Z. Zheng. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Trans. Knowl. Data Eng.*, 16(8):980–991, 2004.
42. P. Zhang, Z. Zhang, W. Zhang, and C. Wu. Semantic similarity computation based on multi-feature combination using hownet. *JSW*, 9(9):2461–2466, 2014.