Combining semantic and lexical measures to evaluate medical terms similarity*

Silvio Domingos Cardoso[1,2], Marcos Da Silveira[1], Ying-Chi Lin[3], Victor Christen[3], Erhard Rahm[3], Chantal Reynaud-Delaître[2], and Cédric Pruski[1]

[1] LIST, Luxembourg Institute of Science and Technology, Luxembourg
{silvio.cardoso,marcos.dasilveira,cedric.pruski}@list.lu
[2] LRI, University of Paris-Sud XI, France
chantal.reynaud@lri.fr
[3] Department of Computer Science, Universität Leipzig, Germany
{lin,christen,rahm}@informatik.uni-leipzig.de

Abstract. The use of similarity measures in various domains is cornerstone for different tasks ranging from ontology alignment to information retrieval. To this end, existing metrics can be classified into several categories among which lexical and semantic families of similarity measures predominate but have rarely been combined to complete the aforementioned tasks. In this paper, we propose an original approach combining lexical and ontology-based semantic similarity measures to improve the evaluation of terms relatedness. We validate our approach through a set of experiments based on a corpus of reference constructed by domain experts of the medical field and further evaluate the impact of ontology evolution on the used semantic similarity measures.

Keywords: Similarity measures · Ontology evolution · Semantic Web · Medical terminologies

1 Introduction

Measuring the similarity between terms is at the heart of many research investigations. In ontology matching, similarity measures are used to evaluate the relatedness between concepts from different ontologies [9]. The outcomes are the mappings between the ontologies, increasing the coverage of domain knowledge and optimize semantic interoperability between information systems. In information retrieval, similarity measures are used to evaluate the relatedness between units of language (e.g., words, sentences, documents) to optimize search [39]. The literature of this domain reveals that several families of similarity measures can be distinguished [17, 14] such as string-based, corpus-based, knowledge-based metrics, etc. Lexical Similarity Measures (LSM) regroups the similarity families that rely on syntactic or lexical aspects of the units of languages [29]. Such metrics are efficient to compare strings such as "*Failure of the kidney*" with "*Kidney*

* This work is supported by FNR Luxembourg and DFG Germany through the ELISA project

failure". However, they do not capture very well the semantic similarity. For instance, "*Cancer*" and "*malignancy*" can be totally disjointed from the lexical point of view despite their closely related semantics. To overcome this barrier, Semantic Similarity Measures (SSM) have been introduced. They exploit meaning of terms to evaluate their similarity. This is done using two broad types of semantic proxies: corpora of texts and ontologies.

The corpora proxy uses Information Content (IC-based) to observe the usage of terms and determine the similarity based on the distribution of the words or the co-occurrence of terms [25, 26]. The ontology proxy uses the *intrinsic* Information Content (iIC-based) [17], where the structure of the ontology allows calculating some semantic similarities [6]. Both proxies have been use in several domains, but we are working mainly with ontologies and we focus our analysis on SSM that are iIC-based. Although single similarity measures have been successfully used in many works, their combination remains under explored, especially the couple LSM/iIC-based SSM. The goal of this work is not to propose another similarity measure, but to demonstrate that weighted combination of existing ones can improve the outcomes.

Our motivation for this work came from the observations, in our previous work on semantic annotations and mappings adaptation [2, 3, 11], that few information is made available to understand how mappings and semantic annotations were generated and how they are maintained over time. In order to propose an automatic maintenance method for mappings and annotations [4], we search for patterns that allow reasonable explanations for the selection of terms and their relations. The similarity metrics became an essential tool for our approach. We deal with datasets of mapping and annotations that were generated based on very different methods (automatically and/or manually). We are interested on finding a combination of methods that can better explain the reasoning behind the generation/maintenance of mappings or annotations. The single method (LSM or SSM) that we evaluated did not represent well the patterns that we are looking for. Thus, we empirically evaluated the SSM×LSM combination and we are presenting the outcomes of our research in this paper. Differently from other comparative approaches that look for unifying methods or automatically select the best single method for a specific task, the goal of our research was to look for combinations of methods. Our ultimate goal is to define a procedure to analyze the evolution of ontologies and collect relevant information to be used to preserve the validity of annotations and mappings.

In this paper, we present a weighting method that combines LSM and ontology-based SSM to evaluate the relatedness between terms (word or multi-token terms) in order to improve the characterization of changes occurring in the ontology at evolution time. We based our solution on existing well known similarity measures and evaluate it using a Gold standard corpus. Our iterative method allows to find the best weights to associate to LSM and SSM respectively. In our experiments we used datasets constructed by experts from the medical domain [32]. It gathers scores given by domain experts on the relatedness between terms. We proved the validity of our measure by first showing the correlation between

the obtained values and the scores given by domain experts on the data of the reference corpus using the Spearman's rank correlation metric. We then use the Fisher's Z-Transformation to evaluate the added value of our metric with respect to state-of-the-art similarity measures. Through this work, we are able to show:

- The added value of combining LSM and ontology-based SSM for measuring term relatedness.
- The validity of the combination SSM×LSM with respect to experts score.
- The impact of the evolution of ontologies on the used SSM.
- The most suitable metrics and weights for SNOMED CT and MeSH.

The remainder of this article is structured as follows. Section 2 introduces the various concepts needed to understand our approach. This includes the definition of existing lexical and semantic similarity measures as well as the methods we have followed to evaluate our work. Section 3 presents the related work. Section 4 describes our approach for combining lexical and semantic similarity measures as well as our evaluation methodology while results are presented in Section 5. Section 6 discuss the results. Finally, Section 7 wraps up with concluding remarks and outlines future work.

2 Background

In this section, we provide the necessary background information to understand the notion tackled in this paper. We start by listing the studied LSM and SSM. We then explain the Spearman's rank correlation and the Fisher's Z-transformation formulas we have used in our experiments.

2.1 Lexical Similarity Measures

In our work, we introduced lexical similarity measures through various string-based approaches. It consists in the analysis of the composition of two strings to determine their similarity. Two types of approach can be distinguished: character-based and term-based. The former denotes the comparison of two strings and the quantification of the identified differences. The latter compares the differences between words composing the string. In our experiments, we have used the 12 following LSM: Levenshtein, Smith-Waterman, Jaccard, Cosine, Block Distance, Euclidean Distance, Longest Common Substring, Jaro-Winkler, LACP, TF/IDF, AnnoMap [5] and Bigram.

2.2 Semantic Similarity Measures

Semantic similarity measures denote a family of metrics that rely on external knowledge to evaluate the distance between terms from their meaning point of view. It encompasses corpus-based metrics and ontology-based which [15]. In this work, we put the stress on ontology-based approaches (iIC-based). We have retained 11 SSMs following a deep literature survey. Table 1 hereafter contains

the various semantic similarity measures that have been tested in our work. Table 2 refers to iIC-based methods. Note that the SSMs methods from Table 1 use as input the outcomes of iIC-based methods. Thus, when presenting the results we indicate the name of the SSM method as well as the iIC-based method used as input.

Table 1. Used semantic similarity measures

SSM	**Description**
Jiang Conrath [19]	Similar to Resnik, it uses a corpus of documents in addition to an ontology.
Feature Tversky Ratio Model [40]	Considers the features of label to compute similarity between different concepts, but the position of the concept in the ontology is ignored. Common features tend to increase the similarity and other features tend to decrease the similarity.
Tversky iIC Ratio Model [8]	
Lin [21]	Similar to Resnik's measure but uses a ratio instead of a difference
Lin GraSM [7]	
Mazandu [23]	Combination of node and edge properties of Gene Ontology terms.
Jaccard iIC [16]	It consists in the ratio between the intersection of two sets of feature and the union of the same two sets.
Jaccard 3W iIC [27]	
Resnik GraSM [35]	See Table 2
Resnik [35]	
Sim iIC [20]	Exploits iIC of the Most Informative Common Ancestor of the concepts to evaluate.

2.3 Spearman's Rank Correlation

One of the objectives of this work is to experimentally show the complementarity of LSM and ontology-based SSM to better evaluate the relatedness between terms. Since we compared the results obtained experimentally with the score assigned by domain experts, we need a method to evaluate their correlation. Spearman's Rank Correlation (cf. equation 1) is a statistical method that measures the coefficient strength of a linear relationship between paired data [34]. In other words, its verifies whether the values produced by automatic similarity measures and scores given by domain specialists are correlated.

$$r_s = 1 - \frac{6 \sum_i d_i}{n(n^2 - 1)} \tag{1}$$

In equation 1, d_i is the difference between the two ranks of each observation and n is the number of observations.

Table 2. Information Content based measures

***i*IC-based metrics**	**Description**
Resnik (normalized) [12]	Based on the lowest common ancestor.
Sanchez [36]	*i*IC of a concept is directly proportional to its number of taxonomical subsumers and inversely proportional to the amount of leaves of its hyponym tree.
Sanchez adapted [36]	
Seco [38]	*i*IC is computed based on the number of hyponyms a concept has in WordNet. This metric does not rely on corpus.
Zhou [41]	*i*IC considers not only the hyponyms of each word sense in WordNet but also its depth in the hierarchy.
Harispe [16]	Modification of [36] in order to authorize various non uniformity of *i*ICs among the leafs
Max depth non linear [16]	*i*IC of a concept is directly computed based on the depth of the concept.
Max depth linear [16]	
Ancestors Norm [16]	*i*IC of a concept is computed based on the number of ancestors of the concept.

2.4 Fisher's Z-Transformation

Fisher's Z-Transformation is a statistic method that allows us to verify whether two nonzero's Spearman's rank coefficients are statistically different [34]. The corresponding formula is:

$$z = \frac{1}{2} \ln \left(\frac{1 + r_s}{1 - r_s} \right) \tag{2}$$

Through this normalization of Spearman's rank coefficient we can assure whether r_s from an automatic similarity method X_i is better than a r'_s from a method Y_i.

In order to compare various correlations, we have to apply the following three-steps method:

- Conversion of r_s and r'_s to z_1 and z_2 by applying equation 2.
- Compute the probability value $\rho \in 0 \leq \rho \leq 1$ through equation 3, where N_1 and N_2 are the number of elements in our dataset and $erfc$ denotes the complementary error function.
- Test the null assumption $H_0 : r_s = r'_s$ case $\rho > 0.05$ and vice versa. Nevertheless, it only can be performed when N, i.e., the number of paired data is moderately large ($N \geq 10$) to assure the statistical significance.

$$\rho = erfc \left(\frac{|z_1 - z_2|}{\sqrt{2}\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \right) \tag{3}$$

In consequence, a value smaller than 0,05 indicates that the two evaluated measures are statistically different.

3 Related work

SSM and LSM have been widely used in order to evaluate the relatedness between terms specially in the biomedical domain. However, the combination of LSM and ontology-based SSM has rarely been investigated. Relevant initiatives were proposed in [32, 22] where authors have adapted WordNet based similarity measures to the biomedical domain. Lord et al. [22] have focused on Gene Ontology, a domain specific ontology, while Pedersen et al. [32] decided to be more generic and have grounded their work on SNOMED CT and reinforce their metrics with information derived from text corpora.

In the same line, the authors of [18] present a method for measuring the semantic similarity of texts combining a corpus-based measure of semantic word similarity and a normalized and modified version of the longest common subsequence string matching algorithm. They further evaluate the proposed metric on two well-accepted general corpus of text and show the added value of the combination with respect to comparable existing similarity measures.

In [16], the authors have investigated a broad range of semantic similarity measures to identify the core elements of the existing metrics with a particular focus on ontology-based measures. They further came up with a framework aiming at unifying the studied metrics and show the usability of the framework on the same corpus that is used in the work of Petersen et al. [32].

Ben Aouicha and Hadj Taieb [1] exploit the structure of an ontology to achieve a better semantic understanding of a concept. Their Information Content-based semantic similarity measure consists in expressing the IC by weighting each concept pertaining to the ancestors' subgraph modeling the semantics of a biomedical concept. They validated the added value of their work on three datasets including the one we are using in this work [32].

The work presented in [37] classifies ontology-based semantic similarity measures. They distinguish between edge-counting approaches, Feature-based approaches and intrinsic content ones. Moreover, they defined another ontology-based measure. Their metric considers as features the hierarchy of concepts structuring the ontology in order to evaluate the amount of dissimilarity between concepts. In other words, they assume that a term can be semantically different from other ones by comparing the set of concepts that subsume it.

Oliva et al. [29] have defined the SyMSS method consisting in assessing the influence of the syntactic aspect of two sentences in calculating the similarity. Sentences are expressed as a tree of syntactic dependences. It relies on the observation that a sentence is made up of the meaning of the words that compose it as well as the syntactic links among them. The semantics of these words is evaluated on WordNet that may be problematic for the biomedical domain since WordNet does not contain specific medical terms.

Ferreira et al. [13] defined a measure to evaluate the similarity between sentences taking into account syntactic, lexical and semantic aspects of the sentence and of the words composing it. In their work the semantics of words is obtained by querying the FrameNet database and not via ontologies.

Similarity measures have also been used for ontology matching. In [28], the authors have combined three kinds of different similarity measures: lexical-based, structure-based, and semantic-based techniques as well as information in ontologies including names, labels, comments, relations and positions of concepts in the hierarchy and integrating WordNet dictionary to align ontologies.

As shown in this section, existing work rarely consider the couple LSM/ontology-based SSM to measure similarity between terms. Moreover, the only combination that we have found exploit very specific or highly generic ontologies like GO and WordNet which are not tailored to evaluate medical terms. In this work we are proposing a combination of LSM/ontology-based SSM with ontology representing the medical domain at the right level of abstraction.

4 A new metric for measuring medical term similarity

In this section, we introduce the approach we propose to combine LSM and SSM in order to measure the similarity between medical terms. We continue with the description of the experimental setup we have defined to assess the added value of the proposed combination.

4.1 Combining lexical and semantic similarity measures

As illustrated in section 3, ontology-based SSM and LSM have rarely been combined to measure the similarity between medical terms. To this end, we propose a new metric that combines ontology-based SSM and LSM as a weighted arithmetic mean, see equation 4. It determines the similarity between labels of two concept c_i and c_j by applying the mentioned similarity measures over two respective concepts, e.g., *C0035078:Renal failure* $\leftrightarrow$ *C0035078:Kidney failure* and attributing weights to each similarity.

In equation 4, the values LSM_{score} and SSM_{score} represent the normalized similarity scores given by metrics like Levenshtein and Resnik 1995 GraSM. The variables α and τ are the weights, varying in the interval of [0.1, 1] with an incremental step of 0.1. It allows to change the contribution of each measure to calculate the final similarity. For instance, the configuration $\alpha = 0.8$ and $\tau = 0.3$ describe a situation where the semantic metrics are more precise than the Lexical one, but the Lexical one also contributes to the final similarity value.

$$simi(c_i, c_j) = \frac{(SSM_{score}(c_i, c_j) * \alpha) + (LSM_{score}(c_i, c_j) * \tau)}{\alpha + \tau} \tag{4}$$

4.2 Experimental assessment

To conduct an experimental evaluation of our new metric, we have designed a method that is based on the use of standard terminologies and existing benchmarks in order to compare our results with those generated using related work.

Terminologies

In our experiments, we have used Medical Subject Headings (MeSH) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) to test the SSM. These terminologies were extracted from the UMLS. Our experiments have been done using the versions 2009AA to 2014AA (excluding the AB versions). In contrast to existing comparable approaches, we consider the evolution of concepts.

Benchmarks

We have used the three datasets suggested by [24] to evaluate our approach. We first used *MayoSRS* [31]. It contains 101 pairs of concept labels together with a score assigned to each pair denoting their relatedness. The value of the score, ranging from 0 to 10, is determined by domain experts. 0 represents a low correlation while 10 denotes a strong one.

The second dataset we have used is a subset of *MayoSRS* [31] made up of 30 pairs of concept labels. For this dataset, a distinction is made between the two categories of experts: coders and physicians and the values of the relatedness score is ranging from 1 (unrelated) to 4 (almost synonymous).

The third dataset is the UMNSRS described in [30]. Bigger than the two previous ones, it is composed of 725 concept label pairs whose similarity was evaluated by four medical experts. The similarity score of each pair was given experimentally by users based on a continuous scale ranging from 0 to 1500.

Experimental configuration

Our aims are twofold. First, to evaluate the capacity of our approach to improve the similarity between pairs of concepts and second assess the stability of SSMs over time (with respect to the evolution of implemented ontologies). In consequence, we defined the three different configurations described hereafter:

- **Setup 1** aims at verifying the stability of semantic measures over time. To do so, we follow 3 steps: i) we prepared the gold standard and semantic measures to be used for our experiments, e.g., dataset: *MiniMayoSRS*, SSM: *Jiang Conrath* and *i*IC: *Sanchez* ii) we compute the similarity results using consecutive versions of MeSH and SNOMED CT and, iii) We computed and compared Fisher's Z-Transformation to verify if the obtained results are statistically different.

- **Setup 2** verifies the number of combinations (LSM × SSM) that outperforms the single use of LSM and SSM by making α and τ vary. For this configuration we fixed the ontologies and then we grouped the results from all datasets to verify how many combinations outperformed the single measures. This setup (dataset × ontology version × measures) has produced 25920 combinations. For the sake of readability, we only highlight the overall results and the top-10 cases in the following sections.
- **Setup 3** aims at pointing out the best combinations of metrics over the three datasets. To do so, we have tested two possibilities i) ranking with respect to the ontology. In this case, we fixed the ontology and then we analyzed the performance from all combined measures across the datasets. Here we combined all results and rank them[4]. ii) Overall ranking regardless of the ontology and datasets. In this step we combined the previous rank and verified what measures have higher rank with lowest standard deviation.

5 Results

The results regarding the influence of ontology evolution on SSMs i.e., **setup 1**, can be observed in Table 3. For this experiments we only used the UMNSRS dataset because, among all the datasets, UMNSRS was the only one to have at least one Z-Fisher transformation value $\rho \leq 0.05$, which is our threshold for considering statistical difference between the SSMs over time. The first column represent the iIC/SSM combination, the third column shows the versions of the ontology that have been tested. To build this column, we have considered all possible values of the set

$$\{(i,j)|i,j \in \{2009,2010,2011,2012,2013,2014\}, i<j\}$$

The last column contains the Z-fisher transformation values obtained by comparing the computed iIC/SSMs and the similarity score between two terms given by domain experts.

For a sake of readability we only show in the table the combinations for which we obtained the highest Z-Fisher values (in green) as well as the lowest ones (in red). As we never obtain a value below the 0.05 threshold, we can conclude that there is no statistical difference between the value generated by any of the combination which, in turn, demonstrate a stability of equation 4 with respect to the used ontology versions. In consequence, we can conclude that SSMs are not impacted by the evolution of the underlying ontology.

Regarding **setup 2**, i.e., the percentage of combinations that outperformed the single SSMs, we observed that 5939 combinations from the 25920 possibilities (23%) outperformed the single SSMs using SNOMED CT as ontology. Concerning MeSH, only 5280 combinations from the 25920 possibilities (20%) are better. For this set of experiments, we have used the three datasets as well

[4] https://pandas.pydata.org/pandas-docs/version/0.21/generated/pandas.Series.rank.html

Table 3. Stability of *i*IC/SSMs over time using UMNSRS dataset. We are considering the $\rho < 0.05$ as statistical significance. The red color indicates the lowest Z-Fisher values obtained in our experiments and the green indicates the highest ones.

***i*IC / SSM Measures**	**Years**	**Z-Fisher**
Seco / Jiang Conrath	2009 - 2010	0.519871
Seco / Jiang Conrath	2010 - 2011	0.880821
Seco / Jiang Conrath	2010 - 2014	0.277042
Seco / Jiang Conrath	2011 - 2012	0.991348
Seco / Jiang Conrath	2012 - 2013	0.991341
Seco / Jiang Conrath	2013 - 2014	0.356598
Ancestors Norm / Resnik GraSM	2009 - 2010	0.69417
Ancestors Norm / Resnik GraSM	2010 - 2011	0.832429
Ancestors Norm / Resnik GraSM	2011 - 2012	1.0
Ancestors Norm / Resnik GraSM	2012 - 2013	1.0
Ancestors Norm / Resnik GraSM	2013 - 2014	0.793019

as all the mentioned ontology versions. This reveals a relatively low added value of the random combination of LSM and SSM with respect to the single SSM. However, when we analyzed the metrics separately, as depicted in Table 4, we can observe that for few specific combinations, the results clearly outperform the single use of SSM. This is for instance the case for the combination $AnnoMap \times Zhou/ResnikGraSM$ that is better in 91.667% of the case showing a clear added value of combining LSM and SSM. Our experiments also show that AnnoMap was the most frequent LSM that appears in the most valuable combination. The similarity computed by AnnoMap [5], see equation 5, is based on the combined similarity score from different string similarity functions, in particular TF/IDF, Trigram and LCS (longest common substring). The definition of AnnoMap can explain our observations.

$$sim_{AnnoMap} = MAX(TF/IDF, TriGram, LCS) \tag{5}$$

Table 5 shows combinations that do not improve the single use of SSMs at all. We observed these poor results when we combined techniques that are not complementary. For instance, Block distance, Jaccard and TF/IDF consider strings as orthogonal spaces. When combined with *i*IC measures focused only in the positioning of concepts in an ontology, the results are not improved (compared with SSMs). Note that we are not pointing good or bad techniques, but we are looking for good combination. A typical example is *Sanchez (Normalized)* that is present in both tables 4 and 5, showing that, for instance, Block distance and Lin do not improve the outcomes, but AnnoMap and Resnik do.

Regarding **setup 3**, i.e., the overall rank for the best combinations, we experimentally verified that our approach performed better than the single SSMs regardless of the ontologies (here MeSH and SNOMED CT). We verified that the best performing combination for MeSH is (AnnoMap × Seco/Jiang Conrath)

Table 4. Percentage of Combinations that outperforms the classic SSMs

LSM	***i*IC / SSM**	**%**
AnnoMap	Zhou / Resnik GraSM	91.6667
	Resnik (Normalized) / Tversky *i*IC Ratio Model	91.6667
	Seco / Tversky *i*IC Ratio Model	91.6667
	Resnik (Normalized) / Resnik GraSM	87.5
	Sanchez (Normalized) / Resnik	87.5
	Seco / Resnik	87.5
	Harispe / Jiang Conrath	87.5
	Zhou / Resnik	87.5
	Seco / Resnik GraSM	87.5
	Sanchez (Normalized) / Resnik GraSM	87.5
Longest Common Substring	Sanchez (Normalized) / Tversky *i*IC Ratio Model	87.5
AnnoMap	Resnik (Normalized) / Resnik	87.5
Longest Common Substring	Harispe / Jiang Conrath	83.3333
LACP	Sanchez / Jian Conrath	83.3333

with $\alpha \in \{0.8, 1\}$ and $\tau \in \{0.4, 0.5\}$. We also observed that this combination is ranked in the top 3 best combinations but with different values for α and τ. For SNOMED CT, another combination is ranked as the most performing one. In the results the combination: (AnnoMap $\times$ Sanchez (Normalized)/Jiang Conrath) with $\alpha = 1$ and $\tau = 0.9$ was ranked first. The same behavior was observed for MeSH, where the top measure (AnnoMap $\times$ Seco/Jiang Conrath) with $\alpha = 0.8$ and $\tau = 0.5$ also appears in the top results.

The good performance of our approach is also observed when we combine all the ontologies and dataset to produce the overall rank. The final rank remains the same as we aimed at minimizing sum, average and standard deviation. In our results, we observed that (AnnoMap $\times$ Seco/Jiang Conrath) with $\alpha = 0.8$ and $\tau = 0.5$ is ranked in the top-8 in MeSH. In our experiments, the combination (AnnoMap $\times$ Seco/Jiang Conrath) with $\alpha = 0.8$ and $\tau = 0.5$ is therefore the best one.

The main difference we have observed is regarding the UMNSRS dataset, when we applied the combination (AnnoMap $\times$ Seco/Jiang Conrath) with $\alpha = 0.8$ and $\tau = 0.5$, the obtained similarity values were not greater than the single SSMs. It is due to the low Spearman's coefficient value obtained from the lexical measure [-0.140, -0.113]. We observed that combinations using other measures, for example, (LACP $\times$ Ancestors Norm/Lin GraSM) with $\alpha = 0.8$ and $\tau = 0.1$ show a Spearman's score of 0.462, and performs better than the single best SSM (0.456).

Table 5. Combined measures that failed to outperform the classic ones

LSM	*i*IC / SSM
Block distance	Resnik (Normalized) / Sim *i*IC
	Sanchez (Normalized) / Lin
Levenshtein	Max Linear / Mazandu
Bigram	Ancestors Norm / Resnik
TF/IDF	Ancestors Norm / Resnik GraSM
AnnoMap	Ancestors Norm / Jiang Conrath
Jaccard	Sanchez / Jiang Conrath
	Harispe / Mazandu
Longest Common Substring	Ancestors Norm / Tversky *i*IC Ratio Norm
JaroWinkler	Ancestors Norm / Resnik GraSM
LACP	Ancestors Norm / Sim *i*IC

6 Discussion

The results of our experimental framework presented in section 5 demonstrated that the combination of similarity measures, $LSM \times SSM$ formalized in equation 4, allows a better evaluation of medical terms relatedness. As explained in section 3, very few existing work proposed to combine LSM and ontology-based SSM. In this paper, we bridge this gap by showing experimentally that the couple LSM/ontology-based SSM is of added value for measuring the similarity of medical terms. Our proposal even allow to tune the importance of both measures (LSM and SSM) with the α and τ parameters depending on the context or on the used ontologies. As a result, when calculated using single SSMs, the relatedness between *Pain* and *Morphine* (CUIs: C0030193 and C0026549) we obtain a similarity score of 0.27 but with our approach the similarity score increases to 0.56 which better correspond to the score given by domain specialists in UMNSRS dataset.

Regarding the ratio of combinations that outperformed the single SSMs, we verified that when utilizing LSMs which compare strings as a orthogonal plane, like *TF/IDF*, *Jaccard* or *Block distance*; the Spearman's Rank Correlation is low. We believe that the reason for this lies in the loss of information contained in the prefixed term, e.g.,"Renal failure" $\leftrightarrow$ "Kidney failure". When we verified the scores obtained for the MiniMayoSRS dataset, i.e., (4.0), these terms were classified as strongly related. Therefore, similarity measures should compute a higher score for this pair. However, the mentioned methods only hits a maximum similarity of 0.5 for Cosine and 0.33 for Jaccard. On the other hand, methods like LACP, provides a similarity of 0.77 that matches the scores given by the domain specialists and increases the Spearman's Rank Correlation value. Similar behavior was observed when using *Ancestors Norm* as *i*IC. It computes scores according to the number of ancestors from a concept divided by the total number of concepts of an ontology, i.e., $iIC = nbAncestors(v)/nbConceptInOnto$. Thus, concepts with the same number of ancestors, but in different ontology regions will

have the same iIC. This limitation can be overcome if such metrics also consider sibling concepts. It plays a key role to determine the region of a concept in an ontology and is widely utilized in other domains, e.g., ontology prediction, mapping alignment as demonstrated in [33, 10].

Regarding the overall rank, we observed a significant difference in the rank of the top measures for both ontologies. When we changed the dataset, the top measures substantially dropped their rank from a dataset to another. Since we verified that the three datasets do not contain many concepts having the same label, UMNSRS is the one which has the most divergence between our scores and those given by domain experts. We explain our observations as following: i) the amount of cases to match with the domain specialties scores, around 175 in UMNSRS and 30 in the others dataset; ii) as discussed in [30] and also verified in our experiments, the relation *similarity* $\rightarrow$ *relatedness* is directional, i.e., the terms that are similar are also related but not the opposite, e.g., the semantic similarity of *Sinemet*$\leftrightarrow$*Sinemet* CUIs: C0023570 and C0006982 is 0.93, while *Pain*$\leftrightarrow$*Morphine* CUIs: C0030193 and C0026549 is 0.27.

Finally, we verified that the used SSMs are not significantly impacted by the evolution of underlying ontologies over time. However, the size of the datasets and the number of impacted concepts they contain may moderate our conclusion. We have seen that the percentage of impacted concepts in the dataset is 2,8%, while the percentage of impacted concepts in an ontology region, i.e., *subClass*, *superClass* and *Siblings* is 5.53%. Furthermore, the top-k combinations in our overall rank, implement the measures most impacted by the ontology evolution in **setup 1**. This result highlights that the evolution of the ontologies has a role during the process of calculating the SSMs similarity. Thus, future work on semantic similarity between ontology terms has to include other pairs of impacted concepts in their dataset to verify if the stability of these measures and the obtained rank will remain the same.

7 Conclusion

In this paper, we have introduced a method that combine lexical and ontology-based semantic similarity measures to better evaluate medical terms relatedness. We have evaluated it on three different and well-known datasets and have shown that it outperformed single use of semantic similarity measure and contribute to state-of-the-art as one of the first attempt to combine lexical and ontology-based semantic similarity measures. We also demonstrated that our proposal is not significantly affected by the evolution of underlying ontologies. In our future work, we will further evaluate our approach using larger datasets and put this metric in situation for maintaining semantic annotation impacted by ontology evolution valid over time.

References

1. Aouicha, M.B., Taieb, M.A.H.: Computing semantic similarity between biomedical concepts using new information content approach. Journal of biomedical informat-

ics **59**, 258–275 (2016)
2. Cardoso, S.D., Pruski, C., Da Silveira, M., Lin, Y.C., Groß, A., Rahm, E., Reynaud-Delaître, C.: Leveraging the impact of ontology evolution on semantic annotations. In: Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016. pp. 68–82. Springer (2016)
3. Cardoso, S.D., Reynaud-Delaître, C., Da Silveira, M., Pruski, C.: Combining rules, background knowledge and change patterns to maintain semantic annotations. In: AMIA annual symposium, Washington DC, USA, November 2017 (2017)
4. Cardoso, S.D., Reynaud-Delaître, C., Silveira, M.D., Lin, Y., Groß, A., Rahm, E., Pruski, C.: Evolving semantic annotations through multiple versions of controlled medical terminologies. Health and Technology (2018), http://dx.doi.org/10.1007/s12553-018-0261-3
5. Christen, V., Groß, A., Varghese, J., Dugas, M., Rahm, E.: Annotating medical forms using UMLS. In: International Conference on Data Integration in the Life Sciences. pp. 55–69. Springer (2015)
6. Couto, F., Pinto, S.: The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of Bioinformatics and Computational Biology **11**(5) (2013)
7. Couto, F.M., Silva, M.J., Coutinho, P.M.: Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In: Proceedings of the 14th ACM international conference on Information and knowledge management. pp. 343–344. ACM (2005)
8. Cross, V.: Tversky's parameterized similarity ratio model: A basis for semantic relatedness. In: Fuzzy Information Processing Society, 2006. NAFIPS 2006. Annual meeting of the North American. pp. 541–546. IEEE (2006)
9. Cross, V., Silwal, P., Chen, X.: Experiments varying semantic similarity measures and reference ontologies for ontology alignment. In: Extended Semantic Web Conference. pp. 279–281. Springer (2013)
10. Da Silveira, M., Dos Reis, J.C., Pruski, C.: Management of dynamic biomedical terminologies: Current status and future challenges. Yearbook of Medical informatics **10**(1), 125–133 (2015)
11. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: DyKOSMap: A framework for mapping adaptation between biomedical knowledge organization systems. Journal of biomedical informatics **55**, 153–173 (2015)
12. Faria, D., Pesquita, C., Couto, F.M., Falcão, A.: Proteinon: A web tool for protein semantic similarity. Department of Informatics, University of Lisbon (2007)
13. Ferreira, R., Lins, R.D., Simske, S.J., Freitas, F., Riss, M.: Assessing sentence similarity through lexical, syntactic and semantic analysis. Computer Speech & Language **39**, 1–28 (2016)
14. Garla, V.N., Brandt, C.: Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics **13**(1), 261 (Oct 2012)
15. Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. International Journal of Computer Applications **68**(13), 13–18 (2013)
16. Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., Montmain, J.: A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. Journal of biomedical informatics **48**, 38–53 (2014)
17. Harispe, S.: Knowledge-based Semantic Measures: From Theory to Applications. Ph.D. thesis (2014)

18. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data **2**(2), 10:1–10:25 (2008)
19. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 (1997)
20. Li, B., Wang, J.Z., Feltus, F.A., Zhou, J., Luo, F.: Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. arXiv preprint arXiv:1001.0958 (2010)
21. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 296–304. ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998), http://dl.acm.org/citation.cfm?id=645527.657297
22. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics **19**(10), 1275–1283 (2003)
23. Mazandu, G.K., Mulder, N.J.: A topology-based metric for measuring term similarity in the gene ontology. Advances in bioinformatics **2012** (2012)
24. McInnes, B.T., Pedersen, T.: Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of Biomedical Informatics **46**(6), 1116 – 1124 (2013), special Section: Social Media Environments
25. Mihalcea, R., Corley, C., Strapparava, C., et al.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI. vol. 6, pp. 775–780 (2006)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
27. Morris, J.F.: A quantitative methodology for vetting dark network intelligence sources for social network analysis. Tech. rep., Air Force Inst of Tech Wright-Patterson AFB OH Graduate School of Engineering and Management (2012)
28. Nguyen, T.T., Conrad, S.: Ontology matching using multiple similarity measures. In: 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K). vol. 01, pp. 603–611 (Nov 2015), doi.ieeecomputersociety.org/
29. Oliva, J., Serrano, J.I., del Castillo, M.D., Iglesias, Á.: SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering **70**(4), 390–405 (2011)
30. Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., Melton, G.B.: Semantic similarity and relatedness between clinical terms: An experimental study. AMIA. Annual Symposium proceedings. AMIA Symposium **2010**, 572–6 (2010)
31. Pakhomov, S.V., Pedersen, T., McInnes, B., Melton, G.B., Ruggieri, A., Chute, C.G.: Towards a framework for developing semantic relatedness reference standards. Journal of Biomedical Informatics **44**(2), 251 – 265 (2011)
32. Pedersen, T., Pakhomov, S., Patwardhan, S., Chute, C.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics **40**, 288–299 (2007)
33. Pesquita, C., Couto, F.M.: Predicting the extension of biomedical ontologies. PLoS Comput Biol **8**(9), e1002630 (2012), http://dx.doi.org/10.1371/journal.pcbi.1002630
34. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, New York, NY, USA (1988)

35. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1. pp. 448–453. Morgan Kaufmann Publishers Inc. (1995)
36. Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Knowledge-Based Systems **24**(2), 297–303 (2011)
37. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications **39**(9), 7718–7728 (2012)
38. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in wordnet. In: ECAI. vol. 16, p. 1089 (2004)
39. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: Workshop on artificial intelligence for web search (AAAI 2000). vol. 58, p. 64 (2000)
40. Tversky, A.: Features of similarity. Psychological review **84**(4), 327 (1977)
41. Zhou, Z., Wang, Y., Gu, J.: A new model of information content for semantic similarity in wordnet. In: Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. vol. 3, pp. 85–89. IEEE (2008)