Measuring Conceptual Similarity in Ontologies: How Bad is a Cheap Measure?

Tahani Alsubait, Bijan Parsia, and Uli Sattler

School of Computer Science, The University of Manchester, United Kingdom
{alsubait,bparsia,sattler}@cs.man.ac.uk

Abstract. Several attempts have been made to develop similarity measures for ontologies. Motivated by finding problems in existing measures, we design a new family of measures to address these problems. We carry out an empirical study to explore how good the new measures are and to investigate how likely it is to encounter specific task-oriented problems when using a bad similarity measure.

1 Introduction

The process of assigning a numerical value reflecting the degree of resemblance between two ontology concepts or the so called conceptual similarity measurement is a core step in many ontology-related applications (e.g., ontology alignment [7], ontology learning [2]). Several attempts have been made to develop methods for measuring conceptual similarity in ontologies [22, 32, 23, 20, 4]. In addition, the problem of measuring similarity is well-founded in psychology and a number of similarity models have been already developed [6, 30, 26, 19, 29, 8, 12]. Rather than adopting a psychological model for similarity as a foundation, we noticed that some existing similarity measures for ontologies are ad-hoc and unprincipled. This can negatively affect the application in which they are used. However, in some cases, depending on how simple the ontology/task is, using a computationally expensive "good" similarity measure is no better than using a cheap "bad" measure. Thus, we need to understand the computational cost of the similarity measure and the cases in which it succeeds/fails. Unfortunately, to date, there has been no thorough investigation of similarity measures with respect to these issues.

For this investigation, we use an independently motivated corpus of ontologies (BioPortal[1] library) which contains over 300 ontologies that are used by the biomedical community which is a community that has a high interest in the similarity measurement problem [24, 31].

To understand the major differences between similarity measures w.r.t. the task in which they are involved in, we structure the discussion around the following three tasks:

- Task1: Given a concept C, retrieve all concepts D s.t. Similarity$(C, D) > 0$.
- Task2: Given a concept C, retrieve the N most similar concepts.

[1] http://bioportal.bioontology.org/

– Task3: Given a concept C and some threshold Δ, retrieve all concepts D s.t. Similarity$(C, D) > \Delta$.

We expect most similarity measures to behave similarly in the first task because we are not interested in the particular similarity values nor any particular ordering among the similar concepts. However, the second task gets harder as N gets smaller. In this case, a similarity measure that underestimates the similarity of some very similar concepts and overestimates the similarity of others can fail the task. In the third task, the actual similarity values matter. Hence, using the most accurate similarity measure is essential.

2 Preliminaries

We assume the reader to be familiar with DL ontologies. In what follows, we briefly introduce the relevant terminology. For a detailed overview, the reader is referred to [1]. The set of terms, i.e., concept, individual and role names, in an ontology $\mathcal{O}$ is referred to as its signature, denoted $\widetilde{\mathcal{O}}$. Throughout the paper, we use N_C, N_R for the sets of concept and role names respectively and $C_{\mathcal{L}}$ to denote a set of possibly complex concepts of a concept language $\mathcal{L}(\Sigma)$ over a signature Σ and we use the usual entailment operator $\models$.

3 Desired properties for similarity measures

Various psychological models for similarity have been developed (e.g., Geometric [26, 19], Transformational [17, 8] and Features [29] models). Due to the richness of ontologies, not all models can be adopted when considering conceptual similarity in ontologies. This is because many things are associated with a concept in an ontology (e.g., atomic subsumers/subsumees, complex subsumers/subsumees, instances, referencing axioms). Looking at existing approaches for measuring similarity in DL ontologies, one can notice that approaches which aim at providing a numerical value as a result of the similarity measurement process are mainly founded on feature-based models [29], although they might disagree on which features to consider.

In what follows, we concentrate on feature-based notions of similarity where the degree of similarity S_{CD} between objects C, D depends on features common to C and D, unique features of C and unique features of D. Considering both common and distinguishing features is a vital property of the features model.

Looking at existing approaches for measuring similarity in ontologies, we find that some of these approaches consider common xor unique features (rather than both) and that some approaches consider features that some instances (rather than all) of the compared concepts have. To account for all the features of a concept, we need to look at all (possibly complex) *entailed* subsumers of that concept. To understand these issues, we present the following example:

Example 1 *Consider the ontology:*

$$\{Animal \sqsubseteq Organism \sqcap \exists eats.\top, \quad Plant \sqsubseteq Organism,$$
$$Carnivore \sqsubseteq Animal \sqcap \forall eats.Animal, \quad Herbivore \sqsubseteq Animal \sqcap \forall eats.Plant,$$
$$Omnivore \sqsubseteq Animal \sqcap \exists eats.Animal \sqcap \exists eats.Plant\}$$

Please note that our "Carnivore" is also known as *obligate* carnivore. A good similarity function $Sim(\cdot)$ is expected to derive that Sim(Carnivore, Omnivore) > Sim(Carnivore, Herbivore) because the first pair share more **common** subsumers and have fewer **distinguishing** subsumers. On the one hand $Carnivore$, $Herbivore$ and $Omnivore$ are all subsumed by the following **common** subsumers (abbreviated for readability): $\{\top, Org, A, \exists e.\top\}$. In addition, $Carnivore$ and $Omnivore$ share the following **common** subsumer: $\{\exists e.A\}$. On the other hand, they have the following **distinguishing** subsumer: $\{\exists e.P\}$ while $Carnivore$ and $Herbivore$ have the following **distinguishing** subsumers: $\{\exists e.P, \forall e.P, \exists e.A, \forall e.A\}$. Here, we have made a choice to ignore (infinitely) many subsumers and only consider a select few. Clearly, this choice has an impact on $Sim(\cdot)$. Details on such design choices are discussed later. Note also that we only considered subsumers rather than subsumees. This is because common subsumees do not necessarily reflect commonalities. For example, consider the concept $\exists digests.Insect$ which is a subsumee of both $Animal$ and $Plant$. However, this concept does not reflect commonalities of animals and plants.

We refer to the property of accounting for both common and distinguishing features as rationality. In addition, the related literature refer to some other properties for evaluating similarity measures (e.g., equivalence closure, symmetry, triangle inequality, monotonicity, subsumption preservation, structural dependence). For a detailed overview, the reader is referred to [4, 16].

4 Overview of existing approaches

We classify existing similarity measures into two dimensions as follows.

Taxonomy vs. ontology based measures Taxonomy-based measures [22, 32, 23, 18, 14] only consider the taxonomic representation of the ontology (e.g., for DLs, we *could use* the inferred class hierarchy); hence only atomic subsumptions are considered (e.g., $Carnivore \sqsubseteq Animal$). In fact, this can be considered an approximated solution to the problem which might be sufficient in some cases. However, the user must be aware of the limitations of such approaches. For example, direct siblings are always considered equi-similar although some siblings might share more features/subsumers than others.

Ontology-based measures [4, 13, 16] take into account more of the knowledge in the underlying ontology (e.g., $Carnivore \sqsubseteq \forall eats.Animal$). These measures can be further classified into (a) structural measures, (b) interpretation-based measures or (c) hybrid. Structural measures [13, 16] first transform the compared concepts into a normal form (e.g., $\mathcal{EL}$ normal form or $\mathcal{ALCN}$ disjunctive normal form) and then compare the syntax of their descriptions. To avoid being purely syntactic, they first unfold the concepts w.r.t. the $TBox$ which limits the applicability of such measures to cyclic terminologies. Some structural measures [16] are applicable only to inexpressive DLs (e.g., $\mathcal{EL}$) and it is unclear how they can be extended to more expressive DLs. Interpretation-based measures mainly depend on the notion of canonical models (e.g., in [4] the canonical model based on the $ABox$ is utilised) which do not always exist (e.g., consider disjunctions).

Intensional vs. extensional based measures Intensional-based measures [22, 32, 13, 16] exploit the terminological part of the ontology while extensional-based measures [23, 18, 14, 4] utilise the set of individual names in an $ABox$ or instances in an external corpus. Extensional-based measures are very sensitive to the content under consideration; thus, adding/removing an individual name would change similarity measurements. These measures might be suitable for specific content-based applications but might lead to unintuitive results in other applications because they do not take concept definitions into account. Moreover, extensional-based measures cannot be used with pure terminological ontologies and always require representative data.

5 Detailed inspection of some existing measures

After presenting a general overview of existing measures, we examine in detail some measures that can be considered "cheap" options and explore their possible problems. In what follows, we use $S_{\text{Atomic}}(C)$ to denote the set of atomic subsumers for concept C. We also use $\text{Com}_{\text{Atomic}}(C, D), \text{Diff}_{\text{Atomic}}(C, D)$ to denote the sets of common and distingushing atomic subsumers respectively.

Rada et al. This measure utilises the length of the shortest path [22] between the compared concepts in the inferred class hierarchy. The essential problem here is that the measure takes only distinguishing features into account and ignores any possible common features.

Wu and Palmer. To account for both common and distinguishing features, Wu & Palmer [32] presented a different formula for measuring similarity, as follows:

$S_{\text{Wu \& Palmer}}(C, D) = \frac{2 \cdot |\text{Com}_{\text{Atomic}}(C,D)|}{2 \cdot |\text{Com}_{\text{Atomic}}(C,D)| + |\text{Diff}_{\text{Atomic}}(C,D)|}$

Although this measure accounts for both common and distinguishing features, it only considers atomic concepts and it is more sensitive to commonalties.

Resnik and other IC measures. In information theoretic notions of similarity, the information content $IC_C = -logP_C$ of a concept C is computed based on the probability (P_C) of encountering an instance of that concept. For example, $P_\top = 1$ and $IC_\top = 0$ since $\top$ is not informative. Accordingly, Resnik [23] defines similarity $S_{\text{Resnik}}(C, D)$ as:

$S_{\text{Resnik}}(C, D) = IC_{LCS}$

where LCS is the least common subsumer of C and D (i.e., the most specific concept that subsumes both C and D). IC measures take into account features that some instances of C and D have, which are not necessarily neither common nor distinguishing features of all instances of C and D. In addition, Resnik's measure in particular does not take into account how far the compared concepts are from their least common subsumer. To overcome this problem, two [18, 14] other IC-measures have been proposed:

$$S_{\text{Lin}}(C, D) = \frac{2 \cdot IC_{LCS}}{IC_C + IC_D}$$

$$S_{\text{Jiang\&Conrath}}(C, D) = 1 - IC_C + IC_D - 2 \cdot IC_{LCS}$$

6 A new family of similarity measures

Following our exploration of existing measures and their associated problems, we present a new family of similarity measures that addresses these problems. The new measures adopt the features model where the features under consideration are the subsumers of the concepts being compared. The new measures are based on Jaccard's similarity coefficient [11] which has been proved to be a proper metric (i.e., satisfies the properties: equivalence closure, symmetry and triangle inequality). Jaccard's coefficient, which maps similarity to a value in the range [0,1], is defined as follows (for sets of "features" A',B' of A,B, i.e., subsumers of A and B):

$$J(A,B) = \frac{|(A' \cap B')|}{|(A' \cup B')|}$$

We aim at similarity measures for general OWL ontologies and thus a naive implementation of this approach would be trivialised because a concept has infinitely many subsumers. To overcome this issue, we present some refinements for the similarity function in which we do not simply count all subsumers but consider subsumers from a set of (possibly complex) concepts of a concept language $\mathcal{L}$. More precisely, for concepts C, D an ontology $\mathcal{O}$ and a concept language $\mathcal{L}$, we set:

$$\begin{aligned} S(C,\mathcal{O},\mathcal{L}) &= \{D \in \mathcal{L}(\widetilde{\mathcal{O}}) \mid \mathcal{O} \models C \sqsubseteq D\} \\ \mathrm{Com}(C,D,\mathcal{O},\mathcal{L}) &= S(C,\mathcal{O},\mathcal{L}) \cap S(D,\mathcal{O},\mathcal{L}) \\ \mathrm{Union}(C,D,\mathcal{O},\mathcal{L}) &= S(C,\mathcal{O},\mathcal{L}) \cup S(D,\mathcal{O},\mathcal{L}) \\ \mathrm{Sim}(C,D,\mathcal{O},\mathcal{L}) &= \frac{|Com(C,D,\mathcal{O},\mathcal{L})|}{|Union(C,D,\mathcal{O},\mathcal{L})|} \end{aligned}$$

To design a new measure, it remains to specify the set $\mathcal{L}$. In what follows, we present some examples:

$$\begin{aligned} AtomicSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\mathrm{Atomic}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\mathrm{Atomic}}(\widetilde{\mathcal{O}}) = \widetilde{\mathcal{O}} \cap N_C. \\ SubSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\mathrm{Sub}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\mathrm{Sub}}(\widetilde{\mathcal{O}}) = Sub(\mathcal{O}). \\ GrSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\mathrm{G}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\mathrm{G}}(\widetilde{\mathcal{O}}) = \{E \mid E \in Sub(\mathcal{O}) \\ &\quad \text{or } E = \exists r.F, \text{for some } r \in \widetilde{\mathcal{O}} \cap N_R \text{ and } F \in Sub(\mathcal{O})\}. \end{aligned}$$

where $Sub(\mathcal{O})$ is the set of concept expressions in $\mathcal{O}$. $AtomicSim(\cdot)$ captures taxonomy-based measures since it considers atomic concepts only. The rationale of $SubSim(\cdot)$ is that it provides similarity measurements that are sensitive to the modeller's focus. It also provides a cheap (yet principled) way for measuring similarity in expressive DLs since the number of candidates is linear in the size of the ontology. To capture more possible subsumers, one can use $GrSim(\cdot)$. We have chosen to include only grammar concepts which are subconcepts or which take the form $\exists r.F$ to make experiments in the Empirical inspection Section more manageable. However, the grammar can be extended easily.

7 Approximations of similarity measures

Some of the presented examples for similarity measures might be practically inefficient due to the large number of candidate subsumers. For this reason, it would be nice if we can explore and understand whether a "cheap" measure can be a good approximation for a more expensive one. We start by characterising the properties of an approximation in the following definition.

Definition 1 *Given two similarity functions $Sim(\cdot)$,$Sim'(\cdot)$, and an ontology $\mathcal{O}$, we say that:*

- $Sim'(\cdot)$ preserves the order of $Sim(\cdot)$ *if* $\forall A_1, B_1, A_2, B_2 \in \widetilde{\mathcal{O}}$: $Sim(A_1, B_1) \leq Sim(A_2, B_2) \implies Sim'(A_1, B_1) \leq Sim'(A_2, B_2)$.
- $Sim'(\cdot)$ approximates $Sim(\cdot)$ from above *if* $\forall A, B \in \widetilde{\mathcal{O}}$: $Sim(A, B) \leq Sim'(A, B)$.
- $Sim'(\cdot)$ approximates $Sim(\cdot)$ from below *if* $\forall A, B \in \widetilde{\mathcal{O}}$: $Sim(A, B) \geq Sim'(A, B)$.

Consider $AtomicSim(\cdot)$ and $SubSim(\cdot)$. The first thing to notice is that the set of candidate subsumers for the first measure is actually a subset of the set of candidate subsumers for the second measure ($\widetilde{\mathcal{O}} \cap N_C \subseteq Sub(\mathcal{O})$). However, we need to notice also that the number of entailed subsumers in the two cases need not to be proportionally related. For example, if the number of atomic candidate subsumers is n and two compared concepts share $\frac{n}{2}$ common subsumers. We cannot conclude that they will also share half of the subconcept subsumers. They could actually share all or none of the complex subsumers. Therefore, the order-preserving property need not be always satisfied. As a concrete example, let the number of common and distinguishing atomic subsumers for C and D to be 2 and 4 respectively (out of 8 atomic concepts) and let the number of their common and distinguishing subsoncept subsumers to be 4 and 6 respectively (out of 20 subconcepts). Let the number of common and distinguishing atomic subsumers for C and E to be 4 and 4 respectively and let the number of their common and distinguishing subsoncept subsumers to be 4 and 8 respectively. In this case, $AtomicSim(C, D) = \frac{2}{6} = 0.33$, $SubSim(C, D) = \frac{4}{10} = 0.4$, $AtomicSim(C, E) = \frac{4}{8} = 0.5$, $SubSim(C, E) = \frac{4}{12} = 0.33$. Notice that $AtomicSim(C, D) < AtomicSim(C, E)$ while $SubSim(C, D) > SubSim(C, E)$. Here, $AtomicSim(\cdot)$ is not preserving the order of $SubSim(\cdot)$ and$AtomicSim(\cdot)$ underestimates the similarity of C,D and overestimates the similarity of C,E compared to $SubSim(\cdot)$.

A similar argument can be made to show that entailed subconcept subsumers are not necessarily proportionally related to the number of entailed grammar-based subsumers. We conclude that the above examples of similarity measures are, theoretically, none-approximations of each other. In the next section, we are interested in knowing the relation between these measures in practice.

8 Empirical inspection

Following our conceptual discussion on similarity measures in the previous sections, we explore the behaviour of some similarity measures in practice. Given a

range of similarity measures with different costs, we want to know how good an expensive measure is, its cost and the cases in which we are required to pay that cost to get a reasonable similarity measurement. Also we want to know how bad a cheap measure is, the specific problems associated with it and how likely it is for a cheap measure to be a good substitute for more expensive measures.

The empirical inspection constitutes two parts. First, we carry out a comparison between the three measures $GrSim(\cdot)$, $SubSim(\cdot)$ and $AtomicSim(\cdot)$ against human experts-based similarity judgments. In [21], IC-measures along with Rada measure [22] has been compared against human judgements using the same data set which is used in the current study. The previous study [21] has found that IC-measures are worse than Rada measure so we only include Rada measure in our comparison and exclude IC-measures. We also include another path-based measure with is Wu & Palmer [32]. Secondly, we further study in detail the behaviour of our new family of measures in practice. $GrSim(\cdot)$ is considered as the expensive and most precise measure in this study. We use $AtomicSim(\cdot)$ as the cheap measure as it only considers atomic concepts as candidate subsumers. Studying this measure can allow us to understand the problems associated with taxonomy-based measures as they all consider atomic subsumers only. Recall that taxonomy-based measures suffer from other problems that were presented in the conceptual inspection section. Hence, $AtomicSim(\cdot)$ can be considered the best candidate in its class since it does not suffer from these problems. We also consider $SubSim(\cdot)$ as a cheaper measure than $GrSim(\cdot)$ and more precise than $AtomicSim(\cdot)$ and we expect it to be a better approximation for $GrSim(\cdot)$ compared to $AtomicSim(\cdot)$. We excluded from the study instance-based measures since they require representative data which is not guaranteed to be present in our corpus of ontologies.

We have shown in the previous section that the above three measures are not proper approximations of each other. However, this might be not the case in practice as we will explore in the following experiment. To study the relation between the different measures in practice, we examine the following properties: (1) order-preservation, (2) approximation from above (3) approximation from below, (4) correlation and (5) closeness. With respect to these five properties, we study the relation between $AtomicSim(\cdot)$ and $SubSim(\cdot)$ and refer to this as AS, the relation between $AtomicSim(\cdot)$ and $GrSim(\cdot)$ and refer to this as AG, the relation between $SubSim(\cdot)$ and $GrSim(\cdot)$ and refer to this as SG. Properties 1-3 are defined in Definition 1. For correlations, we calculate Pearson's coefficient for the relation between each pair of measures. Finally, two measures are considered close if the following property holds: $|Sim_1(C,D) - Sim_2(C,D)| \leq \Delta$ where $\Delta = 0.1$ in the following experiment. We also compare the measures to human-based similarity judgements to confirm that the expensive measures can be more precise than the cheap ones.

8.1 Infrastructure

With respect to hardware, we used the following machine: Intel Quad-core i7 2.4GHz processor, 4 GB 1333 MHz DDR3 RAM, running Mac OS X 10.7.5.

As for the software we use OWL API v3.4.4 [9]. To avoid runtime errors caused by using some reasoners with some ontologies, a stack of freely available reasoners were utilised: FaCT++ [28], HermiT [25], JFact,[2] and Pellet [27].

8.2 Test data

The BioPortal corpus: The BioPortal library of biomedical ontologies has been used for evaluating different ontology-related tools such as reasoners [15], module extractors [5], justification extractors [10], to name a few. The corpus contains 365 user contributed ontologies (as in October 2013) with varying characteristics such as axiom count, concept name count and expressivity.

Ontology selection: A snapshot of the BioPortal corpus from November 2012 was used. It contains a total of 293 ontologies. We excluded 86 ontologies which have only atomic subsumptions as for such ontologies the behaviour of the considered measures will be identical, i.e., we already know that $AtomicSim(\cdot)$ is good and cheap. We also excluded 38 more ontologies due to having no concept names or due to run time errors. This has left us with a total of 169 ontologies.

Sampling: Due to the large number of concept names (565,661) and difficulty of spotting interesting patterns by eye, we calculated the pairwise similarity for a sample of concept names from the corpus. The size of the sample is 1,843 concept names with 99% confidence level. To ensure that the sample encompasses concepts with different characteristics, we picked 14 concepts from each ontology. The selection was not purely random. Instead, we picked 2 random concept names and for each random concept name we picked some neighbour concept names (i.e., 3 random siblings, atomic subsumer, atomic subsumee, sibling of direct subsumer). This choice was made to allow us to examine the behaviour of the considered similarity measures even with special cases such as measuring similarity among direct siblings.

8.3 Experiment workflow

Module extraction: After classifying the ontology, we pick a sample of 14 concept names. The selected 14 concept names are used as a seed signature for extracting a $\bot$-module [3]. For optimisation, rather than working on the whole ontology, the following steps are performed on the extracted module. One of the important properties of $\bot$-modules is that they preserve almost all the related subsumptions. There are 3 cases in which a $\bot$-module would miss some subsumers. The first case occurs when $\mathcal{O} \models C \sqsubseteq \forall s.X$ and $\mathcal{O} \models C \sqsubseteq \forall s.\bot$. The second case occurs when $\mathcal{O} \models C \sqsubseteq \forall s.X$ and $\mathcal{O} \models \forall s.X \equiv \top$. The third case occurs when $\mathcal{O} \models C \sqsubseteq \forall s.X$ and $\mathcal{O} \not\models C \sqsubseteq \exists s.X$. Since in all three cases $\forall s.X$ is a vacuous subsumer of C, we chose to ignore these, i.e., use $\bot$-modules without taking special measures to account for them.

Candidate subsumers extraction: In addition to extracting all atomic concepts in the $\bot$-module we recursively use the method getNestedClassExpressions() to extract all subconcepts from all axioms in the $\bot$-module. The extracted subconcepts are used to generate grammar-based concepts. For practical reasons,

[2] http://jfact.sourceforge.net/

we only generate concepts taking the form $\exists r.D$ s.t. $D \in Sub(\mathcal{O})$ and r a role name in the signature of the extracted $\bot$-module. Focusing on existential restrictions is justifiable by the fact that they are dominant in our corpus (77.89% of subconcepts) compared to other complex expression types (e.g., universal restrictions: 2.57%, complements: 0.14%, intersections: 13.89, unions: 2.05%).

Testing for subsumption entailments: For each concept C_i in our sample and each candidate subsumer S_j, we test whether the ontology entails that $C_i \sqsubseteq S_j$. If the entailment holds, subsumer S_j is added to the set of C_i's subsumers.

Calculating pairwise similarities: The similarity of each distinct pair in our sample is calculated using the three measures.

Comparison to human judgements: We picked one ontology (SNOMED-CT) from BioPortal corpus to carry out a comparison between the three measures against human experts-based similarity judgments. The reason for choosing this particular ontology is the availability of data showing experts' judgements for the similarity between some concepts from that ontology. In [21], the similarity of 30 pairs of clinical terms is rated by medical experts. We include in our study 19 pairs out of the 30 pairs after excluding pairs that have at least one concept that has been described as an ambiguous concept in the ontology (i.e., is a subsumee of the concept ambiguous_concept). In [21], similarity values for two groups of experts (physicians and coders) are presented. We consider the average of physicians and coders similarity values in the comparison. For details regarding the construction of this dataset, the reader is referred to [21].

8.4 Results and discussion

How good is the expensive measure? Not surprisingly, $GrSim$ and $SubSim$ had the highest correlation values with experts' similarity (Pearson's correlation coefficient $r = 0.87, p < 0.001$). Secondly comes $AtomicSim$ ($r = 0.86$). Finally comes Wu & Palmer then Rada ($r = 0.81$, $r = 0.64$ respectively). Clearly, the new expensive measures are more correlated with human judgements which is expected as they consider more of the information in the ontology. The differences in correlation values might seem to be small but this is expected as SNOMED is an $\mathcal{EL}$ ontology and we expect differences to grow as expressivity increases.

Cost of the expensive measure: One of the main issues we want to explore in this study is the cost (in terms of time) for similarity measurement in general and the cost of the most expensive similarity measure in particular.

The average time per ontology taken to calculate grammar-based pairwise similarities was 2.3 minutes (standard deviation $\sigma = 10.6$ minutes, median $m = 0.9$ seconds) and the maximum time was 93 minutes for the Neglected Tropical Disease Ontology which is a $\mathcal{SRIQ}$ ontology with 1237 logical axioms, 252 concept names and 99 role names. For this ontology, the cost of $AtomicSim(\cdot)$ was only 15.545 sec and 15.549 sec for $SubSim(\cdot)$. 9 out of 196 ontologies took over 1 hour to be processed. One thing to note about these ontologies is the high number of logical axioms and role names. However, these are not necessary conditions for long processing times. For example, the Family Health History Ontology has 431 role names and 1103 logical axioms and was processed in less than 13 sec. Clearly, $GrSim(\cdot)$ is far more costly than the other two measures.

This is why we want to know how good/bad a cheaper measure can be. These reported times include module extraction and ontology classification times.

Approximations and correlations: Regarding the relations (AS, AG, SG) between the three measures, we want to find out how frequently can a cheap measure be a good approximation for/have a strong correlation with a more expensive measure. Recall that we have excluded all ontologies with only atomic subsumptions from the study. However, in 12% of the ontologies the three measures were perfectly correlated ($r = 1, p < 0.001$) mostly due to having only atomic subsumptions in the extracted module (except for three ontologies which have more than atomic subsumptions). In addition to these perfect correlations for all the three measures, in 11 more ontologies the relation SG was a perfect correlation ($r = 1, p < 0.001$) and AS and AG were very highly correlated ($r \geq 0.99, p < 0.001$). These perfect correlations indicate that, in some cases, the benefit of using an expensive measure is totally neglectable.

In about fifth of the ontologies (20.47%), the relation SG was a very high correlation ($1 > r \geq 0.99, p < 0.001$) within which 5 ontologies were 100% order-preserving and approximating from below. In this category, in 22 ontologies the relation SG was 100% close. As for the relation AG, in only 8% of the ontologies the correlation was very high.

In nearly half of the ontologies (48.54%), the correlation for SG was considered medium ($0.99 > r \geq 0.90, p < 0.001$). And in 11% of the ontologies, the correlation for SG was considered low ($r < 0.90, p < 0.001$) with ($r = 0.63$) as the lowest correlation value. In comparison, the correlation for AG was considered medium in 38% of the ontologies and low in 32.75% of the ontologies.

As for the order-preservations, approximations from above/below and closeness for the relations AG and SG, we summarise our findings in the following table. Not surprisingly, $SubSim(\cdot)$ is more frequently a better approximation to $GrSim(\cdot)$ compared to $AtomicSim(\cdot)$. Although one would expect that the

	Order-preservations	Approx. from below	Approx. from above	Closeness
AG	32	32	37	28
SG	44	49	42	56

Table 1: Ontologies satisfying properties of approximation

properties of an ontology have an impact on the relation between the different measures used to compute the ontology's pairwise similarities, we found no indicators. With regard to this, we categorised the ontologies according to the degree of correlation (i.e., perfect, high, medium and low correlations) for the SG relation. For each category, we studied the following properties of the ontologies in that category: expressivity, number of logical axioms, number of concept names, number of role names, length of the longest axiom, number of subconcept expressions. For ontologies in the perfect correlation category, the important factor was having a low number of subconcepts. In this category, the length of the longest axiom was also low (≤ 11, compared to 53 which is the maximum length of the longest axiom in all the extracted modules from all ontologies). In addition, the expressivity of most ontologies in this category was $\mathcal{AL}$. Apart from this category, there were no obvious factors related to the other categories.

How bad is a cheap measure? To explore how likely it is for a cheap measure to encounter problems (e.g., fail one of the tasks presented in the intro-

duction), we examine the cases in which a cheap measure was not an approximation for the expensive measure. AG and SG were not order-preserving in 80% and 73% of the ontologies respectively. Also, they were not approximations from above nor from below in 72% and 64% of the ontologies respectively and were not close in 83% and 66% of the ontologies respectively.

If we take a closer look at the African Traditional Medicine ontology for which the similarity curves are presented in Figure 1, we find that SG is 100% order-preserving while AG is only 99% order-preserving. Note that for presentation purposes, only part of the curve is shown. Both relations were 100% approximations from below. As for closeness, SG was 100% close while AG was only 12% close. In order to determine how bad are $AtomicSim(\cdot)$ and $SubSim(\cdot)$ as cheap approximations for $GrSim(\cdot)$, we study the behaviour of these measures w.r.t. the three tasks presented in the introduction. Both cheap measures would succeed in performing task 1 while only $SubSim(\cdot)$ can succeed in task 2 (1% failure chance for $AtomicSim(\cdot)$). For task 3, there is a higher failure chance for $AtomicSim(\cdot)$ since closeness is low (12%).

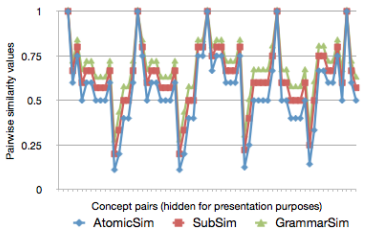

Fig. 1: African Traditional Medicine

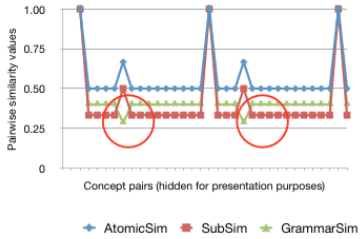

Fig. 2: Platynereis Stage

As another example, we examine the Platynereis Stage Ontology for which the similarity curves are presented in Figure 2. In this ontology, both AG and SG are 75% order-preserving. However, AG was 100% approximating from above while SG was 85% approximating from below (note the highlighted red spots). In this case, both $AtomicSim(\cdot)$ and $SubSim(\cdot)$ can succeed in task 1 but not always in tasks 2 & 3 with $SubSim(\cdot)$ being worse as it can be overestimating in some cases and underestimating in other cases.

In general, both measures are good cheap alternatives w.r.t. task 1. However, $AtomicSim(\cdot)$ would fail more often than $SubSim(\cdot)$ when performing tasks 2/3.

9 Conclusion and future research directions

In conclusion, no obvious indicators were found to inform the decision of choosing between a cheap or expensive measure based on the properties of an ontology. However, the task under consideration and the error rate allowed in the intended application can help. In general, $SubSim(\cdot)$ seems to be a good alternative to the expensive $GrSim(\cdot)$. First, it is restricted in a principled way to the modeller's focus. Second, it has less failure chance in practise compared to $AtomicSim(\cdot)$.

As for our future research directions, we aim to extend the study by looking deeply at the possible causes of failure and run the measures on some ontologies as they are instead of some modules of them to see how well they scale.

References

1. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. (eds.) Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, second edition, 2007.
2. T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390405, 2010.
3. B. Cuenca Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular reuse of ontologies: Theory and practice. *J. of Artificial Intelligence Research*, 31:273318, 2008.
4. C. d'Amato, S. Staab, and N. Fanizzi. On the Inuence of Description Logics Ontologies on Conceptual Similarity. In *EKAW '08 Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns*, 2008.
5. C. Del Vescovo, P. Klinov, B. Parsia, U. Sattler, T. Schneider, and D. Tsarkov. Syntactic vs. semantic locality: How good is a cheap approximation? In *WoMO 2012*, 2012.
6. W. K. Estes. Statistical theory of distributional phenomena in learning. *Psychological Review*, 62:369–377, 1955.
7. J. Euzenat and P. Shvaiko. Ontology matching. *Springer-Verlag*, 2007.
8. U Hahn, N Chater, and LB Richardson. Similarity as transformation. *COGNITION*, 87 (1):1 – 32, 2003.
9. M. Horridge and S. Bechhofer. The owl api: A java api for working with owl 2 ontologies. In *In Proceedings of the 6th International Workshop on OWL: Experiences and Directions (OWLED)*, 2009.
10. M. Horridge, B. Parsia, and U. Sattler. Extracting justifications from bioportal ontologies. *International Semantic Web Conference*, 2:287–299, 2012.
11. P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901.
12. W. James. The principles of psychology. dover: New york. (original work published 1890), 1890/1950.
13. K. Janowicz. Sim-dl: Towards a semantic similarity measurement theory for the description logic alcnr in geographic information retrieval. In *SeBGIS 2006, OTM Workshops 2006, pages 1681-1692*, 2006.
14. J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics, Taiwan*, 1997.
15. Y.-B. Kang, Y.-F. Li, and S. Krishnaswamy. Predicting reasoning performance using ontology metrics. In *ISWC 2012 Lecture Notes in Computer Science. Volume 7649*, 2012.
16. K. Lehmann and A. Turhan. A framework for semantic-based similarity measures for ELH-concepts. *JELIA 2012*, pages 307–319, 2012.
17. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
18. D. Lin. An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*, San Francisco, CA, 1998. Morgan Kaufmann.
19. R. M. Nosofsky. Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43:25–53, 1992.
20. R. Othman, S. Deris, and R. Illias. A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of BiomedInform*, 23, 2007.

21. T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 30(3):288–299, 2007.
22. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. In *IEEE Transaction on Systems, Man, and Cybernetics*, volume 19, page 1730, 1989.
23. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI95)*, volume 1, pages 448–453, 1995.
24. A. Schlicker, FS. Domingues, J. Rahnenfu hrer, and T. Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7, 2006.
25. R. Shearer, B. Motik, and I. Horrocks. HermiT: A highly-efficient OWL reasoner. In *Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED-08EU)*, 2008.
26. R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.
27. E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2), 2007.
28. D. Tsarkov and I. Horrocks. FaCT++ description logic reasoner: System description. In *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR)*, 2006.
29. A. Tversky. Features of similarity. *Psycological Review by the American Psycological Association, Inc.*, 84(4), July 1977.
30. A.R. Wagner. Evolution of an elemental theory of pavlovian conditioning. *Learning and Behavior*, 36:253–265, 2008.
31. JZZ. Wang, Z. Du, R. Payattakool, PSS. Yu, and CFF. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 2007.
32. Z. Wu and MS. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, page 133138, 1994.