A Framework for Semantic-based Similarity Measures for $\mathcal{ELH}$-Concepts

Karsten Lehmann[1,2] and Anni-Yasmin Turhan[3]⋆

[1] Optimisation Research Group, NICTA; Karsten.Lehmann@nicta.com.au
[2] Artificial Intelligence Group, Australian National University
[3] Institute for Theoretical Computer, TU Dresden, Germany;
turhan@tcs.inf.tu-dresden.de

Abstract. Similarity measures for concepts written in Description Logics (DLs) are often devised based on the syntax of concepts or simply by adjusting them to a set of instance data. These measures do not take the semantics of the concepts into account and can thus lead to unintuitive results. It even remains unclear how these measures behave if applied to new domains or new sets of instance data.

In this paper we develop a framework for similarity measures for $\mathcal{ELH}$-concept descriptions based on the semantics of the DL $\mathcal{ELH}$. We show that our framework ensures that the measures resulting from instantiations fulfill fundamental properties, such as equivalence invariance, yet the framework provides the flexibility to adjust measures to specifics of the modelled domain.

1 Introduction

Concept similarity measures map a pair of concepts from an ontology to a value between 0 and 1 indicating how similar the concepts are. These measures are an important means to discover similar concepts in ontologies. In bio-medical ontology-based applications, for example the Gene ontology [5], they are employed to discover functional similarities of genes. Furthermore, concept similarity measures are used in ontology alignment algorithms [9].

A common approach to find and evaluate similarity measures is to have test data and to tune a similarity measure until it matches the results of a human expert. The disadvantage of this approach is that the behavior of such a measure is hard to predict when applied to new test data, or when used for ontologies modeling a different domain. As a consequence an ontology developer cannot competently decide whether a measure obtained in this way is suitable for a particular task.

Description Logics (DLs) are a family of knowledge representation formalisms with formal semantics. A good similarity measure for DL concepts should take the semantics of the underlying formalism into account, instead of assessing

⋆ Partially supported by the German Research Foundation (DFG) in the Collaborative Research Center 912 "Highly Adaptive Energy-Efficient Computing".

similarity in a purely syntactical way. Similarity measures are often tailored for particular applications. Thus, one similarity measure will hardly meet the needs of all applications.

In [8] the intended behavior of a measure was discussed and partially captured in terms of properties. These properties were adapted from metric spaces which are related to similarity measures. We follow this approach to address the problems mentioned above. We extend this set of properties by including DL specific ones and mathematically describe those from [8] in terms of DL. The formalization of the properties allows us to prove whether or not an obtained measure has the desired properties. Additionally, we investigate existing DL similarity measures to determine which of the properties they fulfill. We then propose the framework *simi* for similarity measures for $\mathcal{ELH}$-concepts. If instantiated with the right functions and operators as building blocks, *simi* yields measures for which (most of) the formalized properties can be guaranteed. At the same time the framework retains flexibility as it allows users to choose from the list which properties the resulting measure should have and to build their measure accordingly. Furthermore, the resulting similarity measures can be computed efficiently, provided that functions employed can be computed efficiently as well.

Our choice for the DL $\mathcal{ELH}$ is motivated by the fact that large, well-known biomedical ontologies such as the Gene Ontology [5] or Snomed [21] are written in (extensions of) $\mathcal{ELH}$. Furthermore, $\mathcal{ELH}$ is a fragment of the DL that corresponds to the OWL 2 EL profile, which is part of the W3C standard for an ontology language for the Semantic Web [23, 19].

The paper is structured as follows: we start with preliminaries on DLs. In Section 3, we introduce the set of properties desirable for similarity measures and in Section 4 we devise a framework for constructing similarity measures that fulfill (most of) the introduced properties. The paper ends with conclusions and directions for future work.

2 Preliminaries

In this section we introduce the basic notions of DLs. For a thorough introduction see [1]. Starting from a finite set of concept names N_C and a finite set of role names N_R, complex concepts can be defined using *concept constructors.* Let A, $B \in N_C$, then $\mathcal{EL}$*-concepts* are formed according to the following syntax rule:

$$C ::= \top \mid A \mid C \sqcap D \mid \exists r.C$$

where $r \in N_R$ and C, D denote arbitrary $\mathcal{EL}$-concepts. A concept of the form $\exists r.C$ is called an *existential restriction* and one of the from $C \sqcap D$ is called a *conjunction.* We call the DL, that only offers conjunction as a concept constructor, $\mathcal{L}_0$. The semantics of concepts is given in terms of interpretations. An *interpretation* $\mathcal{I} = (\Delta, \cdot)$ consists of the *interpretation domain* $\Delta^{\mathcal{I}}$ a non-empty set and an *interpretation function* $\cdot^{\mathcal{I}}$ that assigns role names to binary relations on $\Delta^{\mathcal{I}}$ and concepts to subsets of $\Delta^{\mathcal{I}}$. The top-concept $\top$ is mapped to $\Delta^{\mathcal{I}}$. The extension of the interpretation function to conjunctions is $(C \sqcap D)^{\mathcal{I}} := C^{\mathcal{I}} \cap D^{\mathcal{I}}$ and to existential restrictions $(\exists r.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} \mid \exists e \in \Delta^{\mathcal{I}} : (d, e) \in r^{\mathcal{I}}$ and $e \in C^{\mathcal{I}}\}$.

A *concept definition* assigns a concept name to a complex concept. We call $A = C$ a *concept definition* and $A \sqsubseteq C$ a *primitive concept definition.* A finite set of (possibly primitive) concept definitions is a *TBox* $\mathcal{T}$. If the (primitive) definitions in a TBox are acyclic and do not contain multiple definitions we call the TBox *unfoldable.* Concept names occurring on the left-hand side of a definition are called *defined concepts.* All other concept names are called *primitive concepts.* Let $s, r \in N_R$. A *role inclusion axiom* (RIA) is a statement of the form: $r \sqsubseteq s$. The DL that extends $\mathcal{EL}$ by RIAs is called $\mathcal{ELH}$. An interpretation is a model for $s \sqsubseteq r$ iff $s^{\mathcal{I}} \subseteq r^{\mathcal{I}}$. A finite set of RIAs is called *RBox* $\mathcal{R}$. An interpretation $\mathcal{I}$ is a model of the TBox $\mathcal{T}$ (RBox $\mathcal{R}$) iff it satisfies all its concept definitions (RIAs). We write $s \sqsubseteq_{\mathcal{R}} r$, if $s^{\mathcal{I}} \subseteq r^{\mathcal{I}}$ holds in all models of $\mathcal{R}$ and $s \equiv_{\mathcal{R}} r$, if $s \sqsubseteq_{\mathcal{R}} r$ and $r \sqsubseteq_{\mathcal{R}} s$ hold.

A DL *knowledge base* (KB) $\mathcal{K}$ consists of the *TBox* and the *RBox* and we say that an interpretation $\mathcal{I}$ is a *model* of $\mathcal{K}$, if it is a model for the corresponding TBox and RBox.

Based on the semantics of concepts, reasoning problems can be defined. The concept C is *subsumed* by the concept D w.r.t. the KB $\mathcal{K}$ ($C \sqsubseteq_{\mathcal{K}} D$) iff $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds for all models $\mathcal{I}$ of $\mathcal{K}$. C and D are *equivalent* w.r.t. $\mathcal{K}$ ($C \equiv_{\mathcal{K}} D$) iff $C \sqsubseteq_{\mathcal{K}} D$ and $D \sqsubseteq_{\mathcal{K}} C$.

For a given concept C, *expansion* replaces exhaustively all occurrences of defined concepts in C by the right-hand sides of their concept definitions. For unfoldable TBoxes all reasoning problems can be reduced to reasoning for concepts by using expansion of concepts w.r.t. the TBox [1].

We denote the set of concepts for a specific DL $\mathcal{L}$ with $\mathcal{C}(\mathcal{L})$, e.g., $\mathcal{C}(\mathcal{EL})$ is the set of all $\mathcal{EL}$-concepts. We call concepts that are either concept names or existential restrictions *atoms* and denote the set of atoms by N_A.

For $\mathcal{EL}$-concepts a unique normal form (modulo associativity and commutativity), was given in [2], which we extend to $\mathcal{ELH}$-concepts in presence of RBoxes. To treat equivalent roles, we define $[r] = \{s \in N_R \mid r \equiv_{\mathcal{R}} s\}$ and fix a function f that picks one role r_i from each equivalence class and replaces each occurrence of a role from $[r_i]$ with r_i. Given an RBox $\mathcal{R}$ and an $\mathcal{ELH}$-concept C, C is in *$\mathcal{ELH}$-normal form*, if the following 4 rules have been applied exhaustively to the concept C and its subconcepts:

1. $A \sqcap \top \longrightarrow A$,
2. $A \sqcap A \longrightarrow A$,
3. $\exists r.C' \longrightarrow \exists f([r]).C'$,
4. $\exists r.C' \sqcap \exists s.D' \longrightarrow \exists r.C'$ if $r \sqsubseteq_{\mathcal{R}} s$ and $C' \sqsubseteq D'$

The transformation of $\mathcal{ELH}$-concepts into $\mathcal{ELH}$-normal form can be done in polynomial time.

3 Properties for Concept Similarity Measures

Formally, a *concept similarity measure sim* is a function mapping from pairs of $\mathcal{ELH}$-concepts to the interval $[0, 1]$. To identify properties of similarity measures for concepts, [8] used *metric spaces* as a starting point, which was also done in other areas of similarity research (see [22, 16, 17, 20]). A metric can be interpreted as a *dissimilarity measure*. The distance represents the dissimilarity between two objects—the lower their distance, the higher the similarity. Using a metric d, we can obtain a similarity function s by defining $s(a, b) := 1 - d(a, b)$. If we adapt the properties of a metric accordingly, we obtain the following properties for similarity functions.

Definition 1. *Let D be a set. A function $s : D \times D \longrightarrow [0, 1]$ is called a* similarity function *for D iff for all $a, b, c \in D$ holds*

1. $s(a, b) = 1 \iff a = b$, (identity of indiscernibles)
2. $s(a, b) = s(b, a)$, *and* (symmetry)
3. $1 + s(a, b) \geq s(a, c) + s(c, b)$ (triangle inequality).

Next we present definitions of properties of concept similarity measures and the underlying intuitions for these properties. We start with a formal definition of the properties and discuss each of them afterwards.

Definition 2. *Let $C, D, E \in \mathcal{C}(\mathcal{ELH})$. A similarity measure sim is*

1. symmetric *iff* $sim(C, D) = sim(D, C)$.
2. *fulfilling the* triangle inequality *property iff*
$$1 + sim(D, E) \geq sim(D, C) + sim(C, E).$$
3. equivalence invariant *iff* $C \equiv D \implies sim(C, E) = sim(D, E)$.
4. equivalence closed *iff* $sim(C, D) = 1 \iff C \equiv D$.
5. subsumption preserving *iff* $C \sqsubseteq D \sqsubseteq E \implies sim(C, D) \geq sim(C, E)$.
6. reverse subsumption preserving *iff* $C \sqsubseteq D \sqsubseteq E \implies sim(C, E) \leq sim(D, E)$.
7. structurally dependent *iff for all sequences $(C_n)_n$ of atoms with $\forall i, j \in \mathbb{N}, i \neq j : C_i \not\sqsubseteq C_j$ the concepts*
$$D_n := \bigsqcap_{i \leq n} C_i \sqcap D \quad \text{and} \quad E_n := \bigsqcap_{i \leq n} C_i \sqcap E$$
fulfill the condition $\lim_{n \to \infty} sim(D_n, E_n) = 1$.

The properties 1. to 4. are adopted from the literature, whereas to the best of our knowledge the properties 5. to 7. are introduced for DLs in this paper.

Symmetry is a rather controversial property for similarity in general—while some consider it essential [18], cognitive sciences seems to favor an asymmetric notion of similarity [22, 4]. Even for DL concepts Janowicz et al. [13, 12] prefer asymmetry (but devise symmetric measures), whereas most [3, 7, 6, 10, 8] consider it a fundamental property of similarity of concepts.

Triangle property is inherited from metrics. Two papers mentioned triangle inequality in the context of DLs: [8] argues in favor, while [12] argue against it, because of Tversky's [22] work.

DLs allow the same thing to be described in different ways. Two concepts can be syntactically different and yet semantically equivalent. A similarity measure for complex concepts should depend on the semantics rather than the syntax of the concepts to measure.

Equivalence invariance ensures that two equivalent concepts have the same similarity towards a third concept. Equivalence invariance is widely accepted as a necessary property for measures for DL concepts ([13, 12, 6, 8]). Yet we found that the methods used to ensure equivalence invariance were not always sound (see Section 3.1).

Equivalence closure holds for a similarity measure if and only if two concepts are totally similar if and only if they are equivalent. This corresponds with the idea that indiscernible things are identical. Equivalence closure is considered to be a basic property for concept similarity measures [8, 12] especially since it is inherited from metrics.

One asset of DLs is their reasoning services. An intuitive idea is to characterize similarity of concepts in terms of these services. The subsumption relation yields a total partial order on concepts. Consider the case where $C, D, E \in \mathcal{C}(\mathcal{ELH})$ and $C \sqsubseteq D \sqsubseteq E$. A natural requirement of similarity measures is to reflect this constellation.

Subsumption preservation expresses that the similarity of C and D is higher than the one of C and E because C is 'closer' to D than to E.

Reverse subsumption preservation states likewise that the similarity of D and E is higher than the similarity of C and E, since E is 'closer' to D than to C.

In [15] we also employ the reasoning service least common subsumer to capture the characteristics of total dissimilarity of concept similarity.

Tversky [22] presents the *feature model*, where an object is described by a set of features. The similarity of two objects is measured by a relation between the number of common features of both objects and the number of unique features of each object. The basic rule is that if

1. the number of common features increases and
2. the number of uncommon features is constant

then the similarity must increase.

Structural dependence reflects this basic rule. Concepts are our objects to compare and the atoms of a conjunction represent the features of the object. The intuition is that the more features (atoms) two complex concepts share, the higher their similarity should be.

For a more detailed explanation of the last property and for a presentation of examples illustrating the above properties see [15].

Table 1. Overview of similarity measures and their properties

	symm.	triang.	eq. inv.	eq. cl.	subs.	rev. subs.	struc. dep.	DL
simi	✓	-	✓	✓	✓	-	✓	$\mathcal{ELH}$
Jacc [11]	✓	✓	✓	✓	✓	✓	✓	$\mathcal{L}_0$
[13]	✓	-	-	-	-	-	✓	$\mathcal{SHI}$
[12]	✓	-	-	-	-	-	✓	$\mathcal{ALCHQ}$
[7]	-	-	-	-	-	-	-	$\mathcal{ALC}$
[10]	✓	-	✓	-	✓	✓	-	$\mathcal{ALN}$
[6]	✓	-	✓	-	✓	✓	-	$\mathcal{ALC}$
[8]	✓	-	✓	-	✓	✓	-	$\mathcal{ALE}$

3.1 Inspecting Existing Concept Similarity Measures

We distinguish two kinds of concept similarity measures: structural measures and interpretation based measures. *Structural measures* are defined using the syntax of the concepts to measure. Since conjunction and disjunction are commutative and associative, these measures are invariant to the order of the atoms in a conjunction or disjunction. The measures differ regarding the similarity of primitive concepts: [12] uses the TBox whereas [7] and [10] use the canonical interpretation which takes the set of ABox individuals as the interpretation domain (for an introduction to ABoxes see [1]).

Interpretation based measures are defined using interpretations and cardinality, instead of the syntax of the (possibly) complex concepts to measure. Therefore, they are trivially equivalence invariant. The two interpretation based measures we investigated [6, 8] are using the canonical interpretation $\mathcal{I}_{\mathcal{A}}$. These measures need a populated and representative ABox as a significant domain.

Table 1 presents an overview of similarity measures for concepts written in different DLs (including our measure *simi* to be defined in Section 4) and whether or not they fulfill the properties from Definition 2. The proofs can be found in [15]. The first four measures are purely structural measures. The next two are structural measures which use the canonical interpretations to measure primitives. The last two are purely interpretation based measures.

We included the Jaccard index [11], which is originally a set measure, here adapted to $\mathcal{L}_0$. Interestingly, this is the only measure of those investigated that fulfills the triangle inequality.

Our thorough investigation of the similarity measures defined in the literature showed that defining a similarity measure that fulfills most of the properties from Definition 2 is by no means a trivial task—in particular if the DL allows the use of roles, as the lightweight DL $\mathcal{ELH}$ already does.

4 Developing Concept Similarity Measures for $\mathcal{ELH}$

We present *simi*, a framework for similarity measures for concepts written in the DL $\mathcal{ELH}$ based on the semantics of the logic. It operates on (complex) concepts and an RBox $\mathcal{R}$, which contains role inclusion axioms. If concepts to be processed contain concepts defined in an unfoldable TBox $\mathcal{T}$, we assume that these concepts are expanded w.r.t. $\mathcal{T}$, i.e., all concept names occurring in them are primitive names.

Another preprocessing step is to transform the concepts into $\mathcal{ELH}$-normal form (defined in Section 2). Concepts in this normal form are unique (modulo associativity and commutativity), which ensures that *simi* (and any other measure processing concepts in this normal form) is equivalence invariant. We assume for the remainder of the paper that the concepts are in $\mathcal{ELH}$-normal form.

The framework *simi* constructs similarity measures from several free parameters, i.e., it allows functions to be combined in such a way that, if these functions fulfill certain properties, then the resulting similarity measure can be shown to fulfill all properties from Definition 2 except reverse subsumption preserving and the triangle inequality. Furthermore, it can be computed efficiently.

Simi is inspired by the Jaccard index and it is a conservative extension of the Jaccard index, in the sense that $\forall C, D \in \mathcal{C}(\mathcal{L}_0) : simi(C, D) = Jacc(C, D)$ (proven in [15]). Another inspiration is the equivalence of concepts, which can be regarded as a trivial similarity measure: the similarity of two concepts is 1 if they are equivalent and 0 otherwise. To determine if $C \equiv D$ is true, one can use the subsumption test to find out whether or not $C \sqsubseteq D$ and $D \sqsubseteq C$ are true. We generalize this approach in *simi* by introducing a generalization of the subsumption operator. Since such an operator is in general an asymmetric function, we call it *directed simi* and denote it with $simi_d$ (to be introduced in Section 4.1). Now, once the values $simi_d(C, D)$ and $simi_d(D, C)$ are computed, we have to combine them with an operator to obtain a value for *simi*. Instead of fixing a specific operator, we identify the properties such an operator needs to provide such that *simi* fulfills as many of the properties as possible. We call such an operator a *fuzzy connector* (denoted with $\otimes$). A fuzzy connector $\otimes$ is an operator on the interval $[0, 1]$, $\otimes : [0, 1]^2 \longrightarrow [0, 1]$ such that for all $x, y \in [0, 1]$ the following properties are true.

- Commutativity: $x \otimes y = y \otimes x$,
- Equivalence closed: $x \otimes y = 1 \iff x = y = 1$,
- Weak monotonicity: $x \leq y \implies 1 \otimes x \leq 1 \otimes y$,
- Bounded: $x \otimes y = 0 \implies x = 0$ or $y = 0$ and
- Grounded: $0 \otimes 0 = 0$.

Using a fuzzy connector, *simi* is simply defined as

$$simi(C, D) := simi_d(C, D) \otimes simi_d(D, C)$$

where C and D are arbitrary $\mathcal{ELH}$-concepts.

The commutativity of a fuzzy connector ensures that *simi* is symmetric, the property equivalence closed provides the same property for the resulting similarity measure and weak monotonicity is sufficient to prove that *simi* fulfills subsumption preserving. Examples for fuzzy connectors are the average and triangular norms (t-norms, $\otimes$) [14] which fulfill the property that for all $x, y \in [0, 1] : x \otimes y = 0 \implies x = 0$ or $y = 0$ as shown in [15].

4.1 A Directed Similarity Measure: $simi_d$

To formulate $simi_d$, we need a bit of notation. If convenient, we treat concepts as sets of atoms. Let $C \in \mathcal{C}(\mathcal{ELH})$, then it can be written as $C = \bigsqcap_{i \leq n} C_i$ where $\forall i \leq n : \; C_i \in N_A$. The function $\widehat{(\cdot)}$ maps concepts to sets of atoms, so for C, $\widehat{C} := \{C_1, C_2, \ldots, C_n\}$. Now, the starting point for the derivation of $simi_d$ is the function

$$d(C, D) := \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|}$$

which is inspired by the Jaccard Index. This function can be used to measure the similarity of sets of concept names. In order to be able to incorporate existential restrictions, we rewrite the numerator of d to

$$|\widehat{C} \cap \widehat{D}| = \sum_{C' \in \widehat{C}} \max_{D' \in \widehat{D}} f(C', D'), \tag{1}$$

where the function $f : N_C \longrightarrow \{0, 1\}$ is defined as $f(C', D') := 1$ if $C' = D'$ and 0 otherwise.

The simplifying assumption for f is that two different concept names denote always totally dissimilar concepts. However, this assumption may not be correct in all cases. Therefore, we generalize f by introducing a measure for concept names. In order to work for existential restrictions, this measure has to be able to deal with role names, too. In addition, we have to ensure some properties for this measure to guarantee properties for *simi*. We call this measure for (concept or role) names a *primitive measure* and denote it with *pm*. More formally, it is a function of type $pm : N_C^2 \cup N_r^2 \longrightarrow [0, 1]$ with the property that for all $A, B \in N_C$ and $r, s, t \in N_r$ the following holds:

- $pm(A, B) = 1 \iff A = B$,
- $pm(r, s) = 1 \iff s \sqsubseteq r$,
- $s \sqsubseteq_{\mathcal{R}} r \implies pm(s, r) > 0$, and
- $t \sqsubseteq_{\mathcal{R}} s \implies pm(r, s) \leq pm(r, t)$.

The first two properties are sufficient to ensure that *simi* fulfills equivalence closed and the last one is needed to prove that *simi* fulfills subsumption preserving. Note that *pm* does not need to be symmetric.

To incorporate existential restrictions into d we have three cases to consider. Namely, we need to be able to compute the similarity of two concept names, of a concept name and an existential restriction and of two existential restrictions. The first case is handled directly by the primitive measure *pm*. In the second case, we assert that a concept name and an existential restriction are always totally dissimilar and thus their similarity is 0. For the third case, let $\exists r.C^*$ and $\exists s.D^*$ be the two existential restrictions. To compute the similarity of both atoms, we proceed component-wise. The similarity of the role names is computed using the primitive measure *pm* and the similarity of the concepts C^* and D^*

is computed by a recursive call to d. Then, to combine both values we use a number $w \in (0,1)$ and the formula

$$d(\exists r.C^*, \exists s.D^*) := pm(r,s) \cdot [w + (1-w) \cdot d(C^*, D^*)].$$

Forcing $w > 0$, enables us for $d(C^*, D^*) = 0$ to distinguish between the cases where the roles are similar and where they are not. In the first case, the similarity is w, whereas in the second one, the similarity is 0.

As a suitable w, we suggest the value n where one would say that the concepts

$$C := \underbrace{\exists r. \cdots \exists r.}_{n} A \text{ and } D := \underbrace{\exists r. \cdots \exists r.}_{n} B$$

with $pm(A,B) = 0$ are regarded (almost) totally similar.

In Equation 1, we search for each atom of C for that atom of D with the highest similarity value. This method does not always yield satisfactory results. Consider the case, where $pm(A, B_1) = 0.5$ and $pm(A, B_2) = 0.5$ and we want to measure A towards $B_1 \sqcap B_2$, then the current version of function d does not take into account that A is 'known to be similar' to each of B_1 and B_2 alone and thus should even be more similar to their combination. The function chooses *one* 'best matching partner' instead of combining the two sources of similarity.

To deal with this effect, we propose to replace the maximum operator with a triangular conorm (t-conorm, $\oplus$) [14] which is *bounded*, meaning that for all $x, y \in [0,1] : x \oplus y = 1 \implies x = 1$ or $y = 1$. There are several reasons for the use of a t-conorm. First, the operator *max* is an instance of a bounded t-conorm. Second, all t-conorms yield values greater or equal to those of *max* which is consistent with our expectation that the value should be higher or equal to the maximum. Also, 0 acts as neutral element for t-conorms. Therefore, all atoms from D that are totally dissimilar do not influence the value. If we use the probabilistic sum ($x \oplus_{sum} y = x + y - xy$) instead of the maximum for our example above, then we obtain the value 0.75 instead of 0.5, since the measure takes both similarity values (towards B_1 and B_2) into account.

Another parameter of $simi_d$ is the *weighting function* (denoted g). It weights the atoms by assigning each of them a value greater than 0, so $g : N_A \longrightarrow \mathbb{R}_{>0}$. The effect is that some atoms can 'contribute more' to the similarity than others, thus a part of the vocabulary can be picked by g to supply a context under which the concepts from the KB are assessed. Let's assume we are interested in similarity regarding Anatomy and our KB, say SNOMED, contains atoms from two different subject areas like Anatomy and medical procedures. Now, weighting the atoms related to Anatomy higher would result in their similarity having a greater influence on the overall similarity value between concepts.

Note, that the KB does not need to be changed or adapted to achieve this. Several different such weighting functions can easily be employed for the same KB. To incorporate the weighting function we generalize the cardinality of a set of atoms to the sum of the weights of its elements. To obtain a well-defined measure, the weight needs to be added to the numerator of d as well.

By combining the above presented parts, we can already obtain a definition of $simi_d$ except for some corner cases involving $\top$. If we want to be formally correct, then the type of the function $simi_d$ depends on the used parameters as well as on the concepts to be measured. However, for better readability, we omit writing these parameters.

Definition 3 **($simi_d$).** *Let $C, D \in \mathcal{C}(\mathcal{ELH}) \setminus \{\top\}$, $E, F \in \mathcal{C}(\mathcal{ELH})$, $A, B \in N_C$ and $r, s \in N_R$.* Directed simi *is the function $simi_d : \mathcal{C}(\mathcal{ELH})^2 \longrightarrow [0,1]$ defined (w.r.t. a bounded t-conorm $\oplus$, a primitive measure pm, a weighting function g and $w \in (0,1)$) by*

$$\begin{aligned} simi_d(\top,\top) := simi_d(\top, D) \ &:= \ 1, \\ simi_d(C,\top) := 0, & \end{aligned}$$

$$simi_d(C,D) := \frac{\sum\limits_{C' \in \widehat{C}} [g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_a(C', D')]}{\sum\limits_{C' \in \widehat{C}} g(C')},$$

where $simi_a$ measures the similarity of two atoms and is defined as

$$\begin{aligned} simi_a(A,B) &:= pm(A,B), \\ simi_a(\exists r.E, A) := simi_a(A, \exists r.E) \quad &:= \quad 0, \\ simi_a(\exists r.E, \exists s.F) &:= pm(r,s) \cdot [w + (1-w) simi_d(E,F)]. \end{aligned}$$

4.2 Properties of $simi_d$ and $simi$

We present the lemma needed to prove various properties of *simi*. The proofs can be found in [15] (p. 67 ff). In the following we assume that the primitive measure is *pm*, the weighting function is *g*, the t-conorm is $\oplus$ and the fuzzy connector is $\otimes$.

Lemma 1. *Let $C, D, E \in \mathcal{C}(\mathcal{ELH})$. Then*

1. *$simi_d(C,D) = 1 \iff D \sqsubseteq C$.*
2. *$D \sqsubseteq E \implies simi_d(C,E) \leq simi_d(C,D)$.*

Proof. We present a proof sketch for the left-to-right implication of the first statement. Let $simi_d(C,D) = 1$. If $C = \top$ then $D \sqsubseteq C = \top$ is true. Let $C \neq \top$. To prove $D \sqsubseteq C$ we have to show that $\forall C' \in \widehat{C}\ \exists D' \in \widehat{D} :\ D' \sqsubseteq C'$. Let C' be an arbitrary atom of C. $simi_d(C,D) = 1$ implies that

$$\sum_{C' \in \widehat{C}} g(C') = \sum_{C' \in \widehat{C}} [g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_a(C', D')].$$

Because of $g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_a(C', D') \leq g(C')$ we derive that for all $C' \in \widehat{C} :\ \bigoplus_{D' \in D} simi_a(C', D') = 1$. Since the t-conorm is bounded, $\exists D' \in D$ such that $simi_a(C', D') = 1$. The rest of the proof uses structural induction and case distinction.

If $C' = A$ then $simi_a(C', D') = 1$ leads to $D' = A$ which implies $D' \sqsubseteq C'$. Next, let $C' = \exists r.C^*$. $simi_a(C', D') = 1$ implies that D' is of the form $D' = \exists s.D^*$ and $1 = pm(r, s) \cdot [w + (1 - w) simi_d(C^*, D^*)]$. This leads to $pm(r, s) = 1$ which according to the definition of the primitive measure implies $s \sqsubseteq r$. Since $w < 1$, $simi_d(C^*, D^*) = 1$. Using the induction hypothesis we can derive $D^* \sqsubseteq C^*$, therefore $D' \sqsubseteq C'$.

Recall, $simi(C, D) := simi_d(C, D) \otimes simi_d(D, C)$. The resulting function has the following properties.

Theorem 1. *The function simi fulfills*

1. *symmetry,*
2. *equivalence invariance,*
3. *equivalence closed,*
4. *subsumption preserving.*

Let g' be a weighting function with $inf\{g(C') \mid C' \in \mathcal{C}(\mathcal{ELH})\} > 0$. Furthermore, let $\otimes'$ be a fuzzy connector s.t. for all sequences $(x_n)_n$ and $(y_n)_n$ ($x_i, y_i \in [0, 1]$) with $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = 1$ and $\lim_{n\to\infty} x_n \otimes' y_n = 1$. Then simi together with $\otimes'$ and g' fulfills structural dependence.

The main reason why *simi* neither fulfills the triangle inequality nor reverse subsumption preserving is that the computation of $simi_d(C, D)$ does not use the similarity values between the atoms of C (and between the atoms of D). Consider $C := A \sqcap \bigsqcap_{i \le n} B_i$, where the B_i are very similar to each other, $D := A \sqcap B_0$ and $E := A$ then the similarity of D and E is approximately 0.5, the similarity of C and D is close to 1 (since each B_i is very similar to B_0) but the similarity of C and E converges to 0 with increasing n. For the proofs of other properties of *simi* and further details see [15].

An important property of *simi* is that it can be computed efficiently, provided that the involved parameter functions can be computed efficiently as well.

Lemma 2. *If the specific fuzzy connector, the bounded t-conorm, the primitive measure and the weighting function can be computed in polynomial time, then simi can be computed in time polynomial in the size of the concepts to measure.*

5 Conclusions

Similarity measures are important procedures for central ontology management tasks such as alignment of ontologies. Often these measures are built in an ad-hoc way by simply tuning them to test data.

In this paper we have proposed a different approach to construct a whole range of such measures for $\mathcal{ELH}$-concepts. Our starting point was a set of formally defined properties for concept similarity measures, which make use of the semantics of DL concepts and of DL reasoning services. We devised a framework that, if instantiated with appropriate functions and operators as discussed in this paper, allows to generate similarity measures that have 5 of the proposed 7 properties (reverse subsumption preservation and triangle inequality are missing). In that sense one could claim that our framework for similarity measures is not only semantics-based, but also provides the measures with semantics. Moreover, our approach does not restrict users to a single similarity measure, but allows them to design their own measures by selecting the functions and operators appropriate to yield the needed individual similarity measure. If the selected functions conform to the framework described in this paper, the resulting similarity measure is equipped with the properties.

Similarity is often perceived as a context-dependent characteristic. Even in this case our framework can offer support, in the sense that the directed measure $simi_d$ allows atoms appearing in the concept to be weighted differently using the weighting function g. Different instantiations of g allow different thematic subdomains of the domain of discourse to be highlighted.

To test our framework empirically is a non-trivial task, since each application may require a different instantiation of *simi* with functions and operators. To aquire such instantiations suitable for each application requires profound knowledge of the application in question. Thus for now it remains future work to compare the outcome of *simi* instantiations with other well-accepted similarity measures.

On the theoretical side it would be interesting to investigate such frameworks for more expressive DLs and for the concepts defined w.r.t. general TBoxes. Since a unique normal form is the main means to achieve an equivalence invariant similarity measure, it is not obvious how to extend *simi* to these more expressive scenarios.

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[2] F. Baader, R. Küsters, and R. Molitor. Computing least common subsumers in description logics with existential restrictions. In T. Dean, editor, *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI-99)*, pages 96–101, Stockholm, Sweden, 1999. Morgan Kaufmann, Los Altos.

[3] A. Borgida, T. J. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In *Proceedings of the International Workshop on Description Logics (DL2005)*, 2005.

[4] B. Bowdle and D. Gentner. Informativity and asymmetry in comparisons. *Cognitive Psychology*, 34(3):244–286, 1997. PMID: 9466832.

[5] T. G. O. Consortium. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[6] C. d'Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In *Convegno Italiano di Logica Computazionale (CILC 2005)*, 2005.

[7] C. d'Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for $\mathcal{ALC}$ concept descriptions. In *Proceedings of the ACM symposium on Applied computing*, SAC '06, pages 1695–1699, 2006.

[8] C. d'Amato, S. Staab, and N. Fanizzi. On the influence of description logics ontologies on conceptual similarity. In *Proceedings of the 16th Knowledge Engineering Conference (EKAW2008)*, volume 5268, pages 48–63, 2008.

[9] J. Euzenat and P. Valtchev. Similarity-based ontology alignment in OWL-lite. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04)*, pages 333–337. IOS Press, 2004.

[10] N. Fanizzi and C. d'Amato. A similarity measure for the $\mathcal{ALN}$ description logic. In *Convegno Italiano di Logica Computazionale (CILC 2006)*, 2006.

[11] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[12] K. Janowicz. SIM-DL: Towards a semantic similarity measurement theory for the description logic $\mathcal{ALCNR}$ in geographic information retrieval. *SeBGIS 2006, OTM Workshops 2006*, pages 1681–1692, 2006.

[13] K. Janowicz and M. Wilkes. SIM-DLA: a novel semantic similarity measure for description logics reducing Inter-Concept to Inter-Instance similarity. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web Research and Applications*, pages 353–367, 2009.

[14] E. P. Klement, R. Mesiar, and E. Pap. *Triangular Norms.* SV, 2000.

[15] K. Lehmann. A framework for semantic invariant similarity measures for $\mathcal{ELH}$ concept descriptions. Master's thesis, TU Dresden, 2012. Available from: http://lat.inf.tu-dresden.de/research/mas.

[16] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 863—872, 2003.

[17] M. Li and M. R. Sleep. Melody classification using a similarity metric based on Kolmogorov complexity. In *Proceedings of the Sound and Music Computing Conference (SMC'04)*, 2004.

[18] D. Lin. An Information-Theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, 1998.

[19] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 web ontology language profiles. W3C Recommendation, 27 October 2009. http://www.w3.org/TR/2009/REC-owl2-profiles-20091027/.

[20] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity - a review. *QSAR & Combinatorial Science*, 22:1006–1026, 2003.

[21] K. Spackman. Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with snomed-rt. *Journal of the American Medical Informatics Assoc.*, 2000. Fall Symposium Special Issue.

[22] A. Tversky. Features of similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.

[23] W3C OWL Working Group. OWL 2 web ontology language document overview. W3C Recommendation, 27th October 2009. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.