Computing and Informatics, Vol. 36, 2017, 733–764, doi: 10.4149/cai_2017_4_733

ALGORITHMS FOR MEASURING SIMILARITY BETWEEN $\mathcal{ELH}$ CONCEPT DESCRIPTIONS: A CASE STUDY ON SNOMED CT

Suwan Tongphu, Boontawee Suntisrivaraporn

School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology
Thammasat University, Thailand
e-mail: stongphu@gmail.com, sun@siit.tu.ac.th

Abstract. In Description Logics, subsumption is regarded as one of the most prominent reasoning services. It checks, relative to the logical definitions in the ontology, whether one concept is more general/specific than another. When no subsumption relationship is identified, however, no information about the two concepts can be given. In several realistic Semantic Web applications, knowing the level of similarity between two concepts, though lacking the subsumption relationship, is beneficial. This work introduces a new method for measuring the degree of similarity between two concept descriptions in the DL $\mathcal{ELH}$, despite not being in a subsumption relation. Two algorithms are devised based on the known homomorphism-based structural subsumption characterization. The first algorithm employs the top-down approach, whereas the second is carried out in the reverse direction. A bottom-up algorithm has better efficiency, making it more suitable to large-scale ontologies developed using an inexpressive DL in the $\mathcal{EL}$ family, such as the renowned medical ontology Snomed ct. The computational performance of the proposed measure is intensively studied, and interesting findings in Snomed ct are reported.

Keywords: Similarity measure, Snomed ct, semantic web ontology, concept matching

1 INTRODUCTION

Description Logics (DLs) [3] are a family of logic-based knowledge representation formalisms, which can be used to develop ontologies in a formally well-founded way.

This is true both for expressive DLs, which are the logical basis of the Web Ontology Language OWL 2, and for lightweight DLs of the $\mathcal{ELH}$ family [2], which are used in the design of large-scale medical ontologies such as *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED CT) [26] and *Gene Ontology* [20], and form one of the W3C-recommended tractable OWL profiles, OWL 2 EL [21]. One of the main advantages of employing a logic-based ontology language is that reasoning services can be used to derive implicit knowledge from one explicitly represented. DL systems can, for example, classify a given ontology, i.e. compute all the subsumption (i.e. subclass–superclass) relationships between the concepts defined in the ontology and arrange these relationships as a hierarchical graph. The advantage of using a lightweight DL in the $\mathcal{ELH}$ family is that classification is tractable, i.e. a subsumption hierarchy of a given ontology can be computed in polynomial time. Though inevitably useful in ontology design, the reasoning service of subsumption merely gives a crisp response, i.e. a positive response concluding that one concept is subsumed by the other or a negative response otherwise. In virtually every domain, certain concepts may share commonality and as such can be perceived as similar to one another, despite the fact that they are out of the subsumption relation.

Moreover, in several realistic Semantic Web applications, knowing the level of similarity between two concepts, though lacking a subsumption relationship, is beneficial. Examples include a health decision support system retrieving similar treatment cases in the past as guidelines to treat the current patient, an object detection scenario [31] that tries to identify an object of interest from its parts, and a natural language processing tool in which technical terms are extracted from a full text. In these applications, some information or keywords may be missing, but the existing information still forms a level of relevance relative to a concept in question.

There has been a good number of works on similarity measures. The majority, however, are either ineffective or lack desirable properties for similarity measure [15, 14, 8, 11, 7, 9]. For example, in [13], the author presents a simple similarity measure. Despite fulfilling all the desirable properties (see more detail in Section 6), it supports a language with concept conjunction while ontologies in practice are modeled using at least $\mathcal{ELH}$. An extended work of [13] by Lehman and Turhan [19] proposes a generic framework for the DL $\mathcal{ELH}$ that satisfies most of the properties for similarity. However, as it is a very generic framework, the functions and operators needed for the computation are left in question and rather described by means of promising properties. In other words, the framework does not specify implementation details and, to the best of our knowledge, has not been implemented yet.

Our approach to similarity measure is driven by the structural subsumption characterization by means of tree homomorphism. This is first proposed in [29] for $\mathcal{EL}$, and its desirable properties are investigated for the extended DL $\mathcal{ELH}$ in [30]. A preliminary study on this usability is reported in our proceedings paper [16]. Though similar to the measure in [19], our definition is more practical and suited for implementation since homomorphism on trees can be computed in a bottom-up fashion which is in practice three times faster than the counterpart recursive-based (top-down) method. With our extensive experiments of the implemented algorithm on Snomed ct, not only can we ensure its practical computability, but we also learn important characteristics in the design of this and other medical ontologies.

The rest of the paper is organized in the following order. The background on the DL $\mathcal{ELH}$, unfoldable TBoxes, and the structural subsumption algorithm is presented in the next section. Sections 3, 4, and 5 introduce the notions of homomorphism degree and the top-down algorithm, the bottom-up reasoning algorithm, and $\mathcal{ELH}$ semantic similarity measure, respectively. Section 6 lists major similarity properties and provides mathematical proofs. Experiments and their results are explained in Section 7. The usability evaluation of the proposed reasoning system is further explained in the subsequent section. The last section gives some concluding remarks.

2 BACKGROUND

In DLs, *concept descriptions* are inductively defined with the help of a set of *constructors*, starting with a set CN of *concept names* and a set RN of *role names*. $\mathcal{ELH}$ concept descriptions are formed using the constructors shown in the upper part of Table 1. Conventionally, r, s possibly with subscripts are used to range over RN, A, B to range over CN, and C, D to range over concept descriptions.

Name	Syntax	Semantics
Top	$\top$	$\Delta^{\mathcal{I}}$
Concept name	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
Conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
Existential restriction	$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}} : (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
Primitive concept def	$B \sqsubseteq D$	$B^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
Full concept def	$B \equiv D$	$B^{\mathcal{I}} = D^{\mathcal{I}}$
Primitive role def	$r \sqsubseteq s$	$r^{\mathcal{I}} \subseteq s^{\mathcal{I}}$

Table 1. Syntax and semantics of the Description Logic $\mathcal{ELH}$

Concept names appearing on the left-hand side of a definition are called *primitively or fully defined concept names* (in symbol $\mathsf{CN}^{\mathsf{def}}$). Other concept names are called *primitive concept names* (denoted by $\mathsf{CN}^{\mathsf{pri}}$). Therefore, $\mathsf{CN} = \mathsf{CN}^{\mathsf{pri}} \cup \mathsf{CN}^{\mathsf{def}}$. An *$\mathcal{ELH}$ terminology box* or *TBox* is a finite set of concept definitions, whose syntax is shown in the lower part of Table 1. A TBox is called *unfoldable* if it is definitorial (i.e. $\mathsf{CN}^{\mathsf{pri}}$ uniquely define $\mathsf{CN}^{\mathsf{def}}$) and acyclic (i.e., it does not refer to itself neither directly nor indirectly).

Figure 1 depicts an example $\mathcal{ELH}$ unfoldable terminology, hitherto referred to as $\mathcal{O}_{\mathsf{med}}$, that is extracted from Snomed ct. For convenience of further references, an axiom ID is assigned to each concept definition in $\mathcal{O}_{\mathsf{med}}$. Primitively defined concepts and primitive concept names are commonly found in realistic terminologies. By introducing a set of fresh concept names, such concept names can easily be transformed into semantically equivalent fully defined ones using the following steps:

$\mathsf{NeonatalAspirationOfAmnioticFluid}$	$\equiv$	$\mathsf{NeonatalAspirationSyndromes}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{causativeAgent}.\mathsf{AmnioticFluid})$
$\mathsf{NeonatalAspirationOfMucus}$	$\equiv$	$\mathsf{NeonatalAspirationSyndromes}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{causativeAgent}.\mathsf{Mucus})$
$\mathsf{NeonatalAspirationOfMilk}$	$\equiv$	$\mathsf{NeonatalAspirationSyndromes} \sqcap \mathsf{InhalationOfLiquid}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{causativeAgent}.\mathsf{Milk})$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{associatedWith}.\mathsf{Milk})$
$\mathsf{NeonatalAspirationOfMeconium}$	$\equiv$	$\mathsf{NeonatalAspirationSyndromes}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{causativeAgent}.\mathsf{MeconiumStool})$
$\mathsf{Hypoxia}$	$\equiv$	$\mathsf{DisorderOfRespiratorySystem} \sqcap \mathsf{DisorderOfBloodGas}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{interprets}.\mathsf{OxygenDelivery})$
$\mathsf{Hypoxemia}$	$\equiv$	$\mathsf{DisorderOfRespiratorySystem} \sqcap \mathsf{DisorderOfBloodGas}$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{interprets}.\mathsf{OxygenDelivery})$ $\sqcap\exists\mathsf{roleGroup}.(\exists\mathsf{findingSite}.\mathsf{ArterialSystemStructure})$
$\mathsf{BodySecretion}$	$\sqsubseteq$	$\mathsf{BodySubstance}$
$\mathsf{BodySubstance}$	$\sqsubseteq$	$\mathsf{Substance}$
Milk	$\sqsubseteq$	$\mathsf{DairyFoods}$
$\mathsf{DairyFoods}$	$\sqsubseteq$	$\mathsf{FoodAllergen} \sqcap \mathsf{Foods}$
$\mathsf{BodyFluid}$	$\sqsubseteq$	$\mathsf{BodySubstance} \sqcap \mathsf{LiquidSubstance}$
$\mathsf{FoodAllergen}$	$\sqsubseteq$	$\mathsf{AllergenClass}$
$\mathsf{AllergenClass}$	$\sqsubseteq$	$\mathsf{Substance}$
$\mathsf{AmnioticFluid}$	$\sqsubseteq$	$\mathsf{BodyFluid}$
Mucus	$\sqsubseteq$	$\mathsf{BodySecretion}$
$\mathsf{MeconiumStool}$	$\sqsubseteq$	$\mathsf{DigestiveSystemFluid}$
$\mathsf{causativeAgent}$	$\sqsubseteq$	$\mathsf{associatedWith}$

Figure 1. Examples of $\mathcal{ELH}$ concept descriptions defined in $\mathcal{O}_{\mathsf{med}}$

1. For each $A \in \mathsf{CN}^{\mathsf{pri}}$, add $A \equiv F$ to $\mathcal{O}$ with F a fresh concept name.
2. For each $B \in \mathsf{CN}^{\mathsf{def}}$ with $B \sqsubseteq D \in \mathcal{O}$, $B \sqsubseteq D$ is replaced by $B \equiv G \sqcap D$ with G, a fresh concept name.

For example, we can transform the concept $\mathsf{AllergenClass}$ as $\mathsf{AllergenClass} \equiv F_{13} \sqcap \mathsf{Substance} \equiv G_{13} \sqcap F_8$ where G_{13} and F_8 are fresh concept names. Note that every primitive concept A is subsumed by $\top$ (i.e. $A \sqsubseteq \top$). Hence, without loss of generality, A can be likewise replaced by F.

Like any DLs, the semantics of $\mathcal{ELH}$ are defined in terms of *interpretations* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where the domain $\Delta^{\mathcal{I}}$ is a non-empty set of individuals, and the interpretation function $\cdot^{\mathcal{I}}$ maps each concept name $A \in \mathsf{CN}$ to a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and each role name $r \in \mathsf{RN}$ to a binary relation $r^{\mathcal{I}}$ on $\Delta^{\mathcal{I}}$. The extension of $\cdot^{\mathcal{I}}$ to arbitrary concept descriptions is inductively defined, as shown in the semantics column of Table 1. An interpretation $\mathcal{I}$ is a *model* of a TBox $\mathcal{O}$ if, for each concept definition and role hierarchy axiom in $\mathcal{O}$, the conditions given in the semantics column of Table 1 are satisfied. The main inference problem in DL is a concept subsumption.

Definition 1 (concept subsumption)**.** Given two $\mathcal{ELH}$ concept descriptions C, D and an $\mathcal{ELH}$ TBox $\mathcal{O}$, C is subsumed by D w.r.t. $\mathcal{O}$ (written $C \sqsubseteq_{\mathcal{O}} D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$

ω_1	NeonatalAspirationOfAmnioticFluid	ω_2	NeonatalAspirationOfMucus
ω_3	NeonatalAspirationSyndromes	ω_4	NeonatalAspirationOfMilk
ω_5	NeonatalAspirationOfMeconium	ω_6	BodySecretion
ω_7	BodySubstance	ω_8	Substance
ω_9	InhalationOfLiquid	ω_{10}	Milk
ω_{11}	DairyFoods	ω_{12}	BodyFluid
ω_{13}	FoodAllergen	ω_{14}	AllergenClass
ω_{15}	AmnioticFluid	ω_{16}	LiquidSubstance
ω_{17}	Mucus	ω_{18}	MeconiumStool
ω_{19}	DigestiveSystemFluid	ω_{20}	Foods
ω_{21}	Hypoxia	ω_{22}	DisorderOfRespiratorySystem
ω_{23}	DisorderOfBloodGas	ω_{24}	OxygenDelivery
ω_{25}	ArterialSystemStructure	β_1	associatedWith
β_2	causativeAgent	β_3	interprets

Figure 2. List of the concept and role names in $\mathcal{O}_{\mathsf{med}}$

in every model $\mathcal{I}$ of $\mathcal{O}$. Moreover, C, D are equivalent w.r.t. $\mathcal{O}$ (written $C \equiv_{\mathcal{O}} D$) if $C \sqsubseteq_{\mathcal{O}} D$ and $D \sqsubseteq_{\mathcal{O}} C$.

Providing that the TBox is unfoldable (i.e. acyclic and definitorial), any $\mathcal{ELH}$ concept can be expanded to an equivalent one $\hat{C}$, consisting only of fresh concept names. Hence, unless stated otherwise, we assume without loss of generality that an $\mathcal{ELH}$ concept C is expanded and has the following form:

$$P_1 \sqcap \cdots \sqcap P_m \sqcap \exists r_1.C_1 \sqcap \cdots \sqcap \exists r_n.C_n \qquad (1)$$

where $P_i \in \mathsf{CN}^{\mathsf{pri}}$, $r_j \in \mathsf{RN}$, and C_j are defined in a similar manner, for $1 \leq i \leq m$ and $1 \leq j \leq n$. A terminology $\mathcal{O}$ can be dispensed with $C \sqsubseteq_{\mathcal{O}} D$ iff $\hat{C} \sqsubseteq \hat{D}$. For convenience, we denote by $\mathcal{P}_C$ and $\mathcal{E}_C$ the set of top-level concept names $\{P_1, \ldots, P_m\}$ and the set of top-level existential restrictions $\{\exists r_1.C_1, \ldots, \exists r_n.C_n\}$, respectively. Also, we denote by $\mathcal{R}_r = \{s | r \sqsubseteq^* s\}$ the set of all r's super roles where r and s are role names and $r \sqsubseteq^* s$ if $r = s$ or $r_i \sqsubseteq r_{i+1} \in \mathcal{O}$ where $1 \leq i \leq n$, $r_1 = r$, $r_n = s$, and $*$ is a transitive closure. In [10, 1], a characterization of subsumption in $\mathcal{ELH}$ w.r.t. an unfoldable TBox using homomorphism has been proposed. Instead of directly considering concept descriptions, the characterization considers so-called *$\mathcal{ELH}$ description trees* that structurally correspond to $\mathcal{ELH}$ concept descriptions and can be constructed using Algorithm 1. In essence, the root v of the $\mathcal{ELH}$ concept description tree $\mathcal{T}$ for the concept description in Formula (1) has $\{P_1, \ldots, P_m\}$ as its label, and has n outgoing edges, each labeled with $\mathcal{R}_{r_j}$ to a vertex v_j, for $1 \leq j \leq n$. Then, the subtree $\mathcal{T}|_{v_j}$ with the root v_j is defined inductively based on the nested concept C_j. The subsumption is then characterized by means of an existence of a homomorphism in the reverse direction.

Algorithm 1 $\mathcal{ELH}$ description tree construction

```
Input: $\mathcal{P}_C$ and $\mathcal{E}_C$
Output: The description tree $\mathcal{T}$
function build-tree($\mathcal{P}_C, \mathcal{E}_C$)
  Create a new tree $\mathcal{T}$
  Create a new vertex $v \in V$
  $\ell(v) \leftarrow \mathcal{P}_C$
  for each $\exists r.C' \in \mathcal{E}_C$ do
      build-child-node($v, r, \mathcal{P}_{C'}, \mathcal{E}_{C'}$)
  return $\mathcal{T}$
function build-child-node($v, r, \mathcal{P}_C, \mathcal{E}_C$)
  Create a new vertex $w \in V$
  $\ell(w) \leftarrow \mathcal{P}_C$
  Add a new edge $(v, w)$ to $E$
  $\rho(v, w) \leftarrow \{r\}$
  for each $\exists s.C' \in \mathcal{E}_C$ do
      build-child-node($w, s, \mathcal{P}_{C'}, \mathcal{E}_{C'}$)
```

Definition 2 (Homomorphism). Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted $\mathcal{ELH}$ concept description trees. For any $\mathcal{T} = (V, E, rt, \ell, \rho)$ where V is a set of nodes, $E \subseteq V \times V$ is a set of edges, rt is the root, $\ell : V \rightarrow 2^{\mathsf{CN}^{\mathsf{pri}}}$ is a node labeling function, and $\rho : E \rightarrow 2^{\mathsf{RN}}$ is an edge labeling function. Given $v \in V$ and $v' \in V'$, there exists a homomorphism h from $\mathcal{T}$ to $\mathcal{T}'$, written $h : \mathcal{T} \rightarrow \mathcal{T}'$ iff the following conditions are satisfied:

1. $h(rt) = rt'$ and $\ell(v) \subseteq \ell'(h(v))$
2. For each successor w of v in $\mathcal{T}$, $h(w)$ is a successor of $h(v)$ with $\rho(v, w) \subseteq \rho'(h(v), h(w))$

Theorem 3 ([18]). Let C, D be expanded $\mathcal{ELH}$ concept descriptions and $\mathcal{T}_C$ and $\mathcal{T}_D$ are $\mathcal{ELH}$ description trees w.r.t. C and D. Then, $C \sqsubseteq D$ iff there exists a homomorphism $h : \mathcal{T}_D \rightarrow \mathcal{T}_C$ that maps the root of $\mathcal{T}_D$ to the root of $\mathcal{T}_C$.

Corollary 4. Let C and D be concept names. Then $C \sqsubseteq D$ iff $\mathcal{P}_D \subseteq \mathcal{P}_C$ and for each $\exists r.D' \in \mathcal{E}_D$ there exists $\exists s.C' \in \mathcal{E}_C$ such that $s \sqsubseteq^* r$ and $C' \sqsubseteq D'$.

Corollary 5. Let C and D be concept names, then $\mathcal{E}_D \cong \mathcal{E}_C$ iff for each $\exists r.D' \in \mathcal{E}_D$ there exists $\exists s.C'$ such that $s \sqsubseteq^* r$, $r \sqsubseteq^* s$, $C' \sqsubseteq D'$, and $D' \sqsubseteq C'$. Moreover, $C \equiv D$ iff $\mathcal{P}_D = \mathcal{P}_C$ (i.e. $\mathcal{P}_D \subseteq \mathcal{P}_C$ and $\mathcal{P}_C \subseteq \mathcal{P}_D$) and $\mathcal{E}_D \cong \mathcal{E}_C$.

Theorem 3 describes that if C is subsumed by D then there is a homomorphism mapping in the reverse direction. Corollary 4 suggests that if C is subsumed by D then, for each top-level concept name in D, it must also appear in C. Besides, each existential restriction $(\exists r.D')$ in D has a counterpart existential restriction $(\exists s.C')$ in C such that $s \sqsubseteq^* r$ and $C' \sqsubseteq D'$. Corollary 5 suggests further that $\mathcal{E}_D$ is congruent to $\mathcal{E}_C$ (in symbol $\mathcal{E}_D \cong \mathcal{E}_C$) then, each $\exists r.D'$ has a counterpart $\exists s.C'$ such that $s \sqsubseteq^* r$, $r \sqsubseteq^* s$, $C' \sqsubseteq D'$, and $D' \sqsubseteq C'$. By using Theorem 3 together with properties of homomorphism mapping defined in Definition 2, Corollary 4 and Corollary 5 hold due to an associativity and commutativity of concept conjunction.

Consider the concept description $\mathsf{NeonatalAspirationSyndromes}$ (Nas), $\mathsf{NeonatalAspirationOfMilk}$ (Nam), and $\mathsf{NeonatalAspirationOfAmnioticFluid}$ (Naaf) defined in $\mathcal{O}_{\mathsf{med}}$. Let $\widehat{\mathsf{Nas}}$, $\widehat{\mathsf{Nam}}$ and $\widehat{\mathsf{Naaf}}$ be the expanded concepts w.r.t. Nas, Nam, and Naaf, respectively. By Theorem 3, this ensures that $\widehat{\mathsf{Naaf}} \sqsubseteq \widehat{\mathsf{Nas}}$, and that $\widehat{\mathsf{Naaf}} \not\sqsubseteq \widehat{\mathsf{Nam}}$. Though sharing some common features, the classical reasoning of subsumption does not suffice to tell how similar they are.

Our similarity measure is based on the structural characterization. Given two concept descriptions C and D, instead of merely giving either a positive or negative result, the similarity measure computes a numerical value, suggesting their degree of similarity such that $0 \leq \mathsf{sim}(C, D) = \mathsf{sim}(D, C) \leq 1$. Intuitively, the larger the number, the more similar the two concepts are. In particular, if the similarity degree is 1, then the two concepts are logically equivalent.

3 HOMOMORPHISM DEGREE

Theorem 3 suggests that an existence of a homomorphism between $\mathcal{ELH}$ description trees implies the subsumption relationship between the corresponding concept descriptions. We extend this idea to the case where *no* such homomorphism exists, but there is some commonality.

Let C, D be unfolded $\mathcal{ELH}$ concept descriptions, $\mathcal{P}_C$, $\mathcal{P}_D$, $\mathcal{E}_C$, $\mathcal{E}_D$ be as defined in the previous section, $\mathcal{T}_C, \mathcal{T}_D$ be the corresponding $\mathcal{ELH}$ description trees, $\mathcal{R}_r$, $\mathcal{R}_s$ be sets of super roles w.r.t. r, s, respectively. Then, the degree of having a homomorphism from $\mathcal{T}_D$ to $\mathcal{T}_C$ is defined by Definition 6.

Definition 6 (Homomorphism degree). Let $\mathbf{T}^{\mathcal{ELH}}$ be the set of all $\mathcal{ELH}$ description trees. The *homomorphism degree function* $\mathsf{hd} : \mathbf{T}^{\mathcal{ELH}} \times \mathbf{T}^{\mathcal{ELH}} \to [0, 1]$ is inductively defined as follows:

$$\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) := \mu \cdot \mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) + (1 - \mu) \cdot \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C) \tag{2}$$

where $0 \leq \mu \leq 1$;

$$\mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) := \begin{cases} 1 & \text{if } \mathcal{P}_D = \emptyset, \\ \frac{|\mathcal{P}_D \cap \mathcal{P}_C|}{|\mathcal{P}_D|} & \text{otherwise} \end{cases} \tag{3}$$

where $|\cdot|$ represents the set cardinality;

$$\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C) := \begin{cases} 1 & \text{if } \mathcal{E}_D = \emptyset, \\ 0 & \text{if } \mathcal{E}_D \neq \emptyset \text{ and } \mathcal{E}_C = \emptyset, \\ \sum_{\epsilon_i \in \mathcal{E}_D} \frac{max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_C\}}{|\mathcal{E}_D|} & \text{otherwise} \end{cases} \quad (4)$$

where ϵ_i, ϵ_j are existential restrictions; and

$$\mathsf{e\text{-}hd}(\exists r.X, \exists s.Y) := \gamma(\nu + (1-\nu) \cdot \mathsf{hd}(\mathcal{T}_X, \mathcal{T}_Y)) \quad (5)$$

where $\gamma = \frac{|\mathcal{R}_r \cap \mathcal{R}_s|}{|\mathcal{R}_r|}$ and $0 \leq \nu < 1$.

Intuitively, the homomorphism degree Formula (2) is defined as the weighted sum of the degree of the label set inclusion (p-hd) and the degree of the edge condition matching (e-set-hd). Formula (3) calculates the proportion of the matched primitive concepts to all the primitive concepts in the top level conjunction. Formula (5) measures the degree of an edge mapping in a potential homomorphism. If the edge-labeling roles are the same or share some superroles, then part of the edge matching conditions is satisfied, but the successors' labels and structures have yet to be checked. This is defined recursively as $\mathsf{hd}(\mathcal{T}_X, \mathcal{T}_Y)$. The values computed in Formula (5) are collectively used to determine the degree of the edge matching condition. Formula (4) calculates the maximum degree for each edge in $\mathcal{E}_D$ representing the best possible mapping and returns the average.

The weight μ in Formula (2) indicates how important the primitive concept names are to be considered for similarity. It is recommended that $\mu = \frac{|\mathcal{P}_D|}{|\mathcal{P}_D \cup \mathcal{E}_D|}$, i.e. the ratio of the primitive concept names to that of all the terms in the top level conjunction. For the special case where $D = \top$ (i.e. $\mathcal{P}_D = \mathcal{E}_D = \emptyset$), the value of μ is irrelevant as $\mathcal{T}_\top$ is the smallest $\mathcal{ELH}$ description tree with only one node and $\mathsf{hd}(\mathcal{T}_\top, \mathcal{T}_C) = 1$ for all concepts C. The weight ν determines how important the roles are to be considered for similarity between two existential restrictions.

The value of γ in Formula (5) is the ratio of a number of common superroles to a number of all the supperroles of r. For the case where $\gamma = 0$, this means there is no commonality between r and s, i.e., further computations for the degree of similarity among their successors is irrelevant. Thus, the two existential restrictions are regarded as dissimilar. If $0 < \gamma \leq 1$, this reveals that there exists some commonality. However, the case where $\gamma = 1$ implies that r and s are the same role name.

Let $\hat{\mathcal{O}}_{\mathsf{med}}$ be an unfolded version of the example TBox from Figure 1. The following demonstrates the computation of the homomorphism degree from NeonatalAspirationOfAmnioticFluid (Naaf) to NeonatalAspirationOfMilk (Nam) and vice versa using the top-down approach (see e.g. their definitions and relevant name IDs in Figure 1 and Figure 2, respectively).

Algorithm 2 $\mathcal{ELH}$ similarity measure

Input: Two $\mathcal{ELH}$ description trees $\mathcal{T}_D$ and $\mathcal{T}_C$
Output: The homomorphism degree from $\mathcal{T}_D$ to $\mathcal{T}_C$

```
function hd($\mathcal{T}_D, \mathcal{T}_C$)

return $\mu \cdot \mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) + (1-\mu)\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_D, \mathcal{E}_C)$

function p-hd($\mathcal{P}_D, \mathcal{P}_C$)

if $\mathcal{P}_D \leftarrow \emptyset$ then
    return 1
else
    return $\frac{|\mathcal{P}_D \cap \mathcal{P}_C|}{|\mathcal{P}_D|}$

function e-set-hd($\mathcal{E}_D, \mathcal{E}_C$)

$sum \leftarrow 0$
for each $e_i \in \mathcal{E}_D$ do
    $max \leftarrow 0$
    for each $e_j \in \mathcal{E}_C$ do
        if $\mathsf{e\text{-}hd}(e_i, e_j) > max$ then
            $max \leftarrow \mathsf{e\text{-}hd}(e_i, e_j)$
    $sum \leftarrow sum + max$
return $\frac{sum}{|\mathcal{E}_D|}$

function e-hd($\exists r.X, \exists s.Y$)

$\gamma \leftarrow \frac{|\mathcal{R}_r \cap \mathcal{R}_s|}{|\mathcal{R}_r|}$
if $\gamma = 0$ then
    return 0
else
    return $\gamma(\nu + (1-\nu) \cdot \mathsf{hd}(\mathcal{T}_X, \mathcal{T}_Y))$
```

Example Consider the expansion of the concept $\mathsf{N\hat{a}af}$ and Nam defined in $\hat{\mathcal{O}}$:

$$\begin{aligned}
\mathsf{N\hat{a}af} &\equiv F_3 \sqcap \exists \mathsf{rG}.(\exists \mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16})),\\
\mathsf{N\hat{a}m} &\equiv F_3 \sqcap F_9 \sqcap \exists \mathsf{rG}.(\exists \mathsf{cA}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}))\\
&\quad \sqcap \exists \mathsf{rG}.(\exists \mathsf{aW}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20})).
\end{aligned}$$

Using $\nu = 0.4$, the homomorphism degree from $\mathsf{N\hat{a}af}$ to $\mathsf{N\hat{a}m}$ can be computed using Algorithm 2. The following shows the computation step by step[1].

[1] Obvious abbreviations are used here for the sake of succinctness.

$$\begin{aligned}
\mathsf{hd}(\mathcal{T}_{\mathsf{N\hat{a}af}},\mathcal{T}_{\mathsf{N\hat{a}m}}) &:= \frac{1}{2}\mathsf{p\text{-}hd}(\mathcal{P}_{\mathsf{N\hat{a}af}},\mathcal{P}_{\mathsf{N\hat{a}m}}) + \frac{1}{2}\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}af}},\mathcal{E}_{\mathsf{N\hat{a}m}})\\
&:= \frac{1}{2}[1] + \frac{1}{2}\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}af}},\mathcal{E}_{\mathsf{N\hat{a}m}})\\
&:= \frac{1}{2}[1] + \frac{1}{2}\left[\frac{89}{125}\right] \quad := 0.856.
\end{aligned}$$

The computation for the sub-descriptions, corresponding with $\epsilon_i = \exists\mathsf{rG}.(\exists\mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16}))$ and $\epsilon_j^1 = \exists\mathsf{rG}.(\exists\mathsf{cA}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}))$, is as follows:

$$\begin{aligned}
\mathsf{e\text{-}hd}(\epsilon_i,\epsilon_j^1) &:= [1]\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{\exists\mathsf{cA}.(G_{14}\sqcap G_{11}\sqcap G_8\sqcap F_8\sqcap F_{16})},\mathcal{T}_{\exists\mathsf{cA}.(G_9\sqcap G_{10}\sqcap G_{12}\sqcap G_{13}\sqcap F_8\sqcap F_{20})})\right]\\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[[1]\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{G_{14}\sqcap G_{11}\sqcap G_8\sqcap F_8\sqcap F_{16}},\mathcal{T}_{G_9\sqcap G_{10}\sqcap G_{12}\sqcap G_{13}\sqcap F_8\sqcap F_{20}})\right]\right]\right]\\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[[1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{5}\right]\right]\right]\right]\\
&:= \frac{89}{125}.
\end{aligned}$$

Another possibility is to map ϵ_i to $\epsilon_j^2 = \exists\mathsf{rG}.(\exists\mathsf{aW}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}))$ where $\gamma = \frac{|\mathcal{R}_{\mathsf{rG}} \cap \mathcal{R}_{\mathsf{rG}}|}{|\mathcal{R}_{\mathsf{rG}}|} = 1$. The computation in such a case is as follows:

$$\begin{aligned}
\mathsf{e\text{-}hd}(\epsilon_i,\epsilon_j^2) &:= [1]\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{\exists\mathsf{cA}.(G_{14}\sqcap G_{11}\sqcap G_8\sqcap F_8\sqcap F_{16})},\mathcal{T}_{\exists\mathsf{aW}.(G_9\sqcap G_{10}\sqcap G_{12}\sqcap G_{13}\sqcap F_8\sqcap F_{20})})\right]\\
&\qquad // \text{ where } \gamma = \frac{|\ \mathcal{R}_{\mathsf{cA}} \cap \mathcal{R}_{\mathsf{aW}}\ |}{|\ \mathcal{R}_{\mathsf{cA}}\ |} = \frac{|\ \{\mathsf{aW}\}\ |}{|\ \{\mathsf{cA},\mathsf{aW}\}\ |} = \frac{1}{2}\\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{2}\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{G_{14}\sqcap G_{11}\sqcap G_8\sqcap F_8\sqcap F_{16}},\mathcal{T}_{G_9\sqcap G_{10}\sqcap G_{12}\sqcap G_{13}\sqcap F_8\sqcap F_{20}})\right]\right]\right]\\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{2}\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{5}\right]\right]\right]\right]\\
&:= \frac{139}{250}.
\end{aligned}$$

Since $\frac{89}{125} > \frac{139}{250}$ and ϵ_i is the only existential restriction in $\mathcal{E}_{\mathsf{N\hat{a}af}}$, $\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}af}},\mathcal{E}_{\mathsf{N\hat{a}m}})$ returns $\frac{89}{125}$. For the reverse direction, it can be computed by:

$$
\begin{aligned}
\mathsf{hd}(\mathcal{T}_{\mathsf{N\hat{a}m}}, \mathcal{T}_{\mathsf{N\hat{a}af}}) &:= \frac{2}{4}\mathsf{p\text{-}hd}(\mathcal{P}_{\mathsf{N\hat{a}m}}, \mathcal{P}_{\mathsf{N\hat{a}af}}) + \frac{2}{4}\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}m}}, \mathcal{E}_{\mathsf{N\hat{a}af}}) \\
&:= \frac{2}{4}\left[\frac{1}{2}\right] + \frac{2}{4}\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}m}}, \mathcal{E}_{\mathsf{N\hat{a}af}}) \\
&:= \frac{2}{4}\left[\frac{1}{2}\right] + \frac{2}{4}\left[\frac{7}{10}\right] \\
&:= 0.6.
\end{aligned}
$$

The computation for the sub-descriptions, corresponding with $\epsilon_i^1 = \exists \mathsf{rG}.(\exists \mathsf{cA}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}))$, $\epsilon_j = \exists \mathsf{rG}.(\exists \mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16}))$, and $\gamma = \frac{|\mathcal{R}_{\mathsf{rG}} \cap \mathcal{R}_{\mathsf{rG}}|}{|\mathcal{R}_{\mathsf{rG}}|} = 1$ is as follows:

$$
\begin{aligned}
\mathsf{e\text{-}hd}(\epsilon_i^1, \epsilon_j) &:= [1]\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{\exists \mathsf{cA}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20})}, \mathcal{T}_{\exists \mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16})})\right] \\
&\qquad // \text{ where } \gamma = \frac{\mid \mathcal{R}_{\mathsf{cA}} \cap \mathcal{R}_{\mathsf{cA}} \mid}{\mid \mathcal{R}_{\mathsf{cA}} \mid} = \frac{\mid \{\mathsf{cA}, \mathsf{aW}\} \mid}{\mid \{\mathsf{cA}, \mathsf{aW}\} \mid} = 1 \\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{2}{2}\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}}, \mathcal{T}_{G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16}})\right]\right]\right] \\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{2}{2}\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{6}\right]\right]\right]\right] \\
&:= \frac{7}{10}.
\end{aligned}
$$

The other edge matching w.r.t. $\epsilon_i^2 = \exists \mathsf{rG}.(\exists \mathsf{aW}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}))$ and $\epsilon_j = \exists \mathsf{rG}.(\exists \mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16}))$, can be computed by:

$$
\begin{aligned}
\mathsf{e\text{-}hd}(\epsilon_i^2, \epsilon_j) &:= [1]\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{\exists \mathsf{aW}.(G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20})}, \mathcal{T}_{\exists \mathsf{cA}.(G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16})})\right] \\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{1}\left[\frac{2}{5} + \frac{3}{5}\mathsf{hd}(\mathcal{T}_{G_9 \sqcap G_{10} \sqcap G_{12} \sqcap G_{13} \sqcap F_8 \sqcap F_{20}}, \mathcal{T}_{G_{14} \sqcap G_{11} \sqcap G_8 \sqcap F_8 \sqcap F_{16}})\right]\right]\right] \\
&:= [1]\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{1}\left[\frac{2}{5} + \frac{3}{5}\left[\frac{1}{6}\right]\right]\right]\right] \\
&:= \frac{7}{10}.
\end{aligned}
$$

The $\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{\mathsf{N\hat{a}m}}, \mathcal{E}_{\mathsf{N\hat{a}af}}) := \frac{\frac{7}{10} + \frac{7}{10}}{2} := \frac{7}{10}$ is thus the average of the maximum.

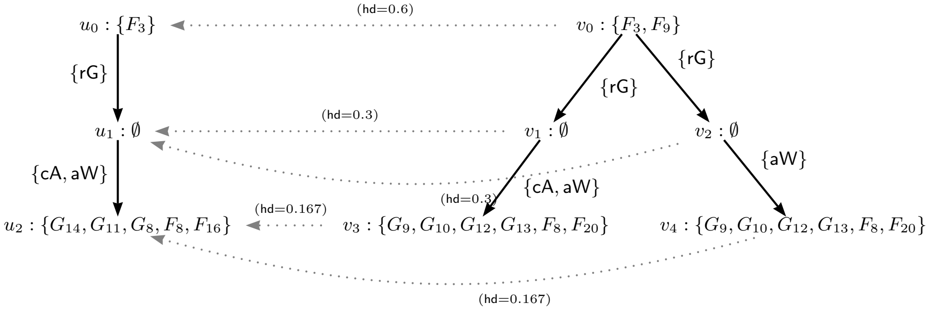

Figure 3. A homomorphism degree that maps the root of $\mathcal{T}_{\mathsf{N\hat{a}m}}$ (v_0) to the root of $\mathcal{T}_{\mathsf{N\hat{a}af}}$ (u_0) (see dotted arrows)

Hence, the degree of having a homomorphism from $\mathsf{N\hat{a}af}$ to $\mathsf{N\hat{a}m}$ is 0.856, and that from the opposite direction is 0.6. The hd values for other pairs of concepts in $\mathcal{O}_{\mathsf{med}}$ can be obtained in an analogous manner and are shown in Table 2. Figure 3 shows the example of the homomorphism degree mapping the root of $\mathcal{T}_{\mathsf{N\hat{a}m}}$ to that of $\mathcal{T}_{\mathsf{N\hat{a}af}}$.

	Neonatal Aspiration of					
hd(↓, →)	Milk	AF	Mucus	Mec	HPX	HPM
Milk	1.00	0.60	0.60	0.57	0.20	0.20
AmnioticFluid	0.85	1.00	0.89	0.82	0.20	0.20
Mucus	0.86	0.91	1.00	0.82	0.20	0.20
Meconium	0.82	0.82	0.82	1.00	0.20	0.20
Hypoxia	0.13	0.13	0.13	0.13	1.00	1.00
Hypoxemia	0.20	0.20	0.20	0.20	0.85	1.00

Table 2. Homomorphism degree among the defined concepts in $\mathcal{O}_{\mathsf{med}}$

Using a proof by induction, together with Theorem 3 [10, 1], it is not difficult to obtain the correspondence between the homomorphism degree and subsumption. Proposition 7 describes the property of a concept subsumption. We say that C is a subconcept of D if the homomorphism degree of the corresponding description tree $\mathcal{T}_D$ to $\mathcal{T}_C$ is equal to 1 and vice versa (see the proof in Appendix).

Proposition 7. Let C, D be expanded $\mathcal{ELH}$ concept descriptions, and $\mathcal{T}_C$, $\mathcal{T}_D$ be their corresponding description tree, respectively. Then, the following are equivalent:

1. $C \sqsubseteq D$,
2. $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$.

In fact, the closer the $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C)$ value is to 1, the more likely the corresponding subsumption may hold. More precisely, the label and edge constraints in $\mathcal{T}_D$ can likely be simulated by those in $\mathcal{T}_C$.

Consider the case where the two description trees are identical and all the successors have similar edges of the same role names. In such a case, the complexity of the function hd is $O(|V||V|)$ where V is the set of vertices of the description tree. Intuitively, with the recursive implementation detail of the function hd (see Algorithm 2), vertices of the description trees at the same level have to be recursively compared. The similarity degree obtained from each node is then propagated and returned back to the caller once the algorithm reaches the leaves of the two description trees. Though, it is suited for implementation, the time required for similarity measure using Algorithm 2 is sometimes prohibitive. This is due to the nature of recursion which is usually slow since all function calls need to be pushed onto a stack (i.e., a block of memory has to be newly allocated for each call) and popped when returned back.

The homomorphism degree function hd we introduced is quite similar to $simi_a$ proposed in [19] in the sense that they are both recursive definitions for the same DL $\mathcal{ELH}$. In fact, the operators that represent the t-conorm, and fuzzy connector are relatively used but differently defined. However, unlike the work proposed by [19], the use of μ and the way it is weighted, which determines how important the primitive concepts are to be considered, is defined. One another potential drawback of [19] can be an abstract framework that solely describes the desirable properties of their proposed similarity measure. For example, rather than providing exact definitions of the operators required for computation, merely promising properties are given. This is unclear and makes it difficult to implement. The other is obviously the distinction of their inspirations. While $simi_a$ is inspired by the Jaccard Index [13], the function hd proposed in this work is, on the other hand, motivated by the homomorphism-based structural subsumption characterization. With hd, concept names are transformed into an $\mathcal{ELH}$ concept description trees. Taking this as an advantage, in the next section, a bottom-up approach, which is in practice three times faster than the counterpart top-down based algorithm, is introduced (see more detail about the average computation time reported in Section 7.2). In addition to the advantage, the proposed bottom-up algorithm does away with costly recursive calls by making use of solutions to subproblems. As a natural step of enhancement, dynamic programming that conceptually takes benefit of overlapping subproblems, can be invented.

4 A BOTTOM-UP ALGORITHM

This section describes how the computation of the homomorphism degree based on Definition 6 can be alternatively achieved in a bottom-up manner. Let $\mathcal{T}$ and $\mathcal{T}'$ be the corresponding $\mathcal{ELH}$ description trees for two concepts of interest, and $\mathsf{hdtab} : V \times V' \to [0, 1]$ be a memory table storing the homomorphism degree from

$v \in V$ to $v' \in V'$. Then, a computation for the homomorphism degree mapping the root of $\mathcal{T}$ to the root of $\mathcal{T}'$ can be achieved using Algorithm 3. For any $\mathcal{ELH}$ description trees, we denote by $\mathsf{children}(\cdot)$ the function that returns all children of a given node, $\mathsf{nodesAtDepth}(\cdot)$ the function that returns a set of all nodes at a certain depth, and $\mathsf{depth}(\cdot)$ the function that returns the depth of a tree.

In contrast to the recursive approach as previously described in Section 3, Proposition 9 suggests that the bottom-up algorithm can begin by finding the lower depth d between the two description trees in question. Further computation for nodes below the level $d+1$, i.e. a computation for $\mathsf{e\text{-}set\text{-}hd}$, is irrelevant (see a proof in Appendix). In addition, at level d of both trees, only the computation for the degree of the label set inclusion $\mathsf{p\text{-}hd}$ is relevant. The computed value is then kept in the memory table hdtab and lately referred to by a parent node. With this technique, the homomorphism degrees of parent nodes can be iteratively computed by reusing the computed value of their children.

In what follows, $\mathcal{T}_{\text{cutoff}=c}$ is the rooted tree obtained from $\mathcal{T}$ by removing the vertices at the depth $> c$.

Definition 8. Let $\mathcal{T} = (V, E, rt, \ell, \rho)$ be a rooted tree of depth d. A *rooted cutoff tree* of depth c of $\mathcal{T}$, with $c \leq d$, is a rooted tree $\mathcal{T}_{\text{cutoff}=c} = (V', E', rt', \ell', \rho')$ where $rt' = rt$, $V' = \{v \in V \mid v \text{ is of depth} \leq c\}$, $E' = E \cap (V' \times V')$, ℓ' and ρ' are induced functions from ℓ, ρ w.r.t. V' and E', respectively.

For any pair of $\mathcal{ELH}$ description trees, the degree of similarity can be computed in a bottom-up manner starting from the depth of the shortest description tree.

Proposition 9. Let $\mathcal{T}$, $\mathcal{T}'$ be the rooted $\mathcal{ELH}$ description trees of a depth d and d', respectively, where $d \leq d'$. Then,

1. $\mathsf{hd}(\mathcal{T}, \mathcal{T}') = \mathsf{hd}(\mathcal{T}, \mathcal{T}'_{\text{cutoff}=d+1})$ and
2. $\mathsf{hd}(\mathcal{T}', \mathcal{T}) = \mathsf{hd}(\mathcal{T}'_{\text{cutoff}=d+1}, \mathcal{T})$.

Runtime Analysis. As shown in Algorithm 3, the complexity of $\mathsf{bu\text{-}e\text{-}set\text{-}hd}(v, v')$ is proportional to $O(\mathsf{bu\text{-}e\text{-}set\text{-}hd}(v, v')) = O(|\mathsf{children}(v)| \cdot |\mathsf{children}(v')|)$. Since all nodes except the root are child nodes, it always holds that $O(\mathsf{bu\text{-}e\text{-}set\text{-}hd}(v, v')) \leq O(|V| \cdot |V'|)$. Let d be as defined in Algorithm 3, and V_i,V_i' be a set of all vertices at level i of $\mathcal{T}$,$\mathcal{T}'$, respectively. The time complexity required for computing degree among the nodes at level i of the two given $\mathcal{ELH}$ description trees is $O(\sum_{v \in V_i} \sum_{v' \in V_i'} (|\mathsf{children}(v)| \cdot |\mathsf{children}(v')|))$. Therefore, the complexity of the entire algorithm $O(\mathsf{bu\text{-}hd}(\mathcal{T}, \mathcal{T}'))$ is proportional to $O(\sum_{i=0}^{d} \sum_{v \in V_i} \sum_{v' \in V_i'} (|\mathsf{children}(v)| \cdot |\mathsf{children}(v')|))$. Consider the following:

$$\begin{aligned}
&\textstyle\sum_{i=0}^{d}\Big(\sum_{v\in V_i}\sum_{v'\in V_i'}(|\mathsf{children}(v)|\cdot|\mathsf{children}(v')|)\Big)\\
&= \textstyle\sum_{i=0}^{d}\Big(\sum_{v\in V_i}(|\mathsf{children}(v)|\cdot\sum_{v'\in V_i'}(|\mathsf{children}(v')|))\Big)
\end{aligned}$$

//since children of all vertices at level i are the vertices at level $i+1$, this infers

$$= \textstyle\sum_{i=1}^{d+1}(|V_i|\cdot|V_i'|).$$

Therefore, if $d = \mathsf{depth}(\mathcal{T}) \leq \mathsf{depth}(\mathcal{T}')$, the complexity is $O(|V||V'_{\mathsf{cutoff}=d+1}|)$ where $V'_{\mathsf{cutoff}=d+1}$ is a set of all vertices of $\mathcal{T}'_{\mathsf{cutoff}=d+1}$. Intuitively, it is possible to say that only the set of edges and vertices at the depth $\leq d+1$ is relevant to the complexity of the bottom-up approach.

5 $\mathcal{ELH}$ SEMANTIC SIMILARITY

The homomorphism degree function provides a numerical value that represents structural similarity of one concept description when compared against another concept description. As illustrated by the example in Section 3, the direction of the homomorphism degree matters, viz., $\mathsf{hd}(\mathcal{T}_{\mathsf{Naaf}},\mathcal{T}_{\mathsf{Nam}}) = 0.856$, whereas $\mathsf{hd}(\mathcal{T}_{\mathsf{Nam}},\mathcal{T}_{\mathsf{Naaf}})$ $= 0.6$. Since both directions constitute the degree of the two concepts being equivalent, our similarity measure for $\mathcal{ELH}$ concept descriptions is defined by means of these values.

Definition 10. Let C, D be expanded $\mathcal{ELH}$ concept descriptions. The *degree of similarity between C and D* is defined as:

$$\mathsf{sim}(C,D) := \frac{\mathsf{hd}(\mathcal{T}_C,\mathcal{T}_D)+\mathsf{hd}(\mathcal{T}_D,\mathcal{T}_C)}{2}. \tag{6}$$

Intuitively, the degree of similarity between two concepts is the average of the degree of having homomorphisms in both directions, thus $\mathsf{sim}(C,D) = \mathsf{sim}(D,C)$ as required. Note that one could adopt an alternative definition, e.g. based on the multiplication $\mathsf{sim}^{mult}(C,D)$ or the root mean square $\mathsf{sim}^{rms}(C,D)$ [29]. These would give rather unsatisfactory values for extreme cases such as the concepts A and $\top$, where $\mathsf{sim}^{mult}(A,\top) = 0$ and $\mathsf{sim}^{rms}(A,\top) = 0.856$. Since $\mathsf{sim}^{mult}(C,D) \leq \mathsf{sim}(C,D) \leq \mathsf{sim}^{rms}(C,D)$, we believe that the average-based definition given above is most appropriate.

Based on the homomorphism degree values in Table 2, the degrees of similarity among the defined concepts in the example ontology $\mathcal{O}_{\mathsf{med}}$ can be obtained; see Table 3. Observe that there are two mutually exclusive clusters of similar concepts $\{\mathsf{Hypoxia},\mathsf{Hypoxemia}\}$ and a set of neonatal aspiration causing by different agents including {NA-Milk, NA-AmnioticFluid, NA-Mucus, NA-Meconium}. Observe that concepts from the same clusters are relatively similar (i.e. $\mathsf{sim} \geq 0.69$) and those

Algorithm 3 $\mathcal{ELH}$ similarity measure using the bottom-up approach

Input: Two $\mathcal{ELH}$ description trees $\mathcal{T}$ and $\mathcal{T}'$
Output: The homomorphism degree from $\mathcal{T}$ to $\mathcal{T}'$

```
function bu-hd($\mathcal{T}, \mathcal{T}'$)

$hd \leftarrow 0$
$d \leftarrow i \leftarrow \mathsf{min}(\mathsf{depth}(\mathcal{T}), \mathsf{depth}(\mathcal{T}'))$
while $i \geq 0$ do
    $V_i \leftarrow \mathcal{T}.\mathsf{nodesAtDepth}(i)$
    $V_i' \leftarrow \mathcal{T}'.\mathsf{nodesAtDepth}(i)$
    for each $v \in V_i$ do
        for each $v' \in V_i'$ do
            if $i = d$ then
                $hd \leftarrow phd \leftarrow \frac{|\ell(v) \cap \ell(v')|}{|\ell(v)|}$
                $\mathsf{hdtab}(v, v') \leftarrow hd$
            else
                if $| \ell(v) | = 0$ then
                    $phd \leftarrow 1$
                else
                    $phd \leftarrow \frac{|\ell(v) \cap \ell(v')|}{|\ell(v)|}$
                $ehd \leftarrow \mathsf{bu\text{-}e\text{-}set\text{-}hd}(v, v')$
                $\mu \leftarrow \frac{|\ell(v)|}{|\ell(v)| + |\mathsf{children}(v)|}$
                $hd \leftarrow \mu \cdot phd + (1 - \mu) \cdot ehd$
                $\mathsf{hdtab}(v, v') \leftarrow hd$
    $i \leftarrow i - 1$
return $hd$
function bu-e-set-hd($v, v'$)
$W \leftarrow \mathsf{children}(v)$
$W' \leftarrow \mathsf{children}(v')$
$sumEHD \leftarrow 0$
for each $w \in W$ do
    $maxEHD \leftarrow 0$
    $curEHD \leftarrow 0$
    for each $w' \in W'$ do
        $\gamma \leftarrow \frac{|\rho(v,w) \cap \rho(v',w')|}{|\rho(v,w)|}$
        $curEHD \leftarrow \gamma(v + (1 - v) \cdot \mathsf{hdtab}(w, w'))$
        if $curEHD > maxEHD$ then
            $maxEHD \leftarrow curEHD$
    $sumEHD \leftarrow sumEHD + maxEHD$
return $\frac{sumEHD}{|W|}$
```

$\mathsf{hd}(\downarrow, \rightarrow)$	Neonatal Aspiration of				HPX	HPM
	Milk	AF	Mucus	Mec		
Milk	1.00	0.72	0.73	0.69	0.16	0.20
AmnioticFluid	–	1.00	0.90	0.82	0.16	0.20
Mucus	–	–	1.00	0.82	0.16	0.20
Meconium	–	–	-	1.00	0.16	0.20
Hypoxia	–	–	–	–	1.00	0.92
Hypoxemia	–	–	–	–	–	1.00

Table 3. Similarity degrees among the defined concepts in $\mathcal{O}_{\mathsf{med}}$

from different clusters are apparently dissimilar (i.e. $\mathsf{sim} \leq 0.20$). Note that there is some similarity between these two clusters due to their share of the special-purpose role called roleGroup [24, 27] (more detail in Subsection 7.1). Note that, though not included in Table 2 and 3, the similarity involving primitive concepts like Substance and Foods can also be computed. Nevertheless, the pairwise similarity degree between any two primitive concepts is zero by our definition since there is absolutely no commonality between them apart from both being subsumed by $\top$.

6 DESIRABLE PROPERTIES FOR CONCEPT SIMILARITY MEASURE

This section describes desirable properties for concept similarity measure and provides corresponding mathematical proofs. At the end of the section, a comparison of satisfactory properties between our similarity measure and other works is presented.

Definition 11 summarizes important properties for concept similarity measure introduced in [19]. These are believed to be desirable features.

Definition 11. Let C, D and E be $\mathcal{ELH}$ concept, the similarity measure is

1. *symmetric* iff $\mathsf{sim}(C, D) = \mathsf{sim}(D, C)$,
2. *equivalence closed* iff $\mathsf{sim}(C, D) = 1 \iff C \equiv D$,
3. *equivalence invariant* if $C \equiv D$ then $\mathsf{sim}(C, E) = \mathsf{sim}(D, E)$,
4. *subsumption preserving* if $C \sqsubseteq D \sqsubseteq E$ then $\mathsf{sim}(C, D) \geq \mathsf{sim}(C, E)$,
5. *reverse subsumption preserving* if $C \sqsubseteq D \sqsubseteq E$ then $\mathsf{sim}(C, E) \leq \mathsf{sim}(D, E)$,
6. *structurally dependent* Let C_i and C_j be atoms in C where $C_i \not\sqsubseteq C_j$, the concept $D' := \bigsqcap_{i \leq n} C_i \sqcap D$ and $E' := \bigsqcap_{i \leq n} C_i \sqcap E$ satisfies the condition $\lim_{n \to \infty} \mathsf{sim}(D', E') = 1$,
7. satisfying *triangle inequality* iff $1 + \mathsf{sim}(D, E) \geq \mathsf{sim}(D, C) + \mathsf{sim}(C, E)$.

Theorem 12 states the characteristics of sim.

Theorem 12. The similarity-measure sim is 1) symmetric, 2) equivalence closed, 3) equivalence invariant, 4) subsumption preserving, 5) structurally dependent, 6) not reverse subsumption preserving, and 7) not satisfying triangle inequality.

Proof.

1. By Definition 10, it is obvious that $\mathsf{sim}(C, D) = \mathsf{sim}(D, C)$.
2. ($\Longrightarrow$) By Definition 6, $\mathsf{sim}(C, D) = 1$ iff $\mathsf{hd}(\mathcal{T}_C, \mathcal{T}_D) = 1$ and $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$. By Proposition 7, these imply that $C \sqsubseteq D$ and $D \sqsubseteq C$. Therefore, $C \equiv D$. ($\Longleftarrow$) Assume $C \equiv D$, then $C \sqsubseteq D$ and $D \sqsubseteq C$. Using the same proposition, this ensures that $\mathsf{hd}(\mathcal{T}_C, \mathcal{T}_D) = 1$, and $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$, which means $\mathsf{sim}(C, D) = 1$.
3. $C \equiv D$ iff $C \sqsubseteq D$ and $D \sqsubseteq C$. By using Corollary 5, we have $\mathcal{P}_C = \mathcal{P}_D$ and $\mathcal{E}_C \cong \mathcal{E}_D$. Therefore, $\mathcal{T}_C = \mathcal{T}_D$, and this implies $\mathsf{hd}(\mathcal{T}_C, \mathcal{T}_E) = \mathsf{hd}(\mathcal{T}_D, \mathcal{T}_E)$ and $\mathsf{hd}(\mathcal{T}_E, \mathcal{T}_C) = \mathsf{hd}(\mathcal{T}_E, \mathcal{T}_D)$, such that $\mathsf{sim}(C, E) = \mathsf{sim}(D, E)$.
4. We need to show that $\frac{\mathsf{hd}(\mathcal{T}_C,\mathcal{T}_D)+\mathsf{hd}(\mathcal{T}_D,\mathcal{T}_C)}{2} \geq \frac{\mathsf{hd}(\mathcal{T}_C,\mathcal{T}_E)+\mathsf{hd}(\mathcal{T}_E,\mathcal{T}_C)}{2}$.
   Since $C \sqsubseteq D$ and $D \sqsubseteq E$, then $C \sqsubseteq E$. By Proposition 7, $\mathsf{hd}(\mathcal{T}_E, \mathcal{T}_C) = 1$ and $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$, we need to show that $\mathsf{hd}(\mathcal{T}_C, \mathcal{T}_D) \geq \mathsf{hd}(\mathcal{T}_C, \mathcal{T}_E)$. That means it is adequate to show that $\mathsf{p\text{-}hd}(\mathcal{P}_C, \mathcal{P}_D) \geq \mathsf{p\text{-}hd}(\mathcal{P}_C, \mathcal{P}_E)$ and $\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_C, \mathcal{E}_D) \geq \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_C, \mathcal{E}_E)$. For the first part, we show that $\frac{|\mathcal{P}_C \cap \mathcal{P}_D|}{|\mathcal{P}_C|} \geq \frac{|\mathcal{P}_C \cap \mathcal{P}_E|}{|\mathcal{P}_C|}$. By Corollary 4, $C \sqsubseteq D \sqsubseteq E$ means $\mathcal{P}_E \subseteq \mathcal{P}_D \subseteq \mathcal{P}_C$. Therefore $\mid \mathcal{P}_D \mid \geq \mid \mathcal{P}_E \mid$ and $\frac{|\mathcal{P}_C \cap \mathcal{P}_D|}{|\mathcal{P}_C|} \geq \frac{|\mathcal{P}_C \cap \mathcal{P}_E|}{|\mathcal{P}_C|}$ is true. For the second part, we show that
   $$\sum_{\epsilon_i \in \mathcal{E}_C} \frac{\max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_D\}}{\mid \mathcal{E}_C \mid} \geq \sum_{\epsilon_i \in \mathcal{E}_C} \frac{\max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_E\}}{\mid \mathcal{E}_C \mid}, \qquad (7)$$
   $$\sum_{\epsilon_i \in \mathcal{E}_C} \max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_D\} \geq \sum_{\epsilon_i \in \mathcal{E}_C} \max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_E\}.$$
   Let $\hat{\epsilon}_j \in \mathcal{E}_E$ such that $\mathsf{e\text{-}hd}(\epsilon_i, \hat{\epsilon}_j) = max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_E\}$, but since $\hat{\epsilon}_j \in \mathcal{E}_E \subseteq \mathcal{E}_D$, then $max\{\mathsf{e\text{-}hd}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_D\} \geq \mathsf{e\text{-}hd}(\epsilon_i, \hat{\epsilon}_j)$. Therefore, Equation (7) is true.
5. Let $D' := \bigsqcap_{i \leq n} C_i \sqcap D$, $E' := \bigsqcap_{i \leq n} C_i \sqcap E$, and $n = n_{\mathcal{P}} + n_{\mathcal{E}}$ be the number of all atom sequences in C where $n_{\mathcal{P}}$ and $n_{\mathcal{E}}$ be the number of primitive concepts and the number existential restrictions, respectively. To prove this, we consider the following case distinction.
   (a) if $n_{\mathcal{P}} \to \infty$ and $n_{\mathcal{E}}$ is finite, it suffices to show that $\lim_{n_{\mathcal{P}} \to \infty} \mu = 1$ and $\lim_{n_{\mathcal{P}} \to \infty} \mathsf{p\text{-}hd}(\mathcal{P}_{D'}, \mathcal{P}_{E'}) = 1$. Therefore, $\mathsf{hd}(D', E') = \mathsf{hd}(E', D') = 1$ and these imply that $\mathsf{sim}(D', E') = 1$. From Equation (3), we have:
   $$\mu = \frac{\mid \mathcal{P}_{D'} \mid}{\mid \mathcal{P}_{D'} \cup \mathcal{E}_{D'} \mid} = \frac{\mid \mathcal{P}_{D'} \mid}{\mid \mathcal{P}_{D'} \mid + \mid \mathcal{E}_{D'} \mid} = \frac{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid}{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid + \mid \mathcal{E}_{D'} \mid}$$
   $$= \frac{n_{\mathcal{P}} + \mid \mathcal{P}_D \mid}{n_{\mathcal{P}} + \mid \mathcal{P}_D \mid + \mid \mathcal{E}_{D'} \mid}. \qquad (8)$$

Since $\mid \mathcal{P}_D \mid$ and $\mid \mathcal{E}_{D'} \mid$ are constant, $\lim_{n_\mathcal{P}\to\infty} \mu = \lim_{n_\mathcal{P}\to\infty} \frac{n_\mathcal{P}+|\mathcal{P}_D|}{n_\mathcal{P}+|\mathcal{P}_D|+|\mathcal{E}_{D'}|} = 1$. For the second part, we have:

$$\mathsf{p\text{-}hd}(\mathcal{P}_{D'}, \mathcal{P}_{E'}) = \frac{\mid \mathcal{P}_{D'} \cap \mathcal{P}_{E'} \mid}{\mid \mathcal{P}_{D'} \mid} = \frac{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \cap \mathcal{P}_E \mid}{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid} = \frac{n_\mathcal{P} + \mid \mathcal{P}_D \cap \mathcal{P}_E \mid}{n_\mathcal{P} + \mid \mathcal{P}_D \mid}$$

where $\mid \mathcal{P}_D \cap \mathcal{P}_E \mid$ and $\mid \mathcal{P}_D \mid$ are constant. Thus,

$$\lim_{n_\mathcal{P}\to\infty} \mathsf{p\text{-}hd}(\mathcal{P}_{D'}, \mathcal{P}_{E'}) = \lim_{n_\mathcal{P}\to\infty} \frac{n_\mathcal{P} + \mid \mathcal{P}_D \cap \mathcal{P}_E \mid}{n_\mathcal{P} + \mid \mathcal{P}_D \mid} = 1. \quad (9)$$

(b) if $n_\mathcal{E} \to \infty$ and $n_\mathcal{P}$ is finite, it suffices to show that $\lim_{n_\mathcal{E}\to\infty} \mu = 0$ and $\lim_{n_\mathcal{E}\to\infty} \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) = 1$ which implies $\mathsf{hd}(D', E') = \mathsf{hd}(E', D') = 1$, and $\mathsf{sim}(D', E') = 1$. From Equation (8), the value of μ is as follows:

$$\mu = \frac{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid}{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid + \mid \mathcal{E}_{D'} \mid} = \frac{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid}{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid + n_\mathcal{E} + \mid \mathcal{E}_D \mid}.$$

Since $\mid \mathcal{P}_C \mid$, $\mid \mathcal{P}_D \mid$ and $\mid \mathcal{E}_D \mid$ are constant, by taking limit, we have:

$$\lim_{n_\mathcal{E}\to\infty} \mu = \lim_{n_\mathcal{E}\to\infty} \frac{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid}{\mid \mathcal{P}_C \mid + \mid \mathcal{P}_D \mid + n_\mathcal{E} + \mid \mathcal{E}_D \mid} = 0.$$

To show that $\lim_{n_\mathcal{E}\to\infty} \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) = 1$, we have:

$$\begin{aligned}
&\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) \\
&= \sum_{e_i \in \mathcal{E}_{D'}} \frac{\max\{\mathsf{e\text{-}hd}(e_i, e_j) : e_j \in \mathcal{E}_{E'}\}}{\mid \mathcal{E}_{D'} \mid} \\
&= \frac{\sum_{e_i \in \mathcal{E}_{D'}} \max\{\mathsf{e\text{-}hd}(e_i, e_j) : e_j \in \mathcal{E}_{E'}\}}{\mid \mathcal{E}_{D'} \mid} \\
&= \frac{\sum_{e_i \in \mathcal{E}_C} \max\{\mathsf{e\text{-}hd}(e_i, e_j) : e_j \in \mathcal{E}_{E'}\} + \sum_{e_i \in \mathcal{E}_D} \max\{\mathsf{e\text{-}hd}(e_i, e_j) : e_j \in \mathcal{E}_{E'}\}}{\mid \mathcal{E}_C \cup \mathcal{E}_D \mid}.
\end{aligned}$$

Since $\mathcal{E}_C \subseteq \mathcal{E}_{E'}$, for each $\epsilon_i \in \mathcal{E}_C$ there is $\epsilon_j \in \mathcal{E}_{E'}$ such that $\epsilon_i = \epsilon_j$. Thus,

$$\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) = \frac{n_\mathcal{E} + p}{\mid \mathcal{E}_C \mid + \mid \mathcal{E}_D \mid} = \frac{n_\mathcal{E} + p}{n_\mathcal{E} + \mid \mathcal{E}_D \mid}$$

where $p = \sum_{e_i \in \mathcal{E}_D} \max\{\mathsf{e\text{-}hd}(e_i, e_j) : e_j \in \mathcal{E}_{E'}\}$, and $p \leq\mid \mathcal{E}_D \mid$. Therefore,

$$\lim_{n_\mathcal{E}\to\infty} \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) = \lim_{n_\mathcal{E}\to\infty} \frac{n_\mathcal{E} + p}{n_\mathcal{E} + \mid \mathcal{E}_D \mid} = 1. \quad (10)$$

(c) if $n_{\mathcal{P}} \to \infty$ and $n_{\mathcal{E}} \to \infty$, it suffices to show that $\lim_{n_{\mathcal{P}}\to\infty} \mathsf{p\text{-}hd}(\mathcal{P}_{D'}, \mathcal{P}_{E'}) = 1$ and $\lim_{n_{\mathcal{E}}\to\infty} \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_{D'}, \mathcal{E}_{E'}) = 1$. These follow from Equations (9) and (10).

6. Consider a counter example defined in Figure 4. It is obvious that $C \sqsubseteq D \sqsubseteq E$. By definition, $\mathsf{sim}(C, E) = 0.7125$ and $\mathsf{sim}(D, E) = 0.6667$. Apparently, there exists the case $\mathsf{sim}(C, E) \not\leq \mathsf{sim}(D, E)$.

$$\boxed{\begin{array}{rcl} \mathsf{E} & \equiv & \exists \mathsf{r}.(\mathsf{F} \sqcap \mathsf{G}) \\ \mathsf{D} & \equiv & \exists \mathsf{r}.(\mathsf{F} \sqcap \mathsf{G}) \sqcap \exists s.\mathsf{F} \ \sqcap \exists s.\mathsf{G} \\ \mathsf{C} & \equiv & \exists \mathsf{r}.(\mathsf{F} \sqcap \mathsf{G}) \sqcap \exists s.\mathsf{F} \sqcap \exists s.\mathsf{G} \sqcap \exists \mathsf{r}.(\mathsf{F} \sqcap \mathsf{H}) \end{array}}$$

Figure 4. Examples of $\mathcal{ELH}$ concept descriptions

7. Providing the concept description C, D, and E defined in Figure 4, the following demonstrates the case $1 + \mathsf{sim}(D, E) \not\geq \mathsf{sim}(D, C) + \mathsf{sim}(C, E)$. Here, we have $\mathsf{sim}(D, E) = 0.6667$, $\mathsf{sim}(D, C) = 0.9625$ and $\mathsf{sim}(C, E) = 0.7125$. By applying a summation, it is obvious that $1.6667 \not\geq 1.675$.

□

To ensure that our proposed method reaches the performance, Table 4 compares desirable properties of our similarity measure sim against those previously reported in other works. Besides the work proposed by [13], which allows only concept conjunction, our approach and the one proposed in [19] apparently hold significant features.

Similarity Measure	DL	symmetric	equi. closed	equi invariant	sub. preserving	struc. dependent	rev. sub. preserving	triangle inequality
Our measure sim	$\mathcal{ELH}$	✓	✓	✓	✓	✓		
d'Amato et al. [8]	$\mathcal{ALC}$							
d'Amato et al. [7]	$\mathcal{ALC}$	✓		✓		✓	✓	
d'Amato et al. [9]	$\mathcal{ALE}$	✓		✓		✓	✓	
Fanizzi and d'Amato [11]	$\mathcal{ALN}$	✓		✓		✓	✓	
Jaccard [13]	$\mathcal{L}_0$	✓	✓	✓	✓	✓	✓	✓
Janowicz and Wilkes [15]	$\mathcal{SHI}$	✓						✓
Janowicz [14]	$\mathcal{ALCHQ}$	✓						✓
Lehmann and Turhan [19]	$\mathcal{ELH}$	✓	✓	✓	✓	✓		

Table 4. A comparison on concept-similarity properties [19]

7 EXPERIMENTS AND RESULTS

To measure the reasoning performance, in this work, one of the most well-known life science ontologies that has been developed based on an inexpressive DLs of the $\mathcal{ELH}$ family, SNOMED CT is selected. Among a variety of clinical ontologies, SNOMED CT provides a standard terminology such that clinical and medical concepts are formally defined and are related to each other not only by the subsumption relation but also by other domain-specific relations by means of roles. In our experiments, we employ a DL version released in January 2005 which contains 379 691 concept names and 62 role names, hitherto referred to as $\mathcal{O}_{\text{SNOMED}}$. In the version of SNOMED CT, the defined concepts are broadly categorized as subconcepts of one of 18 mutually exclusive top-level concepts. Hence, in a sense of subsumption relation, the concepts that belong to different categories are likely more dissimilar than those that belong to the same category.

The experiments are performed on a laptop with a 1.7 GHz Intel Core i5 CPU, and 4 GB of memory. Though our optimized bottom-up algorithm takes less computation time, the overall number of pairs of concepts in $\mathcal{O}_{\text{SNOMED}}$ is approximately 10^{11}. Assuming the similarity measure between each pair of concepts can be computed in a millisecond, we still need about 1 157 days. For this reason, we consider well-representative samples of pairs of concepts in SNOMED CT [27]. For each top-level category C_i where $1 \leq i \leq 18$, 0.05 % of all concepts in category C_i are sampled (the way we sample is similar to that reported in [5]). Hitherto, they are called *stem* concepts and denoted by *c-sample*(C_i). For each category C_i and each stem concept D in *c-sample*(C_i), we do the following:

1. Extract all stated and inferred superconcepts[2] E of D, i.e. $D \sqsubseteq_{\mathcal{O}_{\text{SNOMED}}} E$. Let k be the number of superconcepts E.
2. Sample k non-subsumers F of D, i.e. $D \not\sqsubseteq_{\mathcal{O}_{\text{SNOMED}}} F$ and $F \not\sqsubseteq_{\mathcal{O}_{\text{SNOMED}}} D$.
3. Sample k non-subsumers G from category C_j, where $j \neq i$ and j is randomized for every sampling.

For each category C_i, we systematically and semantically measure the degrees of similarity between the following concepts:

1. D and E (i.e. similarity between two concepts in the subsumption relation).
2. D and F (i.e. similarity between two concepts that belong to the same category but are not in the subsumption relation).
3. D and G (i.e. similarity between concepts from different categories, and thus known not to be in the subsumption relation).

For later references, let δ_i^1, δ_i^2 and δ_i^3 denote sets of concept pairs defined in (1), (2) and (3), respectively. We denote by *test-set*$(C_i) = (\delta_i^1, \delta_i^2, \delta_i^3)$ the test set for top-level category C_i in $\mathcal{O}_{\text{SNOMED}}$.

[2] All superconcepts E used in the experiments are obtained by exploiting the CEL reasoner [4].

7.1 Snomed ct-Specific Similarity Measure

The use of relation group was inherent in Snomed ct. However, by using the so-called roleGroup [24, 25] to represent a group of existential restrictions, this would unintentionally increase the degree of similarity due to role commonality (i.e. the increasing of γ). For example, although there is neither implicit nor explicit relationship between the concept Hypoxia and the concept NeonatalAspirationOfMucus (see Table 3), there still exists some degrees of similarity between the two concepts (0.167 degrees), but since roleGroup occurs preceding every existential restriction, it is meaningless to regard co-occurrence of this role in two Snomed concepts as being similar. The similarity measuring of roleGroup in such a case is therefore ignored. If roleGroup is found, we set $\nu = 0$ (i.e., the similarity of roleGroup is not considered).

hd(↓, →)	Neonatal Aspiration of				HPX	HPM
	Milk	AF	Mucus	Mec		
Milk	1.0	0.63	0.6375	0.575	0.0	0.0
AmnioticFluid	–	1.0	0.835	0.7	0.0	0.0
Mucus	–	–	1.0	0.7	0.0	0.0
Meconium	–	–	–	1.0	0.0	0.0
Hypoxia	–	–	–	–	1.0	0.875
Hypoxemia	–	–	–	–	–	1.0

Table 5. Similarity degree among the defined concepts in $\mathcal{O}_{\mathsf{med}}$ using the augmented algorithm for Snomed ct

Another point to take into account is a share of the Snomed top concept SCT-Top in every expanded description. In Snomed ct, defined concepts are categorized under the 18 top-level concepts. Intuitively, without loss of commonality, SCT-Top can be presumably neglected and rather we can treat all those 18 top-level concepts as direct sub-concepts of $\top$. Based on this justification and the omission of roleGroup, the adjustment of Algorithm 2 for Snomed ct can be adopted.

Table 5 depicts the new degrees of similarity among the defined concepts in $\mathcal{O}_{\mathsf{med}}$ after applying the augmented algorithm for Snomed ct. Note that the degrees of similarity among the concepts within the same cluster whose descriptions are nestedly defined using roleGroup are slightly reduced and likewise for those from mutually exclusive clusters (i.e, the concepts are totally dissimilar). Further to the augmented algorithm, the illustrative concept NeonatalAspirationOfMucus are completely dissimilar (i.e. having a degree of 0) to both the concept Hypoxia and the concept Hypoxemia, as desired.

7.2 Experimental Results on Snomed ct

In this section, we describe results obtained from a test set *test-set*(C_i) where $1 \leq i \leq 18$. Table 6 shows extensive results from the described experiments. The first two columns show the top-level concept categories and the number of sampled stem concepts D. The third column shows the number of superconcepts. The next three columns depict the degrees of similarity between each stem concept D and each of the concepts E, F, G described above (i.e. concept pairs in δ_i^1, δ_i^2, δ_i^3, respectively). In all three columns, the average, median, and max values are provided. The last column describes the number of recursive calls required by Algorithm 2. The overall results for all samples of all categories are shown in the last row of the table. On average, the bottom-up algorithm requires 0.157, 0.297, and 0.047 milliseconds, whereas the top-down algorithm requires 0.554, 0.974, and 0.099 milliseconds for each comparison in δ_i^1, δ_i^2, δ_i^3, respectively. More specifically, the average time required by the bottom-up algorithm for each computation is merely 0.167 milliseconds. This is three times faster than the counterpart top-down algorithm, which requires 0.542 milliseconds on average.

As shown in Table 6, due to a commonality of the top-level concept categories (i.e. a share of top-level primitive concepts), it is not surprising that the degrees of similarity among the concepts within the same category (see e.g. the average, median, and mean of $\mathsf{sim}(D, E)$ and $\mathsf{sim}(D, F)$) are obviously higher than those of the concepts from different categories (see e.g. the average, median, and mean of $\mathsf{sim}(D, G)$); likewise, the similarity between the concepts having more commonality (i.e. subsumees and subsumers) are higher than those between non-subsumption related concepts ($\mathsf{sim}(D, E) > \mathsf{sim}(D, F)$).

Being sampled from different categories and having no commonality, those pairs in δ_i^3 are mostly judged as totally dissimilar. Interestingly, there are few cases where the degree is nonzero. For example, the similarity between the concepts $\mathsf{HemorrhageIntoBladderLumen}$ (HBL) and $\mathsf{IrrigationOfGingivalOperculum}$ (IGO) whose categories are *Body Structure* and *Procedure* respectively (see definitions in Figure 5) can be computed as $\mathsf{sim}(\hat{\mathsf{HBL}}, \hat{\mathsf{IGO}}) = \frac{\mathsf{hd}(\mathcal{T}_{\hat{\mathsf{HBL}}}, \mathcal{T}_{\hat{\mathsf{IGO}}}) + \mathsf{hd}(\mathcal{T}_{\hat{\mathsf{IGO}}}, \mathcal{T}_{\hat{\mathsf{HBL}}})}{2} = \frac{0.56462+0.02892}{2} = 0.29677$. Here the obtained degrees reveal the hidden knowledge that there exists some relationship between the two concepts. However without considering the degrees of similarity, it is still possible to argue that these concepts are related in terms of plain medical definitions. Consider the concept $\mathsf{HemorrhageIntoBladderLumen}$ which is a disorder in the urinary system. As a consequence of bleeding in the bladder lumen, in many cases, there is a chance that the blood will clot and block the flow of urine. A common curing procedure in such a case is to open a urinary catheter and insert a saline into the catheter in order to remove the clot using a sterile fluid, so that the urine can drain from the bladder clearly. This procedure is commonly known as an *irrigation procedure*. Not only is irrigation applicable for hemorrhaging into bladder lumen, but it is also applicable to any disorders where the cleaning of organs using sterile fluid is needed, i.e. ears, gingival operculum, bowel, etc. Based on this supportive argument, it is thus possible to conclude that the concepts $\mathsf{HemorrhageIntoBladderLumen}$ and $\mathsf{IrrigationOfGingivalOperculum}$ are relevant. Hence, the proposed system is capable of revealing hidden knowledge.

As shown in column six and seven in Table 6, the average time required by the bottom-up approach is evidently less than that required by the top-down approach.

SNOMED CT category C_i	♯ Stem concepts D	♯ Super concepts E	Similarity degrees among concepts			♯ Recursive calls of Algorithm 2 (δ_i^1 / δ_i^2 / δ_i^3)
			δ_i^1 (avg/median/max)	δ_i^2 (avg/median/max)	δ_i^3 (avg/median/max)	
Attribute	5	15	0.798 / 0.785 / 0.928	0.467 / 0.450 / 0.625	0.0 / 0.0 / 0.0	30 / 30 / 45
Body structure	158	4 422	0.666 / 0.622 / 0.993	0.272 / 0.265 / 0.923	0.0 / 0.0 / 0.0	10 712 / 14 214 / 11 148
Clinical Finding	557	9 998	0.779 / 0.785 / 0.999	0.367 / 0.333 / 0.965	x_1 / 0.0 / 0.296	3 103 238 / 4 917 882 / 436 426
Environments and geographical locations	8	35	0.788 / 0.785 / 0.937	0.278 / 0.171 / 0.666	0.0 / 0.0 / 0.0	70 / 70 / 70
Events	8	12	0.818 / 0.833 / 0.900	0.459 / 0.491 / 0.666	0.0 / 0.0 / 0.0	24 / 12 / 24
Observable entity	36	179	0.771 / 0.750 / 0.954	0.246 / 0.242 / 0.773	0.0 / 0.0 / 0.0	358 / 358 / 358
Organism	129	1 171	0.752 / 0.750 / 0.975	0.203 / 0.112 / 0.937	0.0 / 0.0 / 0.0	2 342 / 2 342 / 2 342
Pharmaceutical/ biologic product	163	907	0.774 / 0.786 / 0.984	0.312 / 0.313 / 0.916	0.0 / 0.0 / 0.0	59 574 / 108 406 / 17 124
Physical force	1	4	0.770 / 0.750 / 0.916	0.462 / 0.416 / 0.833	0.0 / 0.0 / 0.0	8 / 8 / 8
Physical object	21	124	0.778 / 0.781 / 0.954	0.367 / 0.416 / 0.729	0.0 / 0.0 / 0.0	248 / 248 / 248
Procedure	262	4 659	0.774 / 0.782 / 0.999	0.457 / 0.536 / 0.952	0.0 / 0.0 / 0.0	1 202 330 / 2 363 040 / 194 890
Qualifier value	41	134	0.781 / 0.750 / 0.937	0.227 / 0.198 / 0.857	0.0 / 0.0 / 0.0	268 / 268 / 268
Situation with explicit context	33	142	0.840 / 0.863 / 0.999	0.460 / 0.430 / 0.982	0.0 / 0.0 / 0.0	20 006 / 23 328 / 2 596
Social context	25	154	0.752 / 0.729 / 0.961	0.299 / 0.291 / 0.844	0.0 / 0.0 / 0.0	308 / 308 / 308
Special concept	318	636	0.812 / 0.812 / 0.875	0.602 / 0.500 / 0.750	0.0 / 0.0 / 0.0	1 272 / 1 272 / 1 272
Specimen	5	28	0.782 / 0.757 / 0.954	0.407 / 0.446 / 0.732	0.0 / 0.0 / 0.0	3 754 / 6 868 / 500
Staging and scales	5	10	0.805 / 0.833 / 0.928	0.446 / 0.504 / 0.714	0.0 / 0.0 / 0.0	20 / 20 / 20
Substance	114	1,067	0.709 / 0.678 / 0.983	0.237 / 0.166 / 0.882	0.0 / 0.0 / 0.0	2 128 / 3 201 / 2 128
Overall in SNOMED CT	1 889	23,697	0.753 / 0.75 / 0.9998	0.356 / 0.302 / 0.9829	x_2 / 0.0 / 0.2967	4 406 690 / 7 441 875 / 669 775

Table 6. Empirical results of the proposed similarity measure algorithms on SNOMED CT. For the sake of compactness, x_1 and x_2 stand for 0.0000296 and 0.00001252, respectively.

$$\begin{aligned}
\mathsf{HemorrhageIntoBladderLumen} \sqsubseteq\ & \mathsf{DisorderOfUrinaryBladder} \sqcap \mathsf{BladderHemorrhage} \\
& \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{findingSite}.\mathsf{UrinaryBladderStructure} \\
& \sqcap \exists \mathsf{associatedMorphology}.\mathsf{Hemorrhage}) \\
\mathsf{IrrigationOfGingivalOperculum} \sqsubseteq\ & \mathsf{OralIrrigation} \sqcap \mathsf{PeriodonticProcedure} \sqcap \exists \mathsf{roleGroup}.(\\
& \exists \mathsf{procedureSite}.\mathsf{StructureOfGumAndSupportingStructureOfTooth} \\
& \sqcap \exists \mathsf{method}.\mathsf{IrrigationAction}) \\
& \sqcap \exists \mathsf{roleGroup}.(\exists \mathsf{directSubstance}.\mathsf{IrrigatingSolution})
\end{aligned}$$

Figure 5. Examples of SNOMED CT concept definitions extracted from different categories

As pre-process, it is to be mentioned that the time taken for constructing the $\mathcal{ELH}$ description trees that is required by Algorithm 1 is excluded from those of the execution time reported in the sixth and seventh column.

	Overall	Min	Max	Avg
Time (ms)	256 306	< 1	3 013	3.512

Table 7. Time required by the system in order to construct the $\mathcal{ELH}$ description trees for the *c-sample*(C_i)

Table 7 reports the time taken by the system, which is obviously low. The information about all related factors of the $\mathcal{ELH}$ description trees is reported in Table 8.

Related Factor	Min	Max	Avg
Maximum branching factor	0	231	3.599
Average branching factor	0	4.142	0.787
Maximum label size	1	87	13.068
Average label size	0.667	87	7.538
Tree depth	0	10	0.992
Average branch depth	0	6.363	0.958
Number of nodes	1	917	9.506

Table 8. The minimum, maximum, and average values of each related factor of the $\mathcal{ELH}$ description tree construction

8 RELATED WORKS

In Table 4, we have compared the desirable properties of our method with those achieved in the state-of-the-art works. Apart from the work of [13] that supports mere the concept conjunction, it should be sufficient to remark that our work, together with that of [19], outperforms the others [7, 8, 9, 11, 14, 15]. This should make a comparison of the similarity degree obtained from the two methods sound promising. However, the similarity measure defined in [19] is merely a broad framework with no implementation detail (i.e., there is no actual algorithm suggested by the authors). This makes a comparison obviously impossible.

In [7], a semantic similarity measure for the DL $\mathcal{ALC}$ is proposed. The similarity is calculated based on a number of shared instances of two concepts. One drawback of this method is that it cannot be applied to an ontology without instances, e.g. Snomed ct. A similar work of the same idea is subsequently proposed in [8]. Extensions that support DL $\mathcal{ALE}$ and $\mathcal{ALN}$ are proposed in [9] and [11], respectively.

Apart from the DL-based similarity measures shown in Table 4, there is a semantic-based method suggested in [28], in which the effort for deducing certain relationships such as subsumption can be regarded as a distance between the two concepts in question. This was in contrast to the path distance approach that ignores such information. The limitation of this *reasoning effort* approach is that any pair of concepts out of the subsumption relation are always treated as totally dissimilar.

In [17], a new reasoning service for measuring a probabilistic degree, which specifies how likely one concept is subsumed by one another, is proposed. Despite its usefulness, one major drawback of the method is that it requires a probabilistic terminology. This makes it inapplicable to many ontologies which are usually modeled based on the W3C-recommended OWL.

Another category of similarity measure is a syntactic-based approach. Many of these works can be applied with concepts that mostly emphasize their hierarchy rather than meaning. These works are likely to have an inefficient measure as more complex concepts are defined. However, for an ontology in which a hierarchy of concepts is mandatory, these methods seem to be applicable to several specific scenarios. Herewith, we give a review on major related works.

The Rada distance introduced in [22] is a directed-graph-based similarity measure which represents concepts as a set of nodes. The method averages all possible paths between a pair of concepts and returns a numerical value, which indicates how similar the two concepts are, as a result.

In [23], a distance function called *Ontology Structure based Similarity* (OSS) is introduced. To measure the degree of similarity between concepts, the cost required for a transformation from one concept to one another is computed. The distance is measured based on the cost estimated in the early step. One disadvantage of this method is that it is variant to a number of descendants of the compared concepts. Therefore, a decreasing of the similarity as the deeper the concepts being compared.

Wu and Palmer [32] introduce a scaled metric that allows measuring similarity between a pair of words defined in WordNet. The method, called conceptual similarity, measures a depth of two given concepts in an *is-a* hierarchy. To enable similarity measure, a distance to a least common ancestor (LCA) of the two concepts is computed. The similarity measure, which in this case is a summation of the path lengths from the two concepts to the LCA, is returned as an output. Just like other distance-based similarity, the method has a disadvantage in which it totally relies on a skeleton of a taxonomy rather than semantic descriptions of terms.

A similar approach to that of [32], proposed by Ge and Qiu [12], adjusts the idea and applies it to an ontology. With the adjustment, in an initial step, different numerical weights are first assigned to different is-a edges and uniformly decreased as they are far away from the root. In the next step, all possible paths from the root to the two concepts are investigated. The similarity is calculated based on a summation of the weights of the shortest path.

Caviedes and Cimino [6] introduced a similarity measure called CDist. Similar to [32], the method is invented for a concept that relates to one another via the *is-a* relation. The degree of similarity, which in this case is a distance, is computed based on a number of nodes between the two concepts.

9 CONCLUSIONS AND FUTURE DEVELOPMENT

This paper introduces a new similarity measure between two concepts, w.r.t. an $\mathcal{ELH}$ ontology. It presents top-down and bottom-up algorithm and reports on extensive evaluation results of the implemented similarity reasoner on Snomed ct. Though two concepts are not in the subsumption relation, the measure is capable of informing their relationship based on the common and discriminant features. Our intensive experiments on Snomed ct show that the proposed non-standard reasoner is capable of revealing hidden knowledge that the standard one [27] cannot, and yet merely requires an inexpensive computation time by employing the bottom-up algorithm.

Apart from computability evaluation, we have also conducted a usability evaluation of the proposed measure through manual assessment by medical practitioner and specialists. The idea is to calculate the agreement between concept similarity rankings from the domain experts and such rankings from our implemented reasoner. Of course, completeness and quality of $\mathcal{O}_{\mathsf{snomed}}$ also have a great impact on the ranking results. Nevertheless, the study reveals a promising level of agreement between the system and human expert. For more details about the evaluation, we direct interested readers to our proceedings paper [16].

In [16], a naïve top-down algorithm has been implemented for use with a Web service application, for reporting. In this extended work, several improvements and addendums have been made which can be summarized in order. First, we introduce a more efficient bottom-up algorithm (see Algorithm 3), which is at least three times faster than the naïve algorithm. Second, after an inspection under scrutiny, it is uncovered that fixing the value of ν (in Equation (5)) for all roles in the original measure deems inappropriate due to the presence of roleGroup and some Snomed's modeling discipline (see Section 7.1). This results in a more general similarity measure for the DL $\mathcal{ELH}$. Additionally, proofs of desirable properties, extensive experimental results on Snomed ct, and efficiency analysis regarding the description trees shapes and properties are reported and discussed.

The proposed non-standard reasoning service is believed to be useful in real-world applications, in which concept descriptions may not be formed manually by domain experts but rather automatically from abundant data. For example, one could extract technical terms from a text and use them to create a concept description. This description may not be related via the subsumption relation to a reference concept in the ontology but could hold certain information facets pertinent to the ontology and user's interest. Another promising application of the $\mathcal{ELH}$ semantic similarity measure is visualization. Traditionally, concepts are visualized with equal distances. More intuitive visualization tools could employ the degree of similarity to determine the most appropriate placement of each concept. Classical ontology debugging merely looks at logical inconsistency or unintended subsumption relationship. Such a reasoning service cannot give any useful suggestion regarding two consistent concepts that are out of subsumption relation. With similarity measure, the ontology designers will have more information while authoring concept definitions. For instance, two seemingly different concepts "Hemorrhage into Bladder Lumen" and "Irrigation of Gingival Operculum" with an unusually high similarity degree may be flagged for in-depth inspection by the designers.

There are a few directions for future work. Firstly, it appears to be a natural next step to consider tractable extensions to a terminology with cyclic concepts definitions. Secondly, though regarded as a pre-processing step, the overall computation time for the $\mathcal{ELH}$ description tree reconstruction is high (about 2.6 minutes as reported in the experiments). We therefore aim at reducing such a computation time. One possible way to achieve this is through a representation of an entire TBox as a forest of inter-dependent $\mathcal{ELH}$ partial description trees. Thirdly, we aim at extending more reasoning services, especially for a non-standard instance checking service. We believe that this can be done in a similar manner, i.e. the calculation of the homomorphism degree. However, to carry out this, a small modification of the algorithms appears to be required. One approach is to transform an instance problem to a concept problem, and the other is to represent the ABox as a graph such that the proposed membership homomorphism measure could be applied. Lastly, we also aim at improving the proposed method to support a more expressive DL family, i.e. handling for negation, disjunction, etc.

Acknowledgements

This work was partially supported by the National Research University (NRU) project of Thailand Office for Higher Education Commission and by the Center of Excellence in Intelligent Informatics, Speech and Language Technology, and Service Innovation (CILS), Thammasat University.

REFERENCES

[1] Baader, F.: Terminological Cycles in a Description Logic with Existential Restrictions. In: Gottlob, G., Walsh, T. (Eds.): Proceedings of the 18th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 2003.

[2] BAADER, F.—BRANDT, S.—LUTZ, C.: Pushing the $\mathcal{EL}$ Envelope. Proceedings of the 19th International Conference on Artificial Intelligence (IJCAI '05), Morgan-Kaufmann Publishers, Edinburgh, UK, 2005.

[3] BAADER, F.—CALVANESE, D.—MCGUINNESS, D.—NARDI, D.—PATEL-SCHNEIDER, P. (Eds.): The Description Logic Handbook: Theory, Implementation and Applications. 2nd Ed. Cambridge University Press, 2007, doi: 10.1017/CBO9780511711787.

[4] BAADER, F.—LUTZ, C.—SUNTISRIVARAPORN, B.: CEL: A Polynomial-Time Reasoner for Life Science Ontologies. Proceedings of the Third International Joint Conference on Automated Reasoning (IJCAR '06), Springer-Verlag, Berlin, Heidelberg, 2006, doi: 10.1007/11814771_25.

[5] BAADER, F.—SUNTISRIVARAPORN, B.: Debugging SNOMED CT Using Axiom Pinpointing in the Description Logic $\mathcal{EL}^+$. Proceedings of the 3rd Knowledge Representation in Medicine Conference (KR-MED '08): Representing and Sharing Knowledge Using SNOMED. CEUR-WS, Vol. 410, 2008.

[6] CAVIEDES, J. E.—CIMINO, J. J.: Towards the Development of a Conceptual Distance Metric for the UMLs. Journal of Biomedical Informatics, Vol. 37, 2004, No. 2, pp. 77–85, doi: 10.1016/j.jbi.2004.02.001.

[7] D'AMATO, C.—FANIZZI, N.—ESPOSITO, F.: A Semantic Similarity Measure for Expressive Description Logics. Proceedings of Convegno Italiano di Logica Computazionale (CILC05), 2005.

[8] D'AMATO, C.—FANIZZI, N.—ESPOSITO, F.: A Dissimilarity Measure for $\mathcal{ALC}$ Concept Descriptions. Proceedings of the 2006 ACM Symposium on Applied Computing (SAC '06), ACM, New York, NY, USA, 2006, doi: 10.1145/1141277.1141677.

[9] D'AMATO, C.—STAAB, S.—FANIZZI, N.: On the Influence of Description Logics Ontologies on Conceptual Similarity. Proceedings of the 16th International Conference on Knowledge Engineering: Practice and Patterns (EKAW '08), Springer-Verlag, Berlin, Heidelberg, 2008.

[10] BAADER, F.—KÜSTERS, R.: Matching in Description Logics with Existential Restrictions. In: Cohn, A. G., Giunchiglia, F., Selman, B. (Eds.): Proceedings of the Seventh International Conference on Knowledge Representation and Reasoning (KR2000), Morgan Kaufmann Publishers, San Francisco, CA, 2000.

[11] FANIZZI, N.—D'AMATO, C.: A Similarity Measure for the ALN Description Logic. Proceedings of Convegno Italiano di Logica Computazionale (CILC06), 2006.

[12] GE, J.—QIU, Y.: Concept Similarity Matching Based on Semantic Distance. Proceedings of the 2008 Fourth International Conference on Semantics, Knowledge and Grid (SKG '08), IEEE Computer Society, Washington, DC, USA, 2008, doi: 10.1109/SKG.2008.24.

[13] JACCARD, P.: Étude Comparative de la Distribution Florale Dans une Portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles, Vol. 37, 1901, pp. 547–579.

[14] JANOWICZ, K.: SIM-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic $\mathcal{ALCNR}$ in Geographic Information Retrieval. In: Meersman, R., Tari, Z., Herrero, P. (Eds.): On the Move to Meaningful Internet Systems 2006: OTM

2006 Workshops. Springer Berlin Heidelberg, Lecture Notes in Computer Science, Vol. 4278, 2006, pp. 1681–1692, doi: 10.1007/11915072_74.

[15] Janowicz, K.—Wilkes, M.: SIM-DL$_A$: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-Concept to Inter-Instance Similarity. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (Eds.): The Semantic Web: Research and Applications. Springer, Lecture Notes in Computer Science, Vol. 5554, 2009, pp. 353–367.

[16] Jirathitikul, P.—Nithisansawadikul, S.—Tongphu, S.—Suntisrivaraporn, B.: A Similarity Measuring Service for Snomed-ct: Structural Analysis of Concepts in Ontology. 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2014, doi: 10.1109/ECTICon.2014.6839771.

[17] Klinov, P.: Pronto: A Non-Monotonic Probabilistic Description Logic Reasoner. Proceedings of European Semantic Web Conference (ESWC), 2008, doi: 10.1007/978-3-540-68234-9_66.

[18] Lehmann, J.—Haase, C.: Ideal Downward Refinement in the $\mathcal{EL}$ Description Logic. Technical Report, University of Leipzig, 2009.

[19] Lehmann, K.—Turhan, A.-Y.: A Framework for Semantic-Based Similarity Measures for $\mathcal{ELH}$-Concepts. In: del Cerro, L. F., Herzig, A., Mengin, J. (Eds.): JELIA. Springer, Lecture Notes in Computer Science, Vol. 7519, 2012.

[20] The Gene Ontology Cosortium: Creating the Gene Ontology Resource: Design and Implementation. Genome Research, Vol. 11, 2001, No. 8, pp. 1425–1433.

[21] Motik, B.—Grau, B. C.—Horrocks, I.—Wu, Z.—Fokoue, A.—Lutz, C.: OWL 2 Web Ontology Language Profiles. 2009.

[22] Rada, R.—Mili, H.—Bicknell, E.—Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, 1989, No. 1, pp. 17–30.

[23] Schickel-Zuber, V.—Faltings, B.: OSS: A Semantic Similarity Function Based on Hierarchical Ontologies. In: Veloso, M. M. (Ed.): International Joint Conference on Artificial Intelligence, 2007.

[24] Schulz, S.—Suntisrivaraporn, B.—Baader, F.: Snomed ct's Problem List: Ontologists' and Logicians' Therapy Suggestions. Studies in Health Technology and Informatics, Vol. 129, 2007, No. 1, pp. 802–806.

[25] Spackman, K. A.—Dionne, R.—Mays, E.—Weis, J.: Role Grouping as an Extension to the Description Logic of Ontylog Motivated by Concept Modeling in Snomed. Proceedings of the AMIA Symposium, 2002, pp. 712–716.

[26] Stearns, M.—Price, C.—Spackman, K.—Wang, A.: Snomed Clinical Terms: Overview of the Development Process and Project Status. Proceedings of the 2001 AMIA Annual Symposium, Hanley & Belfus, 2001.

[27] Suntisrivaraporn, B.: Polynomial Time Reasoning Support for Design and Maintenance of Large-Scale Biomedical Ontologies. Ph.D. Thesis. Dresden University of Technology, Institute for Theoretical Computer Science, 2009.

[28] Suntisrivaraporn, B.: Structural Distance Between $\mathcal{EL}^+$ Concepts. Multi-Disciplinary Trends in Artificial Intelligence (MIWAI 2011). Springer, Lecture Notes in Computer Science, Vol. 7080, 2011, pp. 100–111, doi: 10.1007/978-3-642-25725-4_9.

[29] Suntisrivaraporn, B.: A Similarity Measure for the Description Logic $\mathcal{EL}$ with Unfoldable Terminologies. International Conference on Intelligent Networking and Collaborative Systems (INCoS-13), 2013, doi: 10.1109/INCoS.2013.77.

[30] Tongphu, S.—Suntisrivaraporn, B.: On Desirable Properties of the Structural Subsumption-Based Similarity Measure. Joint International Semantic Technology Conference (JIST), Chiang Mai, Thailand, 2014.

[31] Tongphu, S.—Suntisrivaraporn, B.—Uyyanonvara, B.—Dailey, M. N.: Ontology-Based Object Recognition of Car Sides. 9th International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI), 2012, doi: 10.1109/ECTICon.2012.6254268.

[32] Wu, Z.—Palmer, M.: Verbs Semantics and Lexical Selection. Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 1994, pp. 133–138.

Suwan Tongphu received his B.Sc. in computer science from Mae Fah Luang University in 2007 and his M.Sc. in the same field from Asian Institute of Technology in 2009. Currently, he is working towards his Ph.D. at Sirindhorn International Institute of Technology (SIIT), Thammasat University. His research interests include semantic-based object detection, measuring similarity in Description Logics, and non-standard reasoning services.

Boontawee Suntisrivaraporn is Assistant Professor at the School of Information, Computer and Communication Technology (ICT), which belongs to Sirindhorn International Institute of Technology (SIIT) of Thammasat University; and Visiting Associate Professor at the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST). He received his M.Sc. in computer science and D.Eng. in the same field from Dresden University of Technology, Germany. His research interests include description logics, semantic web technologies, graph algorithms, and knowledge representation and reasoning. Dr. Suntisrivaraporn has also been serving as a tutorial chair and PC member for several relevant international conferences, such as ICAART, JIST and KICSS.

APPENDICES

Proof for Proposition 7.

1 $\Longrightarrow$ 2: To prove this, we need to show that for each $v \in V_D$, there exists $h(v) \in V_C$ such that $\mathsf{p\text{-}hd}(\cdot, \cdot) = 1$ and $\mathsf{e\text{-}set\text{-}hd}(\cdot, \cdot) = 1$ (only for those non-leaf nodes). Let d be the depth of $\mathcal{T}_D$. Since $C \sqsubseteq_{\mathcal{O}} D$, by Theorem 3 there exists a homomorphism from $\mathcal{T}_D$ to $\mathcal{T}_C$. For the induction base case where $d = 0$ and $C = P_1 \sqcap \ldots \sqcap P_m$, there exists a mapping from rt_D to rt_C such that $\ell_D(v) \subseteq \ell_C(h(v))$ (i.e. $\mathsf{hd} = \mathsf{p\text{-}hd} = 1$). For the induction step where $C = P_1 \sqcap \ldots \sqcap P_m \sqcap \exists r_1 C_1 \sqcap \ldots \sqcap \exists r_n C_n$ there exists a mapping from each v to $h(v)$ such that $\ell_D(v) \subseteq \ell_C(h(v))$ (i.e. $\mathsf{p\text{-}hd}(\cdot, \cdot) = 1$) and $\rho_D(v, w) \subseteq \rho_C(h(v), h(w))$ (i.e. $\mathsf{e\text{-}set\text{-}hd}(\cdot, \cdot) = 1$) where w and $h(w)$ are successors of v and $h(v)$, respectively. For the case where v is a leaf, this is similar to the base case (i.e. $\mathsf{p\text{-}hd}(\cdot, \cdot) = 1$).

2 $\Longrightarrow$ 1: By Definition 6, $\mathsf{hd}(\mathcal{T}_D, \mathcal{T}_C) = 1$ means $\mathsf{p\text{-}hd}(\mathcal{P}_D, \mathcal{P}_C) = 1$ and $\mathsf{e\text{-}set\text{-}hd}$ $(\mathcal{E}_D, \mathcal{E}_C) = 1$, therefore for each $P \in \mathcal{P}_D$ there exists $P \in \mathcal{P}_C$ (i.e. $\mathcal{P}_D \subseteq \mathcal{P}_C$) and for each $\exists r.D' \in \mathcal{E}_D$ there exists $\exists s.C' \in \mathcal{E}_C$ such that $\frac{|\mathcal{R}_r \cap \mathcal{R}_s|}{|\mathcal{R}_r|} = 1$ and $\mathsf{hd}(\mathcal{T}_{D'}, \mathcal{T}_{C'}) = 1$. The former implies $\mathcal{R}_r \subseteq \mathcal{R}_s$ which means that $s \sqsubseteq^* r$. The latter implies that $C' \sqsubseteq D'$. By Corollary 4, this means $C \sqsubseteq_{\mathcal{O}} D$.

Proof for Proposition 9. Let W, W' be sets of all nodes at level d in $\mathcal{T}$ and $\mathcal{T}'$, and let $w \in W$ and $w' \in W'$. To prove (1) and (2), we need to show that $(1 - \mu) \cdot \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}, \mathcal{E}')$ and $(1 - \mu) \cdot \mathsf{e\text{-}set\text{-}hd}(\mathcal{E}', \mathcal{E})$ are all zero for every mapping from w to w' and from w' to w. At the level d, for the first case, we have $\mu = 1$ and this implies the case. For the latter, we have $\mid \mathsf{children}(w) \mid = 0$. Since $\mathcal{E} = \emptyset$ and $\mathcal{E}' \neq \emptyset$, by Equation (4), we have $\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}', \mathcal{E}) = 0$.