A Similarity Measure for the Description Logic $\mathcal{EL}$ with Unfoldable Terminologies

Boontawee Suntisrivaraporn
School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
Email: sun@siit.tu.ac.th

***Abstract*—Description Logics (DLs) are a family of logic-based knowledge representation formalisms, which can be used to develop ontologies in a formally well-founded way. The standard reasoning service of subsumption has proved indispensable in ontology design and maintenance. This 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. This work presents a new notion of semantic similarity which stems from the known homomorphism-based structural subsumption algorithm. The proposed similarity measure computes a numerical degree of similarity between two $\mathcal{EL}$ concept descriptions despite not being in the subsumption relation.**

***Keywords*-similarity measure; description logic; semantic web ontology; concept matching**

I. Introduction

Description Logics (DLs) [2] 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 $\mathcal{EL}$ family [1], which are used in the design of large-scale medical ontologies such as SNOMED CT [11] and form one of the W3C-recommended tractable OWL profiles, OWL 2 EL [9]. One of the main advantages of employing a logic-based ontology language is that reasoning services can be used to derive implicit knowledge from the 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 of the $\mathcal{EL}$ 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 crisp responses, i.e. a positive response concluding that one concept is subsumed by the other, or a negative response concluding that they are not related that way. In virtually every domain, however, a concept may be more similar to certain concepts than others despite the fact that they are out of the subsumption relation. Consider, for instance, the concepts Grandfather, Father, Uncle and Mother with their natural definitions. It is not hard to be convinced that Mother is more similar to Father than to Grandfather; and that it is more similar to Grandfather than to Uncle. Obviously, subsumption alone is not enough to handle this matter since Mother is in *no* subsumption relationship with the other concepts. The aim of the present paper is to systematically and semantically define a measure for similarity between two $\mathcal{EL}$ concept descriptions w.r.t. a terminology.

The rest of the paper is organized in order. The background on the DL $\mathcal{EL}$, unfoldable TBoxes, and the structural subsumption algorithm is presented in the next section. Section III and IV introduce the notions of homomorphism likelihood and $\mathcal{EL}$ semantic similarity measure, respectively, and exemplify the introduced measure by means of a small yet prototypical medical ontology. Related works are discussed in Section V, and the last section gives some concluding remarks.

II. Background

In DLs, *concept descriptions* are inductively defined with the help of a set of *constructors*, starting with a set CN of *primitive concept names* and a set RN of *role names*. $\mathcal{EL}$ concept descriptions are formed using the constructors shown in the upper part of Table I. An $\mathcal{EL}$ *terminology* or *TBox* is a finite set of concept definitions, whose syntax is shown in the lower part of Table I. A TBox is called *unfoldable* if it contains at most one concept definition for each concept name in CN and does not contain cyclic dependencies. The set $\mathsf{CN}^{\mathsf{def}}$ of *defined concepts* are concept names that appear on the left hand side of a concept definition. Other concepts are called *primitive concepts*, denoted by $\mathsf{CN}^{\mathsf{pri}}$. 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. Primitive concept definitions are commonly found in realistic terminologies to define those concepts, of which only necessary conditions are known; see, e.g., the concept Pericardium and its primitive definition ω_1 in Figure 2. Such a primitive definition $B \sqsubseteq D$ can easily

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 definition	$B \sqsubseteq D$	$A^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
full definition	$B \equiv D$	$A^{\mathcal{I}} = D^{\mathcal{I}}$

Table I
SYNTAX AND SEMANTICS OF THE DESCRIPTION LOGIC $\mathcal{EL}$.

be transformed into a semantically equivalent full definitions $B \equiv X \sqcap D$ where X is a fresh concept name.

The semantics of $\mathcal{EL}$ is 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 I. An interpretation $\mathcal{I}$ is a *model* of a TBox $\mathcal{O}$ if, for each concept definition in $\mathcal{O}$, the conditions given in the semantics column of Table I are satisfied. The main inference problem for $\mathcal{EL}$ is the subsumption problem:

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

Provided that the TBox is unfoldable (i.e. acyclic and definitional), any $\mathcal{EL}$ concept can be expanded to an equivalent one that consists of only primitive concept names in $\mathsf{CN}^{\mathsf{pri}}$ by repeatedly replacing defined concepts by their definitions until no more defined concepts appear in the concept description. Given, for example, a stated definition of Grandfather:

$$\mathsf{Grandfather} \equiv \mathsf{Man} \sqcap \exists \mathsf{child}.\mathsf{Parent}$$

By replacing the defined concept Parent with its description $\mathsf{Person} \sqcap \exists \mathsf{child}.\mathsf{Person}$, and Man with $\mathsf{Male} \sqcap \mathsf{Person}$, the description can be expanded to:

$$\mathsf{Male} \sqcap \mathsf{Person} \sqcap \exists \mathsf{child}.(\mathsf{Person} \sqcap \exists \mathsf{child}.\mathsf{Person}) \qquad (1)$$

where $\mathsf{Person}, \mathsf{Male} \in \mathsf{CN}^{\mathsf{pri}}$. We denote by $\hat{C}$ the expanded equivalence of the concept description C.

Henceforth, we assume without loss of generality that an $\mathcal{EL}$ concept C is of the following form:

$$P_1 \sqcap \cdots \sqcap P_k \sqcap \exists r_1.C_1 \sqcap \cdots \sqcap \exists r_\ell.C_\ell \qquad (2)$$

where $P_i \in \mathsf{CN}^{\mathsf{pri}}, r_j \in \mathsf{RN}$, and C_j are concept descriptions, for $1 \leq i \leq k$ and $1 \leq j \leq \ell$. For convenience, we denote by $\mathcal{P}_C$ and $\mathcal{E}_C$ the set of top-level primitive concepts $\{P_1, \ldots, P_k\}$ and the set of top-level existential restrictions $\{\exists r_1.C_1, \ldots, \exists r_\ell.C_\ell\}$, respectively.

In [4], [3], a characterization of subsumption in $\mathcal{EL}$ w.r.t. an unfoldable TBox using homomorphism has been proposed. Instead of considering concept descriptions directly, the characterization considers so-called $\mathcal{EL}$ description trees that structurally correspond to the $\mathcal{EL}$ concept descriptions. In essence, the root v of the $\mathcal{EL}$ description tree $\mathcal{T}$ for the concept description in Formula 2 has $\{P_1, \ldots, P_k\}$ as its label, and has ℓ outgoing edges, each labeled with R_j to a vertex v_j, for $1 \leq j \leq \ell$. Then, the subtree $\mathcal{T}\,|_{v_j}$ with the root v_j is defined inductively based on C_j. The subsumption is then characterized by means of an existence of a homomorphism in the reverse direction.

Theorem 1 ([4], [3]). *Let C, D be $\mathcal{EL}$ concept descriptions. Then, $C \sqsubseteq D$ iff there exists a homomorphism $h : \mathcal{T}_D \to \mathcal{T}_C$ which maps the root of $\mathcal{T}_D$ to the root of $\mathcal{T}_C$.*

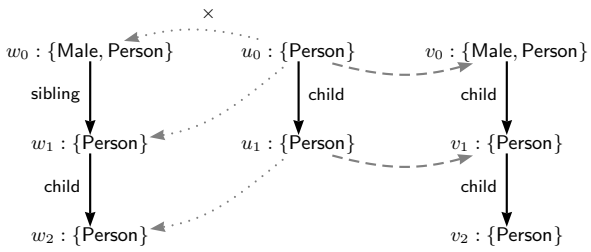

Figure 1. A homomorphism h (dashed arrows) that maps the root of $\mathcal{T}_{\mathsf{Parent}}$ to the root of $\mathcal{T}_{\mathsf{Grandparent}}$; a failed attempt to identify a homomorphism (dotted arrows) that maps the root of $\mathcal{T}_{\mathsf{Parent}}$ to the root of $\mathcal{T}_{\mathsf{Uncle}}$.

Consider the aforementioned descriptions for Parent and Grandfather, and the following description for Uncle:

$$\mathsf{Man} \sqcap \mathsf{Person} \sqcap \exists \mathsf{sibling}.(\mathsf{Person} \sqcap \exists \mathsf{child}.\mathsf{Person})$$

Figure 1 depicts the $\mathcal{EL}$ description trees $\mathcal{T}_{\mathsf{Parent}}$ (center), $\mathcal{T}_{\mathsf{Grandfather}}$ (right), and $\mathcal{T}_{\mathsf{Uncle}}$ (left), and shows a homomorphism h as the dashed arrows that maps the root u_0 of $\mathcal{T}_{\mathsf{Parent}}$ to the root v_0 of $\mathcal{T}_{\mathsf{Grandfather}}$. It also shows a failed attempt to obtain a homomorphism as the dotted arrows from $\mathcal{T}_{\mathsf{Parent}}$ to $\mathcal{T}_{\mathsf{Uncle}}$.

By Theorem 1, it is then ensured that $\mathsf{Grandfather} \sqsubseteq_{\mathcal{O}} \mathsf{Parent}$ and that $\mathsf{Uncle} \not\sqsubseteq_{\mathcal{O}} \mathsf{Parent}$. Though sharing some common feature, the classical reasoning of subsumption does not suffice to tell how similar they are.

Our similarity measure is based on this structural characterization. Instead of merely giving either positive or negative result between two descriptions, rather the similarity measure provides a numerical result 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 measure is 1, then the two concepts are equivalent.

III. HOMOMORPHISM LIKELIHOOD

Theorem 1 suggests that an existence of a homomorphism between $\mathcal{EL}$ description trees implies a subsumption relationship between the corresponding concept descriptions. We extend this idea also to the case where *no* such homomorphism exists but there is some likelihood.

Let C, D be $\mathcal{EL}$ concept descriptions, $\mathcal{P}_C, \mathcal{P}_D, \mathcal{E}_C, \mathcal{E}_D$ be as defined in the previous section, and $\mathcal{T}_C, \mathcal{T}_D$ be the corresponding $\mathcal{EL}$ description trees. Then, the likelihood of having a homomorphism from $\mathcal{T}_D$ to $\mathcal{T}_C$ is defined as follows:

Definition (homomorphism likelihood) Let $\mathbf{T}^{\mathcal{EL}}$ be the set of all $\mathcal{EL}$ description trees. The *homomorphism likelihood function* $\mathsf{hl} : \mathbf{T}^{\mathcal{EL}} \times \mathbf{T}^{\mathcal{EL}} \rightarrow [0, 1]$ is inductively defined as follows:

$$\mathsf{hl}(\mathcal{T}_D, \mathcal{T}_C) := \mu \cdot \mathsf{p\text{-}hl}(\mathcal{P}_D, \mathcal{P}_C) + (1-\mu) \cdot \mathsf{e\text{-}hl}(\mathcal{E}_D, \mathcal{E}_C), \quad (3)$$

where $0 < \mu < 1$;

$$\mathsf{p\text{-}hl}(\mathcal{P}_D, \mathcal{P}_C) := \begin{cases} 1 & \text{if } \mathcal{P}_D = \emptyset \\ \frac{|\mathcal{P}_D \cap \mathcal{P}_C|}{|\mathcal{P}_D|} & \text{otherwise,} \end{cases} \quad (4)$$

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

$$\mathsf{e\text{-}hl}(\mathcal{E}_D, \mathcal{E}_C) := \sum_{\epsilon_i \in \mathcal{E}_D} \frac{max\{\mathsf{e\text{-}hl}(\epsilon_i, \epsilon_j) : \epsilon_j \in \mathcal{E}_C\}}{\mid \mathcal{E}_D \mid}, \quad (5)$$

where ϵ_i, ϵ_j are existential restrictions; and

$$\mathsf{e\text{-}hl}(\exists r.X, \exists s.Y) := \begin{cases} 0 & \text{if } r \neq s \\ \nu + (1-\nu) \cdot \mathsf{hl}(\mathcal{T}_X, \mathcal{T}_Y) & \text{if } r = s, \end{cases} \quad (6)$$

where $0 \leq \nu < 1$.

Intuitively, the homomorphism likelihood Formula 3 is defined as the weighted sum of the likelihood of the label set inclusion (p-hl) and the likelihood of the edge condition matching (e-hl). Formula 4 calculates the proportion of the matched primitive concepts to all the primitive concepts in the top level. Formula 6 measures the likelihood of an edge mapping in a potential homomorphism. If the edge-labeling roles are the same, then there is some likelihood; but the successors' labels and structures have yet to be checked. This is done recursively by calling the function $\mathsf{hl}(\mathcal{T}_X, \mathcal{T}_Y)$. The values computed in Formula 6 collectively are used to determine the likelihood of the edge condition matching. Formula 5 calculates the maximum likelihood for each edge in $\mathcal{E}_D$ and returns the average thereof.

The weight μ in Formula 3 determines how important the primitive concepts are to be considered for similarity measure. It is recommended to set $\mu = \frac{|\mathcal{P}_D|}{|\mathcal{P}_D \cup \mathcal{E}_D|}$, i.e. the proportion of the primitive concepts to all the terms in the top level. For the special case where $\mathcal{P}_D = \mathcal{E}_D = \emptyset$, the value of μ is irrelevant as $\mathcal{T}_\top$ is the smallest $\mathcal{EL}$ description tree and $\mathsf{hl}(\mathcal{T}_\top, \mathcal{T}_C) = 1$ for all concepts C.

ω_1	Pericardium	$\sqsubseteq$	Tissue $\sqcap \exists$part.Heart
ω_2	Endocardium	$\sqsubseteq$	Tissue $\sqcap \exists$part.Heart
ω_3	Appendicitis	$\equiv$	Inflammation $\sqcap \exists$loc.Appendix
ω_4	Pericarditis	$\equiv$	Inflammation $\sqcap \exists$loc.Pericardium
ω_5	Endocarditis	$\equiv$	Inflammation $\sqcap \exists$loc.Endocardium
ω_6	Inflammation	$\sqsubseteq$	Disease
ω_7	HeartDisease	$\equiv$	Disease $\sqcap \exists$loc.$\exists$part.Heart

Figure 2. An example $\mathcal{EL}$ unfoldable terminology $\mathcal{O}_{\mathsf{med}}$.

The value of ν in Formula 6 determines how important the unqualified existential information, i.e. considering merely roles in an existential restriction, should be considered for similarity measure. For instance, $\exists$child.Male and $\exists$child.Female for dissimilar nested concepts Male and Female should not be regarded as entirely dissimilar themselves. If ν is assigned the values $0.2, 0.3, 0.4$, then e-hl($\exists$child.Male, $\exists$child.Female) is $0.2, 0.3, 0.4$, respectively. Providing more axiomatic information in the unfoldable TBox like Male $\sqsubseteq$ Gender and Female $\sqsubseteq$ Gender, the e-hl figures would be $0.6, 0.65, 0.7$, respectively.

To better understand the notion of homomorphism likelihood, consider a medical ontology $\mathcal{O}_{\mathsf{med}}$ in Figure 2. By introducing fresh concept names X, Y, Z, the primitive definitions ω_1, ω_2 and ω_6 can be transformed to the following full definitions:

ω'_1	Pericardium	$\equiv$	$X \sqcap$ Tissue $\sqcap \exists$part.Heart
ω'_2	Endocardium	$\equiv$	$Y \sqcap$ Tissue $\sqcap \exists$part.Heart
ω'_6	Inflammation	$\equiv$	$Z \sqcap$ Disease

Let $\mathcal{O}'_{\mathsf{med}}$ be the unfoldable TBox obtained from $\mathcal{O}_{\mathsf{med}}$ by replacing ω_1, ω_2 and ω_6, respectively, by their equivalent ω'_1, ω'_2 and ω'_6.

Example Consider the defined concepts HeartDisease and Pericarditis in $\mathcal{O}'_{\mathsf{med}}$; and their expanded descriptions as follows:

$$\mathsf{Disease} \sqcap \exists \mathsf{loc}.(\exists \mathsf{part.Heart}) \quad (7)$$

$$Z \sqcap \mathsf{Disease} \sqcap \exists \mathsf{loc}.(X \sqcap \mathsf{Tissue} \sqcap \exists \mathsf{part.Heart}) \quad (8)$$

Using $\nu = 0.4$, the homomorphism likelihood from

$\mathsf{hl}(\downarrow, \rightarrow)$	Pdm	Edm	Ats	Pts	Ets	Inf	Hds
Pericardium	1.0	0.67	0	0	0	0	0
Endocardium	0.67	1.0	0	0	0	0	0
Appendicitis	0	0	1.0	0.8	0.8	0.67	0.47
Pericarditis	0	0	0.8	1.0	0.93	0.67	0.53
Endocarditis	0	0	0.8	0.93	1.0	0.67	0.53
Inflammation	0	0	1.0	1.0	1.0	1.0	0.5
HeartDisease	0	0	0.70	1.0	1.0	0.5	1.0

Table II
HOMOMORPHISM LIKELIHOOD AMONG DEFINED CONCEPTS IN $\mathcal{O}_{\mathsf{MED}}$.

HeartDisease to Pericarditis can be computed as follows:[1]

$$\begin{aligned}
\mathsf{hl}(\mathcal{T}_{\mathsf{Hds}}, \mathcal{T}_{\mathsf{Pts}}) &:= \tfrac{1}{2}\mathsf{p\text{-}hl}(\mathcal{P}_{\mathsf{Hds}}, \mathcal{P}_{\mathsf{Pts}}) + \tfrac{1}{2}\mathsf{e\text{-}hl}(\mathcal{E}_{\mathsf{Hds}}, \mathcal{E}_{\mathsf{Pts}}) \\
&:= \tfrac{1}{2}[\tfrac{1}{1}] + \tfrac{1}{2}\mathsf{e\text{-}hl}(\epsilon_i, \epsilon_j) \\
\text{// with } \epsilon_i = \exists \mathsf{l}.\exists \mathsf{p}.\mathsf{H} &\text{ and } \epsilon_j = \exists \mathsf{l}.(X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}) \\
&:= \tfrac{1}{2}[\tfrac{1}{1}] + \tfrac{1}{2}[\tfrac{2}{5} + \tfrac{3}{5}\mathsf{hl}(\mathcal{T}_{\exists \mathsf{p}.\mathsf{H}}, \mathcal{T}_{X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}})] \\
&:= \tfrac{1}{2}[\tfrac{1}{1}] + \tfrac{1}{2}[\tfrac{2}{5} + \tfrac{3}{5}\mathsf{e\text{-}hl}(\exists \mathsf{p}.\mathsf{H}, \exists \mathsf{p}.\mathsf{H})] \\
&:= \tfrac{1}{2}[\tfrac{1}{1}] + \tfrac{1}{2}[\tfrac{2}{5} + \tfrac{3}{5}\{\tfrac{2}{5} + \tfrac{3}{5}\mathsf{hl}(\mathcal{T}_{\mathsf{H}}, \mathcal{T}_{\mathsf{H}})\}] \\
&:= \tfrac{1}{2}[\tfrac{1}{1}] + \tfrac{1}{2}[\tfrac{2}{5} + \tfrac{3}{5}\{\tfrac{2}{5} + \tfrac{3}{5} \cdot 1\}] \\
&:= 1
\end{aligned}$$

Intuitively, the homomorphism likelihood of 1 means there is a homomorphism; see the dashed arrows in Figure 3. The reverse direction can be computed as follows:

$$\begin{aligned}
\mathsf{hl}(\mathcal{T}_{\mathsf{Pts}}, \mathcal{T}_{\mathsf{Hds}}) &:= \tfrac{2}{3}\mathsf{p\text{-}hl}(\mathcal{P}_{\mathsf{Pts}}, \mathcal{P}_{\mathsf{Hds}}) + \tfrac{1}{3}\mathsf{e\text{-}hl}(\mathcal{E}_{\mathsf{Pts}}, \mathcal{E}_{\mathsf{Hds}}) \\
&:= \tfrac{2}{3}[\tfrac{1}{2}] + \tfrac{1}{3}\mathsf{e\text{-}hl}(\epsilon_i, \epsilon_j) \\
\text{// with } \epsilon_i = \exists \mathsf{l}.(X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}) &\text{ and } \epsilon_j = \exists \mathsf{l}.\exists \mathsf{p}.\mathsf{H} \\
&:= \tfrac{2}{3}[\tfrac{1}{2}] + \tfrac{1}{3}[\tfrac{2}{5} + \tfrac{3}{5}\mathsf{hl}(\mathcal{T}_{X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}}, \mathcal{T}_{\exists \mathsf{p}.\mathsf{H}})] \\
\text{// where } \mathsf{hl}(\mathcal{T}_{X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}}, \mathcal{T}_{\exists \mathsf{p}.\mathsf{H}}) &\text{ yields } \tfrac{1}{3}\text{; see below} \\
&:= \tfrac{2}{3}[\tfrac{1}{2}] + \tfrac{1}{3}[\tfrac{2}{5} + \tfrac{3}{5}\{\tfrac{1}{3}\}] \\
&:= \tfrac{8}{15} = 0.53
\end{aligned}$$

The computation for the sub-descriptions, corresponding to v_1 and u_1 in Figure 3, is as follows:

$$\begin{aligned}
&\mathsf{hl}(\mathcal{T}_{X \sqcap \mathsf{T} \sqcap \exists \mathsf{p}.\mathsf{H}}, \mathcal{T}_{\exists \mathsf{p}.\mathsf{H}}) \\
&:= \tfrac{2}{3}\mathsf{p\text{-}hl}(\{X, \mathsf{T}\}, \emptyset) + \tfrac{1}{3}\mathsf{e\text{-}hl}(\exists \mathsf{p}.\mathsf{H}, \exists \mathsf{p}.\mathsf{H}) \\
&:= \tfrac{2}{3}[\tfrac{0}{2}] + \tfrac{1}{3}[\tfrac{2}{5} + \tfrac{3}{5}\mathsf{hl}(\mathcal{T}_{\mathsf{H}}, \mathcal{T}_{\mathsf{H}})] \\
&:= \tfrac{2}{3}[\tfrac{0}{2}] + \tfrac{1}{3}[\tfrac{2}{5} + \tfrac{3}{5} \cdot 1] \\
&:= \tfrac{1}{3}
\end{aligned}$$

Hence, the likelihood of having a homomorphism from HeartDisease to Pericarditis is 1, and that for the opposite direction is 0.53. The hl values for other pairs can be obtained in an analogous manner and are shown in Table II.

Using a proof by induction, together with Theorem 1 [4], [3], it is not difficult to obtain the correspondence between the homomorphism likelihood and subsumption.

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

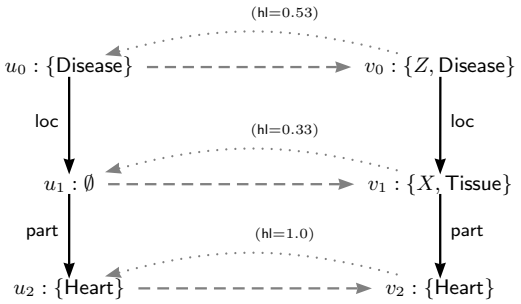

Figure 3. A homomorphism mapping the root of $\mathcal{T}_{\mathsf{HeartDisease}}$ to the root of $\mathcal{T}_{\mathsf{Pericarditis}}$ a failed attempt to identify (dashed arrows); a homomorphism likelihood that could map the root of $\mathcal{T}_{\mathsf{Pericarditis}}$ to the root of $\mathcal{T}_{\mathsf{HeartDisease}}$ (dotted arrows).

Proposition 2. *Let C, D be $\mathcal{EL}$ concept descriptions, and $\mathcal{O}$ an $\mathcal{EL}$ unfoldable TBox. Then, the following are equivalent:*

- $C \sqsubseteq_{\mathcal{O}} D$
- $\mathsf{hl}(\mathcal{T}_{\hat{D}}, \mathcal{T}_{\hat{C}}) = 1$,

where $\hat{X}$ is the equivalent expanded concept description from X w.r.t. $\mathcal{O}$, and $\mathcal{T}_{\hat{X}}$ is its corresponding $\mathcal{EL}$ description tree, with $X \in \{C, D\}$.

In fact, the closer the $\mathsf{hl}(\mathcal{T}_{\hat{D}}, \mathcal{T}_{\hat{C}})$ value is to 1, the more likely the corresponding subsumption may hold. Putting more simply, the label and edge constraints in $\mathcal{T}_{\hat{D}}$ can likely be simulated by those in $\mathcal{T}_{\hat{C}}$.

IV. $\mathcal{EL}$ SEMANTIC SIMILARITY

The homomorphism likelihood 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 the previous section, the direction of the homomorphism likelihood matters, viz., $\mathsf{hl}(\mathcal{T}_{\mathsf{Pts}}, \mathcal{T}_{\mathsf{Hds}}) = 0.53$, whereas $\mathsf{hl}(\mathcal{T}_{\mathsf{Hds}}, \mathcal{T}_{\mathsf{Pts}}) = 1$. Since both directions constitute the likelihood of the two concepts being equivalent (i.e. highest degree of similarity), our similarity measure for $\mathcal{EL}$ concept descriptions is defined by means of these values.

Definition ($\mathcal{EL}$ similarity degree) Let C, D be $\mathcal{EL}$ concept descriptions, and $\mathcal{O}$ an $\mathcal{EL}$ unfoldable TBox. The *degree of similarity between C and D*, in symbols $\mathsf{sim}(C, D)$, is defined as:

$$\mathsf{sim}(C, D) := \frac{\mathsf{hl}(\mathcal{T}_{\hat{C}}, \mathcal{T}_{\hat{D}}) + \mathsf{hl}(\mathcal{T}_{\hat{D}}, \mathcal{T}_{\hat{C}})}{2}, \tag{9}$$

where $\hat{X}$ is the equivalent expanded concept description from X w.r.t. $\mathcal{O}$, and $\mathcal{T}_{\hat{X}}$ is its corresponding $\mathcal{EL}$ description tree, with $X \in \{C, D\}$.

Intuitively, the degree of similarity between two concepts is the average of the likelihood of having homomorphisms in both directions, thus $\mathsf{sim}(C, D) = \mathsf{sim}(D, C)$ as required.

hl($\downarrow, \rightarrow$)	Pdm	Edm	Ats	Pts	Ets	Inf	Hds
Pericardium	1.0	0.67	0	0	0	0	0
Endocardium	-	1.0	0	0	0	0	0
Appendicitis	-	-	1.0	0.80	0.80	0.84	0.59
Pericarditis	-	-	-	1.0	0.93	0.84	0.77
Endocarditis	-	-	-	-	1.0	0.84	0.77
Inflammation	-	-	-	-	-	1.0	0.50
HeartDisease	-	-	-	-	-	-	1.0

Table III
SIMILARITY DEGREE AMONG DEFINED CONCEPTS IN $\mathcal{O}_{\mathsf{MED}}$.

Proposition 3. *Let C, D be $\mathcal{EL}$ concept descriptions, and $\mathcal{O}$ an $\mathcal{EL}$ unfoldable TBox. Then, the following are equivalent:*

- $C \equiv_{\mathcal{O}} D$
- $\mathsf{sim}(C, D) = \mathsf{sim}(D, C) = 1$,

Note that one could adopt an alternative definition, e.g., based on the multiplication $\mathsf{sim}^{mult}(C, D)$ or the root mean square $\mathsf{sim}^{rms}(C, D)$. Alas, these would give rather unsatisfactory values for the extreme cases such as the concepts A and $\top$, where $\mathsf{sim}^{mult}(A, \top) = 0$ and $\mathsf{sim}^{rms}(A, \top) = 0.707$. 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 likelihood values in Table II, the degrees of similarity among the defined concepts in the example ontology $\mathcal{O}_{\mathsf{med}}$ can be obtained; see Table III. Observe that there are two mutually exclusive clusters of similar concepts $\{\mathsf{Pdm}, \mathsf{Edm}\}$ and $\{\mathsf{Ats}, \mathsf{Pts}, \mathsf{Ets}, \mathsf{Inf}, \mathsf{Hds}\}$, where concepts from the same clusters are relatively similar (i.e. $\mathsf{sim} \geq 0.5$) and any from the different clusters are totally dissimilar (i.e. $\mathsf{sim} = 0$). This observation directly matches separate hierarchical structures in the classification results (see Figure 4). Note also that, though not included in Table II and III, the similarity involving primitive concepts like Heart, Tissue and Disease 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 $\top$.

Figure 4 highlights on Pericarditis to exemplify a possible utility of the $\mathcal{EL}$ semantic similarity. On top of the known subsumption relationships, similarity relationships together with their numerical degree can be displayed. It may appear more natural also to place more similar concepts closer to each other in a visualization tool.

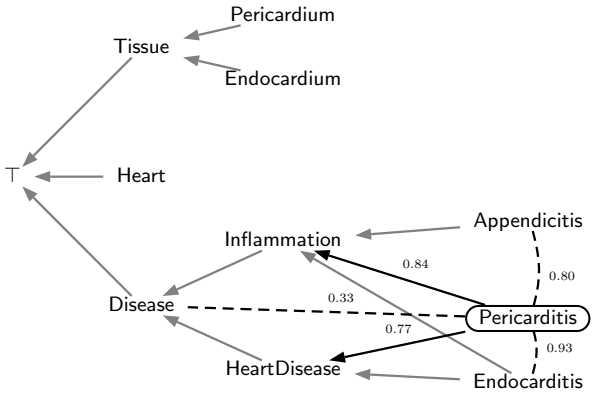

Figure 4. The classification hierarchy of the example medical TBox $\mathcal{O}_{\mathsf{med}}$ (bold arrows) augmented with the similarity relationships revolving the concept Pericarditis.

V. RELATED WORKS

There have been a good number of works on concept similarity in the literature which vary both in terms of algorithmic approaches and representation formalisms.

The *path distance* approach measures the distance between concepts in the pre-computed hierarchical structures of an ontology [6], [7]. Here, the definitions and constraints of concept definitions in the ontology are not used. Merely the placement of concepts in the hierarchy is relevant.

Tversky introduced *feature matching* [13] which suggests that both common and discriminant features between two concepts or concept instances be considered for the computation of a semantic similarity. Although the focus in that paper is on similarity between objects in geometric and dimensional spaces, the idea is so general that it can also be applied to conceptual representation.

It has been suggested in [12] that the effort for deducing certain relationship such as subsumption can be regarded as a distance between the 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.

Another similarity measure for concepts within a hierarchy is defined in terms of variation of the *information content* depicted by concepts of interest and the one depicted by their common parent concept [10].

Fanizzi and d'Amato has proposed a semantic similarity measure specific for the EL $\mathcal{ALN}$ [5]. This measure is based on the structural subsumption algorithm for the logic and computes commonality and distinctness between two $\mathcal{ALN}$ concept descriptions by resorting to counting named individuals (instances) in the knowledge base. One potential disadvantage of this measure is that it cannot be applied to an ontology without instances, for example, SNOMED CT.

In [8], the authors introduced two kernel functions for $\mathcal{EL}^{++}$ (i.e. an extension of the DL $\mathcal{EL}$) concepts which could be used for measuring similarity between concept descriptions. This approach stems from the *kernel methods for pattern analysis*. Unfortunately, neither details nor examples were given in the paper as to how numeric values can actually be computed. Besides, the canonical form, together

with its transformation rules, appears to contain a flaw since a concept may not be expandable w.r.t. expressive axioms allowed in DL $\mathcal{EL}^{++}$.

Our proposed measure is similar to that in [5], but the homomorphism-based structural subsumption algorithm for the DL $\mathcal{EL}$ is considered. The introduced notion of *homomorphism likelihood* does away with counting (named) individuals which allows our similarity measure to be applicable to various ontologies without the need to populate all the concepts with instances.

VI. Discussions and Future Development

This paper is the first attempt of extending the $\mathcal{EL}$ structural subsumption algorithm to calculate the degree of similarity between two concept descriptions. Though the concepts in question are not in the subsumption relation, the measure is capable of informing their relationship based on the common and discriminant features.

This non-standard reasoning service is believed to be useful in real-world applications, in which concept descriptions may be formed not by the domain expert but rather by the 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 $\mathcal{EL}$ 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. Besides, additional similarity links could be added to the graph as illustrated by the dashed edges in Figure 4.

There are a few directions for future work. Firstly, it appears to be a natural next step to consider tractable extensions to $\mathcal{EL}$, especially with role inclusions and terminological cycles. Secondly, we aim at carrying out an efficient implementation of this similarity measure algorithm. Obviously, one needs to ponder on how to efficiently calculate the homomorphism likelihood values in the bottom-up fashion, in order to avoid the possible exponential matching. Finally, it would be interesting to explore the $\mathcal{EL}$ semantic similarity among concepts in SNOMED CT and to compare the usefulness and meaningfulness of different measures w.r.t. this medical ontology.

Acknowledgment

This work is partially supported by the National Research University (NRU) project of Thailand Office for Higher Education Commission.

References

[1] F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proceedings of the 19th International Conference on Artificial Intelligence (IJCAI'05)*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, second edition, 2007.

[3] Franz Baader. Terminological cycles in a description logic with existential restrictions. In Georg Gottlob and Toby Walsh, editors, *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 325–330. Morgan Kaufmann, 2003.

[4] F. Baader and R. Küsters. Matching in description logics with existential restrictions. In A.G. Cohn, F. Giunchiglia, and B. Selman, editors, *Proceedings of the Seventh International Conference on Knowledge Representation and Reasoning (KR2000)*, pages 261–272, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

[5] N. Fanizzi and C. d'Amato. A similarity measure for the $\mathcal{ALN}$ description logic. In *In Proceedings of the Italian Conference on Computational Logic (CILC 2006)*, pages 26–27, 2006.

[6] J. Ge and Y. Qiu. Concept similarity matching based on semantic distance. In Georg Gottlob and Toby Walsh, editors, *Proceedings of the Forth International Conference on Semantics, Knowledge and Grid (SKG 2008)*, pages 380–383. Morgan Kaufmann, 2003.

[7] F. Giunchiglia, M. Yatskevich, and P. Shvaiko. Semantic matching: Algorithms and implementation. *Journal of Data Semantics*, 9:1–38, 2007.

[8] L. Jozefowski, A. Lawrynowicz, J. Jozefowska, J. Potoniec, and T. Lukaszewski. Kernels for $\mathcal{EL}^{++}$ description logic concepts. In *In Proceedings of the 21st International Conference on Inductive Logic Programming (ILP 2011)*, 2011.

[9] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 web ontology language profiles. http://www.w3.org/TR/owl2-profiles/ *(last access on 2 June 2011)*, 2009.

[10] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[11] M.Q. Stearns, C. Price, K.A. Spackman, and A.Y. Wang. SNOMED clinical terms: Overview of the development process and project status. In *Proceedings of the 2001 AMIA Annual Symposium*, pages 662–666. Hanley&Belfus, 2001.

[12] B. Suntisrivaraporn. Structural distance between $\mathcal{EL}^{+}$ concepts. In *Proceedings of the 5th Multi-disciplinary International Workshop in Artificial Intelligence (MIWAI 2011)*, LNAI, pages 100–111, Hyderabad, India, 2011. Springer.

[13] A. Tversky. Features of similarity. *Psychological Reviews*, 84(4):327–352, 1977.