Fakultät für Informatik
Institut für Theoretische Informatik
Lehrstuhl für Automatentheorie

A Framework for Semantic Invariant Similarity Measures for $\mathcal{ELH}$ Concept Descriptions

Diplomarbeit
zur Erlangung des akademischen Grades

Diplom-Informatiker

eingereicht von: Karsten Lehmann

eingereicht am: 07.02.2012

Betreuerin: Dr.-Ing. Anni-Yasmin Turhan

Contents

1 Introduction

The increasing usage of modern technology provides an increasing amount of data. Therefore, knowledge representation and reasoning are gaining increasing attention. Areas like the Semantic Web and scientific research are interested in possibilities for high-level descriptions and the ability to find implicit knowledge and consequences out of their data. With the number of information resources increasing, it becomes necessary to develop tools which support the automated connection and/or the interaction with this resources. In order to do that, similarities between the resources need to be identified. This creates the need for *similarity measures.* In this work we focus on similarity for description logics, a family of knowledge representation languages.

Basically, a similarity measure is a function mapping two objects to a value between 0 and 1. The higher the value the higher the similarity. Zero is interpreted as 'totally dissimilar' and one as 'totally similar'. However, simply mapping pairs of concepts to a value between 0 and 1 is not enough to ensure that the measure is useful. For example the function mapping everything to zero would have no application. The overall goal for similarity measures is to reproduce the intuition of a human expert. Hence the standard approach to find and evaluate a measure is to have a small set of test data and develop a measure where the results are matching the results of an (several) human expert(s). The disadvantage of this approach is that the overall behaviour of the measures is unknown. To deal with this problem, formal and measure-independent statements, called properties, are used. The independence ensures that the statement says something about the general behaviour, whereas the formal description allows to prove it. One source of properties for similarity measures are *metric spaces.* In this thesis, we also present several new properties which express our expectation of how a similarity measure should behave in general.

In relevant literature, one can find two authors who introduce similarity measure for description logics. D'Amato et.al presented four [dFE06, dFE05, FD06, dSF08] and Janowicz et.al presented two measures [JW09, Jan06]. For both authors, it is important that a measure does not depend solely on the syntax of the concepts. Only the semantics should be of importance. This is called *semantic* and/or *equivalence invariance.* For d'Amato it is also of importance that two concepts are totally similar if and only if they are equivalent. We call this property *equivalence closed.* We found out that Janowicz's two measures are not equivalence invariant, but he claims they are. Additionally, we detected that one measure from d'Amato [dFE06] is not equivalence invariant either, but is claimed to be and all four measures are not equivalence closed, despite claim. We also investigated if the measures fulfil the other properties. In some cases we are able to point out general unintuitive behaviour of a measure using the counterexample we found to disprove a property.

The similarity measure *simi*, presented in this thesis, is a measure which is equivalence invariant, equivalence closed and it also fulfils other properties we defined. Some parts of *simi* are flexible and introduced as parameter to allow *simi* to be “tuned“. Hence we call *simi* a framework. The parameters are designed such that the specific choice of a parameter does not influence the properties of *simi*.

This work is structured as follows. Chapter 2 presents the basic notations and definitions of description logics and triangular norms. In Chapter 3 we introduce a formal definition of similarity measures and we present and motivate all properties relevant in this thesis. Chapter 4 contains the analysis of the properties and behaviour of eight description-logic-similarity measures. In Chapter 5 we present our measure *simi* and Chapter 6 contains the conclusion and a sketch of open problems.

2 Preliminaries

This chapter introduces Description Logics and triangular norms and conorms which are used by *simi*.

2.1 Syntax and Semantics of Description Logics

Here we present several description logics. Also, our measure *simi* (Chapter 5) is defined for the description logic $\mathcal{ELH}$, in Chapter 4 we present other similarity measures. Therefore, we introduce more constructors than *simi* can handle.

Description logics represent knowledge through the *knowledge base* which consists of the *TBox*, the *RBox* and the *ABox*. The TBox presents the conceptual knowledge through *concept axioms*. Concept axioms are describing relationships between *concept descriptions* which are built from constructors, *concept names*, *role names* and constants. The RBox describes the relationships between role names and the ABox contains the assertional knowledge. The underlying description logic determines the available constructors and in case of roles the allowed role relationships. In the following section, we first introduce concept descriptions, then the TBox, the RBox and the ABox. Finally, we present a normal form for the description logic $\mathcal{ELH}$ which is used by our measure *simi*.

2.1.1 Concept Descriptions

The main way to express knowledge in a description logic are *concept descriptions*. Concept descriptions consist of constructors related to logic, *concept names*, *role names* and constants. Concept names and role names are finite sets which are denoted as N_C and N_r. In the remainder of this work, we typically use capital letters A and B to refer to concept names, C and D to refer to concept descriptions and r and s to refer to role names. Table 2.1 presents a selection of constructors and the two constants $\top$ and $\bot$, where A is a concept name, r is a role name and C, D are concept descriptions.

Name	Syntax	Semantics
Top	$\top$	$\Delta^{\mathcal{I}}$
Bottom	$\bot$	$\emptyset$
Negation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
Atomic negation	$\neg A$	$\Delta^{\mathcal{I}} \setminus A^{\mathcal{I}}$
Disjunction	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
Conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
Existential quantification	$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in C^{\mathcal{I}} : (x,y) \in r^{\mathcal{I}}\}$
Value restriction	$\forall r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}} : (x,y) \in r^{\mathcal{I}} \Rightarrow y \in C^{\mathcal{I}}\}$
At-least restriction	$(\geq n.r)$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in \Delta^{\mathcal{I}} \mid (x,y) \in r^{\mathcal{I}}\}\rvert \geq n\}$
At-most restriction	$(\leq n.r)$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in \Delta^{\mathcal{I}} \mid (x,y) \in r^{\mathcal{I}}\}\rvert \leq n\}$
Qualified at-least restriction	$(\geq n\, r.C)$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in C^{\mathcal{I}} \mid (x,y) \in r^{\mathcal{I}}\}\rvert \geq n\}$
Qualified at-most restriction	$(\leq n\, r.C)$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in C^{\mathcal{I}} \mid (x,y) \in r^{\mathcal{I}}\}\rvert \leq n\}$

Table 2.1: Concept descriptions constructors

Table 2.2 presents the description logics relevant for this thesis. The set of concept descriptions for a specific description logic $\mathcal{DL}$ (for example the set of $\mathcal{EL}$ concept descriptions) is defined as the smallest set such that all concept names are concept descriptions and all the constructors are satisfied. We denote this set with $\mathcal{C}(\mathcal{DL})$. For example, $\mathcal{C}(\mathcal{EL})$ is the set of all $\mathcal{EL}$ concept descriptions.
Our measure *simi* also allows *role inclusion axioms* which are defined in Section 2.1.2. To denote this extension, the letter $\mathcal{H}$ is added to the description logic. For example, the description logic for *simi* is denoted as $\mathcal{ELH}$.

Constructors and Constants	$\mathcal{L}_0$	$\mathcal{EL}$	$\mathcal{ALN}$	$\mathcal{ALE}$	$\mathcal{ALC}$	$\mathcal{ALCNQ}$
Top	x	x	x	x	x	x
Bottom			x	x	x	x
Negation					x	x
Atomic negation			x	x	x	x
Disjunction					x	x
Conjunction	x	x	x	x	x	x
Existential quantification		x		x	x	x
Value restriction			x	x	x	x
At-least restriction			x			x
At-most restriction			x			x
Qualified at-least restriction						x
Qualified at-most restriction						x

Table 2.2: Description Logics

The semantics of a description logic is given through *interpretations.*

Definition 1 (interpretation)**.** *An* interpretation *is a pair* $\mathcal{I} = (\Delta^{\mathcal{I}}, (\cdot)^{\mathcal{I}})$ *where the* *domain $\Delta^{\mathcal{I}}$ is a non-empty set and $(\cdot)^{\mathcal{I}}$ is a mapping, assigning every concept name $A \in N_C$ a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and every role name $r \in N_r$ a set $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. This mapping is extended to concept descriptions in the way described in Table 2.1.*

Two predicates for concept descriptions depending on interpretations are *subsumption* and *equivalence.*

Definition 2 (subsumed, equivalent)**.** *Let C, D be two concept descriptions. C is* subsumed *by D, denoted as $C \sqsubseteq D$ iff for all interpretations $\mathcal{I}$, $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. C and D are called* equivalent*, denoted as $C \equiv D$ iff $C \sqsubseteq D$ and $D \sqsubseteq C$.*

We introduce a notation where we regard $\mathcal{ELH}$ concept descriptions as sets of *atoms*, where atoms are either concept names or concept descriptions of the form $\exists r.C'$. To distinguish between a concept description an its set representation we use the operator $\widehat{(\cdot)}$.

Definition 3 (atom)**.** *The set of* existential restrictions *of $\mathcal{ELH}$, denoted as N_q is defined as $N_q := \{\exists r.C' \mid r \in N_r, C' \in \mathcal{C}(\mathcal{ELH})\}$. The set of* atoms *of $\mathcal{ELH}$ is $N_A := N_C \cup \{\top\} \cup N_q$. The operator $\widehat{(\cdot)}$ is a function mapping a concept description to a set of atoms, so $\widehat{(\cdot)} : \mathcal{C}(\mathcal{ELH}) \longrightarrow \mathcal{P}(N_A)$. Let $n \in \mathbb{N}_{>0}$ and $C \in \mathcal{C}(\mathcal{ELH})$ with $C = \bigsqcap_{i \leq n} C_i$ where $\forall i \leq n : \; C_i \in N_A$, then $\widehat{C} := \{C_1, C_2, \ldots, C_n\}$.*

As an example, let

$$C := A \sqcap B \sqcap A \sqcap \exists r.(A \sqcap B) \sqcap \exists r.A \sqcap \top$$

then

$$\widehat{C} = \{A, B, \exists r.(A \sqcap B), \exists r.A, \top\}.$$

To formulate some properties of similarity measures we use the *least common subsumer.*

Definition 4 (least common subsumer)**.** *Let $C_1, \ldots, C_n \in N_C$. The* least common subsumer*, denoted as $lcs(C_1, \ldots, C_n)$ is the concept description C such that*

- *$\forall i \leq n : \; C_i \sqsubseteq C$ and*
- *$\forall D \in \mathcal{C}(\mathcal{ELH}) : [\forall i \leq n : C_i \sqsubseteq D] \implies C \sqsubseteq D$.*

2.1.2 TBox and RBox

So far, we described how to build complex concepts (concept descriptions) out of concept names, role names and constants. In this section we introduce *concept axioms* which are used in description logics to describe the relationships between concept descriptions. Additionally, we introduce *role inclusion axioms* which describe relationships between roles.

Definition 5 (axiom). *Let C, D be concept descriptions, $A \in N_C$ and $r, s \in N_r$. A* concept axiom *is either a* concept equivalence axiom*, $C \equiv D$ or a* concept inclusion axiom*, $C \sqsubseteq D$. A concept equivalence axiom of the form $A \equiv D$ is called a* concept definition *defining A and a concept inclusion axiom of the form $A \sqsubseteq D$ is called a* primitive concept definition*.*
A role inclusion axiom *is of the form $r \sqsubseteq s$.*
An interpretation $\mathcal{I}$ satisfies a concept axiom α, denoted as $\mathcal{I} \models \alpha$ if the corresponding semantic condition in Table 2.3 holds.

Name	Syntax	Semantics
concept equivalence axiom	$C \equiv D$	$C^{\mathcal{I}} = D^{\mathcal{I}}$
concept definition	$A \equiv D$	$A^{\mathcal{I}} = D^{\mathcal{I}}$
concept inclusion axiom	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
primitive concept definition	$A \sqsubseteq D$	$A^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
role inclusion axiom	$r \sqsubseteq s$	$r^{\mathcal{I}} \subseteq s^{\mathcal{I}}$

Table 2.3: Concept axioms and role inclusion axiom

It is obvious that a concept equivalence axiom $C \equiv D$ is equivalent to the two concept inclusion axioms $C \sqsubseteq D$ and $D \sqsubseteq C$.
Axioms are grouped together in the *TBox* and the *RBox*.

Definition 6 (TBox). *A finite set of concept axioms is called a* TBox *$\mathcal{T}$. An interpretation $\mathcal{I}$ is called a* model *of $\mathcal{T}$, denoted as $\mathcal{I} \models \mathcal{T}$ iff $\mathcal{I}$ is a model of each element of $\mathcal{T}$. Two TBoxes are called* equivalent *iff they have the same models.*
Let $A, B \in N_C$. We say that A directly uses with *B with respect to $\mathcal{T}$, iff B appears on the right-hand side of a concept axiom of $\mathcal{T}$ defining A. We call* uses *the transitive closure of the relation* directly uses*. $\mathcal{T}$ is called* cyclic *iff there exists a concept name in $\mathcal{T}$ that uses itself. Otherwise, $\mathcal{T}$ is called* acyclic*.*

Definition 7 (RBox). *A finite set of role inclusion axioms is called an RBox $\mathcal{R}$.*
An interpretation $\mathcal{I}$ is called a model of $\mathcal{R}$, denoted as $\mathcal{I} \models \mathcal{R}$, iff $\mathcal{I}$ is a model for each element of $\mathcal{R}$.
A role inclusion axiom $r \sqsubseteq s$ follows *from an RBox $\mathcal{R}$, denoted as $r \sqsubseteq_{\mathcal{R}} s$, iff for all interpretations $\mathcal{I}$ we have $\mathcal{I} \models \mathcal{R} \implies \mathcal{I} \models r \sqsubseteq s$.*
We say that two roles r and s are equivalent with respect to an RBox $\mathcal{R}$, denoted as $r \equiv_{\mathcal{R}} s$ iff $r \sqsubseteq_{\mathcal{R}} s$ and $s \sqsubseteq_{\mathcal{R}} r$.

Note that for all RBoxes $\mathcal{R}$ and all $r \in N_r$ we have $r \sqsubseteq_{\mathcal{R}} r$.

Definition 8 (primitive name). *Let $\mathcal{T}$ be a TBox. A concept name A is called a* primitive name *of $\mathcal{T}$, iff it does not occur on the left-hand side of any concept axiom in $\mathcal{T}$. The set of all primitive names is denoted as N_B.*

In this thesis, we focus on the usage of a special kind of TBoxes called *unfoldable TBoxes*. The reason is that *unfoldable TBoxes* allow us to *expand* the concept descriptions to measure in a way that the knowledge represented by the TBox is not necessary any more in order to measure similarity. The procedure of expansion is described in the end of this section.

Definition 9 (unfoldable TBox). *A TBox $\mathcal{T}$ is called a* unfoldable TBox *iff it is acyclic, consists of concept definitions and primitive concept definitions only and every concept name occurs at most once on a left-hand side of a concept axiom.*

We show how a concept description can be transformed into an equivalent one, such that the knowledge of the concept axioms of the TBox is not necessary any more. Let $\mathcal{T}$ be an arbitrary unfoldable TBox and C be a concept description. We start by transforming the TBox. First, we transform all primitive concept definitions $A \sqsubseteq D$ into concept definitions by introducing a new concept name A' and using the rule

$$A \sqsubseteq D \longrightarrow A \equiv A' \sqcap D.$$

The resulting TBox is called the *normalization* of $\mathcal{T}$. The following corollary from [BCM$^+$03] show that the normalization does not change the semantics on the old concept names.

Corollary 1 ([BCM$^+$03] Proposition 2.10). *Let $\mathcal{T}$ be a unfoldable TBox and $\overline{\mathcal{T}}$ its normalization. Then*

- *every model of $\overline{\mathcal{T}}$ is a model of $\mathcal{T}$ and*
- *every model $\mathcal{I}$ of $\mathcal{T}$ can be extended to a model of $\overline{\mathcal{T}}$ such that both interpretations have the same domain.*

Every acyclic TBox $\mathcal{T}$ can be transformed into a TBox $\mathcal{T}^*$ such that on the right-hand side of each concept axiom, only primitive names occur. This can be achieved by replacing each occurrence of a defined name on the right-hand side by its definition. Since there are no cycles, this process terminates and we obtain the TBox $\mathcal{T}^*$ called *expansion*.

Corollary 2 ([BCM$^+$03], Proposition 2.1). *Let $\mathcal{T}$ be an acyclic TBox and $\mathcal{T}^*$ its expansion. Then $\mathcal{T}$ and $\mathcal{T}^*$ are equivalent.*

When we are measuring two concept descriptions C and D with respect to an unfoldable TBox $\mathcal{T}$, we first normalize and expand $\mathcal{T}$ and then replace every occurrence of a non primitive name in C and D with its definition.
For example, let $N_C := \{A, B, E, F, G\}$, $N_r := \{r, s\}$. The TBox is

$$\mathcal{T} := \{E \sqsubseteq A \sqcap \exists r F,\ F \sqsubseteq B \sqcap G\},$$

$C = E \sqcap B \sqcap \exists r.(F \sqcap B)$ and $D = F \sqcap \exists s.(A \sqcap E)$.
The normalization of $\mathcal{T}$ is

$$\mathcal{T}' = \{E \equiv E' \sqcap A \sqcap \exists r.F,\ F \equiv F' \sqcap B \sqcap G\}$$

and the expansion of $\mathcal{T}$' is

$$\mathcal{T}^* = \{E \equiv E' \sqcap A \sqcap \exists r.(F' \sqcap B \sqcap G),\ F \equiv F' \sqcap B \sqcap G\}.$$

Finally, C and D are expanded to

$$C' = E' \sqcap A \sqcap \exists r.(F' \sqcap B \sqcap G) \sqcap B \sqcap \exists r.(F' \sqcap B \sqcap G \sqcap B)$$

and

$$D' = F' \sqcap B \sqcap G \sqcap \exists s.(A \sqcap E' \sqcap A \sqcap \exists r.(F' \sqcap B \sqcap G)).$$

Note that normalization is polynomial whereas the expansion is in general exponential in the number of concept axioms of $\mathcal{T}$ [Neb90].

Finally, we present a characterisation of subsumption in $\mathcal{ELH}$ with respect to an empty TBox and an RBox.

Lemma 1. *Let $C, D \in \mathcal{C}(\mathcal{ELH})$ and $\mathcal{R}$ be an RBox. Then $C \sqsubseteq D$ with respect to $\mathcal{R}$ iff*

- *$\widehat{C} \cap N_C \subseteq \widehat{D} \cap N_C$ and*
- *for all existential restrictions $\exists r.D' \in \widehat{D}$ there exist an existential restriction $\exists s.C' \in \widehat{C}$ such that $s \sqsubseteq_{\mathcal{R}} r$ and $C' \sqsubseteq D'$.*

2.1.3 ABox

An ABox allows to express knowledge about *individuals* through *assertions*. There are two kinds of assertions. One to express that an individual belongs to a concept description and the other to express that two individuals are related in the context of a role.

Definition 10 (ABox)**.** *Let N_I be a finite set which we call the set of individuals, $x, y \in N_I$, C be a concept description and $r \in N_r$. We call $C(x)$ a* concept assertion *and $r(x, y)$ a* role assertion*. An* ABox*, denoted with $\mathcal{A}$, is a finite set containing concept and role assertions.*

To describe the semantics of ABoxes, the definition of interpretations is extended to individuals. An interpretation $\mathcal{I}$ has to map every individual x to an element of the domain $\Delta^{\mathcal{I}}$. This is denoted as $x^{\mathcal{I}} \in \Delta^{\mathcal{I}}$.

Definition 11. *We say the interpretation $\mathcal{I}$* satisfies the concept assertion *$C(x)$ iff $x^{\mathcal{I}} \in C^{\mathcal{I}}$ and it* satisfies the role assertion *$r(x, y)$ iff $(x, y) \in r^{\mathcal{I}}$. The interpretation $\mathcal{I}$* satisfies *$\mathcal{A}$ iff it satisfies all assertions of $\mathcal{A}$.*

An ABox $\mathcal{A}$ has a special interpretation called *canonical interpretation*.

Definition 12 (canonical interpretation)**.** *Let $\mathcal{A}$ be an ABox and N_I its set of individuals. The interpretation $\mathcal{I}_{\mathcal{A}}$ with*

- $\Delta^{\mathcal{I}_\mathcal{A}} = N_I$,
- $\forall x \in N_I : x^{\mathcal{I}_\mathcal{A}} = x$,
- $C(x) \in \mathcal{A} \iff x \in C^{\mathcal{I}_\mathcal{A}}$ *and*
- $r(x,y) \in \mathcal{A} \iff (x,y) \in r^{\mathcal{I}_\mathcal{A}}$

is called the canonical interpretation*.*

2.1.4 A Normal Form for $\mathcal{ELH}$ Concept Descriptions

To ensure that our measure *simi* is equivalence invariant, we use the unique normal form (with respect to associativity and commutativity) for $\mathcal{EL}$ concept descriptions introduced in [Kü00]. According to [Kü00] (Proposition 6.3.1), this normal form can be obtained by recursively applying the following rules

1. $A \sqcap \top \longrightarrow A$,
2. $A \sqcap A \longrightarrow A$,
3. $\exists r.C' \sqcap \exists r.D' \longrightarrow \exists r.C'$ if $C' \sqsubseteq D'$

to the concept description until no further rule can be applied.
The normal form is built without using the knowledge of a TBox. Our later defined measure *simi* does not use a TBox either. But it uses the knowledge of an RBox $\mathcal{R}$. To incorporate $\mathcal{R}$ and still obtain a unique normal form, we have to make two adjustments to the normal form. First, we build role equivalence classes. For $r \in N_r$ we define

$$[r] := \{s \in N_r \mid r \equiv_\mathcal{R} s\}$$

and

$$A(N_r) := \{[r] \mid r \in N_r\}.$$

Then we randomly pick one role from every equivalence class and replace every occurrence of roles in the same class with this one. More formally, this means we build a function

$$f : A(N_r) \longrightarrow N_r$$

with $f([r]) \in [r]$ and then recursively apply the rule

$$\exists r.C' \longrightarrow \exists f([r]).C'.$$

The second adjustment is that we have to change the third rule from [Kü00] to

$$\exists r.C' \sqcap \exists s.D' \longrightarrow \exists r.C' \text{ if } r \sqsubseteq_\mathcal{R} s \text{ and } C' \sqsubseteq D'.$$

For example, let

$$C := A \sqcap B \sqcap A \sqcap \exists r.(A \sqcap B) \sqcap \exists s_2.A \sqcap \exists t.A \sqcap \top$$

and $\mathcal{R} := \{r \sqsubseteq s_1, s_1 \sqsubseteq s_2\}$ then the normal form of C is $C' = A \sqcap B \sqcap \exists r.(A \sqcap B) \sqcap \exists t.A$.

2.2 Triangular Norms and Conorms

Triangular norms and conorms are a way to generalise the 2-valued propositional logic. They are operators accepting all values between 0 and 1. Our measure *simi* depends on a triangular norm and on a triangular conorm. Therefore, we present a short introduction.
A triangular norm is the generalization of the *and* ($\wedge$) operator from propositional logic.

Definition 13 (t-norm)**.** *A function* $\otimes : [0,1]^2 \longrightarrow [0,1]$ *is called a* triangular norm *(*t-norm*) iff it fulfills the following properties for all* $x, y, z, w \in [0,1]$*:*

- *commutativity:* $x \otimes y = y \otimes x$*,*
- *monotonicity:* $x \leq z$ *and* $y \leq w \implies x \otimes y \leq z \otimes w$*,*
- *associativity:* $(x \otimes y) \otimes z = x \otimes (y \otimes z)$ *and*
- *identity:* $x \otimes 1 = x$*.*

A t-norm is called bounded *iff* $x \otimes y = 0 \implies x = 0$ *or* $y = 0$*.*

Triangular conorms are the generalization of the logical *or* ($\vee$) operator.

Definition 14 (t-conorm)**.** *A function* $\oplus : [0,1]^2 \longrightarrow [0,1]$ *is called a* triangular conorm *(*t-conorm*) iff it fulfills the following properties for all* $x, y, z, w \in [0,1]$*:*

- *commutativity:* $x \oplus y = y \oplus x$*,*
- *monotonicity:* $x \leq z$ *and* $y \leq w \implies x \oplus y \leq z \oplus w$*,*
- *associativity:* $(x \oplus y) \oplus z = x \otimes (y \oplus z)$ *and*
- *identity:* $x \oplus 0 = x$*.*

A t-conorm is called bounded *iff* $x \oplus y = 1 \implies x = 1$ *or* $y = 1$*.*

It is easy to prove that, for any t-norm $\otimes$, the operator $\oplus'$ defined through

$$x \oplus' y := 1 - [(1-x) \otimes (1-y)]$$

is a t-conorm and in duality using a t-conorm $\oplus$, one can construct a t-norm $\otimes'$ by

$$x \otimes' y := 1 - [(1-x) \oplus (1-y)].$$

Here, $1 - x$ acts as a generalisation of the negation operator. The way of obtaining a t-conorm from a t-norm is similar to how the *or* operator in propositional logic can be defined through negation and the *and* operator.
Examples of t-norms and their corresponding t-conorm are presented in Tables 2.4 and 2.5.

Name	Symbol	$x \otimes y =$	bounded?
Minimum t-norm	$\otimes_{min}$	$min\{x, y\}$	yes
Product t-norm	$\otimes_{prod}$	xy	yes
Lukasiewicz t-norm	$\otimes_{luk}$	$max\{0, x + y - 1\}$	no
Drastic t-norm	$\otimes_{dr}$	b if $a = 1, a$ if $b = 1, 0$ othwerwise	no
Hamacher product	$\otimes_{H0}$	0 if $x = y = 0$, $\frac{xy}{x+y-xy}$ otherwise	yes

Table 2.4: Examples of t-norms

Name	Symbol	$x \oplus y =$	bounded?
Maximum t-conorm	$\oplus_{max}$	$max\{x, y\}$	yes
Probabilistic sum	$\oplus_{sum}$	$x + y - xy$	yes
Bounded sum	$\oplus_{luk}$	$min\{x + y, 1\}$	no
Drastic t-conorm	$\oplus_{dr}$	b if $a = 0, a$ if $b = 0, 1$ othwerwise	no
Einstein sum	$\oplus_{H2}$	$\frac{x+y}{1+xy}$	yes

Table 2.5: Examples of t-conorms

Finally, we present basic properties of t-norms and t-conorms.

Lemma 2. *Let $\otimes$ be a t-norm, $\oplus$ be a t-conorm and $x, y \in [0, 1]$. Then*

1. $0 \otimes x = 0$ *and*
2. $1 \oplus x = 1$.

Proof.

1. Monotonicity implies $0 \otimes x \leq 0 \otimes 1$ and identity implies that $0 \otimes 1 = 0$. Therefore $0 \otimes x = 0$.
2. Monotonicity implies $1 \oplus x \geq 1 \oplus 0$ and identity implies that $1 \oplus 0 = 1$. Therefore $1 \oplus x = 1$.

□

Note that, because t-norms and t-conorms are associative and commutative, the notions $\bigotimes_{i \leq n} x_n$ and $\bigoplus_{i \leq n} x_n$ for some $x_0, \ldots, x_n \in [0, 1]$ are well defined.

3 Similarity-Measure Properties

In this chapter we present several properties of similarity measures. None of them depends on the knowledge of a TBox. The reason is that all measures are defined for unfoldable TBoxes and therefore we assume that the concept descriptions are expanded (see Section 2.1.2).

As starting point to identify description-logic-similarity-measure properties, d'Amato et.al. suggested to look at *metric spaces* [dSF08]. Metrics and their properties are of interest in similarity research of several areas like information theory [LCL$^+$03], chemistry [NJ03], music similarity [LS04] and psychology [Tve77, BG97, GS04].

Definition 15 (metric space)**.** *A* metric space (D, d) *consists of a set D and a function $d : D \times D \longrightarrow \mathbb{R}_{\geq 0}$ called* metric *such that for all $a, b, c \in D$*

1. $d(a, b) = 0 \iff a = b$ (identity of indiscernible)

2. $d(a, b) = d(b, a)$ (symmetry)

3. $d(a, b) \leq d(a, c) + d(c, b)$ (triangle inequality).

For $a, b \in D$ the value $d(a, b)$ represents the *distance* of a and b. A metric can be interpreted as a *dissimilarity measure.* The distance represents the dissimilarity between a and b. The lower the distance, the higher the similarity. To obtain a similarity measure from a metric we have to *normalize* it. With normalizing we mean that we bound the maximal distance with 1, so that the metric is a function producing distances between 0 and 1.

Definition 16 (normalized metric space)**.** *A metric space (D, d) is called* normalized *iff for all $a, b \in D :\ d(a, b) \leq 1$.*

From a normalized metric space (D, d), we can define a similarity function $s : D \times D \longrightarrow [0, 1]$ through for all $a, b \in D$

$$s(a, b) := 1 - d(a, b).$$

If we translate the properties of a metric we obtain similar properties for similarity functions.

Definition 17 (similarity function)**.** *Let D be a set. A function $s : D \times D \longrightarrow [0, 1]$ is called a* similarity function *for D iff for all $a, b, c \in D$*

1. $s(a, b) = 1 \iff a = b$, (identity of indiscernible)

2. $s(a,b) = s(b,a)$, *(symmetry)*

3. $1 + s(a,b) \geq s(a,c) + s(c,b)$ *(triangle inequality).*

The derivation for the triangle inequality from the triangle inequality of a metric d is as follows

$$\begin{aligned}
&d(C,D) \leq d(C,E) + d(E,D) && \Longleftrightarrow\\
&1 - d(C,D) \geq 1 - d(C,E) - d(E,D) && \Longleftrightarrow\\
&s(C,D) \geq s(C,E) - d(E,D) && \Longleftrightarrow\\
&s(C,D) \geq s(C,E) - 1 + 1 - d(E,D) && \Longleftrightarrow\\
&1 + s(C,D) \geq s(C,E) + s(E,D).
\end{aligned}$$

The property symmetry is rather easy to achieve and to prove, whereas triangle inequality and identity of indiscernible are more difficult. Therefore, we include symmetry in our definition of similarity measures and present the others as additional properties. Symmetry is also a very controversial property. While some consider it to be essential [Lin98], cognitive sciences favours asymmetric similarity measures because of their founding that human intuition behaves asymmetric [Tve77, BG97]. As for measures for description logics Janowicz et.al [JW09, Jan06] prefers asymmetry (but presented symmetric versions of his measures) where d'Amato et.al [dFE06, dFE05, FD06, dSF08] consider it to be a fundamental property. We believe that the reason why human intuition may not be symmetric is that the knowledge base of a human is not constant. If he reads the first word, it already influences his thoughts and therefore his "TBox“. In contrast to the human brain, in computer science and applications of description logics the knowledge base is always constant. For a computer program, reading a concept description to measure does not change the TBox. Therefore, we assume that the research of Tversky et.al cannot be applied and we vote in favour of symmetry.

Definition 18 (similarity measure)**.** *A similarity measure sim is a function*

$$sim : \mathcal{C}(\mathcal{ELH}) \times \mathcal{C}(\mathcal{ELH}) \longrightarrow [0,1]$$

such that for all $C, D \in \mathcal{C}(\mathcal{ELH})$ $sim(C,D) = sim(D,C)$ *(symmetry).*

In the following sections we present the definition of other properties and the underlying expectation this properties represent. The properties *triangle inequality*, *equivalence invariant* and *equivalence closed* are derived from relevant literature whereas the properties *subsumption preserving*, *reverse subsumption preserving*, *dissimilar closed*, *bounded* and *structural dependent* are new. To find a new property we used the perspective that a property is a formalism of expected behaviour. By implication, this also means that if a property does not hold, then the corresponding counterexample represents a case of unintuitive behaviour. Therefore, properties can be found by generating an example with unintuitive behaviour and then formulating a property such that this behaviour is not possible any more. Then one should try to contradict the new property by identifying an example that contains unintuitive behaviour, yet is consistent with the underlying idea of the property itself.

In the last section, we present the property *monotonicity* which is defined in [dSF08] and we explain why we do not use it.

3.1 Triangle Inequality

Adopting the definition of the similarity-function triangle inequality for description logics leads to the following definition.

Definition 19 (triangle inequality)**.** *A similarity measure sim has the* triangle inequality *property iff for all* $C, D, E \in \mathcal{C}(\mathcal{ELH})$

$$1 + sim(D, E) \geq sim(D, C) + sim(C, E).$$

The following lemma presents a different formulation of the triangle inequality. The version of the definition is easier to prove, where the other version is easier to interpret.

Lemma 3. *Let sim be a similarity measure then the following two statements are equivalent.*

1. $\forall C, D, E \in \mathcal{C}(\mathcal{ELH}) : 1 + sim(D, E) \geq sim(D, C) + sim(C, E)$
2. $\forall C, D, E \in \mathcal{C}(\mathcal{ELH}) : 1 - sim(C, D) \geq |sim(C, E) - sim(E, D)|$.

Proof. Let C, D and E be arbitrary concept descriptions. We distinguish two cases. First, if $sim(C, E) \geq sim(E, D)$ then $|sim(C, E) - sim(E, D)| = sim(C, E) - sim(E, D)$ and we have

$$\begin{aligned}1 + sim(D, E) \geq sim(D, C) + sim(C, E) &\iff\\ 1 \geq sim(D, C) + sim(C, E) - sim(D, E) &\iff\\ 1 - sim(C, D) \geq sim(C, E) - sim(E, D) &\iff\\ 1 - sim(C, D) \geq |sim(C, E) - sim(E, D)|.\end{aligned}$$

Second, if $sim(C, E) < sim(E, D)$ then

$$1 + sim(D, E) > 1 + sim(C, E) \geq sim(D, C) + sim(C, E).$$

To show $1 - sim(C, D) \geq |sim(C, E) - sim(E, D)|$ we use the fact that

$$|sim(C, E) - sim(E, D)| = sim(E, D) - sim(C, E).$$

Since we can assume that statement 1. is true for all C, D and E, we know that $1 + sim(C, E) \geq sim(D, C) + sim(D, E)$ is true. Therefore we derive

$$1 + sim(C, E) \geq sim(D, C) + sim(D, E) \iff$$

$$\begin{aligned} 1 \geq sim(D,C) + sim(D,E) - sim(C,E) \iff \\ 1 - sim(C,D) \geq sim(D,E) - sim(C,E) \iff \\ 1 - sim(C,D) \geq |sim(C,E) - sim(E,D)|. \end{aligned}$$

□

We present a short example. Let $sim(C,D) = 0.9$ and $sim(C,E) = 0.5$. Looking at the second version, the triangle inequality implies that $sim(E,D) \in [0.4, 0.6]$. It can only variate ±0.1 from $sim(C,E)$ because $sim(C,D)$ is only 0.1 less than 1.
In [Tve77], Tversky argued that human similarity reasoning does not fulfill triangle inequality. His argument is based on the following example (note that the paper is from 1977). "Jamaica is similar to Cuba (because of geographical proximity), Cuba is similar to Russia (because of their political affinity), but Jamaica and Russia are not similar at all." We found that this argument is not applicable to description logic similarity. A human who has to measure Jamaica and Cuba tends to give a high value because of the geographical proximity. When measuring Cuba and Russia on the other hand, the focus resides on the political affinity. As for the geographical proximity, it is either weighted down in its influence on the measure or it is not part of the description any more. Therefore, either the weights of the features (political system, geographical proximity) or the concept descriptions are depending on the actual objects to measure. This is never the case in description logics. Both the concept descriptions and the weighting are constant. If we represent the example with description logics and include all features then we have

$$\begin{aligned} \mathit{Cuba} &\equiv \mathit{Communism} \sqcap \mathit{MiddleAmerica}, \\ \mathit{Jamaica} &\equiv \mathit{Democracy} \sqcap \mathit{MiddleAmerica}, \\ \mathit{Russian} &\equiv \mathit{Communism} \sqcap \mathit{Asia}. \end{aligned}$$

Now, reasonable similarities could be

$$\begin{aligned} sim(\mathit{Cuba}, \mathit{Jamaica}) &= \frac{1}{2}, \\ sim(\mathit{Cuba}, \mathit{Russia}) &= \frac{1}{2}, \\ sim(\mathit{Russia}, \mathit{Jamaica}) &= 0 \end{aligned}$$

which is consistent with the triangle inequality.

Neither any of the measures for description logics we found ([JW09, Jan06, dFE06, dFE05, FD06, dSF08]), nor our own measure *simi* fulfills triangle inequality. Two papers mentioned triangle inequality. [dSF08], where triangle inequality is described as a desirable property and [Jan06] where it is argued against the pursue of triangle inequality because of Tverskys [Tve77] work.

3.2 Equivalence Closed and Equivalence Invariant

Description logics make it possible to describe the same thing in different ways. Two concept descriptions can be syntactically different yet semantically equivalent. A similarity measure should depend on the semantics rather than the syntax of the concept descriptions to measure. *Equivalence invariant* expresses this requirement in form of a property. It states that two equivalent concept descriptions must have the same similarity towards a third concept description.

Definition 20 (equivalence invariant)**.** *A similarity measure sim is* equivalence invariant *iff for all* $C, D, E \in \mathcal{C}(\mathcal{ELH})$

$$C \equiv D \implies sim(C, E) = sim(D, E).$$

Equivalence invariant is widely accepted as a necessary property for description logic measures ([JW09, Jan06, dFE06, dFE05, FD06, dSF08]). Yet we found that the methods used to ensure equivalence invariant where not always sound (see Chapter 4).

In addition, we adopt the property "identity of indiscernible" from similarity functions to description logics by replacing the equality with equivalence and call it *equivalence closed.* A similarity measure is equivalence closed iff two concept descriptions are totally similar if and only if they are equivalent.

Definition 21 (equivalence closed)**.** *A similarity measure sim is* equivalence closed *iff for all* $C, D \in \mathcal{C}(\mathcal{ELH})$

$$sim(C, D) = 1 \iff C \equiv D.$$

Equivalence closed is considered to be a basic property for description logic similarity measures [dSF08, Jan06] especially because it is adapted from metrics.

3.3 Subsumption Preserving and Reverse Subsumption Preserving

The properties *subsumption preserving* and *reverse subsumption preserving* aim to connect subsumption and similarity. Let $C, D, E \in \mathcal{C}(\mathcal{ELH})$ with $C \sqsubseteq D \sqsubseteq E$. Subsumption preserving expresses the expectation that the similarity of C and D should be higher than the similarity of C and E because D is 'closer' to C than E. Reverse subsumption preserving is the addition stating the expectation that the similarity of D and E should be higher than the similarity of C and E because D is 'closer' to E than C.

Definition 22 (subsumption preserving, reverse subsumption preserving)**.** *A similarity measure sim is* subsumption preserving *iff for all* $C, D, E \in \mathcal{C}(\mathcal{ELH})$ *with* $C \sqsubseteq D \sqsubseteq E$ *we have* $sim(C, D) \geq sim(C, E)$.
It is reverse subsumption preserving *iff for all* $C, D, E \in \mathcal{C}(\mathcal{ELH})$ *with* $C \sqsubseteq D \sqsubseteq E$ *we have* $sim(C, E) \leq sim(D, E)$.

3.4 Bounded and Dissimilar Closed

Bounded and *dissimilar closed* are properties regarding total dissimilarity. Both are using the least common subsumer. Dissimilar closed states that if the least common subsumer of two concept descriptions is equivalent to top then we expect that they are totally dissimilar (their similarity value is 0). If the least common subsumer is not equivalent to top, then the concept descriptions have something in common and the similarity should be higher than 0. This expectation is summed up in the property bounded.

Definition 23 (dissimilar closed, bounded)**.** *A similarity measure sim is called* dissimilar closed *iff for all $C, D \in \mathcal{C}(\mathcal{ELH})$*

$$C \not\equiv \top, D \not\equiv \top \text{ and } lcs(C, D) \equiv \top \implies sim(C, D) = 0$$

and it is called bounded *iff for all $C, D \in \mathcal{C}(\mathcal{ELH})$*

$$lcs(C, D) \not\equiv \top \implies sim(C, D) > 0.$$

3.5 Structural Dependent

The property structural dependent is motivated through the *feature model* which is presented by Tversky in [Tve77]. There, an object is described through a set of features. The similarity of two objects is measured by a relation between the number of common features of both objects and the number of unique features of each of the objects. The basic rule is that if

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

then the similarity must increase. The property *structural dependent* is derived from this basic rule. We take the perspective that an atom can be regarded as a feature. The atoms of a conjunction (of atoms) represent the features of an object (concept description). Our expectation is that the more features (atoms) two concept descriptions share, the higher their similarity should be. This expectation is expressed as follows.

Definition 24 (structural dependent)**.** *A similarity measure sim is called* structural dependent *iff for all $D, E \in \mathcal{C}(\mathcal{ELH})$ and all sequences $(C_n)_n$ of atoms with $\forall i, j \in \mathbb{N}, i \neq j : C_i \not\sqsubseteq C_j$ the concept descriptions*

$$D_n := \bigsqcap_{i \leq n} C_i \sqcap D$$

and

$$E_n := \bigsqcap_{i \leq n} C_i \sqcap E$$

fulfill the condition

$$\lim_{n\to\infty} sim(D_n, E_n) = 1.$$

In the definition, D and E act as the “sets“ of (possible) uncommon features and as the second condition of the basic rule requires, they are constant. That the first condition of the basic rule is true is not so easy to see. We have to prove that the number of common features really increases infinitely. First, let us observe that $D_1 \sqsupseteq D_2 \sqsupseteq \ldots \sqsupseteq D_n \sqsupseteq \ldots$. The condition $C_i \not\sqsubseteq C_j$ in the definition of structural dependent is there to ensure that every “new“ C_n is different from the others and provides new common information. However, we still have to prove that the sequence $(D_n)_n$ of concept descriptions is really a descending chain of subsumed concept descriptions. More formally, there exists no $m \in \mathbb{N}$ such that for all $n \geq m : D_m \equiv D_n$. To prove this we need the following lemma.

Lemma 4. *Let $C \in \mathcal{C}(\mathcal{ELH})$ and $\mathcal{R}$ be the corresponding RBox then the number of subsumers of C is finite, or more formally*

$$|\{D \in \mathcal{C}(\mathcal{ELH}) \mid C \sqsubseteq D\}| \in \mathbb{N}.$$

Proof. For the rest of the proof we define $S(C) := \{D \in \mathcal{C}(\mathcal{ELH}) \mid C \sqsubseteq D\}$. The proof is conducted by structural induction. If $C \in N_C$ then $S(C) = \{\top, C\}$ which is finite.
Let $C = \exists r.C^*$ for some $r \in N_r$ and we assume that $S(C^*)$ is finite. Let D be a concept description with $C \sqsubseteq D$. Lemma 1 implies that all atoms D' in $\widehat{D}$ must subsume C. Therefore, we only need to prove that there are only finitely many atoms D' with $C \sqsubseteq D'$. For all $A \in N_C$ we have $C \not\sqsubseteq A$. Let $\exists s.D^*$ be a existential restriction with $\exists r.C^* \sqsubseteq \exists s.D^*$. This implies that $r \sqsubseteq_{\mathcal{R}} s$ and $D^* \in S(C^*)$. Therefore we can derive that $|S(C)| \leq |N_r| \cdot |S(C^*)|$. Since N_r and $S(C^*)$ are assumed to be finite, $S(C)$ has to be finite as well.
For the final case, let C be a conjunction of atoms where we know that for all $C' \in \widehat{C}$ the set $S(C')$ is finite. Additionally let D be a concept description with $C \sqsubseteq D$. Lemma 1 implies that for all $D' \in \widehat{D}$ there exists a $C' \in \widehat{C}$ such that $D' \in S(C')$. Therefore, $|S(C)| \leq \prod_{C' \in \widehat{C}} |S(C')|$ which implies that $S(C)$ is finite. □

Now we can prove that the definition of structural dependent also fulfills the first condition of the basic rule.

Lemma 5. *Let $D \in \mathcal{C}(\mathcal{ELH})$, $(C_n)_n$ be a sequence of atoms with $\forall i, j \in \mathbb{N}, i \neq j : C_i \not\sqsubseteq C_j$ and*

$$D_n := \bigsqcap_{i \leq n} C_i \sqcap D.$$

There exists no $m \in \mathbb{N}$ such that for all $n \geq m : D_m \equiv D_n$.

Proof. Let us assume that there exists such an m and let n be an arbitrary number with $n > m$. Since $D_m \equiv D_n$ we know that $D_m \sqsubseteq D_n$. For any C_i with $m < i \leq n$ Lemma 1 implies that there exits a $F' \in \hat{D_m}$ such that $F' \sqsubseteq C_i$. Since for all $j \leq m : C_j \not\sqsubseteq C_i$, F' has to be an atom of D which implies that $D \sqsubseteq C_i$. Therefore, all concept descriptions C_i with $m < i$ are subsumers of D. The fact that $C_j \not\sqsubseteq C_i$ implies that all of them are not equivalent to each other. This leads us to the conclusion that D has infinitely many subsumers which is a contradiction to Lemma 4. □

3.6 Towards Monotonicity

In [dSF08] a property called *monotonicity* is presented. It is defined for dissimilarity measures over arbitrary description logics. We present an adjustment of the definition to similarity measures and $\mathcal{ELH}$.

Definition 25 ([dSF08])**.** *Let sim be similarity measure. Sim obeys the* monotonicity *criterion iff given the concept descriptions* $C, D, E, U, L \in \mathcal{C}(\mathcal{ELH})$

1. $C, D \sqsubseteq U$ *and* $C, D \sqsubseteq L$*,*
2. $E \sqsubseteq U$*,* $E \not\sqsubseteq L$ *and*
3. $\nexists H \in \mathcal{C}(\mathcal{ELH})$ *such that* $C, E \sqsubseteq H$ *and* $D \not\sqsubseteq H$*.*

imply that $sim(C, D) > sim(C, E)$*.*

The underling idea is that "if given the concepts C, D and E, the concept generalizing C and D is more specific (w.r.t. the subsumption relationship) than the concept generalizing C and E, then C and D are more similar to each other w.r.t. C and E."

The first part states that C and D have two (different) common subsumers. The second part states that E is subsumed only by one of this subsumers and the third property ensures that C and E do not have another common subsumer which is not a subsumer of D. The implication is that because C and D have (at least) two common subsumers where C and E have one, the similarity of C and D should be higher. In a more abstract point of view, monotonicity states that the similarity of C and D is higher than the similarity of C and E because C and D have more in common. However, in our point of view, the problem of this property is that it does not forces any condition on the differences of C and D. This allows to construct an example which fulfills all three conditions, but in our intuition, the similarity between C and D should be lower than the similarity of C and E. Let $N_C := \{A_1, \ldots, A_n, B, X, W, Y_1, \ldots, Y_m\}$ where all concept names are different from each other and

- $C := \bigsqcap_{i \leq n} A_i \sqcap B \sqcap X$,
- $D := \bigsqcap_{i \leq n} A_i \sqcap B \sqcap \bigsqcap_{j \leq m} Y_j$,
- $E := \bigsqcap_{i \leq n} A_i \sqcap W$,
- $U := \bigsqcap_{i \leq n} A_i$ and
- $L := B$.

Then we have $C, D, E \sqsubseteq U$, $C, D \sqsubseteq L$ and $E \not\sqsubseteq L$. Since $lcs(C, E) = U$ and $D \sqsubseteq U$, every H with $C, E \sqsubseteq H$ would imply that $D \sqsubseteq H$. Therefore condition three is true. If n is a high number then the similarity between C and E should be very high because they only differ in the concept names B, X and W. On the other hand, the greater m is, the lower the similarity between C and D should be. If we set m to a value which is much higher than n then the similarity of C and D should be lower than the similarity of C and E. This is in contradiction to monotonicity which requires the opposite.

4 Known Similarity Measures and their Properties

In this chapter we investigate several measures towards which of the properties from Chapter 3 they fulfill and in some case what kind of general unintuitive behaviour we can derive from this investigation. We distinguish between two kinds of measures, structural measures and interpretation based measures. *Structural measures* are defined using the syntax of the concept descriptions to measure. Since conjunction and disjunction are commutative and associative, these measures are invariant to the order of the atoms in a conjunction and disjunction meaning that the order of atoms does not influence the result. One difference between the measures arises on the computation of the similarity of primitive concepts. [Jan08] uses the TBox whereas the measures [dFE06] and [FD06] use the canonical interpretation for this purpose. The Jaccard Index and Dice's Coefficient assume that two concept names are totally dissimilar iff they are different. To achieve equivalence invariance, the concept descriptions need to be transformed into a unique normal form before measuring similarity. The uniqueness ensures that for two different yet equivalent concept descriptions the resulting normal forms are syntactically the same (with respect to commutativity and associativity). [JW09] uses the *negation normal form* where [Jan08] and [dFE06] are using the $\mathcal{ALCN}$ normal form presented in Section 4.1. Both normal forms are not unique and therefore the measures are not equivalence invariant (yet they claim to be). In [FD06], a restriction of the $\mathcal{ALCN}$ normal form to $\mathcal{ALN}$ is used. The two $\mathcal{L}_0$ measures Jaccard Index and Dice's Coefficient do not need a normal form because the description logic is to simple.

Interpretation based measures are defined using interpretations and cardinality. They do not use the syntax of the concept descriptions to measure. Therefore, they are trivially equivalence invariant. The two interpretation based measures we present [dFE05, dSF08] are using the canonical interpretation $\mathcal{I}_\mathcal{A}$. That is why these measures need a populated and representative domain (an ABox).

In the first section we define the $\mathcal{ALCN}$ normal form used by several structural measures. Then, in the Sections 4.2 and 4.3, we present an analysis of the properties of six structural measures and two interpretation based measures. Since not every measure has a name, we use the names of the papers as titles of our sections and denote the measures with *sim* (if it does not have a name). All measures are working with unfoldable TBoxes. Therefore, we assume that the concepts to measure are expanded. Note that we are not assuming that the TBox is empty because the measure [Jan08] uses the knowledge of the TBox to expand the concepts "on the fly". The analysis is done as follows. First, we shortly explain if and what kind of preconditions (for example a normal form) are needed. Then we present a definition of the measure (except for [JW09]). Afterwards, we point to unintuitive behaviour we observed while investigating the properties of the measure. Finally, we present proofs of all fulfilled properties and counterexamples for the non-fulfilled ones.

In the last section we present an tabular overview that contains all measures (including our measure *simi* from Chapter 5) and properties.

4.1 A Normal Form for $\mathcal{ALCN}$ Concept Descriptions

We present a normal form for the description logic $\mathcal{ALCN}$ from [dFE06] (originally [BKT02]). To describe the normal form we introduce the following notations.

Definition 26. *Let $r \in N_r$. We define*

- $N_\forall(r) := \{\forall r.D \mid D \in \mathcal{C}(\mathcal{ALCN})\}$,
- $N_\exists(r) := \{\exists r.D \mid D \in \mathcal{C}(\mathcal{ALCN})\}$,
- $N_\leq(r) := \{(\leq n.r) \mid n \in \mathbf{N}\}$,
- $N_\geq(r) := \{(\geq n.r) \mid n \in \mathbf{N}\}$ *and*
- $N_p := N_C \cup \{\neg A \mid A \in N_C\}$.

Let $n \in \mathbb{N}_{>0}$, $C_1, \dots C_n$ be concept descriptions with

$$\{C_1, \dots, C_n\} \subseteq N_\forall \cup N_\exists \cup N_\leq \cup N_\geq \cup N_p,$$

$C := \bigsqcap_{i \leq n} C_i$ and $r \in N_r$. We define

- $prim(C) := \{C_1, \dots, C_n\} \cap N_p$,
- $ex_r(C) := \{D \in \mathcal{C}(\mathcal{ALCN}) \mid \exists r.D \in \{C_1, \dots, C_n\}\}$,
- $val_r(C) := \begin{cases} \top & \text{if } \{C_1, \dots, C_n\} \cap N_\forall(r) = \emptyset \\ \bigsqcap_{\forall r.D \in \{C_1,\dots,C_n\}} D & \text{otherwise} \end{cases}$
- $min_r(C) := max\{n \in \mathbf{N} \mid C \sqsubseteq (\geq n.r)\}$
- $max_r(C) := \begin{cases} min\{n \in \mathbf{N} \mid C \sqsubseteq (\leq n.r)\} & \text{if } \{n \in \mathbf{N} \mid C \sqsubseteq (\leq n.r)\} \neq \emptyset \\ \infty & \text{othwerwise} \end{cases}$

The $\mathcal{ALCN}$ normal form is as follows.

Definition 27 ($\mathcal{ALCN}$ Normal Form)**.** *Let $C \in \mathcal{C}(\mathcal{ALCN})$. C is in* normal form *iff $C = \bot$, $C = \top$ or C is of the form $C = \bigsqcup_{i \leq n} C_i$ with*

$$C_i := \bigsqcap_{D \in prim(C_i)} D \sqcap \bigsqcap_{r \in N_r} \Big(\bigsqcap_{D \in ex_r(C_i)} \exists r.D \sqcap \forall r.val_r(C_i) \sqcap (\leq max_r.C_i) \sqcap (\geq min_r.C_i)\Big)$$

where $val_r(C_i)$ and all $D \in ex_r(C_i)$ are in $\mathcal{ALCN}$-Normal Form. We also call C_i to be in conjunctive normal form*(CNF).*

Note that this normal form is not unique. For example $A \equiv (A \sqcap B) \sqcup (A \sqcap \neg B)$ and $\exists r.A \sqcap \forall r.B \equiv \exists r(A \sqcap B) \sqcap \forall r.B$ where all concept descriptions are in normal form. The restriction to the description logic $\mathcal{ALN}$ leads to the following normal form which is used by the measure defined in [FD06], Subsection 4.2.6.

Definition 28 ($\mathcal{ALN}$ Normal Form)**.** *Let $C \in \mathcal{C}(\mathcal{ALN})$. C is in normal form iff $C = \bot$, $C = \top$ or C is of the form*

$$C := \bigsqcap_{D \in prim(C)} D \sqcap \bigsqcap_{r \in N_r} (\bigsqcap_{D \in ex_r(C)} \exists r.D \sqcap \forall r.val_r(C))$$

where $val_r(C)$ and all $D \in ex_r(C)$ are in $\mathcal{ALN}$-Normal Form.

4.2 Structural Measure

Here we present six structural measures. The Jaccard Index, *Dice* and the measures presented in [Jan08, FD06, JW09, dFE06].

4.2.1 The Jaccard Index

The Jaccard Index is an adaption of the set similarity measure with the same name ([Jac01]) where it regards $\mathcal{L}_0$ concept descriptions as sets of concept names . The first suggestion to use it as similarity measure for the GeneOntology was in [Gen07].

Definition 29 (Jaccard Index)**.** *Let $C, D \in \mathcal{C}(\mathcal{L}_0)$. The* Jaccard Index *is defined as*

$$Jacc(C, D) := \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C} \cup \widehat{D}|}.$$

As the following lemma shows, the Jaccard Index fulfills all properties we defined in Chapter 3.

Lemma 6. *The Jaccard Index is*

1. *symmetric,*
2. *equivalence closed,*
3. *equivalence invariant,*
4. *subsumption preserving,*
5. *reverse subsumption preserving,*
6. *structural dependent,*
7. *dissimilar closed,*

8. *bounded and*

9. *has the triangle inequality property.*

Proof. Let $C, D, E \in \mathcal{C}(\mathcal{L}_0)$ which implies that $\widehat{C}, \widehat{D}, \widehat{E} \subseteq N_C$.

1. The symmetry of the $Jacc$ is obvious.

2. Equivalence Closed: We have
$$Jacc(C, D) = 1 \iff |\widehat{C} \cap \widehat{D}| = |\widehat{C} \cup \widehat{D}| \iff \widehat{C} = \widehat{D} \iff C \equiv D.$$

3. Equivalence Invariant: If $C \equiv D$, then $\widehat{C} = \widehat{D}$ because there are no roles ($N_r = \emptyset$). This implies $Jacc(C, E) = Jacc(D, E)$.

4. Subsumption Preserving: Let $C \sqsubseteq D \sqsubseteq E$. This implies that $\widehat{E} \subseteq \widehat{D} \subseteq \widehat{C}$. Therefore, $|\widehat{E}| \leq |\widehat{D}|$. Also, $|\widehat{E} \cap \widehat{C}| = |\widehat{E}|$ and $|\widehat{D} \cap \widehat{C}| = |\widehat{D}|$. We start with $|\widehat{E}| \leq |\widehat{D}|$ which is equivalent to $|\widehat{E} \cap \widehat{C}| \leq |\widehat{D} \cap \widehat{C}|$ and leads us to
$$\frac{|\widehat{E} \cap \widehat{C}|}{|\widehat{C}|} \leq \frac{|\widehat{D} \cap \widehat{C}|}{|\widehat{C}|} \iff \frac{|\widehat{E} \cap \widehat{C}|}{|\widehat{E} \cup \widehat{C}|} \leq \frac{|\widehat{D} \cap \widehat{C}|}{|\widehat{D} \cup \widehat{C}|} \iff Jacc(\widehat{C}, \widehat{E}) \leq Jacc(\widehat{C}, \widehat{D}).$$

5. Reverse Subsumption Preserving: Let $C \sqsubseteq D \sqsubseteq E$. This implies that $\widehat{E} \subseteq \widehat{D} \subseteq \widehat{C}$. Since $|\widehat{D}| \leq |\widehat{C}|$ and $|\widehat{E} \cap \widehat{C}| = |\widehat{E} \cap \widehat{D}| = |\widehat{E}|$ we derive
$$Jacc(C, E) = \frac{|\widehat{E}|}{|\widehat{C}|} \leq \frac{|\widehat{E}|}{|\widehat{D}|} = Jacc(D, E).$$

6. Structural Dependent: Let $(C_n)_n$ be a sequence of pairwise different concept names,
$$D_n := \bigsqcap_{i \leq n} C_i \sqcap D$$
and
$$E_n := \bigsqcap_{i \leq n} C_i \sqcap E.$$
We have
$$1 \geq \lim_{n \to \infty} Jacc(D_n, E_n) \geq \lim_{n \longrightarrow \infty} \frac{n}{n + |\widehat{D}| + |\widehat{E}|} = 1$$
which implies
$$\lim_{n \to \infty} Jacc(D_n, E_n) = 1.$$

7. Dissimilar Closed: If $lcs(C, D) = \top$, then $|\widehat{C} \cap \widehat{D}| = 0$ which implies $Jacc(C, D) = 0$.

8. Bounded: If $lcs(C, D) \neq \top$, then $|\widehat{C} \cap \widehat{D}| > 0$ which implies $Jacc(C, D) > 0$.

9. Triangle Inequality: That the set version of Jaccard Index fulfills triangle inequality is proven in [Lip99]. Since our version is not significantly different, the proof can easily be adapted.

□

4.2.2 Dice's Coefficient

Dice's Coefficient is an adaption of the set similarity measure with the same name ([Dic45]). It is defined for the description logic $\mathcal{L}_0$. A $\mathcal{L}_0$ concept description is a conjunction of concept names. Viewing the concept description as a set of concept names one can apply Dice's Coefficient.

Definition 30 (Dice's Coefficient)**.** *Let* $C, D \in \mathcal{C}(\mathcal{L}_0)$*. The* Dice's Coefficient *is defined as*

$$Dice(C, D) := 2\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}| + |\widehat{D}|}.$$

As the following lemma shows, the Dice's Coefficient fulfills all properties we defined in Chapter 3 except the triangle inequality.

Lemma 7. *The Dice's Coefficient is*

1. *symmetric,*
2. *equivalence closed,*
3. *equivalence invariant,*
4. *subsumption preserving,*
5. *reverse subsumption preserving,*
6. *structural dependent,*
7. *dissimilar closed and*
8. *bounded.*

Proof. Let $C, D, E \in \mathcal{C}(\mathcal{L}_0)$ which implies that $\widehat{C}, \widehat{D}, \widehat{E} \subseteq N_C$.

1. The symmetry of $Dice$ is obvious.
2. Equivalence Closed: We have

$$Dice(C, D) = 1 \iff \frac{2 \cdot |\widehat{C} \cap \widehat{D}|}{|\widehat{C}| + |\widehat{D}|} \iff 2 \cdot |\widehat{C} \cap \widehat{D}| = |\widehat{C}| + |\widehat{D}|.$$

   Since $|\widehat{C} \cap \widehat{D}| \leq |\widehat{C}|, |\widehat{D}|$, the last statement is equivalent to

$$|\widehat{C} \cap \widehat{D}| = |\widehat{C}| \text{ and } |\widehat{C} \cap \widehat{D}| = |\widehat{D}| \iff \widehat{C} = \widehat{D} \iff C \equiv D.$$

3. Equivalence Invariant: If $C \equiv D$, then $\widehat{C} = \widehat{D}$ because there are no roles ($N_r = \emptyset$). This implies $Dice(C, E) = Dice(D, E)$.

4. Subsumption Preserving: Let $C \sqsubseteq D \sqsubseteq E$. This implies that $\widehat{E} \subseteq \widehat{D} \subseteq \widehat{C}$. Therefore, $|\widehat{E}| \leq |\widehat{D}|$. Also, $|\widehat{E} \cap \widehat{C}| = |\widehat{E}|$ and $|\widehat{D} \cap \widehat{C}| = |\widehat{D}|$. Starting with $|\widehat{E}| \leq |\widehat{D}|$, we derive

$$|\widehat{E}| \leq |\widehat{D}| \iff |\widehat{E}||\widehat{C}| \leq |\widehat{D}||\widehat{C}| \iff |\widehat{E}||\widehat{D}| + |\widehat{E}||\widehat{C}| \leq |\widehat{E}||\widehat{D}| + |\widehat{D}||\widehat{C}|$$

which is equivalent to

$$|\widehat{E}|(|\widehat{D}| + |\widehat{C}|) \leq |\widehat{D}|(|\widehat{E}| + |\widehat{C}|) \iff \frac{|\widehat{E}|}{|\widehat{E}| + |\widehat{C}|} \leq \frac{|\widehat{D}|}{|\widehat{D}| + |\widehat{C}|}$$

$$\iff \frac{2|\widehat{E} \cap \widehat{C}|}{|\widehat{E}| + |\widehat{C}|} \leq \frac{2|\widehat{D} \cap \widehat{C}|}{|\widehat{D}| + |\widehat{C}|} \iff \mathit{Dice}(\widehat{C}, \widehat{E}) \leq \mathit{Dice}(\widehat{C}, \widehat{D}).$$

5. Reverse Subsumption Preserving: Let $C \sqsubseteq D \sqsubseteq E$. This implies that $\widehat{E} \subseteq \widehat{D} \subseteq \widehat{C}$. Since $|\widehat{D}| \leq |\widehat{C}|$ and $|\widehat{E} \cap \widehat{C}| = |\widehat{E} \cap \widehat{D}| = |\widehat{E}|$ we derive

$$\mathit{Dice}(C, E) = \frac{2|\widehat{E}|}{|\widehat{C}| + |\widehat{E}|} \geq \frac{2|\widehat{E}|}{|\widehat{D}| + |\widehat{E}|} = \mathit{Dice}(D, E).$$

6. Structural Dependent: Let $(C_n)_n$ be a sequence of pairwise different concept names,

$$D_n := \bigsqcap_{i \leq n} C_i \sqcap D$$

and

$$E_n := \bigsqcap_{i \leq n} C_i \sqcap E.$$

We have

$$1 \geq \lim_{n \to \infty} \mathit{Dice}(D_n, E_n) \geq \lim_{n \longrightarrow \infty} \frac{2n}{2n + |\widehat{D}| + |\widehat{E}|} = 1$$

which implies

$$\lim_{n \to \infty} \mathit{Dice}(D_n, E_n) = 1.$$

7. Dissimilar Closed: If $\mathit{lcs}(C, D) = \top$, then $|\widehat{C} \cap \widehat{D}| = 0$ which implies $\mathit{Dice}(C, D) = 0$.

8. Bounded: If $\mathit{lcs}(C, D) \neq \top$, then $|\widehat{C} \cap \widehat{D}| > 0$ which implies $\mathit{Dice}(C, D) > 0$.

□

To show that Dice's Coefficient does not fulfill the **triangle inequality**, we use the following counterexample. Let $N_C := \{A, B\}$, $C := A \sqcap B$, $D := A$ and $E := B$. Then we have

$$
\begin{aligned}
1 = 1 + 2\frac{|\emptyset|}{|\{A\}| + |\{B\}|} = 1 + \mathit{Dice}(D, E) &< \mathit{Dice}(D, C) + \mathit{Dice}(C, E) \\
&= 2\frac{|\{A\}|}{|\{A\}| + |\{A, B\}|} + 2\frac{|\{B\}|}{|\{A, B\}| + |\{B\}|} \\
&= \frac{4}{3}.
\end{aligned}
$$

4.2.3 Computing Semantic Similarity Among Geographic Feature Types Represented in Expressive Description Logics [Jan08]

In the PhD thesis [Jan08], a measure called *simdl* is defined. The journal article can be found in [Jan06]. *Simdl* is defined for the description logic $\mathcal{ALCHQ}$ and it uses an adopted version of the $\mathcal{ALCN}$ normal form presented in Section 4.1.

Overall, we found it hard to understand how the measure is defined in detail. The reason is that *simdl* is mostly presented through text rather than precise mathematical formulation. To fill our gaps of knowledge, we analysed the source code of an implementation of *simdl* called *SimCat*. From this analysis and our understanding of [Jan08], we formulated a mathematical model of *simdl* which is presented below. The model is necessary to be able to analyse which properties hold. For simplicity, our model does not cover the full description logic for which the original *simdl* is defined. Our version works for $\mathcal{ALC}$-concept descriptions only. We found that this is already enough to disprove most of the properties.

We also noticed that the implementation does not respect the commutativity of conjunction, meaning that it is possible to construct an example such that $\mathit{simdl}(A \sqcap B, C) \neq \mathit{simdl}(B \sqcap A, C)$. We assume that this is simply a bug because such behaviour is not stated in [Jan08] (yet it is also not denied). Therefore, our version respects commutativity. Also, note that in the paper and the implementation, the measure has an asymmetric version. We analyse the symmetric version because we believe that symmetry is a vital property (see Chapter 3).

In order to define *simdl*, we present the notation of *valid* relations. Additionally, we expand the definition of the set of atoms of a concept description C in CNF (denoted as $\widehat{C}$) to include value restrictions.

Definition 31. *Let $C = \bigsqcap_{C' \in \widehat{C}} C'$ and $D = \bigsqcap_{D' \in \widehat{D}} D'$ be two concept descriptions where the $\widehat{C}$ and $\widehat{D}$ are atoms. A relation $A \subseteq \widehat{C} \times \widehat{D}$ is called* valid *iff it fulfills the properties*

1. $\forall C' \in \widehat{C}\ \forall D_1, D_2 \in \widehat{D} : (C', D_1), (C', D_2) \in A \implies D_1 = D_2$,
2. $\forall D' \in \widehat{D}\ \forall C_1, C_2 \in \widehat{C} : (C_1, D'), (C_2, D') \in A \implies C_1 = C_2$ *and*
3. $|A| = \min\{|\widehat{C}|, |\widehat{D}|\}$.

The set of all valid relations of C and D is denoted as $\mathfrak{A}(C, D)$.

Note that *simdl* does not expand the concept descriptions to measure in advance. The expansion is done “on the fly“ while measuring similarity. In the following definition, we use $exp_{\mathcal{T}}(C)$ to express the concept description we obtain by expanding the concept name C with respect to the TBox $\mathcal{T}$.

Definition 32 ([Jan08])**.** *Let $\mathcal{T}$ be a TBox and C and D be two concept descriptions in normal form, so $C := \bigsqcup_{i \leq n} C_i$, $D := \bigsqcup_{j \leq m} D_j$ where the C_i and D_j are conjunction of atoms. Additionally, let $S : N_B \longrightarrow 2^{\mathcal{C}(\mathcal{ALC})}$ be the function defined through for all $A \in N_B$:*

$$S(A) := \{E \in N_C \setminus \{A\} \mid E \sqsubseteq_{\mathcal{T}} A\}.$$

The function simdl : $\mathcal{C}(\mathcal{ALC})^2 \longrightarrow [0, 1]$ is defined as follows:

$$\begin{aligned}
simdl(C, D) &:= sim_{\sqcup}(C, D) \\
sim_{\sqcup}(C, D) &:= \max_{i \leq n, j \leq m} sim_{\sqcap}(C_i, D_j) \\
sim_{\sqcap}(C_i, D_j) &:= \max_{A \in \mathfrak{A}(C_i, D_j)} \frac{\sum_{(\widehat{C}, \widehat{D}) \in A} sim_a(C', D')}{|A|} \\
sim_a(C', D') &:= \begin{cases} 1 & \textit{if } C' \equiv D' \\ 0 & \textit{if } C' \equiv \neg D' \\ sim_R(r, s) \cdot sim_{\sqcup}(E, F) & \textit{if } C' = \exists r.E \textit{ and } D' = \exists s.F \\ sim_R(r, s) \cdot sim_{\sqcup}(E, F) & \textit{if } C' = \forall r.E \textit{ and } D' = \forall s.F \\ sim_P(C', D') & \textit{if } C', D \in N_B \\ sim_{\sqcap}(exp_{\mathcal{T}}(C'), exp_{\mathcal{T}}(D')) & \textit{if } C' \in N_C \setminus N_B \textit{ or } D \in N_C \setminus N_B \\ 0 & \textit{otherwise} \end{cases} \\
sim_R(r, s) &:= \begin{cases} 1 & \textit{if } r = s \\ 0 & \textit{otherwise} \end{cases} \\
sim_P(A, B) &:= \begin{cases} \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|} & \textit{if } |S(A) \cup S(B)| \neq 0 \\ 0 & \textit{otherwise.} \end{cases}
\end{aligned}$$

In the asymmetric case of *simdl*, the condition of the first case of the function sim_a is changed to $D' \sqsubseteq C'$ and the function $sim_{\sqcup}$ gets a second case where $sim_{\sqcup}(C, D) = 1$ if $D \sqsubseteq C$. Everything else remains the same.

The normal form is used together with a set of rewriting rules which are supposed to ensure that *simdl* is equivalence invariant. We noticed that to achieve this goal, at least two rules are missing,

$$C \sqcup D \longrightarrow D \text{ if } C \sqsubseteq D$$

and

$$\forall r.C \sqcap \exists r.D \longrightarrow \forall r.C \sqcap \exists r.(D \sqcap C).$$

Even with this rules, the resulting normal form is not unique and therefore *simdl* is not equivalence invariant.

Another problem are the valid relations. Every atom has at most one partner in a valid relation. If we have two concept description, $\exists r.(A \sqcap B)$ and $\exists r.A \sqcap \exists r.B$, then all valid relations have one element only. The measure cannot take into account that $\exists r.(A \sqcap B)$ is related to both $\exists r.A$ and $\exists r.B$. The "on the fly" expansion can also lead to unintuitive behaviour and a way to disprove equivalence invariant (see below).

In the remainder of the section, we prove that *simdl* is symmetric and structural dependent. Additionally we show that *simdl* is not equivalence closed, equivalence invariant, dissimilar closed, bounded, subsumption preserving, reverse subsumption preserving and it does not fulfill the triangle inequality.

Lemma 8. *The measure simdl is*

1. *symmetric and*
2. *structural dependent.*

Proof. 1. Symmetry of *simdl* is obvious because all involved functions, $sim_{\sqcup}$, $sim_{\sqcap}$, sim_P and sim_R are symmetric.

2. Let $(C_n)_n$ be a sequence of pairwise different atoms, E, D be concept descriptions in CNF,

$$D_n := \prod_{i \le n} C_i \sqcap D$$

and

$$E_n := \prod_{i \le n} C_i \sqcap E.$$

Since $sim_a(C_i, C_i) = 1$ for all $i \le n$ we know that

$$simdl(D_n, C_n) = sim_{\sqcap}(D_n, C_n) \ge \frac{n}{n + min\{|\widehat{C}|, |\widehat{D}|\}}$$

which implies

$$1 \ge \lim_{n \to \infty} simdl(D_n, C_n) \ge \lim_{n \to \infty} \frac{n}{n + min\{|\widehat{C}|, |\widehat{D}|\}} = 1$$

and therefore $\lim_{n\to\infty} simdl(D_n, C_n) = 1$.

□

- The measure is not **equivalence closed**. The reason is that at on the disjunction level, the maximum of all pairs of conjunction is chosen. Let C and D be concept descriptions in normal form and $A \in N_C$. Then

$$\begin{aligned} simdl(A \sqcup C, A \sqcup D) &= max\{sim_{\sqcap}(A, A), sim_{\sqcap}(A, D), sim_{\sqcap}(C, A), sim_{\sqcap}(C, D)\} \\ &= 1. \end{aligned}$$

Another way to disprove equivalence closed is as follows. Let $N_C := \{A, B, C\}$ and $\mathcal{T} := \{C \equiv A \sqcap B\}$ then

$$simdl(A, B) = sim_P(A, B) = \frac{|\{C\}|}{|\{C\}|} = 1$$

but $A \not\equiv B$. Note that this behaviour is intended by the author. He states that "The comparison of two primitives yields 1, if they cannot be differentiated" and the definition of sim_P expresses his expectation.
The definition of valid relations yields another way of disproving equivalence closed. Let N_C be as above and C be an arbitrary N_C in CNF with $A \not\sqsubseteq C$. Then we have

$$simdl(C \sqcap A, A) = sim_\sqcap(C \sqcap A, A) = \frac{sim_a(A, A)}{1} = 1$$

where $C \sqcap A \not\equiv A$.

- The measure is not **equivalence invariant** because the used normal form is not unique. Let $N_C := \{A, B, G\}$, $N_r := \{r\}$ and $\mathcal{T} := \emptyset$. Then

  $$simdl(\exists r.(B \sqcap G), \exists r.(B \sqcap A) \sqcap \forall r.A) = sim_\sqcap(B \sqcap G, B \sqcap A) = \frac{1}{2}$$

  and

  $$simdl(\exists r.(B \sqcap G), \exists r.B \sqcap \forall r.A) = sim_\sqcap(B \sqcap G, B) = 1$$

  but $\exists r.(B \sqcap A) \sqcap \forall r.A \equiv \exists r.B \sqcap \forall r.A$.
  The "on the fly" expansion yield another way to disprove equivalence invariant. Let $N_C := \{A, B_1, B_2, B_3, C\}$ and $\mathcal{T} := \{C \equiv B_1 \sqcap B_2 \sqcap B_3\}$. Then

  $$simdl(A, B_1) = simdl(A, B_2) = simdl(A, B_3) = 0.$$

  That implies

  $$\begin{aligned} simdl(A \sqcap C, B_1 \sqcap B_2 \sqcap B_3) &= \frac{\max_{i \leq 3} sim_\sqcap(A, B_i) + \max_{i \leq 3} sim_\sqcap(B_1 \sqcap B_2 \sqcap B_3, B_i)}{2} \\ &= \frac{0+1}{2} = \frac{1}{2} \\ &\neq simdl(A \sqcap B_1 \sqcap B_2 \sqcap B_3, B_1 \sqcap B_2 \sqcap B_3) = 1 \end{aligned}$$

  which contradicts equivalence invariant.

- The measure is not **dissimilar closed**. Let $N_C := \{A, B, C\}$ and $\mathcal{T} := \{C \equiv B \sqcap \neg A\}$. We have $lcs(A \sqcap B, \neg A \sqcup \neg B) = \top$ yet

  $$\begin{aligned} simdl(A \sqcap B, \neg A \sqcup \neg B) &= max\{sim_\sqcap(A \sqcap B, \neg A), sim_\sqcap(A \sqcap B, \neg B)\} \\ &= max\{1, 0\} = 1. \end{aligned}$$

- The measure is not **bounded**. Let $N_C := \{A, B, G\}$ and $\mathcal{T} := \emptyset$ then $lcs(A \sqcap B, A \sqcap G) \equiv A$ yet $simdl(A \sqcap B, A \sqcap G) = 0$.

- The measure is not **subsumption preserving**. Let $N_C := \{A, B, G\}$, $N_r := \{r\}$ and $\mathcal{T} := \emptyset$. Using the aberrations

  $$\begin{aligned} C &:= \exists r.(B \sqcap G) \sqcap \forall r.A \sqcap A, \\ D &:= \exists r.(B \sqcap A) \sqcap \forall r.A \sqcap A \end{aligned}$$

  we have $C \sqsubseteq D \sqsubseteq A$ but

  $$simdl(C, D) = \frac{\frac{1}{2} + 1 + 1}{3} = \frac{5}{6} < simdl(C, A) = 1.$$

- The measure is not **reverse subsumption preserving**. Let $N_C := \{A, B, G\}$, $N_r := \{r\}$ and $\mathcal{T} := \emptyset$. Using the aberrations

  $$\begin{aligned} C &:= \exists r.(B \sqcap G) \sqcap \forall r.A \sqcap \exists r.(B \sqcap A), \\ D &:= \exists r.(B \sqcap G) \sqcap \forall r.A, \\ E &:= \exists r.(B \sqcap A) \sqcap \forall r.A \end{aligned}$$

  we have $C \sqsubseteq D \sqsubseteq E$ where C, D and E are in CNF but

  $$simdl(C, E) = 1 > simdl(D, E) = \frac{\frac{1}{2} + 1 + 1}{3} = \frac{5}{6}.$$

- *Simdl* does not fulfill the **triangle inequality**. Let $N_C := \{A, B, G\}$ and $\mathcal{T} := \emptyset$. We have

  $$1 + simdl(A \sqcap B, A \sqcap G) = \frac{3}{2} < simdl(A \sqcap B, A) + simdl(A, A \sqcap G) = 2$$

  which contradicts triangle inequality.

4.2.4 $SIM - DL_A$: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-Concept to Inter-Instance Similarity [JW09]

The measure $simdl_A$ is defined for the description logic $\mathcal{SHI}$ which is like $\mathcal{ALCH}$ plus it also allows for inverse and transitive roles. As normal form it uses the *negation normal form* (NNF). A concept description is in NNF if all occurring negations are atomic negations.

The measure is not presented with a full mathematical model. Instead, a mixture of mathematical syntax and text is used to informally describe the behaviour of $simdl_A$. The textual description is not always precise. For example, it is unclear what the result of the concepts A and $\exists r.A$ is. To fill our gaps of understanding, we analysed the implementation from the *SimCat* project. There we found that the result of A and $\exists r.A$ is 0. However, we omit presenting a full mathematical model for simdl_A. The reason is that such a model would force us to present a lot of new syntax which does not provide much gain of knowledge since the measure does not fulfill most of our properties. Here we only present a short description of how simdl_A works. For further details we refer to [JW09].

The measure has three stages. First, both concept descriptions are transformed in NNF and a changed version of the tableau algorithm is used to generate a completion tree for each concept description. In contrast to the original tableau algorithm, the $\sqcup$-rule is modified and another $\forall$-rule is added. In the next stage, the completion tree is used to generate a set of so called *proxy models*. A proxy model is a tree where the edges are labelled with role names and the nodes are labelled with sets of concept names. Finally, both sets of proxy models are used to measure the similarity. This is done by measuring the similarity of all pairs of proxy models (using tree similarity) and then either choosing the maximum, minimum or computing the average of all results.

We noticed that the measure is not well defined. The problem is that while building the completion tree, there are cases where it is possible to apply two different tableau rules at the same time which leads to different proxy models. For example, for $C := \forall r.A \sqcap \exists r.B$, after using the $\sqcap$-rule, two rules can be applied, the $\exists$-rule and the new $\forall$-rule (which can be applied if there is no r-neighbour). Therefore, the result of the measuring depends on the order of how the rules are applied. In the paper, no order is specified and in the implementations an order is chosen without comment.

In the following we present examples to show that simdl_A is not equivalence closed, equivalence invariant, subsumption preserving, reverse subsumption preserving, dissimilar closed, bounded and it does not fulfill the triangle inequality. However, the measure is symmetric (but has also an asymmetric version) and we assume that it is structural dependent. Yet, because of the absence of a mathematical model, we do not prove this here.

- The measure is not **equivalence closed** in all versions, symmetric or asymmetric and minimum or maximum approach. Let $N_C := \{A, B\}$ and $N_r := \{r\}$. We have

  $$\mathit{simdl}_A(A \sqcap \exists r.B, (A \sqcap \exists r.B) \sqcup (A \sqcap \forall r.\neg B)) = 1$$

  independently from the version, but $A \sqcap \exists r.B \not\equiv (A \sqcap \exists r.B) \sqcup (A \sqcap \forall r.\neg B) \equiv A$.

- The measure is not **equivalence invariant**. Let $N_C := \{A, B\}$, $N_r := \{r\}$, $C := A$, $D := (A \sqcap \exists r.B) \sqcup (A \sqcap \forall r.\neg B)$, $E := A \sqcap \exists r.B$. The concept descriptions C, D and E are in normal form and $C \equiv D$. If we use the maximum approach for the Model Level Matrix then

  $$\mathit{simdl}_A(C, E) = 0.75 < \mathit{simdl}_A(D, E) = 1.$$

- The measure is not **subsumption preserving**. Let $N_C := \{A, B\}$ and $N_r := \{r\}$. Additionally, let $C := A \sqcap \exists r.B$, $D := A$ and $E := (A \sqcap \exists r.B) \sqcup (A \sqcap \forall r.\neg B)$. Then

$C \sqsubseteq D \sqsubseteq E$ and

$$simdl_A(C, D) = 0.75 < simdl_A(C, E) = 1.$$

- The measure is not **triangle inequality**. Let $N_C := \{A, B, G\}$ then

$$1 + simdl_A(A \sqcap G, B \sqcap G) = \frac{3}{2} < simdl_A(A \sqcap G, G) + simdl_A(G, B \sqcap G) = 2.$$

- The measure is not **reverse subsumption preserving**. Let $N_C := \{A, B\}$, $N_r := \{r\}$ and we use the abbreviations $C := A \sqcap \exists r.(A \sqcap B)$, $D := A \sqcap \exists r.A \sqcap \exists r.B$ and $E := A$. Then we have $C \sqsubseteq D \sqsubseteq E$ but $simdl_A(C, E) = \frac{1}{2} > simdl_A(D, E) = \frac{1}{3}$.
- The measures is not **dissimilar closed**. Let $N_C := \{A\}$. Then $simdl_A(A, \not A) = 0.5$ but $lcs(A, \not A) \equiv \top$.

4.2.5 A Dissimilarity Measure for $\mathcal{ALC}$ Concept Descriptions [dFE06]

This measure is defined for the description logic $\mathcal{ALC}$. Before two concepts are measured, they are converted into the $\mathcal{ALC}$ normal form presented in Section 4.1 Definition 27. To evaluate primitive concepts it uses the *Information Content* that is depending on the canonical interpretation $\mathcal{I}_\mathcal{A}$.

Definition 33 (Information Content, *IC*, [dFE06])**.** *Let $\mathcal{A}$ be an ABox and $\mathcal{I}_\mathcal{A}$ be the canonical interpretation. For $C \in \mathcal{C}(\mathcal{ALC})$ we define*

$$pr(C) := \frac{|C^{\mathcal{I}_\mathcal{A}}|}{|\Delta^{\mathcal{I}_\mathcal{A}}|}.$$

The Information Content *is defined as*

$$IC(C) := -log\, pr(C).$$

We noticed that the function “information content“ is not well defined. It is unclear what the result of $pr(C)$ is if $|\Delta^{\mathcal{I}_\mathcal{A}}| = 0$. Additionally, if $|C^{\mathcal{I}_\mathcal{A}}| = 0$ then $pr(C) = 0$ (if $|\Delta^{\mathcal{I}_\mathcal{A}}| \neq 0$) but the logarithm is defined only for values greater than 0. In such a case, we assume that $IC(C) = \infty$ since infinity is used also in the definition of the measure.

Definition 34 ([dFE06])**.** *Let $\mathcal{A}$ be an ABox and $\mathcal{I}_\mathcal{A}$ be the canonical interpretation. The similarity measure sim is a function $sim : \mathcal{C}(\mathcal{ALC})^2 \longrightarrow [0, 1]$ defined as follows. Let C and D be $\mathcal{ALC}$ concept descriptions in normal form, so $C = \bigsqcup_{i=1}^{n} C_i$, $D = \bigsqcup_{j=1}^{m} D_j$*

then

$$sim(C,D) := \begin{cases} 1 & \textit{if } d_\sqcup(C,D) = 0 \\ 0 & \textit{if } d_\sqcup(C,D) = \infty \\ \frac{1}{d_\sqcup(C,D)} & \textit{otherwise} \end{cases}$$

$$d_\sqcup(C,D) := \begin{cases} 0 & \textit{if } C \equiv D \\ \infty & \textit{if } C \sqcap D \equiv \bot \\ \max\limits_{\substack{i \le n \\ j \le m}} d_\sqcap(C_i, D_j) & \textit{otherwise} \end{cases}$$

$$d_\sqcap(C_i, D_j) := d_P(C_i, D_j) + d_\forall(C_i, D_j) + d_\exists(C_i, D_j)$$

$$d_P(C_i, D_j) := \begin{cases} \infty & \textit{if } prim(C_i) \sqcap prim(D_j) \equiv \bot \\ \frac{IC(prim(C_i) \sqcap prim(D_j))+1}{IC(lcs(prim(C_i), prim(D_j)))+1} & \textit{otherwise} \end{cases}$$

$$d_\forall(C_i, D_j) := \sum_{r \in N_r} d_\sqcup(val_r(C_i), val_r(D_j))$$

$$d_\exists(C_i, D_j) := \sum_{r \in N_r} d_R(r, C_i, D_j)$$

$$d_R(r, C_i, D_j) := \begin{cases} \sum\limits_{C' \in ex_r(C_i)} \max\limits_{D' \in ex_r(D_j)} d_\sqcup(C', D') & \textit{if } |ex_r(C_i)| \ge |ex_r(D_i)| \\ \sum\limits_{D' \in ex_r(D_j)} \max\limits_{C' \in ex_r(C_i)} d_\sqcup(D', C') & \textit{otherwise} \end{cases}$$

We noticed that the measure is not well defined. First, the function *prim* is defined as the set of primitive concepts on the current role level. However, in the definition it is treated as the conjunction of this concepts. This raises the question of what happens if there are no primitive concepts. Usually, the empty conjunction is interpreted as $\top$ and we assume that the same is true here. Also, if one argument has no existential restriction and the other has, then we have to build the maximum over an empty set in function d_R. Since in general this is assumed to be 0, we do the same. Another point is that the similarity of $\bot$ and $\bot$ is double-defined. We have $d_\sqcup(\bot, \bot) = 0$ because $\bot \equiv \bot$ and $d_\sqcup(\bot, \bot) = \infty$ because $\bot \sqcap \bot \equiv \bot$. Additionally, since the information content is not defined for concept descriptions with an empty canonical interpretation, it is unclear what the value of $sim(A, B)$ with $(A \sqcap B)^{\mathcal{I}_\mathcal{A}} = \emptyset$ is. As we wrote above, we assume that the result is ∞.

We found that an overall problem with these measures is that the value of the function

$d_\sqcup$ is the sum of all comparisons from all levels. If at one level, the value is infinite then the result at the top level is infinite as well. Another problem arises when one argument has no existential restriction and the other has. Since the empty maximum is 0, the similarity is 1. For example $sim(\top, \exists r, A) = 1$. Note that the example also uses the fact that if *prim* is empty, then it is interpreted as $\top$. Using these problems, we create counterexamples for all of the properties defined in Chapter 3. This includes symmetry which was part of the definition of similarity measures.

- In [dFE06], Proposition 4.1, it is claimed that the measure is **symmetric**. However we found that this claim is incorrect. The problem lies in the definition of $d_\exists$ as the following counterexample illustrates. Let $N_C := \{A, B, G\}$, $N_r := \{r\}$, $\mathcal{A} := \{A(x), A(y), B(x), B(z), G(x), G(y), G(z)\}$, $C := \exists r.A \sqcap \exists r.B$ and $D := \exists r.A \sqcap \exists r.G$. First we observe that $IC(lcs(A, B)) = IC(lcs(A, G)) = IC(lcs(B, G)) = 0$. We have
$$\begin{aligned} d_\sqcup(C, D) = d_\exists(C, D) &= \\ &= max\{0, IC(A \sqcap G) + 1\} + max\{IC(B \sqcap A) + 1, IC(B \sqcap G) + 1\} \\ &= (-log(2/3) + 1) + (-log(2/3) + 1) \end{aligned}$$
where as
$$\begin{aligned} d_\sqcup(D, C) = d_\exists(D, C) &= \\ &= max\{0, IC(A \sqcap B) + 1\} + max\{IC(G \sqcap A) + 1, IC(G \sqcap B) + 1\} \\ &= (-log(1/3) + 1) + (-log(2/3) + 1) \\ &< sim(C, D). \end{aligned}$$

- The measure is not **equivalence invariant**. The problem lies in the fact that the normal form is not unique. It does not compensate for de'Morgan operations. We present a counterexample. Let $N_C := \{A, B, G\}$ and
$$\mathcal{A} := \{A(x), A(y), B(x), B(z), G(x)\}.$$
We have $A \equiv (A \sqcap B) \sqcup (A \sqcap \neg B)$. However,
$$\begin{aligned} d_\sqcup(G, (A \sqcap B) \sqcup (A \sqcap \neg B)) &= max\{d_p(G, A \sqcap B), d_p(G, A \sqcap \neg B)\} \\ &= max\{\frac{-log(1/3) + 1}{-log(1/3) + 1}, \infty\} \\ &= \infty \\ &\neq d_\sqcup(G, A) = \frac{-log(1/3) + 1}{-log(2/3) + 1}. \end{aligned}$$

- The measure does not fulfill **triangle inequality**. Using the example presented above and the definitions $C := A, D := G$ and $E := (A \sqcap B) \sqcup (A \sqcap \neg B)$ we have
$$1 + sim(D, E) = 1 + 0 = 1 < sim(D, C) + sim(C, E) = \frac{-log(2/3) + 1}{-log(1/3) + 1} + 1$$

- The measure is not **equivalence closed**. If $C \equiv D$ then by definition $sim(C, D) = 1$. However, the other direction is not true. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(x)\}$. Then

  $$sim(A, B) = \frac{1}{d_{\sqcup}(A, B)} = \frac{1}{d_P(A, B)} = \frac{IC(lcs(A, B))}{IC(A \sqcap B)} = 1$$

  but $A \not\equiv B$.

- The measure is not **subsumption preserving**. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(x)\}$. Additionally, we use the aberrations $C := A \sqcap B$, $D := (A \sqcap B) \sqcup (A \sqcap \neg B)$ and $E := A$. We have $C \sqsubseteq D \sqsubseteq E$. However

  $$\begin{aligned} d_{\sqcup}(C, D) &= max\{d_P(A \sqcap B, A \sqcap B), d_P(A \sqcap B, A \sqcap \neg B)\} \\ &= max\{1, \infty\} \\ &= \infty \end{aligned}$$

  and therefore $sim(C, D) = 0$ where as

  $$d_{\sqcup}(C, E) = d_P(A \sqcap B, A) = \frac{IC(A \sqcap B) + 1}{IC(A) + 1} = 1$$

  which implies $sim(C, E) = 1$.

- The measure is not **reverse subsumption preserving**. Let $N_C := \{A, B\}$, $N_r := \{r, s\}$ and $\mathcal{A} := \{A(x), A(y), B(x)\}$. Additionally, we define the concept descriptions

  $$\begin{aligned} &C := \forall r.B \sqcap \exists r.A \sqcap \exists s.(A \sqcap B), \\ &D := \forall r.B \sqcap \exists r.(A \sqcap B) \sqcap \exists s.A, \\ &E := \forall r.B \sqcap \exists r.A. \end{aligned}$$

  Since $\exists r.(A \sqcap B) \sqcap \forall r.B \equiv \exists r.A \sqcap \forall r.B$,we have $C \sqsubseteq D \sqsubseteq E$,

  $$\begin{aligned} d_{\sqcup}(E, D) &= d_{\sqcap}(E, D) \\ &= d_P(E, D) + d_{\forall}(E, D) + d_{\exists}(E, D) \\ &= d_P(E, D) + d_{\forall}(E, D) + d_R(r, E, D) + d_R(s, E, D) \\ &= 0 + 0 + (-log(\frac{1}{2}) + 1) + 0 = -log(\frac{1}{2}) + 1 \end{aligned}$$

  and

  $$\begin{aligned} d_{\sqcup}(E, C) &= d_{\sqcap}(E, C) \\ &= d_P(E, C) + d_{\forall}(E, C) + d_{\exists}(E, C) \\ &= d_P(E, C) + d_{\forall}(E, C) + d_R(r, E, C) + d_R(s, E, C) \\ &= 0 + 0 + 0 + 0 = 0. \end{aligned}$$

Therefore, we have

$$sim(E, C) = 1 > sim(E, D) = \frac{1}{-log(\frac{1}{2}) + 1}$$

which contradicts reverse subsumption preserving.

- The measure is not **bounded**. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(x)\}$. Then $lcs(A \sqcap B, A \sqcap \neg B) = A \not\equiv \top$ but

$$d_{\sqcup}(A \sqcap B, A \sqcap \neg B) = d_P(A \sqcap B, A \sqcap \neg B) = \infty$$

which implies $sim(A \sqcap B, A \sqcap \neg B) = 0$.

- The measure is not **structural dependent**. Let $A_1, \ldots A_n, B$ be arbitrary concept names, and r be a role name. The similarity of the concept descriptions

$$C := \prod_{i \leq n} A_i \sqcap \exists r.B$$
$$D := \prod_{i \leq n} A_i \sqcap \exists r.\neg B.$$

is always 0 because $d_P(B, \neg B) = \infty$.

- The measure is not **dissimilar closed**. Let $N_C := \{A, B\}$, $N_r := \{r\}$ and $\mathcal{A} := \{A(y), B(x)\}$. Using the definitions $C := \exists r A$ and $D := \forall r.(\neg A \sqcup B)$ we have $C \not\equiv \top$ and $D \not\equiv \top$. Additionally, we show that $lcs(C, D) = C \sqcup D \equiv \top$. First, we observe that $D \equiv \neg(\exists r.(A \sqcap \neg B))$. Let $\mathcal{I}'$ be an arbitrary interpretation and $a \in \Delta^{\mathcal{I}'}$. If $a \notin (\exists r.A)^{\mathcal{I}'}$ then $\not\exists b \in A^{\mathcal{I}'} : (a, b) \in r^{\mathcal{I}'}$ which implies $\not\exists b \in (A \sqcap \neg B)^{\mathcal{I}'} : (a, b) \in r^{\mathcal{I}'}$ and finally $a \notin (\exists r.(A \sqcap \neg B))^{\mathcal{I}'}$. Therefore $lcs(C, D) \equiv \top$. Since $C \sqcap D \not\equiv \bot$ we have

$$\begin{aligned} d_{\sqcup}(C, D) &= d_P(C, D) + d_{\forall}(C, D) + d_{\exists}(C, D) \\ &= 0 + d_{\sqcup}(val_r(C), val_r(D)) + 0 \\ &= max\{d_P(\top, \neg A), d_P(\top, B)\} = -log(1/2) + 1 \end{aligned}$$

and therefore $sim(C, D) = \frac{1}{-log(1/2)+1} \neq 0$.

4.2.6 A Similarity Measure for the $\mathcal{ALN}$ Description Logic [FD06]

This measure is defined for $\mathcal{ALN}$ and it uses the $\mathcal{ALC}$ normal form presented in Section 4.1. To measure primitive concepts it uses the canonical interpretation. Therefore, it depends on the existence of an ABox.

Definition 35. *Let $\mathcal{A}$ be an ABox, $\mathcal{I}_{\mathcal{A}}$ its canonical interpretation and $\lambda \in]0, \frac{1}{3}]$ The similarity measure measure $sim : \mathcal{C}(\mathcal{ALN})^2 \longrightarrow [0,1]$ is defined as follows,*

$$\begin{aligned} sim(C,D) := \lambda[&sim_P(C,D) + \frac{1}{|N_r|}\sum_{r \in N_r} sim(val_r(C), val_r(D)) \\ &+ \frac{1}{|N_r|}\sum_{r \in N_r} sim_N((min_r(C), max_r(C)), (min_r(D), max_r(D)))] \end{aligned}$$

$$sim_P(C,D) := \frac{|\bigcap\limits_{B \in prim(C)} B^{\mathcal{I}_{\mathcal{A}}} \cap \bigcap\limits_{B \in prim(D)} B^{\mathcal{I}_{\mathcal{A}}}|}{|\bigcap\limits_{B \in prim(C)} B^{\mathcal{I}_{\mathcal{A}}} \cup \bigcap\limits_{B \in prim(D)} B^{\mathcal{I}_{\mathcal{A}}}|}$$

$$sim_N((i_C, a_C), (i_D, a_D)) := \begin{cases} \frac{min\{a_C, a_D\} - max\{i_C, i_D\} + 1}{max\{a_C, a_D\} - min\{i_C, i_D\} + 1} & \textit{if } min\{a_C, a_D\} > max\{i_C, i_D\} \\ 0 & \textit{otherwise.} \end{cases}$$

The measure is not well defined. First, the case $N_r = \emptyset$ is not covered. Additionally, the value of sim_P is unclear if $|\bigcap_{B \in prim(C)} B^{\mathcal{I}_{\mathcal{A}}} \cup \bigcap_{B \in prim(D)} B^{\mathcal{I}_{\mathcal{A}}}| = 0$. Since in this case $|\bigcap_{B \in prim(C)} B^{\mathcal{I}_{\mathcal{A}}} \cap \bigcap_{B \in prim(D)} B^{\mathcal{I}_{\mathcal{A}}}| = 0$ as well, we assume that $sim_P(C,D) = 1$. Also, since $max_r(C)$ can be ∞ we need to expand the arithmetic of real numbers to include ∞ in order to compute sim_N. Finally, the similarity value of any two concept descriptions is unknown because the computation of this value never terminates. It ends up in an infinite recursive loop. The reason is that for a concept description C with no value restrictions which uses the role r, $val_r(C)$ is defined to be $\top$. For example, let $N_C := \emptyset$ and $N_r := \{r\}$ then

$$\begin{aligned} sim(\top,\top) &= \lambda[sim_P(\top,\top) + \frac{1}{1}sim(val_r(\top), val_r(\top)) + sim_N((0,\infty),(0,\infty))] \\ &= \lambda[1 + sim(\top,\top) + 1] \\ &= \lambda[2 + sim_P(\top,\top) + \frac{1}{1}sim(val_r(\top), val_r(\top)) + sim_N((0,\infty),(0,\infty))] \\ &= \ldots. \end{aligned}$$

In the following, we assume that the definition is extended with the missing base-case $sim(\top,\top) := 1$.

In our investigation of the properties, we assume that the free parameter λ is $\frac{1}{3}$. The first problem we noticed is that the result depends on the number of role names, even the one that do not appear anywhere. For example for $N_r := \{r\}$ we have $sim(A,B) = \frac{1}{3}[sim_P(A,B) + 1 + 1]$ whereas for $N_r := \{r, s\}$ we have $sim(A,B) = \frac{1}{3}[sim_P(A,B) + 1 + 1 + 1 + 1]$. This is unintuitive because the measure should depend only on information provided by the concept descriptions.

Another point is that the measure uses the canonical interpretation to measure concept names. Therefore, it cannot distinguish between concept names that have the same interpretation which implies that the measure is not equivalence closed.

We also found it problematic that all concept names are measures together in one function. If we measure $\bigsqcap_{i\leq n} A_i \sqcap \forall r.A$ and $\bigsqcap_{i\leq n} A_i \sqcap \forall s.B$ then the similarity is independent form n. This is contradictory to the perspective that every atom is regarded as a feature and that the similarity should increase the more features two concept descriptions share.

In the following lemma we prove that the measure is symmetric, equivalence invariant and subsumption preserving. Additionally, we provide counterexamples to show that the properties equivalence closed, structural dependent, bounded, dissimilar closed and triangle inequality are not fulfilled.

Lemma 9. *The similarity measure sim is*

1. *symmetric,*
2. *equivalence invariant,*
3. *subsumption preserving,*
4. *reverse subsumption preserving.*

Proof.

1. It is easy to see that the functions sim_P and sim_N are symmetric and therefore, sim is symmetric as well.
2. The measure is equivalence invariant because the normal form used is unique with respect to commutativity and associativity.
3. Let $C, D, E \in \mathcal{C}(\mathcal{ALN})$ be in $\mathcal{ALC}$ normal form with $C \sqsubseteq D \sqsubseteq E$. First we prove that $C \sqsubseteq D$ implies
$$\bigcap_{B\in prim(C)} B^{\mathcal{I}_\mathcal{A}} \subseteq \bigcap_{B\in prim(D)} B^{\mathcal{I}_\mathcal{A}}. \tag{4.1}$$
In order to fulfill $C \sqsubseteq D$, there are only three cases possible.
   a) $prim(D) = \emptyset$. This implies that $\bigcap_{B\in prim(D)} B^{\mathcal{I}_\mathcal{A}} = \Delta^{\mathcal{I}_\mathcal{A}}$ which satisfies Equation 4.1.
   b) There exists a concept name A such that $A, \neg A \in prim(C)$. In this case we have $\bigcap_{B\in prim(C)} B^{\mathcal{I}_\mathcal{A}} = \emptyset$ which satisfies Equation 4.1.
   c) The final case is $prim(C) \subseteq prim(D)$ which satisfies Equation 4.1 trivially.

   Using the same arguments for E and D, we obtain
$$\bigcap_{B\in prim(C)} B^{\mathcal{I}_\mathcal{A}} \subseteq \bigcap_{B\in prim(D)} B^{\mathcal{I}_\mathcal{A}} \subseteq \bigcap_{B\in prim(E)} B^{\mathcal{I}_\mathcal{A}}.$$
   Since the function sim_P is related to the Jaccard Index and we have proven that the Jaccard Index is subsumption preserving, we can derive that $sim_P(C, D) \leq$

$sim_P(D, E)$.

For the next part of the proof, let r be an arbitrary role name. Also, we use the following abbreviations: $i_C := min_r(C)$, $a_C := max_r(C)$, $i_D = min_r(D)$, $a_D = max_r(D)$, $i_E = min_r(E)$ and $a_E = max_r(E)$. We are going to prove that the function sim_N respects subsumption preserving for an arbitrary role name r, meaning we show that

$$sim_N((i_C, a_C), (i_D, a_D)) \geq sim_N((i_C, a_C), (i_E, a_E)). \tag{4.2}$$

First we observe that $C \sqsubseteq D \sqsubseteq E$ implies $i_E \leq i_D \leq i_C$ and $a_C \leq a_D \leq a_E$. To prove Equation 4.2, we have to consider three cases.

The first case is $min\{a_C, a_E\} \leq max\{i_C, i_E\}$. This implies

$$sim_N((i_C, a_C), (i_E, a_E)) = 0$$

and Equation 4.2 is true.

The second case is $min\{a_C, a_D\} \leq max\{i_C, i_D\}$. Since $min\{a_C, a_D\} = a_C = min\{a_C, a_E\}$ and $max\{i_C, i_D\} = i_C = max\{i_C, i_E\}$ we know that

$$sim_N((i_C, a_C), (i_E, a_D)) = sim_N((i_C, a_C), (i_E, a_E)) = 0.$$

The final case is $min\{a_C, a_D\} > max\{i_C, i_D\}$ and $min\{a_C, a_E\} > max\{i_C, i_E\}$. This implies $a_E - i_E + 1 > 0$ and $a_D - i_D + 1 > 0$. The facts $i_E \leq i_D \leq i_C$ and $a_C \leq a_D \leq a_E$ allow us to derive

$$\begin{aligned}
a_E \geq a_D &\iff a_E - i_E + 1 \geq a_D - i_D + 1\\
&\iff \frac{1}{a_D - i_D + 1} \geq \frac{1}{a_E - i_E + 1}\\
&\iff \frac{a_C - i_C + 1}{a_D - i_D + 1} \geq \frac{a_C - i_C + 1}{a_E - i_E + 1}\\
&\iff \frac{min\{a_C, a_D\} - max\{i_C, i_D\} + 1}{max\{a_C, a_D\} - min\{i_C, i_D\} + 1} \geq \frac{min\{a_C, a_E\} - max\{i_C, i_E\} + 1}{max\{a_C, a_E\} - min\{i_C, i_E\} + 1}\\
&\iff sim_N((i_C, a_C), (i_D, a_D)) \geq sim_N((i_C, a_C), (i_E, a_E)).
\end{aligned}$$

So far, we have prove that the functions sim_P and sim_N respect subsumption preserving. These are the base case of a structural induction. The last step is the induction step, showing that the part of $simi$ which deals with value restrictions respect subsumption preserving, too. This can simply be derived from the two facts that $C \sqsubseteq D \sqsubseteq E$ implies that for all $r \in N_r : \forall r.val_r(C) \sqsubseteq \forall r.val_r(D) \sqsubseteq \forall r.val_r(E)$ which implies $val_r(C) \sqsubseteq val_r(D) \sqsubseteq val_r(E)$. Using the induction hypothesis we derive

$$sim(val_r(C), val_r(D)) \geq sim(val_r(C), val_r(E)).$$

Every part of the sum which defines sim respects subsumption preserving. Therefore, $sim(C, D) \geq sim(C, E)$.

4. We use the same assumptions and abbreviations as in the proof of subsumption preserving. From

$$\bigcap_{B \in prim(C)} B^{\mathcal{I}_\mathcal{A}} \subseteq \bigcap_{B \in prim(D)} B^{\mathcal{I}_\mathcal{A}} \subseteq \bigcap_{B \in prim(E)} B^{\mathcal{I}_\mathcal{A}}.$$

and the fact that the Jaccard Index is reverse subsumption preserving, we derive that the function sim_P is reverse subsumption preserving for every role name r.

To prove that sim_N is reverse subsumption preserving we use a similar approach as in the proof above. We have to show that

$$sim_N((i_C, a_C), (i_E, a_E)) \leq sim_N((i_D, a_D), (i_E, a_E)). \tag{4.3}$$

First we observe that $C \sqsubseteq D \sqsubseteq E$ implies $i_E \leq i_D \leq i_C$ and $a_C \leq a_D \leq a_E$. To prove Equation 4.3, we have to consider three cases.

The first case is $min\{a_C, a_E\} \leq max\{i_C, i_E\}$. This implies

$$sim_N((i_C, a_C), (i_E, a_E)) = 0$$

and Equation 4.3 is true.

The second case is $min\{a_D, a_E\} \leq max\{i_D, i_E\}$. The two facts $min\{a_C, a_E\} \leq min\{a_D, a_E\}$ and $max\{i_D, i_E\} \leq max\{i_C, i_E\}$ allow us to derive

$$min\{a_C, a_E\} \leq min\{a_D, a_E\} \leq max\{i_D, i_E\} \leq max\{i_C, i_E\}$$

and therefore

$$sim_N((i_C, a_C), (i_E, a_D)) = sim_N((i_C, a_C), (i_E, a_E)) = 0.$$

The final case is $min\{a_C, a_E\} > max\{i_C, i_E\}$ and $min\{a_D, a_E\} > max\{i_D, i_E\}$. This implies $a_E - i_E + 1 > 0$. The facts $i_E \leq i_D \leq i_C$ and $a_C \leq a_D \leq a_E$ allow us to derive

$$\begin{aligned}
a_C \leq a_D &\iff a_C - i_C + 1 \leq a_D - i_D + 1 \\
&\iff \frac{a_C - i_C + 1}{a_E - i_E + 1} \leq \frac{a_D - i_D + 1}{a_E - i_E + 1} \\
&\iff \frac{min\{a_C, a_E\} - max\{i_C, i_E\} + 1}{max\{a_C, a_E\} - min\{i_C, i_E\} + 1} \leq \frac{min\{a_D, a_E\} - max\{i_D, i_E\} + 1}{max\{a_D, a_E\} - min\{i_D, i_E\} + 1} \\
&\iff sim_N((i_C, a_C), (i_E, a_E)) \leq sim_N((i_D, a_D), (i_E, a_E)).
\end{aligned}$$

For the rest of the proof, we can use the same arguments as in the proof above deriving that for every role name r

$$sim(val_r(C), val_r(E)) \leq sim(val_r(D), val_r(E))$$

which finally implies $sim(C, E) \leq sim(D, E)$.

□

- The measure is not **equivalence closed**. *sim* evaluates the concept names using the canonical interpretation, therefore we can construct a counterexample. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(x)\}$. Then $sim(A, B) = 1$ but $A \not\equiv B$. Additionally, the other direction of equivalence closed is not true either. Let $C :=\leq 3.r \sqcap \geq 3.r$ then $s(C, C) = \frac{1}{3}[1 + 1 + 0] = \frac{2}{3}$.

- The measure is not **structural dependent**. Let $N_C := \{A_1, \dots A_n, B\}$, $N_r := \{r\}$ and $\mathcal{A}$ be and arbitrary ABox. The similarity of the concept descriptions

$$C_n := \bigsqcap_{i \leq n} A_i \sqcap \forall r.B$$

$$D_n := \bigsqcap_{i \leq n} A_i \sqcap \forall r.\neg B$$

is

$$sim(C_n, D_n) = \frac{1}{3}[1 + sim(B, \neg B) + 1] = \frac{1}{3}[1 + 0 + 1] = \frac{2}{3}$$

for all $n \geq 1$. Therefore, *sim* is not structural dependent.

- The measure is not **bounded**. Let $N_C := \{A, B\}$, $N_r := \{r\}$, $\mathcal{A} := \{A(x), B(y)\}$,

$$C := A \sqcap B \sqcap \forall r.A \sqcap (\geq 0.r) \sqcap (\leq 0.r)$$

and

$$D := B \sqcap \forall r.\neg A \sqcap (\geq 1.r) \sqcap (\leq 1.r).$$

We have

$$sim(C, D) = \frac{1}{3}[sim_P(C, D) + sim(A, \neg A) + sim_N((0, 0), (1, 1))] = \frac{1}{3}[0 + 0 + 0] = 0$$

but $lcs(C, D) \not\equiv \top$.

- The measure is not **dissimilar closed**. Let $N_C := \{A\}$ and $\mathcal{A} := \{A(x)\}$. Then $lcs(A, \neg A) \equiv \top$ but $sim(A, \neg A) = \frac{1}{3}[0 + 1 + 1] = \frac{2}{3}$.

- The measure does not fulfill **triangle inequality**. Let $N_C := \{A, B, G, F\}$, $N_r := \{r\}$, $\mathcal{A} := \{A(x), B(y), G(x), G(y), F(z), r(y, x)\}$,

$$C := G \sqcap (\geq 1.r) \sqcap (\leq 4.r),$$

$$D := A \sqcap \forall r.A \sqcap (\geq 1.r) \sqcap (\leq 2.r)$$

and

$$E := B \sqcap \forall r.\neg A \sqcap (\geq 3.r) \sqcap (\leq 4.r).$$

Then

$$\begin{aligned} 1 + sim(D, E) &= 1 + \frac{1}{3}[sim_P(D, E) + s(A, \neg A) + sim_N((1, 2), (3, 4))] \\ &= 1 + \frac{1}{3}[0 + 0 + 0] = 1 \end{aligned}$$

and

$$\begin{aligned} sim(D,C)+sim(C,E) &= \frac{1}{3}[sim_P(D,C)+sim(A,\top)+sim_N((1,2),(1,4))] \\ &+\frac{1}{3}[sim_P(C,D)+sim(\top,\neg A)+sim_N((3,4),(1,4))] \\ &= \frac{1}{3}[\frac{1}{2}+\frac{1}{2}+\frac{2-1+1}{4-1+1}]+\frac{1}{3}[\frac{1}{2}+\frac{2}{3}+\frac{4-3+1}{4-1+1}] \\ &= \frac{1}{2}+\frac{5}{9}=1+\frac{1}{18} \\ &> 1+sim(D,E) \end{aligned}$$

which contradicts triangle inequality.

4.3 Interpretation Based Measure

Here we present two interpretation based measures from [dFE05] and [dSF08].

4.3.1 A Semantic Similarity Measure for expressive Description Logics [dFE05]

The measure is defined for the description logic $\mathcal{ALC}$ and it uses the canonical interpretation. Therefore, it is easy to compute and it depends on a populated and representative domain.

Definition 36 ([dFE05])**.** *Let $\mathcal{A}$ be an ABox with canonical interpretation $\mathcal{I}_\mathcal{A}$. The semantic similarity measure sim is a function*

$$sim : \mathcal{C}(\mathcal{ALC})^2 \longrightarrow [0,1]$$

defined as follows

$$sim(C,D) := \frac{|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|+|D^{\mathcal{I}_\mathcal{A}}|-|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}\cdot \max\{\frac{|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|},\frac{|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|D^{\mathcal{I}_\mathcal{A}}|}\}.$$

The measure suffers from the problem that two concept names A and B are not distinguishable when their canonical interpretation is the same ($A^{\mathcal{I}_\mathcal{A}} = B^{\mathcal{I}_\mathcal{A}}$). Therefore, this measure is not equivalence closed. We find this unintuitive because the designer of the knowledge base intentionally used two different concept names. If he wants them to be totally similar, then he would have used only on concept name. Therefore, the measure should be able to distinguish the concept names.

Another problem arises when we measure $A \sqcap \bigsqcap_{i\leq n} B_i$ and $\neg A \sqcap \bigsqcap_{i\leq n} B_i$. The similarity is always zero, despite the fact that both concept descriptions share allot of features/atoms and only two atoms are totally different.

In the following lemma we prove that the measure is symmetric, equivalence invariant, subsumption preserving and reverse subsumption preserving. Additionally, we provide examples to show that the measure is not equivalence closed, bounded, dissimilar closed, structural dependent and does not fulfill triangle inequality.

Lemma 10. *The measure sim is*

1. *symmetric,*
2. *equivalence invariant,*
3. *subsumption preserving and*
4. *reverse subsumption preserving.*

Proof. Let $C, D, E \in \mathcal{C}(\mathcal{ALC})$.

1. The symmetry of *sim* is obvious.
2. Since the measure uses the canonical interpretation, it is equivalence invariant.
3. From $C \sqsubseteq D \sqsubseteq E$ we can derive that
   - $|(C \sqcap D)^{\mathcal{I}_\mathcal{A}}| = |C^{\mathcal{I}_\mathcal{A}}|$,
   - $|(C \sqcap E)^{\mathcal{I}_\mathcal{A}}| = |C^{\mathcal{I}_\mathcal{A}}|$,
   - $|D^{\mathcal{I}_\mathcal{A}}| \leq |E^{\mathcal{I}_\mathcal{A}}|$ and
   - $\max\{\frac{|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|}, \frac{|(C\sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|D^{\mathcal{I}_\mathcal{A}}|}\} = \max\{\frac{|(C\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|}, \frac{|(C\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|}\} = 1$.

   Therefore,
   $$sim(C, D) = \frac{|C^{\mathcal{I}_\mathcal{A}}|}{|D^{\mathcal{I}_\mathcal{A}}|} \geq \frac{|C^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|} = sim(C, E).$$
4. From $C \sqsubseteq D \sqsubseteq E$ we can derive that
   - $|(C \sqcap E)^{\mathcal{I}_\mathcal{A}}| = |C^{\mathcal{I}_\mathcal{A}}|$,
   - $|(D \sqcap E)^{\mathcal{I}_\mathcal{A}}| = |D^{\mathcal{I}_\mathcal{A}}|$,
   - $|C^{\mathcal{I}_\mathcal{A}}| \leq |D^{\mathcal{I}_\mathcal{A}}|$ and
   - $\max\{\frac{|(C\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|}, \frac{|(C\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|}\} = \max\{\frac{|(D\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|D^{\mathcal{I}_\mathcal{A}}|}, \frac{|(D\sqcap E)^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|}\} = 1$.

   Therefore,
   $$sim(E, D) = \frac{|D^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|} \geq \frac{|C^{\mathcal{I}_\mathcal{A}}|}{|E^{\mathcal{I}_\mathcal{A}}|} = sim(C, E).$$

□

- The measure is not **equivalence closed**. If two different atoms have the same evaluation under the canonical interpretation, then the measures cannot distinguish them anymore. The following counterexamples illustrates this. Let $N_C := \{A, B, G\}$, $N_r := \{r\}$ and $\mathcal{A} := \{A(x), B(y), r(x, y), G(y)\}$. Then $sim(A, \exists r.B) = 1$ but $A \not\equiv \exists r.B$.

- The measure is not **structural dependent** because if the intersection of the evaluation of two concept descriptions is empty, then the measure has a value of 0. Let $A_1, \dots A_n, B$ be arbitrary concept names. The similarity of the concept descriptions

$$
\begin{aligned}
C_n &:= \bigsqcap_{i \leq n} A_i \sqcap B \\
D_n &:= \bigsqcap_{i \leq n} A_i \sqcap \neg B
\end{aligned}
$$

  is 0 for all $n \geq 0$ because $B \sqcap \neg B \equiv \bot$ implies $|(C_n \sqcap D_n)^{\mathcal{I}_\mathcal{A}}| = 0$.

- The measure does not fulfill the **triangle inequality**. Also it looks related to the Jaccard Index, the additional term

$$
\max\{\frac{|(C \sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|C^{\mathcal{I}_\mathcal{A}}|}, \frac{|(C \sqcap D)^{\mathcal{I}_\mathcal{A}}|}{|D^{\mathcal{I}_\mathcal{A}}|}\}
$$

  makes it possible to construct a counterexample to the triangle inequality. Let $N_C := \{C, D, E\}$ and

$$
\mathcal{A} := \{C(x), C(y), C(z), D(x), D(y), E(x), E(z)\}.
$$

  Then

$$
\begin{aligned}
1 + sim(D, E) &= 1 + \frac{1}{2+2-1} \cdot \frac{1}{2} \\
&= 1 + \frac{1}{6}
\end{aligned}
$$

  and

$$
\begin{aligned}
sim(D, C) + sim(C, E) &= \frac{2}{3+2-2} \cdot \frac{2}{2} + \frac{2}{3+2-2} \cdot \frac{2}{2} \\
&= \frac{2}{3} + \frac{2}{3} \\
&= 1 + \frac{2}{6}
\end{aligned}
$$

  which implies $1 + sim(D, E) < sim(D, C) + sim(C, E)$.

- The measure is not **bounded**. Let $N_C := \{A\}$ and $\mathcal{A} := \emptyset$ then $sim(A, A) = 0$ but $lcs(A, A) \not\equiv \top$.

- The measure is not **dissimilar closed**. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(x)\}$ then $sim(A, B) = \frac{1}{1+1-1} \cdot max\{1, 1\} = 1$.

4.3.2 On the influence of description logics ontologies on conceptual similarity [dSF08]

This measure is defined for $\mathcal{ALE}$ using the canonical interpretation. Additionally, it uses the *good common subsumer*(GCS) [BST07]. Note that since we restrict our investigation to unfoldable TBoxes and expanded concept descriptions, the GCS is the same as the least common subsumer.

Definition 37. *Let $\mathcal{A}$ be an ABox and $\mathcal{I}_\mathcal{A}$ its canonical interpretation. The similarity measure $s : \mathcal{C}(\mathcal{ALE})^2 \longrightarrow [0,1]$ is defined as follows:*

$$s(C,D) := \frac{min\{|C^{\mathcal{I}_\mathcal{A}}|, |D^{\mathcal{I}_\mathcal{A}}|\}}{|GCS(C,D)^{\mathcal{I}_\mathcal{A}}|}(1 - \frac{|GCS(C,D)^{\mathcal{I}_\mathcal{A}}|}{|\Delta^{\mathcal{I}_\mathcal{A}}|}(1 - \frac{min\{|C^{\mathcal{I}_\mathcal{A}}|, |D^{\mathcal{I}_\mathcal{A}}|\}}{|GCS(C,D)^{\mathcal{I}_\mathcal{A}}|})).$$

A problem with this measure is that the cardinality of the canonical interpretation of the GCS does not depend on the number of common features/atoms the concept descriptions share. For example, if $C := \bigsqcap_{i \leq n} A_i \sqcap A$ and $D := \bigsqcap_{j \leq m} B_j \sqcap A$ where $A^{\mathcal{I}_\mathcal{A}} \subseteq A_i^{\mathcal{I}_\mathcal{A}}, B_j^{\mathcal{I}_\mathcal{A}}$ for all $i \leq n, j \leq m$ then

$$|GCS(C,D)^{\mathcal{I}_\mathcal{A}}| = |C^{\mathcal{I}_\mathcal{A}}| = |D^{\mathcal{I}_\mathcal{A}}| = |A^{\mathcal{I}_\mathcal{A}}|$$

and therefore $s(C,D) = 1$. If the atoms A_i and B_j are totally different to each other then this behaviour is unintuitive. We would expect that the similarity decrease with a growing n and m.

In the following lemma we prove that the measure is symmetric, equivalence invariant, subsumption preserving and reverse subsumption preserving. Additionally, we provide counterexamples for the properties equivalence closed, structural dependent, dissimilar closed, bounded and triangle inequality.

Lemma 11. *The similarity measure s is*

1. *symmetric,*
2. *equivalence invariant,*
3. *subsumption preserving and*
4. *reverse subsumption preserving.*

Proof. 1. The symmetry of s is obvious.

2. As a interpretation based measure, it is easy to see that s is equivalence invariant.

3. Let $C, D, E \in \mathcal{C}(\mathcal{ALE})$ with $C \sqsubseteq D \sqsubseteq E$. In the following, we use the abbreviations $t := |\Delta^{\mathcal{I}_\mathcal{A}}|$, $c := |C^{\mathcal{I}_\mathcal{A}}|$, $d := |D^{\mathcal{I}_\mathcal{A}}|$ and $e := |E^{\mathcal{I}_\mathcal{A}}|$. We know that $GCS(C,D) = lcs(C,D) = D$, $GCS(C,E) = lcs(C,E) = E$, $c \leq d \leq e$, $min\{c,d\} = c$ and $min\{c,e\} = c$. Therefore, we have

$$s(C,D) = \frac{c}{d}(1 - \frac{d}{t}(1 - \frac{c}{d})) = c \cdot \frac{t + c - d}{td}$$

and

$$s(C, E) = \frac{c}{e}(1 - \frac{e}{t}(1 - \frac{c}{e})) = c \cdot \frac{t + c - e}{te}.$$

and can derive $s(C, E) \leq s(C, D)$ using the following chain:

$$\begin{aligned} d \leq e &\iff d(t + c) \leq e(t + c) \iff dt + dc - ed \leq et + ec - ed \\ &\iff \frac{t + c - e}{e} \leq \frac{t + c - d}{d} \iff c \cdot \frac{t + c - e}{te} \leq c \cdot \frac{t + c - d}{td} \\ &\iff s(C, E) \leq s(C, D). \end{aligned}$$

4. Let C, D, E, c, d, e and t be as above. We know that $c \leq d \leq e \leq t$. Therefore $(t - e)(d - c) \geq 0$. Using this fact we derive

$$\begin{aligned} (t - e)(d - c) \geq 0 &\iff td - ed \geq tc - ec \iff td + d^2 - ed \geq tc + c^2 - ec \\ &\iff d(t + d - e) \geq c(t + c - e) \\ &\iff d \cdot \frac{t + d - e}{te} \geq c \cdot \frac{t + c - e}{te} \\ &\iff s(D, E) \geq s(C, E). \end{aligned}$$

□

- The measure is not **equivalence closed**. Let $N_C := \{A, B\}$, $N_r := \{r\}$ and $\mathcal{A} := \{A(x), B(y), r(a, x), r(a, y)\}$. Then $GCS(\exists r.A, \exists r.B) = \exists r.\top$ and $|(\exists r.\top)^{\mathcal{I}_\mathcal{A}}| = 1$ which implies

$$s(\exists r.A, \exists r.B) = \frac{min\{1, 1\}}{|(\exists r.\top)^{\mathcal{I}_\mathcal{A}}|}(1 - \frac{|(\exists r.\top)^{\mathcal{I}_\mathcal{A}}|}{3}(1 - \frac{min\{1, 1\}}{|(\exists r.\top)^{\mathcal{I}_\mathcal{A}}|})) = 1$$

but $\exists r.A \not\equiv \exists r.B$.

- The measure is not **structural dependent** as the following example proves. Let $N_C := \{A_1, \ldots A_n, B, G\}$ and $\mathcal{A} := \{A_1(x), A_1(y), \ldots A_n(x), A_n(y), B(x)\}$. The similarity of the concept descriptions

$$\begin{aligned} C_n &:= \bigsqcap_{i \leq n} A_i \sqcap B \\ D_n &:= \bigsqcap_{i \leq n} A_i \sqcap \neg B \end{aligned}$$

is

$$\begin{aligned} s(C_n, D_n) &= \frac{min\{1, 1\}}{|(\bigsqcap_{i \leq n} A_i)^{\mathcal{I}_\mathcal{A}}|}(1 - \frac{|(\bigsqcap_{i \leq n} A_i)^{\mathcal{I}_\mathcal{A}}|}{2}(1 - \frac{min\{1, 1\}}{|(\bigsqcap_{i \leq n} A_i)^{\mathcal{I}_\mathcal{A}}|})) \\ &= \frac{1}{2}(1 - \frac{2}{2}(1 - \frac{1}{2})) \\ &= \frac{1}{4} \end{aligned}$$

for all $n \geq 1$ and therefore s is not structural dependent.

- The measure is not **dissimilar closed**. Let $N_C := \{A, B\}$ and $\mathcal{A} := \{A(x), B(y)\}$ then $lcs(A, B) = \top$ but

$$s(A, B) = \frac{1}{2}(1 - \frac{2}{2}(1 - \frac{1}{2}) = \frac{1}{4}.$$

- The measure is not **bounded** because the similarity of a concept descriptions C with $|C^{\mathcal{I}_\mathcal{A}}| = 0$ and any other arbitrary concept description is always 0.

- The measure does not fulfill the **triangle inequality**. Let

$$\begin{aligned} N_C &:= \{A_C, A_D, A_E, A_{CD}, A_{CE}, A_{DE}\}, \\ \mathcal{A} &:= \{A_C(x), A_D(x), A_E(x), A_{CD}(x), A_{CE}(x), A_{DE}(x), A_{CD}(y)\} \end{aligned}$$

  and we define $C := A_C \sqcap A_{CD} \sqcap A_{CE}$, $D := A_D \sqcap A_{CD} \sqcap A_{DE}$ and $E := A_E \sqcap A_{CE} \sqcap A_{DE}$. The similarity values are

$$s(C, D) = \frac{1}{2}(1 - \frac{1}{2}(1 - \frac{1}{2})) = \frac{1}{4}$$

  and

$$s(C, E) = s(E, D) = \frac{1}{1}(1 - \frac{1}{2}(1 - \frac{1}{1})) = 1.$$

  Therefore we have

$$1 + s(C, D) = \frac{5}{4} < s(C, E) + s(E, D) = 2$$

  which contradicts triangle inequality.

4.4 Tabular Overview

Table 4.1 presents an overview that contains all measures (including our measure *simi* from Chapter 5) and properties. The first five measures are pure structural measures. The next two are structural measures which use the canonical interpretations to measure primitives and the last two are pure interpretation based measures. The shortcuts used at the head of the table are

- sym = symmetric,
- tring = triangle inequality,
- eqcl = equivalence closed,
- eqinv = equivalence invariant,
- sub = subsumption preserving,
- resub = reverse subsumption preserving,

- diss = dissimilar closed,
- bound = bounded,
- struc = structural dependent.

Measure	sym	tring	eqcl	eqinv	sub	resub	diss	bound	struc	DL
simi	x	o	x	x	x	o	x	x	x	$\mathcal{ELH}$
Jacc	x	x	x	x	x	x	x	x	x	$\mathcal{L}_0$
Dice	x	o	x	x	x	x	x	x	x	$\mathcal{L}_0$
[JW09]	x	o	o	o	o	o	o	o	x	$\mathcal{SHI}$
[Jan06]	x	o	o	o	o	o	o	o	x	$\mathcal{ALCHQ}$
[dFE06]	o	o	o	o	o	o	o	o	o	$\mathcal{ALC}$
[FD06]	x	o	o	x	x	x	o	o	o	$\mathcal{ALN}$
[dFE05]	x	o	o	x	x	x	o	o	o	$\mathcal{ALC}$
[dSF08]	x	o	o	x	x	x	o	o	o	$\mathcal{ALE}$

Table 4.1: Overview of similarity measures and their properties

5 The Similarity Measure Simi

In this chapter we present *simi*, a structural measure for $\mathcal{ELH}$ concept descriptions and unfoldable TBoxes. It has parameter which allow tuning, is computable in time polynomial in the size of the concept descriptions and it fulfills all properties (equivalence closed, equivalence invariant, dissimilar closed, bounded, structural dependent, subsumption preserving) except reverse subsumption preserving and triangle inequality.

Simi does not use the concept axioms stored in the TBox. As presented in Chapter 2, unfoldable TBoxes can be normalized and extended. After both steps, we can extend the concepts to measures, so that all occurring concept names are primitive names. By doing so, the knowledge of the concept definitions stored in the TBox is not necessary to measure similarity. This is because a TBox does not provide any knowledge about primitive names. If one wants to use *simi* to measure concept descriptions with respect to an unfoldable TBox, *simi* has to be used after the normalization and extension steps. Therefore, for the rest of the chapter, we assume that the TBox is empty. Note that this assumption influences our notations. The expanded concept descriptions consist of primitive names only. In our definitions and proofs, we use the term concept names because in the absence of concept axioms, they are the same. However, *simi* uses the knowledge stored in an RBox that is denoted with $\mathcal{R}$.

Another preprocessing step is the transformation of the concept descriptions to measure into the $\mathcal{ELH}$ normal form presented in 2.1.4. The uniqueness of this normal form (with respect to associativity and commutativity) ensures that *simi* (and any other measure using this normal form) is equivalence invariant. For the rest of this chapter we assume that the concepts involved are all in normal form.

Simi depends on several parameters. Therefore, we consider *simi* to be a *framework* rather than just one measure. The choice of values for the parameters does not influence the properties (except dissimilar closed and structural dependent), but can be used to tune *simi* towards ones needs.

The appearance of simi is partially inspired by the equivalence operator. Equivalence can be regarded as a very trivial similarity measure. The similarity of two concept descriptions is one if they are equal and zero otherwise. To determine if $C \equiv D$ is true, one can use the subsumption operator to find out whether or not $C \sqsubseteq D$ and $D \sqsubseteq C$ are true. We generalize this approach in *simi* by introducing a generalization of the subsumption operator. Since such an operator is in general an asymmetric function, we call it *directed simi* and denote it with $simi_d$. As a similarity measure should be 1 if and only if both concept descriptions to measure are equivalent, a generalization of the subsumption operator should be 1 if and only if the first argument is subsumed by the second one. Note that in $simi_d$, we reverse the order of the arguments compared to the subsumption operator. Therefore, formally, for all concept descriptions C and D we need that

$$simi_d(C, D) = 1 \iff D \sqsubseteq C$$

in order to generalize the subsumption operator.
Once we have computed the values $simi_d(C, D)$ and $simi_d(D, C)$ we have to combine them with an operator to obtain a value for *simi*. Instead of using an specific operator, we identified the properties such a operator should have so that *simi* fulfills as much properties as possible. We call such an operator (with sufficient properties) a *fuzzy connector* and denote it with $\otimes$. Using fuzzy connectors, *simi* is simply defined as

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

Another inspiration for *simi* (and $simi_d$) is the Jaccard Index (see 4.2.1). The Jaccard Index can be regarded as a $\mathcal{L}_0$ similarity measure which, as we proved in 4.2.1, fulfills all properties defined in Chapter 3. Also, it is used to measure GeneOntology-based protein semantic similarity and did well compared to other measures in this area [PFB+07]. Therefore, we aimed to generalize it in a way that it can deal with existential restrictions and a role hierarchy.

In the following Section we present the definition of $simi_d$ and its derivation from the Jaccard Index. Section 5.2 introduces fuzzy connectors which enable us to define *simi* in Section 5.3. Section 5.4 contains the proofs that independently from the choice of values for the parameters, *simi* is symmetric, equivalence invariant, equivalence closed, subsumption preserving, bounded, structural dependent and under some circumstances dissimilar closed. Additionally, it contains the proofs that *simi* can be computed in time polynomial in the size of the extended concept descriptions to measure, *simi* generalizes the Jaccard Index and in general, it is not reverse subsumption preserving and it does not fulfill the triangle inequality.

5.1 The Function $simi_d$

In this section we present a derivation and the definition of $simi_d$. We use the average as fuzzy connector to present examples. It is denoted with $\otimes_{avg}$ and the proof that the average is a fuzzy connector can be found in Section 5.2.

The starting point for the derivation of $simi_d$ is the function

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

which represents

How much information of C is shared with D.

In the following, we use the abbreviation

$$s(C, D) := d(C, D) \otimes_{avg} d(D, C).$$

This function can be used to measure sets of concept names. In order to be able to incorporate existential restrictions, we rewrite the numerator of d. For every $A \in C$, we search for the atom in D which has the highest similarity to A. If we then take the sum of all this similarity values, we have the same result as in the numerator of d. Using the function $f : N_C \longrightarrow [0,1]$ defined as

$$f(C', D') := \begin{cases} 0 & \text{if } C' \neq D' \\ 1 & \text{if } C' = D' \end{cases}$$

we can express the numerator through

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

and redefine d (by regarding f as a parameter) with

$$d[f](C,D) := \frac{\sum_{C' \in \widehat{C}} \max_{D' \in \widehat{D}} f(C', D')}{|\widehat{C}|}.$$

Here, the function f is the similarity-measure version of the equivalence operator restricted to concept names. The underlying assumption for f is that two different concept names are always totally dissimilar. However, this assumption may not be correct in all cases. Therefore, we generalize f by introducing a measure for concept names called *primitive measure*. To be able to work with existential restrictions, a primitive measure has to be able to deal with role names too. Also we have to force some properties to ensure properties of *simi*.

Definition 38 (primitive measure)**.** *A function $pm : N_C^2 \cup N_r^2 \longrightarrow [0,1]$ with the properties that for all $A, B \in N_C$ and $r, s, t \in N_r$*

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

is called a primitive measure.

The first two properties are necessary to ensure that *simi* is equivalence closed, the third one ensures that *simi* is bounded and the last one is needed to prove that *simi* is subsumption preserving. Note that *pm* does not have to be symmetric.

We present a short examples of a use-case of a primitive measure. Let

$$N_C := \{Brown, Green, Yellow, Black, Mammal, Reptile,$$

$$BrownSnake, GreenFrog, FireSalamander, BrownBear\}$$

and

- $BrownSnake \equiv Brown \sqcap Reptile$,
- $GreenFrog \equiv Green \sqcap Reptile$,
- $FireSalamander \equiv Yellow \sqcap Reptile$,

For the concept descriptions $BrownSnake$, $GreenFrog$ and $FireSalamander$ we have

$$s(BrownSnake, GreenFrog) =$$

$$s(BrownSnake, FireSalamander) = \frac{1}{2} \otimes_{avg} \frac{1}{2} = \frac{1}{2}.$$

One could argue that $Brown$ and $Yellow$ are more similar than $Brown$ and $Green$ and therefore the $BrownSnake$ should be more similar to the $FireSalamander$ as to the $GreenFrog$. To express the higher similarity of $Brown$ and $Yellow$ compared to $Green$ we could define a primitive measure pm' for all $A, B \in N_C$ through

- $pm'(Brown, Yellow) = pm'(Yellow, Brown) := 0.5$,
- $pm'(A, A) := 1$ and
- $\{A, B\} \neq \{Brown, Yellow\}$ and $A \neq B \implies pm'(A, B) := 0$.

Looking at our example above and using pm' we obtain

$$s[pm'](BrownSnake, GreenFrog) =$$

$$\frac{pm'(Reptile, Reptile) + pm'(Brown, Green)}{2} \otimes_{avg} \frac{1+0}{2} =$$

$$\frac{1+0}{2} \otimes_{avg} \frac{1+0}{2} = \frac{1}{2}$$

and

$$s[pm'](BrownSnake, FireSalamander) =$$

$$\frac{pm'(Reptile, Reptile) + pm'(Brown, Yellow)}{2} \otimes_{avg} \frac{1+0.5}{2} =$$

$$\frac{1+0.5}{2} \otimes_{avg} \frac{1+0.5}{2} = \frac{3}{4}.$$

Now $BrownSnake$ and $FireSalamander$ have a higher similarity value than $GreenFrog$ and $BrownSnake$.

The extension of f to a primitive measure is called *default primitive measure*. Whenever no concrete primitive measure is provided, we assume that the default primitive measure is used.

Definition 39 (default primitive measure)**.** *The function* $pm_{def} : N_C^2 \cup N_r^2 \longrightarrow [0,1]$ *defined for all* $A, B \in N_B$ *and* $r, s \in N_r$ *by*

$$pm_{def}(A,B) := \begin{cases} 0 & \textit{if } A \neq B \\ 1 & \textit{if } A = B \end{cases}$$

and

$$pm_{def}(r,s) := \begin{cases} 1 & \textit{if } r = s \textit{ or } s \sqsubseteq r \\ 0 & \textit{if } s \not\sqsubseteq r \textit{ and } r \not\sqsubseteq r \\ 0.01 & \textit{if } r \sqsubseteq s \textit{ and } s \not\sqsubseteq r \end{cases}$$

is called the default primitive measure*.*

To identify a suitable primitive measure, we propose to start with the default primitive measure and assign values to pairs of concept names where refinement is required. If an ABox is present and has a representative domain, then one could also use the canonical interpretation to measures concept names. For example we could define a primitive measure $pm_{\mathcal{I}_\mathcal{A}}$ through

$$pm_{\mathcal{I}_\mathcal{A}}(A,B) := \begin{cases} 1 & \text{if } A^{\mathcal{I}_\mathcal{A}} = B^{\mathcal{I}_\mathcal{A}} = \emptyset \\ \frac{|A^{\mathcal{I}_\mathcal{A}} \cap B^{\mathcal{I}_\mathcal{A}}|}{|A^{\mathcal{I}_\mathcal{A}} \cup B^{\mathcal{I}_\mathcal{A}}|} & \text{otherwise.} \end{cases}$$

However, note that this is not a valid primitive measure because the first property ($pm(A,B) = 1 \iff A = B$) is not true in general since the measure cannot distinguish concept names that have the same individuals. The consequence would be that *simi* would not be equivalence closed. Any other property would not be effected.

A way to measure the similarity of two roles in a role hierarchy can be found in [Jan08] (Section 5.5.2 page 46). In short, it is a network-based approach [RMBB89] where "Similarity is expressed as the ratio between the shortest path from [the role] r to [the role] s and the maximum path within the graph representation of the role hierarchy."

To incorporate existential restrictions we have three different cases to consider. Namely, we need to be able to compute the similarity of

1. two concept names,
2. a concept name and an existential restriction and
3. two existential restrictions.

The first case is handled directly by the primitive measure. In the second case, we assign that a concept name and a existential restriction are always totally dissimilar and therefore the similarity is 0. For the third case, let $\exists r.C^*$ and $\exists s.D^*$ be the two existential restrictions. To compute the similarity of both atoms, we work with two components. The similarity of the role names is computed using the primitive measure and the similarity of the concept descriptions C^* and D^* which is computed by a recursive call. Then, to combine both values we could build the product. Using the notation d', we can express this formally through

$$d'(\exists r.C^*, \exists s.D^*) = pm(r,s) \cdot d'(C^*, D^*).$$

However, this approach has a disadvantage. Let $A, B \in N_C$ with $pm(A,B) = 0$ and $r, s \in N_r$ with $pm(r,s) = 0$. Then

$$d'(\exists r.A, \exists r.B) = d'(\exists r.A, \exists s.B) = 0.$$

In the first case, we measure two existential restrictions where the roles are equal and in the second case both roles are totally dissimilar. The measure does not take into account that both cases are different and it is also not bounded because $lcs(\exists r.A, \exists r.B) = \exists r.\top \neq \top$. By searching a solution for this problem, we discussed several possibilities and found that almost all of them can be described generally using a number $w \in (0,1)$ and the formula

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

In this case, we have

$$d'(\exists r.A, \exists r.B) = pm(r,r) \cdot [w + (1-w)pm(A,B)] = w$$

and

$$d'(\exists r.A, \exists s.B) = pm(r,s) \cdot [w + (1-w)pm(A,B)] = 0.$$

Since we require $w > 0$, both cases are now distinguished and it can be proven (see 5.4) that this is enough to ensure that $simi$ is bounded.

To find a suitable w, we suggest that one should try to identify the value n where one would say that the concept descriptions

$$C := \underbrace{\exists r. \ldots \exists r.}_{n} A$$

and

$$D := \underbrace{\exists r. \ldots \exists r.}_{n} B$$

are (nearly) totally similar. In Table 5.1 we present examples of w and the corresponding n where $simi(C,D) > 0.99$ using different fuzzy connectors.

$w =$	n for $\otimes_{Dice}$	n for $\otimes_{H0}$	n for $\otimes_{avg}$
0.8	4	5	4
0.7	5	6	5
0.6	7	7	7
0.5	8	9	8
0.4	11	12	11
0.3	14	16	14
0.25	18	20	18
0.2	22	25	22
0.1	45	52	45
0.05	91	105	91
0.01	460	528	450

Table 5.1: Examples of w and n where $simi(C, D) > 0.99$

As default value we suggest 0.1.

Putting all pieces together we obtain the following function d' which now has the primitive measure as a third argument. We denote this special argument using curved parentheses. Whenever it is clear what primitive measure to use, we omit writing the curved parentheses.

$$d'[pm](C, D) := \begin{cases} \dfrac{\sum\limits_{C' \in \widehat{C}} \max\limits_{D' \in \widehat{D}} d'[pm](C', D')}{|\widehat{C}|} & \text{if } |\widehat{C}| > 1 \text{ or } |\widehat{D}| > 1, \\ pm(A, B) & \text{if } C, D \in N_C, \\ pm(r, s) \cdot [w + (1-w) d'[pm](E, F)] & \text{if } C = \exists r.E \text{ and } D = \exists s.F, \\ 0 & \text{otherwise.} \end{cases}$$

5.1.1 Weighting Atoms

Currently all atoms are weighted equally in the function d'. However, there are cases where one wants to prioritise some concept names over others. For this purpose, we introduce the possibility to add weights to atoms by using a *weighting function*.

Definition 40 (weighting function). *A function $g : N_A \longrightarrow \mathbb{R}_{>0}$ is called a* weighting function.

To incorporate the weighting function into d' we generalize the cardinality of a set by the sum of the weights of its elements. For example, if $g(\mathit{Reptile}) = 2$ and $g(\mathit{Brown}) = 1$ then the new cardinality of *BrownSnake* is $g(\mathit{Reptile}) + g(\mathit{Brown}) = 3$. To obtain a well-defined measure, we have to add the weights to the numerator of d' as well. We use the notion d^* to refer to the generalization of d' which includes the weighting function.

As like in d', we write the input arguments of d^* for the weighting function and the primitive measure with curved parentheses. Formally d^* is defined by

$$d^*[pm,g](C,D) := \begin{cases} \dfrac{\sum\limits_{C' \in \widehat{C}} g(C') \max\limits_{D' \in \widehat{D}} d^*[pm,g](C',D')}{\sum\limits_{C' \in \widehat{C}} g(C')} & \text{if } |\widehat{C}| > 1 \text{ or } |\widehat{D}| > 1 \\ pm(A,B) & \text{if } C, D \in N_C \\ pm(r,s)[w + (1-w)d^*[pm,g](E,F)] & \text{if } C = \exists r.E \text{ and } D = \exists s.F \\ 0 & \text{otherwise.} \end{cases}$$

Correspondingly, we define

$$s^*[pm,g](C,D) := d^*[pm,g](C,D) \otimes_{avg} d^*[pm,g](D,C).$$

Since N_A is an infinite set we simply cannot write down an entire weighting function. It has to be defined in a more abstract way. The simplest approach is to define a function f which weights primitive names and role names and then use f to define g. For an arbitrary $f : N_B \cup N_r \longrightarrow \mathbb{R}_{>0}$, we could define g through

$$g(C) := \begin{cases} f(C) & \text{if } C \in N_B \\ f(r) & \text{if } C \text{ is of the form } \exists r.D \end{cases}.$$

If $f(N_B \cup N_r) \subseteq [0,1]$, one could also define g recursively through

$$g(C) := \begin{cases} f(C) & \text{if } C \in N_B \\ f(r) \cdot \prod\limits_{D' \in \widehat{D}} g(D') & \text{if } C \text{ is of the form } \exists r.D \end{cases} \cdot$$

To find a suitable weighting function, it is best to start with the *default weighting function* which weighs everything equal.

Definition 41 (default weighting function)**.** *The function $g_{def} : N_A \longrightarrow \mathbb{R}_{>0}$ with for all $C' \in N_A$ $g_{def}(C') := 1$, is called the* default weighting function*.*

The next step would be to identify concept names and role names which should have a higher (lower) impact on the similarity value and increase (decrease) their weights. We now present a short example of a use-case for a weighting function. We use the terminology defined in the example above and add the concept description

$$BrownBear \equiv Brown \sqcap Mammal.$$

If we measure *BrownSnake* and *BrownBear* using the default primitive measure, we have

$$s(BrownSnake, BrownBear) =$$

$$s(BrownSnake, GreenFrog) = \frac{1}{2} \otimes_{avg} \frac{1}{2} = \frac{1}{2}.$$

One could argue that for similarity, the animal class is more important than the color of an animal and therefore the similarity between *BrownSnake* and *GreenFrog* should be higher than between *BrownSnake* and *BrownBear*. We can address this argument with the *weighting function* $\widehat{g}$ defined by

- $\widehat{g}(Reptile) = \widehat{g}(Mammal) := 2$ and
- $\widehat{g}(Brown) = \widehat{g}(Green) = \widehat{g}(Yellow) := 1.$

Then, using s^* we obtain

$$s^*[pm_{def}, \widehat{g}](BrownSnake, BrownBear) =$$
$$\frac{\widehat{g}(Brown) \cdot 1 + \widehat{g}(Reptile) \cdot 0}{\widehat{g}(Brown) + \widehat{g}(Reptile)} \otimes_{avg} \frac{1 \cdot 1 + 2 \cdot 0}{3} =$$
$$\frac{1}{3} \otimes_{avg} \frac{1}{3} = \frac{1}{3}$$

and

$$s^*[pm_{def}, \widehat{g}](BrownSnake, GreenFrog) =$$
$$\frac{\widehat{g}(Brown) \cdot 0 + \widehat{g}(Reptile) \cdot 1}{\widehat{g}(Brown) + \widehat{g}(Reptile)} \otimes_{avg} \frac{1 \cdot 0 + 2 \cdot 1}{3} =$$
$$\frac{2}{3} \otimes_{avg} \frac{2}{3} = \frac{2}{3}.$$

Now the concepts *BrownSnake* and *GreenFrog* are more similar than *BrownSnake* and *BrownBear*.

5.1.2 Using more Knowledge

So far, for every atom of C we search for the atom of D with the highest similarity value to compute s^*. This approach is not always sufficient as the following example illustrates. If we measure *Brown* and $Yellow \sqcap Black$ with the default weighting function and a primitive measure pm^* with

$$pm^*(Brown, Yellow) = pm^*(Yellow, Brown)$$
$$= pm^*(Brown, Black) = pm^*(Black, Brown) = 0.5$$

then

$$s^*[pm^*](Brown, Yellow \sqcap Black) =$$
$$\frac{max\{0.5, 0.5\}}{1} \otimes_{avg} \frac{0.5 + 0.5}{2} = \frac{1}{2}.$$

The way the similarity is computed does not take into account that *Brown* is related to *Yellow* and *Black*. It chooses the "best matching partner". To deal with this problem we propose to exchange the maximum operator with a t-conorm. The choice for a t-conorm comes from several facts. First, the operator max is a t-conorm. Secondly, all t-conorms ($\oplus$) are greater or equal than max ($max\{x, y\} \leq x \oplus y$) which is consistent with our expectation that the value should be higher or equal than the maximum. Also, 0 acts as neutral element for t-conorms. Therefore, all atoms from D that are totally dissimilar do not influence the value.

If we use the probabilistic sum ($x \oplus_{sum} y = x + y - xy$) instead of the maximum for our example above then

$$s^*[pm^*](Brown, Yellow \sqcap Black) =$$

$$\frac{0.5 \oplus_{sum} 0.5}{1} \otimes_{avg} \frac{0.5 + 0.5}{2} = \frac{0.75}{1} \otimes_{avg} \frac{1}{2} = \frac{5}{8}.$$

The probabilistic sum takes credit to the fact that $Brown$ is related to $Yellow$ and $Black$ by having a higher value (0.75) then the maximum (0.5).

To ensure that $simi$ is equivalence closed, the t-conorm has to be bounded ($x \oplus y = 1 \implies x = 1$ or $y = 1$). The t-conorm has to ensure that if the result is one, then there was an atom in the other set such that the similarity is already one. If this is not the case, then we can construct a counterexample where the similarity is one but the concept descriptions are not equivalent. We present a short example with a t-conorm which is not bounded to illustrate this. Let $N_C := \{A, B, C, D\}$ and

$$\forall x, y \in N_C : \overline{pm}(x, y) := \begin{cases} 1 & \text{if } x = y \\ 0.5 & \text{if } x \neq y \end{cases}.$$

Using the bounded sum ($x \oplus_{luk} y = min\{x + y, 1\}$) as t-conorm, we obtain

$$s^*[\overline{pm}](A \sqcap B, C \sqcap D) =$$

$$\frac{min\{\overline{pm}(A, C) + \overline{pm}(A, D), 1\} + min\{\overline{pm}(B, C) + \overline{pm}(B, D), 1\}}{2} \otimes_{avg} \frac{1 + 1}{2} = 1$$

but $A \sqcap B$ and $C \sqcap D$ are not equivalent.

5.1.3 Definition of $simi_d$

We know present a formal definition of $simi_d$ which is basically a summary of the derivation presented above. The only additions are the cases involving the concept $\top$. To fulfill the initially stated property that

$$simi_d(C, D) = 1 \iff D \sqsubseteq C$$

we have to ensure that $simi_d(\top, D) = 1$ for an arbitrary concept description D. Additionally, to ensure that $simi$ is dissimilar closed, we define that for all concept descriptions $C \neq \top$, $simi_d(C, \top) = 0$. This expectation is covered in the 'otherwise' case of the definition.

If we want to be mathematically correct, then the type of the function $simi_d$ depends on the used parameters (the primitive measure, the weighting function, w and the t-conorm) as well as on the concept descriptions to be measured. However, for simplicity, we omit writing the parameters in the type because most of the lemmas later will refer to the case of arbitrary parameters. If we want to make the parameters explicit, we use curved parentheses. The order of parameters is

$$simi_d[\text{t-conorm}, \text{primitive measure}, \text{weighting function}, w].$$

For example, if the t-conorm is $\oplus_{max}$, $w = 0.5$ and the other parameters are free, we write $simi_d[max, \cdot, \cdot, 0.5]$.

Definition 42 ($simi_d$)**.** *Let $C, D, E, F \in \mathcal{C}(\mathcal{ELH})$, $A, B \in N_C$ and $r, s \in N_r$.* Directed simi *is the function $simi_d : \mathcal{C}(\mathcal{ELH})^2 \longrightarrow [0,1]$ defined (with respect to a bounded t-conorm $\oplus$, a primitive measure pm, a weighting function g and $w \in (0,1)$) as follows*

$$simi_d(C,D) := \begin{cases} \dfrac{\sum\limits_{C' \in \widehat{C}} [g(C') \cdot \bigoplus\limits_{D' \in \widehat{D}} simi_d(C', D')]}{\sum\limits_{C' \in \widehat{C}} g(C')} & \textit{if } C \neq \top \textit{and } (|\widehat{C}| > 1 \textit{ or } |\widehat{D}| > 1) \\ 1 & \textit{if } C = \top \\ pm(A,B) & \textit{if } C, D \in N_C \\ pm(r,s)[w + (1-w)simi_d(E,F)] & \textit{if } C = \exists r.E \textit{ and } D = \exists s.F \\ 0 & \textit{otherwise.} \end{cases}$$

5.2 The Fuzzy Connector

A fuzzy connector is a function which combines the values $simi_d(C, D)$ and $simi_d(D, C)$ to obtain a similarity value for C and D. Since $simi_d$ is a function producing values between 0 and 1, a fuzzy connector has to be an operator mapping $[0,1]^2$ to $[0,1]$. The properties of a fuzzy connector are necessary to ensure some of the properties of *simi*.

Definition 43 (fuzzy connector)**.** *A* fuzzy connector *is an operator on the interval* $[0,1]$*,* $\otimes : [0,1]^2 \longrightarrow [0,1]$ *such that for all $x, y \in [0,1]$ the following properties are true.*

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

The commutativity of a fuzzy connector ensures that *simi* is symmetric, the properties equivalence closed and bounded are connected to the properties of similarity measures with the same name, weak monotonicity is necessary to prove that *simi* is subsumption preserving and grounded ensure that *simi* is dissimilar closed.
As mentioned above, the average is an example of a fuzzy connector. To find other interesting fuzzy connectors, we identified the fuzzy connectors which are necessary to resemble the set similarity measures Jaccard Index and Dice's Coefficient. Both set measures are well established and applied to a wide range of applications. Therefore, they are considered to be a good starting point for generalization. By using a t-norm called *Hamacher product* we obtain the Jaccard Index. The Hamacher product is a bounded t-norm and, as we prove later, all bounded t-norms are fuzzy connectors. To obtain the Dice's Coefficient, one has to use the *Dice's Connector* which is defined as follows.

Definition 44 (Dice's Connector)**.** *The* Dice's Connector *is a function* $\otimes_{Dice} : [0,1]^2 \longrightarrow [0,1]$ *defined through*

$$x \otimes_{Dice} y := \begin{cases} 0 & \textit{if } x = y = 0 \\ \frac{2xy}{x+y} & \textit{otherwise} \end{cases}.$$

The following lemma proves our claims that the average, bounded t-norms and the Dice's Connector are fuzzy connectors.

Lemma 12.

1. *Bounded t-norms* ($\otimes$)*,*
2. *the Dice's Connector and*
3. *the average*

are fuzzy connectors.

Proof.

1. Bounded t-norms are defined to be commutative and bounded. Their monotonicity implies weak monotonicity. Therefore, we just have to prove that $\otimes$ is equivalence closed and grounded.
   - Equivalence Closed:
     $\Rightarrow$: Let $x \otimes y = 1$. Because $y \leq 1$ and $\otimes$ is monotonic, we can derive $x \otimes y \leq x \otimes 1 = x \leq 1$. Therefore $1 \leq x \leq 1 \implies x = 1$. With similar arguments, we can derive that $y = 1$.
     $\Leftarrow$: Let $x = y = 1$. Because 1 is an identity element, we know that $x \otimes 1 = x$ and so for $x = 1$ we derive $1 \otimes 1 = 1$.
   - Grounded: Monotonicity implies that $0 = 0 \otimes 1 \geq 0 \otimes 0 \geq 0$ and therefore $0 \otimes 0 = 0$.

2. Dice's Connector:
   - The commutativity of Dice's Connector is obvious.
   - Equivalence Closed:
     $\Rightarrow$: If $x = y = 1$ then $x \otimes_{Dice} y = \frac{2*1*1}{1+1} = 1$.
     $\Leftarrow$: Let $x \otimes_{Dice} y = 1$. Since we know that x and y are both not zero, we can transform the equation $\frac{2xy}{x+y} = 1$ into $2xy = x+y$. We prove that this equation has only two solutions $(x, y) \in [0, 1]^2$, namely $(0, 0)$ and $(1, 1)$. Since the tuple $(0, 0)$ is not a valid solution for the original equation, because $0 \otimes_{Dice} 0 = 0$, we can derive that $x = y = 1$.
     We transform $2xy = x + y$ into $f(x) = y = \frac{x}{2x-1}$ and prove that $f((0, 1)) \cap [0, 1] = \emptyset$. At the point $x = 1/2$, f is not defined. For $0 < x < 1/2$, $2x < 1$ and therefore $2x - 1 < 0$ which implies that $f(x) < 0$. For the last case, $1/2 < x < 1$, we observe $x + 1 > x + x$ which can be transformed into $x > 2x - 1 > 0$. This implies that $f(x) > 1$.
   - Weak Monotonicity: Let $x \leq y$. Then

$$x \leq y \iff 2x \leq 2y \iff 2xy + 2x \leq 2xy + 2y \iff 2x(y + 1) \leq 2y(x + 1)$$

$$\iff \frac{2x}{x+1} \leq \frac{2y}{y+1} \iff 1 \otimes_{Dice} x \leq 1 \otimes_{Dice} y.$$

   - Bounded: Let $x \otimes_{Dice} y = 0$. This implies $2xy = 0$ which implies that $x = 0$ or $y = 0$.
   - Grounded: $0 \otimes_{Dice} 0 = 0$ by definition.

3. Average:
   - Commutativity of the average is obvious.
   - Equivalence Closed:
     $\Rightarrow$: Let $x = y = 1$ then $avg(x, y) = \frac{1+1}{2} = 1$.
     $\Leftarrow$: If $avg(x, y) = 1$ then $x + y = 2$. Since $0 \leq x, y \leq 1$, we derive $x = y = 1$.
   - Weak Monotonicity: Let $x \leq y$. Then

$$x \leq y \iff x + 1 \leq y + 1 \iff \frac{x+1}{2} \leq \frac{y+1}{2} \iff avg(1, x) \leq avg(1, y).$$

   - Bounded: $(x + y)/2 = 0$ implies that $x = y = 0$ because $x, y \geq 0$.
   - Grounded: $(0 + 0)/2 = 0$.

□

The following lemma is an aid to help deciding which fuzzy connector to choose. It proves that the fuzzy connectors we presented can be ordered. To find a suitable fuzzy connector we suggest to start with the minimum. If one has the feeling that the results for some test data are to high then the Hamacher product should be used, whereas if the values are to low then the Dice's Connector might fit better. The inspection of the distributions of the presented fuzzy connector also provides knowledge that can aid to choose.

Lemma 13. *Let $x, y \in [0,1]$, then*

$$x \otimes_{prod} y \;\le\; x \otimes_{H0} y \;\le\; x \otimes_{min} y \;\le\; x \otimes_{Dice} y \;\le\; x \otimes_{avg} y.$$

Proof. First we observe that if $x = y = 0$ then all fuzzy connectors are 0 by definition and therefore the inequality is true. Assuming that x and y are not both 0 and using the definitions we can reformulate the inequality to

$$xy \le \frac{xy}{x+y-xy} \le min\{x,y\} \le \frac{2xy}{x+y} \le \frac{x+y}{2}.$$

- $xy \le \frac{xy}{x+y-xy}$: We have

$$y \le 1 \iff (1-x)y \le (1-x) \iff y - xy \le 1 - x \iff$$
$$x + y - xy \le 1 \iff 1 \le \frac{1}{x+y-xy} \iff xy \le \frac{xy}{x+y-xy}.$$

- $\frac{xy}{x+y-xy} \le min\{x,y\}$: W.l.o.g. we assume that $x \le y$. We have

$$xy \le y \iff 0 \le x - xy \iff y \le x + y - xy \iff$$
$$\frac{y}{x+y-xy} \le 1 \iff \frac{xy}{x+y-xy} \le x = min\{x,y\}.$$

- $min\{x,y\} \le \frac{2xy}{x+y}$: W.l.o.g. we assume that $x \le y$ so we have

$$x \le y \iff x + y \le 2y \iff$$
$$1 \le \frac{2y}{x+y} \iff min\{x,y\} = x \le \frac{2xy}{x+y}.$$

- $\frac{2xy}{x+y} \le \frac{x+y}{2}$: We have

$$0 \le (x-y)^2 \iff 0 \le x^2 + y^2 - 2xy \iff 2xy \le x^2 + y^2$$
$$\iff 4xy \le x^2 + y^2 + 2xy \iff 4xy \le (x+y)^2$$
$$\iff \frac{4xy}{x+y} \le x + y \iff \frac{2xy}{x+y} \le \frac{x+y}{2}.$$

□

5.3 Definition of Simi

The similarity measure *simi* is defined using $simi_d$ and a fuzzy connector. As $simi_d$, we will denote the specific fuzzy connector, as well as the other parameters passed on to $simi_d$ using curved parentheses. The order of parameters is

$$simi[\text{fuzzy connector}, \text{t-conorm}, \text{primitive measure}, \text{weighting function}, w].$$

For example, if the fuzzy connector is $\otimes_{prop}$, the t-conorm is $\oplus_{max}$, $w = 0.5$ and the other parameters are free, we write $simi[prop, max, \cdot, \cdot, 0.5]$.

Definition 45 (*simi*)**.** *Let $\otimes$ be a fuzzy connector. The function simi : $\mathcal{C}(\mathcal{ELH})^2 \longrightarrow [0,1]$ is defined through*

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

5.4 Properties

Here we present the proofs that *simi* is *symmetric* (and therefore indeed a similarity measure according to our definition of similarity measures), *equivalence invariant*, *equivalence closed*, *subsumption preserving*, *bounded*, *structural dependent* and if one uses the default primitive measure then it is also *dissimilar closed*. Then we show that *simi* can be computed in time polynomial in the size of the concept descriptions to measure. Additionally, we prove that in general, *simi* is not reverse subsumption preserving and it does not fulfill the triangle inequality. Finally, we show that *simi* can be used to generalize the Jaccard Index and Dice's Coefficient.
In the following we omit writing the parameters explicitly and assume that the primitive measure is *pm*, the weighting functions is g, the t-conorm is $\oplus$ and the fuzzy connector is $\otimes$.
The following lemma is used later to prove that *simi* is equivalence invariant and equivalence closed.

Lemma 14. *Let $C, D \in \mathcal{C}(\mathcal{ELH})$, then*

$$simi_d(C, D) = 1 \iff D \sqsubseteq C.$$

Proof.

- $\Rightarrow$: Let $simi_d(C, D) = 1$. If $C = \top$ then $D \sqsubseteq C = \top$ is true. Let $C \neq \top$. According to Lemma 1, to prove $D \sqsubseteq C$ we have to show that $\forall C' \in \widehat{C}\ \exists D' \in \widehat{D} : D' \sqsubseteq C'$. Let C' be an arbitrary atom of C. $simi_d(C, D) = 1$ implies that

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

  Because

  $$g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_d(C', D') \leq g(C')$$

we derive that for all $C' \in \widehat{C}$: $\bigoplus_{D' \in D} simi_d(C', D') = 1$. Since the t-conorm is bounded, there $\exists D' \in D$ such that $simi_d(C', D') = 1$. The rest of the proof will be done using structural induction and case.
If $C' = A$ then $simi_d(C', D') = 1$ leads to $D' = A$ which implies $D' \sqsubseteq C'$.
Let $C' = \exists r.C^*$. According to the definition of $simi_d$, the only case possible such that $simi_d(C', D') = 1$ is D' is of the form $D' = \exists r.D^*$. Here

$$simi_d(C', D') = 1 = [w + (1 - w) simi_d(C^*, D^*)].$$

Since $w < 1$, $simi_d(C^*, D^*)$ has to be 1. The induction hypothesis implies that $D^* \sqsubseteq C^*$ which implies $D' \sqsubseteq C'$.

- $\Leftarrow$: Let $D \sqsubseteq C$. If $C = \top$ then by definition of $simi_d$, $simi_d(C, D) = 1$. Let $C \neq \top$. We have to show that

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

  which is equivalent to

  $$\forall C' \in \widehat{C} : \; g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_d(C', D') = g(C')$$

  and

  $$\forall C' \in \widehat{C} : \bigoplus_{D' \in D} simi_d(C', D') = 1.$$

  $D \sqsubseteq C$ and Lemma 1 imply that $\forall C' \in \widehat{C} \; \exists D' \in \widehat{D} : D' \sqsubseteq C'$. Let C' be an arbitrary atom of C and $D' \in \widehat{D}$ such that $D' \sqsubseteq C'$. Lemma 2 implies that it is enough to prove that $\exists D' \in D$ such that $simi_d(C', D') = 1$ to derive $\bigoplus_{D' \in D} simi_d(C', D') = 1$. The rest of the proof will be done using structural induction and case.
  If $C' = A$ then $D' \sqsubseteq C'$ implies that $D' = A$ which leads to

  $$simi_d(C', D') = pm(A, A) = 1.$$

  Let $C' = \exists r.C^*$. $D' \sqsubseteq C'$ implies that D' is of the form $D' = \exists s.D^*$ with $s \sqsubseteq_{\mathcal{R}} r$ and $D^* \sqsubseteq C^*$. The definition of primitive measures leads us to $pm(r, s) = 1$ and the induction hypothesis implies that $simi_d(C^*, D^*) = 1$. Therefore,

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

□

The following lemma will later be used to show that $simi$ is subsumption preserving.

Lemma 15. *Let C, D and E be concept descriptions such that $D \sqsubseteq E$. Then*

$$simi_d(C, E) \leq simi_d(C, D).$$

Proof. If $C = \top$ then $simi_d(C, E) = 1 \leq simi_d(C, D) = 1$ is true. Let $C \neq \top$. We are going to show that for all $C' \in \widehat{C}$

$$\bigoplus_{E' \in \widehat{E}} simi_d(C', E') \leq \bigoplus_{D' \in \widehat{D}} simi_d(C', D').$$

Since t-conorms are monotonic and 0 is the neutral element, it is enough to prove that for all $E' \in \widehat{E}$ there exists a $D' \in \widehat{D}$ such that $simi_d(C', E') \leq simi_d(C', D')$. $D \sqsubseteq E$ implies that there exists a $D' \in D$ with $D' \sqsubseteq E'$ For the rest of the proof, let E' be an arbitrary atom from E and $D' \in \widehat{D}$ with $D' \sqsubseteq E'$. We are going to prove that

$$simi_d(C', E') \leq simi_d(C', D'). \tag{5.1}$$

Let $C' = A \in N_C$. If $E' \notin N_C$ then $simi_d(C', E') = 0$ and Equation 5.1 is fulfilled. If $E' \in N_C$ then $D' \sqsubseteq E'$ implies $D' = E'$ and Equation 5.1 is true.

Let $C' = \exists r.C^*$. If $E' \in N_C \cup \{\top\}$ then $simi_d(C', E') = 0$ and Equation 5.1 is fulfilled. Let E' be of the form $E' = \exists s.E^*$. Since $D' \sqsubseteq E'$, D' has to be of the form $D' = \exists t.D^*$ with $t \sqsubseteq_{\mathcal{R}} s$ and $D^* \sqsubseteq E^*$. The induction hypothesis implies $simi_d(C^*, E^*) \leq simi_d(C^*, D^*)$. This is equivalent to

$$[w + (1 - w)simi_d(C^*, E^*)] \leq [w + (1 - w)simi_d(C^*, D^*)].$$

Because pm is subsumption preserving we have $pm(r, s) \leq pm(r, t)$ which implies

$$pm(r, s)[w + (1 - w)simi_d(C^*, E^*)] \leq pm(r, t)[w + (1 - w)simi_d(C^*, D^*)].$$

and therefore

$$simi_d(C', E') \leq simi_d(C', D').$$

□

The following lemma will be used to prove that *simi* is bounded and dissimilar closed if the default primitive measure is used.

Lemma 16. *Let $C, D \in \mathcal{C}(\mathcal{ELH})$. Then*

1. $simi_d(C, D) = 0 \implies lcs(C, D) = \top$ *and*

2. $C \not\equiv \top$ *and* $lcs(C, D) \equiv \top \implies simi_d[\cdot, pm_{def}](C, D) = 0$.

Proof. 1. If $C = \top$ then $simi_d(C, D) = 1$ which is a contradiction to our assumption and if $D = \top$ then $lcs(C, D) = \top$ is clear. Let $C \neq \top$ and $D \neq \top$. In general $simi_d(C, D) = 0$ implies

$$\sum_{C' \in \widehat{C}} [g(C') \cdot \bigoplus_{D' \in \widehat{D}} simi_d(C', D')] = 0$$

which implies that for all $C' \in \widehat{C}$, $\bigoplus_{D' \in \widehat{D}} simi_d(C', D') = 0$. Since 0 is the neutral element for t-conorms, the latter one is equivalent to $simi_d(C', D') = 0$ for all

$D' \in \widehat{D}$. We are going to use this fact to prove that $lcs(C', D') = \top$ which implies $lcs(C, D) = \top$.
Let $C' \in \widehat{C}$ and $D' \in \widehat{D}$ be arbitrary atoms. If $C' \in N_C$ and $D' \in N_A \setminus N_C$ or $C' \in N_A \setminus N_C$ and $D' \in N_C$ then $lcs(C', D') = \top$.
If $C', D' \in N_C$ then

$$simi_d(C', D') = 0 \implies pm(C', D') = 0 \implies C' \neq D' \implies lcs(C', D') = \top.$$

Otherwise, let $C' = \exists r.C^*$ and $D' = \exists s.D^*$. Then

$$0 = simi_d(C', D') = pm(r, s)[w + (1 - w)simi_d(C^*, D^*)].$$

If $w + (1 - w)simi_d(C^*, D^*)$ would be 0, then

$$simi_d(C^*, D^*) = -\frac{w}{1 - w}$$

which is not possible because $1 > w > 0$ and $simi_d(C^*, D^*) > 0$. Therefore, $pm(r, s) = 0$. This implies that $s \not\sqsubseteq_{\mathcal{R}} r$ and $r \not\sqsubseteq_{\mathcal{R}} s$. Therefore, $lcs(C', D') = \top$.

2. Since we assume that C is in $\mathcal{ELH}$ normal form, $C \not\equiv \top$ implies $C = \top$. If $D = \top$ then $simi_d[\cdot, pm_{def}](C, D) = 0$ by definition. Let $D \neq \top$. To prove $simi_d[\cdot, pm_{def}](C, D) = 0$, we show that for all $C' \in \widehat{C}$ and $D' \in \widehat{D}$, $simi_d(C', D') = 0$.
If $C' \in N_C$ and $D' \in N_A \backslash N_C$ or $C' \in N_A \backslash N_C$ and $D' \in N_C$ then $simi_d(C', D') = 0$ by definition.
Let $C', D' \in N_C$, then $lcs(C, D) = \top$ implies that $C' \neq D'$ which brings us to $simi_d(C', D') = pm_{def}(C', D') = 0$.
Otherwise, if $C' = \exists r.C^*$ and $D' = \exists s.D^*$ then

$$lcs(C', D') = \top \implies s \not\sqsubseteq_{\mathcal{R}} r \implies pm_{def}(s, r) \neq 1$$

$$\implies pm_{def}(s, r) = 0 \implies simi_d(C', D') = 0.$$

□

Next, we are proving a lemma which is used to show that *simi* is structural dependent. This property is not true for arbitrary weighting functions. However, we are proving it for a set of weighting functions which includes the default weighting function.

Lemma 17. *Let $D, E \in \mathcal{C}(\mathcal{ELH})$ and $(C_n)_n$ be a sequence of atoms with $\forall i, j \in \mathbb{N}, i \neq j : C_i \not\sqsubseteq C_j$. Furthermore, let g' be a weighting function where the infimum of the set of all weights (which is the image of the weighting function) is greater than zero, so*

$$inf\{g(C') \mid C' \in \mathcal{C}(\mathcal{ELH})\} > 0,$$

then

$$\lim_{n \to \infty} simi_d[\cdot, \cdot, g'](\bigsqcap_{i \leq n} C_i \sqcap D, \bigsqcap_{i \leq n} C_i \sqcap E) = 1.$$

Proof. Before we can use *simi*, the concept descriptions $\bigsqcap_{i\leq n} C_i \sqcap D$ and $\bigsqcap_{i\leq n} C_i \sqcap E$ need to be transformed into $\mathcal{ELH}$ normal form. Since $C_i \not\sqsubseteq C_j$, the part $\bigsqcap_{i\leq n} C_i$ remains the same. Only D and E would be syntactically changed (or even eliminated). However, we continue to use the notations D and E in the proof because we assumed that D and E are arbitrary concept descriptions and the fact whether or not they are in normal form does not matter in this proof.

For an arbitrary $i \leq n$, Lemma 14 implies that $simi_d(C_i, C_i) = 1$. Together with Lemma 2 we derive that

$$\bigoplus_{E'\in\widehat{E}} simi_d(C_i, E') \oplus \bigoplus_{j\leq n} simi_d(C_i, C_j) = 1.$$

Let $f : \widehat{D} \longrightarrow [0, 1]$ be a function defined by

$$f(D') := \bigoplus_{E'\in\widehat{E}} simi_d[\cdot,\cdot,g'](D', E') \oplus \bigoplus_{i\leq n} simi_d[\cdot,\cdot,g'](D', C_i).$$

Additionally, let $d := |\widehat{D}|$ and

$$g'_{inf} := inf\{g(C') \mid C' \in \mathcal{C}(\mathcal{ELH})\}.$$

Note that according to our premise, $g'_{inf} > 0$. Using the fact that for all $D' \in \widehat{D}$: $f(D') \in [0, 1]$ we obtain

$$\begin{aligned} simi_d[\cdot,\cdot,g'](\bigsqcap_{i\leq n} C_i \sqcap D, \bigsqcap_{i\leq n} C_i \sqcap E) &= \frac{\sum\limits_{i\leq n} g'(C_i) + \sum\limits_{D'\in\widehat{D}} g'(D')\cdot f(D')}{\sum\limits_{i\leq n} g'(C_i) + \sum\limits_{D'\in\widehat{D}} g'(D')} \\ &\geq \frac{\sum\limits_{i\leq n} g'(C_i)}{\sum\limits_{i\leq n} g'(C_i) + \sum\limits_{D'\in\widehat{D}} g'(D')} \\ &\geq \frac{n\cdot g'_{inf}}{(n+d)\cdot g'_{inf}} = \frac{n}{n+d}. \end{aligned}$$

Since $simi_d[\cdot,\cdot,g'](\bigsqcap_{i\leq n} C_i \sqcap D, \bigsqcap_{i\leq n} C_i \sqcap E) \leq 1$, we have

$$1 \geq \lim_{n\to\infty} simi_d[\cdot,\cdot,g'](\bigsqcap_{i\leq n} C_i \sqcap D, \bigsqcap_{i\leq n} C_i \sqcap E) \geq \lim_{n\to\infty} \frac{n}{n+d} = 1.$$

which implies

$$\lim_{n\to\infty} simi_d[\cdot,\cdot,g'](\bigsqcap_{i\leq n} C_i \sqcap D, \bigsqcap_{i\leq n} C_i \sqcap E) = 1.$$

□

It follows the main theorem, describing the properties of *simi*.

Theorem 1. *Simi is*

1. *symmetric,*
2. *equivalence invariant,*
3. *equivalence closed,*
4. *subsumption preserving,*
5. *bounded and*
6. $simi[\cdot, \cdot, pm_{def}]$ *is dissimilar closed.*
7. *Let* g' *be a weighting function with* $inf\{g'(C') \mid C' \in \mathcal{C}(\mathcal{ELH})\} > 0$. *Furthermore, let* $\otimes'$ *be a fuzzy connector such that for all sequences* $(x_n)_n$ *and* $(y_n)_n$ $(x_i, y_i \in [0,1])$ *with* $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = 1$, $\lim_{n\to\infty} x_n \otimes' y_n = 1$. *Then* $simi[\otimes', \cdot, \cdot, g']$ *is structural dependent.*

Proof. Let C, D, E be concept descriptions in $\mathcal{ELH}$ normal form.

1. Fuzzy connectors are defined to be commutative which implies that *simi* is symmetric.
2. Equivalence invariant: Let $C \equiv D$. The goal is to prove that $simi(C, E) = simi(D, E)$. Since we are working with a unique $\mathcal{ELH}$ normal form, $C \equiv D$ implies that C and D (or their corresponding normal forms) are syntactically equal. This trivially implies $simi(C, E) = simi(D, E)$.
3. Equivalence closed: With the usage of Lemma 14 and the fact that fuzzy connectors are equivalence closed we derive the chain
$$C \equiv D \iff C \sqsubseteq D \text{ and } D \sqsubseteq C \iff simi_d(C, D) = simi_d(D, C) = 1$$
$$\iff simi_d(C, D) \otimes simi_d(D, C) = 1 \iff simi(C, D) = 1.$$
4. Subsumption preserving: Let $C \sqsubseteq D \sqsubseteq E$. Using Lemma 14 and the weak monotonicity of fuzzy connectors, we derive
$$simi(C, D) = simi_d(C, D) \otimes simi_d(D, C) = simi_d(C, D) \otimes 1 = simi_d(C, D)$$
and analogues $simi(C, E) = simi_d(C, E)$. Lemma 15 implies that
$$simi(C, E) = simi_d(C, E) \leq simi_d(C, D) = simi(C, D).$$
5. Bounded: Let $simi(C, D) = 0$. Since fuzzy connectors are bounded, this implies that either $simi_d(C, D) = 0$ or $simi_d(D, C) = 0$. W.l.o.g. we assume that $simi_d(C, D) = 0$. The first part of Lemma 16 implies that $lcs(C, D) = \top$.

6. Dissimilar closed: Let $C \not\equiv \top, D \not\equiv \top$ and $lcs(C, D) \equiv \top$. The second part of Lemma 16 implies that $simi_d[\cdot, pm_{def}](C, D) = simi_d[\cdot, pm_{def}](D, C) = 0$. Since fuzzy connectors are grounded, we derive $simi(C, D) = 0 \otimes 0 = 0$.

7. Structural dependent: Let $D', E' \in N_A$ and $(C_n)_n$ be a sequence of atoms with $\forall i, j \in \mathbb{N}, i \neq j : C_i \not\sqsubseteq C_j$. Furthermore, let

$$simi_d^D(n) := simi_d[\cdot, \cdot, g'](\bigsqcap_{i \leq n} C_i \sqcap D', \bigsqcap_{i \leq n} C_i \sqcap E')$$

and

$$simi_d^E(n) := simi_d[\cdot, \cdot, g'](\bigsqcap_{i \leq n} C_i \sqcap E', \bigsqcap_{i \leq n} C_i \sqcap D').$$

Lemma 17 implies that

$$\lim_{n \to \infty} simi_d^D(n) = 1 \text{ and } \lim_{n \to \infty} simi_d^E(n) = 1.$$

Therefore,

$$\lim_{n \to \infty} simi[\otimes', \cdot, \cdot, g'](\bigsqcap_{i \leq n} C_i \sqcap D', \bigsqcap_{i \leq n} C_i \sqcap E') = simi_d^D(n) \otimes' simi_d^E(n) = 1.$$

□

An important property of *simi* is that it can be computed in time polynomial in the size of the concept descriptions to measure if all involved parameter functions are polynomial.

Lemma 18. *Simi can be computed in time polynomial in the size of the concept descriptions to measure if the specific fuzzy connector, the bounded t-conorm, the primitive measure and the weighting function can be computed in polynomial time.*

Proof. Let $C, D \in \mathcal{C}(\mathcal{ELH})$. First, let us observe that if $simi_d$ is polynomial, then *simi* is polynomial as well because we assume that the fuzzy connector is polynomial. We now take a look at the complexity of computing $simi_d(C, D)$. All atomic cases are polynomial because we assume that the primitive measure is polynomial. As for the case

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

it is easy to see that we have to compute $|\widehat{C}| \cdot |\widehat{D}|$ many similarity values. Since the depth of the recursion calls in the case of $C' = \exists r.E$ and $D' = \exists s.F$ is bounded by the size of C and D, the overall complexity of *simi* is bounded by $\mathcal{O}(poly(|C||D|))$ where $|C|$ is the number of occurrences of role names and concept names in C. □

5.4.1 Towards Triangle Inequality and Reverse Subsumption Preserving

In general *simi* is not reverse subsumption preserving and it does not fulfill the triangle inequality. The main reason for not fulfilling triangle inequality is that $simi_d(C, D)$ does not use the similarity values between the atoms of C and between the atoms of D. Before we present a formal proof which covers most parameter settings, we illustrate the problem by example. Let $n \geq 1$, $G, F_0, \ldots F_n$ be some atoms with $\forall i, j \leq n, i \neq j$: $simi_d(F_i, F_j) = \frac{n-1}{n}$ and $simi_d(G, F_i) = simi_d(F_i, G) = 0$. Furthermore, let

$$\begin{aligned} C &:= G \sqcap \prod_{i \leq n} F_i, \\ D &:= G \sqcap F_0, \\ E &:= G. \end{aligned}$$

In the following we use the default primitive measure, the default weighting function and the Hamacher product as fuzzy connector.
Since $C \sqsubseteq D \sqsubseteq E$, Lemma 15 implies that

$$simi_d(E, C) = simi_d(E, D) = simi_d(D, C) = 1.$$

Since the Hamacher product is a t-norm and 1 acts as neutral element for t-norms we have $simi(C, E) = simi_d(C, E)$, $simi(C, D) = simi_d(C, D)$ and $simi(D, E) = simi_d(D, E)$. Independent from the choice of t-conorm we have $simi_d(D, E) = \frac{1}{2}$, $simi_d(C, E) = \frac{1}{n+2}$ and

$$simi_d(C, D) = \frac{1 + 1 + n\frac{n-1}{n}}{n+2} = \frac{n+1}{n+2}.$$

The triangle inequality requires that

$$1 + simi(C, E) \geq simi(C, D) + simi(D, E).$$

In our case we have,

$$1 + \frac{1}{n+2} \geq \frac{n+1}{n+2} + \frac{1}{2}$$

which is false for $n \geq 4$ (for $n = 4$ we have $\frac{7}{6} \not\geq \frac{8}{6}$). Since all F_i are getting more similar the greater n is, the convergence of the similarity of C and D towards 1 is consistent with our intuition. The problem is the similarity value of C and E. It decrease with growing n because the cardinality of $\widehat{C}$ increases. This is in contradiction to our intuition because the F_i are very similar (close to total similar) and therefore the difference between E and C is not as great as the value of $simi_d$ indicates. The reason for the behaviour of $simi_d$ is that the way $simi_d$ is computed does not take usage of the similarity values among the F_i. It does not use the information that they are very similar to each other. An idea to overcome this problem is that a similarity measure could automatically weight the atoms using the similarity values between them so that for examples the weighted cardinality of $\widehat{C}$ is just a little higher than to 2 instead of $n + 2$. Also, we would need to

adjust the numerator in the definition of $simi_d$ as well. Further research is required to find a solution for this problem.
We now prove formally that $simi$ does not have the triangle inequality property. The counter example defined in the proof is derived from the example presented above. Additionally, the counter example can be used to prove that $simi$ is not reverse subsumption preserving.

Lemma 19. *Let $\otimes_z$ be a fuzzy connector.*

1. *If there exists an $n \in \mathbb{N}$ such that*
$$1 - [1 \otimes_z \frac{2n-1}{2n}] \; < \; [1 \otimes_z \frac{1}{2}] - [1 \otimes_z \frac{1}{n+1}] \tag{5.2}$$
then $simi[z, \cdot, \cdot, g_{def}]$ has not the triangle inequality property.

2. *If there exists an $n \in \mathbb{N}$ such that*
$$1 \otimes_z \frac{1}{n+1} < 1 \otimes_z \frac{1}{2} \tag{5.3}$$
then $simi[z, \cdot, \cdot, g_{def}]$ is not reverse subsumption preserving.

Proof.

1. We have to prove that there are concept descriptions C,D and E such that
$$1 + simi[z, \cdot, \cdot, g_{def}](D, E) < simi[z, \cdot, \cdot, g_{def}](D, C) + simi[z, \cdot, \cdot, g_{def}](C, E).$$
For easier reading, we use the notions $simi^*$ to refer to $simi[z, \cdot, \cdot, g_{def}]$ and $simi^*_d$ instead of $simi_d[\cdot, \cdot, g_{def}]$. Let $N_C := \{A_1, \ldots, A_n, B\}$, $N_r := \{r\}$ and we define
$$C := B \sqcap \exists r. \bigsqcap_{i \leq n} A_i,$$
$$D := B \sqcap \bigsqcap_{i \leq n} (\exists r. \bigsqcap_{j \leq n, j \neq i} A_j)$$
and $E := B$. The fact $C \sqsubseteq D \sqsubseteq E$ and Lemma 14 implies
$$simi^*_d(D, C) = simi^*_d(E, C) = simi^*_d(E, D) = 1.$$
In the following we omit writing weights because we are using the g_{def}. For $simi^*_d(C, E)$ and $simi^*_d(D, E)$ we have
$$simi^*_d(C, E) = \frac{simi^*_d(B, B) + simi^*_d(\exists r. \bigsqcap_{i \leq n} A_i, B)}{2} = \frac{1+0}{2} = \frac{1}{2}$$

and

$$simi_d^*(D,E) = \frac{simi_d^*(B,B) + \sum\limits_{i\leq n} simi_d^*(\exists r.\bigsqcap\limits_{j\leq n, j\neq i} A_j, B)}{n+1} = \frac{1+0}{n+1} = \frac{1}{n+1}.$$

As for $simi_d^*(C,D)$, we first take a look at $simi_d^*(\exists r.\bigsqcap_{i\leq n} A_i, \exists r.\bigsqcap_{j\leq n, j\neq i} A_j)$ for some $i \leq n$. With $x := \bigoplus_{j\leq n, j\neq i} pm(A_i, A_j)$, we have

$$simi_d^*(\exists r.\bigsqcap_{i\leq n} A_i, \exists r.\bigsqcap_{j\leq n, j\neq i} A_j) = pm(r,r)[w + (1-w)\frac{n-1+x}{n}]$$

$$= [w + (1-w)\frac{n-1+x}{n}] \geq [w + (1-w)\frac{n-1}{n}] \geq \frac{n-1}{n}.$$

The last inequality is a consequence from the fact that $\forall a,b \in [0,1] : a+(1-a)b \geq b$ which is proven through

$$a \geq ab \iff a - ab \geq 0 \iff a + b - ab \geq b \iff a + (1-a)b \geq b.$$

Using the monotonicity of t-conorms, we obtain

$$simi_d^*(C,D) = \frac{simi_d^*(B,B) + \bigoplus\limits_{i\leq n} simi_d^*(\exists r.\bigsqcap\limits_{i\leq n} A_i, \exists r.\bigsqcap\limits_{j\leq n, j\neq i} A_j)}{2}$$

$$\geq \frac{1 + \bigoplus\limits_{i\leq n}\frac{n-1}{n}}{2} \geq \frac{1+\frac{n-1}{n}}{2} = \frac{2n-1}{2n}.$$

This finally leads us to

$$simi^*(D,E) = 1 \otimes_z \frac{1}{n+1},$$

$$simi^*(C,E) = 1 \otimes_z \frac{1}{2}$$

and

$$simi^*(D,C) \geq 1 \otimes_z \frac{2n-1}{2n}.$$

We have

$$1 - simi^*(D,C) \leq 1 - [1 \otimes_z \frac{2n-1}{2n}]$$

where the initial assumption, Equation 5.2 implies that

$$1 - [1 \otimes_z \frac{2n-1}{2n}] < [1 \otimes_z \frac{1}{2}] - [1 \otimes_z \frac{1}{n+1}] = simi^*(C,E) - simi^*(D,E)$$

and therefore

$$1 - simi^*(D,C) < simi^*(C,E) - simi^*(D,E)$$

which can be reformulated to

$$1 + simi^*(D,E) < simi^*(D,C) + simi^*(C,E).$$

2. We use the same C, D and E as defined above We know that $C \sqsubseteq D \sqsubseteq E$, $simi^*(D, E) = 1 \otimes_z \frac{1}{n+1}$ and $simi^*(C, E) = 1 \otimes_z \frac{1}{2}$. Equation 5.3 directly implies $simi^*(D, E) < simi^*(C, E)$ which contradicts reverse subsumption preserving.

□

Because 1 is the neutral element for t-norms, with $n = 3$ all of them are fulfilling Equation 5.2 and 5.3 as well as the average and the Dice's Connector. Note that the proof could be done using an arbitrary weighting function instead of the default weighting function, however we decided not to prove this version, because it is much more complicated and it does not add to the understanding of the problem. A fuzzy connector not fulfilling Equation 5.2 is $\otimes_{ce}$ which is defined as follows, where $\otimes$ is an arbitrary t-norm and $x, y \in [0, 1]$:

$$x \otimes_{ce} y := \begin{cases} 1 & \text{if } x = y = 1 \\ \frac{1}{2} \cdot [x \otimes y] & \text{otherwise} \end{cases}.$$

Here, for $x \neq 1$ we have $\frac{1}{2} \geq [1 \otimes_{ce} x]$ which implies

$$[1 \otimes_{ce} \frac{1}{2}] - [1 \otimes_z \frac{1}{n+1}] \leq \frac{1}{2} \leq 1 - [1 \otimes_{ce} \frac{2n-1}{2n}].$$

5.4.2 Simi Generalizes the Jaccard Index

One of our aims for *simi* is to be able to generalize the $\mathcal{L}_0$ measure Jaccard Index (see 4.2.1). In this Section we show that this can be accomplished by using the default primitive measure, the default weighting function and the Hamacher product as fuzzy connector. Additionally, we prove that *simi* can generalize Dice's Coefficient (see 4.2.2) by using Dice's Connector as fuzzy connector.

We start by proving that $simi_d$ generalizes the function

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

and then prove that we obtain Dice's Coefficient and the *Jacc* by using Dice's Connector and the Hamacher product as fuzzy connector.

Lemma 20. *Let C, D be concept descriptions with $\widehat{C} \subseteq N_C$ and $\widehat{D} \subseteq N_C$, then*

$$simi_d[\cdot, pm_{def}, g_{def}, \cdot](C, D) = \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|}.$$

Proof. First, we observe that because we are using the default weighting function, $\sum_{C' \in \widehat{C}} g_{def}(C') = |\widehat{C}|$. Let $A \in \widehat{C}$. If $A \in \widehat{D}$ then Lemma 2 implies that

$$\bigoplus_{D' \in \widehat{D}} simi_d[\cdot, pm_{def}, g_{def}, \cdot](A, D') =$$

$$\bigoplus_{D' \in \widehat{D} \setminus \{A\}} simi_d[\cdot, pm_{def}, g_{def}, \cdot](A, D') \oplus pm_{def}(A, A) =$$

$$\bigoplus_{D' \in \widehat{D} \setminus \{A\}} simi_d[\cdot, pm_{def}, g_{def}, \cdot](A, D') \oplus 1 = 1.$$

Otherwise, if $A \notin \widehat{D}$ then $\forall D' \in \widehat{D}:\ simi_d(A, D') = pm_{def}(A, D') = 0$ which implies

$$\bigoplus_{D' \in \widehat{D}} simi_d[\cdot, pm_{def}, g_{def}, \cdot](A, D') = 0.$$

Therefore

$$simi_d[\cdot, pm_{def}, g_{def}, \cdot](C, D) = \frac{\sum_{C' \in \widehat{C}} 1 \cdot |\{C'\} \cap \widehat{D}|}{|\widehat{C}|} = \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|}.$$

□

The next lemma proves our claim that the Hamacher product and Dice's Connector (together with the function d) are the same as the Jaccard Index and Dice's Coefficient. Note that for this prove, concept descriptions are build from concept names only.

Lemma 21. *Let C, D be concept descriptions with $\widehat{C} \subseteq N_C$ and $\widehat{D} \subseteq N_C$, then*

$$Jacc(\widehat{C}, \widehat{D}) = d(C, D) \otimes_{H0} d(D, C)$$

and

$$Dice(\widehat{C}, \widehat{D}) = d(C, D) \otimes_{Dice} d(D, C).$$

Proof.

- Using the definition of the Hamacher product, we obtain

$$d(C, D) \otimes_{H0} d(D, C) = \frac{\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|} \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}|}}{\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|} + \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}|} - \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|} \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}|}} = \frac{\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}||\widehat{D}|}}{\frac{1}{|\widehat{C}|} + \frac{1}{|\widehat{D}|} - \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}||\widehat{D}|}}$$

$$= \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}| + |\widehat{C}| - |\widehat{C} \cap \widehat{D}|} = \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C} \cup \widehat{D}|} = Jacc(\widehat{C}, \widehat{D}).$$

- Using the definition of the Dice's Connector, we obtain

$$d(C, D) \otimes_{Dice} d(D, C) = \frac{2\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|} \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}|}}{\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}|} + \frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{D}|}} = \frac{2\frac{|\widehat{C} \cap \widehat{D}|}{|\widehat{C}||\widehat{D}|}}{\frac{1}{|\widehat{C}|} + \frac{1}{|\widehat{D}|}} = \frac{2|\widehat{C} \cap \widehat{D}|}{|\widehat{D}| + |\widehat{C}|}$$

$$= Dice(\widehat{C}, \widehat{D}).$$

□

6 Conclusion

In this thesis we presented several properties for description logic similarity measures, some that were already present in the relevant literature and some new ones. The new properties model a connection between subsumption and similarity, resulting in the properties subsumption preserving and reverse subsumption preserving, a connection between total dissimilarity and the least common subsumer, resulting in the properties dissimilar closed and bounded and an adoption of the basic rule of Tversky's feature model resulting in the property structural dependent.

An investigation of nine description-logic-similarity measures, towards what properties they fulfil, was conducted. A proof for every fulfilled property and a counterexample for every non-fulfilled property was presented. We used the counterexamples to point to, in our opinion, defects of the corresponding measure underlying our statement that properties can be used to analyse the general behaviour of a measure. As a result we conclude that, except for the $\mathcal{L}_0$ measures Jaccard Index and *Dice*, all analysed measure show cases of unintuitive behaviour.

We presented the $\mathcal{ELH}$ similarity measure *simi*. This measure is a generalization of the Jaccard Index. It can be tuned by parameters and is therefore regarded as a similarity framework. We proved that, almost independently from the choice of values for the parameters, *simi* is equivalence closed, equivalence invariant, dissimilar closed, bounded, structural dependent and subsumption preserving. Additionally, we proved that in general, *simi* is not reverse subsumption preserving and it does not fulfil the triangle inequality. Finally, we showed that *simi* can be computed in time polynomial in the size of the concept descriptions to measure.

6.1 Open Problems

The following list presents some open problems.

- One problem is to find a measure which fulfils the triangle inequality. None of the measures we investigated fulfils triangle inequality, except of the Jaccard Index which is a measure for the inexpressive description logic $\mathcal{L}_0$. As we argued in Chapter 3, triangle inequality is considered to be a natural property and therefore it is worth investigating how such a measure fulfilling triangle inequality would look like. In Section 5.4.1 we described our finding that modifying *simi* such that it fulfils triangle inequality involves using the similarity values between the atoms of every concept description to measure separately.

- Another problem is to deal with more expressive description logics but still try to fulfil as many properties as possible. Especially description logics with disjunction are challenging.

- Cyclic TBoxes are also an open problem. Since in general expansion of concept descriptions is not possible, the knowledge of the TBox needs to be considered when measuring similarity. Except for [dSF08] (which uses the good common subsumer which is computed with respect to a TBox), no measure we investigated is capable of including the knowledge of a cyclic TBox.

- Another question is to identify other general expectations of similarity measures which can be expressed as a property and used to analyse the overall behaviour of a measure.

- Let W be an arbitrary and fixed set of concept descriptions. A use case of similarity measures could be, given a concept description C which we call query concept, to search for the concept description $D \in W$ with the highest similarity regarding C, so $\forall E \in W : \ sim(C, E) \leq sim(C, D)$. The simplest solution to solve the query problem is to compute the similarity between C and all elements of W. Although this approach is linear in W it could be impracticable if W is very large. Therefore, more efficient algorithms are of interest. To find these, one could take advantage of the properties of the measure, either by using the properties defined in Chapter 3 or by identifying new ones. For example, since we know that W is fixed, we could compute the subsumption hierarchy of W in advance. If the measure is subsumption preserving and we have a concept description $D \in W$ with $C \sqsubseteq D$ then we do not need to compute the similarity of C and all subsumers of D.

List of Tables

Bibliography

[BCM$^+$03] Franz Baader, Diego Calvanese, Deborah L McGuinness, Daniele Nardi, and Peter F Patel-Schneider. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[BG97] Brian Bowdle and Dedre Gentner. Informativity and asymmetry in comparisons. *Cognitive Psychology*, 34(3):244–286, 1997. PMID: 9466832.

[BKT02] Sebastian Brandt, Ralf Küsters, and Anni-Yasmin Turhan. Approximating ALCN-Concept descriptions. In *Proceedings of the 2002 International Workshop on Description Logics (DL 2002)*, 2002.

[BST07] Franz Baader, Baris Sertkaya, and Anni-Yasmin Turhan. Computing the least common subsumer w.r.t. a background terminology. *Journal of Applied Logic*, 5(3):392–420, 2007.

[dFE05] Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. A semantic similarity measure for expressive description logics. In *Convegno Italiano di Logica Computazionale (CILC 2005)*, 2005.

[dFE06] Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. A dissimilarity measure for ALC concept descriptions. In *Proceedings of the ACM symposium on Applied computing*, SAC '06, page 1695–1699, 2006.

[Dic45] L.R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[dSF08] Claudia d'Amato, Steffen Staab, and Nicola Fanizzi. On the influence of description logics ontologies on conceptual similarity. In *Proceedings of the 16th Knowledge Engineering Conference (EKAW2008)*, volume 5268, pages 48–63, 2008.

[FD06] N Fanizzi and C D'amato. A similarity measure for the ALN description logic. In *Convegno Italiano di Logica Computazionale (CILC 2006)*, 2006.

[Gen07] R. Gentleman. Visualizing and distances using GO, 2007.

[GS04] Robert L. Goldstone and Ji Yun Son. Similarity. *Cambridge Handbook of Thinking and Reasoning*, pages 13–36, 2004.

[Jac01] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[Jan06] Krzysztof Janowicz. Sim-dl: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. *SeBGIS 2006, OTM Workshops 2006*, pages 1681–1692, 2006.

[Jan08] Krzysztof Janowicz. *Computing Semantic Similarity Among Geographic Feature Types Represented in Expressive Description Logics.* PhD thesis, Institute for Geoinformatics, University of Münster, Germany, 2008.

[JW09] Krzysztof Janowicz and Marc Wilkes. SIM-DLA: a novel semantic similarity measure for description logics reducing Inter-Concept to Inter-Instance similarity. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web Research and Applications*, pages 353–367, 2009.

[Kü00] Ralf Küsters. *Non-Standard Inferences in Description Logics.* PhD thesis, RWTH Aachen, 2000.

[LCL+03] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B Vitányi. The similarity metric. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 863—872, 2003.

[Lin98] Dekang Lin. An Information-Theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, 1998.

[Lip99] A. H. Lipkus. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26(1):263–265, 1999.

[LS04] Ming Li and M. Ronan Sleep. Melody classification using a similarity metric based on kolmogorov complexity. In *Proceedings of the Sound and Music Computing Conference (SMC'04)*, 2004.

[Neb90] Bernhard Nebel. Terminological reasoning is inherently intractable. *Artificial Intelligence*, 43:235—249, 1990.

[NJ03] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity - a review. *QSAR & Combinatorial Science*, 22:1006–1026, 2003.

[PFB+07] Catia Pesquita, Daniel Faria, Hugo Bastos, André O Falcão, and Francisco M Couto. Evaluating GO-based semantic similarity measures. In *Proceedings of the 10th Annual Bio-Ontologies Meeting*, volume 2007, pages 1–4, 2007.

[RMBB89] Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions On Systems Man And Cybernetics*, 19(1):17–30, 1989.

[Tve77] Amos Tversky. Features of similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.

Erklärung

Hiermit erkläre ich, dass ich die am heutigen Tag eingereichte Diplomarbeit mit dem Titel

A Framework for Semantic Invariant Similarity Measures for $\mathcal{ELH}$ Concept Descriptions

unter Betreuung von Dr.-Ing. Anni-Yasmin Turhan selbständig erarbeitet, verfasst und alle Zitate kenntlich gemacht habe. Andere als die von mir angegebenen Hilfsmittel wurden von mir nicht benutzt.

Dresden, den 07.02.2012 Unterschrift