Concept dissimilarity based on tree edit distance and morphological dilation

Dissimilarité entre concepts à partir d'une distance d'édition d'arbres et de dilatations morphologiques

Felix Distel
Jamal Atif
Isabelle Bloch

2014D001

février 2014

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

Concept Dissimilarity based on Tree Edit Distance and Morphological Dilation
Dissimilarité entre concepts à partir d'une distance d'édition d'arbres et de dilatations morphologiques

Felix Distel
Institute of Theoretical Computer Science
Faculty of Computer Science, TU Dresden, Germany
felix@tcs.inf.tu-dresden.de
Jamal Atif
Université Paris Sud, LRI, TAO, Orsay, France
jamal.atif@lri.fr
Isabelle Bloch
Institut Mines Telecom, Telecom ParisTech, CNRS LTCI, Paris, France
isabelle.bloch@telecom-paristech.fr

February 2014

Abstract

Several researchers have developed properties that ensure compatibility of a concept similarity or dissimilarity measure with the formal semantics of Description Logics. While these authors have highlighted the relevance of the triangle inequality, none of their proposed dissimilarity measures satisfy it. In this work we present several dissimilarity measures with this property: first, a simple dissimilarity measure, based on description trees for the lightweight Description Logic $\mathcal{EL}$; second, a general framework based on concept relaxations; third, an instantiation of the general framework using dilation operators from mathematical morphology, exploiting the link between Hausdorff distance and dilations using balls of the ground distance as structuring elements. A comparison between these definitions and their properties is provided as well.

Résumé

Plusieurs chercheurs se sont intéressés aux propriétés qui garantissent la compatibilité entre une mesure de similarité ou dissimilarité entre concepts et la sémantique des logiques de description. Alors que l'intérêt de l'inégalité triangulaire a été souligné, aucune mesure de dissimilarité existante ne la satisfait. Dans ce rapport, nous présentons plusieurs mesures de dissimilarité ayant cette propriété : nous proposons d'abord une mesure de dissimilarité simple, reposant sur les arbres de description pour la logique de description $\mathcal{EL}$; puis nous construisons un cadre général utilisant des opérateurs de dilatation morphologique, en exploitant le lien entre distance de Hausdorff et dilatation avec des éléments structurants définis comme des boules de la distance de base. Enfin, nous comparons ces définitions, ainsi que leurs propriétés.

Keywords: Distances between concepts, triangular inequality, description logics, mathematical morphology, dilation.
Mots clés : Distances entre concepts, inégalité triangulaire, logiques de descriptions, morphologie mathématique, dilatation.

1 Introduction

By nature description logics are well equipped for representing precise knowledge in a formal manner. As ontologies and description logics (DL) reach out to a broader audience some limitations become evident. In practice, it often occurs that two concepts have similar meanings, but no precise logical relationship can be established. Similarity measures, or dually dissimilarity measures are attempts to quantify the differences between concepts, and therefore providing ways to deal with these imprecisions. They are crucial in areas such as information retrieval in ontologies, ontology alignment, inductive logic programming and for some tasks in non-monotonic reasoning such as model-based revision or aggregation.

In a DL setting similarity can be defined between individuals, concepts, or even ontologies. In this work we focus exclusively on concept similarity. A large number of concept similarity measures has been developed, most of which are tailored to the specific needs of a particular field, such as biomedicine [19], or geospatial reasoning [12]. These approaches can be classified according to various criteria, such as the ones given in [8]. Initially, the quality of similarity measures has only been measured in terms of empirical evaluations. Increasingly, researchers are starting to look at theoretical properties that ensure compatibility of a similarity measure with the formal semantics of description logics. Works such as the ones in [9] and [14] list amongst others the properties of a metric, in particular the triangle inequality, as well as soundness with respect to equivalence and subsumption.

Among these criteria the triangle inequality has been somewhat disputed. This dispute originates in [27], where anecdotal evidence is provided that the human perception of similarity violates the triangle inequality. Tversky gives the example of the three countries Cuba, Jamaica and Russia, where Cuba and Russia are perceived to be very similar due to politics, Jamaica and Cuba are perceived to be very similar due to geography, but Jamaica and Russia are believed to be completely dissimilar. In [13] the point is made that this criticism applies only if the weight of the features that are compared is allowed to change. In the example the weight shifts from the politics feature to the feature geography. With the exception of [12], all the aforementioned works on similarity in DL agree that the triangle inequality is a desirable feature. Despite this, none of their proposed similarity measures satisfy it.

In the presence of the triangle inequality it is often more natural to talk about dissimilarity, instead of similarity, since this emphasizes the connection to metrics, as they are known from topology. In this work we present several dissimilarity measures with triangle inequality. In the case of the lightweight DL $\mathcal{EL}$, every concept has a unique representation as a tree. Therefore, any metric on trees yields a metric on concepts. In particular, a simple tree edit distance yields a dissimilarity measure with good theoretical properties.

In a second step, we give a general framework that can be used to construct concept dissimilarity measures with good theoretical properties, including the triangle inequality. The framework is based on concept relaxations, operators that can be used to successively make concepts more general. A directed distance between two concepts C and D can then be defined as the number of times D needs to be relaxed before it subsumes C. We show that the maximum of the two directed distances yields a good dissimilarity measure.

Finally, we instantiate the framework using dilation operators from mathematical morphology. These operators allow us to leverage a tree metric from the level of tree models to the level of DL concepts. This is based on the observation that it is "relatively" easy to define a distance between models [16, 20] or between domain elements [15], whereas on the concept level defining a dissimilarity is much harder. Our approach is based on the Hausdorff distance. In a metric space the Hausdorff distance can be used to leverage a metric between points to a metric between sets of points. We apply this idea to leverage a metric between models to a metric between concepts, by identifying concepts with their sets of models. We then exploit a result from mathematical morphology, that characterizes the Hausdorff distance by dilations using balls of the ground distance as structuring elements. On the concept level this dilation gives rise to a relaxation operator, which we use to instantiate our framework.

2 Preliminaries

2.1 Description Logics

We do not give a complete introduction to description logics, for more information consider [1]. Description logics are a family of knowledge representation formalisms. Every description logic $\mathcal{L}$ provides a set of *concept descriptions* $\mathsf{C}(\mathcal{L})$. Concept descriptions are recursively obtained from a set of *concept names* $\mathcal{N}_C$ and a set of *role names* $\mathcal{N}_R$ using concept constructors. The pair $\Sigma = (\mathcal{N}_C, \mathcal{N}_R)$ is called a *signature*. An overview over some frequently used constructors can be found in Table 1, where A denotes a concept name, while C and D denote arbitrary concept descriptions. The semantics of concept descriptions is defined using interpretations. An *interpretation* $\mathcal{I}$ is a pair $\mathcal{I} = (\Delta_{\mathcal{I}}, \cdot^{\mathcal{I}})$ consisting of an interpretation domain $\Delta_{\mathcal{I}}$ and an interpretation function $\cdot^{\mathcal{I}}$ which maps concept names to subsets of the domain $\Delta_{\mathcal{I}}$ and role names to binary relations on the domain.

Table 1: Semantics of some DL concept constructors.

Constructor	Syntax	Semantics
top concept	$\top$	$\Delta^{\mathcal{I}}$
bottom concept	$\bot$	$\emptyset$
concept name	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
disjunction	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
existential restriction	$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in C^{\mathcal{I}} : (x, y) \in r^{\mathcal{I}}\}$

A concept description C is said to *subsume* a concept description D if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds for every interpretation $\mathcal{I}$. This is denoted by $C \sqsubseteq D$. We say that C and D are *equivalent* (denoted by $C \equiv D$) if both $C \sqsubseteq D$ and $D \sqsubseteq C$ hold. A concept description E is called a *common subsumer* of C and D if it subsumes both C and D.

We call a pair $(\mathcal{I}, x)$ where $\mathcal{I}$ is a DL interpretation and $x \in \Delta^{\mathcal{I}}$ is a domain element a *pointed interpretation*. We denote the set of all pointed interpretations for a given signature Σ by Int_{Σ}. We call $(\mathcal{I}, x)$ a *(pointed) model of* C if $x \in C^{\mathcal{I}}$. For every concept description $C \in \mathcal{L}$ we denote the set of all pointed models of C by $\mathrm{Mod}\,(C)$.

In description logics axioms are typically stored in ontologies, which can be divided into TBoxes and ABoxes. We define our measures in the absence of background ontologies. In the conclusion we give a brief discussion about how they can be adapted to take TBoxes into account.

2.2 Similarity and Dissimilarity on Concepts

When similarity measures were first investigated within the DL community, researcher mainly focused on adaptations of existing measures from other fields (cf. [8] for a survey). The quality of these measures was mainly examined in an empirical way, showing that they perform well in a given setting, but providing little transferable insight. It was only in [9] that qualitative criteria were developed, based on criteria given in [7]. The following definition is slightly adapted to dissimilarity between concepts.

Definition 1 (Dissimilarity [7])**.** *Let $\mathcal{L}$ be a DL language. A function $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called a* dissimilarity measure *if it satisfies the following properties for all $C, D \in \mathsf{C}(\mathcal{L})$.*

- positiveness: $d(C, D) \geq 0$
- reflexivity: $d(C, C) = 0$*, and*
- symmetry: $d(C, D) = d(D, C)$.

These properties can be expected to hold for any dissimilarity measure. In a description logics context it should also be compatible with the semantics of the logic. To ensure this, the additional properties of *equivalence soundness* and *(strict) monotonicity* were introduced in [9].[1]

Definition 2 (Equivalence Soundness)**.** *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called* equivalence sound *if for all $C, D, E \in \mathsf{C}(\mathcal{L})$*

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

[1] Additionally, the properties of *soundness* and *dissimilarity incompatibility* were mentioned, however these were never formally defined.

Notice that in [14] equivalence soundness is referred to as *equivalence invariance*.

Definition 3 ((Strict) Monotonicity)**.** *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called* (strictly) monotone *if for all $C, D, E \in \mathsf{C}(\mathcal{L})$ that satisfy*

- *every common subsumer of C and E also subsumes D,*
- *there is a common subsumer of C and D that does not subsume E,*

it holds that $d(C, D) \leq d(C, E)$, respectively $d(C, D) < d(C, E)$.

The intuition behind monotonicity is that concepts with more common features should be less dissimilar than concepts with few common features, and that common subsumers are a way to extract commonalities from concepts. For example the concepts

$$\begin{aligned} \mathsf{F} &:= \mathsf{Male} \sqcap \exists \mathsf{hasChild}.\top \\ \mathsf{HoJ} &:= \mathsf{Male} \sqcap \exists \mathsf{marriedTo}.(\mathsf{Female} \sqcap \mathsf{Judge}) \end{aligned} \tag{1}$$

share the common feature Male which is also a common subsumer for them. Therefore, the dissimilarity between F and HoJ should be smaller that say the dissimilarity between F and Female, whose only common subsumer is $\top$. An argument against monotonicity is that it is very dependent on the language. In a more expressive language, e.g. a DL with disjunction, F and Female have the common subsumer F $\sqcup$ Female and the property will fail to detect that they do not have common features.

More recently, an extended set of properties has been proposed in [14]. These properties are originally stated for similarity measures, here we present their equivalents for dissimilarity measures.

Definition 4 ([14])**.** *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called*

- equivalence closed *if $d(C, D) = 0 \implies C \equiv D$,*
- subsumption preserving *if $C \sqsubseteq D \sqsubseteq E \implies d(C, D) \leq d(C, E)$*
- reverse subsumption preserving *if $C \sqsubseteq D \sqsubseteq E \implies d(D, E) \leq d(C, E)$*
- structurally dependent *if for all sequences $(C_n)_n$ of atoms with $C_i \not\sqsubseteq C_j$ for all $i, j \in \mathbb{N}$, $i \neq j$ the concepts*

  $$D_n = \bigsqcap_{i \leq n} C_i \sqcap D, E_n = \bigsqcap_{i \leq n} C_i \sqcap E$$

  satisfy $\lim_{n\to\infty} d(D_n, E_n) = 0$ *for all $C, D, E \in \mathsf{C}(\mathcal{L})$.*
- *We say that d fulfills the* triangle inequality *if $d(C, E) \leq d(C, D) + d(D, E)$ for all $C, D, E \in \mathsf{C}(\mathcal{L})$.*

Structural dependence is another attempt to formalize the idea that dissimilarity should decrease as the number of common features increases. In fields such as topology or geometry it is generally accepted that distances should be measured using metrics, or at least pseudo-metrics. A *pseudo-metric* δ_1 is a binary operator that is non-negative, symmetric, reflexive and respects the triangle inequality. A *metric* δ_2, is a pseudo-metric that additionally is strict, i.e. $\delta_2(x, y) = 0$ implies $x = y$. The bottleneck preventing most dissimilarity measures from being metrics is the triangle inequality. Notice, that any equivalence sound dissimilarity measure that fulfills the triangle inequality yields a pseudo-metric on $\mathsf{C}(\mathcal{L})$ modulo equivalence. If, additionally, it is equivalence closed we obtain a metric. Unfortunately, even the measures presented in [14] and [9] with their otherwise good theoretical properties do not satisfy the triangle inequality.

3 $\mathcal{EL}$ and Distances on Trees

3.1 From Concepts to Description Trees

$\mathcal{EL}$ denotes the simple description logic, that allows only for conjunction $\sqcap$, existential restrictions $\exists$ and the top concept $\top$. Despite its limited expressivity, it is a very popular choice among ontology engineers, as it is tractable [2] and forms the basis of the OWL 2 profile OWL 2 $\mathcal{EL}$ [17].

It has been noted early on that $\mathcal{EL}$ concept descriptions can appropriately be represented as labeled trees, often called *$\mathcal{EL}$ description trees* [3]. An $\mathcal{EL}$ description tree is of the form $\mathcal{G} = (V, E, v_0, \ell)$, where $\mathcal{G}$ is a tree with root v_0. The labeling function ℓ associates nodes with sets of concept names from $\mathcal{N}_C$, and edges with role names from $\mathcal{N}_R$. An $\mathcal{EL}$ concept description of the form

$$C \equiv P_1 \sqcap \cdots \sqcap P_n \sqcap \exists r_1.C_1 \sqcap \cdots \sqcap \exists r_m.C_m \tag{2}$$

with $P_i \in \mathcal{N}_C \cup \{\top\}$, can be translated into a description tree by labeling the root node v_0 with $\{P_1, \ldots, P_n\}$, creating an r_j successor, and then proceeding inductively by expanding C_j for the r_j-successor node for all $j \in \{1, \ldots, m\}$. As an example consider the concept description

$$\mathsf{Person} \sqcap \exists \mathsf{c}.\mathsf{Male} \sqcap \exists \mathsf{c}.\exists \mathsf{c}.\mathsf{Female}. \tag{3}$$

Its description tree is depicted in Figure 1.

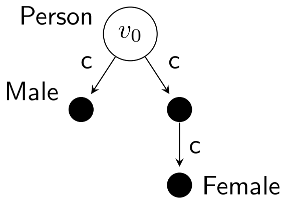

Figure 1: $\mathcal{EL}$-Description Tree for (3).

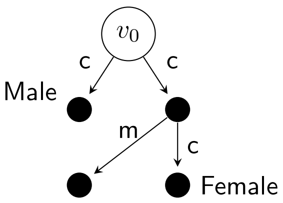

Figure 2: Figure 1 after two edits: Removing the label at v_0 and adding an m-edge.

Due to this tight link between $\mathcal{EL}$-concepts and trees it is natural to use distance measures defined on trees. Examples for existing metrics defined on trees are tree edit distances and tree alignment distances [4]. In the following we assume that we are given a metric δ on the space of all $\mathcal{EL}$-concept descriptions and try to lift it to a dissimilarity measure between $\mathcal{EL}$-concepts. One cannot simply define dissimilarity between two concepts as the distance between their description trees. This would violate equivalence soundness, since a concept can have multiple equivalent representations and thus multiple description trees.

A frequently used workaround is restricting to the normal form, introduced in [3]. An $\mathcal{EL}$-concept is in *normal form* if it is of the form (2) with the additional requirement that no subsumption relation holds between two distinct conjuncts and that all C_j, $j \in \{1, \ldots, m\}$, are also in normal form. The normal form is unique up to reordering of conjuncts, and since reordering of conjuncts does not change the description tree, it yields a unique description tree for each equivalence class of $\mathcal{EL}$-concepts.

Definition 5 (Dissimilarity from Tree Metric)**.** *Let δ be a metric on the space of all $\mathcal{EL}$-description trees. We define a dissimilarity measure $d_\delta^{tree}(C, D) = \delta(T_C, T_D)$ where T_C and T_D are the $\mathcal{EL}$-description trees of the normal form of C and D, respectively.*

It follows immediately from the uniqueness of the $\mathcal{EL}$-description trees for normal forms that d_δ^{tree} is equivalence sound. Since δ is a metric, and thus positive, reflexive, symmetric, strict and satisfying the triangle inequality, we obtain immediately that d_δ^{tree} is positive, reflexive, symmetric, equivalence closed, and satisfies the triangle inequality. However, d_δ^{tree} does not, in general have the properties of monotonicity, structural dependence and (reverse) subsumption preserving.

3.2 Tree Edit Distances

Among the various approaches for defining distances between labeled trees arguably the most widely used are *tree edit distances*, first introduced in [26]. They have been successfully applied in fields as diverse as computer vision, natural language processing, and computational biology (cf. [4] for a survey).

The idea of tree edit distances is simple. One defines a set of edits, each with its associated cost. The *tree edit distance* is then the minimal total cost of transforming one tree into another. If each edit is reversible at the same cost, then the tree edit distance will be a metric. Which set of edit operations is chosen depends on the particular application.

In this paper we use a particularly simple tree edit distance δ^{edit}, allowing for two simple operations *addLabel* and *addNode*, as well as their inverses *delLabel* and *delNode*:

- the operation *addLabel* adds a concept name to a node in the tree,
- *delLabel* removes a concept name from a node,
- for any role r an (unlabeled) r-successor can be added to a node using *addNode*, and
- any unlabeled node without successors can be deleted using *delNode*.

We assign the same cost 1 to each edit. Therefore, the tree edit distance δ^{edit} between two trees T_1 and T_2 is the minimal number of tree edit operations that need to be performed to transform T_1 into T_2. Refer to Figures 1 and 2 for an example.

Using a characterization of subsumption between $\mathcal{EL}$-concepts through homomorphisms between their description graphs, it is straightforward to prove that the dissimilarity measure $d^{\text{tree}}_{\delta^{\text{edit}}}$ is subsumption preserving and reverse subsumption preserving, in addition to the properties that all dissimilarity measures obtained from Definition 5 have.

This shows that for the description logic $\mathcal{EL}$, it is possible to define dissimilarity measures with good theoretical qualities based on metrics on labeled trees. Unlike $\mathcal{EL}$, concepts written in more expressive description logics lack a simple characterization as labeled trees. It is therefore not possible to transfer the ideas from this section to more expressive logics in a straightforward way. In Section 6 we will show how, by working on the space of pointed tree-shaped models instead of the space of concept descriptions, we can still make use of tree distances to define dissimilarity measures.

4 General Framework

In this section we provide a general framework for defining dissimilarity measures. We show that all dissimilarity measures obtained within this framework have all properties from Section 2.2, except monotonicity and structural dependence. The framework is based on *concept relaxation operators*, operators that allow a stepwise generalization of concepts. In Sections 6 and 7 we will instantiate this framework using relaxation operators derived from distances on the model space using mathematical morphology.

Definition 6 (Relaxation)**.** *A* (concept) relaxation *is an operator* $\rho\colon \mathsf{C}(\mathcal{L}) \to \mathsf{C}(\mathcal{L})$ *that satisfies the following three properties for all* $C, D \in \mathcal{L}$*.*

1. ρ *is* non-decreasing*, i.e.* $C \sqsubseteq D$ *implies* $\rho(C) \sqsubseteq \rho(D)$*,*
2. ρ *is* extensive*, i.e.* $C \sqsubseteq \rho(C)$*, and*
3. ρ *is* exhaustive*, i.e.* $\exists k \in \mathbb{N}_0 \colon \top \sqsubseteq \rho^k(C)$*,*
   where ρ^k *denotes* ρ *applied* k *times, and* ρ^0 *is the identity.*

Notice that extensivity and exhaustivity together entail *strong extensivity*, i.e. for all $C \in \mathsf{C}(\mathcal{L})$ it holds that $C \sqsubset \rho(C)$ or $C \equiv \top$. A trivial relaxation is the operator $\rho_\top$ that maps every concept to $\top$. Another relaxation in $\mathcal{EL}$ is the operator ρ_{depth} that reduces the role depth of each concept by 1, simply by pruning the description tree. Figure 3 depicts two applications of ρ_{depth} to (3).

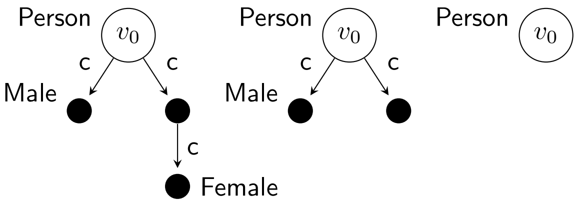

Figure 3: Consecutive application of ρ_{depth} to (3).

A dissimilarity measure that is equivalence sound and closed should have the value $d(C, D) = 0$ if and only if $C \equiv D$, i.e. iff $C \sqsubseteq D$ and $D \sqsubseteq C$. Like in [14] and [25] we first introduce directed measures d^d_ρ that capture how "far" D is from being a subsumer of C, and vice versa. If both $C \sqsubseteq D$ and $D \sqsubseteq C$ hold, then both directed measures will be 0. The directed measure $d^d_\rho(C, D)$ counts how often we need to successively relax D to reach a subsumer of C. If we think of concepts in terms of sets of individuals, then the intuition behind successive relaxations can be visualized as in Figure 4. This also corresponds to the idea of applying successive dilations, as detailed later.

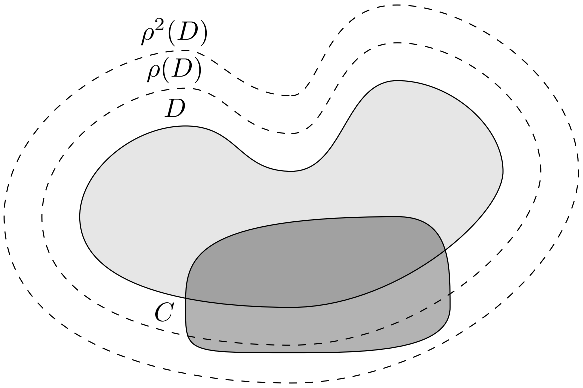

Figure 4: D needs to be relaxed twice before it subsumes C, i.e. $d^d_\rho(C, D) = 2$.

Definition 7 (Directed measure)**.** *Let ρ be a relaxation on $\mathsf{C}(\mathcal{L})$. For $C, D \in \mathsf{C}(\mathcal{L})$ the* directed measure $d^d_\rho(C, D)$ *is defined as*

$$d^d_\rho(C, D) = \min\{k \in \mathbb{N}_0 \mid C \sqsubseteq \rho^k(D)\},$$

where ρ^k denotes ρ applied k times, and ρ^0 is the identity.

The directed measure is always finite because of the exhaustiveness of the ρ operator. We can then define the *relaxation dissimilarity* based on a relaxation operator simply as the maximum of the two directed measures.

Definition 8 (Relaxation Dissimilarity)**.** *Let $\rho \colon \mathcal{L} \to \mathcal{L}$ be a relaxation on $\mathsf{C}(\mathcal{L})$. For two concepts C and D the* relaxation dissimilarity $d_\rho(C, D)$ *is defined as*

$$d_\rho(C, D) = \max\{d^d_\rho(C, D), d^d_\rho(D, C)\}.$$

Lemma 1. *For every relaxation ρ the operator d_ρ is a dissimilarity measure, that is equivalence sound, equivalence closed, subsumption preserving and reverse subsumption preserving, and satisfies the triangle inequality.*

Proof. Positiveness, reflexivity and symmetry follow trivially from Definitions 7 and 8, and therefore d_ρ is a dissimilarity measure.

We have the following chain of equivalences: $C \equiv D$, iff $C \sqsubseteq D$ and $D \sqsubseteq C$, iff $C \sqsubseteq \rho^0(D)$ and $D \sqsubseteq \rho^0(C)$, iff $d^d_\rho(C, D) = d^d_\rho(D, C) = 0$, iff $d_\rho(C, D) = 0$. Thus d_ρ is both equivalence sound and equivalence closed.

To prove the triangle inequality, let C, D, E be concept descriptions and let $d_\rho(C,D) = d_1$, $d_\rho(D,E) = d_2$. Then in particular, $d^d_\rho(C,D) \leq d_1$ and thus $C \sqsubseteq \rho^{d_1}(D)$ by extensivity. Similarly, we obtain $D \sqsubseteq \rho^{d_2}(E)$. Using non-decreasingness of relaxations we obtain from the latter

$$\rho^{d_1}(D) \sqsubseteq \rho^{d_1+d_2}(E)$$

and therefore $C \sqsubseteq \rho^{d_1+d_2}(E)$, i.e. $d^d_\rho(C,E) \leq d_1 + d_2$. Analogously, it can be shown that $d^d_\rho(E,C) \leq d_1 + d_2$ and thus $d_\rho(C,E) \leq d_1 + d_2 = d_\rho(C,D) + d_\rho(D,E)$.

To show subsumption preservingness let $C \sqsubseteq D \sqsubseteq E$ with $d_\rho(C,E) = d$. Then in particular, $E \sqsubseteq \rho^d(C)$ and thus also $D \sqsubseteq \rho^d(C)$. On the other hand, $C \sqsubseteq \rho^0(D) \sqsubseteq \rho^d(D)$ by extensivity, which yields $d_\rho(C,D) \leq k = d_\rho(C,E)$, which proves subsumption preservingness. □

Lemma 1 shows that relaxation operators yield dissimilarity measures with good theoretical properties, even for simple relaxations such as $\rho_\top$ or ρ_{depth}. Obviously, $\rho_\top$ yields a very coarse dissimilarity measure that is 0 iff the concepts are equivalent and 1 otherwise. The dissimilarity measure ρ_{depth} is also very coarse, as it only looks at the first depth-level where the concepts differ, thereby giving higher weight to features at a smaller depth. For example, if we compare F and HoJ from (1) to the concept $\exists\mathsf{hasChild}.\top$ the value will be 2 in both cases, since the change occurs at the lowest level, in the concept name Male. This is counter-intuitive, since F and $\exists\mathsf{hasChild}.\top$ share more common features than HoJ and $\exists\mathsf{hasChild}.\top$. This effect cannot occur with the relaxation obtained from a tree edit distance on models which we will introduce in Section 7, since the tree edit distance puts equal weight on each edit, independent of the depth in the tree at which it occurs.

5 Existing Metrics for Other Logics

Outside of description logics several works have proposed metrics between logical objects. Works such as the one in [18, 21] exploit the fact that is relatively easy to define a metric on ground expressions in first order logic. They extend these ground distances to sets of atoms, or Herbrand interpretations using constructions such as Hausdorff distances or Manhattan distances.

In some cases it is straightforward to define a distance between two terms if one is a generalization of the other. To obtain a distance between two arbitrary terms one can simply use the sum of the distances to their least general common generalization. In a general form the author in [5] has presented this idea as the classical distance in graded lattices. It is used to define a distance between first order literals in [11], who then generalizes it to a distance between clauses using the Hausdorff metric. This idea can also be extended to cases where there is no unique minimally general generalization [10].

6 Relaxation Operators from Dilations

6.1 Mathematical Morphology and the Hausdorff distance

Mathematical Morphology is a theory of spatial transformation, mainly developed in digital image processing [24]. Its deterministic part relies on the algebraic framework of complete lattices [22], thus extending its scope to many domains of information processing, including logics [6]. At the heart of mathematical morphology are two classes of operators: *dilations* and *erosions*. They are defined in the general algebraic setting of complete lattices as operators that commute with the supremum and the infimum, respectively. Particular forms of dilations and erosions involve the notion of "structuring element", representing a binary relation between elements of the underlying space, or a neighborhood of each element. In the case of metric spaces, the structuring element can be a ball of a distance, and dilations and erosions can then be defined as follows. Let (M,δ) be a metric space and $\lambda \in \mathbb{R}$ a real number. The dilation $\mathrm{dil}_{\delta,\lambda}$ and the erosion $\mathrm{ero}_{\delta,\lambda}$ by a ball of δ of radius λ are then defined as operators on the power set of M:

$$\begin{aligned} \mathrm{dil}_{\delta,\lambda}(S) &= \{x \in M \mid \exists y \in S\colon \delta(x,y) \leq \lambda\} \\ \mathrm{ero}_{\delta,\lambda}(S) &= \{x \in M \mid \forall y \in M\colon \delta(x,y) \leq \lambda \implies y \in S\} \end{aligned} \tag{4}$$

for all $S \subseteq M$. For erosions and dilations by a unit ball, i.e. for $\lambda = 1$, we simply write dil_δ and ero_δ. Additionally to the commutativity with the supremum for $\mathrm{dil}_{\delta,\lambda}$, and with the infimum for $\mathrm{ero}_{\delta,\lambda}$, these operations have important properties that will be used in the following: they are increasing with respect to S, $\mathrm{dil}_{\delta,\lambda}$ is increasing and $\mathrm{ero}_{\delta,\lambda}$ is decreasing with respect to λ, $\mathrm{dil}_{\delta,\lambda}$ is extensive (i.e. $S \subseteq \mathrm{dil}_{\delta,\lambda}(S)$) and $\mathrm{ero}_{\delta,\lambda}$ is anti-extensive (i.e. $\mathrm{ero}_{\delta,\lambda}(S) \subseteq S$). Other properties may hold depending on the ground distance δ.

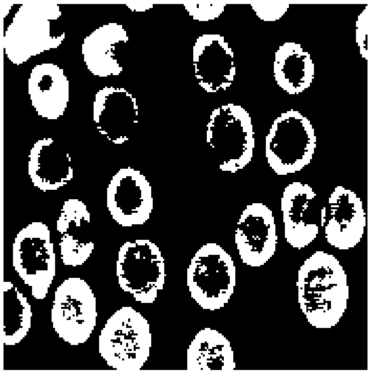

Figure 5: Subset S of binary digital image in white.

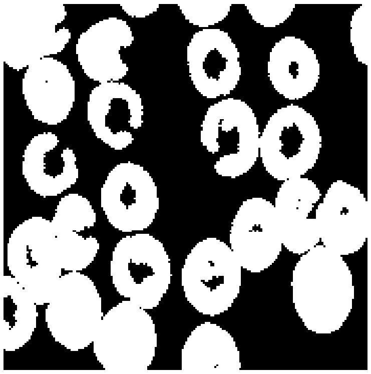

Figure 6: Its dilation by a ball of radius 3 pixels.

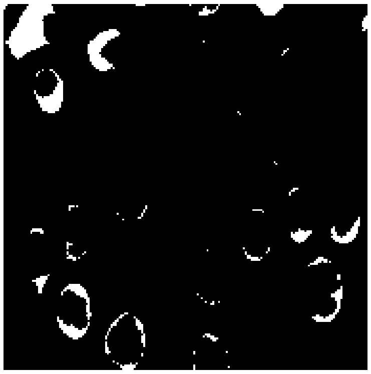

Figure 7: Its erosion by a ball of radius 3 pixels.

Dilations and erosions are illustrated in Figures 5–7 for a binary image. There is an intuitive connection between dilations and relaxations, e.g. both are extensive and monotone. In Section 6.2 we shall exploit this connection for the purpose of defining relaxations based on dilations. In that section we will mostly be interested in discrete metrics, i.e. metrics that only take values from $\mathbb{N}$. For these metrics, arbitrary dilations can be characterized by successive dilations with a unit ball, provided that the betweenness property holds.

Definition 9 (Betweenness Property). *Let δ be a discrete metric on M. We say that δ has the* betweenness property *if for all $x, y \in M$ and all $k \in \{0, 1, \dots, \delta(x, y)\}$ there exists $z \in M$ such that $\delta(x, z) = k$ and $\delta(z, y) = \delta(x, y) - k$.*

Simple induction over λ can be used to prove the following result.

Lemma 2. *If δ is a discrete metric with betweenness property, then for all sets $X \subseteq M$ and all $\lambda \in \mathbb{N}$ it holds that* $\mathrm{dil}_{\delta,\lambda}(X) = (\mathrm{dil}_{\delta,1})^\lambda(X)$.

But first, we point out a connection between dilations and the classical Hausdorff distance mentioned in [24]. Remember that for a metric space (M, δ) the *Hausdorff distance* h is a metric on the power set of M. For $m \in M$ and non-empty compact subsets $X, Y \subseteq M$ let $\delta(x, Y) = \inf\{\delta(x, y) \mid y \in Y\}$, and define

$$h_\delta(X, Y) = \max\left\{\sup_{x \in X} \delta(x, Y), \sup_{y \in Y} \delta(y, X)\right\}.$$

The Hausdorff distance can then be expressed in terms of dilations as described by the following lemma.

Lemma 3 ([24]). *For all non-empty compact sets $X, Y \subseteq M$*

$$h_\delta(X, Y) = \max(h_{d,\delta}(X, Y), h_{d,\delta}(Y, X))$$

where $h_{d,\delta}(X, Y) = \inf\{\lambda \mid X \subseteq \mathrm{dil}_{\delta,\lambda}(Y)\}$ and where $h_{d,\delta}(Y, X)$ is defined analogously.

6.2 From Model Space to Concept Space

The idea behind the Hausdorff distance is to lift a metric defined on points to a metric on sets of points. In a similar way we demonstrate how a metric can be lifted from pointed models to concept descriptions. Let δ be a metric on the space of pointed models Int_Σ. In a logic $\mathcal{L}$ with the tree model property, no two concept descriptions C and D

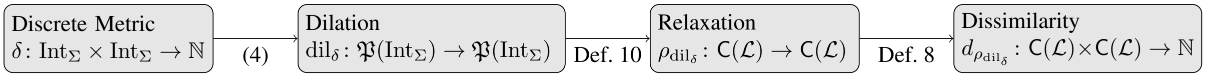
Discrete Metric
$\delta\colon \mathrm{Int}_\Sigma \times \mathrm{Int}_\Sigma \to \mathbb{N}$
(4)
Dilation
$\mathrm{dil}_\delta\colon \mathfrak{P}(\mathrm{Int}_\Sigma) \to \mathfrak{P}(\mathrm{Int}_\Sigma)$
Def. 10
Relaxation
$\rho_{\mathrm{dil}_\delta}\colon \mathsf{C}(\mathcal{L}) \to \mathsf{C}(\mathcal{L})$
Def. 8
Dissimilarity
$d_{\rho_{\mathrm{dil}_\delta}}\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{N}$

Figure 8: From discrete metrics on Int_Σ to dissimilarity measures.

can have the same sets of pointed models $\mathrm{Mod}\,(C)$ and $\mathrm{Mod}\,(D)$, unless they are equivalent. In fact it even holds that

$$C \sqsubseteq D \iff \mathrm{Mod}\,(C) \subseteq \mathrm{Mod}\,(D)\,. \tag{5}$$

In particular, the description logic $\mathcal{ALC}$ and all of its fragments have the tree model property [23]. Therefore, it is natural to use the Hausdorff distance between sets of pointed models to define dissimilarity measures between concepts, and to use dilations of sets to define relaxations on concepts. There are, however, two issues with this approach. First, for two concepts C, D it is usually not possible to compute the suprema and infima in the definition of the Hausdorff distance h_δ, since the sets $\mathrm{Mod}\,(C)$ and $\mathrm{Mod}\,(D)$ are usually infinite. Furthermore, $\mathrm{Mod}\,(C)$ and $\mathrm{Mod}\,(D)$ are not necessarily compact. Secondly, not every set of pointed models can be obtained as the models of a concept description. In order to be able to obtain a relaxation from a dilation in a straightforward way, we need to focus on dilations that are expressible in our logic.

Definition 10 (Expressibility)**.** *Let $\omega\colon \mathfrak{P}(\mathrm{Int}_\Sigma) \to \mathfrak{P}(\mathrm{Int}_\Sigma)$ be a unary operator. We say that ω is* expressible in $\mathcal{L}$ *if for every $C \in \mathsf{C}(\mathcal{L})$ there exists some $D_C \in \mathsf{C}(\mathcal{L})$ such that*

$$\mathrm{Mod}\,(D_C) = \omega(\mathrm{Mod}\,(C)).$$

If $\mathcal{L}$ has the tree model property, then D_C is unique up to equivalence, provided that it exists.

If ω is expressible in $\mathcal{L}$ then we can define an operator $\rho_\omega\colon \mathsf{C}(\mathcal{L}) \to \mathsf{C}(\mathcal{L})$ that maps C to D_C for every concept $C \in \mathsf{C}(\mathcal{L})$.

An example for a Hausdorff-based dilation that is expressible in a description logic will be given in Section 7. The following result is an immediate consequence of the tree model property, the definition of subsumption, and the fact that dilations are non-decreasing and extensive.

Lemma 4. *Let $\mathcal{L}$ be a logic that has the tree-model property. Let* dil *be a dilation on $\mathsf{C}(\mathcal{L})$. If* dil *is expressible in $\mathcal{L}$ then ρ_{dil} is non-decreasing and extensive.*

Notice that in some logics such as $\mathcal{EL}$, a concept can have only finitely many subsumers. For these logics, if ρ_{dil} is strongly extensive, then it is also exhaustive, and therefore a relaxation.

For discrete metrics, we now have all the necessary definitions to obtain a dissimilarity measure on concepts, according to Figure 8. The following theorem shows that the dissimilarity measure obtained in this way can be viewed as a Hausdorff distance, if we identify concepts with their sets of models according to (5).

Theorem 1. *Let δ be a discrete metric on Int_Σ and $C, D \in \mathsf{C}(\mathcal{L})$ concept descriptions such that $\mathrm{Mod}\,(C)$, $\mathrm{Mod}\,(D)$ are compact. Let dil_δ, $\rho_{\mathrm{dil}_\delta}$ and $d_{\rho_{\mathrm{dil}_\delta}}$ be defined as in Equation (4), Definition 10 and Definition 8, respectively. If δ satisfies the betweenness property, dil_δ is expressible in $\mathcal{L}$ and $\rho_{\mathrm{dil}_\delta}$ is exhaustive, then $\rho_{\mathrm{dil}_\delta}$ is a relaxation and*

$$d_{\rho_{\mathrm{dil}_\delta}}(C, D) = h_\delta(\mathrm{Mod}\,(C)\,, \mathrm{Mod}\,(D)).$$

Proof. Lemma 3 states that

$$h_\delta(\mathrm{Mod}\,(C)\,, \mathrm{Mod}\,(D)) = \max(H_{CD}, H_{DC})$$

where $H_{CD} = \inf\{\lambda \mid \mathrm{Mod}\,(C) \subseteq \mathrm{dil}_{\delta,\lambda}(\mathrm{Mod}\,(D))\}$. The betweenness property and expressibility of dil_δ entail

$$\begin{aligned}\mathrm{dil}_{\delta,\lambda}(\mathrm{Mod}\,(D)) &= (\mathrm{dil}_\delta)^\lambda(\mathrm{Mod}\,(D)) \\ &= \mathrm{Mod}\left(\rho^\lambda_{\mathrm{dil}_\delta}(D)\right).\end{aligned}$$

Together with (5) this yields

$$H_{CD} = \inf\left\{\lambda \mid C \sqsubseteq \rho^\lambda_{\mathrm{dil}_\delta}(D)\right\},$$

and finally $H_{CD} = d^d_{\rho_{\mathrm{dil}_\delta}}(C, D)$ from Definition 7. Analogously, one can show $H_{DC} = d^d_{\rho_{\mathrm{dil}_\delta}}(D, C)$, and thus from Definition 8 we obtain $d_{\rho_{\mathrm{dil}_\delta}}(C, D) = \max(H_{CD}, H_{DC}) = h_\delta(\mathrm{Mod}\,(C)\,, \mathrm{Mod}\,(D))$. □

7 A Relaxation from a Tree Edit Distance

In Section 3 we have defined the tree edit distance δ^{edit} on trees with labeled nodes and edges. We have used it on $\mathcal{EL}$-description trees, but since it only requires labeled nodes and edges, it can equally be used as a metric on Int_Σ. In this section, we show how, based on δ^{edit}, a dissimilarity measure can be defined according to the framework depicted in Figure 8.

We consider the logic $\mathcal{EL}^\sqcup$, which allows for disjunction $\sqcup$ in addition to the normal constructors of $\mathcal{EL}$. The extension by disjunction will later be needed to ensure expressibility of the dilation. Note that disjunction commutes with existential restrictions, i.e. for all concepts C, D and all role names r it holds that $\exists r.(C \sqcup D) \equiv \exists r.C \sqcup \exists r.D$. In particular, this means that any complex $\mathcal{EL}^\sqcup$ concept description C can be written as a disjunction of pure $\mathcal{EL}$ concept descriptions $(C_i)_{1 \leq i \leq k}$:

$$C \equiv C_1 \sqcup C_2 \sqcup \cdots C_k. \tag{6}$$

In the later parts of this section, conjunctions over existential restrictions that share the same role name will require special attention. Therefore, we group them when transforming a concept into normal form.

Definition 11. *We say that an $\mathcal{EL}$-concept D is written in* normal form with grouping of existential restrictions *if it is of the form*

$$D = \bigsqcap_{A \in N_D} A \sqcap \bigsqcap_{r \in \mathcal{N}_R} D_r, \tag{7}$$

where $N_D \subseteq \mathcal{N}_C$ is a set of concept names and the concepts D_r are of the form

$$D_r = \bigsqcap_{E \in \mathcal{C}_{D_r}} \exists r.E, \tag{8}$$

where no subsumption relation holds between two distinct conjuncts and $\mathcal{C}_{D_r}$ is a set of complex $\mathcal{EL}$-concepts, that are themselves in normal form with grouping of existential restrictions. *The purpose of D_r terms is simply to group existential restrictions that share the same role name. For an $\mathcal{EL}^\sqcup$-concept C we say that C is in* normal form *if it is of the form (6) and each of the C_i is an $\mathcal{EL}$-concept in normal form with grouping of existential restrictions.*

Given the tree edit distance δ^{edit} we want to apply the framework from Figure 8. Notice that in order to apply Definition 10 we first need to show that the dilation $\text{dil}_{\delta^{\text{edit}}}$ is expressible in $\mathcal{EL}^\sqcup$. Furthermore, in order to apply Definition 8 it is necessary to show that $\rho_{\text{dil}_{\delta^{\text{edit}}}}$ is exhaustive (non-decreasingness and exhaustivity follow from Lemma 4). Our expressibility proof requires the following technical lemma which follows from monotonicity of the $\mathcal{EL}^\sqcup$-constructors.

Lemma 5. *Let $(\mathcal{I}, x)$ is a pointed model of an $\mathcal{EL}^\sqcup$-concept C and let $(\mathcal{I}', x)$ be a model that has been obtained from $(\mathcal{I}, x)$ by either* addLabel *or* addNode. *Then $(\mathcal{I}', x)$ is also a model of C.*

Conversely, if $(\mathcal{I}, x)$ is not a model of D and $(\mathcal{I}'', x)$ is obtained by either delLabel *or* delNode *then $(\mathcal{I}'', x)$ is not a model of D.*

We show that $\text{dil}_{\delta^{\text{edit}}}$, defined as in (4), is expressible in $\mathcal{EL}^\sqcup$, by explicitly giving the operator $\rho_{\text{dil}_{\delta^{\text{edit}}}}$. Given an $\mathcal{EL}^\sqcup$-concept description C we define an operator ρ recursively as follows. For $C = A \in \mathcal{N}_C$ and for $C = \top$ we define

$$\rho(A) = \rho(\top) = \top.$$

For $C = D_r$, where D_r is a group of existential restrictions as in (8), we need to distinguish two cases:

- if $D_r \equiv \exists r.\top$ we define $\rho(D_r) = \top$, and
- if $D_r \not\equiv \exists r.\top$ then we define

$$\rho(D_r) = \bigsqcup_{\mathcal{S} \subseteq \mathcal{C}_{D_r}} \left(\bigsqcap_{E \notin \mathcal{S}} \exists r.E \sqcap \exists r.\rho\Big(\bigsqcap_{F \in \mathcal{S}} F \Big) \right).$$

Notice that in the latter case $\top \notin \mathcal{C}_{D_r}$ since D_r is in normal form. For $C = D$ as in (7) we define

$$\rho(D) = \bigsqcup_{G \in \mathcal{C}_D} \Big(\delta(G) \sqcap \bigsqcap_{H \in \mathcal{C}_D \setminus G} H \Big),$$

where $\mathcal{C}_D = N_D \cup \{D_r \mid r \in \mathcal{N}_R\}$. Finally for $C = C_1 \sqcup C_2 \sqcup \cdots C_k$ we set

$$\rho(C) = \rho(C_1) \sqcup \rho(C_2) \sqcup \cdots \rho(C_k).$$

Theorem 2. *The operator ρ as defined above satisfies*

$$\operatorname{Mod}(\rho(C)) = \operatorname{dil}_{\delta^{edit}}(\operatorname{Mod}(C)).$$

for all concept descriptions $C \in \mathcal{EL}^{\sqcup}$. In particular, this means that $\operatorname{dil}_{\delta^{edit}}$ is expressible in $\mathcal{EL}^{\sqcup}$ and $\rho = \rho_{\operatorname{dil}_{\delta^{edit}}}$.

Proof. Our proof will follow the structure of the definition of ρ. For each case, we need to show that for all pointed models $(\mathcal{I}, x)$ it holds that

$$(\mathcal{I}, x) \in \operatorname{Mod}(\rho(C)) \iff \exists (\mathcal{I}', x) \in \operatorname{Mod}(C) : \delta^{\text{edit}}((\mathcal{I}, x), (\mathcal{I}', x)) \leq 1, \tag{9}$$

i.e. $(\mathcal{I}, x)$ is a model of $\rho(C)$ iff one edit suffices to reach a model $(\mathcal{I}', x)$ of C.

Concept Names and Top: The case $C = \top$ is trivial. We consider the case $C = A \in \mathcal{N}_C$. It is clear that for every pointed model $(\mathcal{I}, x) \in \operatorname{Int}_\Sigma = \operatorname{Mod}(\top) = \operatorname{Mod}(\rho(A))$ at most one edit operation *addLabel* suffices to obtain a model $(\mathcal{I}', x) \in \operatorname{Mod}(A)$, namely adding the label A to the root node x. The other direction is trivial.

Groups of Existential Restrictions: We consider the case where $C = D_r$ as in (8). The case where $D_r \equiv \exists r.\top$ can be treated analogously to the case of concept names. In the case where $D_r \not\equiv \exists r.\top$ the main issue is that an edit performed on an r-branch in a pointed model $(\mathcal{I}, x)$ can affect membership in several of the concepts $E \in \mathcal{C}_{D_r}$ simultaneously. In the definition of ρ this is accounted for by the disjunction over all subsets of $\mathcal{C}_{D_r}$.

We start our formal proof by showing the only-if-direction from (9). That is, we show that every tree-shaped pointed model $(\mathcal{I}, x)$ of $\rho(D_r)$ can be transformed into a model $(\mathcal{I}', x)$ of D_r using one edit. Assume that $(\mathcal{I}, x)$ is a tree-shaped pointed model of

$$\bigsqcup_{\mathcal{S} \subseteq \mathcal{C}_{D_r}} \Bigg(\bigsqcap_{E \notin \mathcal{S}} \exists r.E \sqcap \exists r.\rho\Big(\bigsqcap_{F \in \mathcal{S}} F \Big) \Bigg).$$

By the semantics of the disjunction there must be some set $\mathcal{S} \subseteq \mathcal{C}_{D_r}$ such that $(\mathcal{I}, x)$ is a model of

$$D_{\mathcal{S}} = \bigsqcap_{E \notin \mathcal{S}} \exists r.E \sqcap \exists r.\rho\Big(\bigsqcap_{F \in \mathcal{S}} F \Big).$$

In particular this means that $x \in \big(\exists r.\rho(\bigsqcap_{F \in \mathcal{S}} F)\big)^{\mathcal{I}}$. Thus there is an r-successor y of x, $(x, y) \in r^{\mathcal{I}}$, satisfying $y \in \big(\rho(\bigsqcap_{F \in \mathcal{S}} F)\big)^{\mathcal{I}}$. If we consider the subtree $(\mathcal{I}, y)$ of $(\mathcal{I}, x)$ with root y as a pointed model, then by the induction hypothesis one edit operation suffices to obtain a pointed model $(\mathcal{I}', y)$ that satisfies $y \in (\bigsqcap_{F \in \mathcal{S}} F)^{\mathcal{I}}$. In the larger model $(\mathcal{I}, x)$ we can now simply replace the subtree starting at y by $(\mathcal{I}', y)$ to obtain a model $(\mathcal{I}', x)$ (cf. Figure 1). Then in particular $(\mathcal{I}', x)$ is a model of $\exists r. \bigsqcap_{F \in \mathcal{S}} F$. By Lemma 5 this edit operation must have been either *addLabel* or *addNode*, and therefore $(\mathcal{I}', x)$ is still a model of $\bigsqcap_{D \notin \mathcal{S}} \exists r.D$, again by Lemma 5. Hence, $(\mathcal{I}', x)$ is a model of $D_{\mathcal{S}}$ and thus also of D_r, and it is only one edit removed from $(\mathcal{I}, x)$.

Let us now prove the if-direction from (9). Assume we have performed an edit operation on a pointed tree model $(\mathcal{I}', x)$ of D_r and obtained the interpretation $(\mathcal{I}, x)$.

Suppose first, that an operation *delLabel* or *delNode* has been applied within some subtree starting at an r-successor y of x. Let $\mathcal{S}_y \subseteq \mathcal{C}_{D_r}$ be the set of all concepts $E \in \mathcal{C}_{D_r}$ satisfying $y \in E^{\mathcal{I}'}$. By the induction hypothesis it must then hold that $y \in \rho\big(\bigsqcap_{F \in \mathcal{S}_y} F \big)^{\mathcal{I}}$, and therefore $(\mathcal{I}, x)$ is a model of

$$\bigsqcap_{E \notin \mathcal{S}_y} \exists r.E \sqcap \exists r.\rho\Big(\bigsqcap_{F \in \mathcal{S}_y} F \Big) \sqsubseteq \rho(D_r).$$

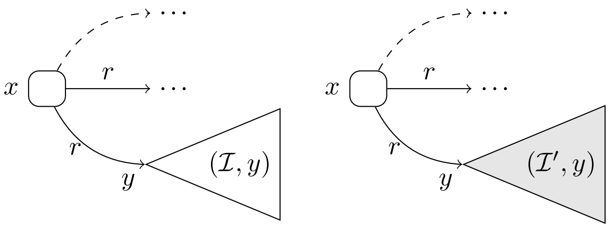

Figure 9: Changes exclusively in the subtree starting at y

For the remaining cases, notice that $D_r \sqsubseteq \rho(D_r)$ holds, since we allow $\mathcal{S}$ to be the empty set. The case where *delLabel* has been used to remove a label of the root node x is trivial, since in D_r concept names only occur behind existential quantifiers, and therefore removing a label at the root node does not affect membership in D_r. The claim then follows from $D_r \sqsubseteq \rho(D_r)$. In the case where *delNode* has been used to remove a direct r-successor y of x we know that y is unlabeled and has no successors, therefore the only concept it belongs to is $\top$. Since $\top \notin \mathcal{C}_{D_r}$ we know that $y \notin E^{\mathcal{I}'}$ for any $E \in \mathcal{C}_{D_r}$ and thus $(\mathcal{I}, x)$ will still be a model of D_r. Again, the claim follows from $D_r \sqsubseteq \rho(D_r)$.

Finally, if the operation was *addLabel* or *addNode* then the claim follows from Lemma 5 and $D_r \sqsubseteq \rho(D_r)$.

Arbitrary conjunctions: Let $D = \bigsqcap_{G \in \mathcal{C}_D} G$ where $\mathcal{C}_D = N_D \cup \{D_r \mid r \in \mathcal{N}_R\}$. We start by proving the only-if direction from (9). Let $(\mathcal{I}, x)$ be a model of $\rho(D)$. By the semantics of disjunction $(\mathcal{I}, x)$ is a model of $\delta(G) \sqcap \bigsqcap_{H \in \mathcal{C}_D \setminus G} H$ for some $G \in \mathcal{C}_D$. We only consider the case where $G = D_r$ for some $r \in \mathcal{N}_R$ is a group of existential restriction, since the case where G is a concept name is similar and slightly easier. Since $(\mathcal{I}, x)$ is a model of $\delta(G)$ we know that performing one edit operation on $(\mathcal{I}, x)$ suffices to obtain a model $(\mathcal{I}', x)$ of G by the induction hypothesis. This edit must have taken place in a subtree starting with the role r. Since we have grouped all r-restrictions into G no concept $H \in \mathcal{C}_D \setminus G$ can contain a restriction on r successors. Hence, after the edit $(\mathcal{I}', x)$ is still a model of $\bigsqcap_{H \in \mathcal{C}_D \setminus G} H$. Therefore, $(\mathcal{I}', x)$ is a model of $G \sqcap \bigsqcap_{H \in \mathcal{C}_D \setminus G} H \equiv D$.

To prove the if-direction from (9) let $(\mathcal{I}', x)$ be a pointed model of D and $(\mathcal{I}, x)$ a model that has been obtained from $(\mathcal{I}', x)$ using one edit. Like in the previous case, only delete operations are interesting. If the delete has occurred inside a subtree starting with r, then $(\mathcal{I}, x)$ will be a model of $\delta(D_r)$ and still be a model of all the other concepts in $H \in \mathcal{C}_D \setminus D_r$. Thus $(\mathcal{I}', x)$ will be a model of $\rho(D)$. The cases where a direct successor of the root node or a label of the root node have been deleted can be treated similarly to the previous case.

Disjunctions: The case where $C = C_1 \sqcup \cdots C_k$ is a simple consequence of the semantics of disjunctions. □

In order for Theorem 1 to be applicable, it only remains to show that ρ is exhaustive.

Lemma 6. *The operator ρ is exhaustive.*

Proof. From (5) it follows that $\rho^k(C) \equiv \top$ iff $\mathrm{Mod}\left(\rho^k(C)\right) = \mathrm{Int}_\Sigma$. By Theorem 2 this is equivalent to $(\mathrm{dil}_{\delta^{\text{edit}}})^k(\mathrm{Mod}\,(C)) = \mathrm{Int}_\Sigma$. Thus in order to show that ρ is exhaustive, it suffices to show that for every $\mathcal{EL}^{\sqcup}$-concept C there exists $k \in \mathbb{N}_0$ such that all $(\mathcal{I}, x) \in \mathrm{Int}_\Sigma$ satisfy $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{Mod}\,(C)) \leq k$. If C is a concept in pure $\mathcal{EL}$ then we can simply take $k = \mathrm{size}(C)$ to be the size of C, i.e. the number of labels and edges in the description tree of C. Then, using k operations *addLabel* and *addNode* we can attach the full description graph of C to the root node x in the model $(\mathcal{I}, x)$. This yields a model $(\mathcal{I}', x)$ of C, and thus $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{Mod}\,(C)) \leq k$. If C is not in pure $\mathcal{EL}$, then it can be written as a disjunction of pure $\mathcal{EL}$ concepts $C_1, \ldots, C_n$. In that case, $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{Mod}\,(C))$ is bounded by

$$\min_{1 \leq j \leq n} \mathrm{size}(C_j).$$

Hence, ρ is exhaustive. □

This finally allows us to apply Theorem 1.

Corollary 1. *The operator $\rho_{\mathrm{dil}_{\delta^{edit}}}$ is a relaxation, and the distance $d_{\rho_{\mathrm{dil}_{\delta^{edit}}}}$ is a dissimilarity measure that corresponds to the Hausdorff distance $h_{\delta^{edit}}$ in the sense of Theorem 1.*

Notice that by Lemma 1 the dissimilarity $d_{\rho_{\mathrm{dil}_{\delta_{\mathrm{edit}}}}}$ is also equivalence sound, equivalence closed, subsumption preserving, reverse subsumption preserving, and satisfies the triangle inequality.

Example 1. *For the relaxation ρ_{depth} we observed that in certain cases it contradicts the intuition that a greater number of common features should yield smaller dissimilarities. If we apply $d_{\rho_{\mathrm{dil}_{\delta_{edit}}}}$ to this example we obtain the dissimilarities*

$$\begin{aligned} d_{\rho_{\mathrm{dil}_{\delta_{edit}}}}(\mathsf{F}, \exists \mathsf{hasChild}.\top) &= 1, \\ d_{\rho_{\mathrm{dil}_{\delta_{edit}}}}(\mathsf{HoJ}, \exists \mathsf{hasChild}.\top) &= 4, \ \textit{and} \\ d_{\rho_{\mathrm{dil}_{\delta_{edit}}}}(\mathsf{HoJ}, \mathsf{F}) &= 3. \end{aligned}$$

as we would expect it by looking at the commonalities between the concepts.

8 Conclusion

In this work, we have presented several dissimilarity measures. Our approach for the Description Logic $\mathcal{EL}$ looks at the unique description tree of a concept's normal form. Then any tree metric can be used to define a dissimilarity. In the general case, such a dissimilarity is not subsumption preserving and reverse subsumption preserving. A special case is the dissimilarity based on the tree edit distance $d^{\mathrm{tree}}_{\delta_{\mathrm{edit}}}$, which satisfies these properties. The problem with this approach is that it is very specific to $\mathcal{EL}$ and cannot easily be adapted to other logics.

For this reason, we have presented a second approach, a framework based on relaxation operators. In order to define a dissimilarity for a new logic, it suffices to find a unary operator on concepts that is non-decreasing, extensive and exhaustive. The relaxation dissimilarity obtained in this way satisfies all the listed properties except monotonicity and structural dependence. We have then instantiated this framework by defining a morphological dilation on the concept space and then expressing it as a relaxation at the concept level. An overview of the properties of the similarity measures that we defined compared to some earlier works can be found in Table 2.

Table 2: Properties of some (dis-)similarity measures.

Measure	Equivalence Sound	Monotone	Equivalence Closed	Subs. Preserving	Rev. Subs. Preserving	Structurally Dependent	Triangle Inequality
[14]	✓	–	✓	✓	✓	✓	–
[9]	✓	✓	–	✓	✓	–	–
$d^{\mathrm{tree}}_{\delta}$	✓	–	✓	–	–	–	✓
$d^{\mathrm{tree}}_{\delta_{\mathrm{edit}}}$	✓	–	✓	✓	✓	–	✓
relaxation dissimilarity	✓	–	✓	✓	✓	–	✓

With respect to the intuition that greater numbers of common features should entail smaller dissimilarity we were only able to provide anecdotal evidence. This is partly due to problems inherent to the two criteria *subsumption preserving* and *monotonicity*, which try to formalize that intuition. Subsumption preservingness is in many cases unnecessarily restrictive, while monotonicity has very limited significance in logics with disjunction.

The similarity measures that we have presented here are defined for concepts without a background terminology. We briefly discuss how they can be adapted to the presence of background ontologies. If the background ontology is an acyclic TBox, then concepts can be unfolded with respect to the TBox. In that case, it is possible to simply compute the dissimilarity with respect to the unfolded concepts. In principle, it is possible to generalize relaxations with respect to general TBoxes, by simply replacing the subsumption relation in their definition by subsumption with respect to a TBox. How to instantiate relaxations with respect to TBoxes is left for future work.

Acknowledgments.

This work was initiated during the stay of Felix Distel at Telecom ParisTech in summer 2012, supported by a grant from this institution. Felix Distel has also been supported by the Collaborative Research Center 912 ‘Highly Adaptive Energy-Efficient Computing’. The authors would like to thank Anni-Yasmin Turhan for fruitful discussions.

References

[1] Franz Baader. Description Logic terminology. In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 485–495. Cambridge University Press, 2003.

[2] Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 364–369. Morgan-Kaufmann, 2005.

[3] Franz Baader, Ralf Küsters, and Ralf Molitor. Computing least common subsumers in description logics with existential restrictions. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 96–101. Morgan-Kaufmann, 1999.

[4] Philip Bille. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1):217–239, 2005.

[5] Garrett Birkhoff. *Lattice theory*, volume 25 of *Colloquium publications*. American Mathematical Society, Providence, Rhode Island, 3rd edition, 1993.

[6] Isabelle Bloch and Jérôme Lang. Towards mathematical morpho-logics. In *Technologies for Constructing Intelligent Systems 2*, pages 367–380. Springer, 2002.

[7] Hans Hermann Bock and Edwin Diday. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer, Berlin, 2000.

[8] Alexander Borgida, Thomas J Walsh, and Haym Hirsh. Towards measuring similarity in description logics. In *Proc. of the 2005 Int. Workshop on Description Logics (DL)*, 2005.

[9] Claudia d’Amato, Steffen Staab, and Nicola Fanizzi. On the influence of description logics ontologies on conceptual similarity. In *Knowledge Engineering: Practice and Patterns*, pages 48–63. Springer, 2008.

[10] Luc De Raedt and Jan Ramon. Deriving distance metrics from generality relations. *Pattern Recognition Letters*, 30(3):187–191, 2009.

[11] Alan Hutchinson. Metrics on terms and clauses. In *Proc. of the European Conf. on Machine Learning (ECML)*, pages 138–145. Springer, 1997.

[12] Krzysztof Janowicz and Marc Wilkes. SIM-DLA: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. In *The Semantic Web: Research and Applications*, pages 353–367. Springer, 2009.

[13] Karsten Lehmann. A framework for semantic invariant similarity measures for $\mathcal{ELH}$ concept descriptions. Master’s thesis, TU Dresden, Germany, 2012.

[14] Karsten Lehmann and Anni-Yasmin Turhan. A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts. In *Proc. of the 13th European Conf. on Logics in Artificial Intelligence (ECAI)*, LNAI, pages 307–319. Springer Verlag, 2012.

[15] Carsten Lutz, Frank Wolter, and Michael Zakharyaschev. A tableau algorithm for reasoning about concepts and similarity. In *Automated reasoning with analytic tableaux and related methods*, volume 2796 of *LNCS*, pages 134–149. Springer, 2003.

[16] Yue Ma and Pascal Hitzler. Distance-based measures of inconsistency and incoherency for description logics. In *Proc. of the 2010 Int. Workshop on Description Logics (DL)*, pages 467–477, 2010.

[17] Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, and Carsten Lutz. OWL 2 web ontology language: Profiles. *W3C recommendation*, 27:61, 2009.

[18] Shan-Hwei Nienhuys-Cheng. Distances and limits on herbrand interpretations. In *Inductive Logic Programming*, pages 250–260. Springer, 1998.

[19] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[20] Guilin Qi and Jianfeng Du. Model-based revision operators for terminologies in description logics. In *Proc. of the 21st Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 891–897. AAAI Press, 2009.

[21] Jan Ramon and Maurice Bruynooghe. A framework for defining distances between first-order logic objects. In *Inductive Logic Programming*, pages 271–280. Springer, 1998.

[22] Christian Ronse. Why mathematical morphology needs complete lattices. *Signal processing*, 21(2):129–154, 1990.

[23] Klaus Schild. A correspondence theory for terminological logics: Preliminary report. In *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 466–471. Morgan-Kaufmann, 1991.

[24] Jean Serra. *Image analysis and mathematical morphology*. London.: Academic Press., 1982.

[25] Boontawee Suntisrivaraporn. A similarity measure for the description logic $\mathcal{EL}$ with unfoldable terminologies. In *Proc. of the 5th Int. Conf. on Intelligent Networking and Collaborative Systems (INCoS)*, pages 408–413. IEEE, 2013.

[26] Kuo-Chung Tai. The tree-to-tree correction problem. *Journal of the ACM (JACM)*, 26(3):422–433, 1979.

[27] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Dépôt légal : 2014 – 1er trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut Mines-Télécom - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.fr

Département TSI