ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-249

Concept Dissimilarity Based on Tree Edit Distances and Morphological Dilations

Felix Distel [1] and **Jamal Atif** [2] and **Isabelle Bloch** [3]

Abstract. A number of similarity measures for comparing description logic concepts have been proposed. Criteria have been developed to evaluate a measure's fitness for an application. These criteria include on the one hand those that ensure compatibility with the semantics, such as equivalence soundness, and on the other hand the properties of a metric, such as the triangle inequality. In this work we present two classes of dissimilarity measures that are at the same time equivalence sound and satisfy the triangle inequality: a simple dissimilarity measure, based on description trees for the lightweight description logic $\mathcal{EL}$; and an instantiation of a general framework, presented in our previous work, using dilation operators from mathematical morphology, and which exploits the link between Hausdorff distance and dilations using balls of the ground distance as structuring elements.

1 INTRODUCTION

The need to quantify similarities or differences between logical objects arises in areas as diverse 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 description logics (DL) one is most often interested in measuring similarity between concepts, while measures for individuals or ontologies also exist. When similarity measures were first investigated within the DL community, researchers mainly focused on adaptations of existing measures from other fields (cf. [8] for a survey). Most of these are tailored to the specific needs of a particular field, such as biomedicine [20], or geospatial reasoning [16]. The quality of these measures was mainly evaluated in an empirical way, showing that they perform well in a given setting, but providing little transferable insight.

As the number of similarity measures grew, the need to formalize criteria for selecting an appropriate measure for a given application arose. Works such as [9] and [17] list on the one hand the properties of a metric, in particular the triangle inequality, as well as properties ensuring compatibility with the semantics of the logic, such as soundness with respect to equivalence and subsumption. Which of these properties are relevant depends on the application. For example, the triangle inequality is deemed irrelevant in [16], but it is crucial in other applications such as metric-based conceptual clustering and distance-based optimization methods [14].

Unfortunately, none of the aforementioned measures satisfy both the properties of a metric and the properties ensuring compatibility with the semantics. Outside of description logics, however, several works have proposed distance measures between logical objects that do satisfy the triangle inequality. Works such as [19, 21] exploit the fact that it 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 Birkhoff has presented this idea as the classical distance in graded lattices [5]. It is used to define a distance between first order literals in [15], which is generalized to a distance between clauses using the Hausdorff metric. The idea can also be extended to cases where there is no unique minimal generalization [10].

This work is based on previous work [12] which introduces a framework for dissimilarity measures based on concept relaxation operators. The resulting dissimilarity measures are at the same time sound with respect to equivalence and subsumption and satisfy the triangle inequality. Unfortunately, simple concept relaxation operators often yield relatively coarse dissimilarity measures.

In this work we start from a tree edit distance in order to obtain more fine grained measures. In the lightweight DL $\mathcal{EL}$ tree edit distances can be applied directly to the concept descriptions as described in Section 3. For more expressive logics this is no longer the case. If the logic has the tree model property, then the tree edit distance can still be used on the model level. In Sections 4 and 5 we show how a relaxation operator can be obtained from a tree edit distance on models. Our approach is inspired by the Hausdorff distance and its links with morphological dilation. 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. Unlike previous approaches our measure is computed in an iterative way. It is defined for two concept descriptions in the absence of a terminology (TBox). In the conclusion we will briefly discuss how, under certain conditions, it can be extended to settings with a background terminology.

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

[1] Institute of Theoretical Computer Science, Faculty of Computer Science, TU Dresden, Germany, email: felix@tcs.inf.tu-dresden.de

[2] Université Paris Sud, LRI, TAO, Orsay, France, email: jamal.atif@lri.fr

[3] Institut Mines Telecom, Telecom ParisTech, CNRS LTCI, Paris, France, email: isabelle.bloch@telecom-paristech.fr

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*. 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 concepts to subsets of the domain $\Delta_{\mathcal{I}}$ and role names to binary relations on the domain.

A concept description C is said to *subsume* a concept description D (denoted by $C \sqsubseteq D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds for every interpretation $\mathcal{I}$. C and D are *equivalent* (denoted by $C \equiv D$) if both $C \sqsubseteq D$ and $D \sqsubseteq C$ hold.

A logic $\mathcal{L}$ is said to have the *tree model property* if $C \sqsubseteq D$ holds for every pair of concept descriptions $C, D \in \mathsf{C}(\mathcal{L})$ that satisfies $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ for all tree-shaped interpretations $\mathcal{I}$.

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, and the set of all pointed tree-shaped interpretations by TInt_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 $\mathrm{Mod}\,(C)$ and the set of all pointed tree-shaped models of C by $\mathrm{TMod}\,(C)$.

2.2 Similarity and Dissimilarity on Concepts

In [9] for the first time qualitative criteria were developed, using the work in [7] as a starting point. 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 is* positive*, i.e. $d(C, D) \geq 0$,* reflexive*, i.e. $d(C, C) = 0$ and* symmetric*, i.e. $d(C, D) = d(D, C)$ for all $C, D \in \mathsf{C}(\mathcal{L})$.*

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 authors in [9] and later [17] introduced additional criteria.[4]

Definition 2 *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called*

- equivalence sound[5] *if $D \equiv E \implies d(C, D) = d(C, E)$,*
- 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 $C \sqsubseteq D \sqsubseteq E \implies d(D, E) \leq d(C, E)$, and*
- *we say that d fulfills the* triangle inequality *if $d(C, E) \leq d(C, D) + d(D, E)$*

for all $C, D, E \in \mathsf{C}(\mathcal{L})$.

A common intuition is that concepts with more common features should be less dissimilar than concepts with fewer 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. Therefore, their dissimilarity should be smaller than the dissimilarity between F and Female. Two attempts to formalize this intuition are *monotonicity* and *structural dependence*. Their definitions can be found in [9] and [17], respectively.

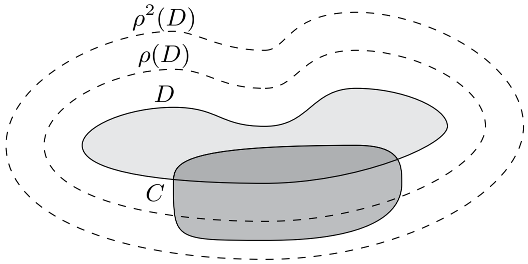

Figure 1. D needs to be relaxed twice before it subsumes C, i.e. $d^d_\rho(C, D) = 2$.

A dissimilarity d is a *metric* if it satisfies the triangle inequality and is additionally strict, i.e. $d(x, y) = 0$ implies $x = y$. The bottleneck preventing most dissimilarity measures from being metrics is the triangle inequality.

2.3 Dissimilarity Based on Relaxations

In [12] we provided a general framework for dissimilarities that 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 (Figure 1).

Definition 3 (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. $\exists k \in \mathbb{N}\colon \top \sqsubseteq \rho^k(C)$,*

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

A trivial relaxation is the operator $\rho_\top$ that maps every concept to $\top$.

Definition 4 (Relaxation Dissimilarity) *Let ρ be a relaxation on $\mathsf{C}(\mathcal{L})$. For two concepts C and D the* relaxation dissimilarity *$d_\rho(C, D)$ is defined as $d_\rho(C, D) = \max\{d^d_\rho(C, D), d^d_\rho(D, C)\}$, where $d^d_\rho(C, D) = \min\{k \in \mathbb{N} \mid C \sqsubseteq \rho^k(D)\}$.*

Using relaxations to define dissimilarity measures can be seen as the dual idea to the one in [13], where similarity measures are used to define relaxations.

Theorem 1 ([12]) *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.*

Theorem 1 shows that relaxation operators yield dissimilarity measures with good theoretical properties. However, as discussed in [12] many simple relaxation operators yield rather coarse dissimilarities. In this work, we instantiate the framework using a relaxation based on a tree edit distance on models on the one hand, and on morphological operators (namely dilations) on the other hand.

[4] The properties of *soundness* and *dissimilarity incompatibility* were also mentioned in [9], however these were never formally defined.

[5] Notice that in [17] equivalence soundness is referred to as *equivalence invariance*.

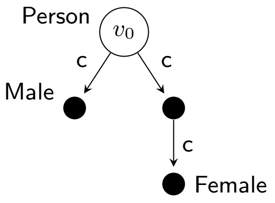

Figure 2. $\mathcal{EL}$-Description Tree for (3).

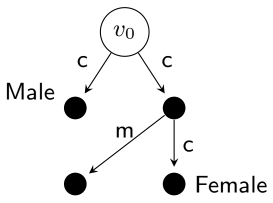

Figure 3. Figure 2 after two edits: Removing the label at v_0 and adding an m-edge.

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 EL [18].

$\mathcal{EL}$ concept descriptions can appropriately be represented as labeled trees, often called *$\mathcal{EL}$ description trees* [3]. An $\mathcal{EL}$ description tree is a tree whose nodes are labeled with sets of concept names and whose edges are labeled with role names. An $\mathcal{EL}$ concept description

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

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 the description tree of

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

is depicted in Figure 2. 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].

We assume that we are given a metric δ on the space of all $\mathcal{EL}$-concept descriptions and try to derive a dissimilarity measure between $\mathcal{EL}$-concepts. Simply defining dissimilarity between two concepts as the distance between their description trees 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 from (2) 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. In general, d_δ^{tree} lacks monotonicity, structural dependence and (reverse) subsumption preservation.

3.2 Tree Edit Distances

Arguably the most widely used among the various approaches for defining distances between labeled trees are *tree edit distances*, first introduced in [25]. They have been successfully applied in fields as diverse as computer vision, natural language processing, and computational biology (cf. [4] for a survey).

To define a tree edit distance, one first 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. The choice of operations depends on the 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
- an unlabeled leaf node 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. Consider Figures 2 and 3 for an example.

Using a characterization of subsumption between $\mathcal{EL}$-concepts from [3] it is straightforward to prove that the dissimilarity measure $d_{\delta^{\text{edit}}}^{\text{tree}}$ is subsumption preserving and reverse subsumption preserving, in addition to the properties shared by all dissimilarity measures obtained from Definition 5.

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 4 we will show how, by working with models instead of concepts, we can still make use of tree distances to define dissimilarity measures.

4 RELAXATION OPERATORS FROM DILATIONS

4.1 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*. Given a metric space (M, δ) and a real number $\lambda \in \mathbb{R}$ the dilation $\mathrm{dil}_{\delta,\lambda}$ and the erosion $\mathrm{ero}_{\delta,\lambda}$ by a ball of δ of radius λ are defined as operators on the power set of M, which we denote by $\mathfrak{P}(M)$:

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

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.

There is an intuitive connection between dilations and relaxations, e.g. both are extensive and monotone. In Section 4.2 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 6 (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 $k \in \{0, 1, \ldots, \delta(x, y)\}$ there exists $z \in M$ such that $\delta(x, z) = k$ and $\delta(z, y) = \delta(x, y) - k$.*

Lemma 1 *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 $\mathrm{dil}_{\delta,\lambda}(X) = (\mathrm{dil}_{\delta,1})^\lambda(X)$.*

A connection between dilations and the classical Hausdorff distance is mentioned in [24]. Remember that for a metric space (M, δ) the *Hausdorff distance* h_δ is a metric between non-empty compact subsets $X, Y \subseteq M$. It is defined as

$$h_\delta(X, Y) = \max\left\{\sup_{x\in X} \delta(x, Y), \sup_{y\in Y} \delta(y, X)\right\}, \tag{5}$$

where $\delta(x, Y) = \inf\{\delta(x, y) \mid y \in Y\}$. The Hausdorff distance can then be expressed in terms of dilations as follows.

Lemma 2 ([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 $h_{d,\delta}(X, Y) = \inf\{\lambda \mid X \subseteq \mathrm{dil}_{\delta,\lambda}(Y)\}$.

4.2 From Model Space to Concept Space

Just like Hausdorff distance generates a metric defined on sets of points from a metric points, we want to lift a metric δ from pointed models to concept descriptions. Let δ be a metric on the space of pointed tree models TInt_Σ. In a logic $\mathcal{L}$ with the tree model property, no two concept descriptions C and D can have the same sets of pointed tree models $\mathrm{TMod}\,(C)$ and $\mathrm{TMod}\,(D)$, unless they are equivalent. In fact it even holds that

$$C \sqsubseteq D \iff \mathrm{TMod}\,(C) \subseteq \mathrm{TMod}\,(D)\,. \tag{6}$$

In particular, the description logic $\mathcal{ALC}$ and all of its fragments have the tree model property [23]. Therefore, one could naively use the Hausdorff distance $h_\delta(\mathrm{TMod}\,(C)\,, \mathrm{TMod}\,(D))$ between sets of pointed tree models to define dissimilarity between the two concepts C and D. In practice, the Hausdorff distance cannot be computed directly from (5), since it is impossible to list all models in $\mathrm{TMod}\,(C)$ and $\mathrm{TMod}\,(D)$. Instead we shall define a dissimilarity from a relaxation, which we obtain from a dilation. This is only possible if the dilation is expressible in the following sense.

Definition 7 (Expressibility) *Let $\omega\colon \mathfrak{P}(\mathrm{TInt}_\Sigma) \to \mathfrak{P}(\mathrm{TInt}_\Sigma)$ 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*

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

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 5. 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 3 *Let $\mathcal{L}$ be a logic that has the tree-model property. Let* dil *be a dilation on TInt_Σ. If* dil *is expressible in $\mathcal{L}$ then ρ_{dil} is non-decreasing and extensive.*

For discrete metrics, we now have all the necessary definitions to obtain a dissimilarity measure on concepts, according to Figure 4. 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 (6).

Theorem 2 *Let δ be a discrete metric on TInt_Σ and let $C, D \in \mathsf{C}(\mathcal{L})$ be concept descriptions such that $\mathrm{TMod}\,(C)$, $\mathrm{TMod}\,(D)$ are compact. Let dil_δ, $\rho_{\mathrm{dil}_\delta}$ and $d_{\rho_{\mathrm{dil}_\delta}}$ be defined as in Equation (4), Definition 7 and Definition 4, 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*

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

Proof: Lemma 2 states that $h_\delta(\mathrm{TMod}\,(C)\,, \mathrm{TMod}\,(D)) = \max(H_{CD}, H_{DC})$ where

$$H_{CD} = \inf\{\lambda \mid \mathrm{TMod}\,(C) \subseteq \mathrm{dil}_{\delta,\lambda}(\mathrm{TMod}\,(D))\}\,.$$

The betweenness property and expressibility of dil_δ entail $\mathrm{dil}_{\delta,\lambda}(\mathrm{TMod}\,(D)) = (\mathrm{dil}_\delta)^\lambda(\mathrm{TMod}\,(D)) = \mathrm{TMod}\,\left(\rho^\lambda_{\mathrm{dil}_\delta}(D)\right)$. Together with (6) 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 4. Analogously, one can show $H_{DC} = d^d_{\rho_{\mathrm{dil}_\delta}}(D, C)$, and thus from Definition 4 we obtain $d_{\rho_{\mathrm{dil}_\delta}}(C, D) = \max(H_{CD}, H_{DC}) = h_\delta(\mathrm{TMod}\,(C)\,, \mathrm{TMod}\,(D))$. □

Applicability of Haussdorf distances is restricted to concepts with compact sets of models. In the following we present a relaxations based approach that can compute dissimilarities between two concepts regardless of compactness of their sets of models. Theorem 2 guarantees that in those cases where they are compact, our approach yields the same result as a Hausdorff distance, i.e. our approach is truly a generalization of a Hausdorff distance.

5 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

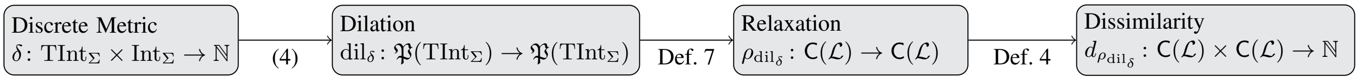
Discrete Metric
$\delta\colon \mathrm{TInt}_\Sigma \times \mathrm{Int}_\Sigma \to \mathbb{N}$
(4)
Dilation
$\mathrm{dil}_\delta\colon \mathfrak{P}(\mathrm{TInt}_\Sigma) \to \mathfrak{P}(\mathrm{TInt}_\Sigma)$
Def. 7
Relaxation
$\rho_{\mathrm{dil}_\delta}\colon \mathsf{C}(\mathcal{L}) \to \mathsf{C}(\mathcal{L})$
Def. 4
Dissimilarity
$d_{\rho_{\mathrm{dil}_\delta}}\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{N}$

Figure 4. From discrete metrics on TInt_Σ to dissimilarity measures.

as a metric on TInt_Σ. In this section, we show how, based on δ^{edit}, a dissimilarity measure can be defined according to the framework depicted in Figure 4.

We consider the logic $\mathcal{ELU}$, 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 $\mathcal{ELU}$ concept description C can be written as a disjunction of pure $\mathcal{EL}$ concept descriptions $(C_i)_{1\le i\le k}$:

$$C \equiv C_1 \sqcup C_2 \sqcup \cdots C_k. \tag{7}$$

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

$$D = \bigsqcap_{A\in N_D} A \sqcap \bigsqcap_{r\in\mathcal{N}_R} D_r, \tag{8}$$

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

$$D_r = \bigsqcap_{E\in\mathcal{C}_{D_r}} \exists r.E, \tag{9}$$

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 $\mathcal{ELU}$-concept C we say that C is in* normal form *if it is of the form (7) 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 4. Notice that in order to apply Definition 7 we first need to show that the dilation $\mathrm{dil}_{\delta^{\text{edit}}}$ is expressible in $\mathcal{ELU}$. Furthermore, in order to apply Definition 4 it is necessary to show that $\rho_{\mathrm{dil}_{\delta^{\text{edit}}}}$ is exhaustive (non-decreasingness and exhaustivity follow from Lemma 3). Our expressibility proof requires the following technical lemma. We omit the proof, which is a straightforward consequence of monotonicity of the $\mathcal{ELU}$-constructors.

Lemma 4 *Let $(\mathcal{I}, x)$ be a pointed tree model of an $\mathcal{ELU}$-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 $\mathrm{dil}_{\delta^{\text{edit}}}$, defined as in (4), is expressible in $\mathcal{ELU}$, by explicitly giving the operator $\rho_{\mathrm{dil}_{\delta^{\text{edit}}}}$. Given an $\mathcal{ELU}$-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 (9), 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 (8) 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 3 *The operator ρ as defined above satisfies*

$$\mathrm{TMod}\,(\rho(C)) = \mathrm{dil}_{\delta^{edit}}(\mathrm{TMod}\,(C)).$$

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

To prove Theorem 3, one needs to show that for all pointed models $(\mathcal{I}, x)$ it holds that

$$\begin{aligned}&(\mathcal{I}, x) \in \mathrm{TMod}\,(\rho(C)) \iff \\ &\qquad \exists (\mathcal{I}', x) \in \mathrm{TMod}\,(C) : \delta^{\text{edit}}((\mathcal{I}, x), (\mathcal{I}', x)) \le 1,\end{aligned} \tag{10}$$

i.e. $(\mathcal{I}, x)$ is a model of $\rho(C)$ iff one edit suffices to reach a model $(\mathcal{I}', x)$ of C. This can be shown using a straightforward but tedious induction on the structure of C that follows the definition of ρ. A complete proof can be found in [11].

In order for Theorem 2 to be applicable, it only remains to show that ρ is exhaustive.

Lemma 5 *The operator ρ is exhaustive.*

Proof: From (6) it follows that $\rho^k(C) \equiv \top$ iff $\mathrm{TMod}\left(\rho^k(C)\right) = \mathrm{TInt}_\Sigma$. By Theorem 3 this is equivalent to $(\mathrm{dil}_{\delta^{\text{edit}}})^k(\mathrm{TMod}\,(C)) = \mathrm{TInt}_\Sigma$. Thus in order to show that ρ is exhaustive, it suffices to show that for every $\mathcal{ELU}$-concept C there exists $k \in \mathbb{N}$ such that all $(\mathcal{I}, x) \in \mathrm{TInt}_\Sigma$ satisfy $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{TMod}\,(C)) \le k$. 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 $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{TMod}\,(C)) \le k$. 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, $\delta^{\text{edit}}((\mathcal{I}, x), \mathrm{TMod}\,(C))$ is bounded by $\min_{1\le j\le n} \mathrm{size}(C_j)$. Hence, ρ is exhaustive. □

This finally allows us to apply Theorem 2.

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 2.*

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 the concepts from (1)*

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

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

6 CONCLUSION

In this work, we have presented two classes of dissimilarity measures. For the Description Logic $\mathcal{EL}$ we looked 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 $d^{\mathrm{tree}}_{\delta_{\mathrm{edit}}}$ satisfying these properties. This approach is specific to $\mathcal{EL}$ and cannot easily be adapted to other logics.

For this reason, we have presented a second approach based on a framework presented in [12]. We have 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 1.

Table 1. Properties of some (dis-)similarity measures.

Measure	Equivalence Sound	Monotone	Equivalence Closed	Subsumption Preserving	Rev. Subs. Preserving	Structurally Dependent	Triangle Inequality
[17]	✓	–	✓	✓	✓	✓	–
[9]	✓	✓	–	✓	✓	–	–
$d^{\mathrm{tree}}_{\delta}$	✓	–	✓	–	–	–	✓
$d^{\mathrm{tree}}_{\delta_{\mathrm{edit}}}$	✓	–	✓	✓	✓	–	✓
d_{ρ}	✓	–	✓	✓	✓	–	✓

The similarity measures that we have presented here are defined for concepts without a background terminology. In the presence of an acyclic TBox, 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 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. Future work should also include a practical evaluation of the measures we defined.

ACKNOWLEDGEMENTS

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] F. Baader, 'Description Logic terminology', in *The Description Logic Handbook: Theory, Implementation, and Applications*, eds., Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, 485–495, Cambridge University Press, (2003).

[2] F. Baader, S. Brandt, and C. Lutz, 'Pushing the $\mathcal{EL}$ envelope', in *IJCAI'05*, pp. 364–369. Morgan-Kaufmann, (2005).

[3] F. Baader, R. Küsters, and R. Molitor, 'Computing least common subsumers in description logics with existential restrictions', in *IJCAI'99*, pp. 96–101. Morgan-Kaufmann, (1999).

[4] P. Bille, 'A survey on tree edit distance and related problems', *Theoretical Computer Science*, **337**(1), 217–239, (2005).

[5] G. Birkhoff, *Lattice theory*, volume 25 of *Colloquium publications*, AMS, Providence, Rhode Island, 3rd edn., 1993.

[6] I. Bloch and J. Lang, 'Towards mathematical morpho-logics', in *Technologies for Constructing Intelligent Systems 2*, 367–380, Springer, (2002).

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

[8] A. Borgida, T.J. Walsh, and H. Hirsh, 'Towards measuring similarity in description logics.', in *DL'05*, (2005).

[9] C. d'Amato, S. Staab, and N. Fanizzi, 'On the influence of description logics ontologies on conceptual similarity', in *Knowledge Engineering: Practice and Patterns*, 48–63, Springer, (2008).

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

[11] F. Distel, J. Atif, and I. Bloch, 'Concept dissimilarity based on tree edit distances and morphological dilations', Technical report, Telecom ParisTech, (2014). `http://perso.telecom-paristech.fr/~bloch/papers/TECHREP/publication-278.pdf`.

[12] F. Distel, J. Atif, and I. Bloch, 'Concept dissimilarity with triangle inequality', in *KR'14*, Vienna, Austria, (2014). AAAI.

[13] A. Ecke, R. Peñaloza, and A.-Y. Turhan, 'Answering instance queries relaxed by concept similarity', in *KR'14*, Vienna, Austria, (2014). AAAI.

[14] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*, MIT Press, 1996.

[15] A. Hutchinson, 'Metrics on terms and clauses', in *ECML'97*, pp. 138–145. Springer, (1997).

[16] K. Janowicz and M. 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*, 353–367, Springer, (2009).

[17] K. Lehmann and A.-Y. Turhan, 'A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts', in *JELIA'12*, LNAI, pp. 307–319. Springer Verlag, (2012).

[18] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz, 'OWL 2 web ontology language: Profiles', *W3C recommendation*, **27**, 61, (2009).

[19] S.-H. Nienhuys-Cheng, 'Distances and limits on herbrand interpretations', in *Inductive Logic Programming*, 250–260, Springer, (1998).

[20] C. Pesquita, D. Faria, A.O. Falcao, P. Lord, and F.M. Couto, 'Semantic similarity in biomedical ontologies', *PLoS computational biology*, **5**(7), (2009).

[21] J. Ramon and M. Bruynooghe, 'A framework for defining distances between first-order logic objects', in *Inductive Logic Programming*, 271–280, Springer, (1998).

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

[23] K. Schild, 'A correspondence theory for terminological logics: Preliminary report', in *IJCAI'91*, pp. 466–471. Morgan-Kaufmann, (1991).

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

[25] K.-C. Tai, 'The tree-to-tree correction problem', *Journal of the ACM (JACM)*, **26**(3), 422–433, (1979).