Reasoning about concepts and similarity

Carsten Lutz, Frank Wolter and Michael Zakharyaschev

Fakultät Informatik, TU Dresden, Germany
email: lutz@tcs.inf.tu-dresden.de
Department of Computer Science, University of Liverpool, U.K.
email: frank@csc.liv.ac.uk
Department of Computer Science, King's College London, U.K.
email: mz@dcs.kcl.ac.uk

1 Introduction

Suppose you want to use description logics (DLs) to develop an ontology of description logics. Such an ontology should contain information about standard DLs such as $\mathcal{FL}_0$, $\mathcal{ALC}$, and $\mathcal{ALCQO}$, description logics extended with temporal, epistemic, and dynamic operators, the computational complexity of DLs, known decision procedures, applications, publications, relevant workshops and conferences, and so on.

A considerable part of such an ontology can straightforwardly be formulated in a sufficiently expressive description logic, say $\mathcal{ALCQO}$ [1]. However, there also exist a number of important concepts that are rather vague and cannot be precisely defined in terms of simpler concepts. Examples of such concepts are 'DL,' 'tableau-algorithm,' 'practical decision procedure,' 'extended description logic,' and others. The vagueness of these concepts is witnessed by the fact that there is often no satisfactory 'yes/no' answer to the question whether a certain formalism is a description logic, whether a certain decision procedure is a tableau algorithm, and so forth. We argue that it is more adequate and informative to define such vague concepts by referring to their prototypical instances. For example, we could define tableau algorithms as algorithms being '*very similar to the standard tableau-algorithm for $\mathcal{ALC}$-concepts relative to general TBoxes, and not similar to structural subsumption algorithms.*' Such an approach to defining vague concepts looks much more promising than squeezing them into crisp, classical DL-style definitions.

This observation suggests that it can be useful to integrate into standard description logics some means for representing and reasoning about *similarities between objects.* Interpretations $\mathcal{I}$ of such an extended description logic should be equipped with *similarity measures* $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_n$ saying that, according to $\boldsymbol{\delta}_j$, objects $x, y \in \Delta^{\mathcal{I}}$ are similar to degree $\boldsymbol{\delta}_j(x, y) \in \mathbb{Q}^+$. For example, $\boldsymbol{\delta}_1$ could measure the similarity between decision procedures for DLs with respect to certain structural features. According to $\boldsymbol{\delta}_1$, resolution-based algorithms would not be very similar to tableau-based algorithms, while the standard tableau-algorithm for $\mathcal{ALC}$-concepts relative to general TBoxes would be rather similar to the tableau-algorithm for $\mathcal{ALC}$-concepts with the universal modality. Another similarity measure $\boldsymbol{\delta}_2$ on the set of implemented decision procedures could compare their performance on certain benchmarks.

As a first step towards a language for representing knowledge using similarity measures we should therefore allow expressions of the form

$$\boldsymbol{\delta}_i(k,\ell) < a, \quad \boldsymbol{\delta}_i(k,\ell) \leq a, \quad \boldsymbol{\delta}_i(k,\ell) > a, \quad \boldsymbol{\delta}_i(k,\ell) \geq a,$$

where k, ℓ are object names (representing, say, description logics or decision procedures) and $a \in \mathbb{Q}^+$. For example, the expression $\boldsymbol{\delta}_i(k,\ell) < a$ means that, according to $\boldsymbol{\delta}_i$, the 'distance' between object k and object ℓ is $< a$, with smaller distances representing a higher degree of similarity.

When designing the DL ontology mentioned above, we obviously cannot assume that the set of *all* possible decision procedures is known to us, and that we know how similar any two of them are. Thus, ontologies using similarities will neither fix domains nor contain complete information about all objects—a property that is shared by ontologies formulated in standard DLs. To deal with this incompleteness, it is desirable to have at our disposal not only the above similarity expressions operating on named objects, but also concept constructors that allow concept defininitions based on similarity measures.

A first idea is to define, for each $\boldsymbol{\delta}_i$ and each $q \in \mathbb{Q}^+$, a role name $\mathsf{similar}_i^{\leq q}$ which is interpreted as follows:

$$(x,y) \in \mathsf{similar}_i^{\leq q} \quad \text{iff} \quad \boldsymbol{\delta}_i(x,y) \leq q.$$

Then the $\mathcal{ALCQO}$-definition

$$\mathsf{tableau\_algorithm} = \mathsf{algorithm} \sqcap \exists \mathsf{similar}_1^{\leq 0.5}.(\mathsf{ta}_1 \sqcup \cdots \sqcup \mathsf{ta}_7), \tag{1}$$

says that tableau algorithms are algorithms which are similar to degree ≤ 0.5 to at least one of the prototypical tableau algorithms $\mathsf{ta}_1, \ldots, \mathsf{ta}_7$ (here the ta_i are nominals). Given a new procedure ta, we can integrate it into the knowledge base by using assertions like

$$\begin{array}{lrcl} & \boldsymbol{\delta}_i(\mathsf{ta},\mathsf{ta}_1) & < & 0.5 \\ \text{or} & \mathsf{ta} & \sqsubseteq & \mathsf{tableau\_algorithm} \\ \text{or} & \mathsf{ta} & \sqsubseteq & \neg\exists \mathsf{similar}_1^{\leq 5}.\mathsf{tableau\_algorithm}. \end{array}$$

The last assertion says that the distance from ta to the 'closest' tableau algorithm is more than 5.

It should be clear that the roles $\mathsf{similar}_i^{\leq q}$ cannot be interpreted by arbitrary relations: in order to describe *natural* similarity measures, some special properties have to be taken into account. We stipulate that similarity measures $\boldsymbol{\delta}_i$ should satisfy the standard axioms of metric spaces, that is

$$\begin{aligned} &\boldsymbol{\delta}_i(x,x) = 0, \\ &\boldsymbol{\delta}_i(x,y) = \boldsymbol{\delta}_i(y,x), \\ &\boldsymbol{\delta}_i(x,z) \leq \boldsymbol{\delta}_i(x,y) + \boldsymbol{\delta}_i(y,z). \end{aligned}$$

Actually, these metric axioms are a standard choice for dealing with similarity measures [3, 5], and we believe that it is adequate for similarity-based description logics as well. Moreover, even if we do not assume all axioms of metric spaces, the 'positive' results presented in this paper still hold true. This applies, for example, to the similarity measures considered in [15, 9, 4] which satisfy the first two axioms of metric spaces only.

Returning back to our initial idea of representing similarity measures in terms of roles, we now face the problem that the axioms of metric spaces, in particular the triangle inequality, *cannot be expressed in standard DLs*. Indeed, it will turn out that it is a good idea to keep roles talking about similarity measures strictly separated from standard roles: as shown in Section 4, there may be strong interactions between standard DL constructors (e.g., qualified number restrictions) and the properties of similarity measures that can lead to undecidability. For this reason, we treat separately the constructors speaking about similarity measures and those required to model conceptual knowledge. In other words, we propose to form the *fusion* [10, 6, 2] of standard DLs with a suitable formalism for reasoning about similarity measures.

The main message of this paper is that such a combined expressive description logic can indeed be devised and, moreover, supported by a tableau-based decision algorithm that is rather similar to tableau algorithms for standard DLs. More precisely, we merge the expressive power of

- the standard description logic $\mathcal{ALCQO}$—i.e., the basic DL $\mathcal{ALC}$ extended with qualified number restrictions, nominals, and general TBoxes [1]—

with

- the logic $\mathcal{MS}$ devised in [17] for reasoning about metric spaces.

Definition (1) can serve as an example of a typical TBox assertion of the resulting 'hybrid' logic that we call *sim-*$\mathcal{ALCQO}$. As another example, consider the following *sim-*$\mathcal{ALCQO}$ ABox assertion, where ha denotes a certain Hilbert-style algorithm:

$$\mathsf{ha} : \mathsf{Algorithm} \sqcap \neg\exists \mathsf{feature}.\mathsf{Termination} \sqcap \forall \mathsf{similar}^{\leq 0.5}.(\exists \mathsf{comprises}.\mathsf{Modus\_ponens}).$$

It says that ha does not necessarily terminate and that all ≤ 0.5 similar algorithms use a kind of *modus ponens* as one of their inference rules.

It may seem more natural to specify similarity in terms of a finite set of *symbolic* similarity measures such as 'close' or 'far' rather than in terms of rational numbers as above. In our approach, however, the user is free to choose either option: one may fix a rational number for each symbolic similarity measure, say, 1 for 'close' and 10 for 'far,' and then work with the symbolic names.

In our opinion, *sim-*$\mathcal{ALCQO}$ provides just the right compromise between expressive power and computational cost:

(1) In *sim-*$\mathcal{ALCQO}$, we can mix constructors of $\mathcal{ALCQO}$ and $\mathcal{MS}$ in order to define concepts based on similarity measures as illustrated above. Moreover, our tableau algorithm shows that reasoning in *sim-*$\mathcal{ALCQO}$ is still decidable. It is of interest to contrast this with the fact that a tighter coupling of $\mathcal{ALCQO}$ and $\mathcal{MS}$ leads to undecidability: as we also show, the extension of $\mathcal{MS}$ with qualified number restrictions such as 'there exists at most 1 point x with property P within distance ≤ 1' results in an undecidable logic. Therefore, the fusion of the two formalisms seems to be a good starting point for investigating the interaction between concepts and similarity measures.

(2) Although there exists a number of general results regarding the transfer of decidability from the components of a fusion to the fusion itself [10, 6, 16, 2, 14], these results do not apply to logics with nominals such as $\mathcal{ALCQO}$. In fact, no transfer result is available from which we could derive the decidability of *sim*-$\mathcal{ALCQO}$ using the decidability of both $\mathcal{ALCQO}$ and $\mathcal{MS}$. Despite the fact that general transfer results are not applicable, our algorithm has an important advantage over algorithms obtained from general transfer theorems: structurally, it is very similar to the tableau algorithms for $\mathcal{SHIQ}$ and $\mathcal{SHOQ}$ proposed in [7, 8]. Since these algorithms have turned out to be implementable in efficient reasoning systems, we hope that our algorithm also has this attractive property.

The reader can find a tableau-based system for *sim*-$\mathcal{ALCQO}$ in [12]. The full version of this paper is available at `http://www.csc.liv.ac.uk/~frank/`.

2 The logic *sim*-$\mathcal{ALCQO}$

In this section, we introduce the combined logic *sim*-$\mathcal{ALCQO}$. To simplify notation, we confine ourselves to the language with a single similarity measure. The reader should not have big problems in extending the language and the decision procedure to cope with a finite set of such measures. The *alphabet* of *sim*-$\mathcal{ALCQO}$ consists of the following elements:

- a countably infinite list of *concept names* $A_1, A_2, \ldots$;
- a countably infinite list of *object names* $\ell_1, \ell_2, \ldots$;
- binary *distance* ($\boldsymbol{\delta}$), *equality* ($\doteq$) and *membership* (:) *predicates*;
- the *Boolean operators* $\sqcap$, $\sqcup$, $\neg$;
- two *distance quantifiers* $\mathsf{E}^{<a}$, $\mathsf{E}^{\leq a}$ and their duals $\mathsf{A}^{<a}$, $\mathsf{A}^{\leq a}$, for every positive rational number a (i.e., $a \in \mathbb{Q}^+$);
- *role names* $R_1, R_2, \ldots$;
- *qualified number restrictions* $(\leq nR.C)$ and $(\geq nR.C)$, for every natural n, every role name R, and every concept C.

Using this alphabet, sim-$\mathcal{ALCQO}$*-concepts* are defined by the formation rule:

$$C ::= A_i \mid \ell_i \mid \neg C \mid C_1 \sqcap C_2 \mid C_1 \sqcup C_2 \mid \mathsf{E}^{<a}C \mid \mathsf{E}^{\leq a}C \mid \mathsf{A}^{<a}C \\ \mid \mathsf{A}^{\leq a}C \mid (\leq nR_i.C) \mid (\geq nR_i.C).$$

As usual, we write $\exists R.C$ for $(\geq 1R.C)$ and $\forall R.C$ for $(\leq 0R.\neg C)$. Object names occurring in concepts are known as *nominals*. We define *sim*-$\mathcal{ALCQO}$*-assertions* as expressions of the following forms:

- $\ell : C$, where ℓ is an object name and C a concept;
- $C_1 \doteq C_2$, where C_1 and C_2 are concepts;

- $\boldsymbol{\delta}(k,\ell) < a$, $\boldsymbol{\delta}(k,\ell) \leq a$, $\boldsymbol{\delta}(k,\ell) > a$, $\boldsymbol{\delta}(k,\ell) \geq a$, where k, ℓ are object names and $a \in \mathbb{Q}^+$.

A *sim-$\mathcal{ALCQO}$ knowledge base* is a finite set of *sim-$\mathcal{ALCQO}$*-assertions.

The semantics of *sim-$\mathcal{ALCQO}$*-concepts is a blend of the semantics for the logic of metric spaces [17] and the usual set-theoretic semantics of description logics. A *concept-distance model* (a *CD-model*, for short) is a structure of the form

$$\mathfrak{B} = \left\langle W, d, A_1^{\mathfrak{B}}, A_2^{\mathfrak{B}}, \dots, R_1^{\mathfrak{B}}, R_2^{\mathfrak{B}}, \dots, \ell_1^{\mathfrak{B}}, \ell_2^{\mathfrak{B}} \dots \right\rangle,$$

where $\langle W, d\rangle$ is a *metric space* with a *distance function* d satisfying, for all $x, y, z \in W$, the axioms

$$d(x,y) = 0 \ \text{ iff } \ x = y, \tag{2}$$

$$d(x,z) \leq d(x,y) + d(y,z), \tag{3}$$

$$d(x,y) = d(y,x), \tag{4}$$

the $A_i^{\mathfrak{B}}$ are subsets of W, the $R_i^{\mathfrak{B}}$ are binary relations on W, and the $\ell_i^{\mathfrak{B}}$ are singleton subsets of W such that $i \neq j$ implies $\ell_i^{\mathfrak{B}} \neq \ell_j^{\mathfrak{B}}$ (unique name assumption). The *extension* $C^{\mathfrak{B}}$ of a *sim-$\mathcal{ALCQO}$*-concept C is computed inductively:

$$\begin{aligned}
(\neg C)^{\mathfrak{B}} &= W - C^{\mathfrak{B}},\\
(C_1 \sqcap C_2)^{\mathfrak{B}} &= C_1^{\mathfrak{B}} \cap C_2^{\mathfrak{B}},\\
(C_1 \sqcup C_2)^{\mathfrak{B}} &= C_1^{\mathfrak{B}} \cup C_2^{\mathfrak{B}},\\
(\mathsf{E}^{\leq a} C)^{\mathfrak{B}} &= \{x \in W \mid \exists y \in W \left(d(x,y) \leq a \ \wedge y \in C^{\mathfrak{B}}\right)\},\\
(\mathsf{E}^{< a} C)^{\mathfrak{B}} &= \{x \in W \mid \exists y \in W \left(d(x,y) < a \ \wedge y \in C^{\mathfrak{B}}\right)\},\\
(\mathsf{A}^{\leq a} C)^{\mathfrak{B}} &= \{x \in W \mid \forall y \in W \left(d(x,y) \leq a \ \rightarrow y \in C^{\mathfrak{B}}\right)\},\\
(\mathsf{A}^{< a} C)^{\mathfrak{B}} &= \{x \in W \mid \forall y \in W \left(d(x,y) < a \ \rightarrow y \in C^{\mathfrak{B}}\right)\},\\
(\leq nR.C)^{\mathfrak{B}} &= \{x \in W \mid \left|\{y \in W \mid (x,y) \in R^{\mathfrak{B}} \wedge y \in C^{\mathfrak{B}}\}\right| \leq n\},\\
(\geq nR.C)^{\mathfrak{B}} &= \{x \in W \mid \left|\{y \in W \mid (x,y) \in R^{\mathfrak{B}} \wedge y \in C^{\mathfrak{B}}\}\right| \geq n\}.
\end{aligned}$$

The *truth-relation* $\models$ between CD-models $\mathfrak{B}$ and *sim-$\mathcal{ALCQO}$*-assertions φ is defined in the natural way by taking:

$$\begin{aligned}
&\mathfrak{B} \models \ell : C \quad \text{iff} \quad \ell^{\mathfrak{B}} \subseteq C^{\mathfrak{B}},\\
&\mathfrak{B} \models C_1 \doteq C_2 \quad \text{iff} \quad C_1^{\mathfrak{B}} = C_2^{\mathfrak{B}},\\
&\mathfrak{B} \models \boldsymbol{\delta}(k,\ell) \leq a \quad \text{iff} \quad d(k^{\mathfrak{B}}, \ell^{\mathfrak{B}}) \leq a,\\
&\mathfrak{B} \models \boldsymbol{\delta}(k,\ell) < a \quad \text{iff} \quad d(k^{\mathfrak{B}}, \ell^{\mathfrak{B}}) < a, \ \text{and similar for } \geq \text{ and } >.
\end{aligned}$$

Finally, a *sim-$\mathcal{ALCQO}$* knowledge base Σ is called *satisfiable* if there exists a CD-model $\mathfrak{B}$ such that $\mathfrak{B} \models \varphi$ for all $\varphi \in \Sigma$. In this case we write $\mathfrak{B} \models \Sigma$.

Let us make some notes on several syntactic and semantic particularities of our logic:

(1) In contrast to the initial idea from Section 1, we do not explicitly introduce a role for the similarity measure—in fact this approach was only taken for didactic purposes in the introduction. Instead, the concept constructors $\mathsf{E}^{\leq a}$, $\mathsf{E}^{<a}$, $\mathsf{A}^{\leq a}$, and $\mathsf{A}^{<a}$ refer directly to distances.

(2) At first sight, it may seem strange to have both strict and non-strict versions of the E and A constructors available. However, this allows us to define the concept $\mathsf{E}^{\leq a}C \sqcap \neg\mathsf{E}^{<a}C$ which states that the most similar object from C is located *precisely* at distance a.

(3) Observe that *sim-$\mathcal{ALCQO}$* knowledge bases subsume both general TBoxes and ABoxes. In particular, the usual ABox assertions of the form $(\ell_1, \ell_2) : R$, where ℓ_1 and ℓ_2 are object names and R a role name, can be viewed as abbreviations for $\ell_1 : \exists R.\ell_2$.

(4) In the semantics, we make the *unique name assumption* (*UNA*), i.e., different object names denote distinct domain elements. The sole purpose of this assumption is to comply with the definition of *sim-$\mathcal{ALCQO}$* given in [12], where a tableau algorithm is devised and the UNA allows a clearer presentation of this algorithm. It is, however, easily seen that the UNA has no influence on decidability, and that the tableau algorithm in [12] can be extended to deal with *sim-$\mathcal{ALCQO}$* without UNA.

(5) Quite often, similarity measures are required to take values from the interval $[0, 1]$, with 0 denoting the lowest degree of similarity and 1 denoting the highest one. There are two main differences to the similarity measures used in *sim-$\mathcal{ALCQO}$*: first, in our approach *small* distance denotes high degree of similarity while *large* distance denotes low degree of similarity. Second, in our logic there is no absolute, lowest degree of similarity. Thus, objects may be 'arbitrarily non-similar.'

(6) We use rational numbers as distances in our language only for simplicity. One could take instead any countable subset of the real numbers on which arithmetical operations can be performed effectively.

3 Tableau algorithm

In [12], we present a tableau algorithm for deciding satisfiability of *sim-$\mathcal{ALCQO}$* knowledge bases. Essentially, this algorithm is a combination of the tableau algorithm for the DL $\mathcal{SHOQ}$ (of which $\mathcal{ALCQO}$ is a fragment) presented in [8], and the tableau algorithm for the logic $\mathcal{MS}$ of metric spaces presented in [17]. Since space limitations make a detailed presentation of the algorithm impossible in this paper, we will only highlight (on a rather abstract level) some of its prominent features.

The tableau algorithm for *sim-$\mathcal{ALCQO}$* attempts to construct a Kripke model for the given input knowledge base. To do this, the algorithm starts with an initial 'completion forest' (having one root for each nominal occurring in the input knowledge base), and then exhaustively applies completion rules which are essentially the ones known from the $\mathcal{SHOQ}$ and $\mathcal{MS}$ algorithms. Let us comment on the $\mathcal{MS}$ part of the algorithm. Apart from the distances that occur explicitly in the input knowledge base Σ, the tableau algorithm may generate new distances during its run. All generated distances are from the closure $M[\Sigma]$, which is the smallest set satisfying the following conditions:

- if $\mathsf{E}^{\leq a}$, $\mathsf{E}^{<a}$, $\mathsf{A}^{\leq a}$, or $\mathsf{A}^{<a}$ occurs in Σ, then $a \in M[\Sigma]$;

- if $a, b \in M[\Sigma]$ and $a + b$ is strictly smaller than the largest distance occurring in Σ, then $a + b \in M[\Sigma]$;
- if $a, b \in M[\Sigma]$ and $a - b > 0$, then $a - b \in M[\Sigma]$.

Each distance a from $M[\Sigma]$ may also occur in the form a^- (e.g., 3.5 becomes 3.5^-) which denotes the distance that is smaller than a by some infinitesimal constant ϵ. A typical tableau rule for dealing with the similarity constructors looks as follows:

$\mathsf{R}_{\mathsf{A}^<}$ If $\mathsf{A}^{<a}C \in S(x)$ and d is the similarity between x and y, then:
if $d = a^-$, then set $S(y) := \{C\} \cup S(y)$;
if $d = b < a$, then set $S(y) := \{\mathsf{A}^{<a-b}C\} \cup S(y)$;
if $d = b^-$ with $b < a$, then set $S(y) := \{\mathsf{A}^{\leq a-b}C\} \cup S(y)$.

Observe that this rule may introduce 'new' distances through subtraction. All these distances are from the closure $M[\Sigma]$.

Since the algorithm does not terminate 'naturally,' we need to use a blocking mechanism. More precisely, we use standard equality blocking with one notable exception: due to the introduction of the additional distances from $M[\Sigma]$, the number of different concepts that may appear in a run of the tableau algorithm is *exponential* in the size of the input knowledge base. This implies that, with a naïve use of equality blocking, the algorithm would generate paths of double exponential length before blocking occurs (as opposed to the exponential length in standard tableau algorithms such as the one for $\mathcal{SHOQ}$). The key observation for curing this defect is that many of the concepts appearing in runs of the tableau algorithm are not independent from one another: 'most' concepts are of the form $\mathsf{A}^{\leq a}D$ and $\mathsf{A}^{<a}D$, and, e.g., $\mathsf{A}^{\leq a}D$ implies $\mathsf{A}^{\leq b}D$ if $b \leq a$. By taking into account such interactions, one can manage to devise a blocking mechanism that guarantees blocking on every path of exponential length.

The soundness proof of the tableau algorithm makes use of an alternative, relational semantics for *sim-*$\mathcal{ALCQO}$. This semantics comprises, for each $a \in \mathbb{Q}^+$, additional binary relations R_a and S_a such that, intuitively, we have uR_av if the distance between u and v is at most a, and uS_av if the distance between u and v is less than a. Formally, a *Kripke model for* Σ is a structure of the form

$$\mathfrak{M} = \left\langle W, A_1^{\mathfrak{M}}, \ldots, R_1^{\mathfrak{M}}, \ldots, (R_a)_{a\in\mathbb{Q}^+}, (S_a)_{a\in\mathbb{Q}^+}, \ell_1^{\mathfrak{M}}, \ldots \right\rangle$$

satisfying, for all $u, v, w \in W$ and all $a, b \in \mathbb{Q}^+$, the following conditions:

(S1$_R$) if uR_av and $a \leq b$, then uR_bv,
(S2$_R$) uR_av iff vR_au,
(S3$_R$) uR_au,
(S4$_R$) if uR_av, vR_bw, then $uR_{a+b}w$,
(S1$_S$) if uS_av and $a \leq b$, then uS_bv,
(S2$_S$) uS_av iff vS_au,
(S3$_S$) uS_au,
(S4$_S$) if uS_av, vS_bw, then $uS_{a+b}w$,
(C1) if uS_av then uR_av,
(C2) if uR_av and $a < b$, then uS_bv,
(C3) if uR_av, vS_bw, then $uS_{a+b}w$,

(C4) if $uS_a v$, $vR_b w$, then $uS_{a+b} w$.

The *value* $C^{\mathfrak{M}}$ of a concept C in $\mathfrak{M}$ and the *truth-relation* $\mathfrak{M} \models C_1 \doteq C_2$ are defined in almost the same way as for CD-models: we only replace $\mathfrak{B}$ with $\mathfrak{M}$ and define the clauses for the distance quantifiers as follows:

$$\begin{aligned}
(\mathsf{E}^{\leq a} C)^{\mathfrak{M}} &= \{x \in W \mid \exists y \in W \big(xR_a y \land y \in C^{\mathfrak{M}}\big)\},\\
(\mathsf{E}^{< a} C)^{\mathfrak{M}} &= \{x \in W \mid \exists y \in W \big(xS_a y \land y \in C^{\mathfrak{M}}\big)\},\\
(\mathsf{A}^{\leq a} C)^{\mathfrak{M}} &= \{x \in W \mid \forall y \in W \big(xR_a y \to y \in C^{\mathfrak{M}}\big)\},\\
(\mathsf{A}^{< a} C)^{\mathfrak{M}} &= \{x \in W \mid \forall y \in W \big(xS_a y \to y \in C^{\mathfrak{M}}\big)\}.
\end{aligned}$$

It can be proved that the alternative Kripke semantics is ‘equivalent’ to the original one:

Theorem 1 *The knowledge base Σ is satisfiable in a CD-model iff it is satisfiable in a Kripke model for Σ.*

This finishes our discussion of the tableau algorithm. We obtain the following main result:

Theorem 2 *The satisfiability problem for* sim-$\mathcal{ALCQO}$ *knowledge bases is decidable.*

As for the complexity of reasoning, we cannot (yet) provide tight bounds. The component logic $\mathcal{MS}$ is ExpTime-complete, even if the distances are encoded in binary [17]. The complexity of the component logic $\mathcal{ALCQO}$ has, to the best of our knowledge, never been formally investigated. There are, however, good reasons to conjecture that it is also ExpTime-complete (in the presence of general TBoxes), even if numbers inside number restrictions are coded in binary. For the combined logic *sim-$\mathcal{ALCQO}$*, we thus clearly inherit ExpTime-hardness from the component logics. The upper bound obtained from our tableau algorithm is a 2-NExpTime one. We conjecture that this upper bound can be improved. We also should like to note that the complexity of the standard tableau algorithms for $\mathcal{MS}$ and $\mathcal{SHOQ}$ is also 2-NExpTime. Surprisingly, despite this high complexity such algorithms can be well-suited for implementation as witnessed by the FaCT and RACER systems.

4 Undecidability

It is natural idea to try a closer integration of the constructors of $\mathcal{MS}$ and $\mathcal{SHOQ}$ by providing concept constructors that resemble qualified number restrictions, but talk about similarity measures. Unfortunately, even very simple variants of this logic are undecidable: denote by *sim$_f$* the language with the following concept formation rule:

$$C ::= A_i \mid \ell_i \mid \neg C \mid C_1 \sqcap C_2 \mid \mathsf{E}^{\leq a} C \mid (\leq^1_a .C),$$

where $(\leq^1_a .C)$ is interpreted in concept-distance models $\mathfrak{B}$ as follows

$$(\leq^1_a .C)^{\mathfrak{B}} = \{x \in W \mid \big|\{y \mid d(x,y) \leq a,\ y \in C^{\mathfrak{B}}\}\big| \leq 1\}.$$

Even this simple form of number restriction on similarity measures suffices to make reasoning undecidable.

Theorem 3 *The satisfiability problem for sim_f knowledge bases in concept-distance models is undecidable.*

Proof (sketch): We can simulate the undecidable $\mathbb{N} \times \mathbb{N}$-tiling problem in almost the same way as in the undecidability proof of [11] for the language $\mathcal{MS}_1$ with the operators $\mathsf{A}^{\leq a}$, $\mathsf{A}^{\geq 0}_{\leq a}$ and their duals: just replace everywhere in the proof of Theorem 3.1 the concept $\mathsf{A}^{\geq 0}_{\leq 80}\neg\chi_{i,j}$ with the concept $(\leq^{1}_{80} \cdot \chi_{i,j})$. □

5 Conclusion

We regard *sim*-$\mathcal{ALCQO}$ only as a first step towards DLs that allow definitions of concepts based on similarity measures. Although we believe that the expressive power provided by *sim*-$\mathcal{ALCQO}$ is quite natural and useful, an in-depth investigation of the expressive means that are relevant for defining vague concepts still has to be performed. Some possible extensions of *sim*-$\mathcal{ALCQO}$ are the following:

(1) New constructors $\mathsf{E}^{<a}R.C$ and $\mathsf{A}^{<a}R.C$, where the former expresses that there exists an R-successor satisfying C at distance smaller than a, and the latter is its dual. Such constructors would, e.g., allow us to say that a person is very similar to his father: $\mathsf{E}^{<0.5}\mathsf{parent.Male}$. The tableau algorithm in [12] should be extendable to this case without any problems.

(2) New constructors $\mathsf{E}^{>a}C$ and $\mathsf{E}^{\geq a}C$ (and their duals) with the obvious semantics. Although these constructors do not seem to be as natural as the variants based on $<$ and $\leq$, they could, e.g., be used to express that a prototypical tableau algorithm pta is very close to *all* other tableau algorithms: $\mathsf{pta} : \mathsf{A}^{>0.5}\neg\mathsf{Tableau\_algorithm}$. While [11] proves the decidability of the metric logic with the operators $\mathsf{E}^{\leq a}C$ and $\mathsf{E}^{>a}C$ (and their duals), nothing is currently known about the extension of $\mathcal{MS}$ with all four possible constructors.

We should add that *sim*-$\mathcal{ALCQO}$ is not the first logic concerned with similarity measures. For example, modal logics for reasoning about similarity have been proposed in [15, 9, 4]. However, there are three main differences to our proposal: first, in the existing approaches similarity measures are usually only required to be reflexive and symmetric; the full set of metric axioms is not treated. Second, the existing approaches do not allow references to concrete distances—one can only say that two objects are similar or not similar. Third, to the best of our knowledge our approach is the first one that uses an integration with description logics by admitting 'free' roles that are not regarded as similarity measures, and by taking into account DL-style constructors such as qualified number restrictions.

Acknowledgements:

The work of the second author was supported by Deutsche Forschungsgemeinschaft (DFG) grant Wo583/3-3. The work of the third author was partially supported by U.K. EPSRC grants no. GR/R45369/01 and GR/R42474/01.

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] F. Baader, C. Lutz, H. Sturm, and F. Wolter. Fusions of description logics and abstract description systems. *J. of Artificial Intelligence Research*, 16:1–58, 2002.

[3] P. Clote and R. Backofen. *Computational Molecular Biology.* John Wiley & Sons, 2000

[4] S. Demri and B. Konikowska. Relative similarity logics are decidable: Reduction to FO^2 with equality. *Lecture Notes in Computer Science*, 1489:279–293, 1998.

[5] M. Dunham. *Data Mining. Introductory and Advanced Topics.* Pearson Education, 2003.

[6] K. Fine and G. Schurz. Transfer theorems for stratified modal logics. In *Logic and Reality, Essays in Pure and Applied Logic. In memory of Arthur Prior*, pages 169–213. Oxford University Press, 1996.

[7] I. Horrocks, U. Sattler, and S. Tobies. Practical reasoning for expressive description logics. In *Proceedings of the 6th International Conference on Logic for Programming and Automated Reasoning (LPAR'99)*, number 1705 in Lecture Notes in Artificial Intelligence, pages 161–180. Springer-Verlag, 1999.

[8] I. Horrocks and U. Sattler. Ontology reasoning in the $\mathcal{SHOQ}(D)$ description logic. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 199–204. Morgan Kaufmann, 2001.

[9] B. Konikowska. A logic for reasoning about relative similarity. *Studia Logica*, 58(1):173–203, 1997.

[10] M. Kracht and F. Wolter. Properties of independently axiomatizable bimodal logics. *J. Symbolic Logic*, 56:1469–1485, 1991.

[11] O. Kutz, H. Sturm, N.-Y. Suzuki, F. Wolter, and M. Zakharyaschev. Logics of metric spaces. *ACM Transactions on Computational Logic*, 2003. In print.

[12] C. Lutz, F. Wolter, and M. Zakhatyaschev. A tableau algorithm for reasoning about concepts and similarity. Proceedings of Tableaux 2003. To appear.

[13] M. Schmidt-Schauß and G. Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48:1–26, 1991.

[14] E. Spaan. *Complexity of Modal Logics.* PhD thesis, Department of Mathematics and Computer Science, University of Amsterdam, 1993.

[15] D. Vakarelov. A modal logic for similarity relations in Pawlak knowledge representation systems. *Fundamenta Informaticae*, 15:61–79, 1991.

[16] F. Wolter. Fusions of modal logics revisited. In M. Kracht, M. De Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 1, pages 361–379. CSLI, Stanford, 1997.

[17] F. Wolter and M. Zakharyaschev. Reasoning about distances. Proceedings of IJCAI 2003. To appear.