A Similarity Measure for the \textit{ALN} Description Logic

Nicola Fanizzi, Claudia d’Amato

\textit{Dipartimento di Informatica \& Università degli Studi di Bari}
Campus Universitario, Via Orabona 4, 70125 Bari, Italy

CILC 2006 \& Bari
Contents

1 Introduction & Motivation
   - Motivations
   - Objectives

2 The Reference Representation Language
   - Knowledge Base & Subsumption
   - Normal Form

3 A Similarity Measure for ALN
   - Definition
   - Similarity Measure: example
   - Measure Involving Individuals
   - Discussion

4 Conclusions and Further Developments
   - Conclusions
   - Future Work
Motivations

- Ontological knowledge
  - Result of a complex process of knowledge acquisition
  - Plays a key role for interoperability in the Semantic Web perspective
  - Is expressed by standard ontology mark-up languages which are supported by well-founded semantics of Description Logics (DLs)
- Need of services able to build knowledge bases automatically or semi-automatically
  - This can be done by the use of inductive inference services
Objectives

- Induction of structural knowledge is known is ML (concept formation).
  - This is generally applied on zero-order representations.
- *our Goal* → to make clusters of concepts or individuals asserted by mean ontological knowledge
- *Problem* → to define a similarity/dissimilarity measure applicable to ontology languages
Why $\mathcal{ALN}$ Logic

- Knowledge representation by mean Description Logic ($\mathcal{ALN}$)
- Description Logic is the counterpart framework of OWL language
  - standard de facto for the knowledge representation in the Semantic Web
The Representation Language

- **Primitive concepts** $N_C = \{C, D, \ldots\}$: subsets of a domain
- **Primitive roles** $N_R = \{R, S, \ldots\}$: binary relations on the domain
- **Interpretation** $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ where
  $\Delta^\mathcal{I}$: *domain* of the interpretation and $\cdot^\mathcal{I}$: *interpretation function*:

<table>
<thead>
<tr>
<th>Name</th>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>top concept</td>
<td>$\top$</td>
<td>$\Delta^\mathcal{I}$</td>
</tr>
<tr>
<td>bottom concept</td>
<td>$\bot$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>primitive concept</td>
<td>$A$</td>
<td>$A^\mathcal{I} \subseteq \Delta^\mathcal{I}$</td>
</tr>
<tr>
<td>primitive negation</td>
<td>$\neg A$</td>
<td>$\Delta^\mathcal{I} \setminus A^\mathcal{I}$</td>
</tr>
<tr>
<td>concept conjunction</td>
<td>$C_1 \sqcap C_2$</td>
<td>$C_1^\mathcal{I} \cap C_2^\mathcal{I}$</td>
</tr>
<tr>
<td>universal restriction</td>
<td>$\forall R. C$</td>
<td>${x \in \Delta^\mathcal{I}</td>
</tr>
<tr>
<td>at-most restriction</td>
<td>$\leq n.R$</td>
<td>${ x \in \Delta^\mathcal{I}</td>
</tr>
<tr>
<td>at-least restriction</td>
<td>$\geq n.R$</td>
<td>${ x \in \Delta^\mathcal{I}</td>
</tr>
</tbody>
</table>

N. Fanizzi, C. d'Amato

A Similarity Measure
Knowledge Base & Subsumption

\[ \mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle \]

- **T-box** \( \mathcal{T} \) is a set of definitions \( C \equiv D \), meaning \( C^\mathcal{I} = D^\mathcal{I} \), where \( C \) is the concept name and \( D \) is a description.
- **A-box** \( \mathcal{A} \) contains extensional assertions on concepts and roles.
  - e.g. \( C(a) \) and \( R(a, b) \), meaning, resp., that \( a^\mathcal{I} \in C^\mathcal{I} \) and \( (a^\mathcal{I}, b^\mathcal{I}) \in R^\mathcal{I} \).

**Subsumption**

Given two concept descriptions \( C \) and \( D \), \( C \) subsumes \( D \), denoted by \( C \sqsupseteq D \), iff for every interpretation \( \mathcal{I} \), it holds that \( C^\mathcal{I} \supseteq D^\mathcal{I} \).
Examples

Instances of concept definitions:
Single \equiv \text{Person} \sqcap \leq 0.\text{isMarriedTo}
Polygamist \equiv \text{Person} \sqcap \forall \text{isMarriedTo}.\text{Person} \sqcap \geq 2.\text{isMarriedTo}
Bigamist \equiv \text{Person} \sqcap \forall \text{isMarriedTo}.\text{Person} \sqcap = 2.\text{isMarriedTo}
MalePolygamist \equiv \text{Male} \sqcap \text{Person} \sqcap \forall \text{isMarriedTo}.\text{Person} \sqcap \geq 2.\text{isMarriedTo}

The following are instances of simple assertions:
Male(Bob), Person(Mary), Single(Jhon), isMarriedTo(Bob, Mary)

It is easy to see that the following relationship holds:
Poligamist \sqsubseteq MalePolygamist.
Other Inference Services

*instance checking* decide whether an individual is an instance of a concept

*retrieval* find all individuals instance of a concept

*realization problem* finding the concepts which an individual belongs to, especially the most specific one, if any:

**most specific concept**

Given an A-Box $\mathcal{A}$ and an individual $a$, the *most specific concept* of $a$ w.r.t. $\mathcal{A}$ is the concept $C$, denoted $\text{MSC}_\mathcal{A}(a)$, such that $\mathcal{A} \models C(a)$ and $C \sqsubseteq D$, $\forall D$ such that $\mathcal{A} \models D(a)$. 
**Normal Form**

A description $C$ is in **ALN normal form** iff $C \equiv \bot$ or $C \equiv \top$ or if

$$C = \bigcap_{P \in \text{prim}(C)} P \cap \bigcap_{R \in N_R} (\forall R. C_R \cap \geq n.R \cap \leq m.R)$$

where:

- $C_R = \text{val}_R(C)$,
- $n = \min_R(C)$ and $m = \max_R(C)$

**primal(C)** set of all (negated) atoms occurring at $C$’s top-level

**val$_R$(C)** conjunction $C_1 \cap \cdots \cap C_n$ in the value restriction on $R$, if any (o.w. $\text{val}_R(C) = \top$);

- $\min_R(C) = \max\{n \in N \mid C \sqsubseteq (\geq n.R)\}$ (always finite number);
- $\max_R(C) = \min\{n \in N \mid C \sqsubseteq (\leq n.R)\}$ (if unlimited $\max_R(C) = \infty$)

For any $R$, every sub-description in $\text{val}_R(C)$ is in normal form.
A Similarity Measure for $\mathcal{ALN}$: Definition

$\mathcal{L} = \mathcal{ALN}/\equiv$ the set of all concepts in $\mathcal{ALN}$ normal form

$I$ canonical interpretation of $\mathcal{A}$ A-Box $s : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$ defined

$\forall C, D \in \mathcal{L}$:

$s(C, D) := \lambda[s_P(\text{prim}(C), \text{prim}(D)) +$

$+ \frac{1}{|N_R|} \sum_{R \in N_R} s(\text{val}_R(C), \text{val}_R(D)) + \frac{1}{|N_R|} \cdot$

$\cdot \sum_{R \in N_R} s_N((\text{min}_R(C), \text{max}_R(C)), (\text{min}_R(D), \text{max}_R(D)))$]

where $\lambda \in ]0, 1]$ (let $\lambda = 1/3$),
A Similarity Measure for $\mathcal{ALN}$: Definition / II

\[
s_P(\text{prim}(C), \text{prim}(D)) := \frac{\bigcap_{P_C \in \text{prim}(C)} P_C \cap \bigcap_{Q_D \in \text{prim}(D)} Q_D}{\bigcap_{P_C \in \text{prim}(C)} P_C \cup \bigcap_{Q_D \in \text{prim}(D)} Q_D}
\]

\[
s_N((m_C, M_C), (m_D, M_D)) := \frac{\min(M_C, M_D) - \max(m_C, m_D) + 1}{\max(M_C, M_D) - \min(m_C, m_D) + 1}
\]

\[
s_N((m_C, M_C), (m_D, M_D)) := 0 \text{ if } \min(M_C, M_D) > \max(m_C, m_D)
\]
Similarity Measure: example...

Let $A$ be the considered ABox

\[
\text{Person(Meg), } \neg \text{Male(Meg), hasChild(Meg,Bob), hasChild(Meg,Pat), }
\text{Person(Bob), Male(Bob), hasChild(Bob,Ann), }
\text{Person(Pat), Male(Pat), hasChild(Pat,Gwen), }
\text{Person(Gwen), } \neg \text{Male(Gwen), }
\text{Person(Ann), } \neg \text{Male(Ann), hasChild(Ann,Sue), marriedTo(Ann,Tom), }
\text{Person(Sue), } \neg \text{Male(Sue), }
\text{Person(Tom), Male(Tom)}
\]

and let $C$ and $D$ be two descriptions in ALN normal form:

\[
C \equiv \text{Person} \sqcap \forall \text{marriedTo.Person} \sqcap \leq 1.\text{hasChild}
\]
\[
D \equiv \text{Male} \sqcap \forall \text{marriedTo.(Person} \sqcap \neg \text{Male}) \sqcap \leq 2.\text{hasChild}
\]
In order to compute $s(C, D)$ let us consider:

- Let be $\lambda := \frac{1}{3}$
- $N_R = \{\text{hasChild, marriedTo}\} \rightarrow |N_R| = 2$

\[
s(C, D) := \frac{1}{3} \left[ s_P(\text{prim}(C), \text{prim}(D)) + \frac{1}{2} \sum_{R \in N_R} s(\text{val}_R(C), \text{val}_R(D)) + \frac{1}{2} \sum_{R \in N_R} s_N((\text{min}_R(C), \text{max}_R(C)), (\text{min}_R(D), \text{max}_R(D))) \right]
\]
In order to compute $s_P$ let us note that:

- $\text{prim}(C) = \text{Person}$
- $\text{prim}(D) = \text{Male}$

$$s_P(\{\text{Person}\}, \{\text{Male}\}) =$$

$$= \frac{|\{\text{Meg},\text{Bob},\text{Pat},\text{Gwen},\text{Ann},\text{Sue},\text{Tom}\} \cap \{\text{Bob},\text{Pat},\text{Tom}\}|}{|\{\text{Meg},\text{Bob},\text{Pat},\text{Gwen},\text{Ann},\text{Sue},\text{Tom}\} \cup \{\text{Bob},\text{Pat},\text{Tom}\}|} =$$

$$= \frac{|\{\text{Bob},\text{Pat},\text{Tom}\}|}{|\{\text{Meg},\text{Bob},\text{Pat},\text{Gwen},\text{Ann},\text{Sue},\text{Tom}\}|} = \frac{3}{7}$$
To compute $s$ for value restrictions, it is important to note that

- $N_R = \{\text{hasChild}, \text{marriedTo}\}$
- $val_{\text{marriedTo}}(C) = \text{Person}$ and $val_{\text{hasChild}}(C) = \top$
- $val_{\text{marriedTo}}(D) = \text{Person} \sqcap \neg \text{Male}$ and $val_{\text{hasChild}}(D) = \top$

$$s(\text{Person}, \text{Person} \sqcap \neg \text{Male}) + s(\top, \top) =$$

$$= \frac{1}{3} \cdot (s_P(\text{Person}, \text{Person} \sqcap \neg \text{Male}) + \frac{1}{2} \cdot (1 + 1) + \frac{1}{2} \cdot (1 + 1)) +$$

$$+ \frac{1}{3} \cdot (1 + 1 + 1) = \frac{1}{3} \cdot (\frac{4}{7} + 1 + 1) + 1 = \frac{13}{7}$$
...Similarity Measure: example...

To compute $s$ for number restrictions it is important to note that

- $N_R = \{\text{hasChild, marriedTo}\}$
- $\min_{\text{marriedTo}}(C) = 0; \quad \max_{\text{marriedTo}}(C) = |\Delta| + 1 = 7 + 1 = 8$
  $\min_{\text{hasChild}}(C) = 0; \quad \max_{\text{hasChild}}(C) = 1$
- $\min_{\text{marriedTo}}(D) = 0; \quad \max_{\text{marriedTo}}(D) = |\Delta| + 1 = 7 + 1 = 8$
  $\min_{\text{hasChild}}(D) = 0; \quad \max_{\text{hasChild}}(D) = 2$
- $\min(M_C, M_D) > \max(m_C, m_D)$

\[
s_N((m_{\text{hasChild}}(C), M_{\text{hasChild}}(C)), (m_{\text{hasChild}}(D), M_{\text{hasChild}}(D))) + \\
+ s_N((m_{\text{marriedTo}}(C), M_{\text{marriedTo}}(C)), (m_{\text{marriedTo}}(D), M_{\text{marriedTo}}(D))) = \\
= \frac{\min(M_{\text{hasChild}}(C), M_{\text{hasChild}}(D)) - \max(m_{\text{hasChild}}(C), m_{\text{hasChild}}(D)) + 1}{\max(M_{\text{hasChild}}(C), M_{\text{hasChild}}(D)) - \min(m_{\text{hasChild}}(C), m_{\text{hasChild}}(D)) + 1} + 1 = \\
= \frac{\min(1,2) - \max(0,0) + 1}{\max(1,2) - \min(0,0) + 1} + 1 = \frac{2}{3} + 1 = \frac{5}{3}
\]
Measure Involving Individuals

Let $c$ and $d$ two individuals in a given A-Box. We can consider $C^* = \text{MSC}^*(c)$ and $D^* = \text{MSC}^*(d)$:

$$s(c, d) := s(C^*, D^*) = s(\text{MSC}^*(c), \text{MSC}^*(d))$$

Analogously:

$$\forall c : s(c, D) := s(\text{MSC}^*(c), D)$$
The similarity value is mainly determined as the amount of overlapping sets of individuals that are extension of the concepts involved, considering also their sub-concepts.

- The influence of sub-concepts in determining similarity value decreases w.r.t. their nesting level.

The similarity measure is defined recursively.

- Its complexity mainly depends on the complexity of the *Instance checking* operator.
  - Limited to primitive concepts.
  - It can be pre-compiled.
Conclusions

- The presented function $s$ is a *Similarity Measure*
  - it is definite positive, symmetric, and has maximal value only when the concepts are equivalent.
- The presented Similarity Measure is based on the A-Box *semantics* and it is applicable also to couples of individuals, or a concepts and an individual.
- $s$ is defined using the set theory and reasoning operators
  - It uses a numerical approach but is applied on symbolic representations.
Further Developments

- **Testing** the Similarity Measure using some *classification* and *clustering algorithms*

- (Ongoing) Extension of the measure for more expressive DL such as *ALCN*

- Definition of new Similarity/Dissimilarity Measures for DLs representations, using *Kernel functions* that are a means to express a notion of similarity in some unknown feature space. Thus it could be possible exploiting the efficiency of kernel methods (e.g. SVMs) in a relational setting
Thanks
For Your Attention