Trimming a consistent DL knowledge base, relying on linguistic evidence

supervisors : Laure Vieu et Nathalie Aussenac-Gilles

Julien Corman

IRIT

DL and OWL

Description Logics (DL):

- Decidable fragments of FOL
- $\mathcal{ALC}$ = "modal fragment" of FOL : unary and binary predicates only (called *atomic concepts* and *roles*), no identity, no function, restrictions on quantification (see appendix).
- Extensions : nominals, cardinality restriction, role subsumption, role composition, inverse roles, ...
- Algorithms and libraries for different tasks/problems : consistency, entailment, modularity, minimal conflicts, ...

DL and OWL

Description Logics (DL):

- Decidable fragments of FOL
- $\mathcal{ALC}$ = "modal fragment" of FOL : unary and binary predicates only (called *atomic concepts* and *roles*), no identity, no function, restrictions on quantification (see appendix).
- Extensions : nominals, cardinality restriction, role subsumption, role composition, inverse roles, ...
- Algorithms and libraries for different tasks/problems : consistency, entailment, modularity, minimal conflicts, ...

OWL 2

- Knowledge representation language, W3C recommendation.
- Equivalent to the DL $\mathcal{SROIQ}^{(\mathcal{D})}$
- Several syntaxes, among which a (hardly readable) RDF serialization.

Example 1

DBPedia

$\texttt{hasKeyPerson}(\textit{Virgin Holidays}, \textit{CEO}).$
$\texttt{hasKeyPerson}(\textit{Caixa Bank}, \textit{CEO}).$
$\texttt{hasOccupation}(\textit{Peter Munk}, \textit{CEO}).$

Example 1

DBPedia

$\mathtt{hasKeyPerson}(\textit{Virgin Holidays}, \textit{CEO})$.
$\mathtt{hasKeyPerson}(\textit{Caixa Bank}, \textit{CEO})$.
$\mathtt{hasOccupation}(\textit{Peter Munk}, \textit{CEO})$.
$\mathtt{hasKeyPerson}(\textit{BrookField Office Properties}, \textit{Peter Munk})$.
$\top \sqsubseteq \forall \mathtt{hasKeyPerson}.\mathtt{Person}$.

Example 1

DBPedia

hasKeyPerson(*Virgin Holidays*, *CEO*).
hasKeyPerson(*Caixa Bank*, *CEO*).
hasOccupation(*Peter Munk*, *CEO*).
hasKeyPerson(*BrookField Office Properties*, *Peter Munk*).
$\top \sqsubseteq \forall$hasKeyPerson.Person.

- Intuitively absurd : violates for instance "No individual (*CEO* here) can be both a person and the occupation of a person".
- More pragmatically, may lead to erroneous inferences : e.g. *Virgin Holidays* and *Caixa Bank* have the same Person as a keyPerson.
- But logically consistent and coherent.

Example 2

DBPedia

$\mathtt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution})$.
$\mathtt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork})$.
$\mathtt{owningCompany}(\textit{Smithsonian Networks}, \textit{Smithsonian Institution})$.
$\top \sqsubseteq \forall \mathtt{doctoralAdvisor}.\mathtt{Person}$.

- Still logically consistent and coherent.

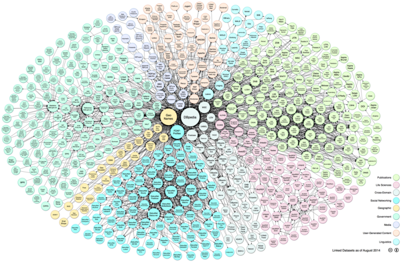
Example 2

DBPedia

$\texttt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution})$.
$\texttt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork})$.
$\texttt{owningCompany}(\textit{Smithsonian Networks}, \textit{Smithsonian Institution})$.
$\top \sqsubseteq \forall \texttt{doctoralAdvisor}.\texttt{Person}$.

- Still logically consistent and coherent.
- These are not just "factual" errors, like $\texttt{director}(\textit{Citizen Kane}, \textit{Woody Allen})$.
- Source of the problem :
  - genuine typos
  - incompatible understandings/uses of a same DL individual/concept/role.

LOD



- Billions of RDF triples, a large part is OWL expressible.
- Sources : handwritten statements, serialized DBs, automatically extracted data, . . .
- Interoperability $\approx$ signatures overlap.
- Low expressiveness overall : e.g. negation is discouraged.
- Consequence : absurd but consistent sets of statements.

OWL data : consistent/coherent by default

- One of the following is necessary for an OWL 2 dataset to be inconsistent/incoherent :
  - `owl:complementOf` or `owl:disjointWith`
  - `owl:negativeObjectPropertyAssertion`
  - `owl:disjointObjectProperties`, `owl:AsymmetricProperty` or `owl:irreflexiveObjectProperty`.
  - `owl:oneOf`
  - `owl:Nothing`
  - `owl:objectMaxCardinality`
  - etc...

OWL data : consistent/coherent by default

- One of the following is necessary for an OWL 2 dataset to be inconsistent/incoherent :
  - `owl:complementOf` or `owl:disjointWith`
  - `owl:negativeObjectPropertyAssertion`
  - `owl:disjointObjectProperties`, `owl:AsymmetricProperty` or `owl:irreflexiveObjectProperty`.
  - `owl:oneOf`
  - `owl:Nothing`
  - `owl:objectMaxCardinality`
  - etc...

Rarely used (source : LODStats (LODCLoud sample))

- `owl:subClassOf` : > 89 000 occ.
- `owl:complementOf` : 2 occ.
- `owl:disjointWith` : 33 occ.

Proposal

- Automatically gathered linguistic evidence in order to detect and repair such violations of common sense.
  - Detect : identify consequences of a set Γ of axioms which are unlikely to hold if the rest of $Cn(\Gamma)$ does.
  - Repair : suggest axioms to be preferably discarded or amended
- Linguistic input : web pages

Proposal

- Automatically gathered linguistic evidence in order to detect and repair such violations of common sense.
  - Detect : identify consequences of a set Γ of axioms which are unlikely to hold if the rest of $Cn(\Gamma)$ does.
  - Repair : suggest axioms to be preferably discarded or amended
- Linguistic input : web pages
- Main hypothesis (distributional evidence) : individuals which share linguistic contexts tend to instantiate the same concepts. Inspiration : ontology population/named entity classification (Tanev and Magnini, . . .)

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\texttt{doctoralAdvisor}(\mathit{Thaddeus\ S.C.\ Lowe}, \mathit{Smithsonian\ Institution})$,
$\texttt{doctoralAdvisor}(\mathit{Nick\ Katz}, \mathit{Bernard\ Dwork})$,
$\top \sqsubseteq \forall \texttt{doctoralAdvisor}.\texttt{Person}, \ldots \}$

- $\Gamma \models \texttt{Person}(\mathit{Smithsonian\ Institution})$
  $\Gamma \models \texttt{Person}(\mathit{Bernard\ Dwork})$

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\texttt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution}),$
$\texttt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork}),$
$\top \sqsubseteq \forall \texttt{doctoralAdvisor.Person}, \dots \}$

- $\Gamma \models \texttt{Person}(\textit{Smithsonian Institution})$
  $\Gamma \models \texttt{Person}(\textit{Bernard Dwork})$
- Assume also that :
  $\Gamma \models \texttt{Person}(\textit{Margaret Atwood})$
  $\Gamma \models \texttt{Person}(\textit{Peter Munk})$
  $\Gamma \models \texttt{Person}(\textit{Thaddeus S.C. Lowe}) \dots$
- Does "the Smithsonian institution" behave like terms denoting other instances of Person according to Γ ?
  Does "Bernard Dwork" behave like terms denoting other instances of Person according to Γ ?

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\texttt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution}),$
$\texttt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork}),$
$\top \sqsubseteq \forall \texttt{doctoralAdvisor}.\texttt{Person}, \ldots \}$

- $\Gamma \models \texttt{Person}(\textit{Smithsonian Institution})$
  $\Gamma \models \texttt{Person}(\textit{Bernard Dwork})$
- "#the Smithsonian Institution was born"
  "Bernard Dwork was born"

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\texttt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution}),$
$\texttt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork}),$
$\top \sqsubseteq \forall \texttt{doctoralAdvisor}.\texttt{Person}$
$\texttt{Organization} \sqsubseteq \neg\texttt{Person}, \ldots\}$

- $\Gamma \models \texttt{Person}(\textit{Smithsonian Institution})$
  $\Gamma \models \texttt{Person}(\textit{Bernard Dwork})$
  $\Gamma \models \neg\texttt{Organization}(\textit{Smithsonian Institution})$
  $\Gamma \models \neg\texttt{Organization}(\textit{Bernard Dwork})$

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\texttt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution}),$
$\texttt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork}),$
$\top \sqsubseteq \forall \texttt{doctoralAdvisor}.\texttt{Person}$
$\texttt{Organization} \sqsubseteq \neg\texttt{Person}, \ldots\}$

- $\Gamma \models \texttt{Person}(\textit{Smithsonian Institution})$
  $\Gamma \models \texttt{Person}(\textit{Bernard Dwork})$
  $\Gamma \models \neg\texttt{Organization}(\textit{Smithsonian Institution})$
  $\Gamma \models \neg\texttt{Organization}(\textit{Bernard Dwork})$
- "the Smithsonian Institution was established", "the Smithsonian Institution's workforce"
  "#Bernard Dwork was established", "#Bernard Dwork's workforce"

Linguistic evidence : intuition

Example

$\Gamma = \{$
$\mathtt{doctoralAdvisor}(\mathit{Thaddeus\ S.C.\ Lowe}, \mathit{Smithsonian\ Institution})$,
$\mathtt{doctoralAdvisor}(\mathit{Nick\ Katz}, \mathit{Bernard\ Dwork})$,
$\top \sqsubseteq \forall \mathtt{doctoralAdvisor.Person}$
$\mathtt{Organization} \sqsubseteq \neg \mathtt{Person}, \ldots \}$

- Linguistic contexts may help identify :
  - plausible consequences of Γ : $\mathtt{Person}(\mathit{Bernard\ Dwork})$, $\neg\mathtt{Organization}(\mathit{Bernard\ Dwork})$
  - implausible consequences of Γ : $\mathtt{Person}(\mathit{Smithsonian\ Institution})$, $\neg\mathtt{Organization}(\mathit{Smithsonian\ Institution})$

Choices

- Focus on Ψ_Γ : consequences of Γ the form $A(e)$ or $\neg A(e)$, with A an atomic concept and e an individual.
- Linguistic terms labeling concepts and roles are never used (only terms labeling individuals).

Individual labels rather than concept labels ?

- Concept labels tend to be more polysemous : e.g. "Group", "Function", "Element", . . .
- Lack of linguistic occurrences for :
  - Ad hoc concepts labels : ex (eGov ontologies) : "Triple path", "Structuring event type" (0 google occ.)
  - Abstract concepts : e.g. "perdurant"

Unary rather than binary predicates ?

- labels already known $\Rightarrow$ lack of linguistic cooccurrences.

Similarity

- Distributional hypothesis : represent a term t by its linguistic contexts
- A context c :
  - sequence of words preceding/surrounding/following an occurrence of the term, possibly lemmatized
  - syntactic dependency, ...
  - ignoring punctuation, determiners, ...
- A terms t is represented as a vector $\mathbf{v}^t$ of frequencies with each observed context.

Similarity

- Distributional hypothesis : represent a term t by its linguistic contexts
- A context c :
  - sequence of words preceding/surrounding/following an occurrence of the term, possibly lemmatized
  - syntactic dependency, ...
  - ignoring punctuation, determiners, ...
- A terms t is represented as a vector $\mathbf{v}^t$ of frequencies with each observed context.
- Weighting observed frequencies :
  - $PMI(c, e) = -\log \frac{p(c,e)}{p(e) \cdot p(c)}$
  - self-information (Giulano and Gliozzo) : $\mathrm{self}(c) = -\log(p(c))$, with $p(c)$ obtained from an external language model...

Similarity

- Distributional hypothesis : represent a term t by its linguistic contexts
- A context c :
  - sequence of words preceding/surrounding/following an occurrence of the term, possibly lemmatized
  - syntactic dependency, ...
  - ignoring punctuation, determiners, ...
- A terms t is represented as a vector $\mathbf{v}^t$ of frequencies with each observed context.
- Weighting observed frequencies :
  - $PMI(c, e) = -\log \frac{p(c,e)}{p(e) \cdot p(c)}$
  - self-information (Giulano and Gliozzo) : $\mathrm{self}(c) = -\log(p(c))$, with $p(c)$ obtained from an external language model...
- Reducing vector dimensions : latent semantic analysis (SVD), latent Dirichlet allocation, skip-gram model, ...
- Similarity $\mathrm{sim}(t_1, t_2)$ given by some distance (cosine, ...) between $\mathbf{v}^{t_1}$ and $\mathbf{v}^{t_2}$.

Plausibility of $A(e) \in \mathrm{Cn}(\Gamma)$

Notation :

- $\mathrm{sim}(e, e')$: similarity between distributional representations of terms denoting e and e'.
- $\mathrm{inst}_\Gamma(A) = \{e' \mid \Gamma \models A(e')\}$
- $S = \mathrm{inst}_\Gamma(A) \setminus \{e\}$: *support set* for $A(e)$.
- $\mathrm{sim}(e, S) \doteq \sum_{e' \in S} \frac{\mathrm{sim}(e,e')}{|S|}$
- $X^\Gamma_{e,|S|}$ (random variable) : expected average similarity between e and $|S|$ random individuals of $\mathrm{inst}_\Gamma(\top) \setminus \{e\}$.

Plausibility score $\mathrm{sc}_\Gamma(A(e))$

- $\mathrm{sc}_\Gamma(A(e)) = p(X^\Gamma_{e,|S|} \leq \mathrm{sim}(e, S))$
- Measures how surprisingly high the similarity between e and individuals of S is.
- Based on the similarity between e and all individuals.

Support set S

- $S = \text{inst}_\Gamma(A) \setminus \{e\}$: *support set* for $A(e)$.
- What about $\text{inst}_\Gamma(\neg A)$?
- Linguistically unrealistic : no reason to think that two instances of $\neg A$ should behave similarly.

Example

- $\Gamma \models \neg\mathtt{Person}(\mathit{WW2})$
- $\Gamma \models \mathtt{Person}(\mathit{Thelonious\ Monk})$
- $\text{sim}(\mathit{Smithsonian\ Institution}, \mathit{WW2}) >$ $\text{sim}(\mathit{Smithsonian\ Institution}, \mathit{Thelonious\ Monk})$???

Support set S

- $S = \text{inst}_\Gamma(A) \setminus \{e\}$: *support set* for $A(e)$.
- What about $\text{inst}_\Gamma(\neg A)$?
- Linguistically unrealistic : no reason to think that two instances of $\neg A$ should behave similarly.

Example

- $\Gamma \models \neg\texttt{Person}(WW2)$
- $\Gamma \models \texttt{Person}(\textit{Thelonious Monk})$
- $\text{sim}(\textit{Smithsonian Institution}, WW2) >$ $\text{sim}(\textit{Smithsonian Institution}, \textit{Thelonious Monk})$???

- Support set for $\neg A(e)$: $S = \text{inst}_\Gamma(A)$

Plausibility of $\neg A(e) \in \mathrm{Cn}(\Gamma)$

Notation :

- $\mathrm{sim}(e, e')$: similarity between distributional representations of e and e'.
- $\mathrm{inst}_\Gamma(A) = \{e' | \Gamma \models A(e')\}$
- $S = \mathrm{inst}_\Gamma(A)$: *support set* for $\neg A(e)$.
- $X^\Gamma_{e,|S|}$ (random variable) : expected average similarity between e and $|S|$ random individuals of $\mathrm{inst}_\Gamma(\top) \setminus \{e\}$.

Plausibility score $\mathrm{sc}_\Gamma(\neg A(e))$

- $\mathrm{sc}_\Gamma(A(e)) = p(X^\Gamma_{e,|S|} \geq \sum_{e' \in S} \frac{\mathrm{sim}(e,e')}{|S|})$
- Measures how surprisingly low the similarity between e and individuals of S is.
- Based on the similarity between e and all individuals.

Expected similarity

- $X^{\Gamma}_{e,|S|}$: expected average similarity between e and $|S|$ random individuals of $\mathrm{inst}_{\Gamma}(\top) \setminus \{e\}$.
- Intuition : *ceteris paribus*, the lower $|S|$, the less informative $\mathrm{sim}(e, S)$ should be.
- The lower $|S|$, the more uniform de distribution of $X^{\Gamma}_{e,|S|}$ should be.

Distribution of $X^{\Gamma}_{e,|S|}$

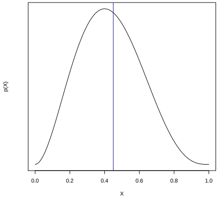
- $m \doteq \mathrm{sim}(e, \mathrm{inst}_{\Gamma}(\top) \setminus \{e\})$
- $X^{\Gamma}_{e,|S|} \sim \mathrm{Beta}(m|S| + 1, (1 - m)|S| + 1)$

Expected plausibility : example

- $\psi_1 = A(e)$, $\psi_2 = B(e)$
- $m = \mathrm{sim}(e, \mathrm{inst}_\Gamma(\top) \setminus \{e\}) = 0.4$

ψ_1

- $S = \mathrm{inst}_\Gamma(A) \setminus \{e\}$
- $|S| = 5$
- $\mathrm{sim}(e, S) = 0.45$
- $\mathrm{sc}_\Gamma(\psi_1) = 0.558$

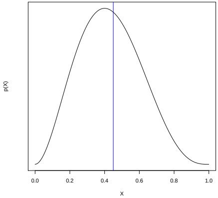


Expected plausibility : example

- $\psi_1 = A(e)$, $\psi_2 = B(e)$
- $m = \mathrm{sim}(e, \mathrm{inst}_\Gamma(\top) \setminus \{e\}) = 0.4$

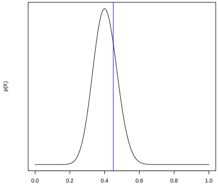
ψ_1

- $S = \mathrm{inst}_\Gamma(A) \setminus \{e\}$
- $|S| = 5$
- $\mathrm{sim}(e, S) = 0.45$
- $\mathrm{sc}_\Gamma(\psi_1) = 0.558$



ψ_2

- $S = \mathrm{inst}_\Gamma(B) \setminus \{e\}$
- $|S| = 50$
- $\mathrm{sim}(e, S) = 0.45$
- $\mathrm{sc}_\Gamma(\psi_2) = 0.754$



Trimming

- An input KB K.
- Objective : use plausibility scores to decide which axioms should be preferebly discarded or amended within K.
- Equivalently, select the optimal $\Gamma_1, .., \Gamma_n \in 2^K$.

Linguistic compliance comp : $2^K \mapsto \mathbb{R}$

$$\text{comp}(\Gamma) = \sum_{\psi \in \Psi_\Gamma} \frac{\text{sc}_\Gamma(\psi)}{|\Psi_\Gamma|}$$

Trimming

- An input KB K.
- Objective : use plausibility scores to decide which axioms should be preferebly discarded or amended within K.
- Equivalently, select the optimal $\Gamma_1, .., \Gamma_n \in 2^K$.

Linguistic compliance comp : $2^K \mapsto \mathbb{R}$

$$\text{comp}(\Gamma) = \sum_{\psi \in \Psi_\Gamma} \frac{\text{sc}_\Gamma(\psi)}{|\Psi_\Gamma|}$$

- $\prec$: strict partial order over 2^K : $\Gamma_1 \prec \Gamma_2$ iff either $\text{comp}(\Gamma_1) < \text{comp}(\Gamma_2)$, or ($\text{comp}(\Gamma_1) = \text{comp}(\Gamma_2)$ and $\Gamma_1 \subset \Gamma_2$).
- Assumption : focus on syntax (see appendix).
- Output $\mathcal{O}$: intersection, or possibly disjunction of the subbases which are maximal wrt $\prec$.

Trimming : practical limits

- Maximizing comp is not trivial :
  - comp(Γ) is not directly a function of Γ, but of Ψ_Γ : so there may be an optimal $\Psi' \subseteq \Psi_K$, and no Γ such that $\Psi_\Gamma = \Psi'$.
  - For $\psi \in \Psi_{\Gamma_1} \cap \Psi_{\Gamma_2}$, $\mathrm{sc}_{\Gamma_1}(\psi) \neq \mathrm{sc}_{\Gamma_2}(\psi)$ in general, because the support sets for ψ differ in Γ_1 and Γ_2.
- The output $\mathcal{O}$ can be very weak, e.g. if $|\mathcal{O}| < 0.5 * |K|$

Trimming : practical limits

- Maximizing comp is not trivial :
  - comp(Γ) is not directly a function of Γ, but of Ψ_Γ : so there may be an optimal $\Psi' \subseteq \Psi_K$, and no Γ such that $\Psi_\Gamma = \Psi'$.
  - For $\psi \in \Psi_{\Gamma_1} \cap \Psi_{\Gamma_2}$, $\mathsf{sc}_{\Gamma_1}(\psi) \neq \mathsf{sc}_{\Gamma_2}(\psi)$ in general, because the support sets for ψ differ in Γ_1 and Γ_2.
- The output $\mathcal{O}$ can be very weak, e.g. if $|\mathcal{O}| < 0.5 * |K|$

More plausible scenarios

- Search space previously circumscribed : e.g. discard at most n axioms.
- (Iteratively) discard the worst axiom (see evaluation).

Alternatives to comp

Linguistic compliance $\mathrm{comp}_K : 2^K \mapsto \mathbb{R}$

$$\mathrm{comp}_K(\Gamma) = \sum_{\psi \in \Psi_\Gamma} \frac{\mathrm{sc}_{\boldsymbol{K}}(\psi)}{|\Psi_\Gamma|}$$

- More amenable to optimizations.
- Ex (trivial) : a subbase Γ_1 with $\max\limits_{\psi \in \Psi_{\Gamma_1}} \mathrm{sc}_K(\psi) < \mathrm{comp}_K(\Gamma_2)$ for some already evaluated subbase Γ_2.
  $\Rightarrow$ No subbase of Γ_1 can be optimal wrt $\prec$.
- Drawback : potentially higher number of optimal subbases.

Alternatives to comp

Lexicographic ordering $\preceq_{lex} \subseteq 2^K \times 2^K$

- Instead of plausibilities mean, penalize subbases whose consequences have a low plausibility (see appendix)

- Then $\prec$ is defined by $\Gamma_1 \prec \Gamma_2$ iff either $\Gamma_1 \prec_{lex} \Gamma_2$, or ($\Gamma_1 =_{lex} \Gamma_2$ and $\Gamma_1 \subset \Gamma_2$).

Lexicographic ordering $\preceq_{lex_K} \subseteq 2^K \times 2^K$

- Identical to $\preceq_{lex}$, but using sc_K instead of sc_Γ for plausibility.

- Closer to traditional KB debugging / belief base revision : identify undesired consequences within K before trimming.

Input 1 : real data

Real input KB

- Source : LOD
- Evaluation procedure : manually verify if consequences with lowest plausibility and discarded axioms are actually erroneous.
- Advantage : plausible data
- Drawback : subjective evaluation (low inter annotator agrement)

- Dataset K_{DBP} : 5721 (logical) axioms automatically extracted from DBPedia (see appendix).
- 1095 individuals
- ABox + TBox
- expressivity : $\mathcal{AL}^{(D)}$

Input 2 : artificially degraded data

Artificially degraded KB

- Source : higher quality KB
- Degrading procedure : randomly select an axiom ϕ of K, and generate ϕ' by replacing *sign*(ϕ) with random elements of *sign*(K). The syntactic structure remains unchanged.
- Requirements : the resulting base $K' = K \cup \{\phi'\}$ must be consistent, and $|\Psi_K| < |\Psi_{K'}|$.
- Assumption : random axioms are very likely be absurd, and so random consequences to be outliers within $\Psi_{K'}$.

Input 2 : artificially degraded data

Artificially degraded KB

- Source : higher quality KB
- Degrading procedure : randomly select an axiom ϕ of K, and generate ϕ' by replacing *sign*(ϕ) with random elements of *sign*(K). The syntactic structure remains unchanged.
- Requirements : the resulting base $K' = K \cup \{\phi'\}$ must be consistent, and $|\Psi_K| < |\Psi_{K'}|$.
- Assumption : random axioms are very likely be absurd, and so random consequences to be outliers within $\Psi_{K'}$.
- Evaluation : automatically retrieve the generated axioms and consequences within K' and $\Psi_{K'}$ respectively.
- Drawback : artificial data
- Advantage : objective evaluation

Input 2 : artificially degraded KB

- Dataset K_F : 1028 axioms automatically extracted from the NEON fisheries ontology (see appendix).
- 71 indivduals
- ABox + TBox
- expressivity : $\mathcal{SI}$

Linguistic input

Corpora

- Web pages retrieved with a search engine, using individals' labels as queries.
- K_{DBP} : $\approx$ 57 000 pages, K_F : $\approx$ 6 300 pages

Linguistic input

Corpora

- Web pages retrieved with a search engine, using individuals' labels as queries.
- K_{DBP} : $\approx$ 57 000 pages, K_F : $\approx$ 6 300 pages

Linguistic contexts

- **LP** : (customized) sequences of surrounding lemma-POS (shifting window), frequencies weighted with PMI
  Limit : "more results about X", "more about X on Twitter", ...
- **NP** : Ngrams preceding or following the term, frequencies weighted with PMI
- **NS** : Ngrams, frequencies weighted with self-information (querying the Microsoft Web N-gram corpus).
- **NPS** : Ngrams + PMI + self-information.

- Similarities : cosine distances

Evaluation : plausibility

- Input : K_F
- Generation of 100 random axioms $\phi_1, \ldots, \phi_{100}$ out of K_F.
- $K_1, \ldots, K_{100}$: 100 input KBs, such that $K_i = K_F \cup \{\phi_i\}$.
- For each K_i, order Ψ_{K_i} by plausibility.
- $\Psi^{rand}_{K_i} = \Psi_{K_i} - \Psi_{K_F}$.

Evaluation : plausibility

- Input : K_F
- Generation of 100 random axioms $\phi_1, \ldots, \phi_{100}$ out of K_F.
- $K_1, \ldots, K_{100}$: 100 input KBs, such that $K_i = K_F \cup \{\phi_i\}$.
- For each K_i, order Ψ_{K_i} by plausibility.
- $\Psi_{K_i}^{rand} = \Psi_{K_i} - \Psi_{K_F}$.

	rank	p-val
LP	4.15 / 216.1	$<$0.001
NP	9.73 / 216.1	$<$0.001
NS	7.33 / 216.1	$<$0.001
NPS	5.59 / 216.1	$<$0.001

Average ranking among Ψ_{K_i} of the lowest-ranked formula of $\Psi_{K_i}^{rand}$, and p-value for the rankings of all formulas of all $\Psi_{K_i}^{rand}$

- For most K_i (75/100), $|\Psi_{K_i}^{rand}| = 1$. In most of theses cases (57/75), the only formula in $\Psi_{K_i}^{rand}$ was also the one with lowest plausibility in Ψ_{K_i}.

Evaluation : trimming

- For each K_i, the set $\Delta_i = \Gamma_{i,1}, \dots, \Gamma_{i,1029}$ of all immediate subbase of K_i was computed.
- Within Δ_i, all $\Gamma_{i,j}$ such that $\Psi_{\Gamma_{i,j}} \neq \Psi_{K_i}$ were ordered according to $\prec$.
- Weighting : LP (lemmaPos + PMI)

	rank	p-val
comp(Γ)	7.86 / 80.03	< 0.001
$\text{comp}_{K_i}(\Gamma)$	8.05 / 80.03	< 0.001
$\preceq_{lex}$	6.51 / 80.03	< 0.001
$\preceq_{lex_{K_i}}$	2.47 / 80.03	< 0.001

Average ranking of the randomly generated statement ϕ_i for each K_i, and p-value for the rankings of all ϕ_i

Evaluation : iterated trimming, K_F

- $K' = K_F$ extended with 20 random axioms
- $|K'| = 1028 + 20 = 1048$

		val.	prec. & rec.	p-val (prop. test)
NPS	comp	9	0.45	< 0.001
	comp_K	9	0.45	< 0.001
	$\preceq_{lex}$	3	0.15	< 0.002
	$\preceq_{lex_{\mathbf{K}}}$	9	0.45	< 0.001
LP	comp	10	0.5	< 0.001
	comp_K	10	0.5	< 0.001
	$\preceq_{lex}$	5	0.25	< 0.001
	$\preceq_{lex_{\mathbf{K}}}$	10	0.5	< 0.001

Table: Randomly generated axioms among the first 20 discarded ones

Evaluation : iterated trimming, K_{DBP}

- $|K_{DBP}| = 5721$

		val.	prec.
NPS	comp	7	0.35
	$\preceq_{lex}$	3	0.15
LP	comp	11	0.55
	$\preceq_{lex}$	5	0.25

Table: Actually erroneous axioms among the 20 first discarded ones

Extensions

Complex concepts

- Most DLs allow the construction of arbitrary complex DL concepts, e.g. $\exists\texttt{doctoralAdvisor}.\top$
- They could (in theory) be used instead of A.
- If Ψ^*_Γ is the set of all resulting consequences, no finite subset Ψ' of Ψ^*_Γ is such that $\Psi^*_\Gamma \subseteq \mathrm{Cn}(\Psi')$.
  $\Rightarrow$ Need to choose among these concepts.
- Some complex concepts are not relevant linguistically, e.g. $(\texttt{Moldavian} \sqcup \texttt{Muslim}) \sqcap \texttt{Lawyer} \sqcap \exists\texttt{hasFather}.\forall\texttt{livesIn}.\texttt{Appartment}$

Extensions

Complex concepts

- Most DLs allow the construction of arbitrary complex DL concepts, e.g. $\exists\mathtt{doctoralAdvisor}.\top$
- They could (in theory) be used instead of A.
- If Ψ^*_Γ is the set of all resulting consequences, no finite subset Ψ' of Ψ^*_Γ is such that $\Psi^*_\Gamma \subseteq \mathrm{Cn}(\Psi')$.
  $\Rightarrow$ Need to choose among these concepts.
- Some complex concepts are not relevant linguistically, e.g. $(\mathtt{Moldavian} \sqcup \mathtt{Muslim}) \sqcap \mathtt{Lawyer} \sqcap \exists\mathtt{hasFather}.\forall\mathtt{livesIn}.\mathtt{Appartment}$

$e \neq e'$

- Set $\Psi_\Gamma = \{\psi = e \neq e' \mid \Gamma \models \psi\}$
- Penalize comp(Γ) if $\sim(e, e')$ is high.

Modal fragment (MF) of FOL ($= \mathcal{ALC}$)

- If A is a unary predicate, then $A(x) \in$ MF.
- MF if closed under boolean operators.
- If $\phi \in$ MF, y does not appear in ϕ, and R is a binary predicate, then :
  - $\exists y(R(x, y) \wedge \phi[x/y]) \in$ MF
  - $\forall y(R(x, y) \rightarrow \phi[x/y]) \in$ MF

Trimming with $\preceq_{lex_K}$

Example

$\Omega = \{$
(1) $\mathtt{doctoralAdvisor}(\textit{Thaddeus S.C. Lowe}, \textit{Smithsonian Institution})$,
(2) $\mathtt{doctoralAdvisor}(\textit{Nick Katz}, \textit{Bernard Dwork})$,
(3) $\top \sqsubseteq \forall \mathtt{doctoralAdvisor}.\mathtt{Person}$
(4) $\mathtt{Organization} \sqsubseteq \neg\mathtt{Person}\}$

- Assume `doctoralAdvisor`, *Bernard Dwork* and *Smithsonian Institution* do not appear in $\Gamma \setminus \Omega$.
- Trimming :
  - discarding axioms in order to give up implausible consequences, but retain plausible ones.
  - no axiom should be unnecessarily discarded
- Only one solution here : discarding (1).

Trimming : assumptions

- $\prec$: strict partial order over 2^K : $\Gamma_1 \prec \Gamma_2$ iff either $comp(\Gamma_1) < comp(\Gamma_2)$, or $(comp(\Gamma_1) = comp(\Gamma_2)$ and $\Gamma_1 \subset \Gamma_2)$.
- Minimize syntactic information loss whenever possible, i.e. Γ_1 and Γ_2 viewed as bases, not theories.
  In particular :
  - If $Cn(\Gamma_1) = Cn(\Gamma_2)$, but $\Gamma_1 \not\subseteq \Gamma_2$ and $\Gamma_2 \not\subseteq \Gamma_1$, then Γ_1 and Γ_2 are not comparable wrt $\prec$.
  - Redundancies should be preserved when possible : if $Cn(\Gamma_1) = Cn(\Gamma_2)$ and $\Gamma_1 \subset \Gamma_2$, then $\Gamma_1 \prec \Gamma_2$ still holds.

Lexicographic ordering $\preceq_{lex}$

- $\boldsymbol{\omega}_\Gamma \doteq \omega_\Gamma^1, .., \omega_\Gamma^{|\Psi_\Gamma|}$: formulas of Ψ_Γ order by increasing score sc_Γ
- $\mathsf{sc}_\Gamma(\boldsymbol{\omega}_\Gamma) = \mathsf{sc}_\Gamma(\omega_\Gamma^1), .., \mathsf{sc}_\Gamma(\omega_\Gamma^{|\Psi_\Gamma|})$
- $\preceq_{lex}$ defined by $\Gamma_1 \preceq_{lex} \Gamma_2$ iff either :
  - $\mathsf{sc}_{\Gamma_1}(\boldsymbol{\omega}_{\Gamma_1}) = \mathsf{sc}_{\Gamma_2}(\boldsymbol{\omega}_{\Gamma_2})$, or
  - there is a $1 \leq i \leq |\Psi_{\Gamma_2}|$ such that $\mathsf{sc}_{\Gamma_1}(\omega_{\Gamma_1}^j) = \mathsf{sc}_{\Gamma_2}(\omega_{\Gamma_2}^j)$ for all $1 \leq j < i$, and either $\mathsf{sc}_{\Gamma_1}(\omega_{\Gamma_1}^i) < \mathsf{sc}_{\Gamma_2}(\omega_{\Gamma_2}^i)$ or $|\Psi_{\Gamma_1}| = i - 1$