Experimental Evaluation of General Concept Inclusions Learned from Textual Data.

Hybris B1: Automatic Generation of Description Logic-based Biomedical Ontologies

Daniel Borchmann, Anas Elghafari, and Yue Ma

TU Dresden

2015-06-08

Goal

Automatically construct biomedical ontologies from text:

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Example (Terminological Knowledge)

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Example (Terminological Knowledge)

- Genes are not protein complexes, and vice versa.

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Example (Terminological Knowledge)

- Genes are not protein complexes, and vice versa.

$$\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$$

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Example (Terminological Knowledge)

- Genes are not protein complexes, and vice versa.

$$\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$$

- Proteins contain amino acids

Goal

Automatically construct biomedical ontologies from text:

- Learn concept definitions from text
- Learn terminological knowledge from text
- Evalutation

Example (Terminological Knowledge)

- Genes are not protein complexes, and vice versa.

$$\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$$

- Proteins contain amino acids

$$\mathsf{ProteinDomain} \sqcap \exists \mathsf{hasPart}.\top \sqsubseteq \exists \mathsf{hasPart}.\, \mathsf{AminoAcid}$$

Looking Back

Looking Back

Previous Approaches

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant$ 2 weeks)

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant$ 2 weeks)
  - GCIs produced mostly nonsense

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant$ 2 weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant$ 2 weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant$ 2 weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge
  - "good" quality, measured with precision and recall

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant 2$ weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge
  - "good" quality, measured with precision and recall
  - only restricted form of concept descriptions (at most 2 conjuncts on the left-hand side, of pre-defined form)

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant 2$ weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge
  - "good" quality, measured with precision and recall
  - only restricted form of concept descriptions (at most 2 conjuncts on the left-hand side, of pre-defined form)

Current Goal

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant 2$ weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge
  - "good" quality, measured with precision and recall
  - only restricted form of concept descriptions (at most 2 conjuncts on the left-hand side, of pre-defined form)

Current Goal

- Learn *all* GCIs that are valid in the text corpus

Looking Back

Previous Approaches

Exploit approach of learning SNOMED definitions from text.

- Generate GCIs and check for their occurrence in the text.
  - GCIs from *attribute exploration* of certain basic concept description, with DL reasoner as expert
  - did not finish ($\geqslant 2$ weeks)
  - GCIs produced mostly nonsense
- Compute *implications* in instance-data generated from annotated text
  - obtained terminological knowledge
  - "good" quality, measured with precision and recall
  - only restricted form of concept descriptions (at most 2 conjuncts on the left-hand side, of pre-defined form)

Current Goal

- Learn *all* GCIs that are valid in the text corpus
- Find a way to evaluate them

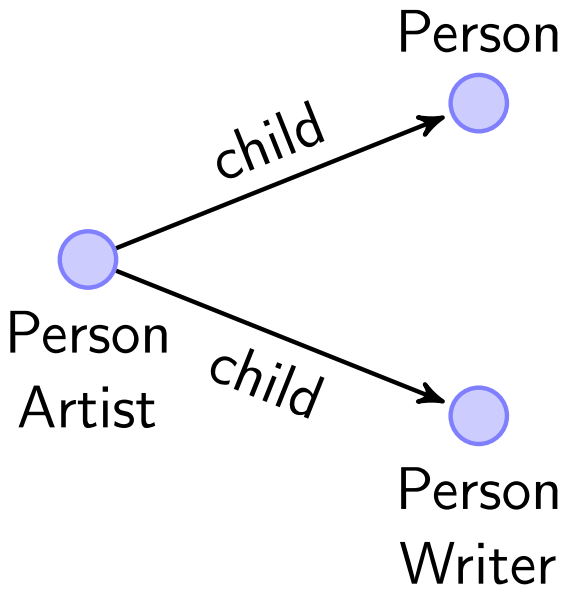
Learning GCIs [Baader and Distel, 2007]

Learning GCIs [Baader and Distel, 2007]

- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*

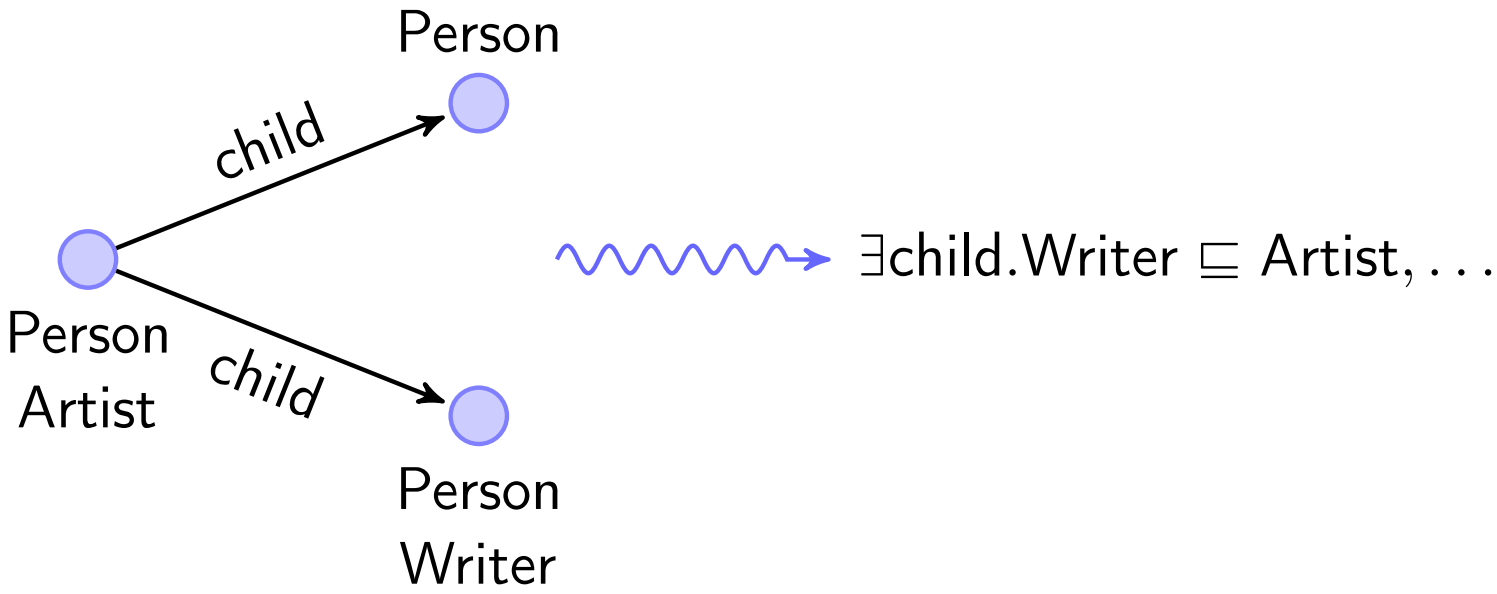
Learning GCIs [Baader and Distel, 2007]

- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



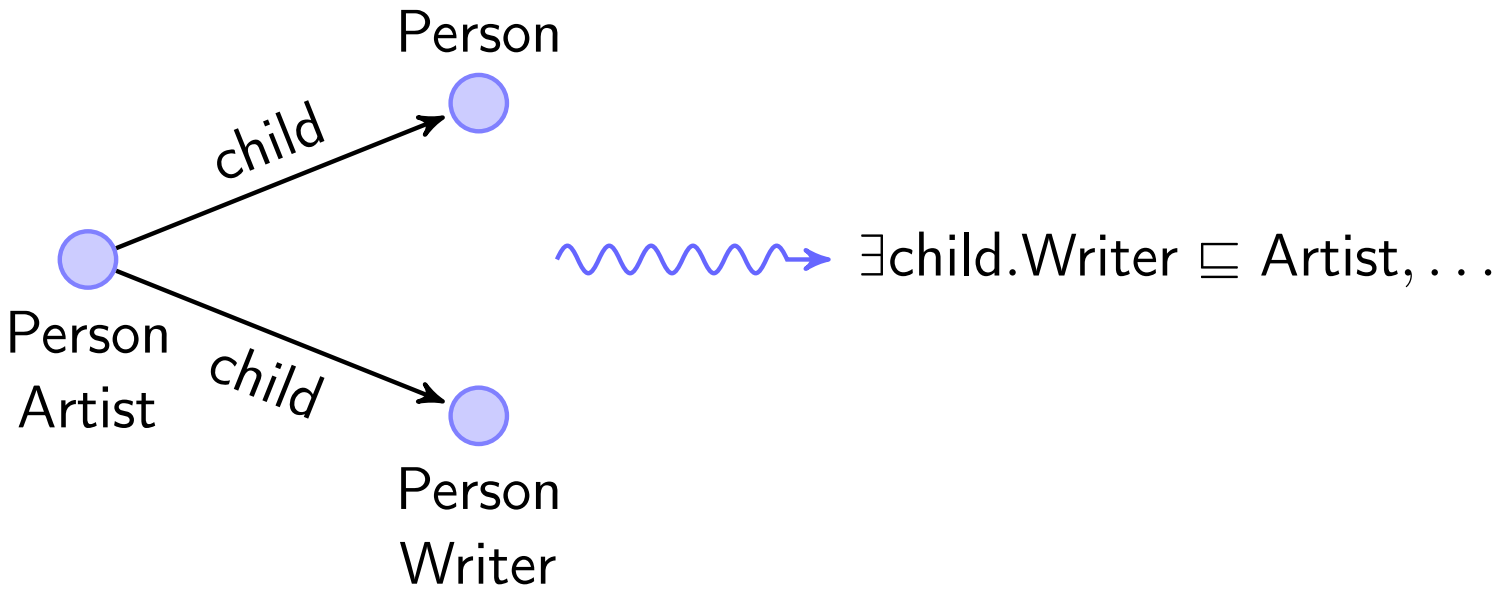
Learning GCIs [Baader and Distel, 2007]

- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



Learning GCIs [Baader and Distel, 2007]

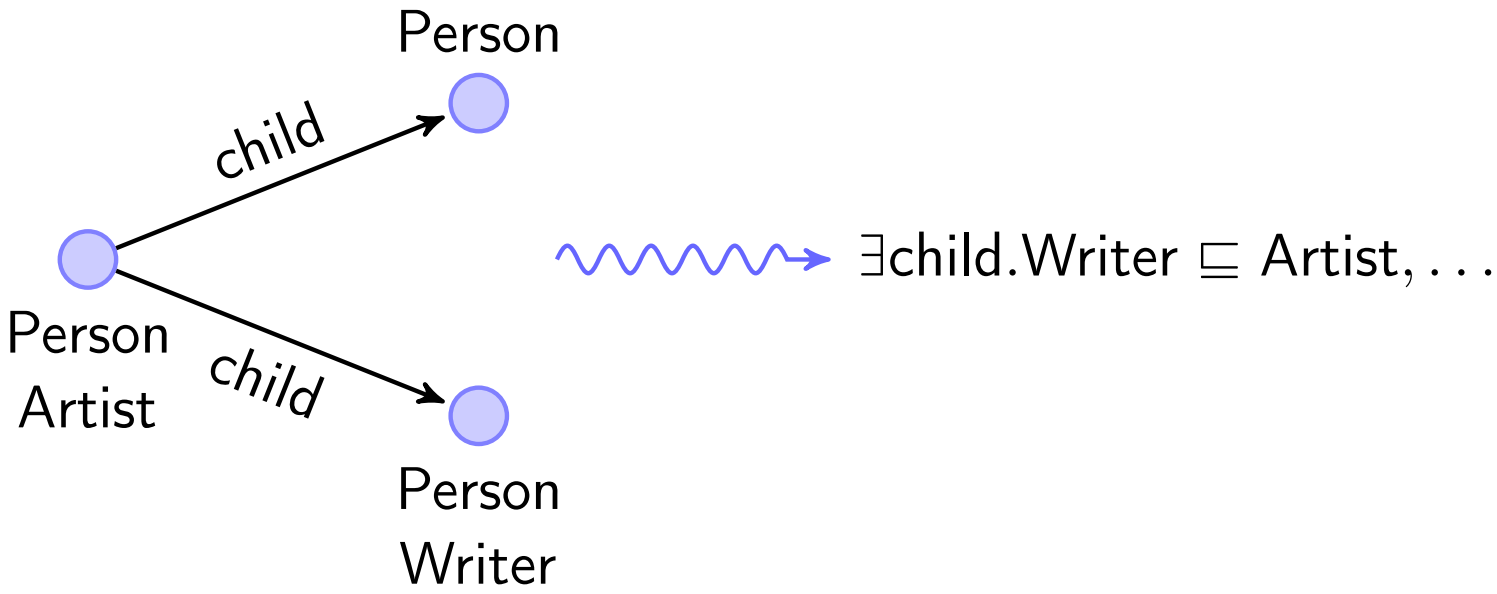
- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



- Computes a *base* of all such GCIs

Learning GCIs [Baader and Distel, 2007]

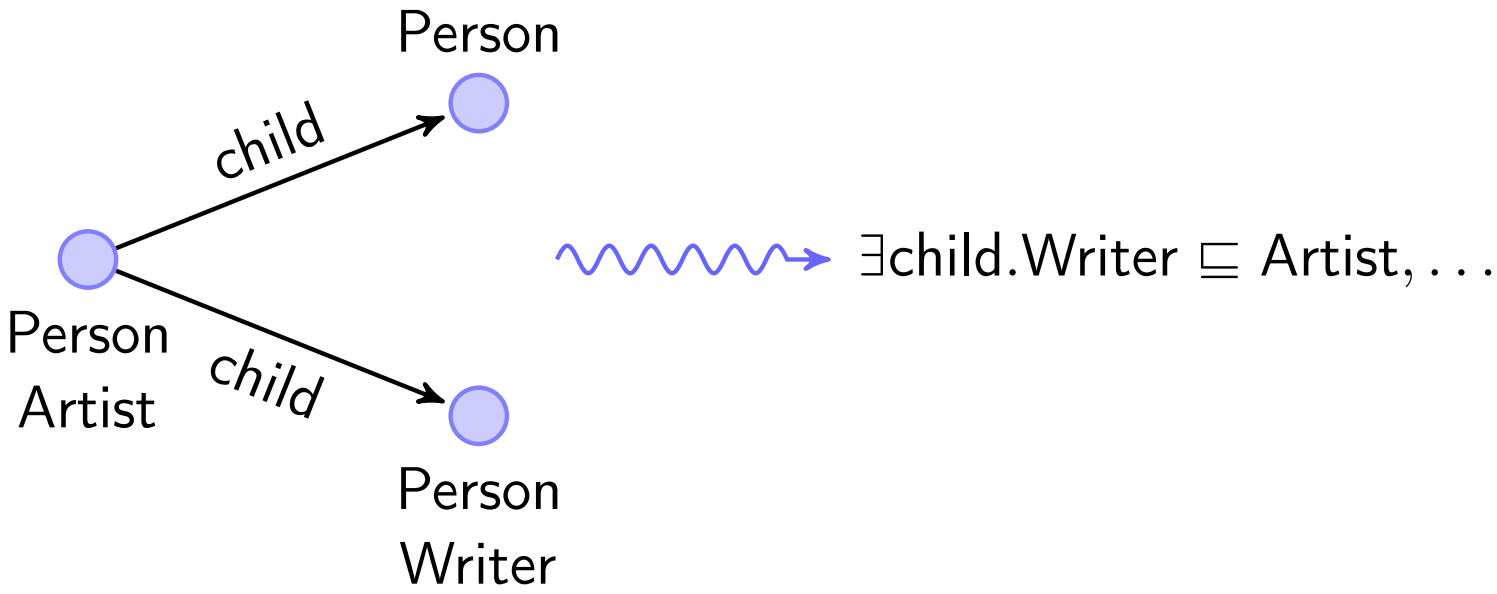
- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



- Computes a *base* of all such GCIs
- Can also compute base of *minimal cardinality*

Learning GCIs [Baader and Distel, 2007]

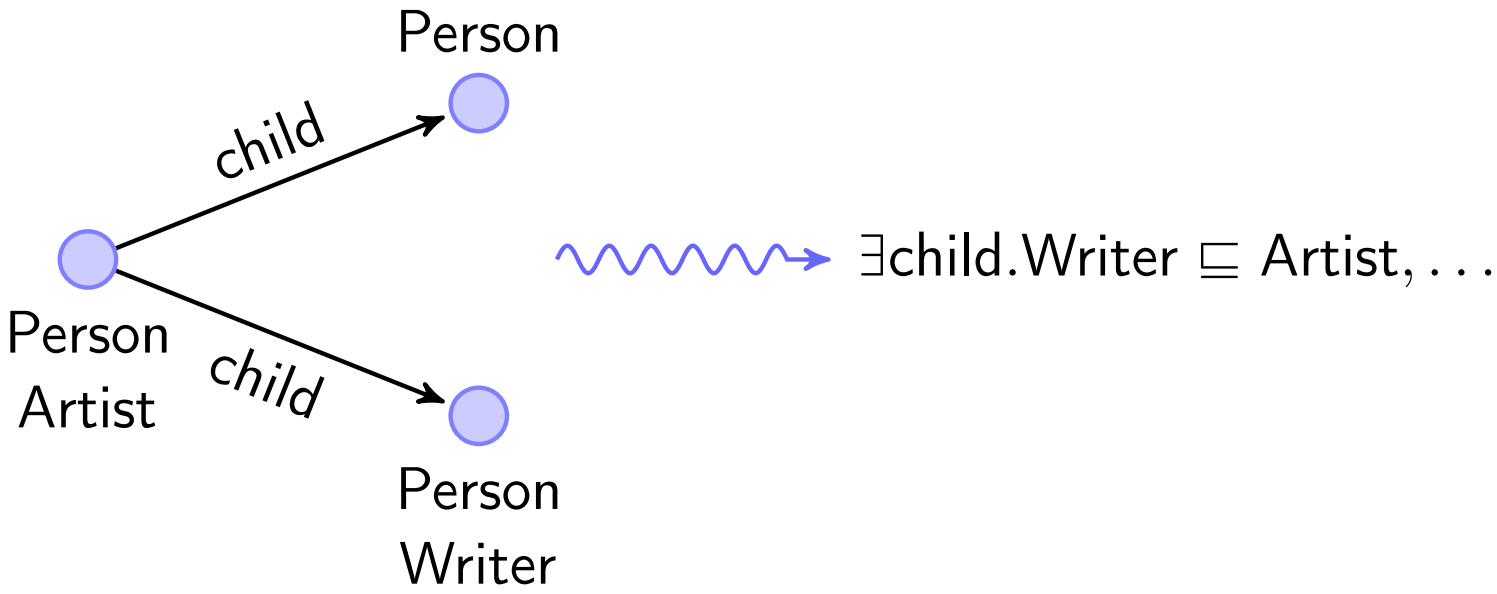
- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



- Computes a *base* of all such GCIs
- Can also compute base of *minimal cardinality*
- Can include *role-depth bounds* [Distel, 2012; Borchmann et.al., 2015]

Learning GCIs [Baader and Distel, 2007]

- Allows to learn *all* valid $\mathcal{EL}$-GCIs from *finite interpretations*



- Computes a *base* of all such GCIs
- Can also compute base of *minimal cardinality*
- Can include *role-depth bounds* [Distel, 2012; Borchmann et.al., 2015]
- Implementations available (prototypes)

Application

Application

Experimental Setup

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)
- How many GCIs cause *inconsistency* or *unsatisfiable classes* in the GRO? (*certainly false positives*)

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)
- How many GCIs cause *inconsistency* or *unsatisfiable classes* in the GRO? (*certainly false positives*)
- How many GCIs of the GRO follow from the GCIs we learned? ("recall")

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)
- How many GCIs cause *inconsistency* or *unsatisfiable classes* in the GRO? (*certainly false positives*)
- How many GCIs of the GRO follow from the GCIs we learned? ("recall")

"Small" Issue

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)
- How many GCIs cause *inconsistency* or *unsatisfiable classes* in the GRO? (*certainly false positives*)
- How many GCIs of the GRO follow from the GCIs we learned? ("recall")

"Small" Issue

- Annotation uses open-world semantics

Application

Experimental Setup

- Take annotated text from the biomedical domain (GRO)
- Turn annotation into relational data
- Learn valid GCIs of a particular role-depth
- Evaluate

Evaluation

- How many GCIs learned follow from the GRO? (*certainly true positives*)
- How many GCIs cause *inconsistency* or *unsatisfiable classes* in the GRO? (*certainly false positives*)
- How many GCIs of the GRO follow from the GCIs we learned? ("recall")

"Small" Issue

- Annotation uses open-world semantics
- Learning uses closed-world semantics

The Data-Set

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)
  - Relations (encodes, locatedIn, fromSpecies, ...)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)
  - Relations (encodes, locatedIn, fromSpecies, ...)

Example (Entities and Events)

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)
  - Relations (encodes, locatedIn, fromSpecies, ...)

Example (Entities and Events)

Activin addition strongly promotes an interaction between these two proteins .

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)
  - Relations (encodes, locatedIn, fromSpecies, ...)

Example (Entities and Events)

Protein Protein

Activin addition strongly promotes an interaction between these two proteins .

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, ...)
  - Events (Mutation, Localization, Experimental Intervention, ...)
  - Relations (encodes, locatedIn, fromSpecies, ...)

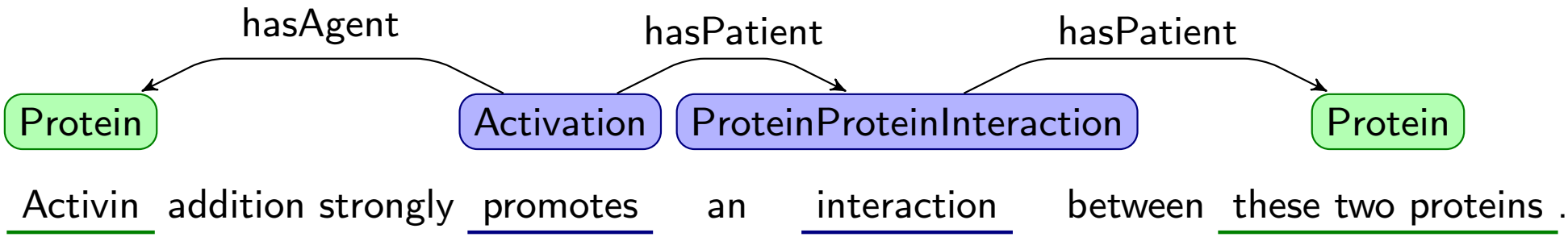
Example (Entities and Events)

Protein		Activation	ProteinProteinInteraction		Protein	
Activin	addition strongly	promotes	an	interaction	between	these two proteins .

The Data-Set

- Gene Regulation Ontology task at BioNLP Shared Task 2013 (`http://2013.bionlp-st.org`)
- 200 manually annotated PubMed abstracts on gene regulation processes
- Annotations from the Gene Regulation Ontology (GRO)
  - Entities (Cell, Protein, Tissue, . . .)
  - Events (Mutation, Localization, Experimental Intervention, . . .)
  - Relations (encodes, locatedIn, fromSpecies, . . .)

Example (Entities and Events)



Evaluation

Evaluation

Experiment

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent
  - has 321 unsatisfiable classes (out of 507)

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent
  - has 321 unsatisfiable classes (out of 507)
- 49 GCIs (each on its own) cause unsatisfiable classes ($\approx 3.2\%$)

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent
  - has 321 unsatisfiable classes (out of 507)
- 49 GCIs (each on its own) cause unsatisfiable classes ($\approx 3.2\%$)
- Removal of 56 GCIs results in no unsatisfiable classes ($\approx 3.6\%$)

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent
  - has 321 unsatisfiable classes (out of 507)
- 49 GCIs (each on its own) cause unsatisfiable classes ($\approx 3.2\%$)
- Removal of 56 GCIs results in no unsatisfiable classes ($\approx 3.6\%$)
- 319 are entailed by the GRO ($\approx 20.6\%$)

Evaluation

Experiment

- considered only 30 most frequent concept-names (reason: performance)
- Resulting interpretation has 7399 elements, 30 concept-names, and 7 role-names
- role-depth bound 1

Results

- 1552 GCIs extracted
  - GRO with these GCIs is still consistent
  - has 321 unsatisfiable classes (out of 507)
- 49 GCIs (each on its own) cause unsatisfiable classes ($\approx 3.2\%$)
- Removal of 56 GCIs results in no unsatisfiable classes ($\approx 3.6\%$)
- 319 are entailed by the GRO ($\approx 20.6\%$)
- Recall not yet available

Certainly correct GCIs

Example

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{hasPart}.\top \sqcap \mathsf{Chromosome} \sqsubseteq \bot$

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{hasPart}.\top \sqcap \mathsf{Chromosome} \sqsubseteq \bot$
- $\exists \mathsf{hasPart}.\top \sqcap \mathsf{Cell} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{CellComponent}$

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{hasPart}.\top \sqcap \mathsf{Chromosome} \sqsubseteq \bot$
- $\exists \mathsf{hasPart}.\top \sqcap \mathsf{Cell} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{CellComponent}$
- $\exists \mathsf{encodes}.\top \sqcap \mathsf{Protein} \sqsubseteq \mathsf{Gene}$

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{hasPart}.\top \sqcap \mathsf{Chromosome} \sqsubseteq \bot$
- $\exists \mathsf{hasPart}.\top \sqcap \mathsf{Cell} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{CellComponent}$
- $\exists \mathsf{encodes}.\top \sqcap \mathsf{Protein} \sqsubseteq \mathsf{Gene}$
- $\exists \mathsf{hasPart}.\top \sqcap \exists \mathsf{locatedIn}.\top \sqcap \mathsf{Gene} \sqcap \mathsf{Protein} \sqsubseteq \exists \mathsf{fromSpecies}.\mathsf{Eukaryote}$

Certainly correct GCIs

Example

- $\mathsf{Gene} \sqcap \mathsf{ProteinComplex} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{hasPart}.\top \sqcap \mathsf{Chromosome} \sqsubseteq \bot$
- $\exists \mathsf{hasPart}.\top \sqcap \mathsf{Cell} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{CellComponent}$
- $\exists \mathsf{encodes}.\top \sqcap \mathsf{Protein} \sqsubseteq \mathsf{Gene}$
- $\exists \mathsf{hasPart}.\top \sqcap \exists \mathsf{locatedIn}.\top \sqcap \mathsf{Gene} \sqcap \mathsf{Protein} \sqsubseteq \exists \mathsf{fromSpecies}.\mathsf{Eukaryote}$
- $\exists \mathsf{encodes}.\top \sqcap \exists \mathsf{fromSpecies}.\mathsf{Eukaryote} \sqcap \exists \mathsf{hasPart}.\mathsf{Peptide} \sqcap \exists \mathsf{hasPart}.\mathsf{ProteinDomain} \sqcap \mathsf{Gene} \sqcap \mathsf{Protein} \sqsubseteq \exists \mathsf{encodes}.\mathsf{MessengerRNA}$

Inconclusive GCIs

Example

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$
- $\mathsf{Cell} \sqcap \mathsf{Virus} \sqsubseteq \bot$

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$
- $\mathsf{Cell} \sqcap \mathsf{Virus} \sqsubseteq \bot$
- $\mathsf{Eukaryote} \sqcap \mathsf{SignalingPathway} \sqsubseteq \bot$

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$
- $\mathsf{Cell} \sqcap \mathsf{Virus} \sqsubseteq \bot$
- $\mathsf{Eukaryote} \sqcap \mathsf{SignalingPathway} \sqsubseteq \bot$

Observation

Two reasons (at least) for inconclusive GCIs

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$
- $\mathsf{Cell} \sqcap \mathsf{Virus} \sqsubseteq \bot$
- $\mathsf{Eukaryote} \sqcap \mathsf{SignalingPathway} \sqsubseteq \bot$

Observation

Two reasons (at least) for inconclusive GCIs

- simply wrong

Inconclusive GCIs

Example

- $\mathsf{Cell} \sqcap \mathsf{Eukaryote} \sqsubseteq \bot$
- $\exists \mathsf{encodes}.\mathsf{Eukaryote} \sqsubseteq \bot$
- $\mathsf{Cell} \sqcap \mathsf{Virus} \sqsubseteq \bot$
- $\mathsf{Eukaryote} \sqcap \mathsf{SignalingPathway} \sqsubseteq \bot$

Observation

Two reasons (at least) for inconclusive GCIs

- simply wrong
- GRO incomplete

Unsatisfiable Classes

Unsatisfiable Classes

Question: Where do they come from?

Unsatisfiable Classes

Question: Where do they come from?

Example

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\mathsf{CellComponent} \sqcap \mathsf{Nucleus} \sqsubseteq \bot$$

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\mathsf{CellComponent} \sqcap \mathsf{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\text{CellComponent} \sqcap \text{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus
- In the GRO, CellComponent is a super-class of Nucleus

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\mathsf{CellComponent} \sqcap \mathsf{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus
- In the GRO, CellComponent is a super-class of Nucleus
- So, the annotation is *incomplete*

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\mathsf{CellComponent} \sqcap \mathsf{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus
- In the GRO, CellComponent is a super-class of Nucleus
- So, the annotation is *incomplete*

Conclusion

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\mathsf{CellComponent} \sqcap \mathsf{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus
- In the GRO, CellComponent is a super-class of Nucleus
- So, the annotation is *incomplete*

Conclusion

- unsatisfiable classes can arise through the *closed-world* interpretation of the *open-world* data-set.

Unsatisfiable Classes

Question: Where do they come from?

Example

$$\text{CellComponent} \sqcap \text{Nucleus} \sqsubseteq \bot$$

- Data-set did not contain any occurrence of an individual that is both CellComponent and Nucleus
- In the GRO, CellComponent is a super-class of Nucleus
- So, the annotation is *incomplete*

Conclusion

- unsatisfiable classes can arise through the *closed-world* interpretation of the *open-world* data-set.
- *all* disjointness axioms containing only concept-names are caused by this

Unsatisfiable Classes

Example

$$\exists \text{locatedIn}.\text{Cell} \sqcap \exists \text{locatedIn}.\text{Nucleus} \sqsubseteq \text{Protein}$$

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \exists \text{locatedIn.Nucleus}$$

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \exists \text{locatedIn.Nucleus}$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \text{ProteinTargeting} \sqsubseteq \exists \text{locatedIn.Cell}$$

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \exists \text{locatedIn.Nucleus}$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \text{ProteinTargeting} \sqsubseteq \exists \text{locatedIn.Cell}$$

- But data-set does not contain any reference to NuclearExportOfmRNA

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \exists \text{locatedIn.Nucleus}$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \text{ProteinTargeting} \sqsubseteq \exists \text{locatedIn.Cell}$$

- But data-set does not contain any reference to NuclearExportOfmRNA
- Approach could not learn this counterexample

Unsatisfiable Classes

Example

$$\exists \text{locatedIn.Cell} \sqcap \exists \text{locatedIn.Nucleus} \sqsubseteq \text{Protein}$$

- Causes the class NuclearExportOfmRNA to become unsatisfiable
- GRO entails

$$\text{NuclearExportOfmRNA} \sqcap \text{Protein} \sqsubseteq \bot$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \exists \text{locatedIn.Nucleus}$$
$$\text{NuclearExportOfmRNA} \sqsubseteq \text{ProteinTargeting} \sqsubseteq \exists \text{locatedIn.Cell}$$

- But data-set does not contain any reference to NuclearExportOfmRNA
- Approach could not learn this counterexample

Idea

Remove concept-names not occurring in the data-set before evaluation?

Further Results

Further Results

- Role depth $\leqslant 1$, top-50 concept-names

Further Results

- Role depth $\leqslant 1$, top-50 concept-names
  - 3101 GCIs extracted
  - consistent
  - remove 130 GCIs to obtain no unsatisfiable classes ($\approx 4.2\%$)
  - 821 entailed by the GRO ($\approx 26.5\%$)

Further Results

- Role depth $\leqslant 1$, top-50 concept-names
  - 3101 GCIs extracted
  - consistent
  - remove 130 GCIs to obtain no unsatisfiable classes ($\approx 4.2\%$)
  - 821 entailed by the GRO ($\approx 26.5\%$)
- Role depth $\leqslant 2$, top-5 concept-names

Further Results

- Role depth $\leqslant 1$, top-50 concept-names
  - 3101 GCIs extracted
  - consistent
  - remove 130 GCIs to obtain no unsatisfiable classes ($\approx 4.2\%$)
  - 821 entailed by the GRO ($\approx 26.5\%$)
- Role depth $\leqslant 2$, top-5 concept-names
  - 473 GCIs extracted
  - consistent
  - removing 20 GCI to obtain no unsatisfiable classes ($\approx 4.2\%$)
  - 39 entailed ($\approx 8.2\%$)

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

$$\mathsf{Protein} \sqcap \exists \mathsf{fromSpecies}.\top \sqcap \mathsf{Gene} \sqsubseteq \exists \mathsf{fromSpecies}.\mathsf{Eukaryote}$$

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

$$\text{Protein} \sqcap \exists \text{fromSpecies}.\top \sqcap \text{Gene} \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$$

  - 16 inconclusive

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

$$\text{Protein} \sqcap \exists \text{fromSpecies}.\top \sqcap \text{Gene} \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$$

  - 16 inconclusive
    - $\exists \text{encodes}.\text{MessengerRNA} \sqsubseteq \text{Gene}$

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

$$\text{Protein} \sqcap \exists \text{fromSpecies}.\top \sqcap \text{Gene} \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$$

  - 16 inconclusive
    - $\exists \text{encodes}.\text{MessengerRNA} \sqsubseteq \text{Gene}$
    - $\exists \text{fromSpecies}.\top \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$

Using Confidence

Idea

Consider also GCIs which are correct in a "large number" of cases.

Experiment

- Role depth $\leqslant 1$, top-30 concept-names, confidence $\geqslant 0.9, \neq 1$
  - 18 GCIs extracted
  - consistent with GRO, no unsatisfiable classes
  - 2 (1) entailed by the GRO:

$$\text{Protein} \sqcap \exists \text{fromSpecies}.\top \sqcap \text{Gene} \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$$

  - 16 inconclusive
    - $\exists \text{encodes}.\text{MessengerRNA} \sqsubseteq \text{Gene}$
    - $\exists \text{fromSpecies}.\top \sqsubseteq \exists \text{fromSpecies}.\text{Eukaryote}$
    - $\exists \text{fromSpecies}.\text{Eukaryote} \sqcap \exists \text{hasPart}.\text{AminoAcid} \sqsubseteq \text{Protein}$

Summing Up

Summing Up

What has been done?

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned
- Open World Assumption vs. Closed World Assumption (inherent?)

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned
- Open World Assumption vs. Closed World Assumption (inherent?)

What's next?

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned
- Open World Assumption vs. Closed World Assumption (inherent?)

What's next?

- Compute recall

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned
- Open World Assumption vs. Closed World Assumption (inherent?)

What's next?

- Compute recall
- Clean up GRO from unknown concept-names before evaluation

Summing Up

What has been done?

- Discussed approach by Baader and Distel to learn GCIs from relational data
- Applied this approach to annotated text from the biomedical domain
- Conducted some first experiments to evaluate the results

Issues

- Hard to evaluate GCIs that have been learned
- Open World Assumption vs. Closed World Assumption (inherent?)

What's next?

- Compute recall
- Clean up GRO from unknown concept-names before evaluation
- Devise evaluation that is "independent" from the data-set?