Semantic similarity and machine learning with ontologies

Ontologies: axioms, not graphs!

Overview | Browse | DLQuery | Download

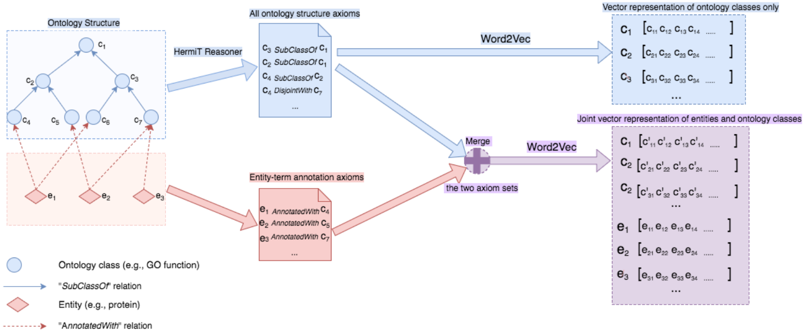
Annotation	Value
label	B cell apoptotic process
definition	Any apoptotic process in a B cell, a lymphocyte of B lineage with the phenotype CD19-positive and capable of B cell mediated immunity.
class	http://purl.obolibrary.org/obo/GO_0001783
ontology	GO-PLUS
Equivalent	apoptotic process and (occurs in some B cell)
SubClassOf	occurs in some B cell, lymphocyte apoptotic process
id	GO:0001783
has_obo_namespace	biological_process

Ontologies: axioms, not graphs!

Gene Ontology:

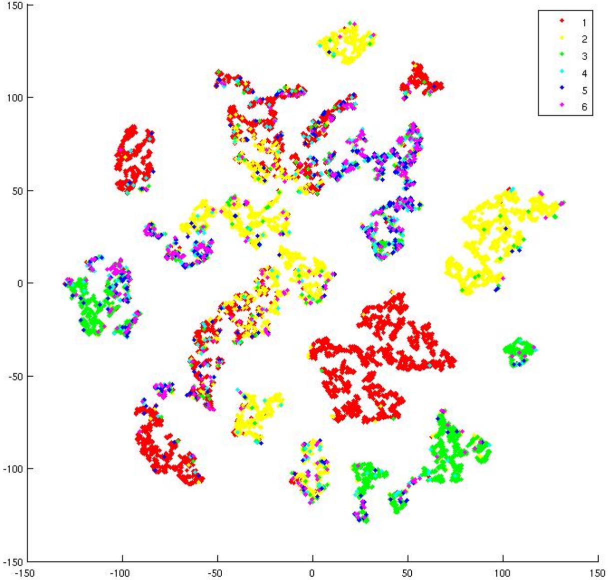
- `behavior DisjointWith: 'developmental process'`
- `behavior SubclassOf: only-in-taxon some metazoa`
- `'cell proliferation' DisjointWith: in-taxon some fungi`
- `'cell growth' EquivalentTo: growth and ('results in growth of' some cell)`
- ...

Onto2Vec



Ontology Structure
HermiT Reasoner
All ontology structure axioms
c3 SubClassOf c1
c2 SubClassOf c1
c4 SubClassOf c2
c4 DisjointWith c7
Word2Vec
Vector representation of ontology classes only
Merge
the two axiom sets
Word2Vec
Joint vector representation of entities and ontology classes
Entity-term annotation axioms
e1 AnnotatedWith c4
e2 AnnotatedWith c5
e3 AnnotatedWith c7
Ontology class (e.g., GO function)
"SubClassOf" relation
Entity (e.g., protein)
"AnnotatedWith" relation

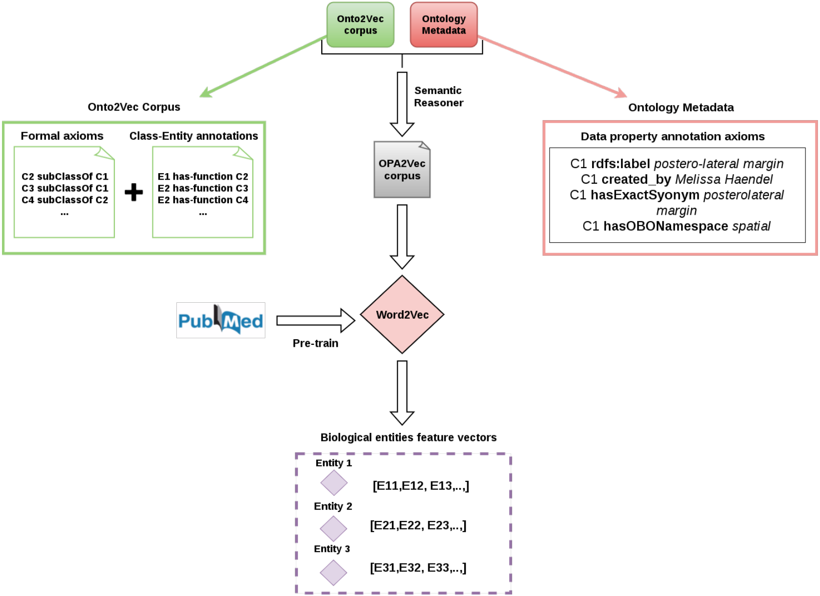
Visualizing embeddings



Combination with text

- ontologies contain more than axioms:
  - labels, synonyms, definitions, authors, etc.
- Description Logic axioms != natural language
- transfer learning: learn on one domain/task, apply to another
  - e.g.: learn on literature, apply to ontologies
  - words have "meaning" in literature, Description Logic symbols have "meaning" in ontology axioms
- Ontologies Plus Annotations 2 Vec (OPA2Vec) combines both

Ontologies Plus Annotations 2 Vec



Onto2Vec and OPA2Vec

- `https://github.com/bio-ontology-research-group/mowl`
- python library
  - input: OWL ontology, set of entities with annotations/associations
  - output: vectors for each class and entity
- Elk reasoner
- limitations: word-based
  - completely ignores any semantics!

How to overcome the semantic gap?

- none of the models discussed above are truly "semantic"
  - all syntactic
  - graph-based or based on axioms

How to overcome the semantic gap?

- none of the models discussed above are truly "semantic"
  - all syntactic
  - graph-based or based on axioms
- what do we actually mean by "semantics"?

How to overcome the semantic gap?

- none of the models discussed above are truly "semantic"
  - all syntactic
  - graph-based or based on axioms
- what do we actually mean by "semantics"?
  - formal definition of "truth" relies on "models"

How to overcome the semantic gap?

- none of the models discussed above are truly "semantic"
  - all syntactic
  - graph-based or based on axioms
- what do we actually mean by "semantics"?
  - formal definition of "truth" relies on "models"
  - universal algebra over formal languages (with signature Σ)

Description Logic EL++

Name	Syntax	Semantics
top	$\top$	$\Delta^{\mathcal{I}}$
bottom	$\bot$	$\emptyset$
nominal	$\{a\}$	$\{a^{\mathcal{I}}\}$
conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
existential restriction	$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}} : (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
generalized concept inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
role inclusion	$r_1 \circ \ldots \circ r_n \sqsubseteq r$	$r_1^{\mathcal{I}} \circ \ldots \circ r_n^{\mathcal{I}} \subseteq r^{\mathcal{I}}$

Models

- Interpretations and Σ-structures
- Model $\mathfrak{A}$ of a formula ϕ: ϕ is true in $\mathfrak{A}$ ($\mathfrak{A} \models \phi$)
- Theory T: set of formulas
- $\mathfrak{A}$ is a model of T if $\mathfrak{A}$ is a model of all formulas in T
- Ontologies are (special kinds of) theories

EL Embeddings

- given a theory/ontology T with signature $\Sigma(T)$
- aim: find $f_e : \Sigma(T) \mapsto \mathbb{R}^n$ s.t. $f_e(\Sigma(T))$ is a model of T $(f_e(\Sigma(T)) \models T)$

EL Embeddings

- given a theory/ontology T with signature $\Sigma(T)$
- aim: find $f_e : \Sigma(T) \mapsto \mathbb{R}^n$ s.t. $f_e(\Sigma(T))$ is a model of T ($f_e(\Sigma(T)) \models T$)
- more general: find an algorithm that maps symbols (signatures) into $\mathbb{R}^n$ so that the *semantics* of the symbol (expressed through axioms and explicit in model structures) is preserved
  - or: the embedding function *is* an interpretation function

Key idea

- for all $r \in \Sigma(T)$ and $C \in \Sigma(T)$, define $f_e(r)$ and $f_e(C)$
- $f_e(C)$ maps to points in an open n-ball such that $f_e(C) = C^{\mathcal{I}}$: $C^{\mathcal{I}} = \{x \in \mathbb{R}^n | \, \|f_e(C) - x\| < r_e(C)\}$
  - these are the *extension* of a class in $\mathbb{R}^n$
- $f_e(r)$ maps a binary relation r to a vector such that $r^{\mathcal{I}} = \{(x, y) | x + f_e(r) = y\}$
  - that's the TransE property for *individuals*
- use the axioms in T as constraints

Algorithm

- normalize the theory:
  - every $\mathcal{EL}^{++}$theory can be expressed using four normal forms (Baader et al., 2005)
- eliminate the ABox: replace each individual symbol with a singleton class: a becomes $\{a\}$
- rewrite relation assertions $r(a, b)$ and class assertions $C(a)$ as $\{a\} \sqsubseteq \exists r.\{b\}$ and $\{a\} \sqsubseteq C$
- normalization rules to generate:
  - $C \sqsubseteq D$
  - $C \sqcap D \sqsubseteq E$
  - $C \sqsubseteq \exists R.D$
  - $\exists R.C \sqsubseteq D$

Algorithm: loss functions

$$
\begin{aligned}
&loss_{C \sqsubseteq D}(c, d) = \\
&\max(0, \|f_\eta(c) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma) \\
&+ |\,\|f_\eta(c)\| - 1| + |\,\|f_\eta(d)\| - 1|
\end{aligned}
\tag{1}
$$

Algorithm: loss functions

$$loss_{C \sqsubseteq \exists R.D}(c, d, r) = \max(0, \|f_\eta(c) + f_\eta(r) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma) + |\,\|f_\eta(c)\| - 1| + |\,\|f_\eta(d)\| - 1| \tag{2}$$

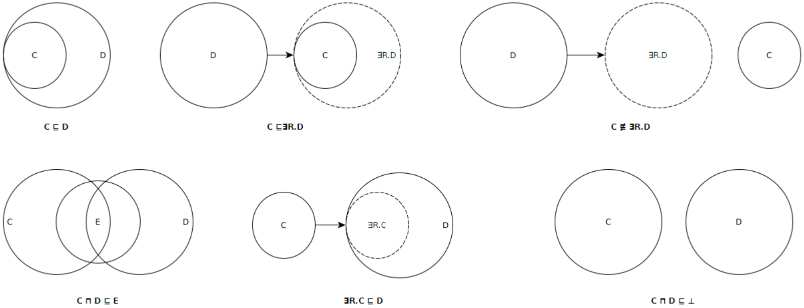
Algorithm: loss functions

$$
\begin{aligned}
&loss_{\exists R.C \sqsubseteq D}(c, d, r) = \\
&\max(0, \|f_\eta(c) - f_\eta(r) - f_\eta(d)\| - r_\eta(c) - r_\eta(d) - \gamma) \qquad (3)\\
&+ |\, \|f_\eta(c)\| - 1| + |\, \|f_\eta(d)\| - 1|
\end{aligned}
$$

Algorithm: loss functions

$$
\begin{aligned}
&loss_{C \sqcap D \sqsubseteq \bot}(c, d, e) = \\
&\max(0, r_\eta(c) + r_\eta(d) - \|f_\eta(c) - f_\eta(d)\| + \gamma) \\
&+ |\,\|f_\eta(c)\| - 1| + |\,\|f_\eta(d)\| - 1|
\end{aligned} \tag{4}
$$

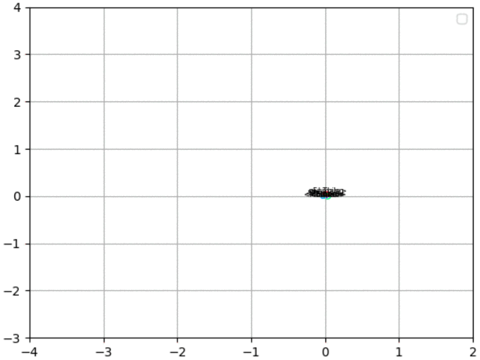
Algorithm: loss functions



EL Embeddings

$$\begin{aligned}
Male &\sqsubseteq Person && (5)\\
Female &\sqsubseteq Person && (6)\\
Father &\sqsubseteq Male && (7)\\
Mother &\sqsubseteq Female && (8)\\
Father &\sqsubseteq Parent && (9)\\
Mother &\sqsubseteq Parent && (10)\\
Female \sqcap Male &\sqsubseteq \bot && (11)\\
Female \sqcap Parent &\sqsubseteq Mother && (12)\\
Male \sqcap Parent &\sqsubseteq Father && (13)\\
\exists hasChild.Person &\sqsubseteq Parent && (14)\\
Parent &\sqsubseteq Person && (15)\\
Parent &\sqsubseteq \exists hasChild.\top && (16)
\end{aligned}$$

EL Embeddings



- model with $\Delta = R^n$
- support quantifiers, negation, conjunction,...