Centre for
Data Analytics

Insight

Putting ontologies to work in NLP

The lemon model and its future

John P. McCrae — National University of Ireland, Galway

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN SFI Science Foundation Ireland

Introduction

- In natural language processing we are doing three main things
  - Understanding natural language
  - Generating natural language
  - Tranformation (translation, summarization)
- These can be typed as:
  - NL $\rightarrow$ Representation
  - Representation $\rightarrow$ NL
  - NL $\rightarrow$ NL

Representation

- We can think of representations as falling into two large classes
  1. Symbolic representations
  2. Numeric representations
- For example: “John sent her a text”
  1. $\mathsf{sent}(\mathsf{John}, x, m, \mathsf{SMS}) \sqcap \mathsf{female}(x) \sqcap \mathsf{Message}(m)$
  2. $(0.664, 0.059, 0.557, 0.906, 0.031)^{\mathsf{T}}$

Symbolic versus numeric representations

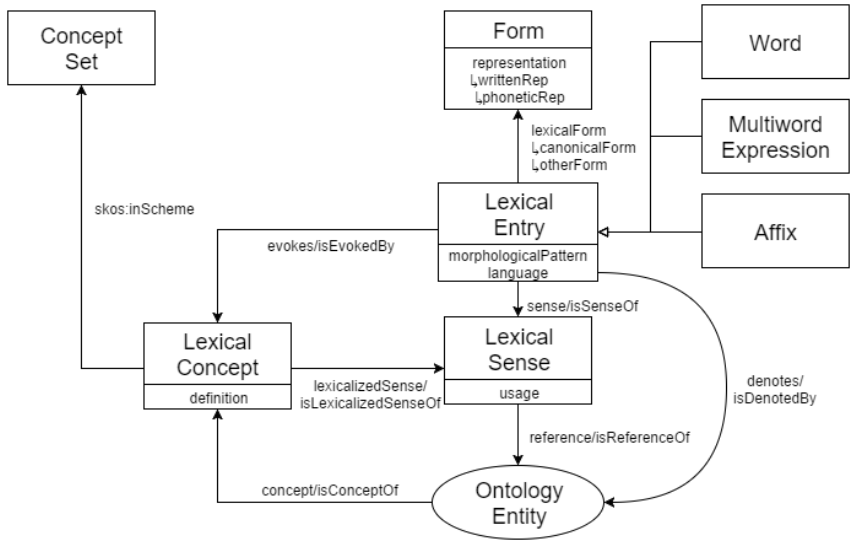
- Numeric representations are:
  - Easy-to-learn from plain text
  - Robust
  - General
- Symbolic representations are:
  - Easier to understand
  - Can make complex inferences
  - Fine-grained

What is an ontology?

- A natural language has a lexicon:
  - A set of words
  - That are combined with rules (syntax)
- A symbolic representation has an ontology:
  - An set of symbols
  - That are combined with rules (logic)
- What is the ontology of a numeric representation?

Ontology-Lexica

- An ontology-lexicon is a model that is both an ontology and a lexicon
- Since 2009 we have been developing *lemon* — The Lexicon Model for Ontologies
- Now (May 2016!) released by the W3C Ontology Lexicon Community Group as a W3C Vocabulary
- `https://www.w3.org/2016/05/ontolex/`



Resources for ontology-lexica

Existing resources

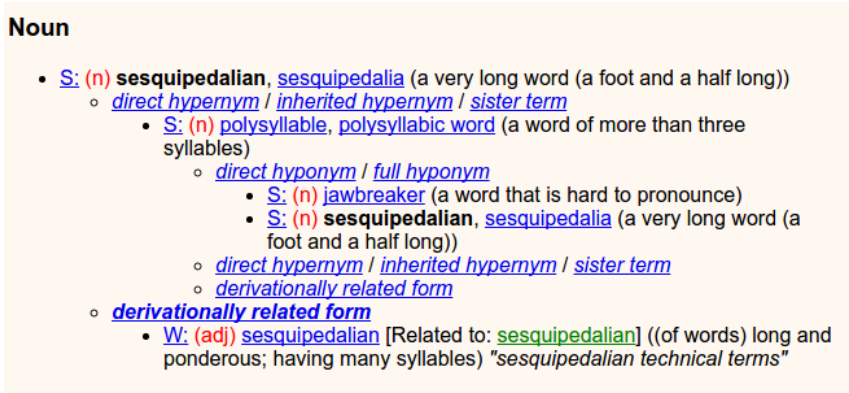
Lexicon: Princeton WordNet

Semantic Network: DBpedia

Ontology: SUMO

Is WordNet an Ontology?

- Provides symbols
- Supports inference, e.g., inverse/symmetric properties
- No frame semantics:
  - WordNet can say "Canberra is a captial city"
  - Cannot say "Canberra is the capital of Australia"
- Words are defined primarily by text



Noun

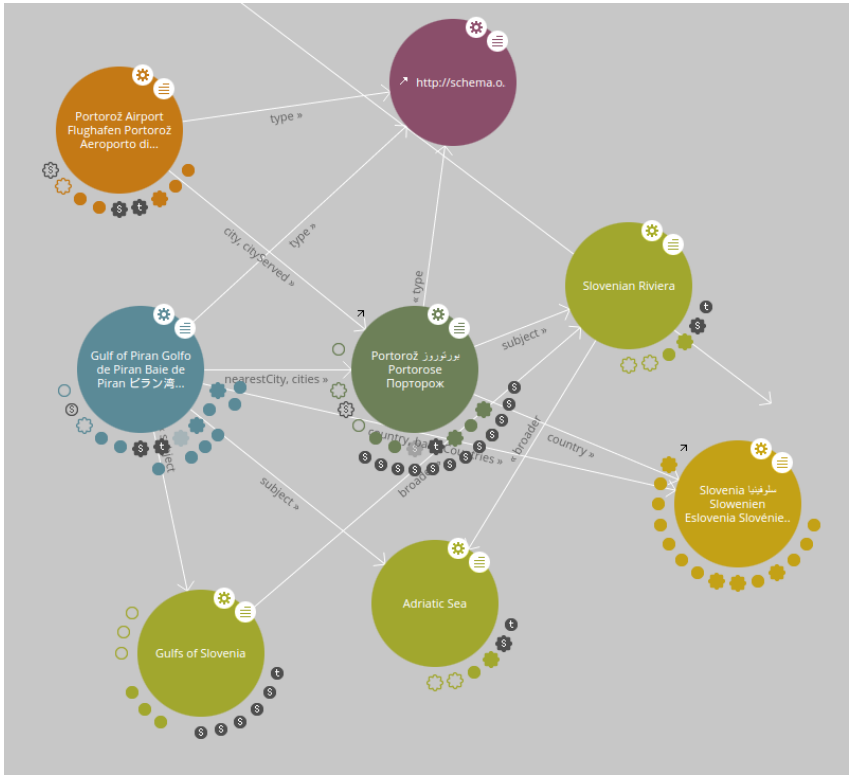
- S: (n) **sesquipedalian**, sesquipedalia (a very long word (a foot and a half long))
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) polysyllable, polysyllabic word (a word of more than three syllables)
      - *direct hyponym* / *full hyponym*
        - S: (n) jawbreaker (a word that is hard to pronounce)
        - S: (n) **sesquipedalian**, sesquipedalia (a very long word (a foot and a half long))
      - *direct hypernym* / *inherited hypernym* / *sister term*
      - *derivationally related form*
  - ***derivationally related form***
    - W: (adj) sesquipedalian [Related to: sesquipedalian] ((of words) long and ponderous; having many syllables) *"sesquipedalian technical terms"*

WordNet and Word Sense Disambiguation

- The sequence of annotations is a formal representation
  - `Canberra[i83245] is a capital_city[i82619]`
- WordNet alone has proven useful for word sense disambiguation (Personalized PageRank - Agirre and Soroa, 2009)
- Produces good performance about 50-70%

DBpedia

- Derived from Wikipedia, so very large
- Has an “ontology” in OWL
- DBpedia can say:
  - “Canberra is a capital city”
  - “Canberra is the capital of Australia”
  - “Canberra is the second largest city in Australia”



SUMO

- Suggested Upper Merged Ontology
- Based on KIF language
- Has definitions in terms of results and consequents

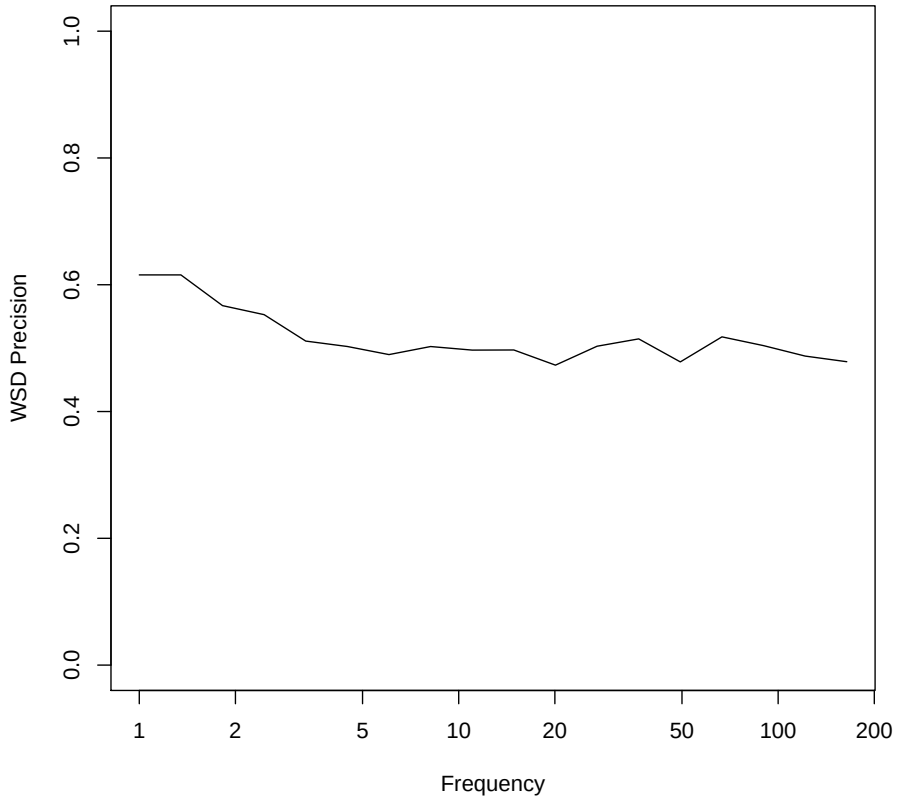
```
(subclass EuropeanNation
          Nation)

(=>
  (instance ?N EuropeanNation)
    (part ?N Europe))
```

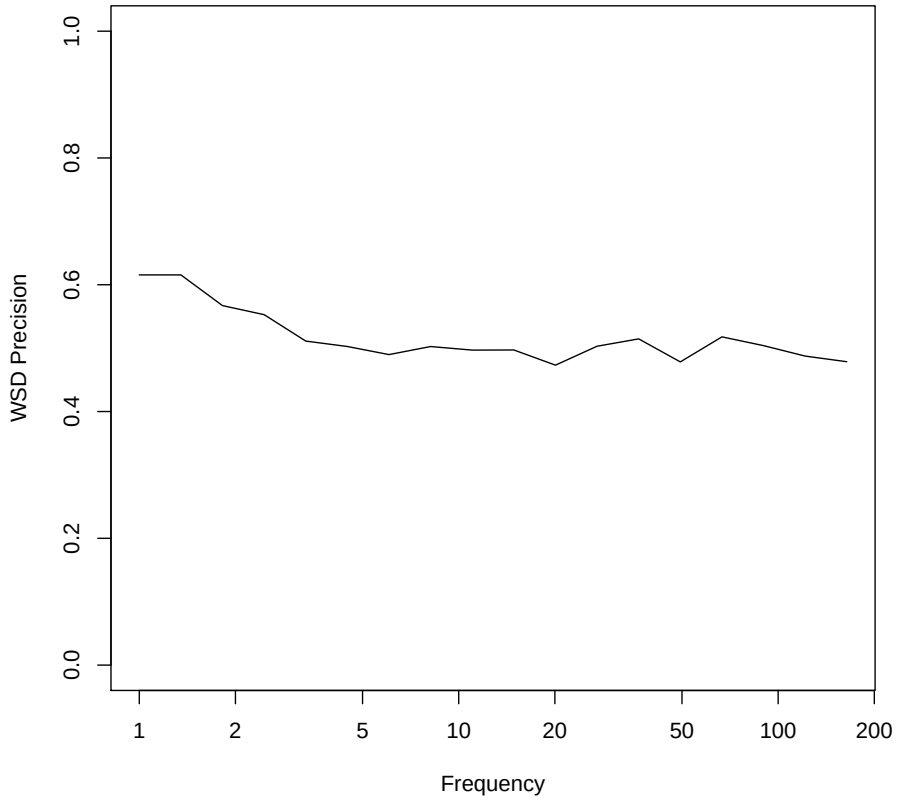
Comparison of these resources

Ontology	Symbols	Links	Ave. Degree
Princeton WordNet	117,791	285,668	2.43
DBpedia-OWL	3,955	4,154	1.05
DBpedia (Infobox EN)	2,866,327	18,328,273	6.39
SUMO	c.25,000	c.80,000	3.2

Does number of symbols matter?

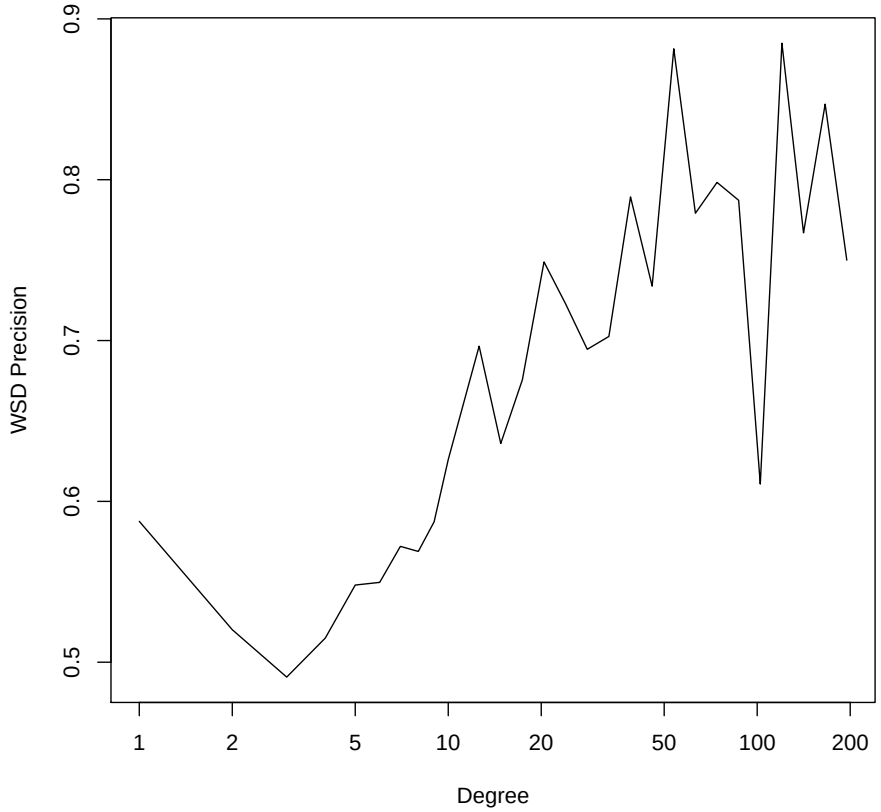


Does number of symbols matter?

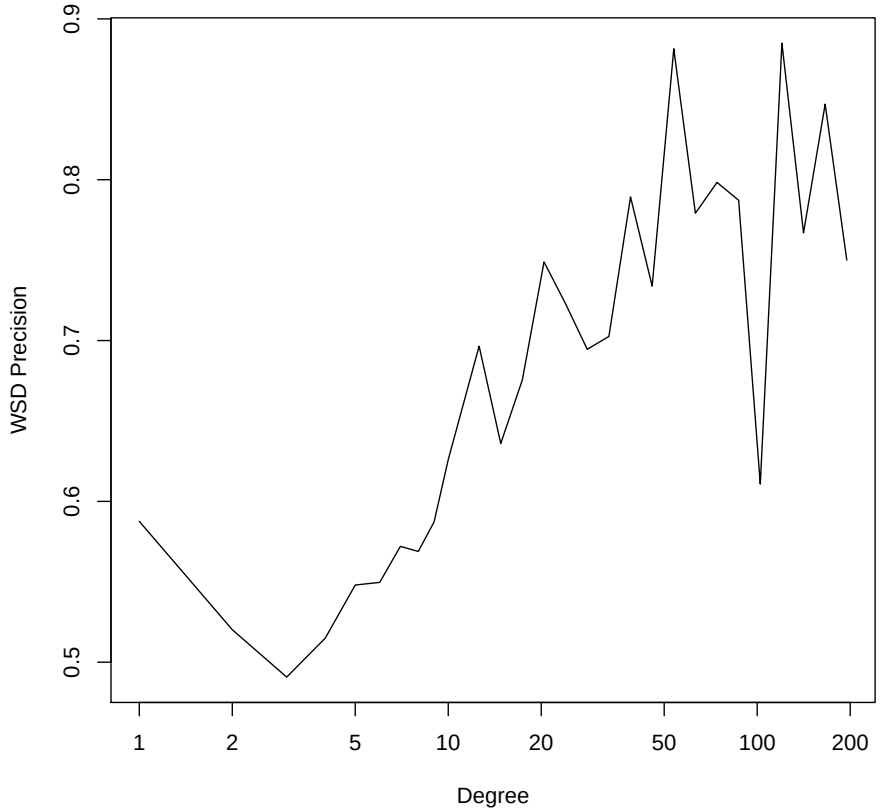


Yes, but exponentially less.

Does degree matter?



Does degree matter?



Yes, quite a lot!

Is one of these resources the best?

- DBpedia is the biggest and densest
- Many basic concepts are missing, e.g., beautiful
- Other collaborative resources (Wiktionary) are of lower density with structural issues
- Combining resources is another approach, e.g., BabelNet, UBY, etc.

The Lexical Gap

The Lexical Gap

- The primary issues with applying ontologies is the lexical gap:
  1. We don’t know all the ways to express the concept in languages
  2. We cannot easily map linguistic structures to formal expressions
  3. These concepts are often insufficiently defined

Lexical Gap 1: Synonym discovery

- Most approaches are based on textual similarity
- Recent models, such as `word2vec` are showing strong performance on term similarity
- Maybe solved soon?

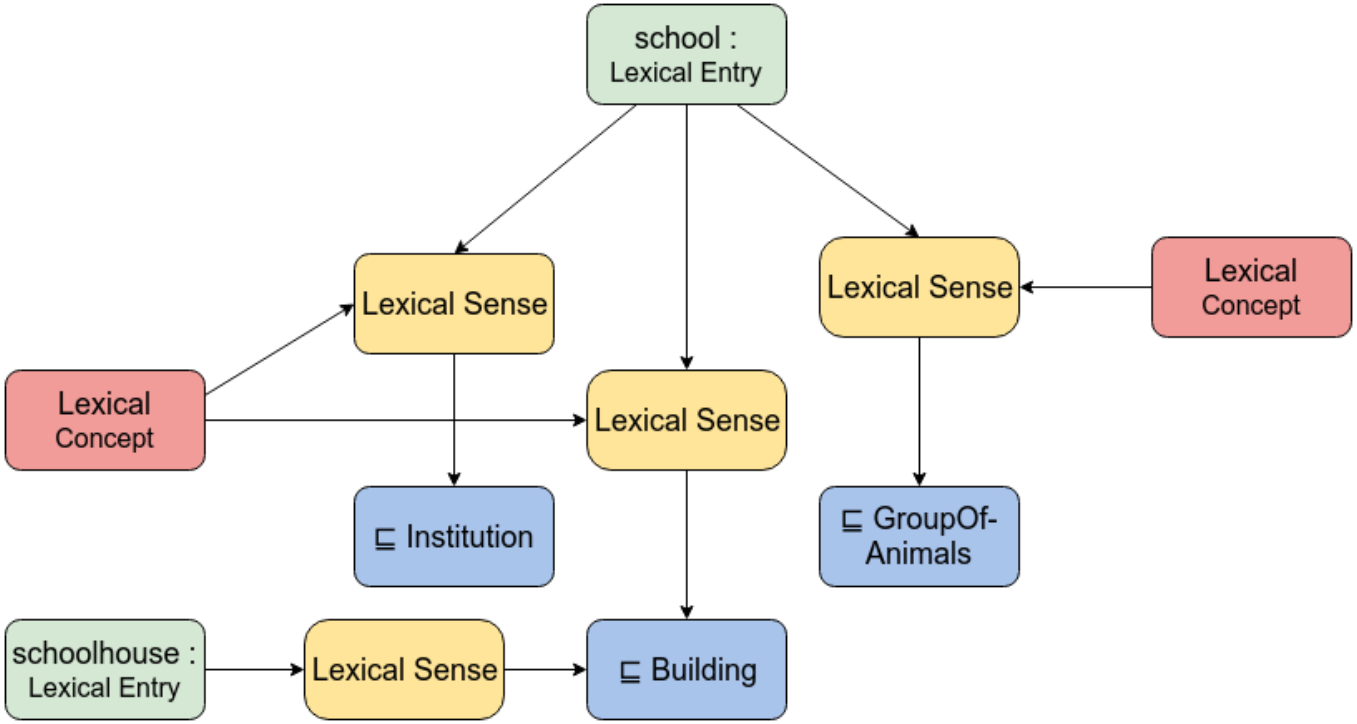
Lexical Gap 2: Mapping

- Word meaning is not exact
- Arguments
- Lexical semantics is not always computable

Systematic polysemy

- “I went to the school”
- “He painted the school”
- “The school announced major changes”

Linguistic Mapping



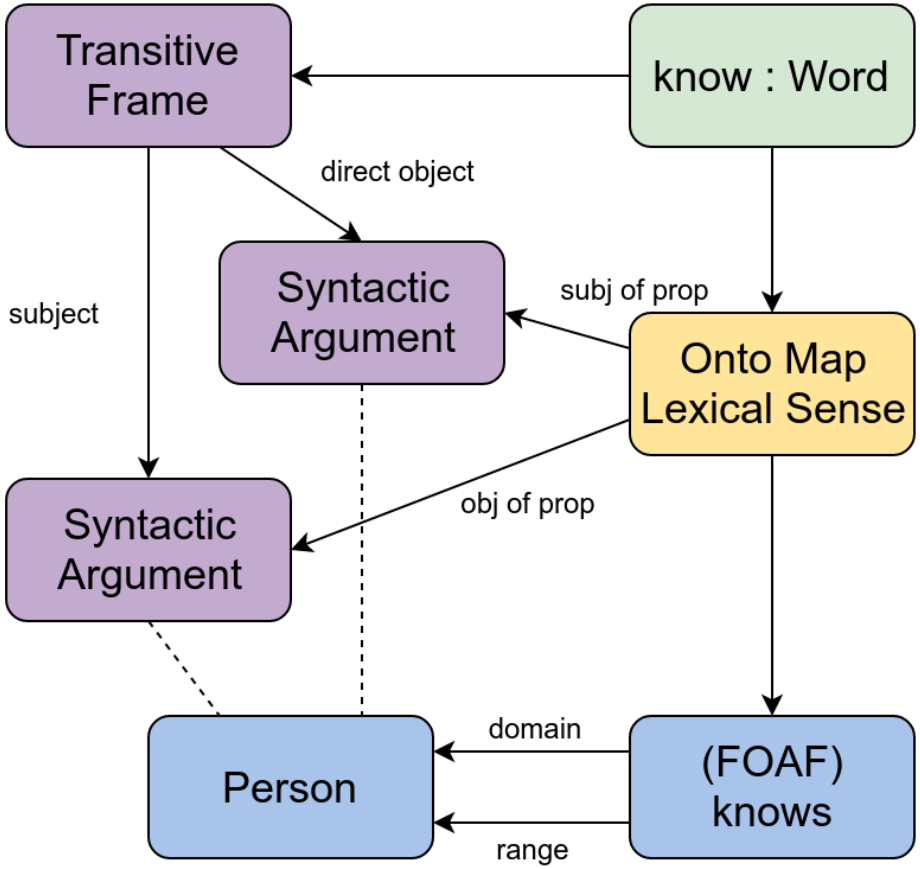
Frames and Correspondence

- The verb “know” is meaningless by itself
  - “John knows Fahad”
- Similarly `foaf:knows` is only used in a triple
  - `insight:jmccrae foaf:knows cnr:fkhan`
- It is necessary to state how these corresponds

Frames and Correspondence

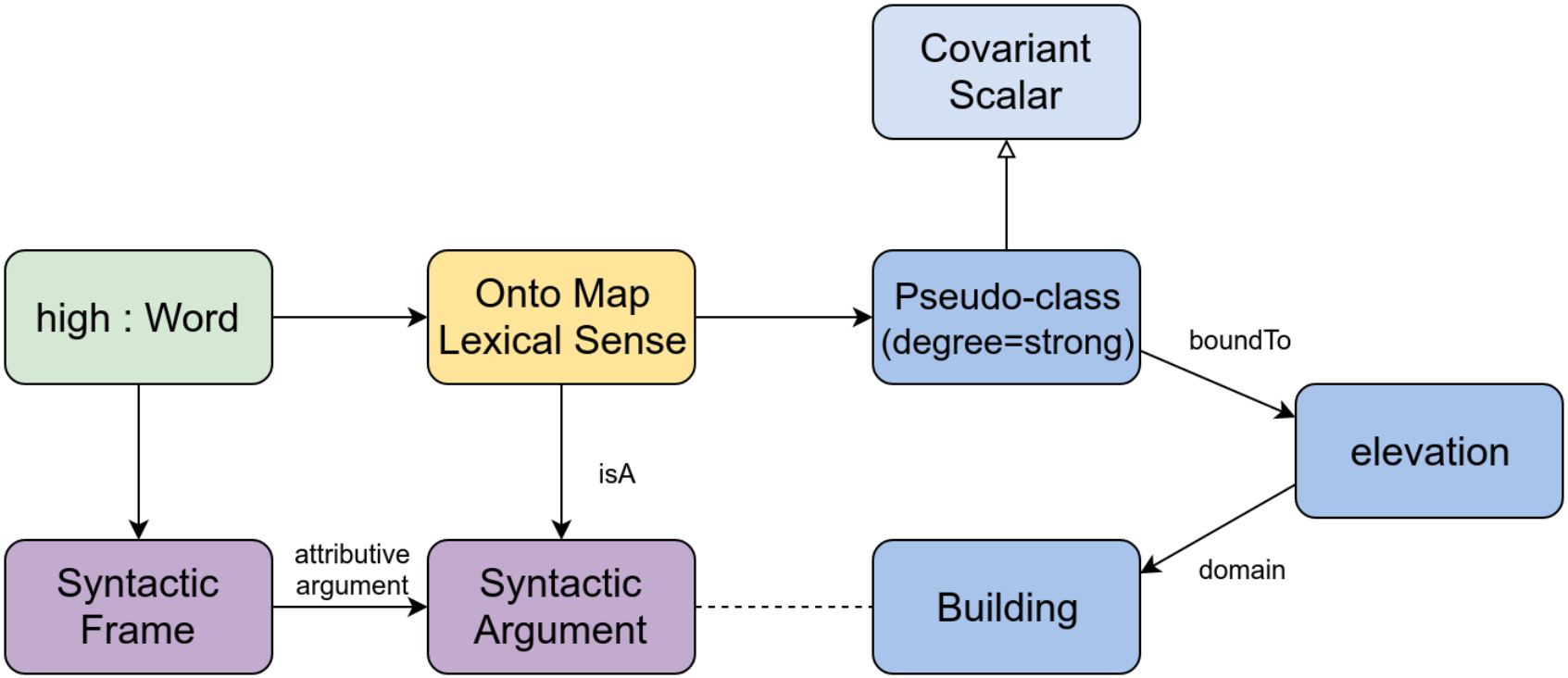
- Linguistically we define each word as having a *subcategorization frame*
  - e.g., "X knows Y"
- Each RDF property has two arguments
  - *Subject*
  - *Object*
- We need to state the correspondence of syntactic arguments and semantic arguments

Frames and Correspondence



Correspondence to Adjectives

- “Many beautiful linguistic theories fail decidely when it comes to adjectives” (Bankston, 2003)
- Especially *scalar* adjectives, such as “high”.
- Scalar adjectives are:
  - Context-sensitive
  - Fuzzy
  - Non-monotonic


Covariant Scalar
high : Word
Onto Map Lexical Sense
Pseudo-class (degree=strong)
boundTo
elevation
isA
attributive argument
Syntactic Frame
Syntactic Argument
Building
domain

Lexcial Gap 3: Defining concepts

- OWL is not a sufficient ontology model
- Interlinkage (graph density) is very important
- We do not need to capture every ‘shade’ of a sense
- Minimum definition of a definition:
  - Given only the machine-readable definition of a concept
  - It should be possible to uniquely distinguish this node

Building resources

Improving an existing resource - Princeton WordNet

- PWN 3.0 was released in 2006.
  - Not in PWN 3.0: netbook, social media, steampunk, Sriracha, hoverboard, fanbase, binge watch, relatable, text (v), spoiler (new sense), trope (new synonym)
- PWN has a low degree
- PWN is only English

Social Media WordNet

- We are working on extending PWN with neologisms
- Analyzing term frequency on Twitter relative to baseline corpus
- Term types:
  - General
  - Novel: affluenza, unboxing
  - Vulgar: chaturbate
  - Abbreviation/Misspelling: finna, idk
  - Names/Non-Lexical: zayn, i love you

The Princeton WordNet gloss corpus

- The adjective ‘Slovenian’ has one link (pertains to ‘Slovenia’)
- But the definition is more detailed and has been tagged:
  - of or relating to or characteristic of Slovenia or its people or language.
- Could we improve the density of WordNet this way?

Multilingual WordNets

- WordNets have been translated into many languages
- Not always easy to translate, e.g., ‘teacher’

  Lehrer A (male) teacher
  Lehrerin A female teacher
- New languages introduce new concepts

The WordNet Interlingual Index

- Each synset now has a Interlingual Identifier
  - `http://globalwordnet.org/ili/i16907`
- Any WordNet can propose a new synset:
  - English definition
  - At least one link

Building a new resource - Lemon Design Patterns

- Many entries have common descriptions
  - Name
  - Class Nouns
  - Object Property Noun
  - Relational Nouns
  - State Verbs
  - Consequence Verbs
  - Intersective Adjectives
  - Relational Adjectives
  - Scalar Adjectives
  - . . .

Lemon Design Patterns

```
ScalarAdjective("hoch",
  [ ontology:elevation > 50 for
    ontology:Building ]) with comparative "höher"
```

Lemon DBpedia

- For 4 Languages: English, German, Spanish, Japanese
- Covers 353 classes and 300 properties
- Finding usage in question answering, ontology engineering

Summary

Summary

- Ontologies are still a relevant target for natural language understanding
- Detail is more important than coverage
  - More semantics
  - Lexical-ontological mapping
  - More models like OntoLex-Lemon