Introduction to NLP

Knowledge Graphs & Chatbots

Dr. Paul Buitelaar

Data Science Institute, NUI Galway

Learning Outcomes of This Lecture

Understand some of the basic ideas behind chatbots

Understand the structure of a knowledge graph and its role in chatbots

Gain insight into developing a taxonomy extraction approach, in particular in regard of term extraction (taxonomy classes) and term pair identification (class hierarchy)

Overview

Chatbots

Knowledge Graphs

Taxonomy extraction

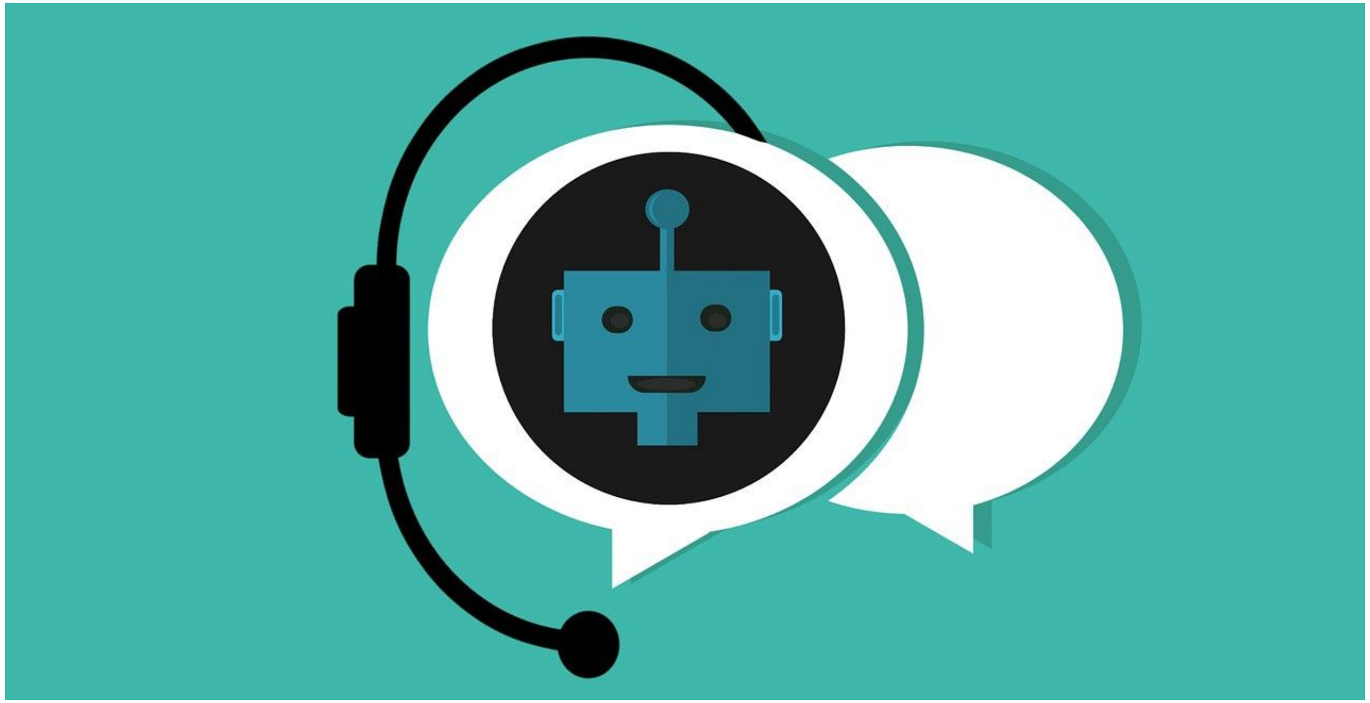
Overview

Chatbots

Knowledge Graphs

Taxonomy extraction

Chatbots



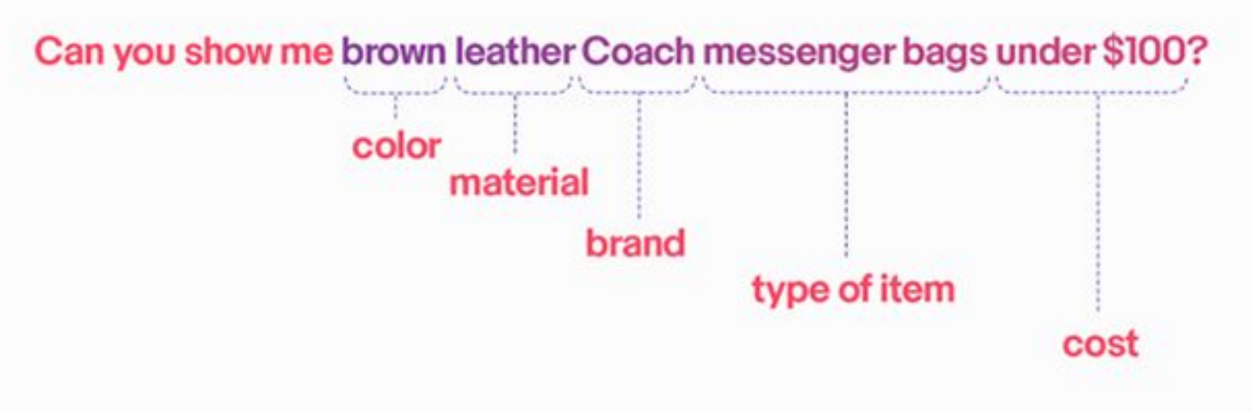
"... software application used to conduct an online chat conversation via text or text-to-speech ..." - Wikipedia

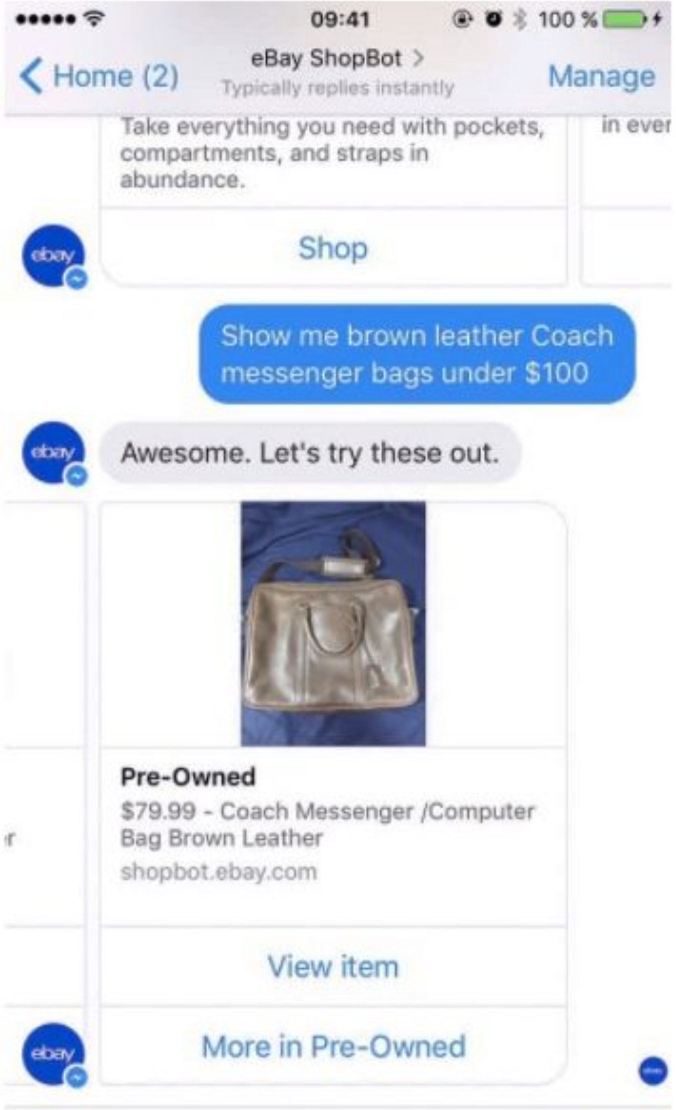
From the viewpoint of NLP a Chatbot or Conversational Agent is a **Dialog System**

For the purpose of this lecture we will refer to such systems however as Chatbots

Image source: pixabay.com

Chatbots - Example (eBay)


Can you show me brown leather Coach messenger bags under $100?
color
material
brand
type of item
cost



Take everything you need with pockets, compartments, and straps in abundance.

in ever

Show me brown leather Coach messenger bags under $100

Awesome. Let's try these out.

Pre-Owned

$79.99 - Coach Messenger /Computer Bag Brown Leather

shopbot.ebay.com

https://www.ebayinc.com/stories/news/cracking-the-code-on-conversational-commerce/

Chatbot Architecture

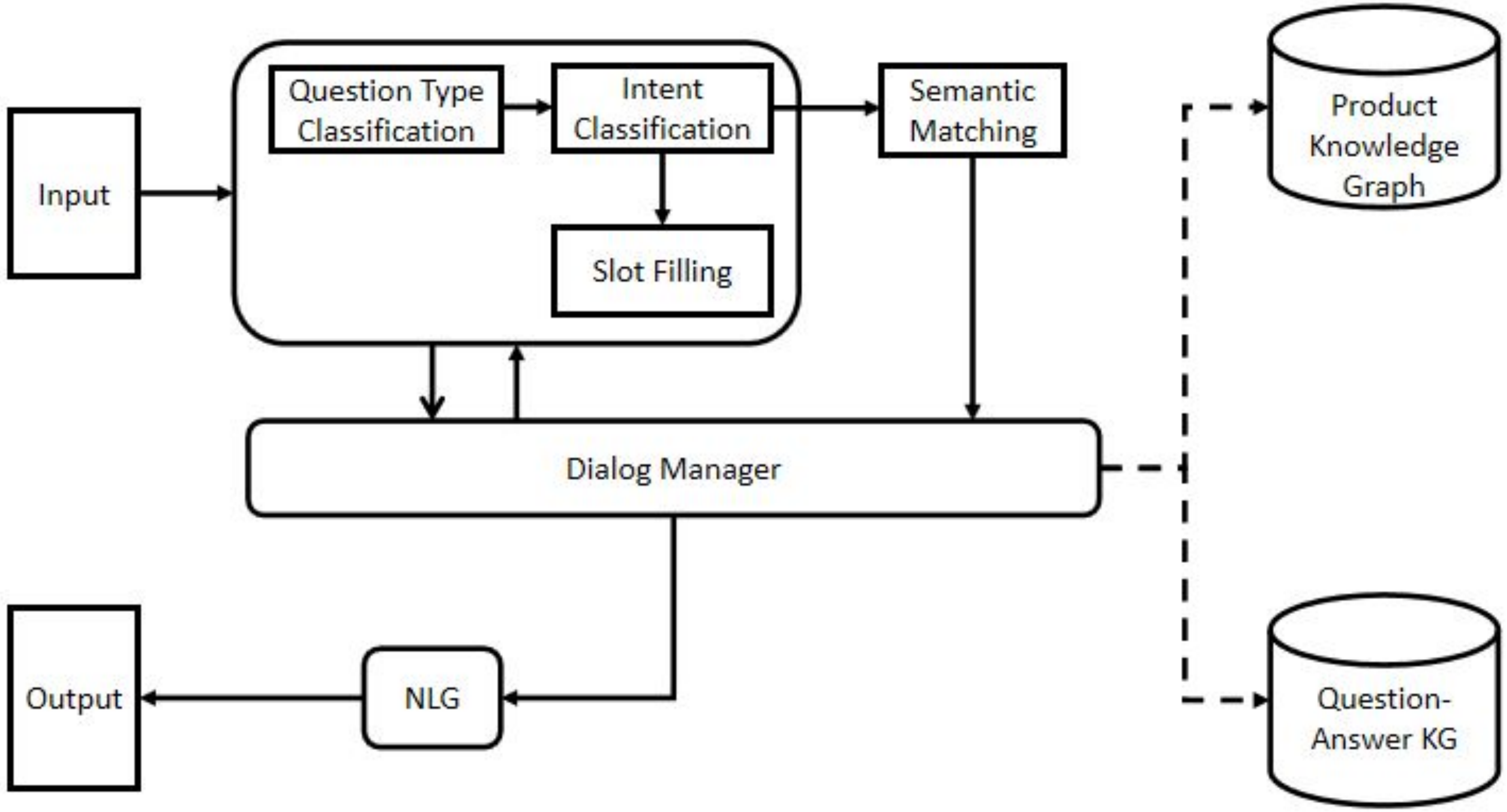
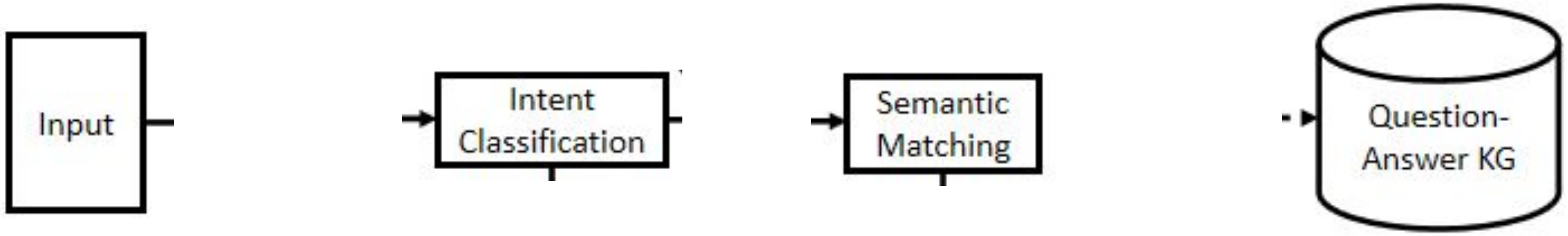


Image from "Zhao, G., Zhao, J., Li, Y., Alt, C., Schwarzenberg, R., Hennig, L., Schaffer, S., Schmeier, S., Hu, C. and Xu, F., 2019. MOLI: Smart Conversation Agent for Mobile Customer Service. Information, 10(2), p.63."

Intent Classification - Question/Answer Pairs



*Q: Do you **sell bags**?*

A: We sell all kinds of leather goods.

...

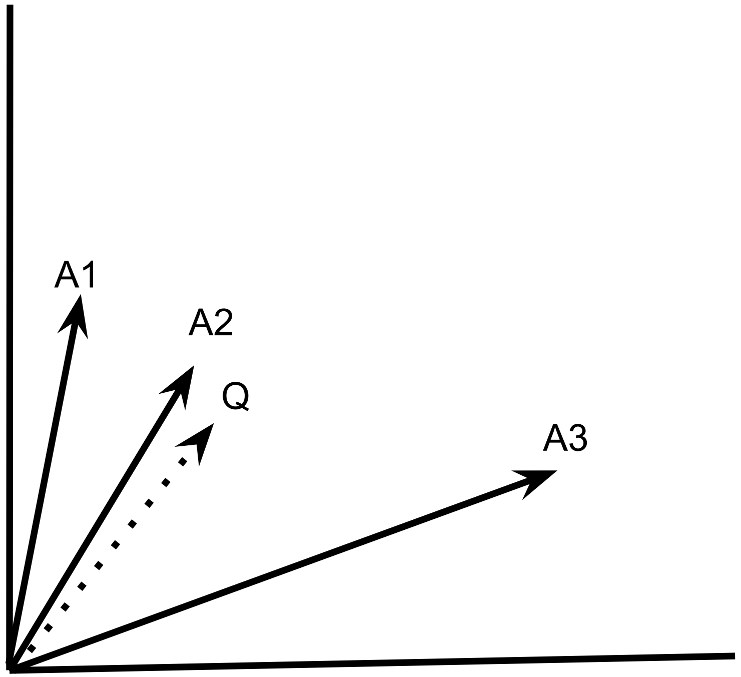
*A: We **sell bags** for all purposes.*

...

A: You can order our products online.

...

Intent Classification - Semantic Matching



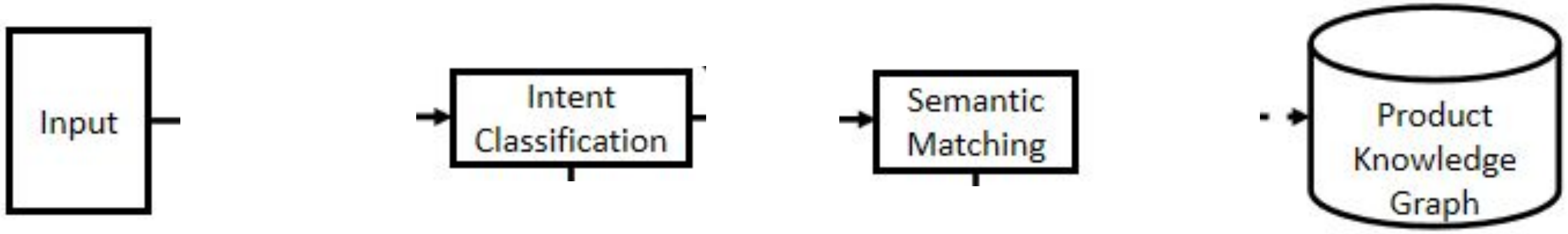
Q: Do you ***sell bags?***

A1: We ***sell*** *all kinds of leather goods****.***

A2: We ***sell bags*** *for all purposes.*

A3: You can order our products online.

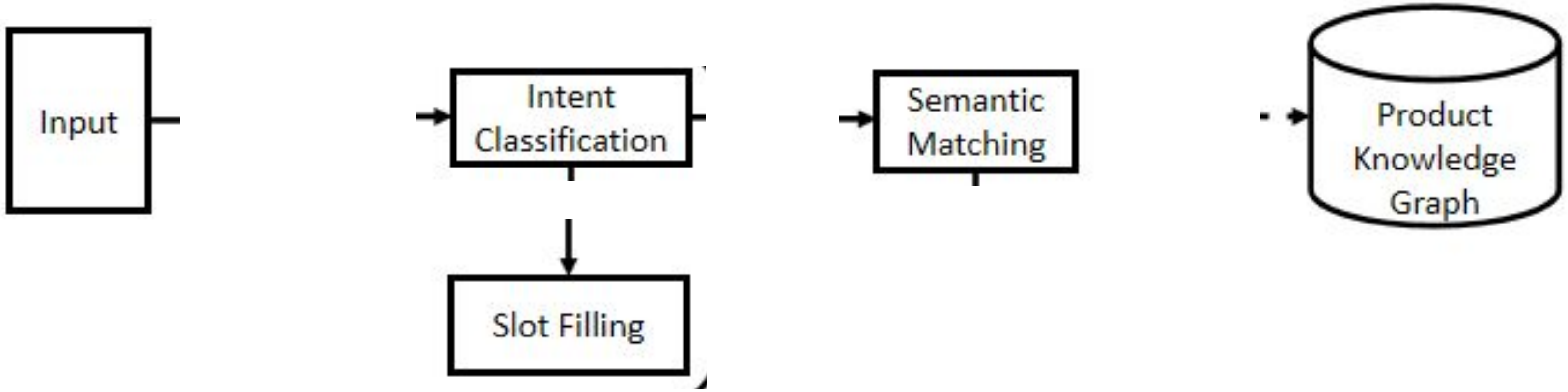
Intent Classification - Knowledge Graph



*What do you have in **shoes**?*

SHOE:	subclass	**BOOT**
	subclass	**SANDAL**
	subclass	**SNEAKER**
	….	

Intent Classification - Knowledge Graph



*How much is the **brown leather bag**?*

“BIG BROWN BAG” :	material	**LEATHER**
	has-color	**BROWN**
	has-price	99.50

Overview

Chatbots

Knowledge Graphs

Taxonomy extraction

What are Knowledge Graphs

Knowledge graphs represent **knowledge** - of the world, of a domain (finance, legal, medical, ...), of an organisation or enterprise - **in a graph structure**

From a data science point of view, knowledge graphs are representations of **semantic metadata** that describe the 'meaning of data'

From an NLP viewpoint, knowledge graphs represent **background knowledge** that can be used in reasoning over text

Consider the knowledge graph example in the following slides

Data Points

2.59

2.59

Data Typing

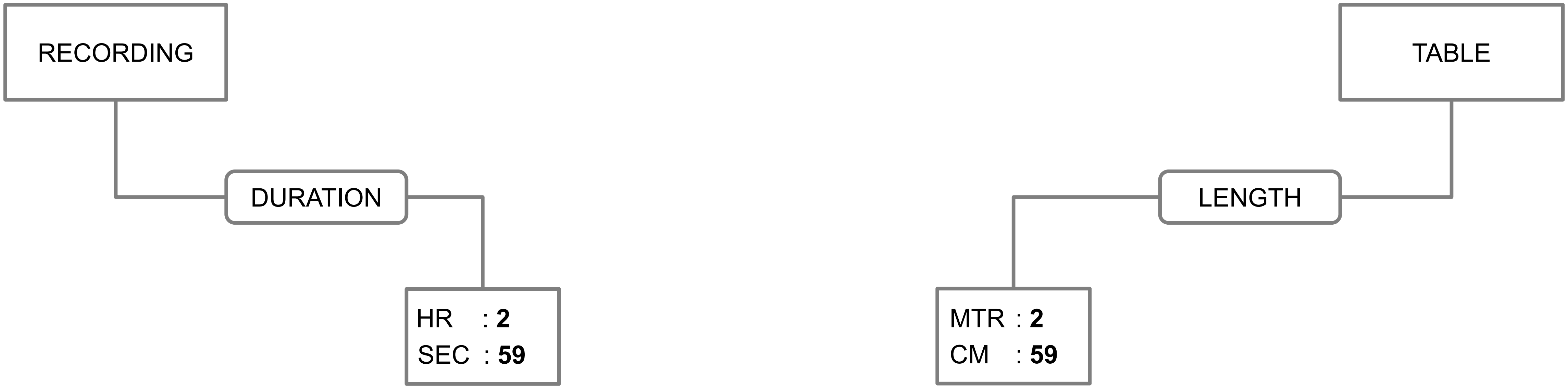
2.59	**REAL**	2.59

Metadata

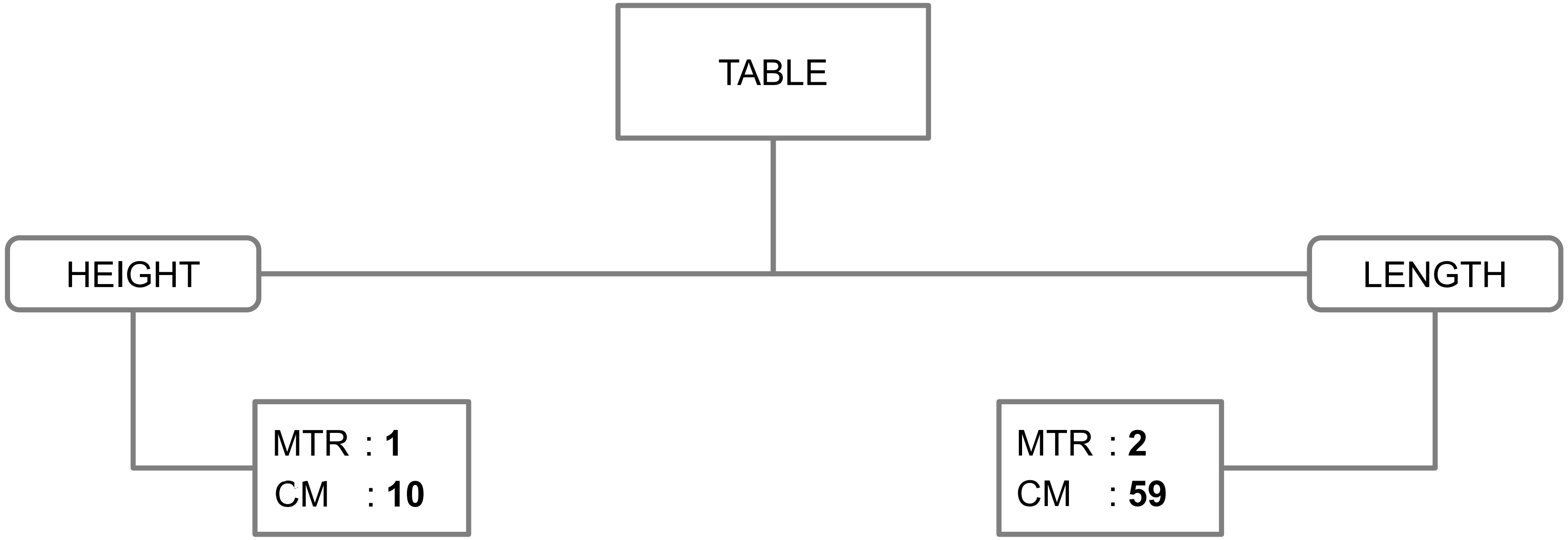
HR	SEC
2	59

MTR	CM
2	59

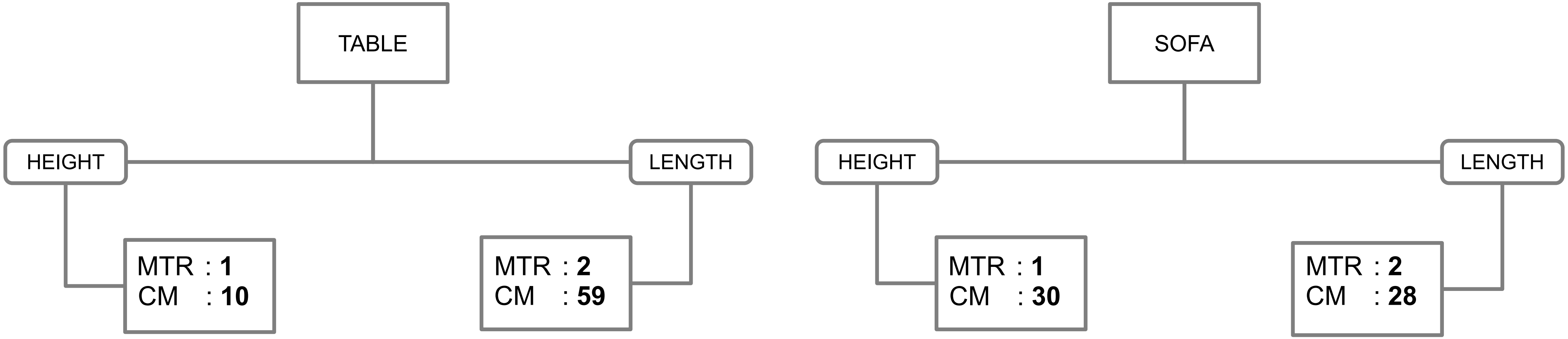
Semantic Metadata



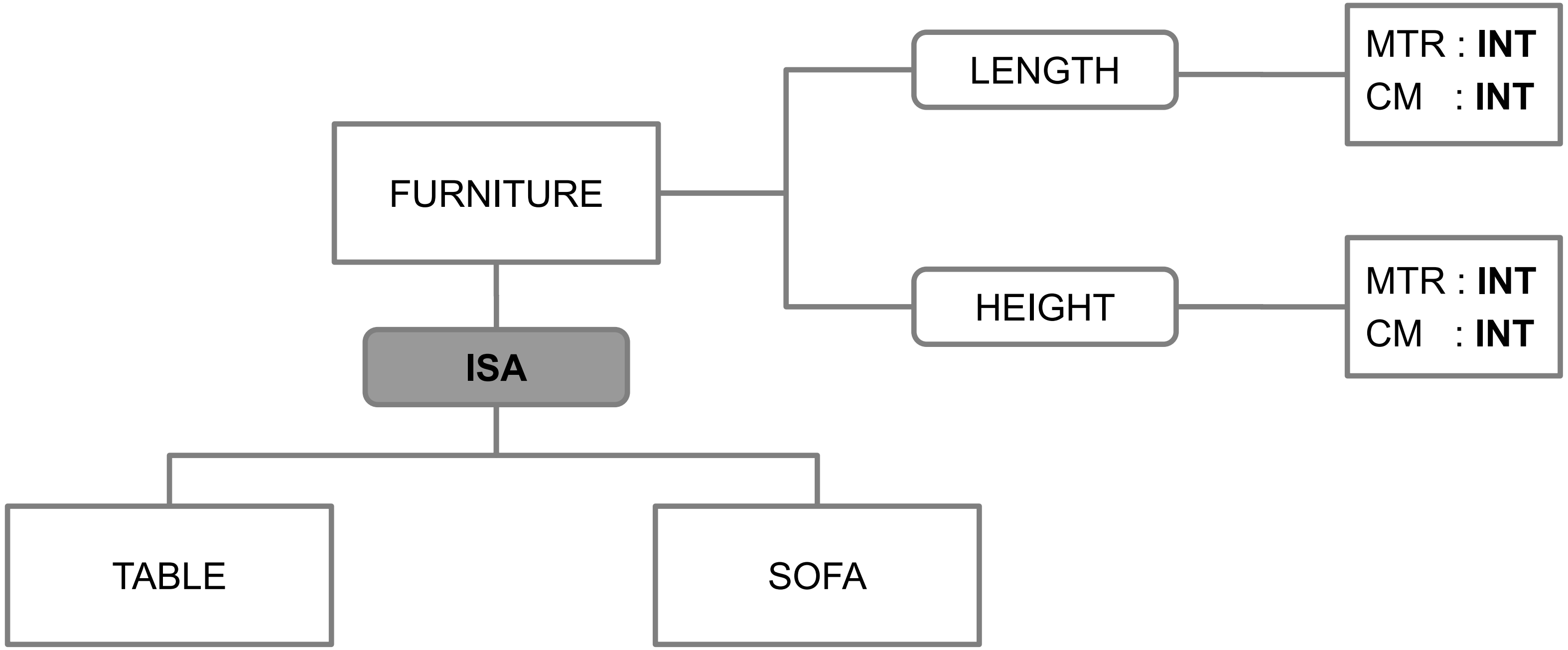
Semantic Metadata: Classes & Properties



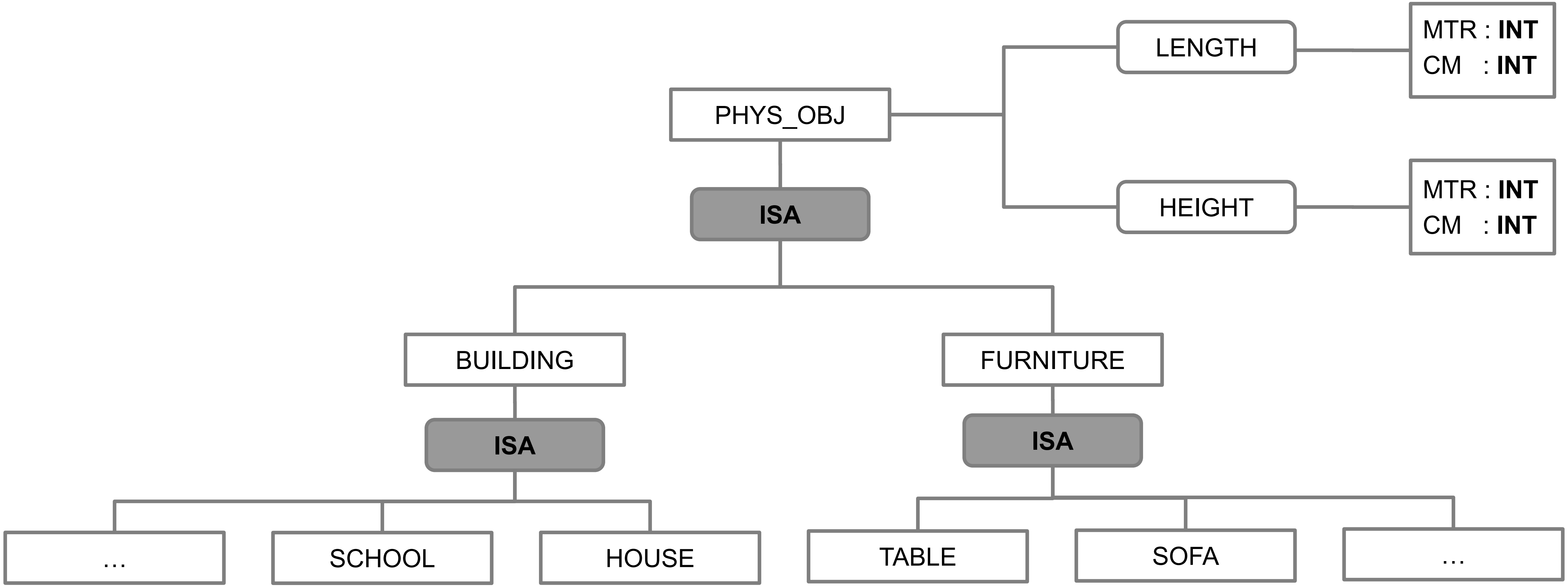
Semantic Metadata: Classes & Properties



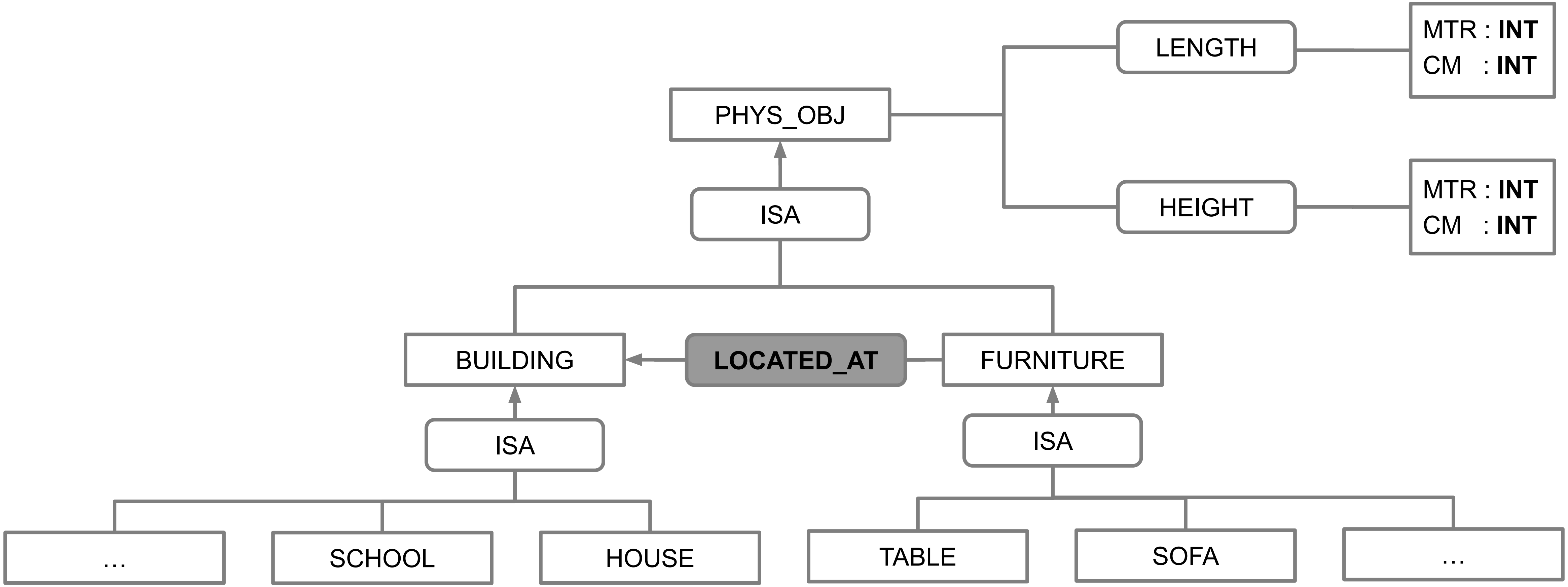
Generalization: Taxonomy



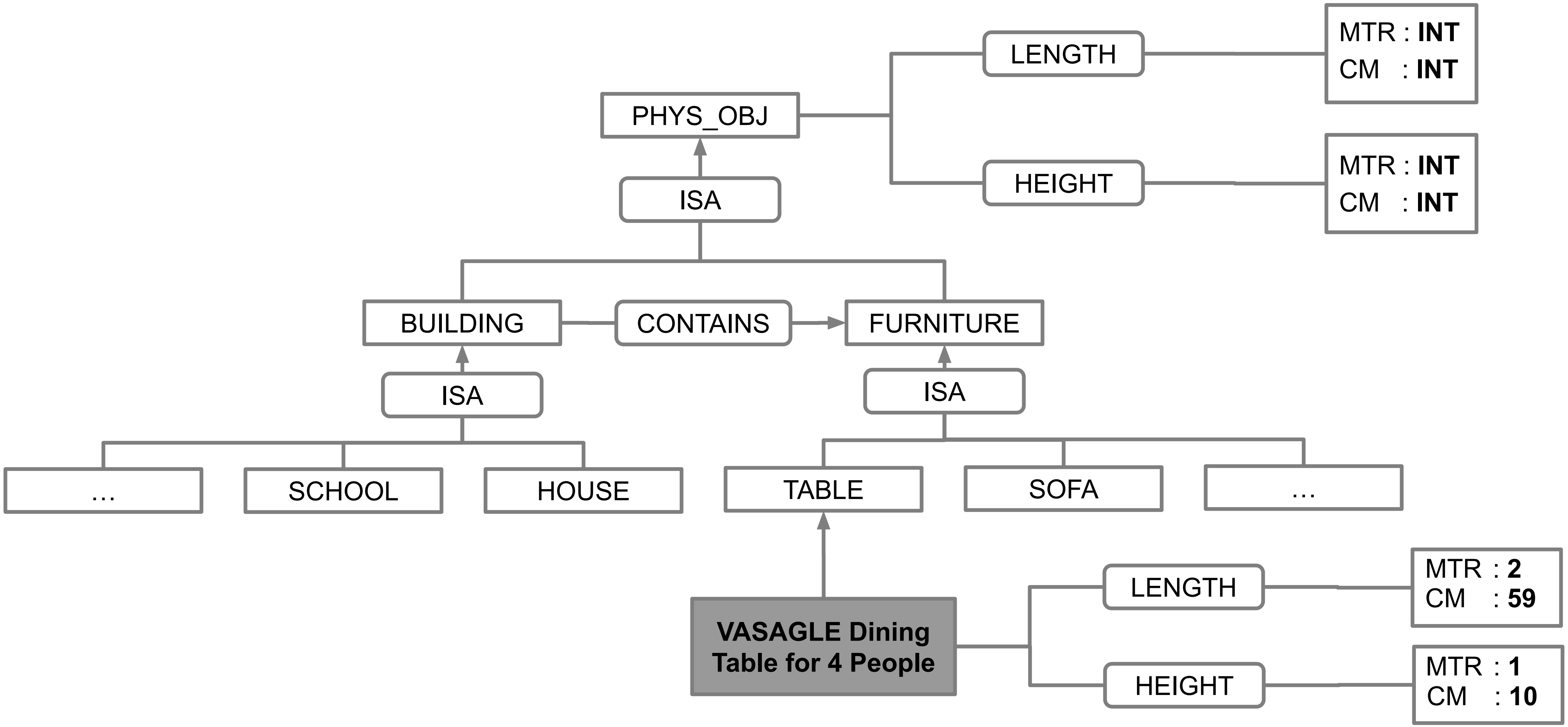
Generalization: Taxonomy



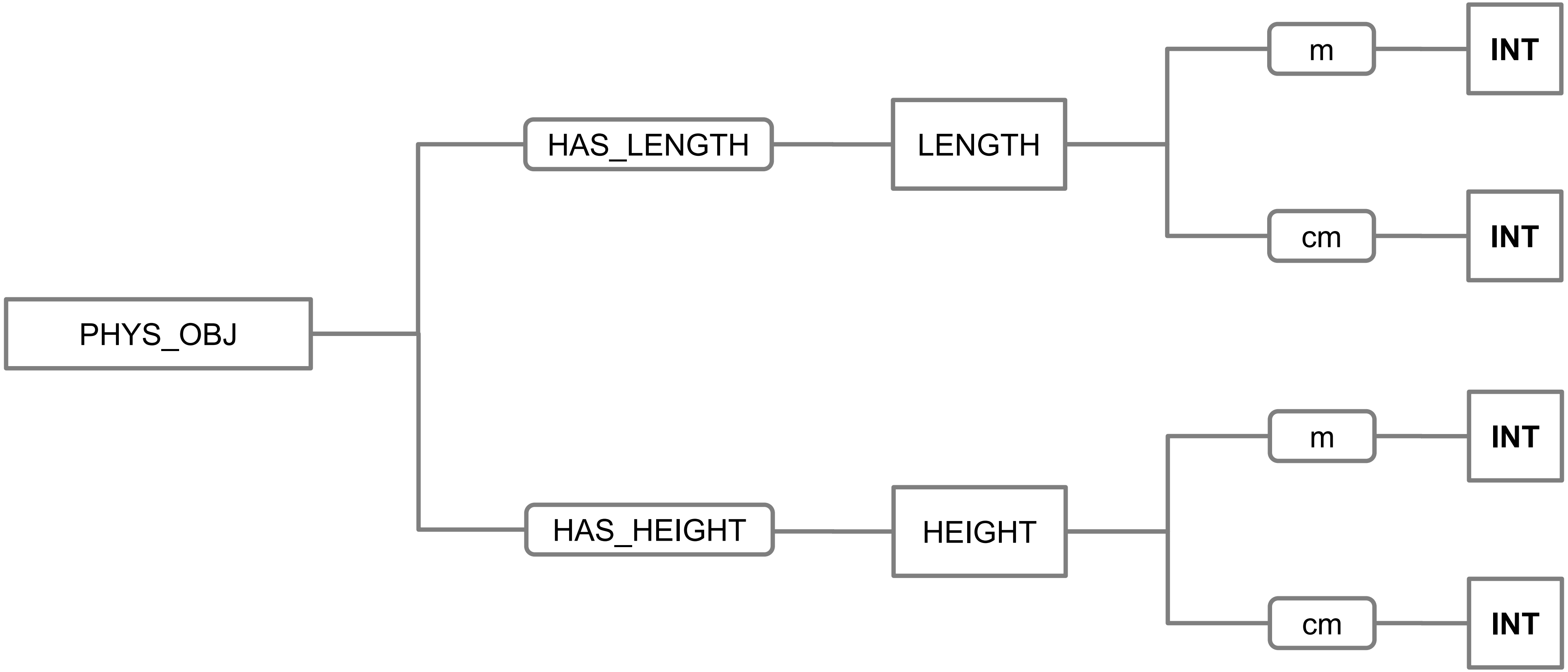
Other Relations



Instances


PHYS_OBJ
LENGTH
MTR : **INT**
CM : **INT**
HEIGHT
MTR : **INT**
CM : **INT**
ISA
BUILDING
CONTAINS
FURNITURE
ISA
ISA
…
SCHOOL
HOUSE
TABLE
SOFA
…
VASAGLE Dining Table for 4 People
LENGTH
MTR : **2**
CM : **59**
HEIGHT
MTR : **1**
CM : **10**

Properties - more detail



Formal Representation (Description Logic)

BUILDING ⊑ PHYS_OBJ
FURNITURE ⊑ PHYS_OBJ
HOUSE ⊑ BUILDING
SCHOOL ⊑ BUILDING
TABLE ⊑ FURNITURE
SOFA ⊑ FURNITURE
(FURNITURE, BUILDING) : LOCATED_AT
(BUILDING, FURNITURE) : CONTAINS
(PHYS_OBJ, LENGTH) : HAS_LENGTH
(PHYS_OBJ, HEIGHT) : HAS_HEIGHT
(LENGTH, INT) : m
(LENGTH, INT) : cm
(HEIGHT, INT) : m
(HEIGHT, INT) : cm

T-BOX

VASAGLE : TABLE
(**VASAGLE**, **VASAGLE LENGTH**) : HAS_LENGTH
(**VASAGLE**, **VASAGLE HEIGHT**) : HAS_HEIGHT
(**VASAGLE LENGTH**, **2**) : m
(**VASAGLE LENGTH**, **59**) : cm
(**VASAGLE HEIGHT**, **1**) : m
(**VASAGLE HEIGHT**, **10**) : cm

A-BOX

https://en.wikipedia.org/wiki/Description_logic#Notation

General Knowledge Graphs

DBpedia

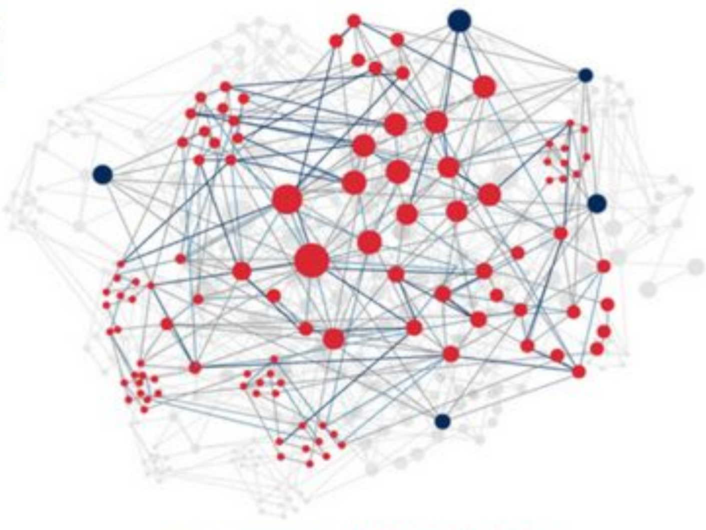
BabelNet

yago select knowledge

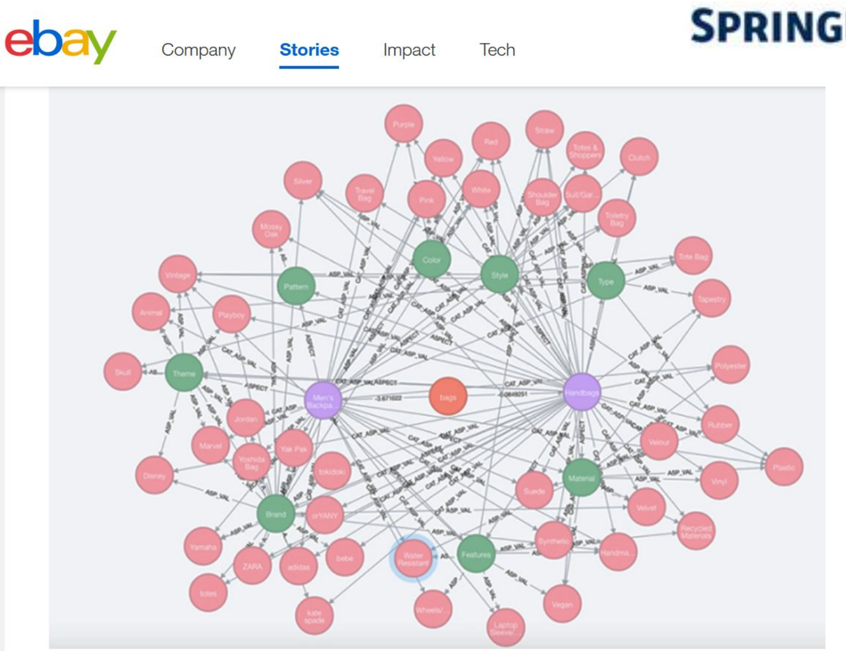
WIKIDATA

Global WordNet Association

Specific Knowledge Graphs



SPRINGER NATURE



Overview

Chatbots

Knowledge Graphs

Taxonomy extraction

Knowledge Graph Extraction

Knowledge **changes rapidly over time and between specific domains or enterprises**

Costly to develop and maintain a knowledge graph manually

NLP techniques can be used in **knowledge graph extraction from text**

A knowledge graph comprises: classes, class instances, class properties, class hierarchy, other relations between classes

Here we focus on the extraction of classes & class hierarchy (taxonomy extraction)

Taxonomy Extraction

A taxonomy class corresponds to a single or multi-word ‘term’: “*graduate degree*” “*academic degree*” “*qualification*” …

- **term extraction**

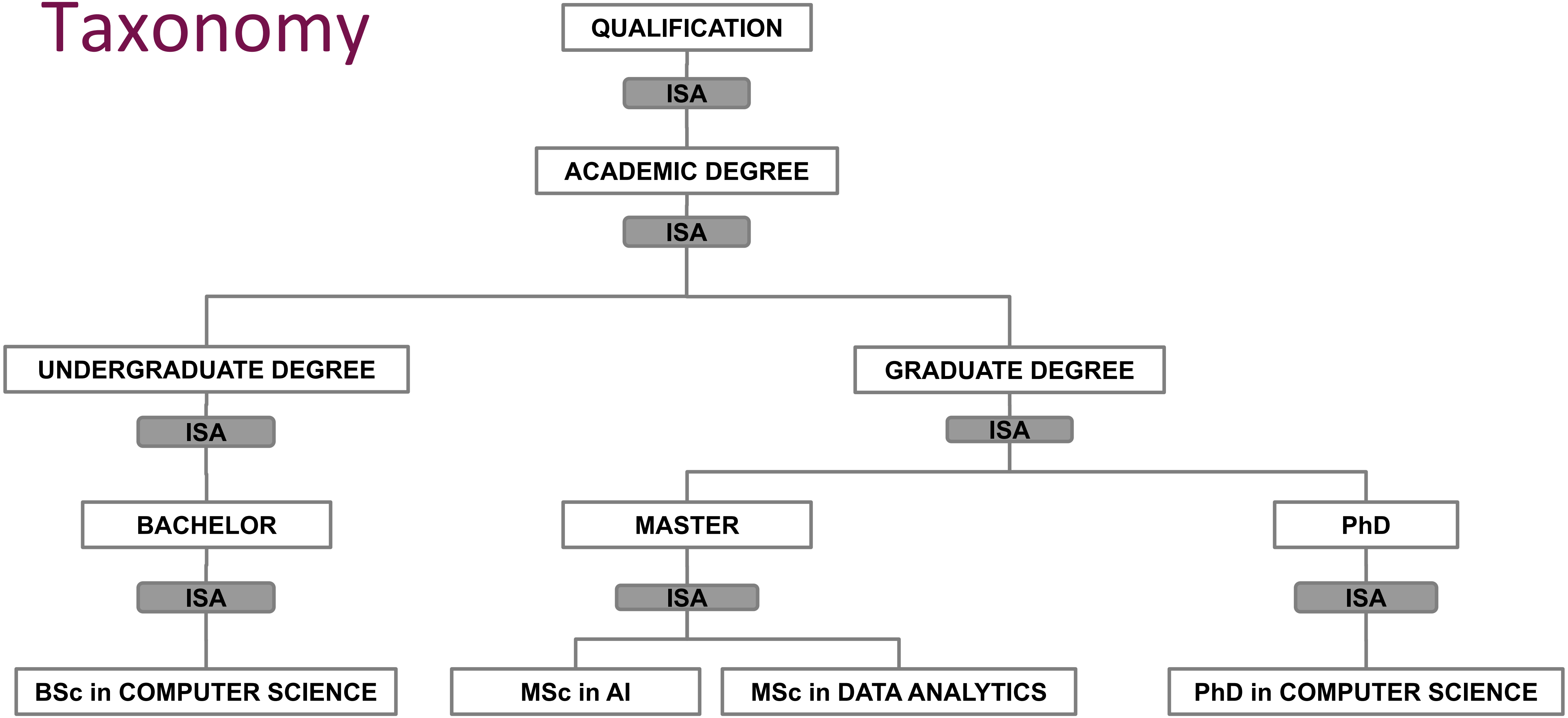
A taxonomic relation corresponds to a pair of terms *{c,d}* where *c* is more specific than *d*: *{“MSc in AI”, ”graduate degree”} {“academic degree”, ”qualification”}* ...

- **term pair identification**

A taxonomy corresponds to an optimal tree constructed from term pairs

- term pair ranking & tree construction

Taxonomy


QUALIFICATION
ISA
ACADEMIC DEGREE
ISA
UNDERGRADUATE DEGREE
GRADUATE DEGREE
ISA
ISA
BACHELOR
MASTER
PhD
ISA
ISA
ISA
BSc in COMPUTER SCIENCE
MSc in AI
MSc in DATA ANALYTICS
PhD in COMPUTER SCIENCE

Taxonomy Elements

Terms

qualification
academic degree
undergraduate degree
graduate degree
bachelor
master
PhD
BSc in Computer Science
MSc in AI
MSc in Data Analytics
PhD in Computer Science

Term Pairs

{academic degree, qualification}
{undergraduate degree, academic degree}
{graduate degree, academic degree}
{bachelor, undergraduate degree]
{master, graduate degree}
{PhD, graduate degree}
{BSc in Computer Science, bachelor}
{MSc in AI, master}
{MSc in Data Analytics, master}
{PhD in Computer Science, PhD}

Term Extraction

Unsupervised

- Extract, rank and filter Noun Phrases

Supervised

- Term annotation

Semi-supervised

- Distant supervision (Wikipedia, termbases)

Unsupervised Term Extraction

Unsupervised term extraction can build on syntactic analysis as a **term often corresponds to a Noun Phrase (NP)**

$[_{NP}\ \mathit{university}_{N}]$

$[_{NP}\ \mathit{teaching}_{JJ}\ \mathit{assistant}_{N}]$

$[_{NP}\ \mathit{postgraduate}_{JJ}\ [_{NP}\ \mathit{taught}_{JJ}\ \mathit{course}_{N}]]$

$[_{NP}\ \mathit{lecturer}_{N}\ [_{PP}\ \mathit{above}_{IN}\ [_{NP}\ \mathit{the}_{DET}\ \mathit{bar}_{n}]]]$

Term Candidates

Not all NPs are equally good term candidates, consider:

undergraduate student

vs.

undergraduate advising

Term Ranking - PMI

Use Pointwise Mutual Information (PMI) to rank NPs based on the correlative strength between the individual words in the NP

Recall the PMI formula used previously

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

PMI Example

	Frequency in Corpus	Co-occurrence with ‘undergraduate’
undergraduate	50	-
student	200	40
advising	500	10

PMI (undergraduate, student)

= $\log_2$ (P(undergraduate, student) / (P(undergraduate) P(student)))
= $\log_2$ (40/750 / (50/750 * 200/750))
= $\log_2$ (0.05 / (0.07 * 0.27)) = $\log_2$ (2.5) = **0.40**

PMI (undergraduate, advising)

= $\log_2$ (P(undergraduate, advising) / (P(undergraduate) P(advising)))
= $\log_2$ (10/750 / (50/750 * 500/750))
= $\log_2$ (0.01 / (0.07 * 0.67)) = $\log_2$ (0.1) = **-1**

Supervised Term Extraction

Supervised term extraction approaches train a classifier on manually labeled data (term annotation)

General challenges in manual annotation apply also here: Inter-Annotator Agreement, cost of manual labour, renewed annotation for different domains

Term Annotation

Term annotation example in the ‘university domain’

“Although each institution is organized differently, nearly all universities have a board of trustees; a president, chancellor, or rector; at least one vice president, vice-chancellor, or vice-rector; and deans of various divisions. Universities are generally divided into a number of academic departments, schools or faculties. Public university systems are ruled over by government-run higher education boards.“ - https://en.wikipedia.org/wiki/University

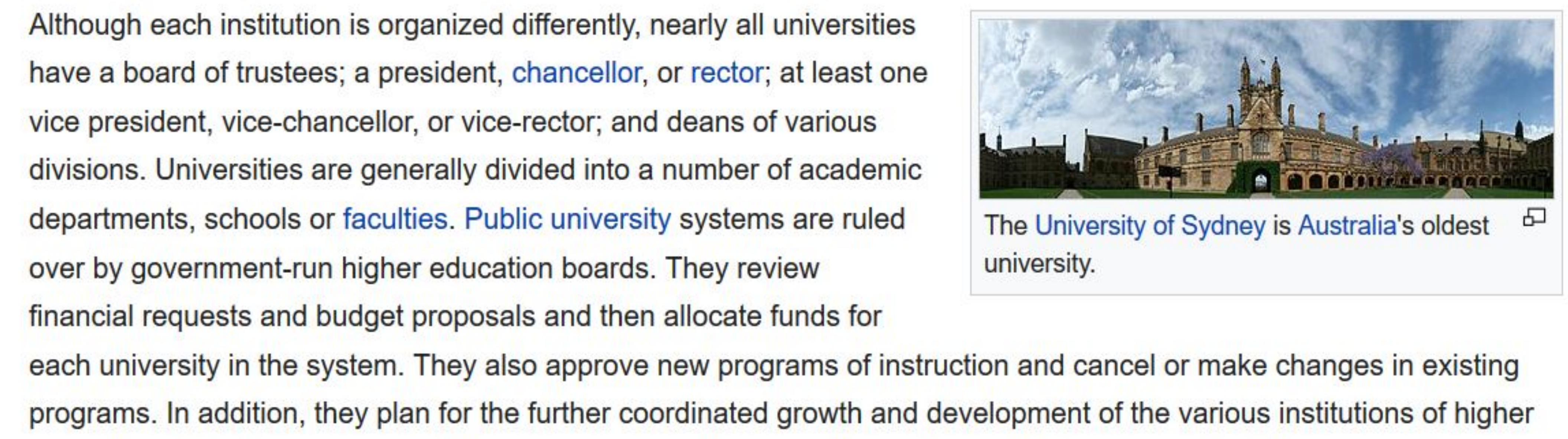
Term Annotation

Term annotation example in the ‘university domain’

*“Although each **[TERM institution]** is organized differently, nearly all **[TERM university]** have a **[TERM board of trustees]**; a **[TERM president]**, **[TERM chancellor]**, or **[TERM rector]**; at least one **[TERM vice president]**, **[TERM vice-chancellor]**, or **[TERM vice-rector]**; and **[TERM dean]** of various divisions. **[TERM university]** are generally divided into a number of **[TERM academic department]**, **[TERM school]** or **[TERM faculty]**. **[TERM public university]** systems are ruled over by government-run **[TERM higher education board]**.“ - https://en.wikipedia.org/wiki/University*

Distant Supervision - Wikipedia

Instead of manual term annotation we can use **Wikipedia ‘anchors’** as indicators for terms, for example:



Although each institution is organized differently, nearly all universities have a board of trustees; a president, chancellor, or rector; at least one vice president, vice-chancellor, or vice-rector; and deans of various divisions. Universities are generally divided into a number of academic departments, schools or faculties. Public university systems are ruled over by government-run higher education boards. They review financial requests and budget proposals and then allocate funds for each university in the system. They also approve new programs of instruction and cancel or make changes in existing programs. In addition, they plan for the further coordinated growth and development of the various institutions of higher

The University of Sydney is Australia's oldest university.

https://en.wikipedia.org/wiki/University

Distant Supervision - Wikipedia

Instead of manual term annotation we can use **Wikipedia ‘anchors’** as indicators for terms, for example:

*Although each institution is organized differently, nearly all universities have a board of trustees; a president, **[[chancellor (education) | chancellor]]**, or **[[Rector (academia) | rector]]**; at least one vice president, vice-chancellor, or vice-rector; and deans of various divisions. Universities are generally divided into a number of academic departments, schools or **[[faculty (division) | faculties]]**. **[[Public university]]** systems are ruled over by government-run higher education boards.*

https://en.wikipedia.org/wiki/University

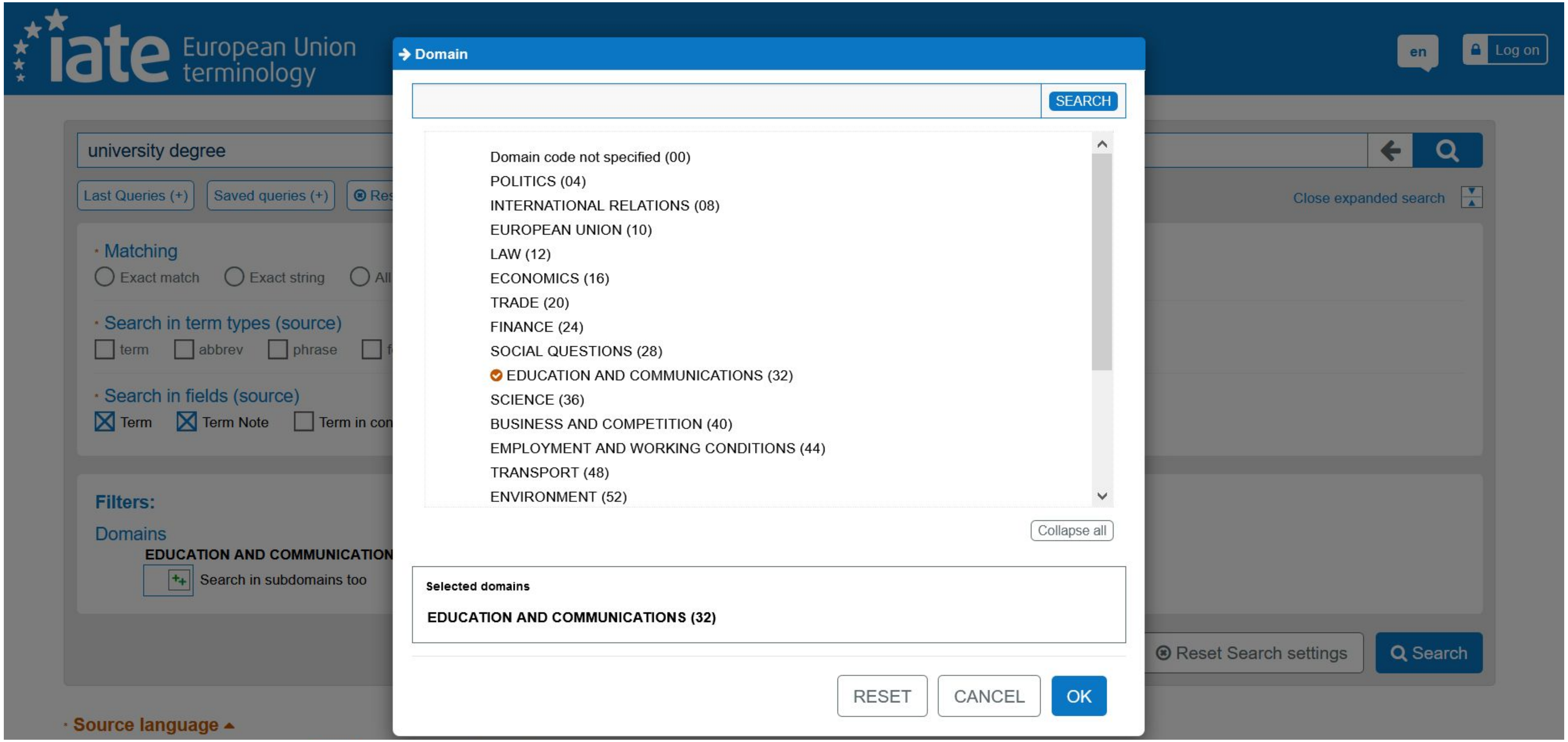
Distant Supervision - Termbases

We can also **use existing terminology (terms lists or termbases)** in developing a term extraction approach

A **termbase (‘terminology database’)** is often developed for translation applications and therefore mostly has terms in multiple languages

Largest termbase publicly available is ‘IATE’, the central multilingual term repository of the European Union

EU Termbase IATE - Terms in Many Domains



Domain code not specified (00)
POLITICS (04)
INTERNATIONAL RELATIONS (08)
EUROPEAN UNION (10)
LAW (12)
ECONOMICS (16)
TRADE (20)
FINANCE (24)
SOCIAL QUESTIONS (28)
✓ EDUCATION AND COMMUNICATIONS (32)
SCIENCE (36)
BUSINESS AND COMPETITION (40)
EMPLOYMENT AND WORKING CONDITIONS (44)
TRANSPORT (48)
ENVIRONMENT (52)

Selected domains

EDUCATION AND COMMUNICATIONS (32)

NUI Galway
OÉ Gaillimh

https://iate.europa.eu/developers

Term Pair Identification

Unsupervised

- Substrings
- Hearst patterns
- Clustering

Supervised

- Taxonomy pairs

Substrings

Class hierarchy often corresponds to substrings, where a **shorter, more general term is embedded in a longer, more specific term**

{graduate degree, degree}
{computer science degree, degree}
{computer science degree, science degree}
{science degree, degree}
{public university, university}
{private university, university}
{higher education board, board}

Substrings - Nominal Head

Substring analysis depends on Noun Phrase analysis (nominal head)

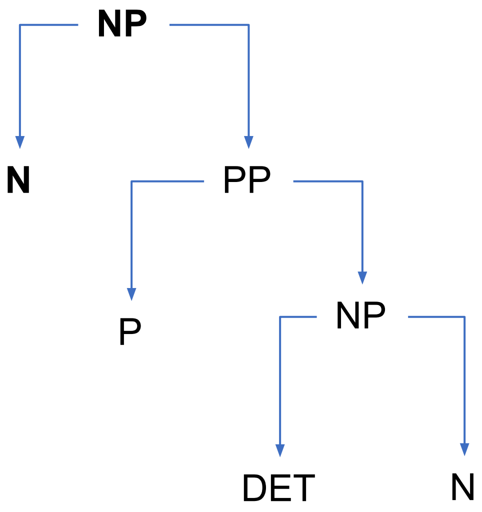
$[_{NP}\ \textit{university}_{N}\ \textit{lecturer}_{N\text{-}head}]$

{university lecturer, lecturer}

but note:

$[_{NP}\ \textit{lecturer}_{N\text{-}head}\ [_{PP}\ \textit{above}_{IN}\ [_{NP}\ \textit{the}_{DET}\ \textit{bar}_{n}]]]$

{lecturer above the bar, lecturer}

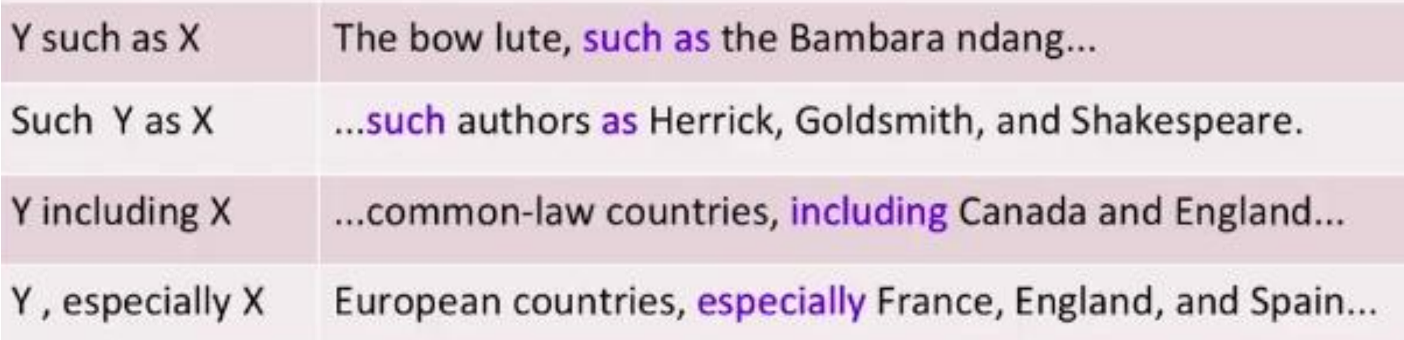


Hearst Patterns

Recall that Hearst Patterns are used for Information Extraction, in particular for the **acquisition (extraction) of hypernyms from text**

Recall that **hypernyms are taxonomic relations between word senses in WordNet**

Hearst patterns can therefore be used also directly in term pair identification



Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

Marti A. Hearst, Automatic acquisition of hyponyms from large text corpora (1992). Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, pp. 539-545

Hearst Patterns

*“These institutions commonly offer **degrees at various levels, usually including bachelor's, master's and doctorates**, often alongside other academic certificates and professional degrees.”* - [*https://en.wikipedia.org/wiki/Academic_degree*](https://en.wikipedia.org/wiki/Academic_degree)

Apply Hearst pattern: **IF *“Y including [X1, …, Xn]”* THEN *{X1 , Y}* … *{Xn , Y}***

“degrees … including bachelor’s, master's and doctorates”-> {bachelor’s , degree}

-> {master's , degree}

-> {doctorate , degree}

Clustering

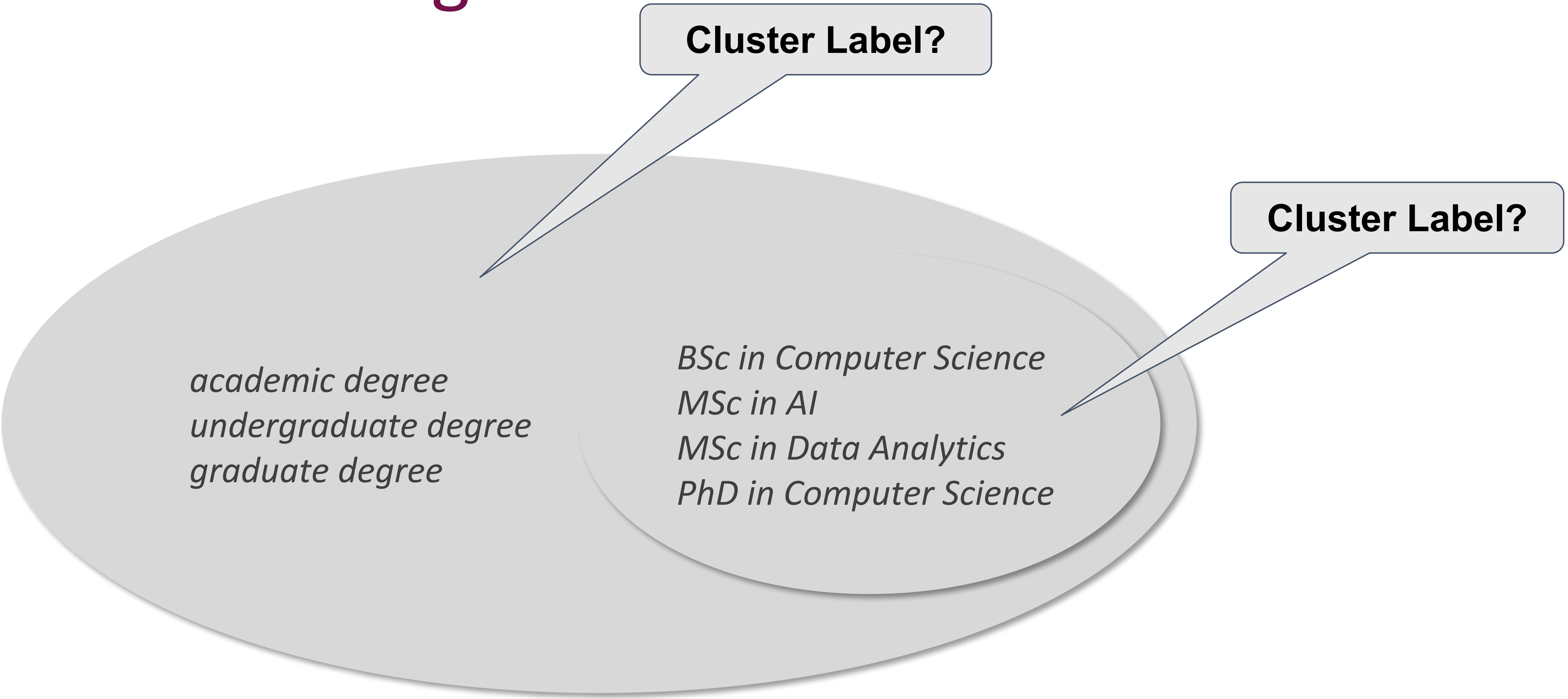
Use vector space model

- Construct vectors for all terms (use pre-trained models)
- Compute cosine similarity between vectors
- Build clusters over similar vectors

Challenges

- Labelling of clusters unclear and therefore difficult to evaluate

Cluster Labeling



Cluster Label?

Cluster Label?

academic degree
undergraduate degree
graduate degree

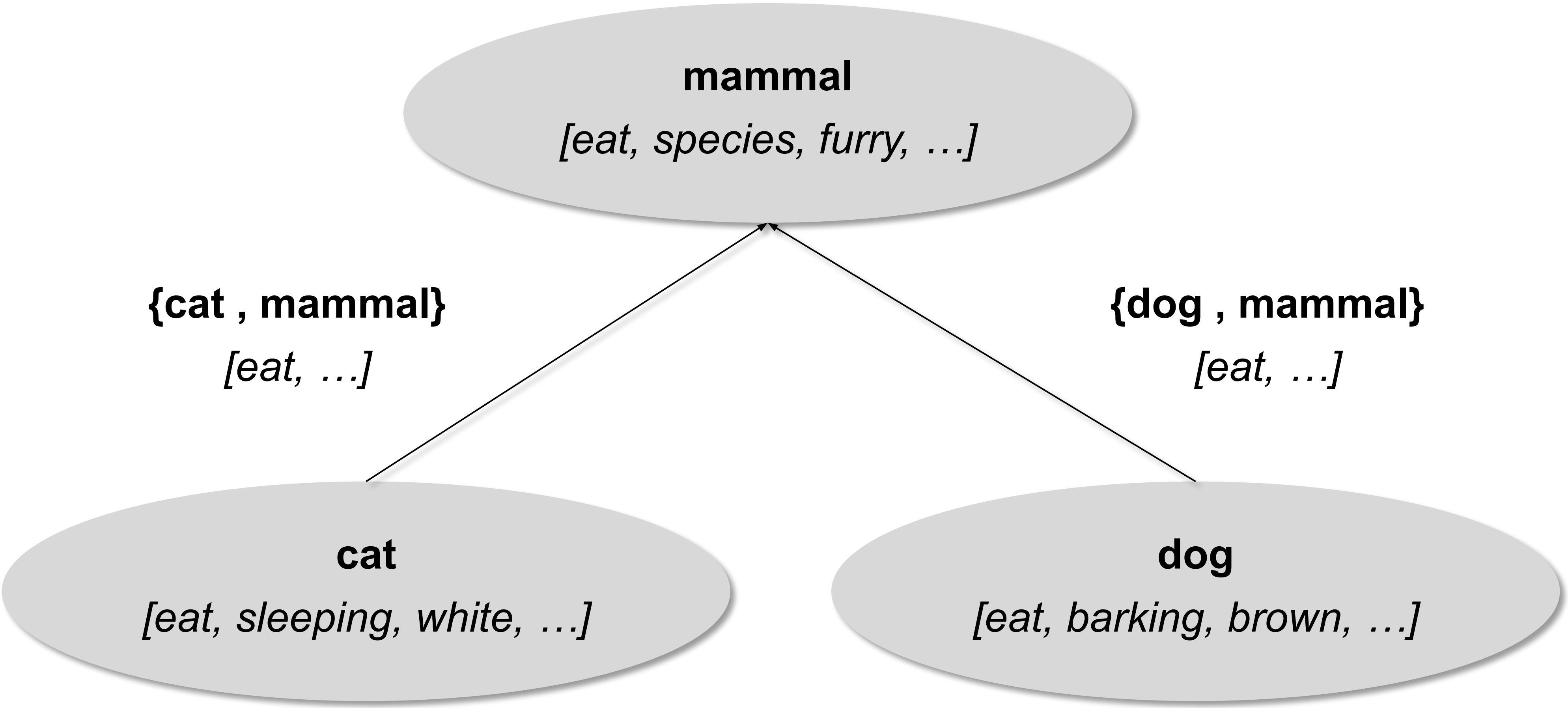
BSc in Computer Science
MSc in AI
MSc in Data Analytics
PhD in Computer Science

Supervised Training

Use vector space model

- Construct vectors for all terms (use pre-trained models)
- Construct combined vectors for term pairs and optimize through supervised training to **predict taxonomic relation between a given term pair**

Supervised Training - Example



mammal

[eat, species, furry, …]

{cat , mammal}

[eat, …]

{dog , mammal}

[eat, …]

cat

[eat, sleeping, white, …]

dog

[eat, barking, brown, …]

Supervised Training on Taxonomy Pairs

Supervised training on taxonomy (term) pairs from **existing taxonomies**

- Domain classifications
- WordNet hypernyms
- Wikipedia categorization

Domain Classifications

{Academic Advising , Academics}

{Academic Calendar , Academics}

…

{Degrees , Academics}

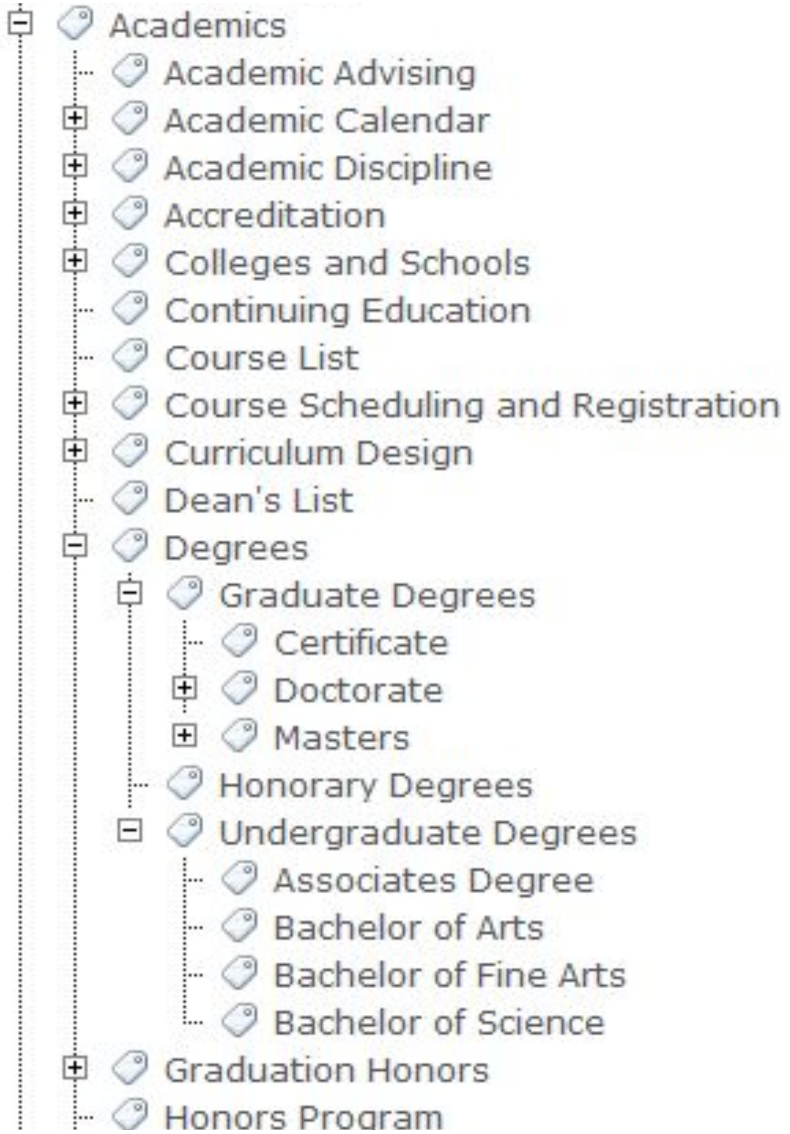
{Graduate Degrees , Degrees}

{Certificate , Graduate Degrees}

…

{Undergraduate Degrees , Degrees}

…



- Academics
  - Academic Advising
  - Academic Calendar
  - Academic Discipline
  - Accreditation
  - Colleges and Schools
  - Continuing Education
  - Course List
  - Course Scheduling and Registration
  - Curriculum Design
  - Dean's List
  - Degrees
    - Graduate Degrees
      - Certificate
      - Doctorate
      - Masters
    - Honorary Degrees
    - Undergraduate Degrees
      - Associates Degree
      - Bachelor of Arts
      - Bachelor of Fine Arts
      - Bachelor of Science
  - Graduation Honors
  - Honors Program

http://www.wandinc.com/wand-higher-education-taxonomy.aspx

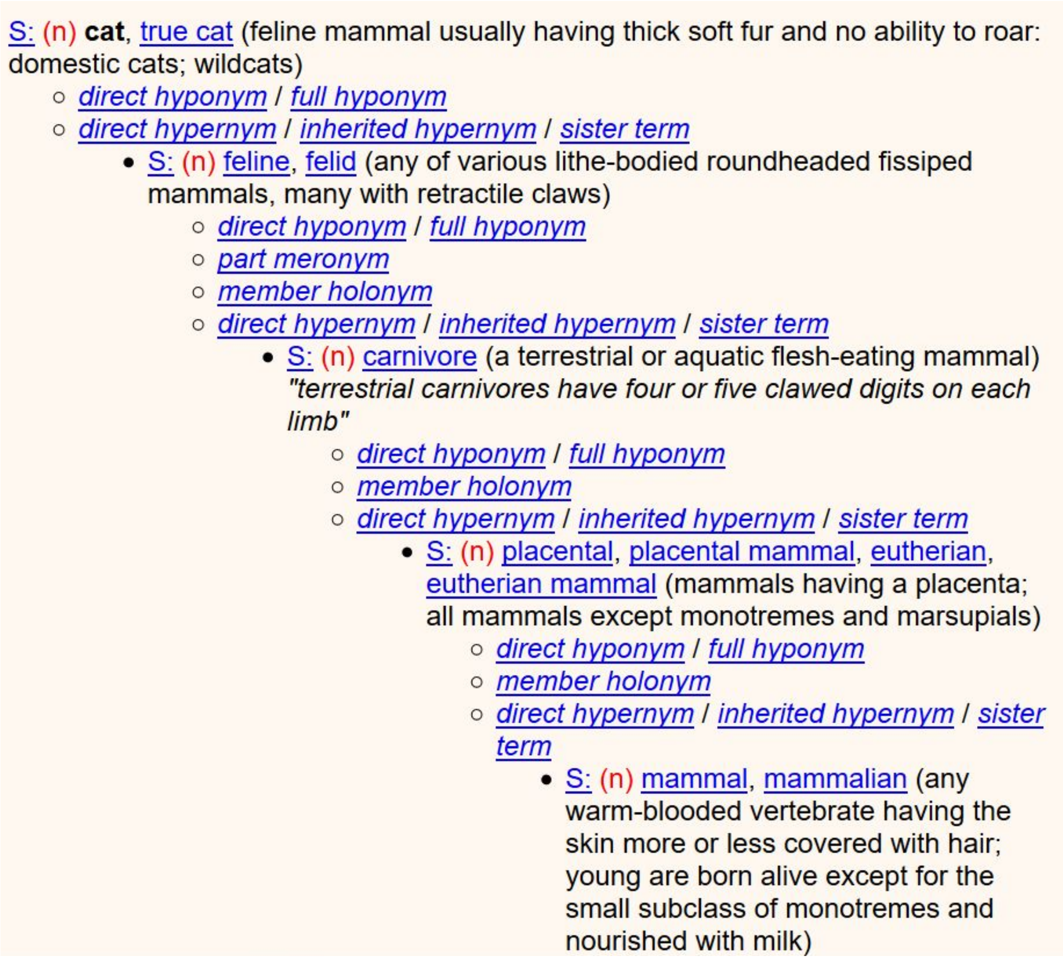
WordNet Hypernyms

{“cat, true cat” , “feline, felid”}

{“feline, felid” , “carnivore”}

{“carnivore” , “placental, placental mammal, eutherian, eutherian mammal”}

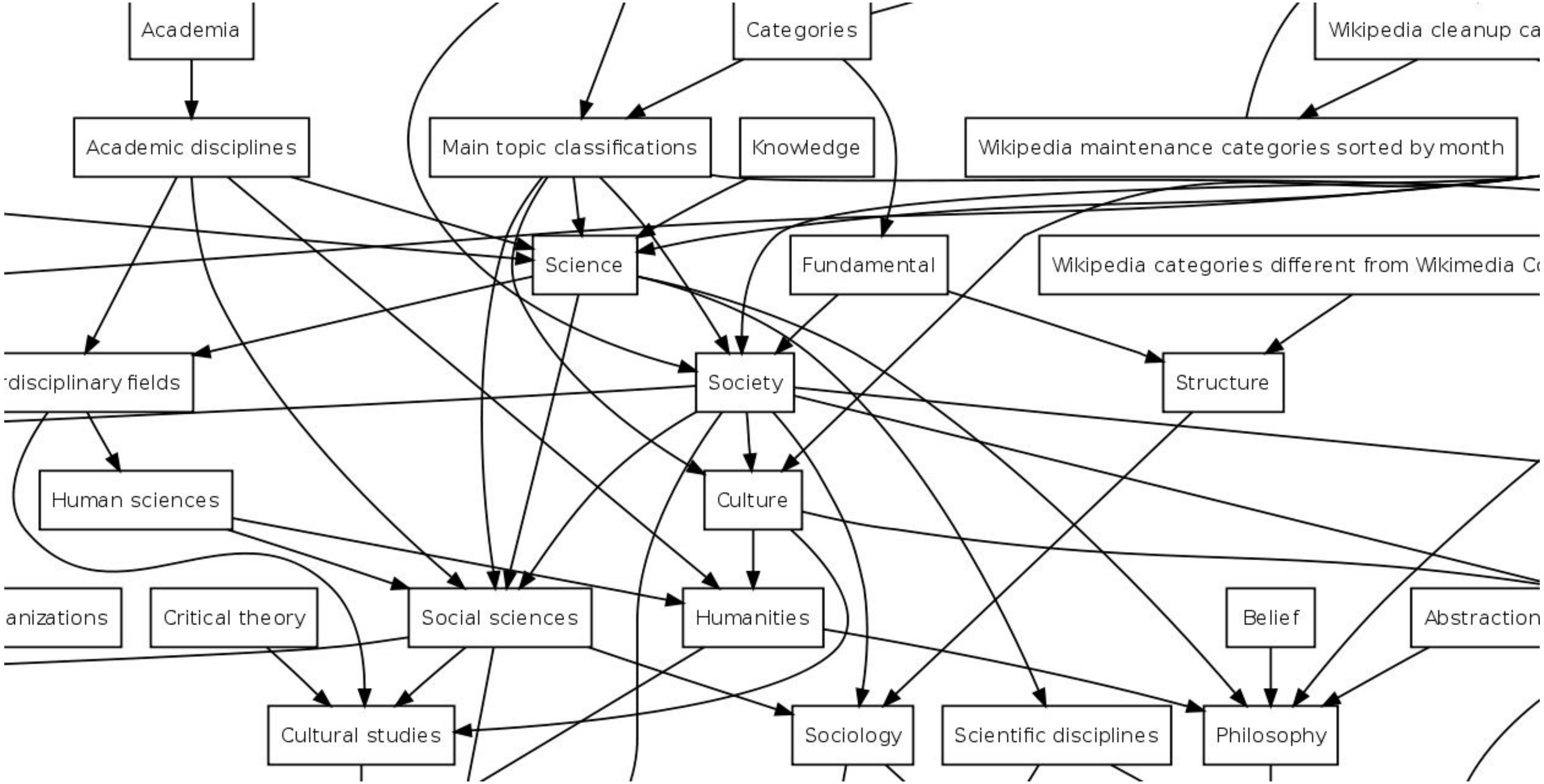
{“placental, placental mammal, eutherian, eutherian mammal” , “mammal, mammalian”}



S: (n) **cat**, true cat (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)

- *direct hyponym* / *full hyponym*
- *direct hypernym* / *inherited hypernym* / *sister term*
  - S: (n) feline, felid (any of various lithe-bodied roundheaded fissiped mammals, many with retractile claws)
    - *direct hyponym* / *full hyponym*
    - *part meronym*
    - *member holonym*
    - *direct hypernym* / *inherited hypernym* / *sister term*
      - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
        - *direct hyponym* / *full hyponym*
        - *member holonym*
        - *direct hypernym* / *inherited hypernym* / *sister term*
          - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
            - *direct hyponym* / *full hyponym*
            - *member holonym*
            - *direct hypernym* / *inherited hypernym* / *sister term*
              - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)

Wikipedia Categorization



NUI Galway
OÉ Gaillimh

https://en.wikipedia.org/wiki/Wikipedia:Categorization

Taxonomy Extraction - Evaluation

Taxonomy extraction is an Information Extraction task

Use IE evaluation metrics as before (Precision/Recall/F-score)

Use an existing taxonomy as Gold Standard (GS)

Compute overlap between GS and extracted taxonomy on terms / pairs

Lab of this Week

Exercises in term extraction

NUI Galway
OÉ Gaillimh

QA