Knowledge Graph extraction from Scholarly Data

-

Construction de graphes de connaissances à partir des publications scientifiques

Davide Buscaldi
LIPN - Laboratoire d'Informatique de Paris Nord
Université Paris 13

davide.buscaldi@lipn.univ-paris13.fr

Journée commune AFIA-ARIA, Paris 2/12/2019

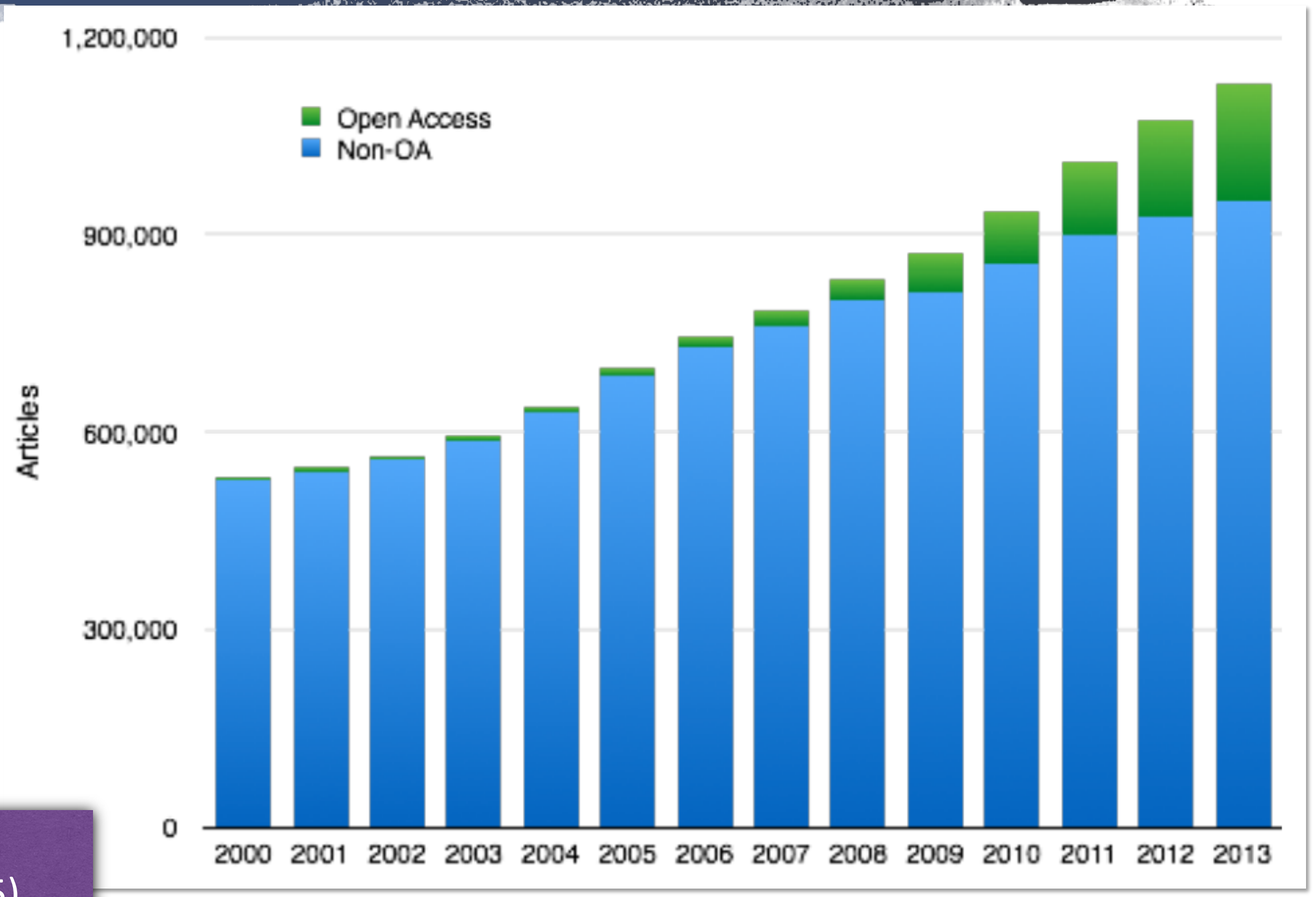
Plan of the talk

- Motivation
- Proposed architecture
- Experiments and evaluation
- Conclusion

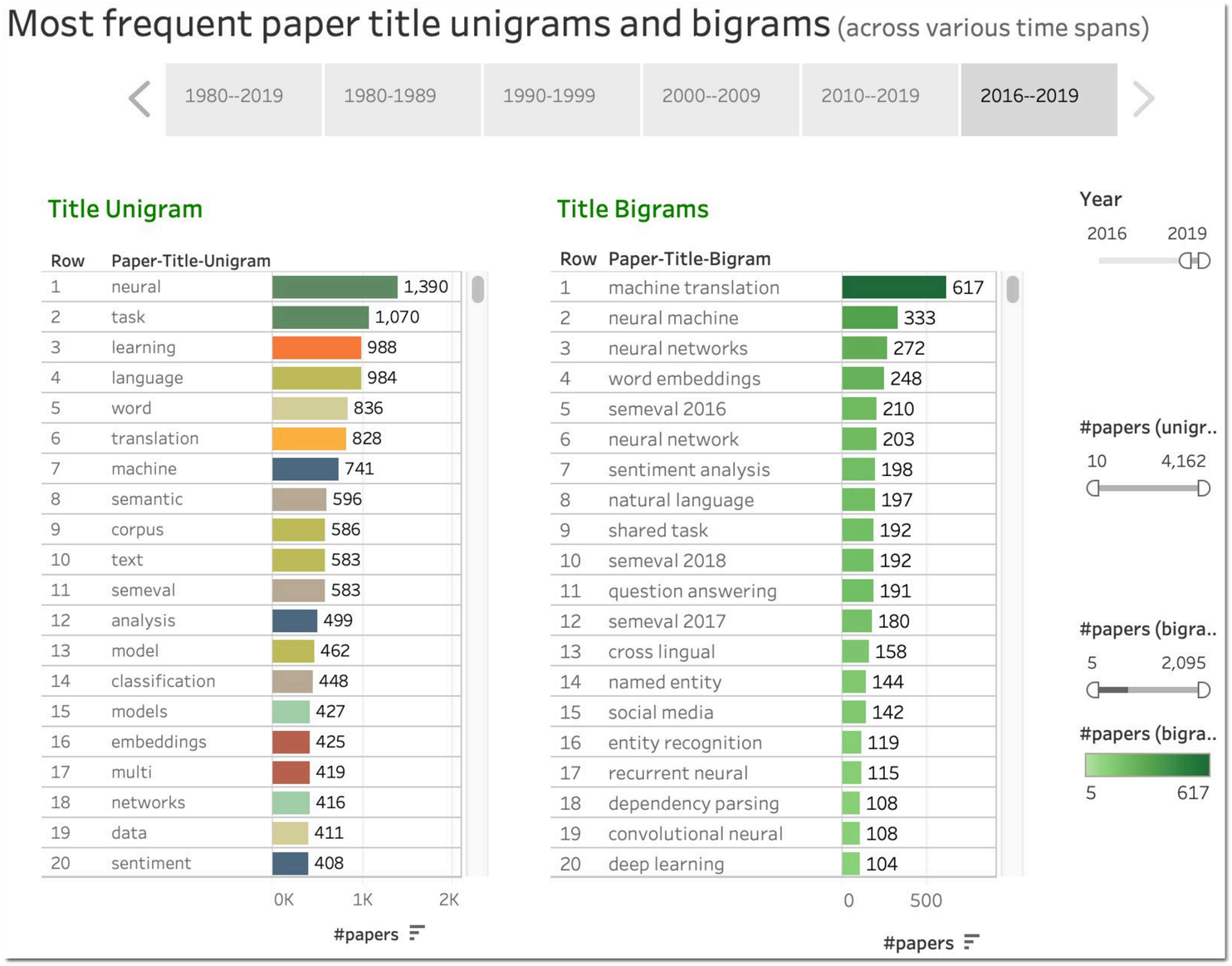
- How difficult it is to …
  - compile a state of the art about a given topic?
  - assess the novelty of a work?
  - find experts to review some papers?

Publication deluge

- The production of scientific literature is growing at an increased pace
  - More conferences, venues, journals, etc.
  - New publishing paradigms that accelerate the process (ex: Open Access)
- Also: new research topics are continuously created



PubMed central growth, 2000-2013
(Ware and Mabe, The STM report 2015)



Elaboration by Saif M. Mohammad on ACL Anthology data

Some (old) problems

Researchers may use the same word to refer to different topics

- Ex: different research domain
  - *lattice* in Physics vs. Computer Science
- Ex: different research communities within the same domain
  - *queries* in relational databases vs. IR systems

Or they may use different words to refer to the same topic

- Ex: reaching the same result in parallel
  - Meucci's "*teletrofono*" vs. G.Bell's *telephone*

Sheng Liang, Paul Hudak, Michael A. Jones • POPL • 1995

We show how a set of building blocks can be used to construct programming language interpreters, and present implementations of such building blocks capable of supporting many commonly known… (More)

56 19 View on ACM Cite Save

Programming Languages

Probabilistic Predicate **Transformers**

Carroll Morgan, Annabelle McIver, Karen Seidel • ACM Trans. Program. Lang. Syst. • 1996

Probabilistic predicates generalize standard predicates over a state space; with probabilistic predicate **transformers** one thus reasons about imperative programs in terms of probabilistic pre- and… (More)

18 10 View on ACM Cite Save

Machine Learning

Universal **Transformers**

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Lukasz Kaiser • ICLR • 2019

Recurrent neural networks (RNNs) sequentially process data by updating their state with each new data point, and have long been the de facto choice for sequence modeling tasks. However, their… (More)

5 23 View PDF on ArXiv Cite Save

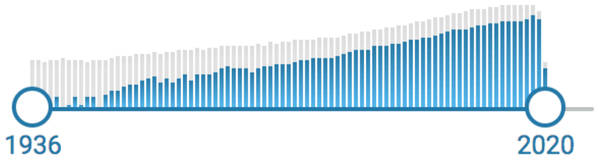
Extensible effects: an alternative to monad **transformers**

Oleg Kiselyov, Amr Sabry, Cameron Swords • Haskell • 2013

We design and implement a library that solves the long-standing problem of combining effects without imposing restrictions on their interactions (such as static ordering). Effects arise from… (More)

6 7 View on ACM View Slides Cite Save

Results by year



This year | Last 5 years | Last 10 years

Transformer

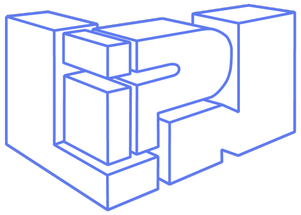
A transformer is an electrical device that transfers electrical energy between two or more circuits through electromagnetic induction. Electromagnetic induction produces… EXPAND

WATCH TOPIC | EXPLORE TOPIC ›

Electrical Engineering

50 Related topics

- Balanced audio
- Balanced line



Google Scholar toponym disambiguation in information retrieval

Articles **Place name** **Resolution**

Date indifférente
Depuis 2018
Depuis 2017
Depuis 2014
Période spécifique...

Trier par pertinence
Trier par date

Toutes les langues
Rechercher les pages en Français

inclure les brevets
inclure les citations

Créer l'alerte

Conseil : Recherchez des résultats uniquement en **Français**. Vous pouvez indiquer votre langue de recherche sur

Toponym disambiguation in **information retrieval**
D Buscaldi - 2010 - riunet.upv.es
In recent years, geography has acquired a great importance in the context of Information Retrieval (IR) and, in general, of the automated processing of information in text. Mobile devices that are able to surf the web and at the same time inform about their position are …
☆ Cité 12 fois Autres articles Les 17 versions

Geographical **information retrieval**
CB Jones, RS Purves - Encyclopedia of Database Systems, 2009 - Springer
… Synonyms Place names; **Toponyms**; Knowledge organization sys- tems; Ontologies … PostGIS Geographic **Information** System relies upon the PostgreSQL GiST implementation for its spatial … GiST-based indexes have been used for applications in image **retrieval**, astronomy data …
☆ Cité 198 fois Autres articles Les 12 versions

Approaches to **disambiguating toponyms**
D Buscaldi - Sigspatial Special, 2011 - dl.acm.org
… In [6], the objective was to **disambiguate toponyms** in a local Italian newspa- per, where the granularity of **toponyms** was intended to be at the level of street … However, there is still room for further work: **toponym disambiguation** may be used to tag ambiguous **toponyms** in Web …
☆ Cité 46 fois Autres articles Les 14 versions

A conceptual density-based approach for the **disambiguation** of **toponyms**
D Buscaldi, P Rosso - … Journal of Geographical **Information** Science, 2008 - Taylor & Francis
… 7. Conclusions and future work **Toponym disambiguation** is an open problem in GIR … The obtained results expose the limits of both WordNet as a resource for the **disambiguation** of **toponyms** and of GeoSemCor as a resource for the testing of **disambiguation** methods …
☆ Cité 89 fois Autres articles Les 5 versions

etrieval

hesis

t en **Français**. Vous pouvez indiquer votre langue de recherche sur

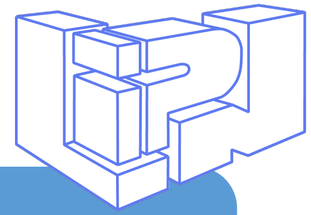
l
ase Systems, 2009 - Springer
gazetteers must support multiple **names** and, ideally …
erspective into how regulation takes **place**, in
ebi.ac.uk/ microarray-as/aer/ GEO http://www …
2 versions

egions for spatial **information retrieval**
e on Spatial **Information** …, 2003 - Springer
imations of **place name** regions … coordinates, there
-referenced through **place names** … important aspects
tic Web are the **resolution** of indi …
versions

and spatial browsing
n [papers presented at …, 1996 - ideals.illinois.edu
aphic **Information** Re- trieval (GIR) in the context …
ndexing and **retrieval** methods appropriate … commonly
Head- ings and **Name** Authorities as …
versions

Name resolution in a directory database

Not only Information Retrieval

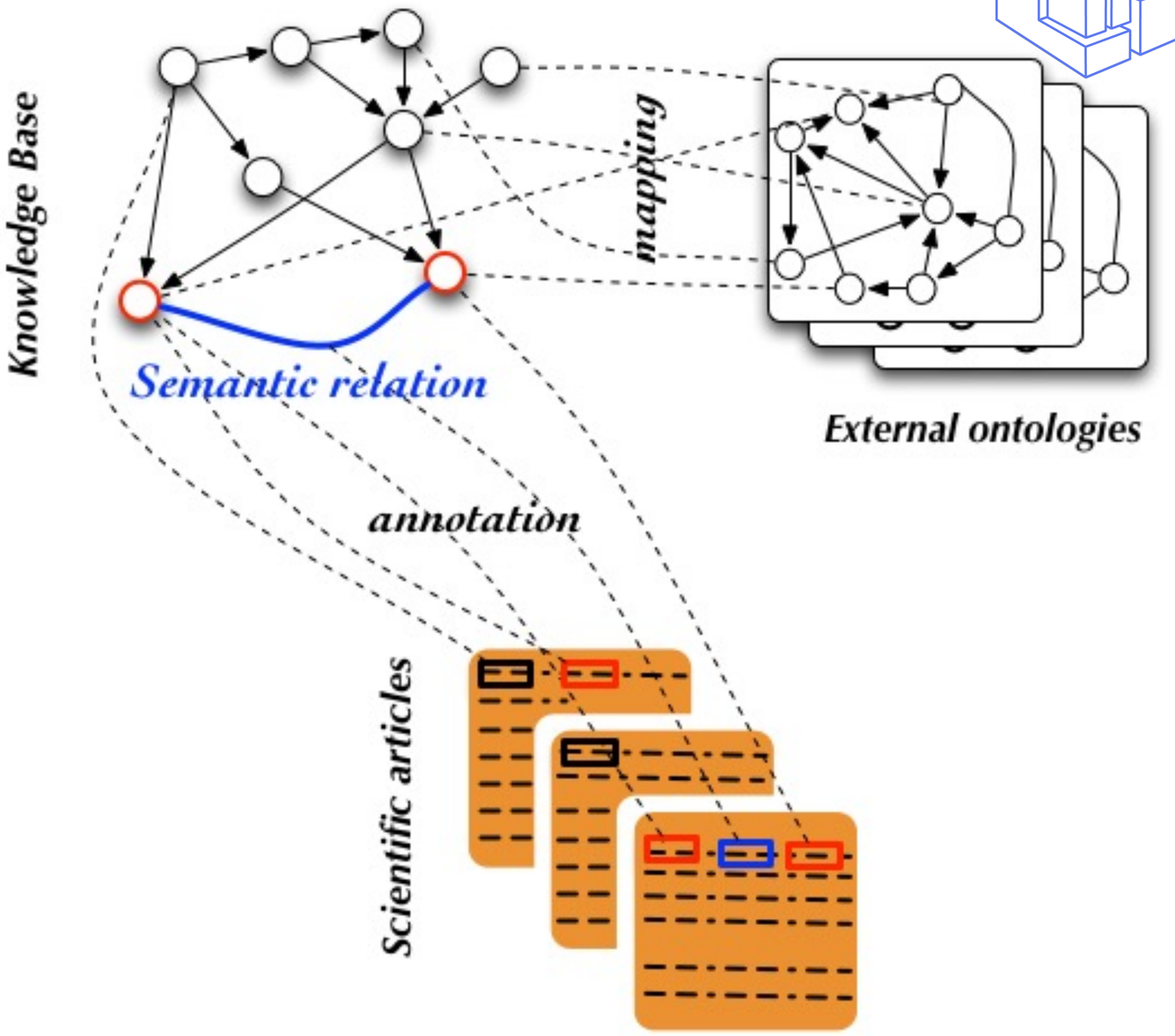


Finding the scientific works related to a given topic is just a part of the problem

- Which works are the most important?
- Which works are the most innovative?
- Who are the leading experts for a given topic?
- What are the emerging topics in a research domain?
- Which methods have been applied to a particular task?

Dedicated search engines currently offer only a partial solution

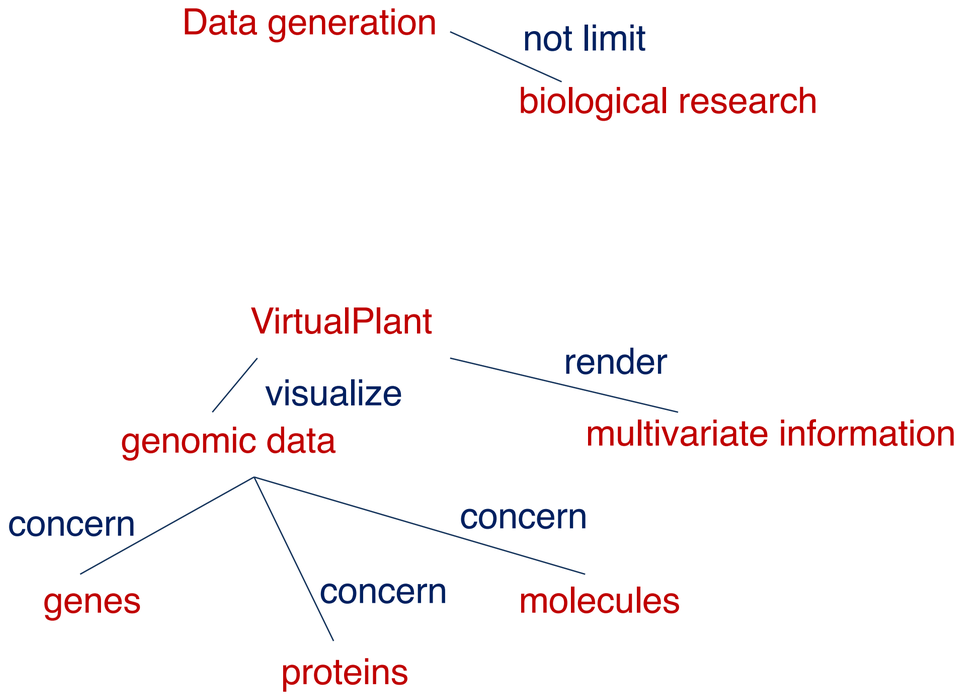
Scientific Knowledge Graphs



Scientific Knowledge Graphs: an example

- Mining textual contents of research papers for build a SKG that represent the underlying knowledge

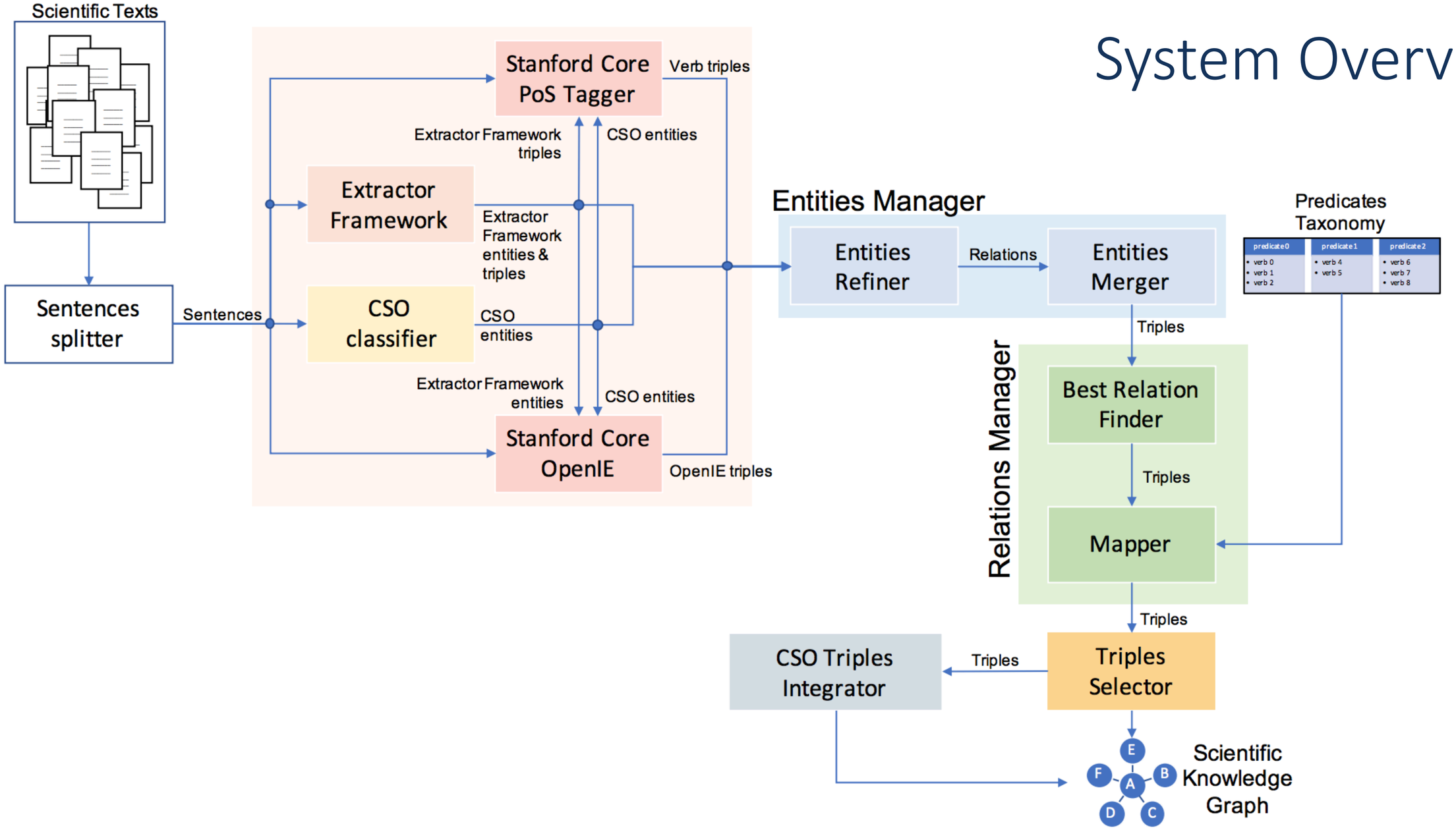
Data generation is no longer the limiting factor in advancing biological research. In addition, data integration, analysis, and interpretation have become key bottlenecks and challenges that biologists conducting genomic research face daily. To enable biologists to derive testable hypotheses from the increasing amount of genomic data, we have developed the VirtualPlant software platform. VirtualPlant enables scientists to visualize, integrate, and analyze genomic data from a systems biology perspective. VirtualPlant integrates genome-wide data concerning the known and predicted relationships among genes, proteins, and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides visualization techniques that render multivariate information in visual formats that facilitate the extraction of biological concepts. Importantly, VirtualPlant helps biologists who are not trained in computer …



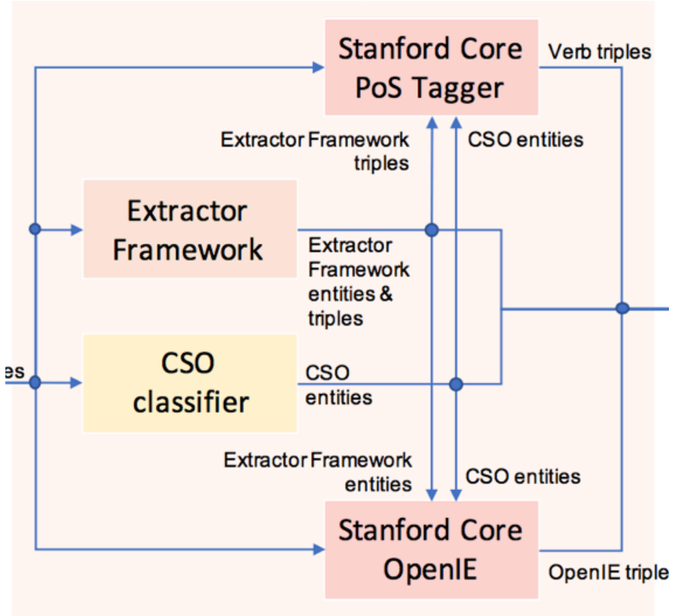
Our proposed architecture

- Formally, given a set of **D = {d_1, . . . , d_n}** scientific documents, we build a model **γ : D → T**, where **T** is a set of triples **(s, p, o)**
  - **s** and **o** belong to a set of entities **E** and **p** belongs to a set of relation labels **L**
- Our framework includes the following steps:
  - 1. *Extraction of entities and triples*, combining various tools
  - 2. *Entity refining*, in which the resulting entities are merged
  - 3. *Triple refining*, in which the triples extracted by the different tools are merged together and the relations are mapped to a common vocabulary
  - 4. *Triple selection*, in which we select the set of «trusted» triples that will be included in the final output

System Overview



Entity + Triple Extraction



Extractor Framework: neural model [Luan Yi et al.] based on SemEval-2018 task 7 winner system

- Six types of entities (Task, Method, Metric, Material, General Scientific, Other)
- Seven types of relations (Compare, Part-of, Conjunction, Evaluate-for, Feature-of, Used-for, Hyponym-Of)
- Problem: precise but the coverage is limited

Compensated by including more extractors into the process:

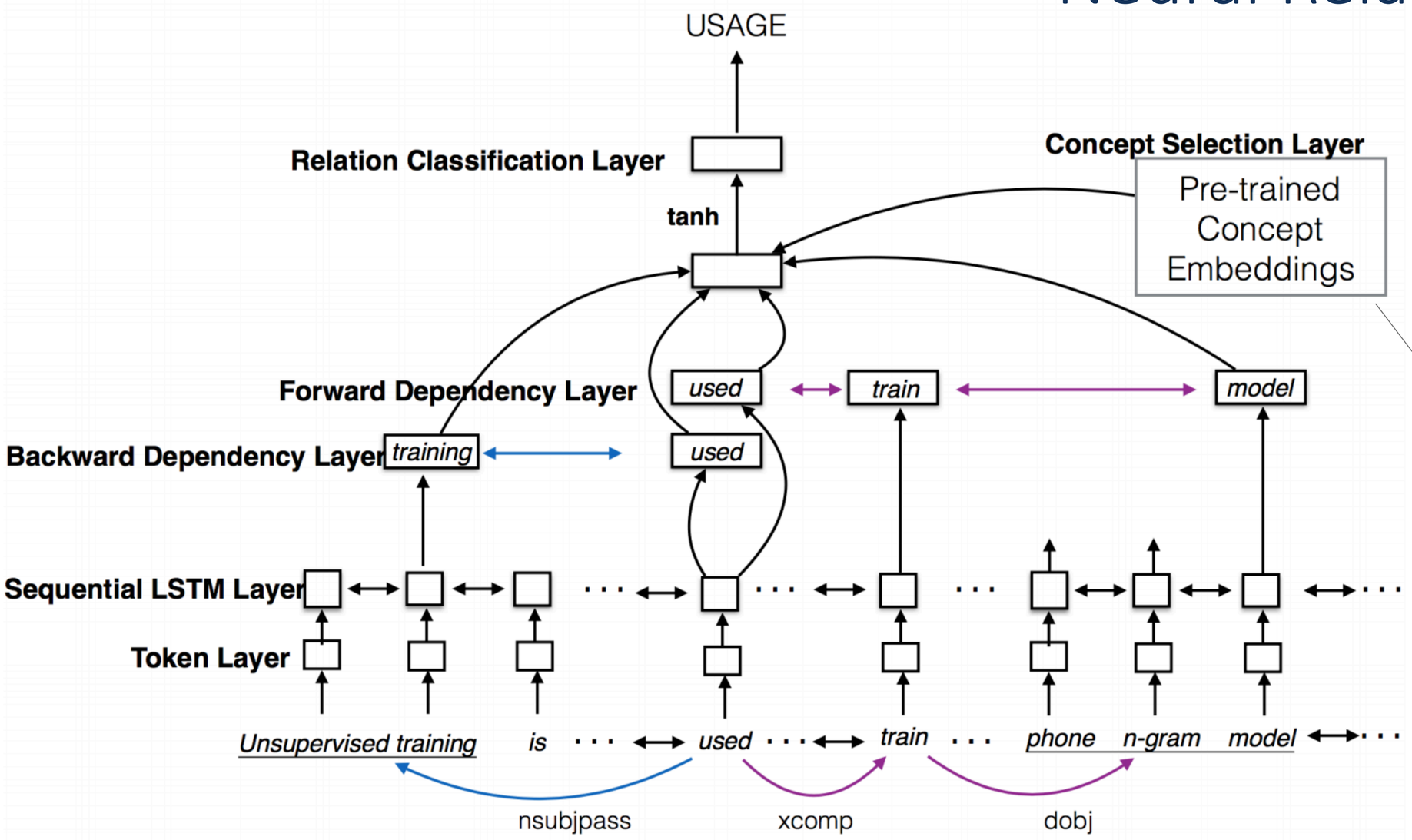
- OpenIE (Stanford Core NLP) extracts open-domain relationships between any noun phrase
- CSO Classifier is a keyword extractor for classifying research papers in the CS domain
  - CSO is paired with the Stanford NLP POS tagger to extract verbs between two keywords extracted by CSO Classifier

Examples of entities:
Semantic Web
knowledge acquisition method
theoretical learning
ODESeW

Examples of extracted relations:
ODESeW, develop, ontology-base portals (OpenIE)
Semantic e-Learning, enable, intelligent operations (OpenIE)
theoretical learning, used-for, learning processes (Deep Learning Extractor)
machine readable information, part-of, Semantic Web (Deep Learning Extractor)

Neural Relation Extractor

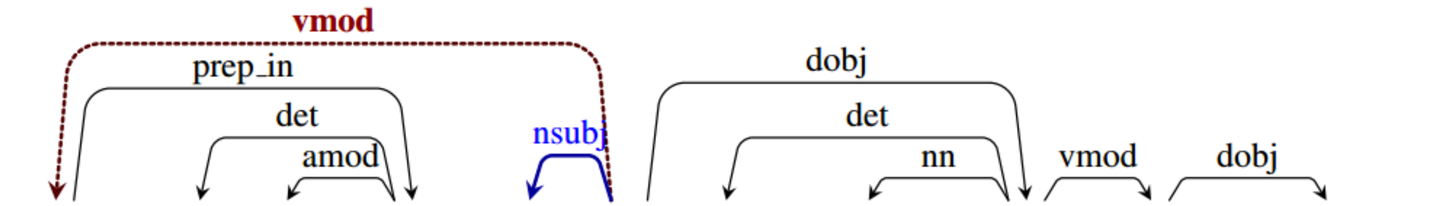
Luan Yi, UW (2018)



ELMo Embeddings

https://allennlp.org/elmo

OpenIE



(input)

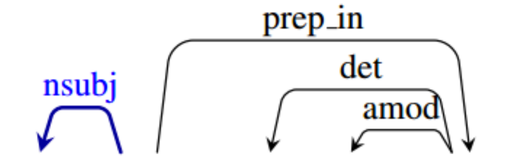
↓

she took the midnight train going anywhere
Born in a small town, she took the midnight train
Born in a town, she took the midnight train

she took the midnight train
she took midnight train
. . .

↓

(she; took; midnight train)



(extracted clause)

↓

she Born in small town
she Born in a town
she Born in town

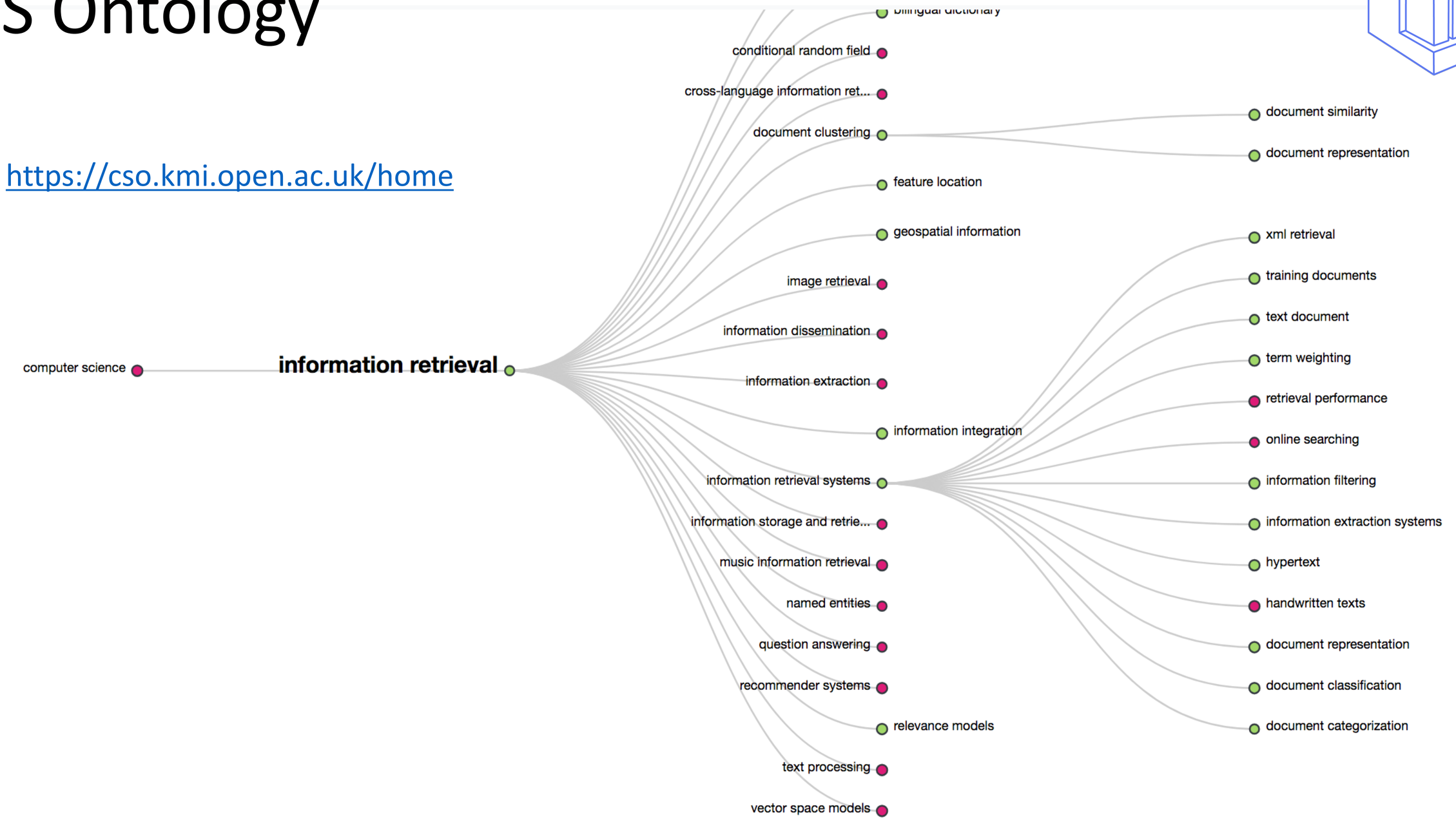
↓

(she; born in; small town)
(she; born in; town)

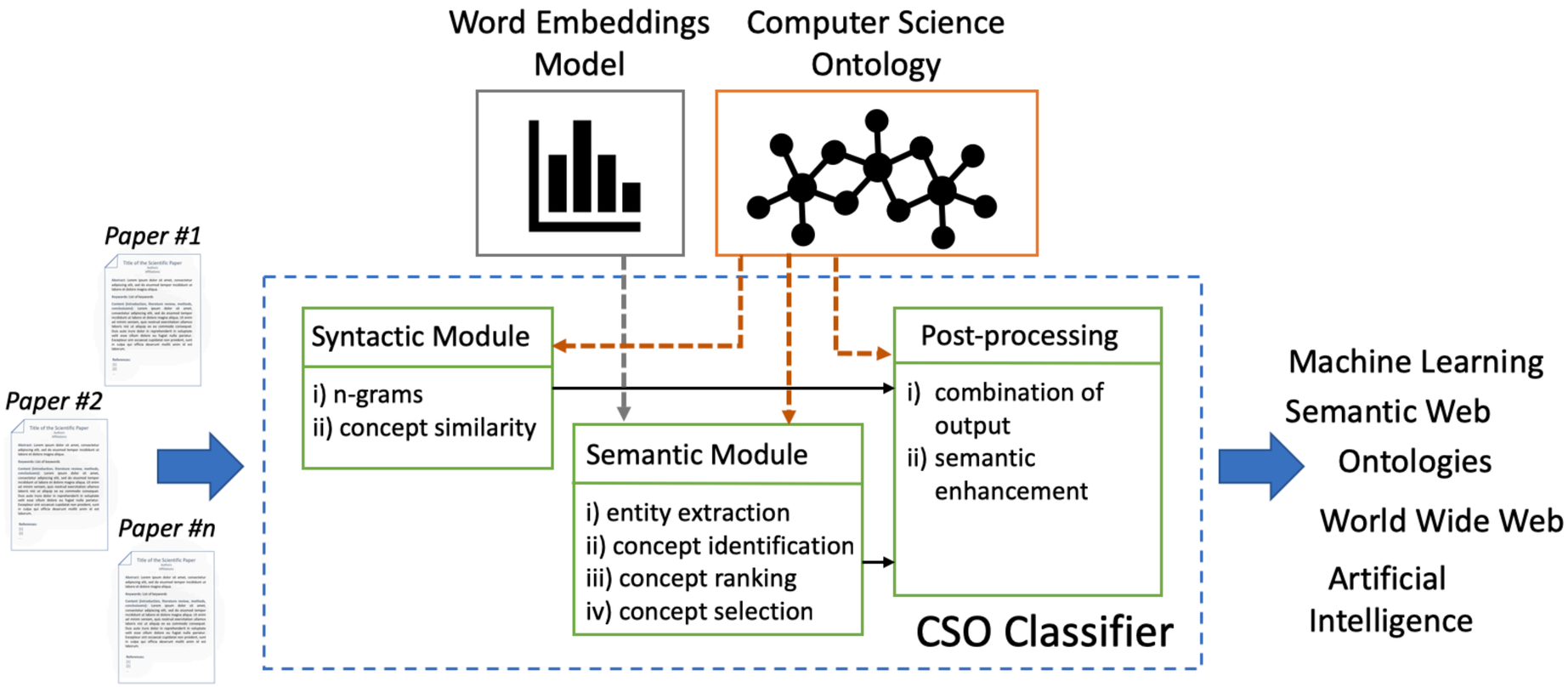
https://nlp.stanford.edu/software/openie.html

CS Ontology

https://cso.kmi.open.ac.uk/home

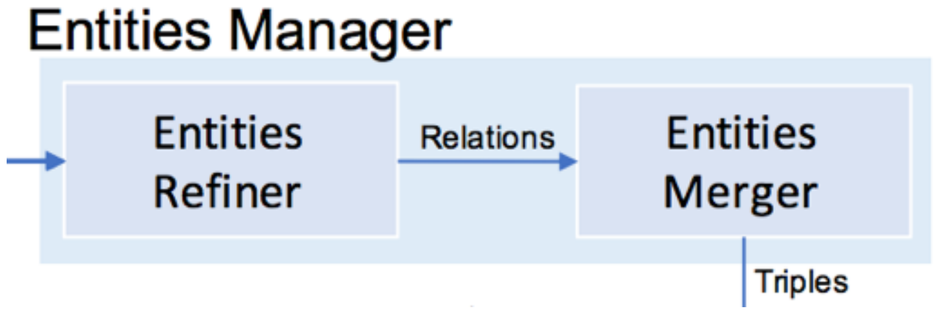


CSO Classifier



https://cso.kmi.open.ac.uk/classify/

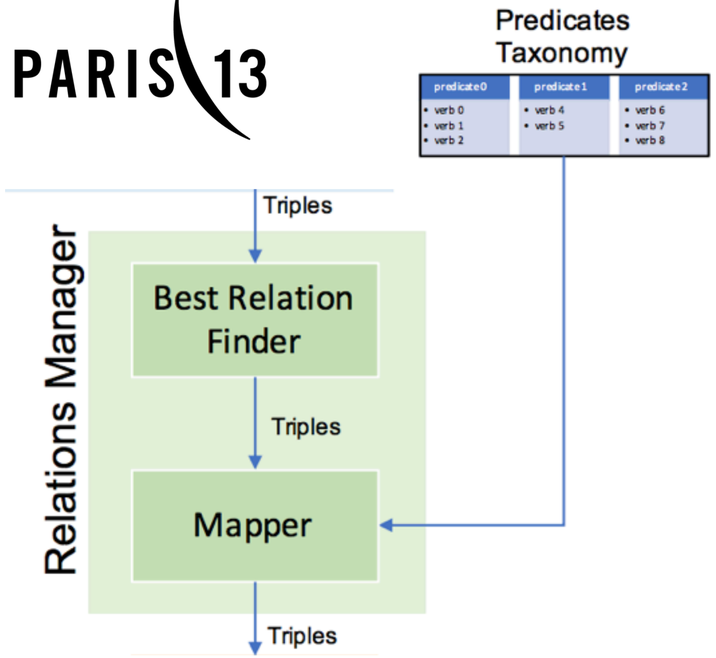
Entity Refining and Merging



- Two or more entities may refer to the same concept for various reasons
- Refining:
  - Acronym detection, punctuation and spaces, lemmatization…
  - Removing “generic” entities based on domain frequency
  - Splitting long entities (if ”and” present
- Merge similar meanings (word embeddings) and synonyms (CSO Ontology)
  - If the cosine similarity is over a given threshold (0.85) the two entities are mapped on the same one

Example of entity merging:
knowledge acquisition method
knowledge acquisition approach → *knowledge acquisition method*

Triples Normalization and Merging



- The same relation may be expressed by different triples depending on the verb connecting the entities
- A predicates taxonomy has been built using word embeddings on the Microsoft Academic Graph and hierarchical clustering, for all possible relation labels

semantic web, combine, markup language *combine* → *w1*

semantic web, analyze, markup language *analyze* → *w2*

semantic web, extend, markup language *extend* → *w3*

semantic web, combine, markup language *combine* → *w1*

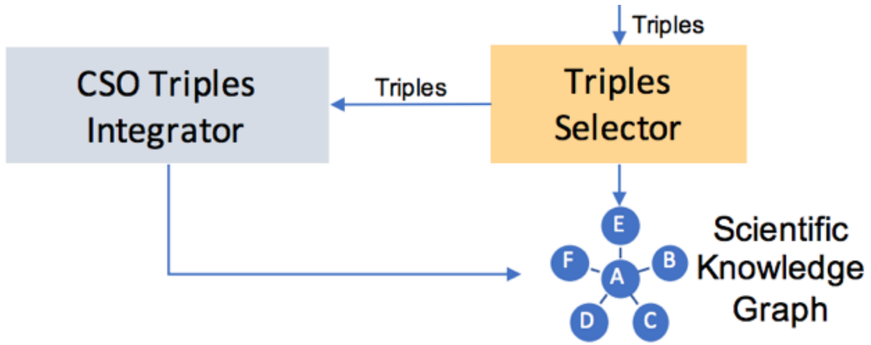
semantic web, use, markup language *use* → *w4*

semantic web, analyze, markup language *analyze* → *w2*

- Every relation is associated with the corresponding word embedding of the verb expressing the relation
- The relation with the word embeddings nearest to the average of all word embeddings is chosen as the most representative relation

W = AVG(w1, w2, w3, w1, w4, w2)
If w1 is the nearest to W the chosen relation will be:
semantic web, combine, markup language

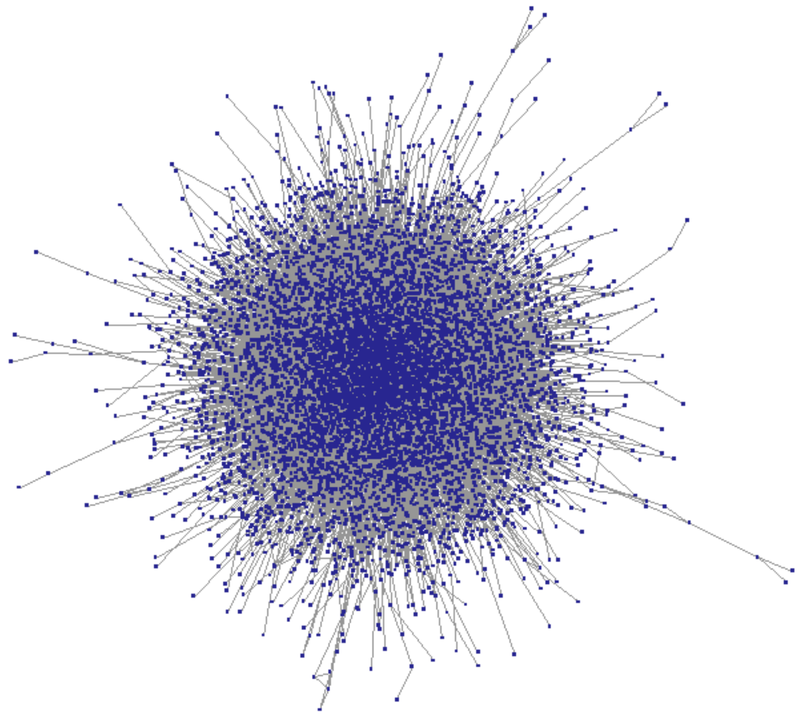
Triples Selection and Enrichment



- All triples coming from the Neural Extractor and OpenIE after the normalization and merging phases are kept
- Triples coming from CSO + « linking » verbs are kept only if their support is large enough (at least seen in 10 documents)

- CSO Triples Integrator: it includes some triples derived with limited inference on CSO:
- given a triple (e1, r, e2), if in CSO the entity e3 is *superTopicOf* of e1, we also infer the triple (e3, r, e2).
- For instance, given the triple («NLP systems», «use», «Dbpedia»)
  - If *hyp*(«Semantic Web Technologies», «Dbpedia»), then we can infer the triple («NLP systems», «use», «Semantic Web Technologies»)

Experiments



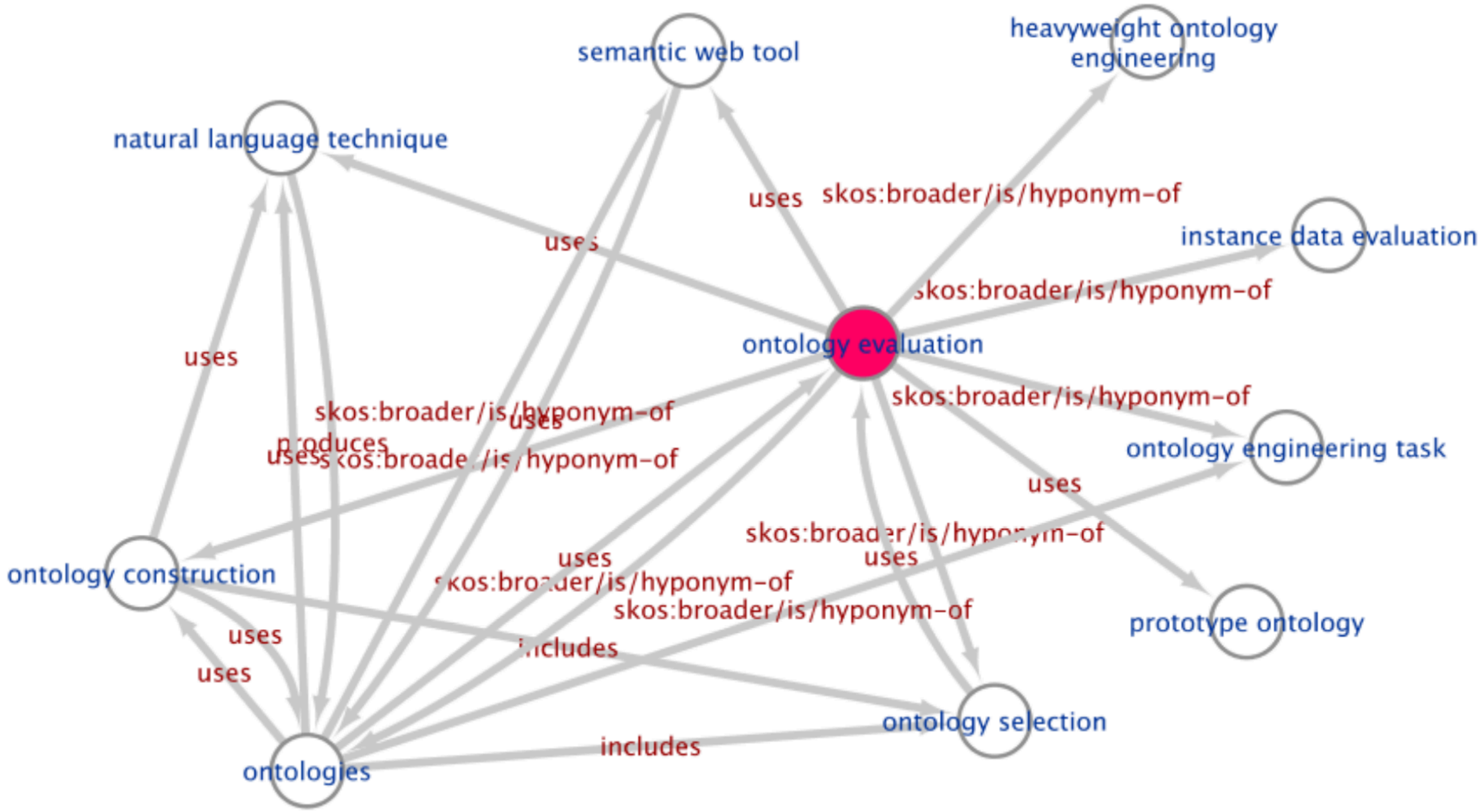
- 26,827 abstracts about the Semantic Web domain from Microsoft Academic Graph

- The resulting knowledge graph contains 109,105 triples
  - 87,030 from the Extractor Framework (TEF)
  - 8,060 from OpenIE (TOIE)
  - 14, 015 from the CSO + linking verbs method (TPoS).

Example of extracted triples

semantic web	extends	WWW
semantic web	supports	information integration
semantic web	relies on	ontology
ontology	provides	semantics
ontology	represents	domain knowledge
RDF	is	data model
SPARQL	is	query language
OWL	is	ontology language
OWL	employs	open world assumption
W3C	recommends	ontology
Protégé	is	ontology editor
Protégé	creates	OWL ontology
semantic web services	extends	web services

Examples of extracted sub-graphs



Evaluation

- We built a Gold Standard composed of 818 triples
  - 401 triples from the Neural Extractor
  - 102 from Stanford Open IE
  - 170 triples from the CSO + linking verbs method
  - 212 randomly selected triples that were discarded by the framework pipeline.

- Five experts in the field of Semantic Web annotated the triples as correct (that is, if they make sense when compared with their knowledge on the subject) or not.
  - The agreement between experts was 0.747 ± 0.036

Evaluation

Approach	Precision	Recall	F-Score
Neural Extractor	0,72	0,55	0,62
OpenIE	0,65	0,13	0,21
CSO + linking verbs	0,73	0,24	0,36
Neural + OpenIE	0,70	0,66	0,68
Full Pipeline	0,70	0,81	0,75

Conclusions

- Trivial knowledge (high-level entities with general relations) either:
  - Does not appear (so trivial that it is not necessary to include into the paper)
  - Appears too often (outweigh all other relationships)
- In any case most (useful) relations appear only once
- Need to introduce a confidence weight over the knowledge when we extract it
  - Is the sentence complex or simple?
  - Is the paragraph in the introduction/conclusions?
- Some parameters may be specific to the considered domain
  - Need more experiments!

Acknowledgements

- Diego Reforgiato, Ass.Prof. - Università di Cagliari
- Danilo Dessì, PhD Student - Università di Cagliari
- Francesco Osborne, Research Fellow - Knowledge Media Institute, The Open University, Milton Keynes