Knowledge Graph Construction, Curation and Its Machine Learning Applications

Jiaoyan Chen

Department of Computer Science, The University of Manchester

MANCHESTER
1824
The University of Manchester

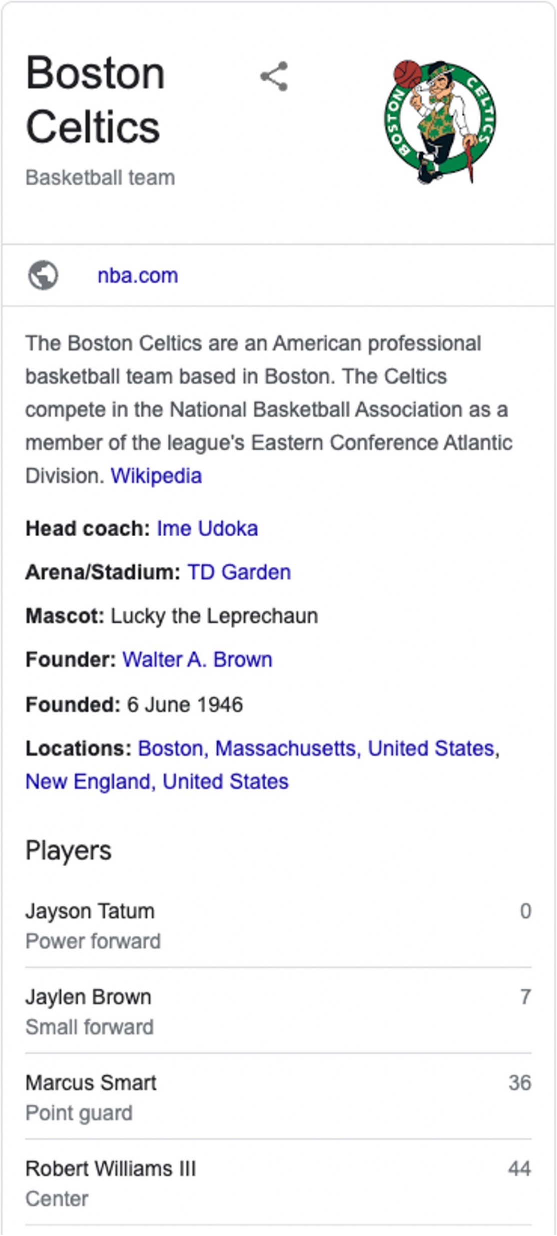
Outline

- **Background**
  - Knowledge Graph and Semantic Web
- **Knowledge Graph Construction and Curation**
  - Semantic Table Annotation
  - Knowledge Integration
  - Knowledge Refinement
  - Ontology Embedding
- **Knowledge Graph Applications**
  - Zero-shot Learning

The Term of "Knowledge Graph"

- The Knowledge Graph is a knowledge base used by **Google** and its services to enhance its search engine's results with knowledge gathered from a variety of sources.
  - Proposed around 2012
  - Knowledge ≈ **Instances + Facts**
  - KG ≈ **Linked Structured Data** (can be regarded as a multi-relational graph)

example



Boston Celtics

Basketball team

nba.com

The Boston Celtics are an American professional basketball team based in Boston. The Celtics compete in the National Basketball Association as a member of the league's Eastern Conference Atlantic Division. Wikipedia

Head coach: Ime Udoka

Arena/Stadium: TD Garden

Mascot: Lucky the Leprechaun

Founder: Walter A. Brown

Founded: 6 June 1946

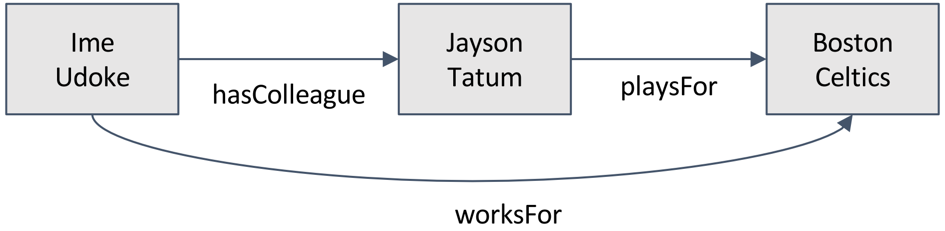
Locations: Boston, Massachusetts, United States, New England, United States

Players

Jayson Tatum Power forward	0
Jaylen Brown Small forward	7
Marcus Smart Point guard	36
Robert Williams III Center	44

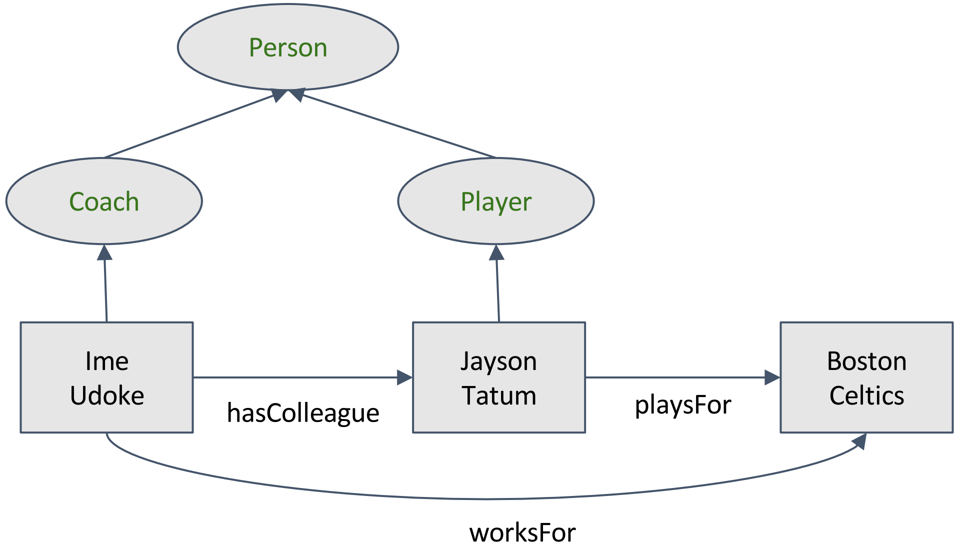
A Semantic Web Perspective

- **RDF** (Resource Description Framework)
  - Triple: <Subject, Predicate, Object>
  - Representing facts:
    - E.g., <Jayson Tatum, playsFor, Boston Celtics>



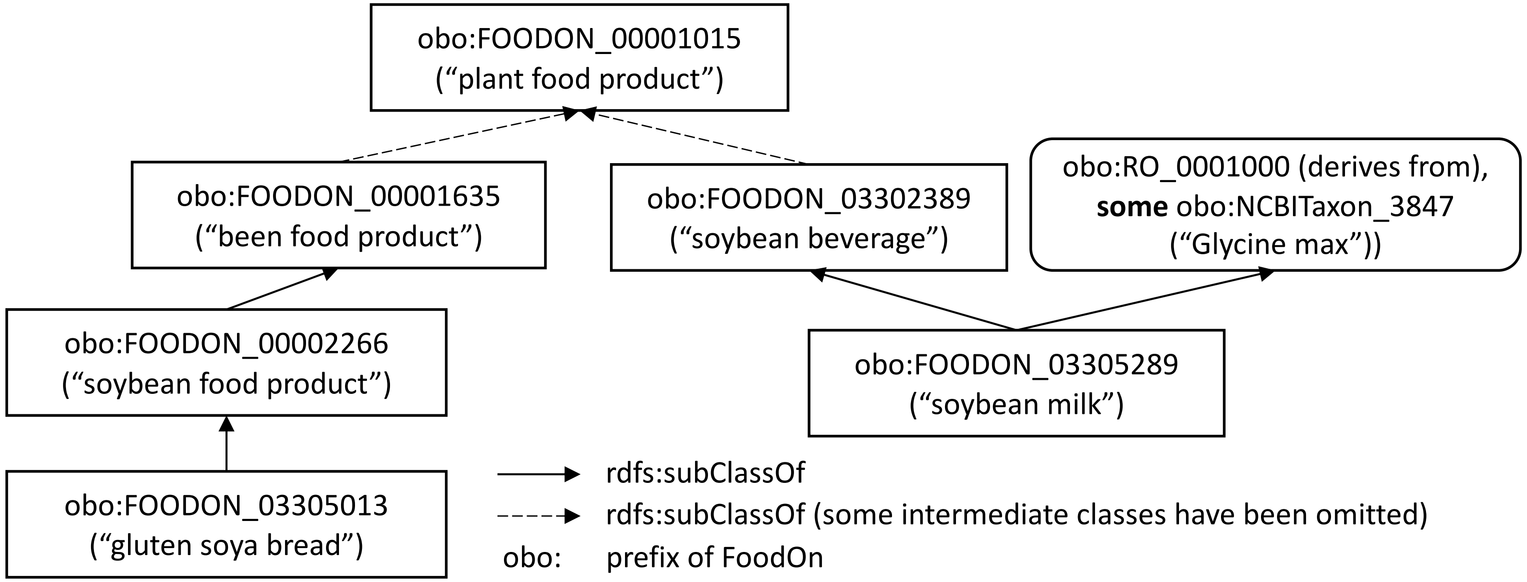
A Semantic Web Perspective

- RDF (Resource Description Framework)
  - Triple: <Subject, Predicate, Object>
  - Representing facts:
    - E.g., <Jayson Tatum, playsFor, Boston Celtics>
- **RDF Schema**
  - Meta data (schema) of instances and facts
    - E.g., class, property domain and range



A Semantic Web Perspective

- **Web Ontology Language** (OWL)
  - Schema, constraints and logical relationships
    - E.g., ‘food material’ ≡ ‘environmental material’ *and* (‘has role’ *some* ‘food’)
    - E.g., the cardinality of “playsFor” is 1
  - Taxonomies and vocabularies


obo:FOODON_00001015 (“plant food product”)
obo:FOODON_00001635 (“been food product”)
obo:FOODON_03302389 (“soybean beverage”)
obo:RO_0001000 (derives from), **some** obo:NCBITaxon_3847 (“Glycine max”))
obo:FOODON_00002266 (“soybean food product”)
obo:FOODON_03305289 (“soybean milk”)
obo:FOODON_03305013 (“gluten soya bread”)
rdfs:subClassOf
rdfs:subClassOf (some intermediate classes have been omitted)
obo: prefix of FoodOn

A segment of the food ontology FoodOn

What is KG?

RDF facts?

RDF facts + schema?

Ontology?

Why use a Knowledge Graph?

- ✓ Intuitive (e.g., no “foreign keys”)
- ✓ Data + schema (ontology)
- ✓ IRI/URI not strings
- ✓ Flexible & extensible
- ✓ Rule language
  - Location + capital → location
  - Parent + brother → uncle
- ✓ Other kinds of query
  - Navigation
  - Similarity & Locality

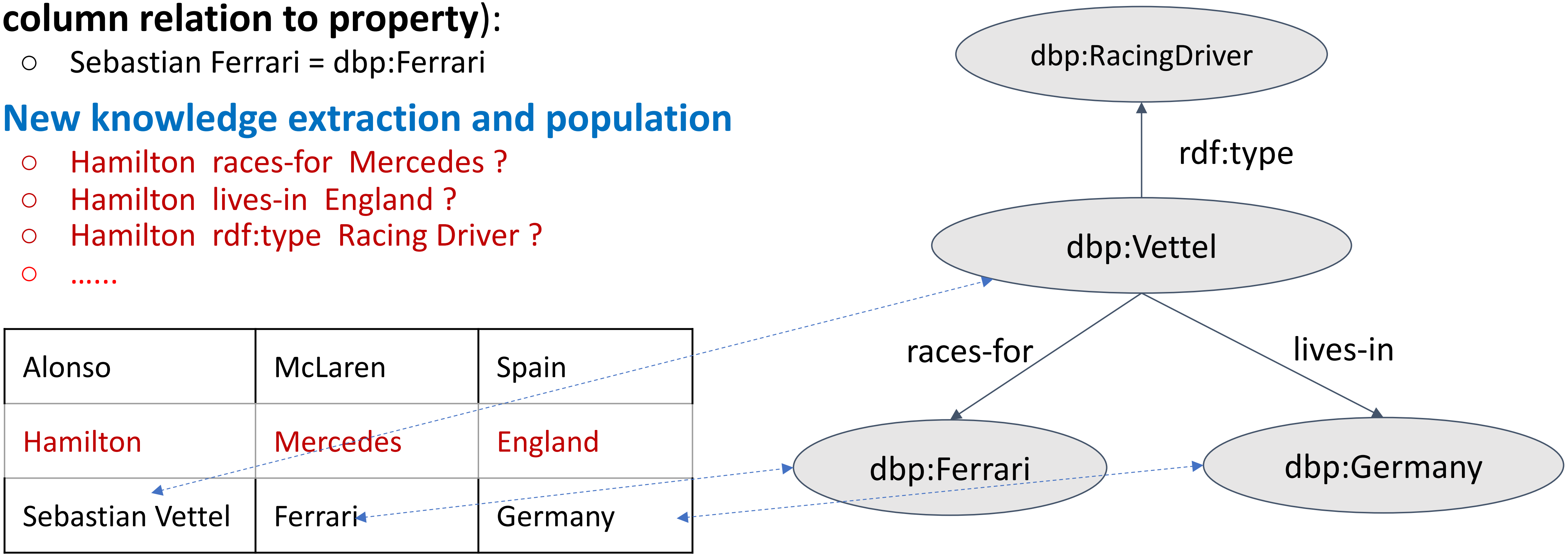
(This slide is from Prof. Ian Horrocks)

Knowledge Graph Construction

- **Crowdsourcing (Encyclopedias) & Domain Experts**
  - DBpedia, Wikidata, Zhishi.me (中文), LinkedGeoData, GeoName
- **The Web Mining**
  - NELL (Never-ending Language Learning)
- **Natural Language Text**
  - Open Information Extraction
- **Tabular data**
  - **DBs, Web Tables, Excel Sheets, CSV files, etc.**
- **KG Alignment (Integration)**

Table for KG Population (Example)

- **Table to KG matching (cell to entity, column type to class, inter-column relation to property):**
  - Sebastian Ferrari = dbp:Ferrari
- **New knowledge extraction and population**
  - Hamilton races-for Mercedes ?
  - Hamilton lives-in England ?
  - Hamilton rdf:type Racing Driver ?
  - ……

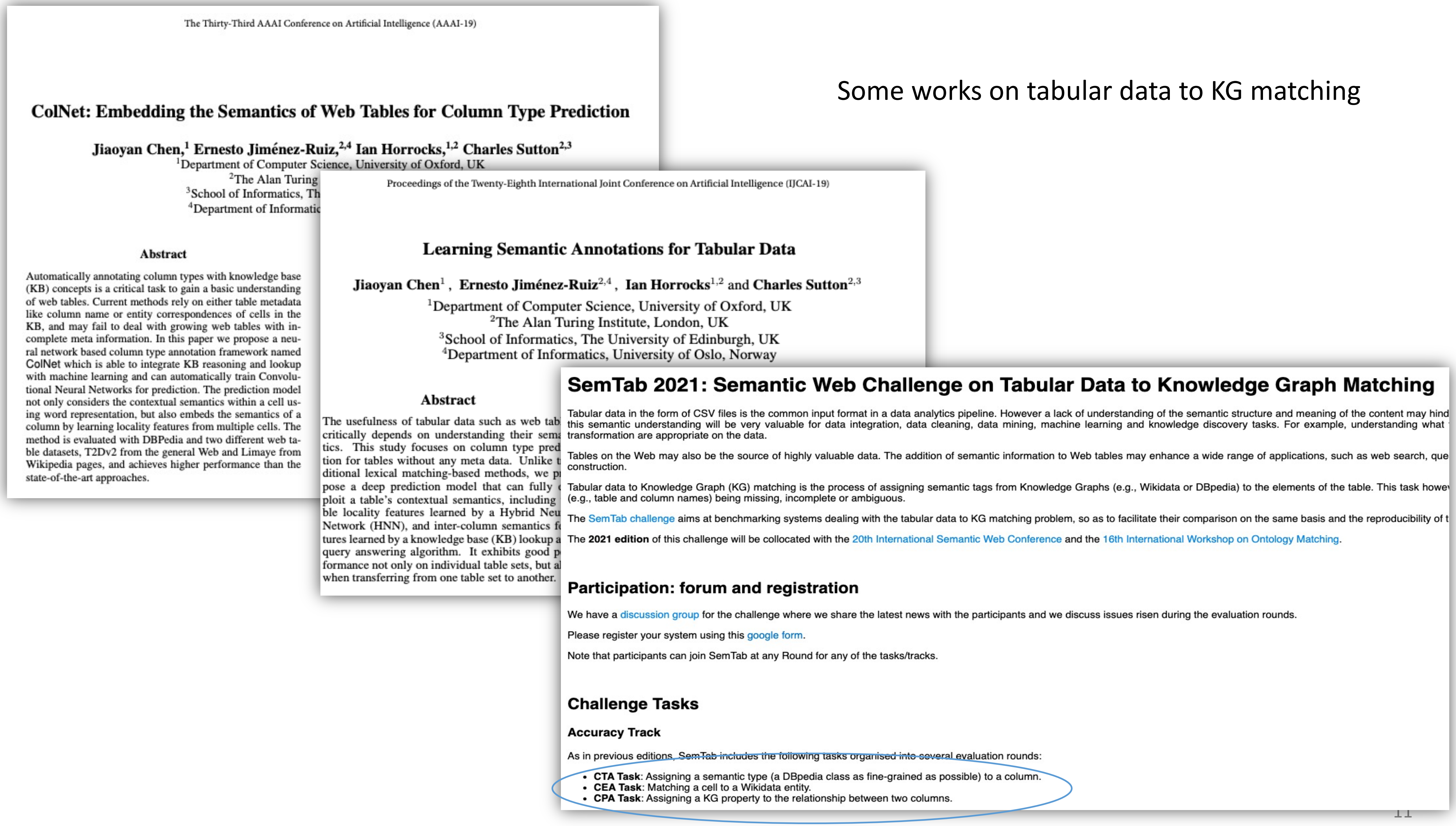


Alonso	McLaren	Spain
Hamilton	Mercedes	England
Sebastian Vettel	Ferrari	Germany

Table on F1

Existing KG on F1

Some works on tabular data to KG matching



The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

ColNet: Embedding the Semantics of Web Tables for Column Type Prediction

Jiaoyan Chen,[1] Ernesto Jiménez-Ruiz,[2,4] Ian Horrocks,[1,2] Charles Sutton[2,3]
[1]Department of Computer Science, University of Oxford, UK
[2]The Alan Turing
[3]School of Informatics, Th
[4]Department of Informatic

Abstract

Automatically annotating column types with knowledge base (KB) concepts is a critical task to gain a basic understanding of web tables. Current methods rely on either table metadata like column name or entity correspondences of cells in the KB, and may fail to deal with growing web tables with incomplete meta information. In this paper we propose a neural network based column type annotation framework named ColNet which is able to integrate KB reasoning and lookup with machine learning and can automatically train Convolutional Neural Networks for prediction. The prediction model not only considers the contextual semantics within a cell using word representation, but also embeds the semantics of a column by learning locality features from multiple cells. The method is evaluated with DBPedia and two different web table datasets, T2Dv2 from the general Web and Limaye from Wikipedia pages, and achieves higher performance than the state-of-the-art approaches.

Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Learning Semantic Annotations for Tabular Data

Jiaoyan Chen[1] , Ernesto Jiménez-Ruiz[2,4] , Ian Horrocks[1,2] and Charles Sutton[2,3]
[1]Department of Computer Science, University of Oxford, UK
[2]The Alan Turing Institute, London, UK
[3]School of Informatics, The University of Edinburgh, UK
[4]Department of Informatics, University of Oslo, Norway

Abstract

The usefulness of tabular data such as web tab
critically depends on understanding their sema
tics. This study focuses on column type pred
tion for tables without any meta data. Unlike t
ditional lexical matching-based methods, we pr
pose a deep prediction model that can fully e
ploit a table's contextual semantics, including
ble locality features learned by a Hybrid Neu
Network (HNN), and inter-column semantics fe
tures learned by a knowledge base (KB) lookup a
query answering algorithm. It exhibits good p
formance not only on individual table sets, but al
when transferring from one table set to another.

SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However a lack of understanding of the semantic structure and meaning of the content may hind
this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what
transformation are appropriate on the data.

Tables on the Web may also be the source of highly valuable data. The addition of semantic information to Web tables may enhance a wide range of applications, such as web search, que
construction.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the table. This task howe
(e.g., table and column names) being missing, incomplete or ambiguous.

The SemTab challenge aims at benchmarking systems dealing with the tabular data to KG matching problem, so as to facilitate their comparison on the same basis and the reproducibility of t

The **2021 edition** of this challenge will be collocated with the 20th International Semantic Web Conference and the 16th International Workshop on Ontology Matching.

Participation: forum and registration

We have a discussion group for the challenge where we share the latest news with the participants and we discuss issues risen during the evaluation rounds.

Please register your system using this google form.

Note that participants can join SemTab at any Round for any of the tasks/tracks.

Challenge Tasks

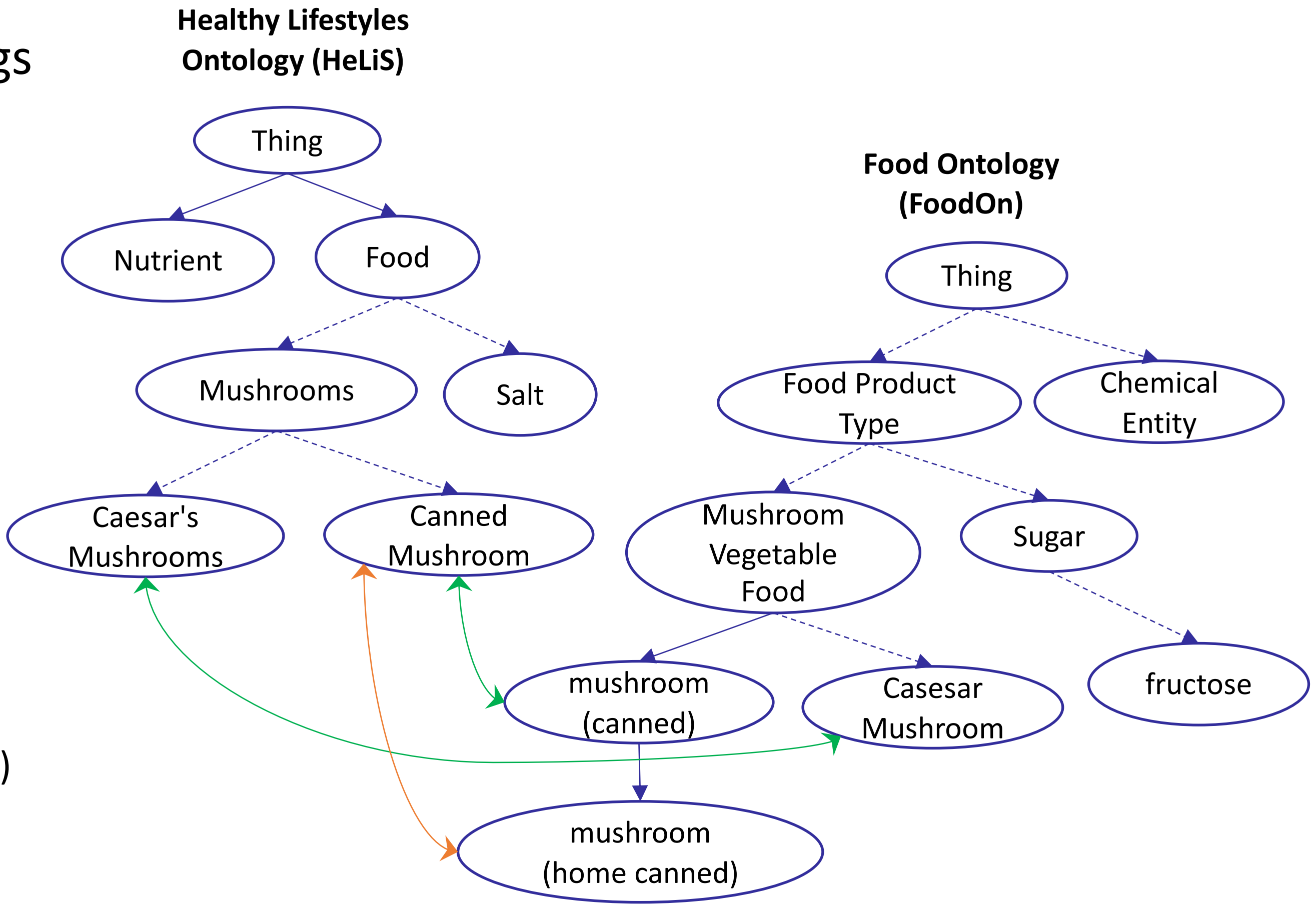
Accuracy Track

As in previous editions, SemTab includes the following tasks organised into several evaluation rounds:

- **CTA Task**: Assigning a semantic type (a DBpedia class as fine-grained as possible) to a column.
- **CEA Task**: Matching a cell to a Wikidata entity.
- **CPA Task**: Assigning a KG property to the relationship between two columns.

Ontology Alignment

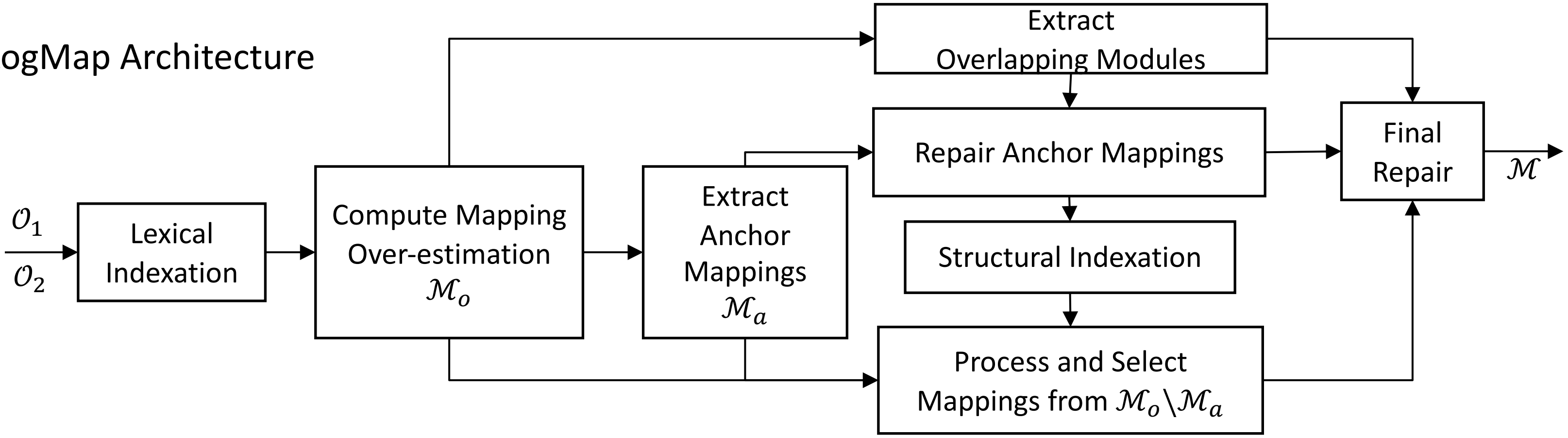
- Discover equivalence or subsumption mappings between classes across two ontologies (often taxonomies)
  - E.g., Canned Mushroom in HeLis vs mushroom (canned) in FoodOn
- Traditional Solutions
  - Lexical index, lexical matching
  - Structure matching
  - Logical reasoning (post-checking and repair)
- Emerging Solutions
  - Machine learning
    - Features (name similarity, neighbor similarity, etc.)
    - Embedding (deep learning)



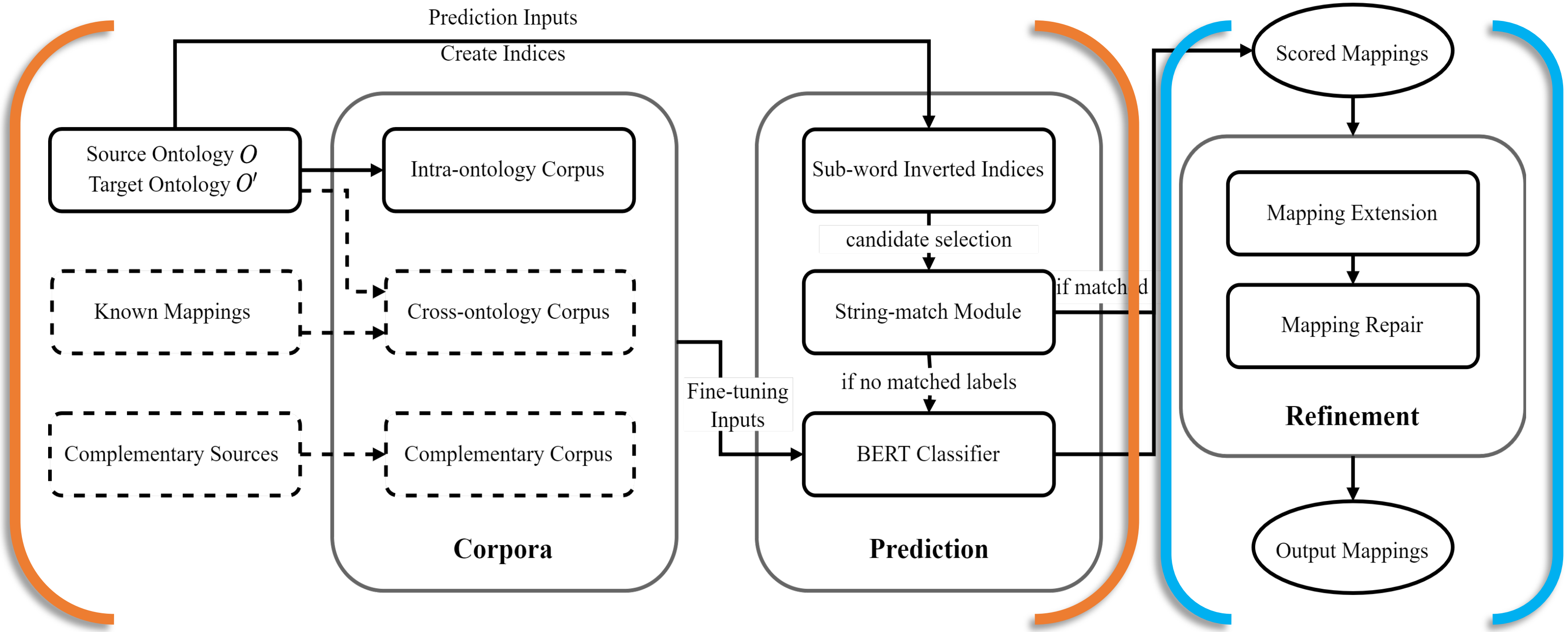
LogMap

- LogMap: lexical index/matching, structure matching, logics-based repair
- Online service: http://krrwebtools.cs.ox.ac.uk/logmap/
- Software: https://github.com/ernestojimenezruiz/logmap-matcher
- Extensions: LogMap-ML (ESWC’21), Distributed LogMap (ECAI’20), etc.

LogMap Architecture


$\mathcal{O}_1$
$\mathcal{O}_2$
Lexical Indexation
Compute Mapping Over-estimation $\mathcal{M}_o$
Extract Anchor Mappings $\mathcal{M}_a$
Extract Overlapping Modules
Repair Anchor Mappings
Structural Indexation
Process and Select Mappings from $\mathcal{M}_o \backslash \mathcal{M}_a$
Final Repair
$\mathcal{M}$

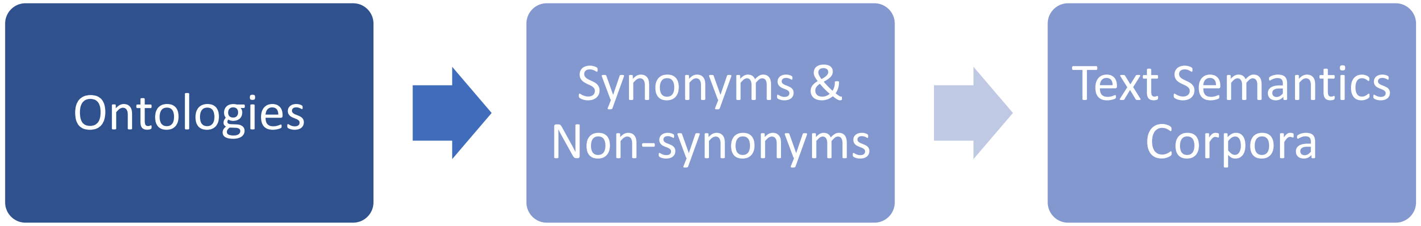
BERTMap (Framework)



Make predictions based on the ensemble results of class label (text) classification

Refine mappings through extension & repair

BERTMap (Corpora for Fine-tuning)



Synonyms: An ontology class could have multiple ***aliases*** defined by some annotational properties, e.g., ***rdfs:label***, ***oboInOwl:hasExactSynonym***.

Note: We considered both ***reversed*** and ***identity*** synonyms

Non-synonyms are retrieved from *either* label pairs of two random classes (***soft***) *or* label pairs of logically disjoint classes (***hard***).

BERTMap (Corpora for Fine-tuning)

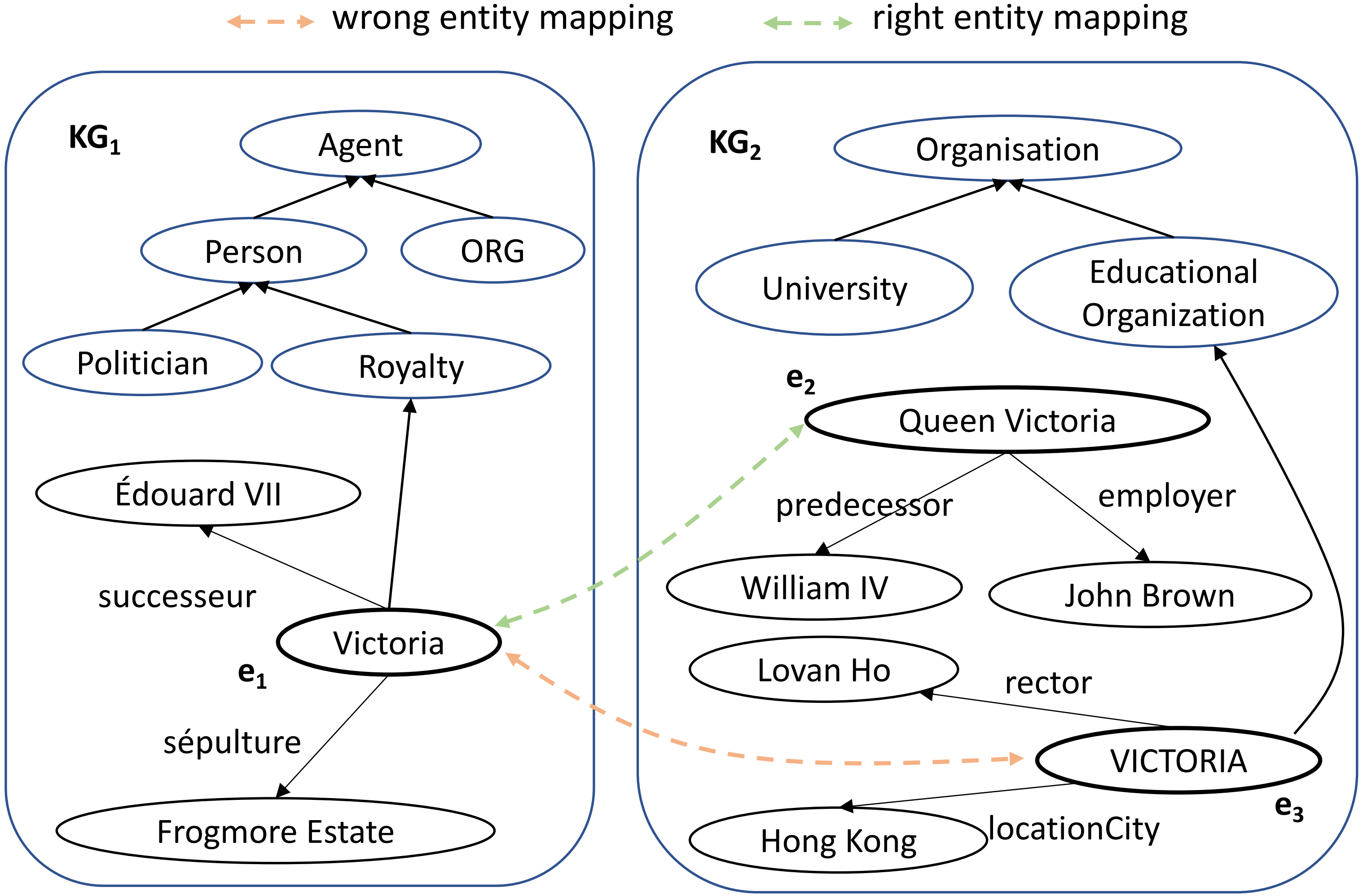
Intra-ontology (from within an ontology)

- As described in the previous slide.

Cross-ontology (from a small portion of given mappings)

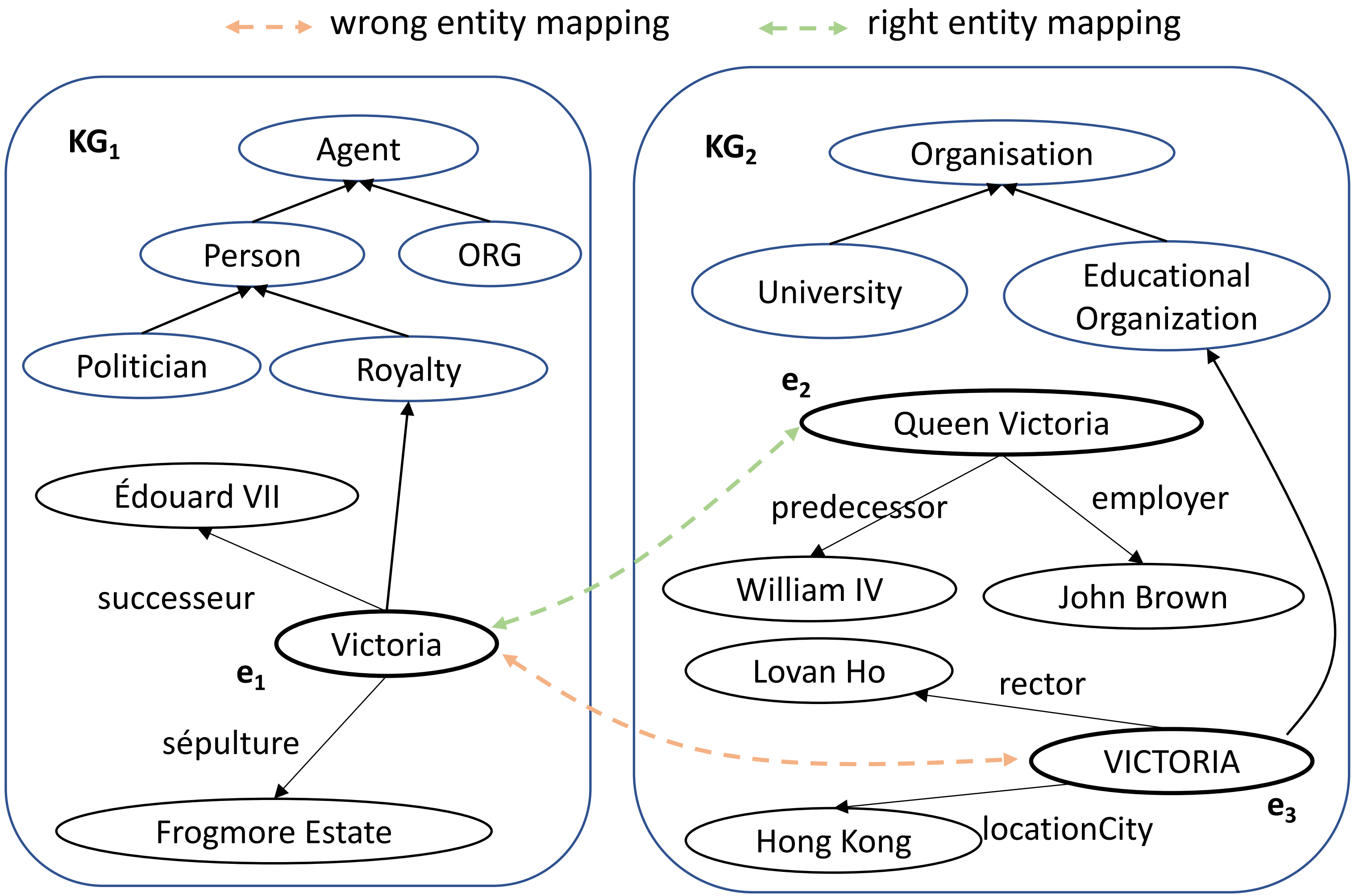
Entity Alignment

- KG entity alignment
  - Equivalent entities/instances across KGs composed of large-scale facts
  - E.g., Victoria vs Queen Victoria in the right figure
- Traditional solutions
  - Lexical index and matching
  - Structure matching
  - Machine learning feature engineering
- Deep learning solutions
  - Learning embeddings in the same vector space
  - See [Su, Zequn et al. VLDB'20] for a survey and benchmarking study



Entity Alignment

- Some recent works:
  - **OntoEA**: utilize the ontology to distinguish entity alignment [Findings of ACL’21]
  - **PARSE**: combing probabilistic reasoning with KG embeddings [IJCAI’21]
    - E.g., utilizing relation functionality to identify equivalent objects
  - **New industrial benchmarks and benchmarking study** with Tencent’s medical KGs [COLING’20]



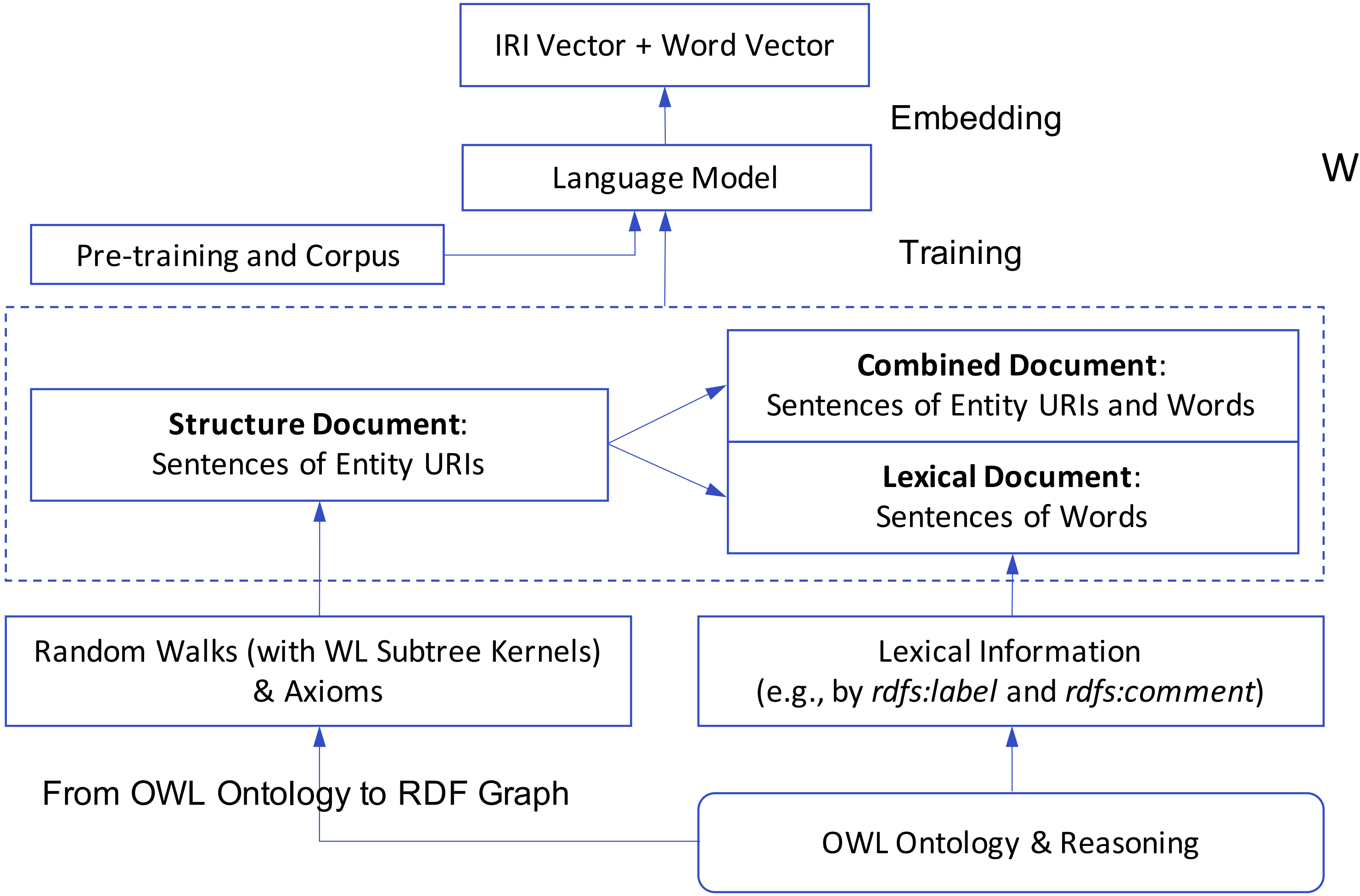
OWL2Vec*

- **Target**
  - OWL ontology embedding: represent entities (classes, individuals and properties) in a vector space, with their semantics (e.g., relationships) kept

- **Follow the pipeline paradigm of KG embedding (another pipeline is end-to-end like TransE)**
  - Segments are transformed into sentences (document) with semantics “kept”
  - Train sequence embedding models e.g., continuous skip-gram and continuous Bag-of-Words
  - E.g., node2vec for undirected graphs, RDF2Vec for graphs of RDF triples

Chen, Jiaoyan, et al. "Owl2vec*: Embedding of owl ontologies." *Machine Learning* 110.7 (2021): 1813-1845.

OWL2Vec*



IRI Vector + Word Vector

Embedding

Language Model

Pre-training and Corpus

Training

Combined Document: Sentences of Entity URIs and Words

Structure Document: Sentences of Entity URIs

Lexical Document: Sentences of Words

Random Walks (with WL Subtree Kernels) & Axioms

Lexical Information (e.g., by *rdfs:label* and *rdfs:comment*)

From OWL Ontology to RDF Graph

OWL Ontology & Reasoning

Word2vec is adopted

OWL2Vec*

- Ontology completion
  - Class membership and subsumption prediction
    - Input: the OWL2Vec* embeddings of two entities (as learned features)
    - Classifiers e.g., Random Forest and Logistic Regression
      - Learn from the known axioms
    - Output: a score in [0, 1]

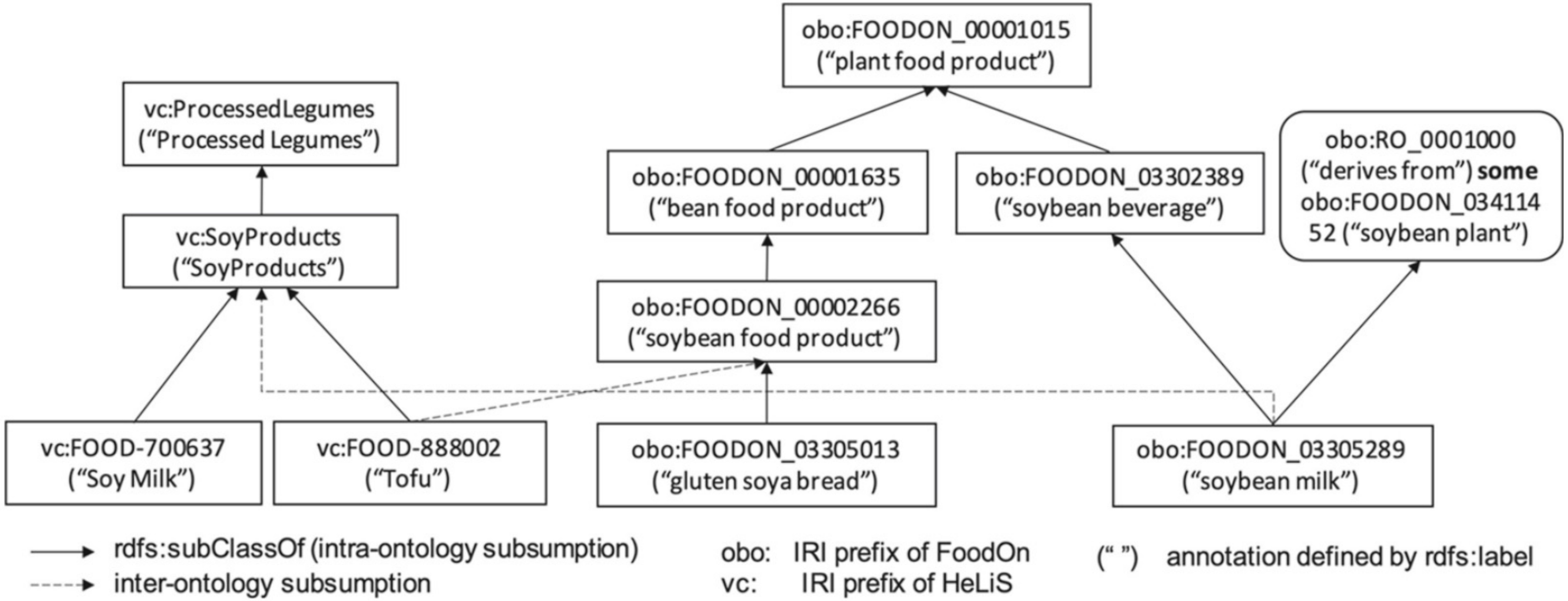
- Others
  - Ontology clustering[1], ontology alignment[2], neural-symbolic AI e.g., ontology-based Low-resource Learning[3], etc.

[1] Ritchie, Ashley, et al. "Ontology Clustering with OWL2Vec." DeepOntoNLP – ESWC Workshop, 2021.

[2] Chen, Jiaoyan, et al. "Augmenting ontology alignment by semantic embedding and distant supervision." ESWC. Springer, Cham, 2021.

[3] Chen, Jiaoyan, et al. "Knowledge-aware Zero-Shot Learning: Survey and Perspective." IJCAI Survey Track (2021).

BERTSubs

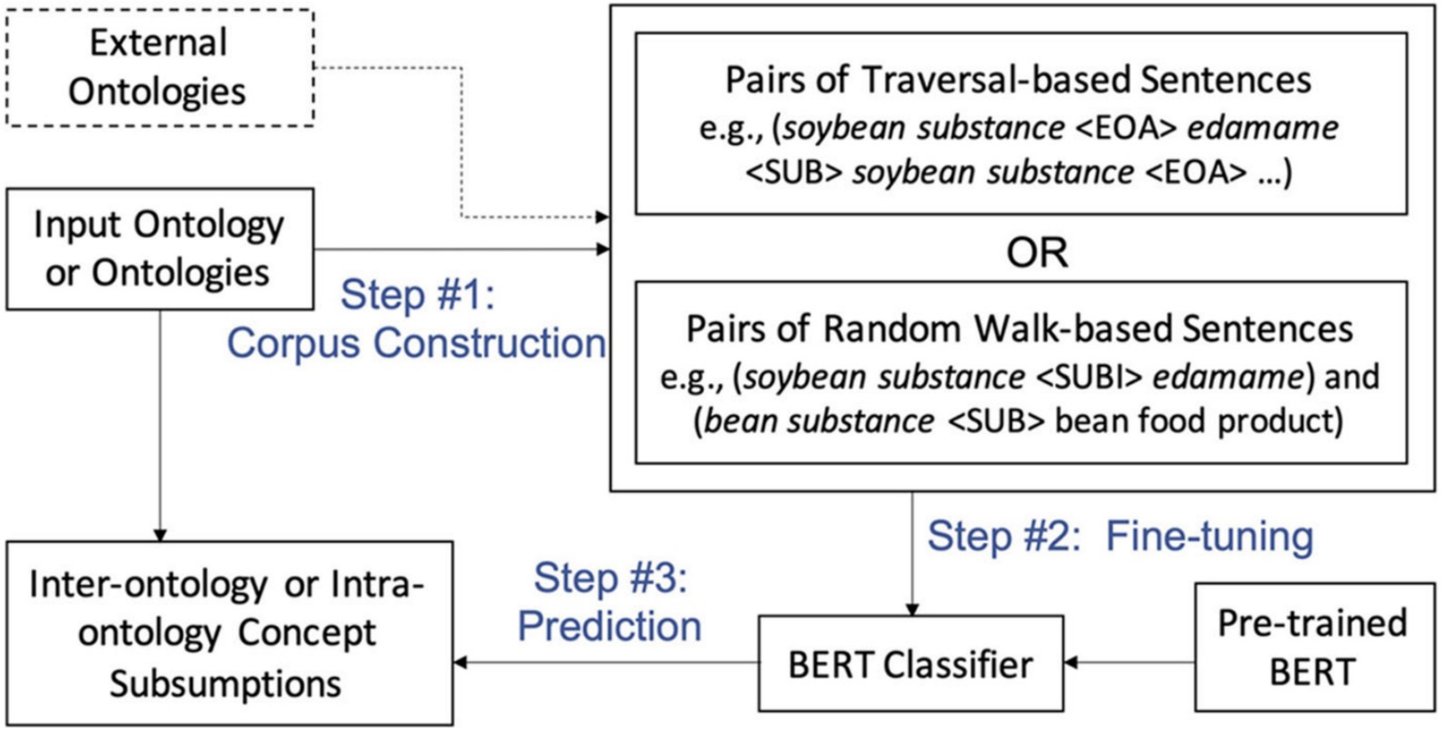


Segments from the healthy lifestyle ontology HeLiS (Left) and the food ontology FoodOn (Right) with examples of inter-ontology and intra-ontology **class subsumptions**

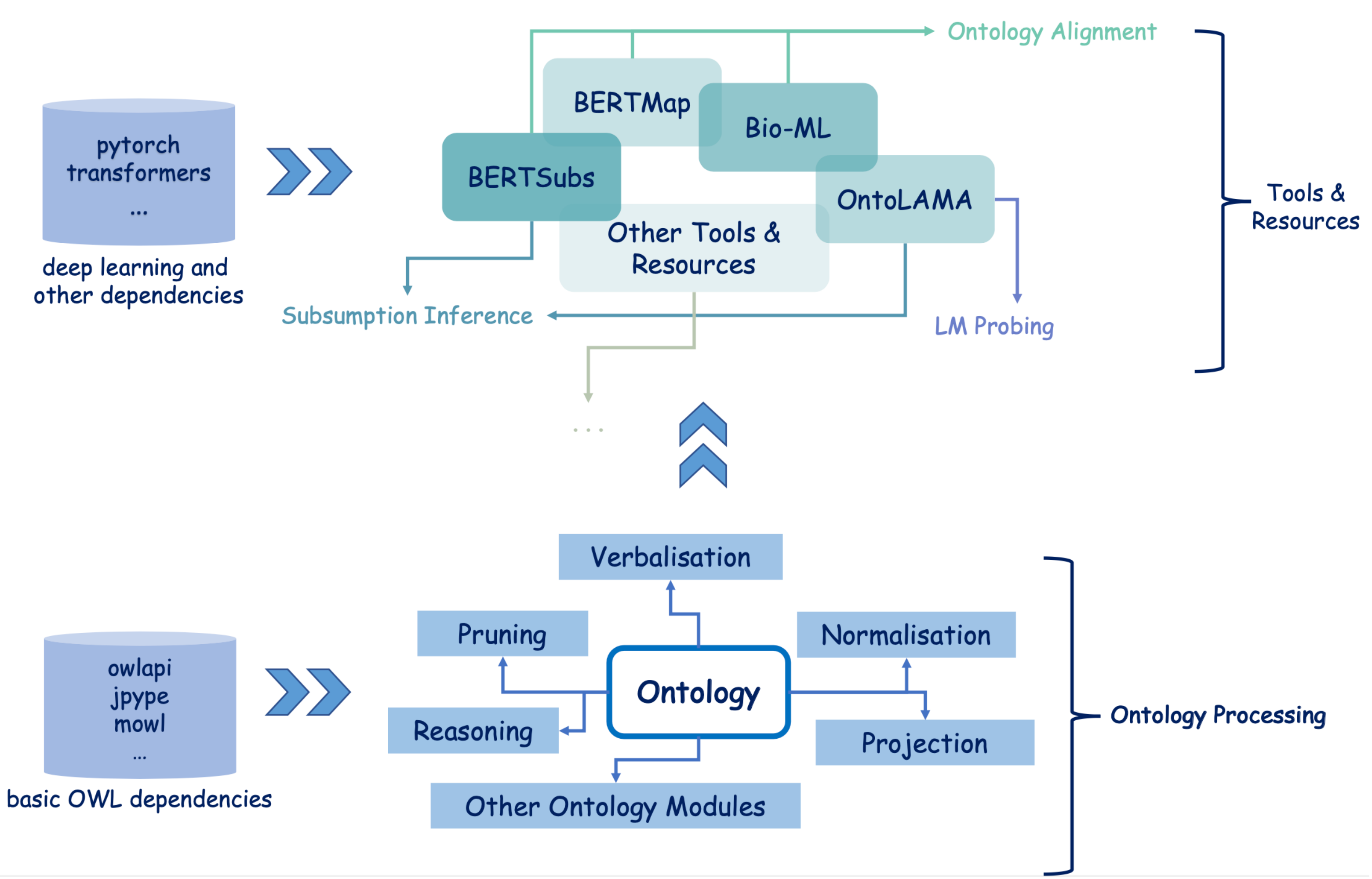
Chen, Jiaoyan, et al. "Contextual semantic embeddings for ontology subsumption prediction." *World Wide Web* (2023): 1-23.

BERTSubs

- Fine-tune a BERT model with subsumptions in the given ontologies
- Different templates for utilizing the context
  - Class label alone
  - Class path
  - Class context (breadth first search)


External Ontologies
Input Ontology or Ontologies
Step #1: Corpus Construction
Pairs of Traversal-based Sentences
e.g., (*soybean substance* <EOA> *edamame* <SUB> *soybean substance* <EOA> ...)
OR
Pairs of Random Walk-based Sentences
e.g., (*soybean substance* <SUBI> *edamame*) and (*bean substance* <SUB> bean food product)
Step #2: Fine-tuning
Pre-trained BERT
BERT Classifier
Step #3: Prediction
Inter-ontology or Intra-ontology Concept Subsumptions

The DeepOnto Library



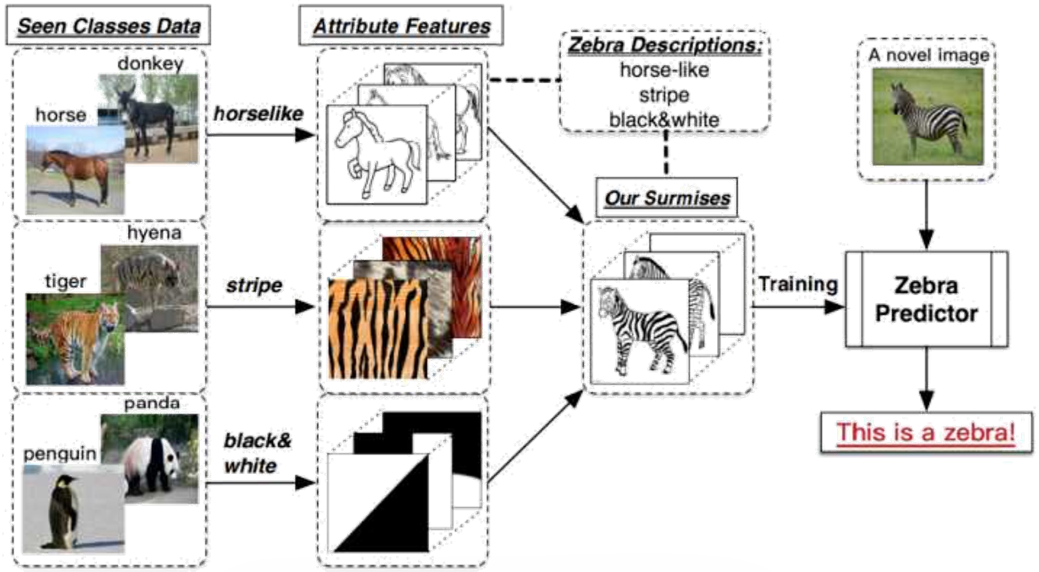
https://github.com/KRR-Oxford/DeepOnto

Applications

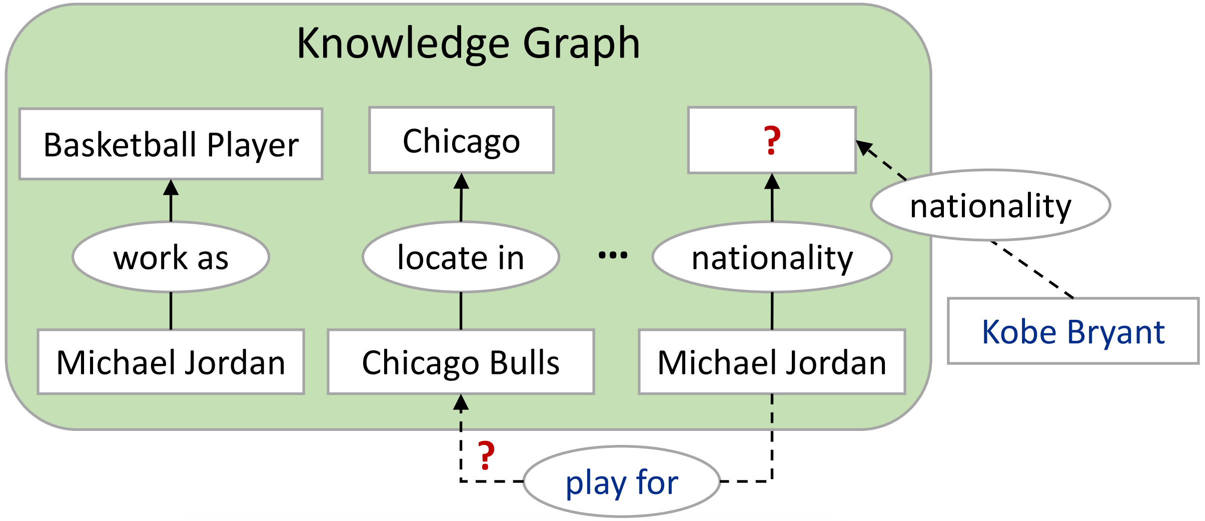
- Search engines (e.g., Google KG)
- Search, browse and recommendation in e-Commerce (e.g., Amazon Product Graph)
- Personal assistants (e.g., Apple Siri, Amazon Alex)
- Clinical AI (e.g., *online doctor*)
- Smart City (e.g., Cities Knowledge Graph)
- …

Knowledge-driven Zero-shot Learning (ZSL)

- What is ZSL?
  - Predict samples with new classes that have never appeared in training
  - Seen classes vs unseen classes



Zero-shot Image Classification


Knowledge Graph

Basketball Player

work as

Michael Jordan

Chicago

locate in

Chicago Bulls

…

?

nationality

Michael Jordan

nationality

Kobe Bryant

?

play for

Zero-shot Knowledge Graph Completion

Knowledge-driven Zero-shot Learning

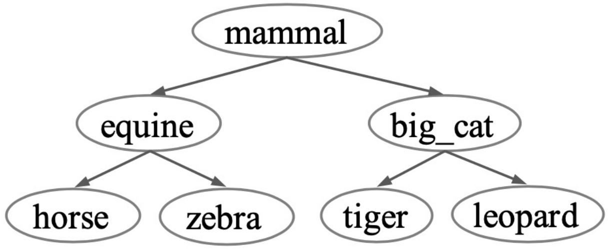
- **External knowledge** (a.k.a. **side information**) model the relationship between classes, thus enabling the transfer of the model from seen classes to unseen classes.

- Textual description:
*"**Zebras** are white animals with black stripes, they have larger, rounder ears than horses ..."*

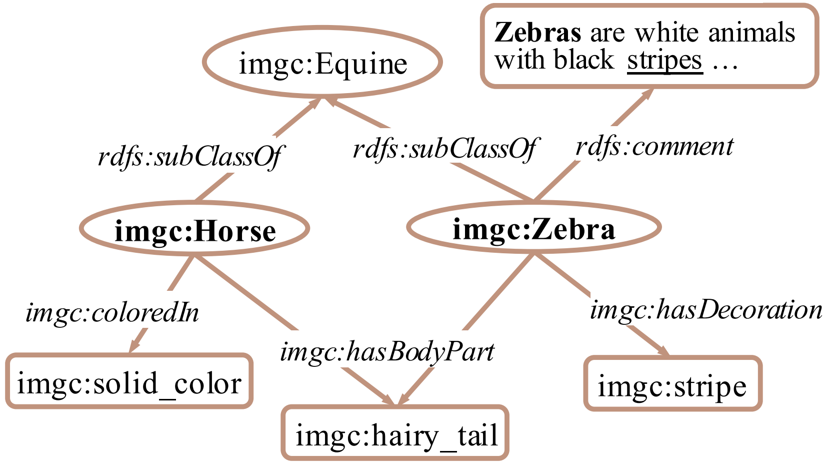
zebra	
black:	yes
white :	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

tiger	
black:	yes
white :	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

- Attribute descriptions, e.g., visual properties of animals



- Taxonomy (category)



- Knowledge Graph (Relational Facts + Categories + Literals)

$$``Zebra \sqsubseteq Equine \sqcap \exists hasTexture.Stripes \sqcap \exists hasHabitat.Meadow \ \ldots"$$

$$``hasUncle \equiv hasParent \circ hasBrother"$$

- Logics & rules

Knowledge-driven Low-resource Learning

- More readings …

Knowledge-aware Zero-Shot Learning: Survey and Perspective

Jiaoyan Chen[1*], Yuxia Geng[2], Zhuo Chen[2], Ian Horrocks[1], Jeff Z. Pan[3] and Huajun Chen[2*]

[1]Department of Computer Science, University of Oxford
[2]College of Computer Science & AZFT Knowledge Engine Lab, Zhejiang University
[3]School of Informatics, The University of Edinburgh

Abstract

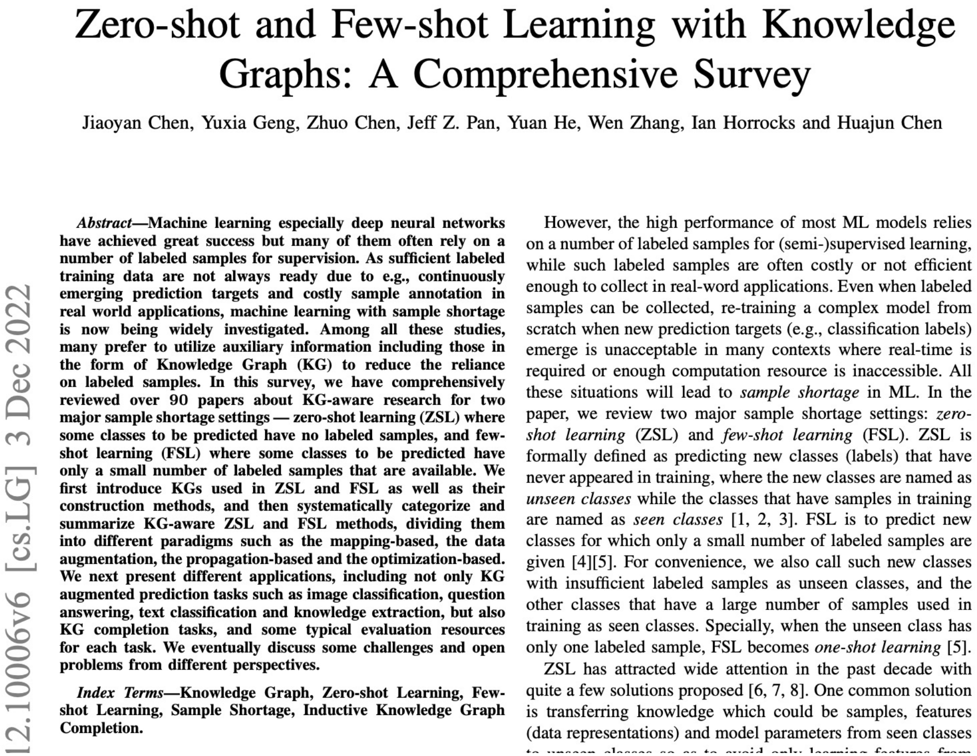
Zero-shot learning (ZSL) which aims at predicting classes that have never appeared during the training using external knowledge (a.k.a. side information) has been widely investigated. In this paper we present a literature review towards ZSL in the perspective of external knowledge, where we categorize the external knowledge, review their methods and compare different external knowledge. With the literature review, we further discuss and outlook the role of symbolic knowledge in addressing ZSL and other machine learning sample shortage issues.

1 Introduction

Normal supervised machine learning (ML) classification trains a model with labeled samples and predicts the classes of subsequent samples using classes that were encountered during the training stage. *Zero-shot learning* (ZSL), however, aims to also predict novel classes that did not occur in the training samples. Such novel classes are known as *unseen classes*, while the classes occurring in training samples are known as *seen classes*. ZSL has been widely investigated as a means of addressing common ML issues such as emerging

based) [Fu *et al.*, 2018; Wang *et al.*, 2019; Xian *et al.*, 2018], and few of them systematically analyze the external knowledge which play a key role in designing ZSL methods and in improving their performance since no samples are given for the unseen classes. In contrast, this paper reviews ZSL studies mainly from the perspective of external knowledge. We categorize the external knowledge into four kinds — text, attribute, KG and ontology & rules, according to their data structures, sources, expressivity, etc. For each kind we review its methods, case studies and benchmarks. We also compare different external knowledge, and discuss the role of symbolic knowledge representation in addressing ZSL and other sample shortage settings. Although [Fu *et al.*, 2018] and [Wang *et al.*, 2019] also introduce the semantic space (i.e., encoding of the external knowledge), no paper analysis is conducted for each external knowledge, and more importantly the external knowledge involved (mainly text and attributes) are quite incomplete — KG and ontology which started to be widely investigated in recent three years are not covered. [Xian *et al.*, 2018] contributes a comprehensive evaluation to multiple ZSL methods, but these methods are limited to image classification, while we consider different tasks in multiple domains including CV, NLP, KG construction and completion, etc.

IJCAI'21 Survey Track



Zero-shot and Few-shot Learning with Knowledge Graphs: A Comprehensive Survey

Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z. Pan, Yuan He, Wen Zhang, Ian Horrocks and Huajun Chen

***Abstract*—Machine learning especially deep neural networks have achieved great success but many of them often rely on a number of labeled samples for supervision. As sufficient labeled training data are not always ready due to e.g., continuously emerging prediction targets and costly sample annotation in real world applications, machine learning with sample shortage is now being widely investigated. Among all these studies, many prefer to utilize auxiliary information including those in the form of Knowledge Graph (KG) to reduce the reliance on labeled samples. In this survey, we have comprehensively reviewed over 90 papers about KG-aware research for two major sample shortage settings — zero-shot learning (ZSL) where some classes to be predicted have no labeled samples, and few-shot learning (FSL) where some classes to be predicted have only a small number of labeled samples that are available. We first introduce KGs used in ZSL and FSL as well as their construction methods, and then systematically categorize and summarize KG-aware ZSL and FSL methods, dividing them into different paradigms such as the mapping-based, the data augmentation, the propagation-based and the optimization-based. We next present different applications, including not only KG augmented prediction tasks such as image classification, question answering, text classification and knowledge extraction, but also KG completion tasks, and some typical evaluation resources for each task. We eventually discuss some challenges and open problems from different perspectives.**

***Index Terms*—Knowledge Graph, Zero-shot Learning, Few-shot Learning, Sample Shortage, Inductive Knowledge Graph Completion.**

However, the high performance of most ML models relies on a number of labeled samples for (semi-)supervised learning, while such labeled samples are often costly or not efficient enough to collect in real-word applications. Even when labeled samples can be collected, re-training a complex model from scratch when new prediction targets (e.g., classification labels) emerge is unacceptable in many contexts where real-time is required or enough computation resource is inaccessible. All these situations will lead to *sample shortage* in ML. In the paper, we review two major sample shortage settings: *zero-shot learning* (ZSL) and *few-shot learning* (FSL). ZSL is formally defined as predicting new classes (labels) that have never appeared in training, where the new classes are named as *unseen classes* while the classes that have samples in training are named as *seen classes* [1, 2, 3]. FSL is to predict new classes for which only a small number of labeled samples are given [4][5]. For convenience, we also call such new classes with insufficient labeled samples as unseen classes, and the other classes that have a large number of samples used in training as seen classes. Specially, when the unseen class has only one labeled sample, FSL becomes *one-shot learning* [5].

ZSL has attracted wide attention in the past decade with quite a few solutions proposed [6, 7, 8]. One common solution is transferring knowledge which could be samples, features (data representations) and model parameters from seen classes

Proceedings of the IEEE (minor revision)

Conclusions

- What are knowledge graphs and why?

- How to construct knowledge graphs?
  - From tabular data
  - Alignment
  - Correction
  - Embedding and refinement

- Knowledge graph applications
  - Zero-shot learning

Thanks!

jiaoyan.chen@Manchester.ac.uk