Information extraction & automated knowledge graph construction

Lecture 4: From IE to AKGC

Lauriane Aufrant (Inria Paris)

Université Paris Cité – UFRL – M2 LI – 30/11/2023

Progress

- ~~09/11 – Overview of information extraction~~
- ~~16/11 – Entities, relations...~~
- ~~23/11 – Coreference, linking...~~
- 30/11 – From IE to automated knowledge graph construction
- 07/12 – IE annotations
- 14/12 – IE in a specialty domain

Today: From IE to AKGC

- Comprehensiveness
- Consistency
- Task interoperability
- Combining tasks

Comprehensiveness

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

Comprehensiveness for entities

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

Comprehensiveness for entities

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Large set of entity types
  $\hookrightarrow$ "a green backpack"

Comprehensiveness for entities

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Large set of entity types
  $\hookrightarrow$ "a green backpack"
- Nested entities
  $\hookrightarrow$ "The dean's secretary"

Comprehensiveness for entities

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Large set of entity types
  $\hookrightarrow$ "a green backpack"
- Nested entities
  $\hookrightarrow$ "The dean's secretary"
- Non-named entities
  $\hookrightarrow$ All of them!

Comprehensiveness for relations

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

Comprehensiveness for relations

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Unforeseen relations: "jogging down"

Comprehensiveness for relations

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Unforeseen relations: "jogging down"
- "with"?

Comprehensiveness for relations

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

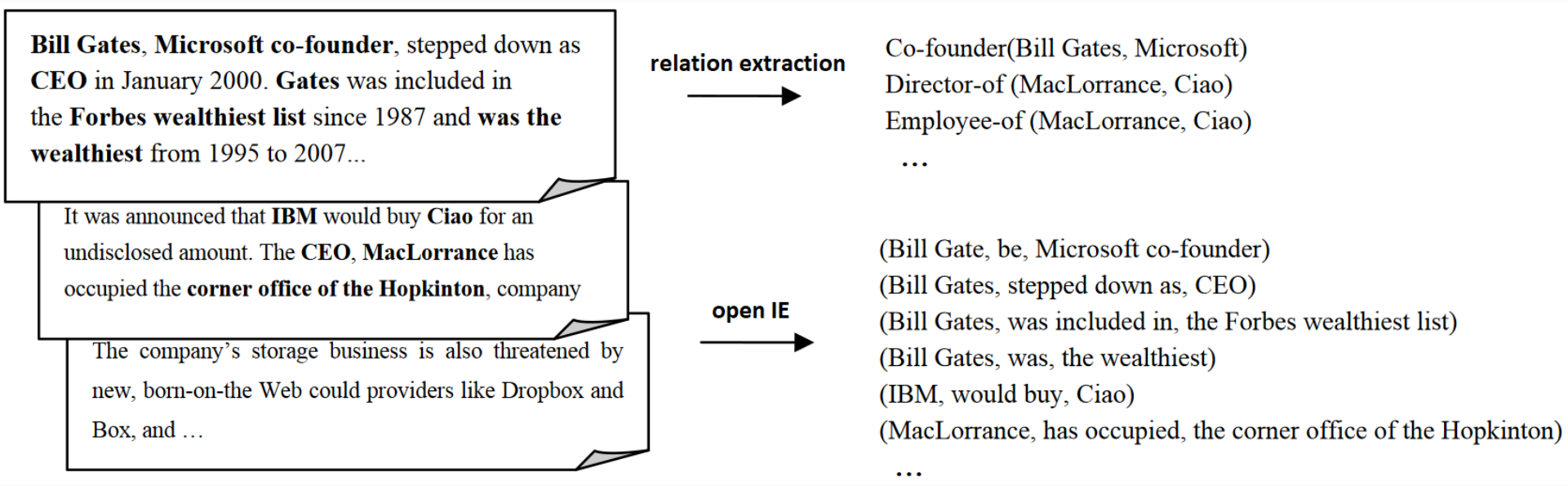
- Unforeseen relations: "jogging down"
- "with"?
- "nearest"?

Comprehensiveness for relations

The dean's secretary saw a guy with a green backpack jogging down the avenue to the nearest restaurant.

- Unforeseen relations: "jogging down"
- "with"?
- "nearest"?
- "saw": "a guy with [...] restaurant"?

Open Information Extraction



Vo & Bagheri, 2016

Corpora for OpenIE

- Often, none: fully unsupervised & manual evaluation
- ORE [Mesquita et al., 2013] (662 binary & 222 n-ary relations + automatic annotations, arguments limited to named entities, 0.6-1 fact / sentence)
- QA-SRL OIE [Stanovsky & Dagan, 2016] (10,359 tuples, automatic extraction from questions & answers, 3 facts / sentence)
- Wire57 [Léchelle et al., 2019] (347 tuples, axiomatic guidelines designed for completeness, 6 facts / sentence)

Evaluating OpenIE

- Exact match F1
- Relaxed containment match
- Greedy matching of tuple pairs & computing token-level overlap
- Penalizing verbosity? Over-specific tuples? Incorrect number of arguments?
- F2 for higher focus on recall

Methods & tools for OpenIE

	Extractions	Matches	Exact matches	Prec. of matches	Recall of matches	Prec.	Recall	F1
ReVerb (Fader et al., 2011)	79	54	13	.83	.77	**.569**	.121	.200
Ollie (Mausam et al., 2012)	145	74	8	.73	.81	.347	.175	.239
ClausIE (Del Corro and Gemulla, 2013)	223	121	**24**	.74	.84	.401	.298	.342
Stanford (Angeli et al., 2015)	371	99	2	.79	.65	.210	.188	.198
OpenIE 4 (Mausam, 2016)	101	74	5	.68	.84	.501	.182	.267
PropS (Stanovsky et al., 2016)	184	69	0	.59	.80	.222	.162	.187
MinIE (Gashteovski et al., 2017)	252	**134**	10	.75	.83	.400	**.323**	**.358**

Léchelle et al., 2019

OpenIE as sequence labeling

Open IE Encoding Examples

(a) *The president claimed that he won the majority vote.*
(The president; **claimed that he won**; the majority vote)

The_{A0-B} president_{A0-I} claimed_{P-B} that_{P-I} he_{P-I} won_{P-I} the_{A1-B} majority_{A1-I} vote_{A1-I}

(b) *Barack Obama, a former U.S president, was born in Hawaii.*
(Barack Obama; **was born in**; Hawaii)
(a former U.S. president; **was born in**; Hawaii)

Barack_{A0-B} Obama_{A0-I} $,_{O}$ a_{A0-B} former_{A0-I} U.S._{A0-I} president_{A0-I} $,_{O}$ was_{P-B} born_{P-I} in_{P-I} Hawaii_{A1-B}

(c) *Theresa May plans for Brexit, on which the UK has voted last June.*
(the UK; **has voted on**; Brexit; last June)

Theresa_{O} May_{O} plans_{O} for_{O} Brexit_{A1-B} $,_{O}$ on_{O} which_{O} the_{A0-B} UK_{A0-I} has_{P-B} voted_{P-I} on_{P-I} last_{A2-B} June_{A2-I}

Stanovsky et al., 2018

OpenIE as sequence generation

"deep learning is a subfield of machine learning"

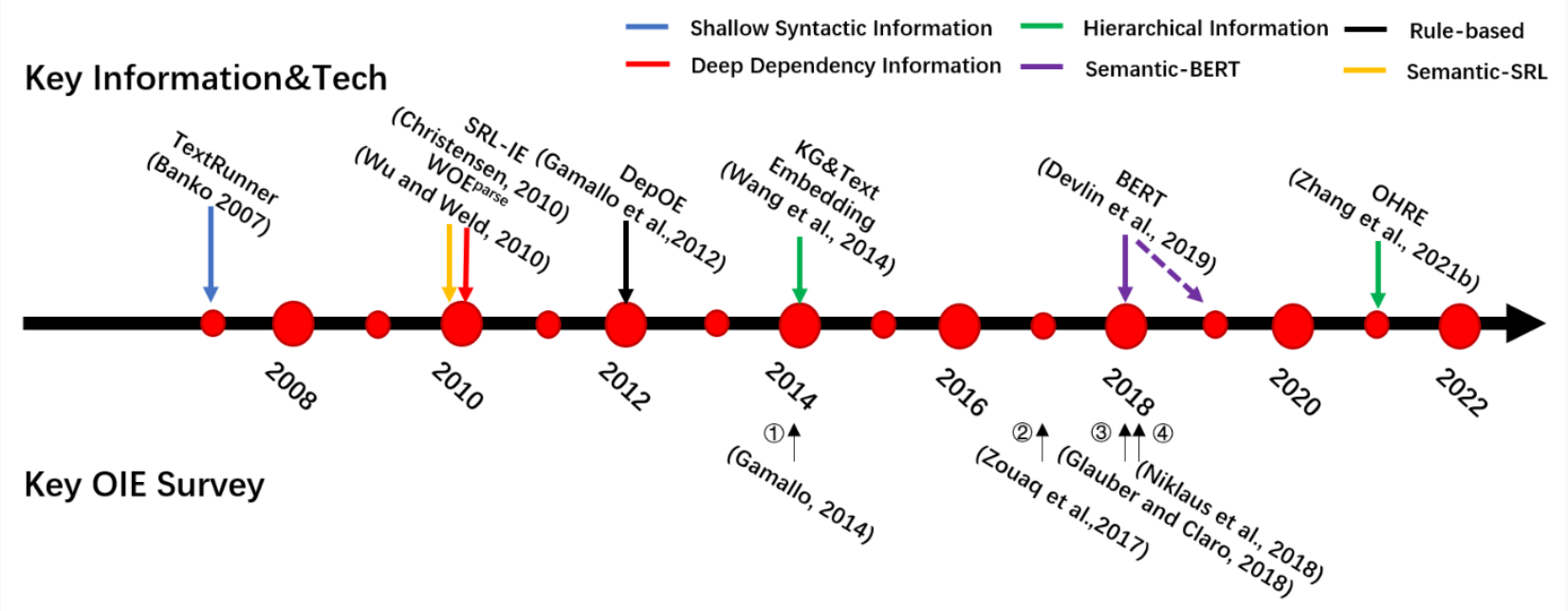
$\Downarrow$

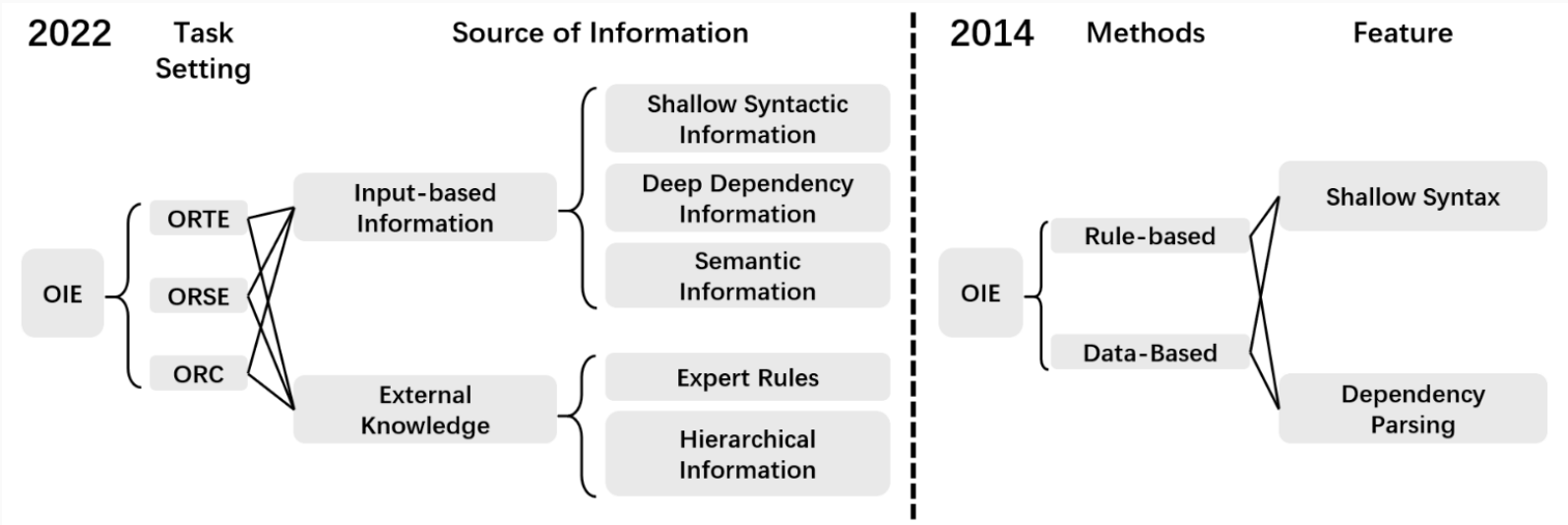
"<arg1> deep learning </arg1> <rel> is a subfield of </rel> <arg2> machine learning </arg2>"

$\rightsquigarrow$ Pure generation or with copy mechanisms

Cui et al., 2018

Overview of OpenIE methods





Liu et al., 2022

Nominal relations

- "Paul's wife turned 35."
- "They met him at the French-German border."
- "The author of this paper is William Léchelle."

  $\hookrightarrow$ "This paper has been written by William Léchelle."

Document-level relations

- "I read an interesting paper about OpenIE. The main author is William Léchelle."

Implicit relations

- "Paris, France"

- "The Turing paper" (*written by*)
  "The Nature paper" (*published in*)

Implicit arguments

- “A second earthquake occurred in Tokyo that day.”

Limits of light inference

Sentence CH 7 – "*His parents are Ashkenazi Jews who had to flee from Hungary during World War II.*"

Annotations

– (His/(Chilly Gonzales's) parents ; are ; Ashkenazi Jews)
– (His/(Chilly Gonzales's) parents ; are ; Jews)
– (His/(Chilly Gonzales's) parents ; had to flee from ; Hungary ; during World War II)
– (His/(Chilly Gonzales's) parents ; [fled] from ; Hungary ; during World War II)
– ([Chilly Gonzales] ; [has] ; parents)

Sentence FI 2 – "*A police statement did not name the man in the boot, but in effect indicated the traveler was State Secretary Samuli Virtanen, who is also the deputy to Foreign Minister Timo Soini.*"

Annotations

– (A police/(Finnish police) statement ; did not name ; (the man in the boot)/(Samuli Virtanen))
– ((the man in the boot)/(Samuli Virtanen) ; was ; Samuli Virtanen) [attributed]
– ((the traveler)/(Samuli Virtanen) ; was ; Samuli Virtanen) [attributed]
– (Samuli Virtanen ; [is] ; State Secretary)
– (Samuli Virtanen ; is ; the deputy to Foreign Minister Timo Soini)
– (Samuli Virtanen ; is ; [a] deputy)
– (Timo Soini ; [is] ; Foreign Minister)
– (Timo Soini ; [has] ; [a] deputy)

Léchelle et al., 2019

Slot filling

- Extracting properties of entities: age, colour, size...

  $\hookrightarrow$ Still triples!

- Known lists of (some) properties for (some) entity types: and the rest?

  $\rightsquigarrow$ Same as relations, but more often implicit

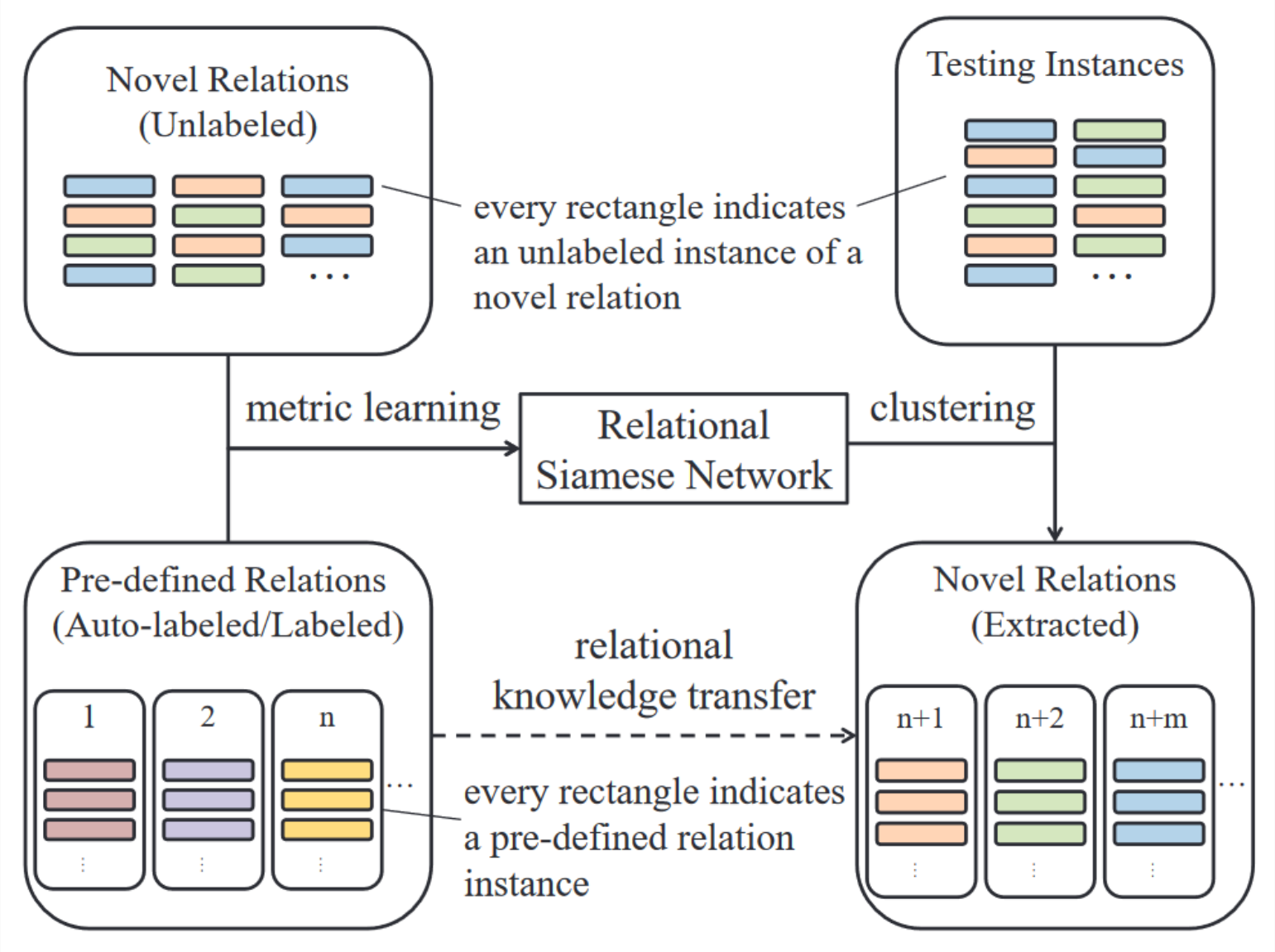
Consistency

Relation discovery

$$\textit{is married to} = \textit{has spouse} \neq \textit{has wife}$$

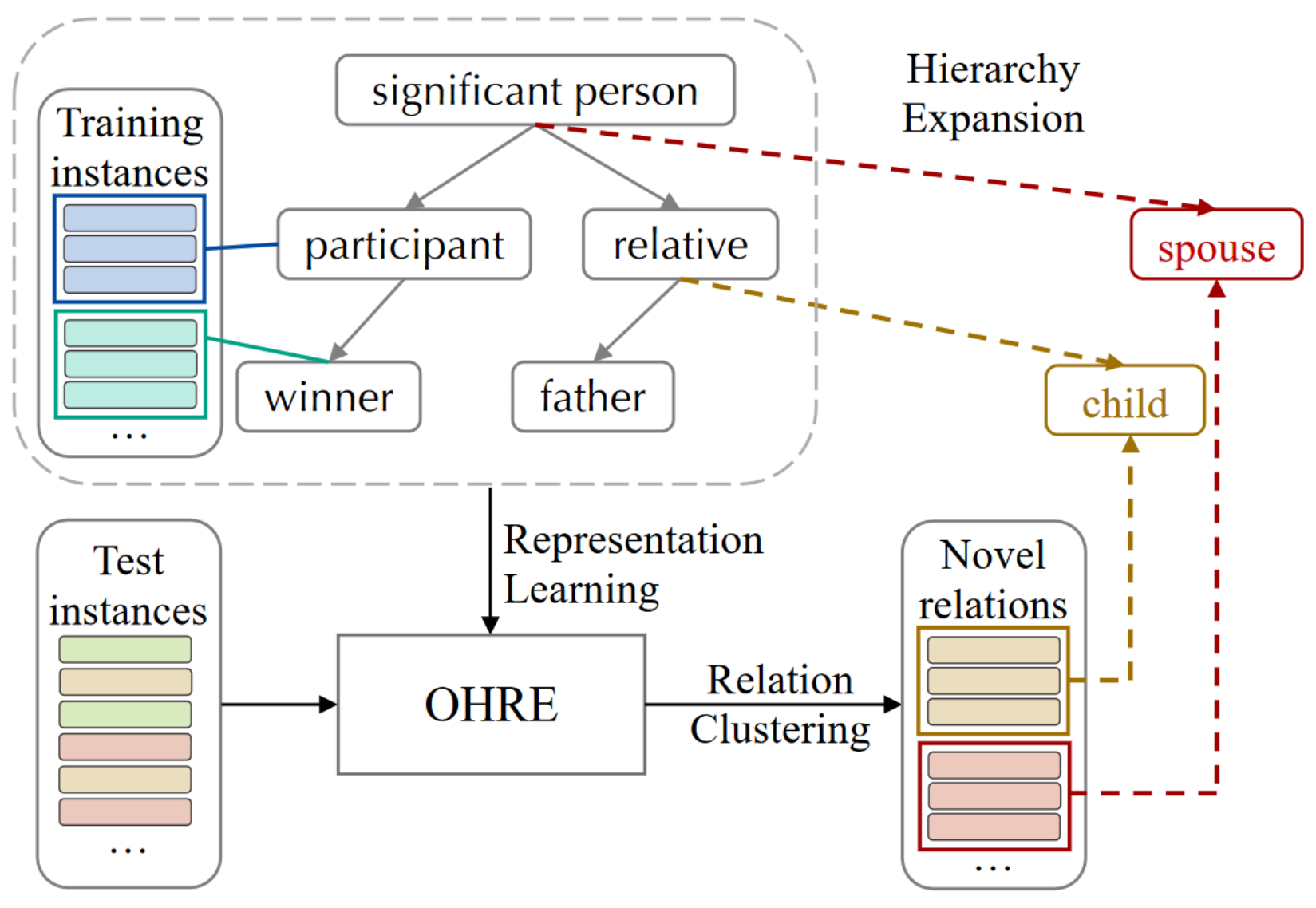
Relation discovery

is married to = *has spouse* ≠ *has wife*



Wu et al., 2019

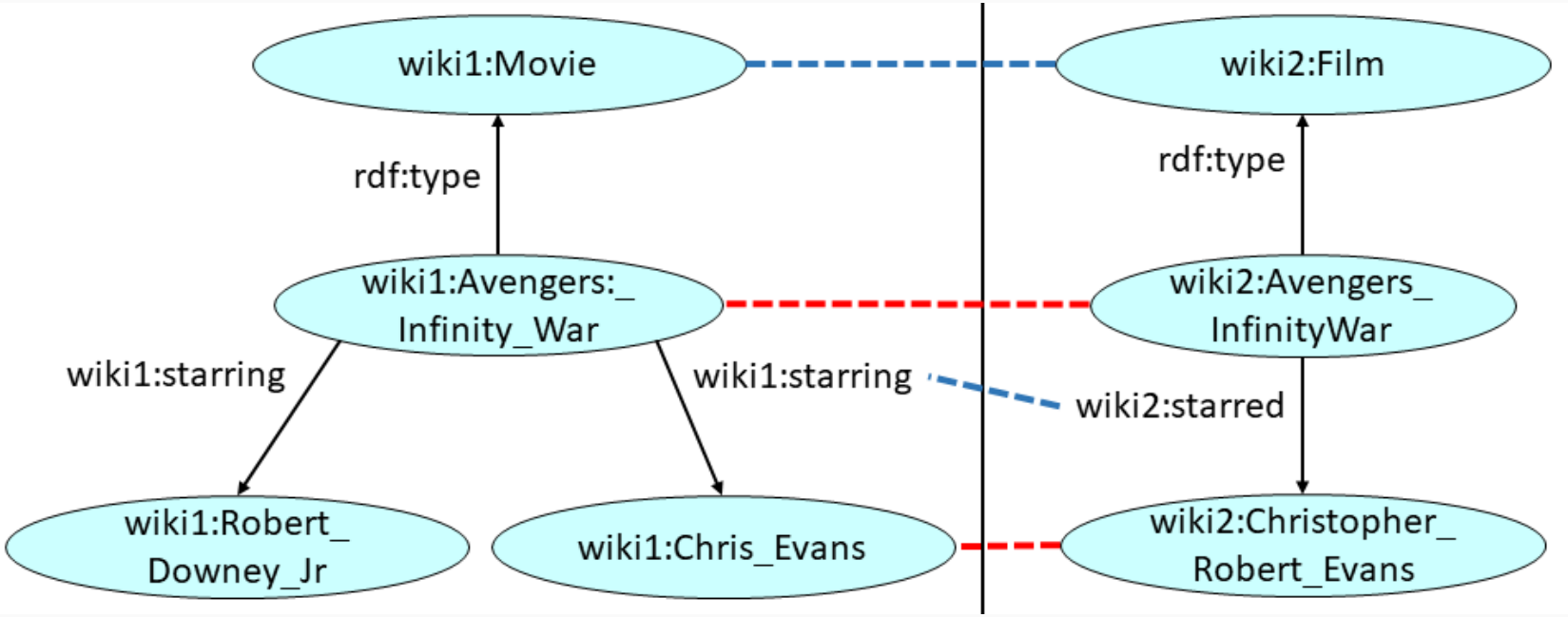
Relation discovery with a base ontology



Zhang et al., 2021

Node deduplication

In case of missed coreference, concurrent entity discovery...



$\rightsquigarrow$ Connections with ontology matching (ontology alignment)

Task interoperability

Different sets of entities across tasks

- Most available models for NER: only named, even flat entities
- Usually in relation extraction: arbitrary noun phrases

Mention detection and entity recognition

Mention detection and entity recognition

- Mention coverage in coreference resolution: non-named, possessives, clauses, unrestricted entity types...

  $\hookrightarrow$ What about NER, EL, RE?

Mention detection and entity recognition

- Mention coverage in coreference resolution: non-named, possessives, clauses, unrestricted entity types...

  $\hookrightarrow$ What about NER, EL, RE?

- But: no type, potentially no singleton...

Mention detection and entity recognition

- Mention coverage in coreference resolution: non-named, possessives, clauses, unrestricted entity types...

  $\hookrightarrow$ What about NER, EL, RE?

- But: no type, potentially no singleton...

$\rightsquigarrow$ Heuristics to reconcile?

Combining tasks

Named entity recognition & relation extraction

Which shall go first?

Named entity recognition & relation extraction

Which shall go first?

- Relation extraction as a classification of entity pairs

Named entity recognition & relation extraction

Which shall go first?

- Relation extraction as a classification of entity pairs
- **Bituyp Jenozs Zoble** was born in Beijing.
  - $\rightsquigarrow$ Probably Person (or maybe Organization)

Entity linking & relation extraction

Entity linking & relation extraction

- Born on 21/12/1977 in Amiens, **Emmanuel** graduated from the *École nationale d'administration* in 2004.

Entity linking & relation extraction

- Born on 21/12/1977 in Amiens, **Emmanuel** graduated from the *École nationale d'administration* in 2004.

- Born on 21/12/1977 in Amiens, **Emmanuel** succeeded to François in 2017.

Coreference resolution & entity linking

Coreference resolution & entity linking

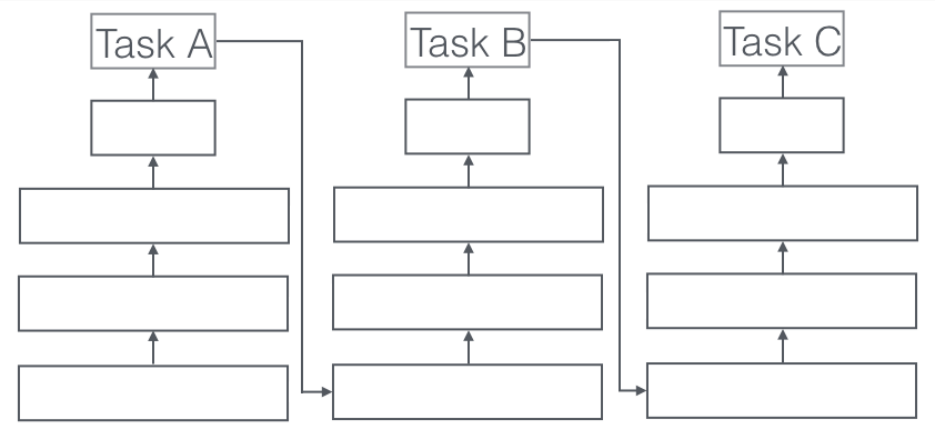
- **Emmanuel** was born on 21/12/1977 in Amiens. **He** graduated from the *École nationale d'administration* in 2004.

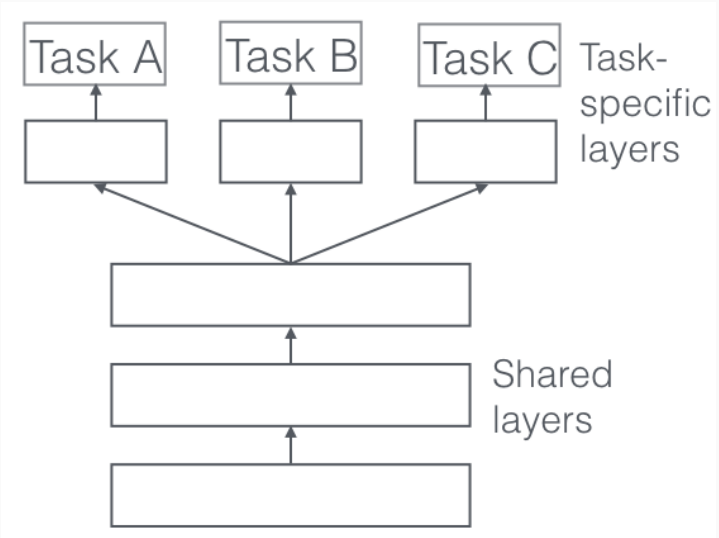
Coreference resolution & entity linking

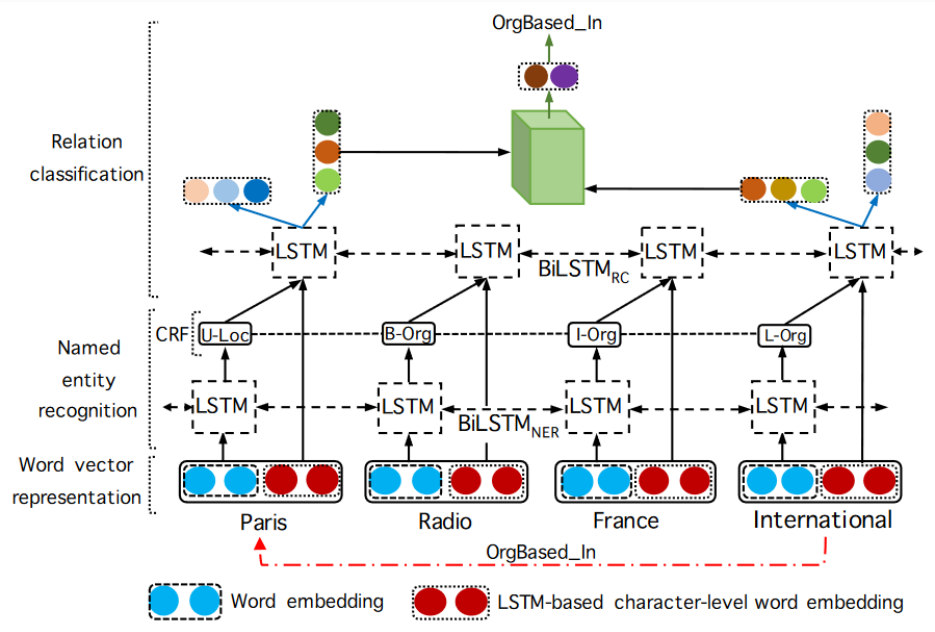
- **Emmanuel** was born on 21/12/1977 in Amiens. **He** graduated from the *École nationale d'administration* in 2004.

- Born on 21/12/1977 in Amiens, **Emmanuel** graduated from the *École nationale d'administration* in 2004. That school was abolished in 2021. It was replaced by the INSP. **Macron** succeeded to François Hollande in 2017.

Interleaving tasks: sequential, joint, multi-task learning

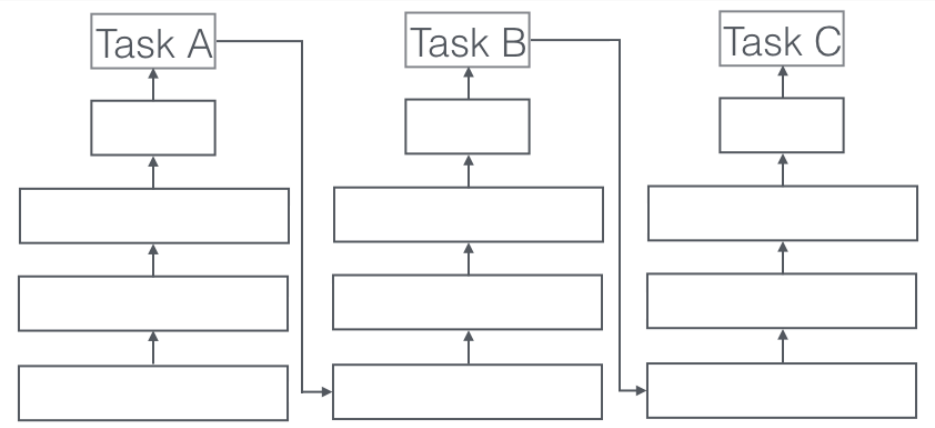


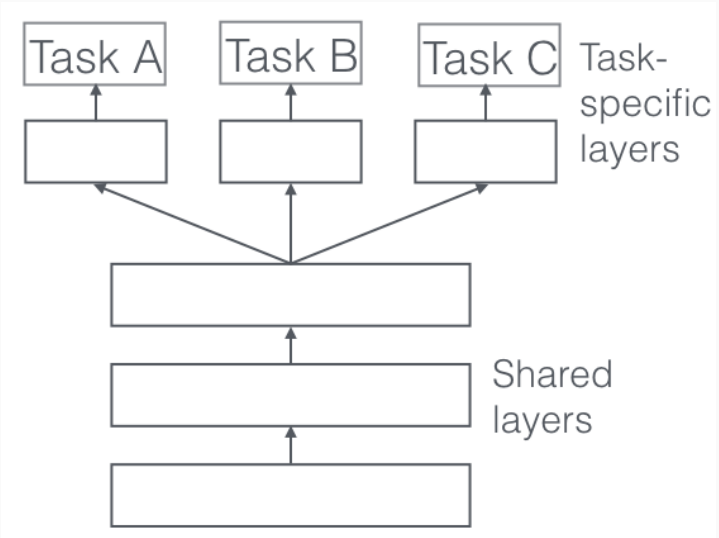


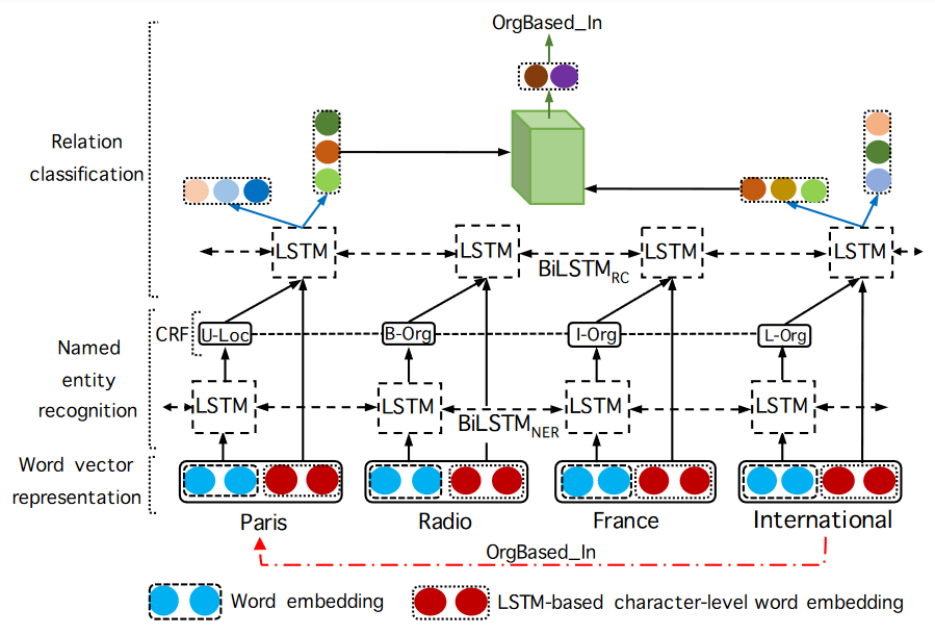


From Ruder. From Nguyen & Verspoor, 2019.

Interleaving tasks: sequential, joint, multi-task learning







And many others:

- Auxiliary inputs
- Auxiliary outputs
- ... stacking? iterations?

From Ruder. From Nguyen & Verspoor, 2019.

Relation extraction & open IE

From Ruder. From Nguyen & Verspoor, 2019.

Relation extraction & open IE

Add open IE triples to seeds and retrain RE? Bootstrapping?

From Ruder. From Nguyen & Verspoor, 2019.

Relation extraction & open IE

Add open IE triples to seeds and retrain RE? Bootstrapping?

- Often lots of noisy triples: how to filter them?
- How to match open relations with predefined ones?

From Ruder. From Nguyen & Verspoor, 2019.

See you next week!

first.last@inria.fr

From Ruder. From Nguyen & Verspoor, 2019.