

(Open) Information Extraction: Where are we going?



SAPIENZA
UNIVERSITÀ DI ROMA

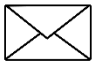
Claudio Delli Bovi
16/10/2015



About me



dellibovi@di.uniroma1.it



<http://wwwusers.di.uniroma1.it/~dellibovi>



bn:17381128n



First-year PhD student

LCL group @ Sapienza

Advisor: prof. Roberto Navigli

Focus (so far): (Open) Information Extraction

Outline

IE and OIE: some background

Outline

IE and OIE: some background

DefIE: OIE from textual definitions

Delli Bovi, Telesca, Navigli: **TACL** (to appear)

Outline

IE and OIE: some background

DefIE: OIE from textual definitions

Delli Bovi, Telesca, Navigli: **TACL** (to appear)

KBUnify: KB disambiguation and unification

Delli Bovi, Espinosa-Anke, Navigli: **EMNLP 2015**

Information Extraction

“A process of getting **structured** data from **unstructured** information in the text”

(Jurafsky and Martin, 2009)

“Identification of instances of a particular class of **relationships** in a natural language text, and the extraction of relevant **arguments** for that relationships”

(Grishman, 1997)

Information Extraction

Why?

Information Extraction

Why?

Machine Reading:

“I hereby offer to bet anyone a lobster dinner that by 2015 we will have a computer program capable of automatically reading at least 80% of the factual content across the entire English speaking web, and placing those facts in a structured knowledge base.”

(T. Mitchell. Reading the Web: A Breakthrough Goal for AI. AI Magazine, 2005)

Information Extraction

What?

Information Extraction

What?

Input:

- large corpus of unstructured text
- set of semantic relations
- labelled training data

Output:

- knowledge base of triples
< **entity, relation, entity** >

Information Extraction

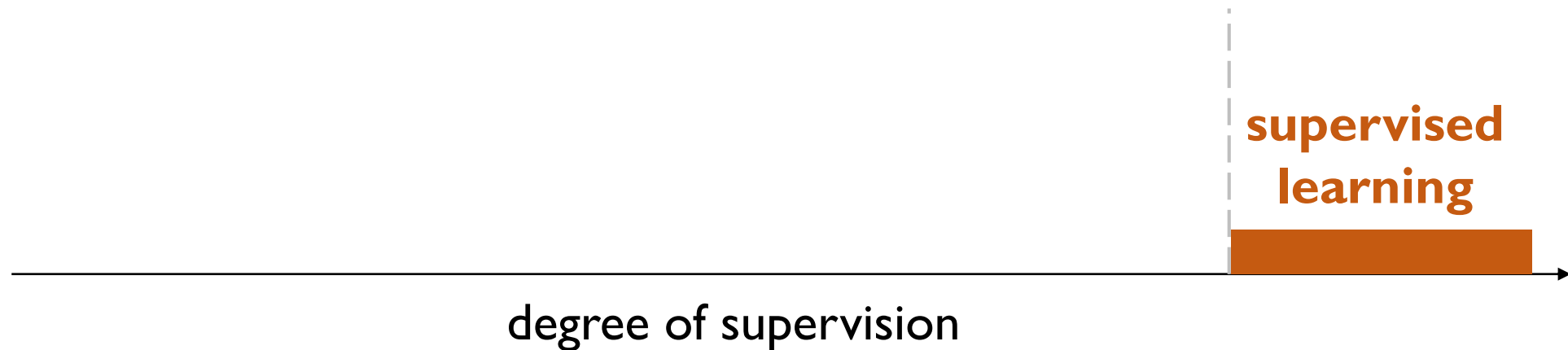
What?

Input:

- large corpus of unstructured text
- set of semantic relations
- labelled training data

Output:

- knowledge base of triples
< **entity, relation, entity** >



Information Extraction

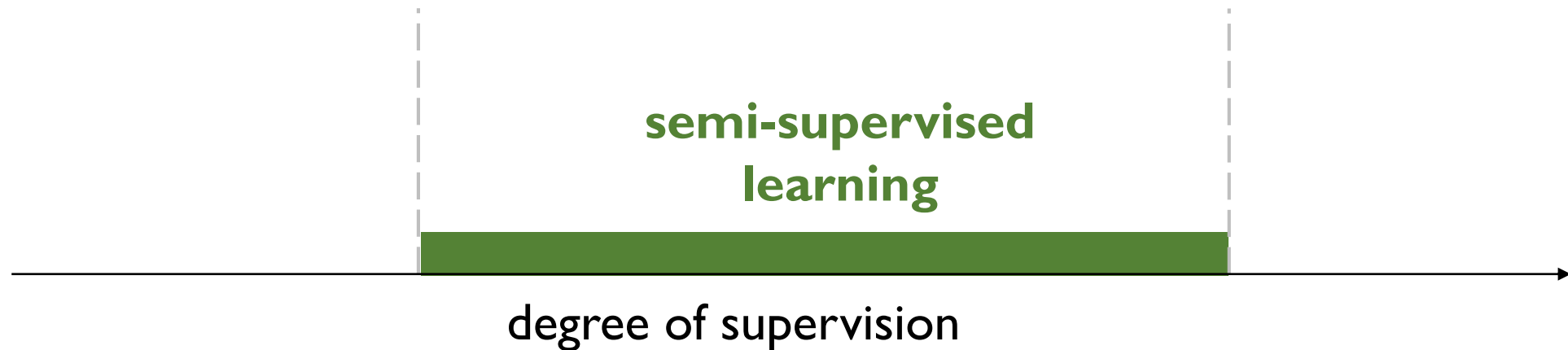
What?

Input:

- large corpus of unstructured text
- set of semantic relations
- high-precision seeds/examples

Output:

- knowledge base of triples
< **entity, relation, entity** >



Information Extraction

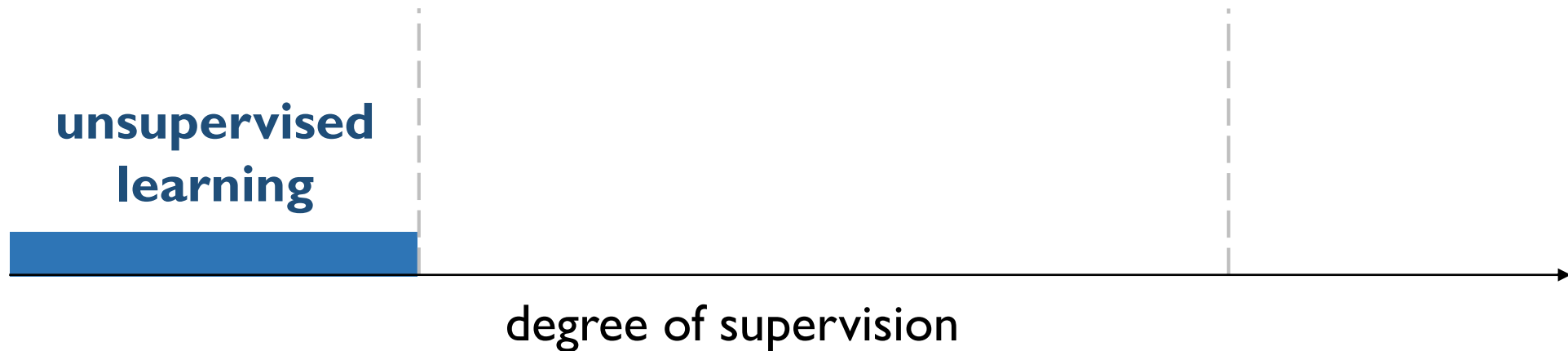
What?

Input:

- large corpus of unstructured text
- ~~set of semantic relations~~

Output:

- knowledge base of triples
< **entity, relation, entity** >
- set of semantic relations

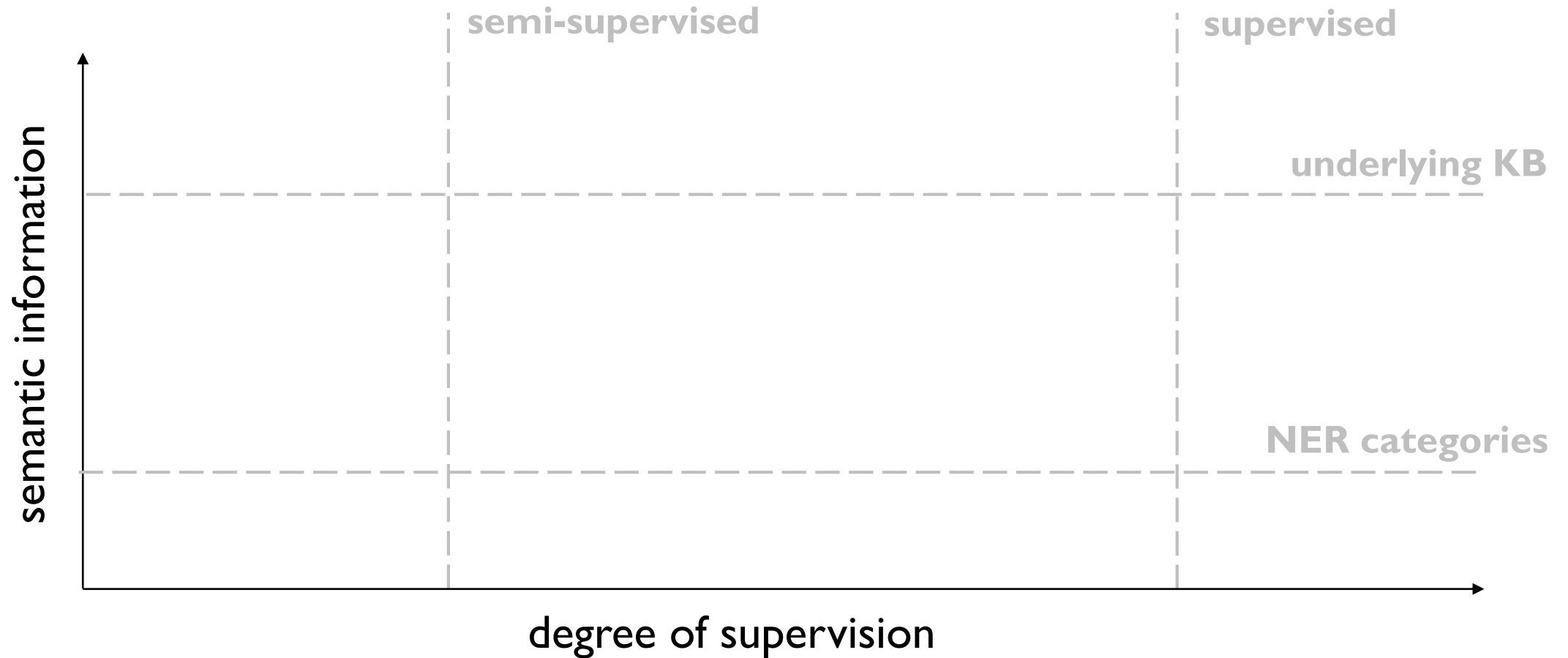


Information Extraction

How?

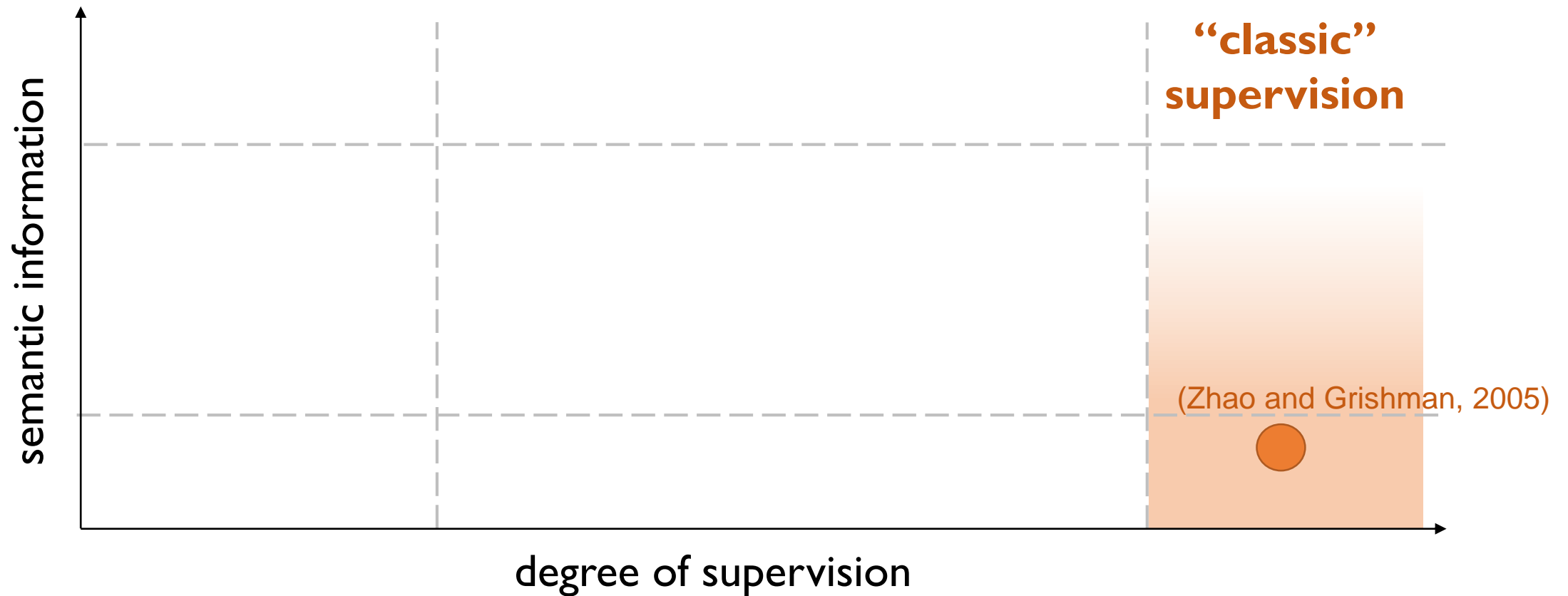
Information Extraction

How?



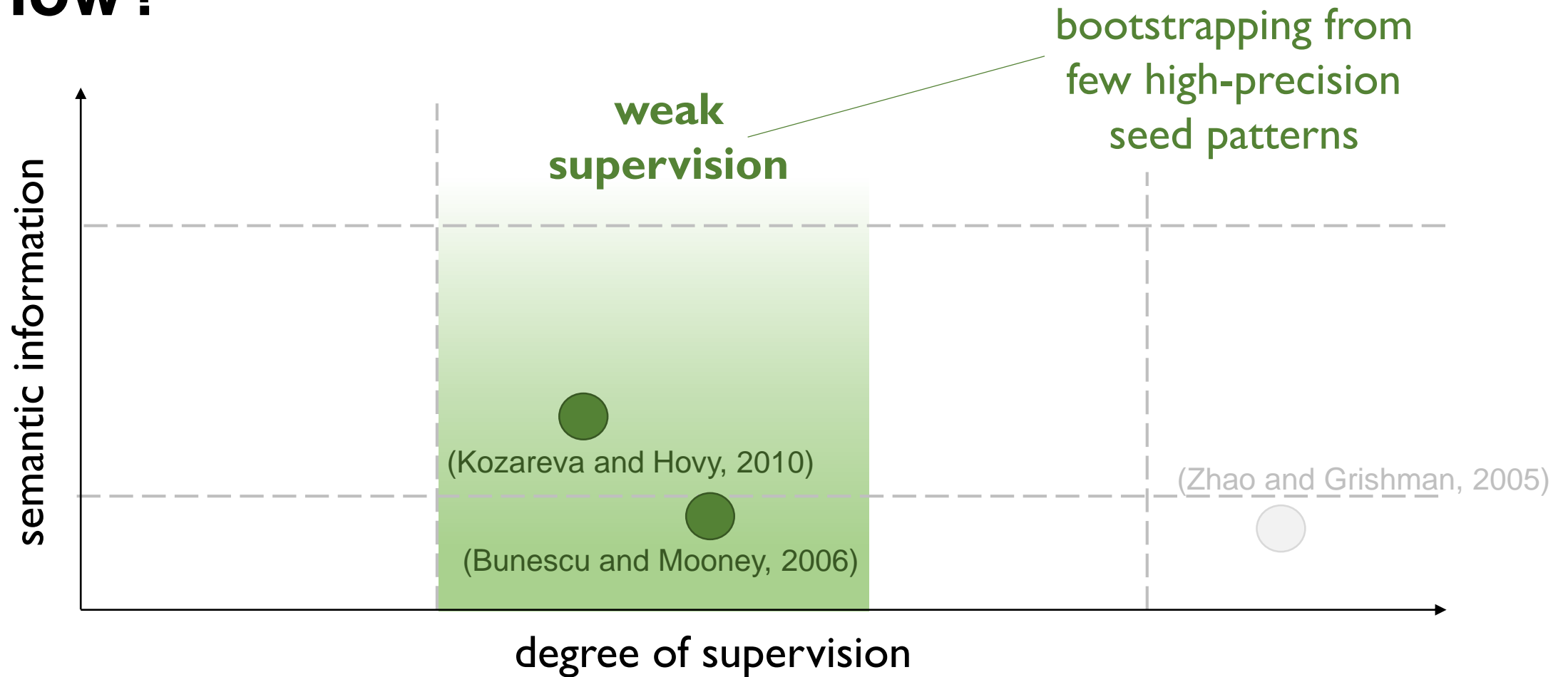
Information Extraction

How?



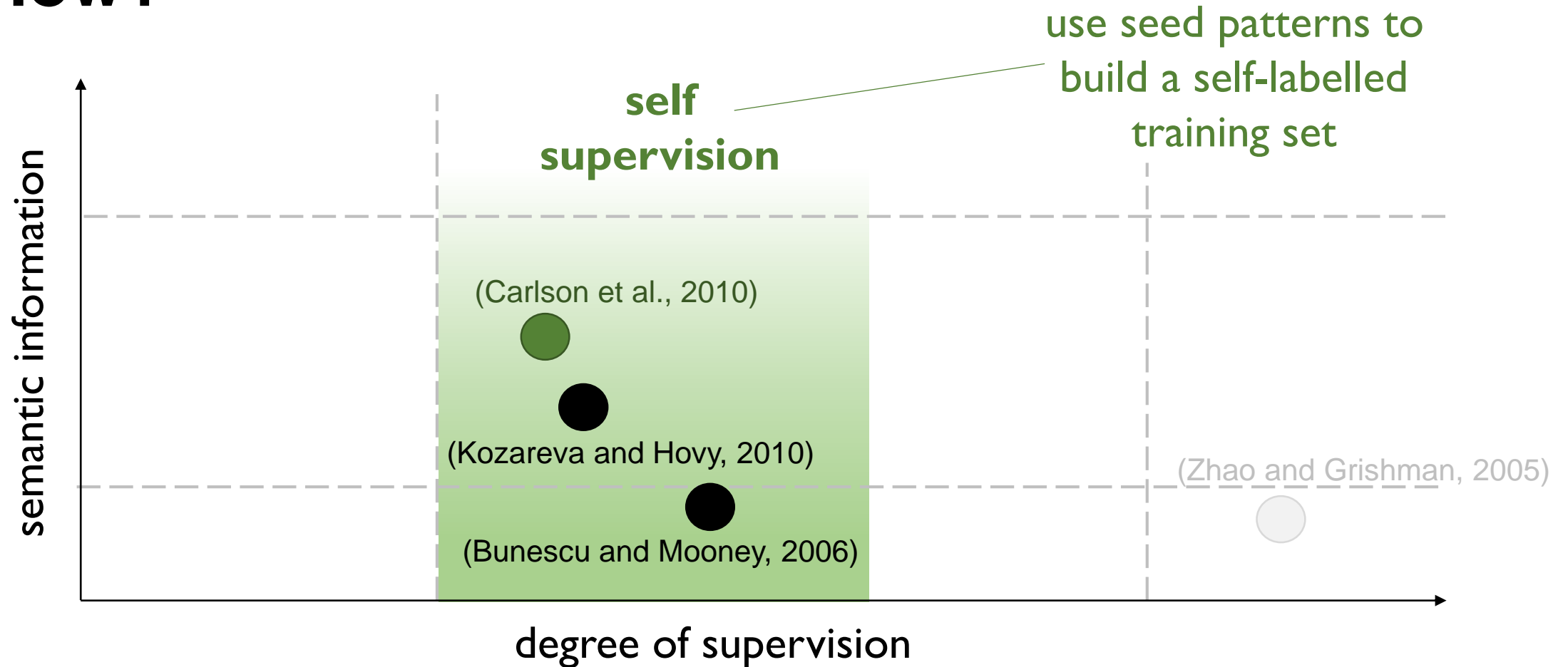
Information Extraction

How?



Information Extraction

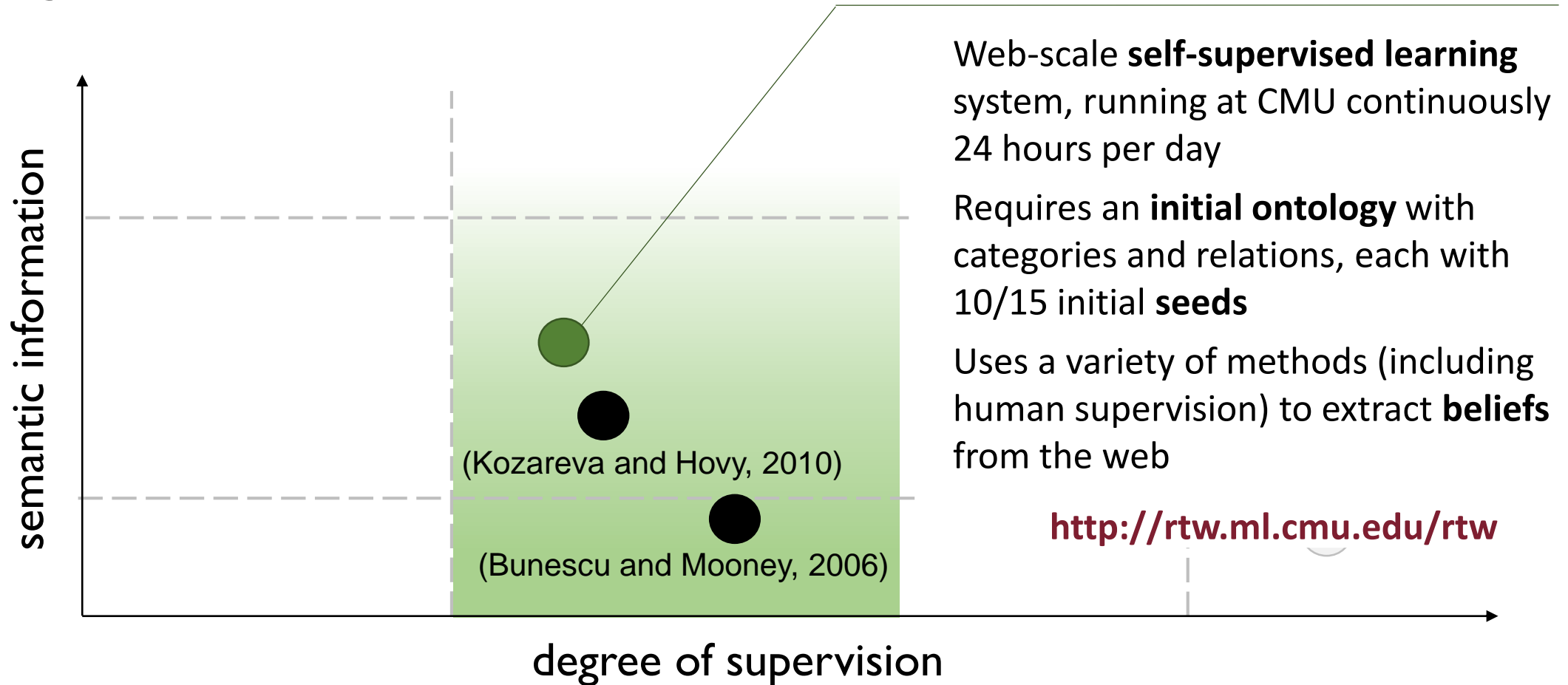
How?



Information Extraction

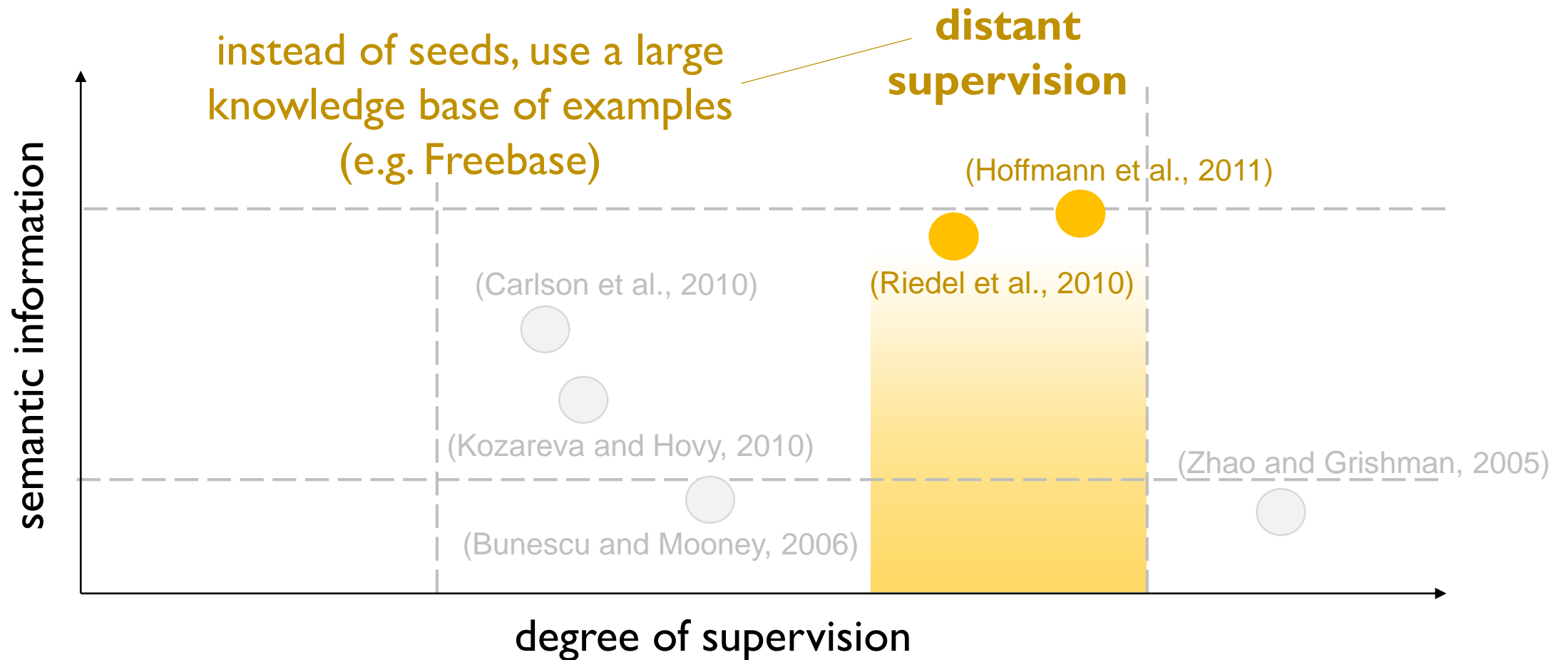
How?

NELL – Never Ending Language Learning (Carlson et al., 2010)



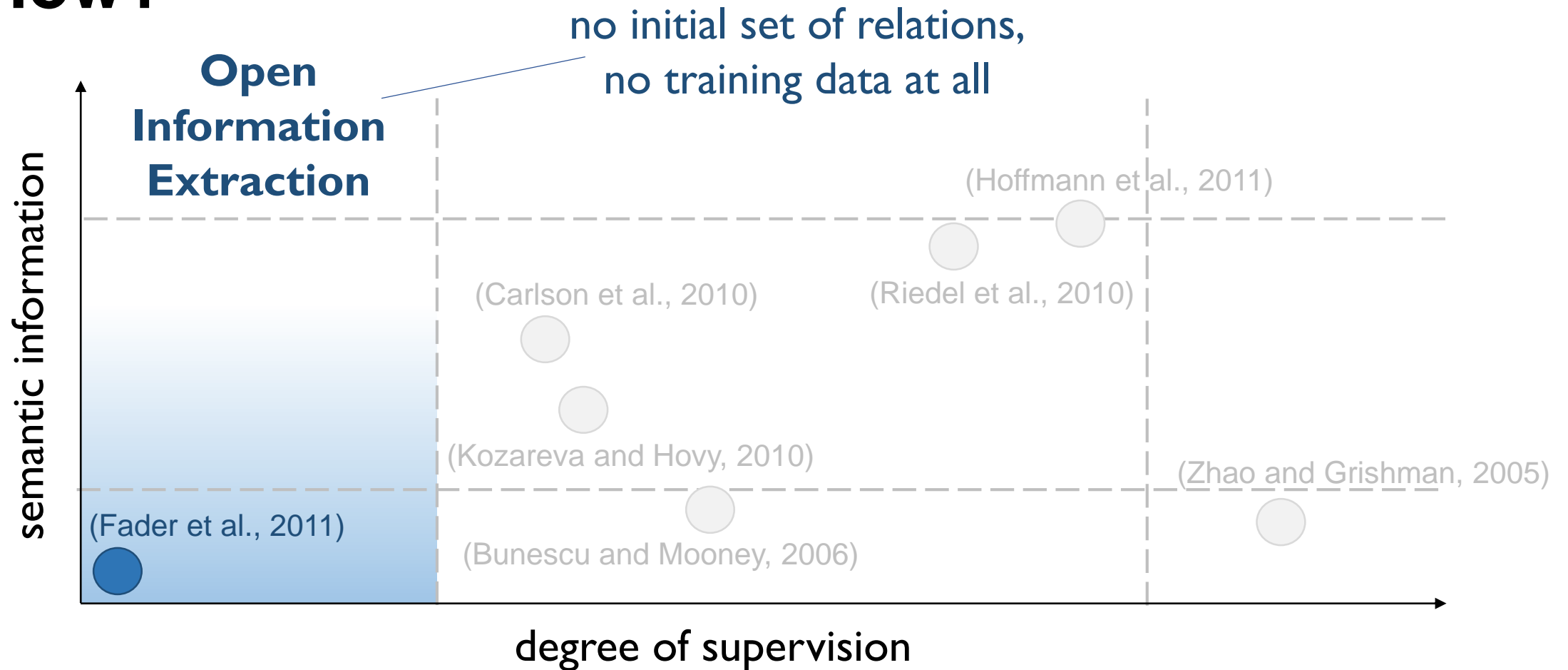
Information Extraction

How?



Information Extraction

How?



Information Extraction

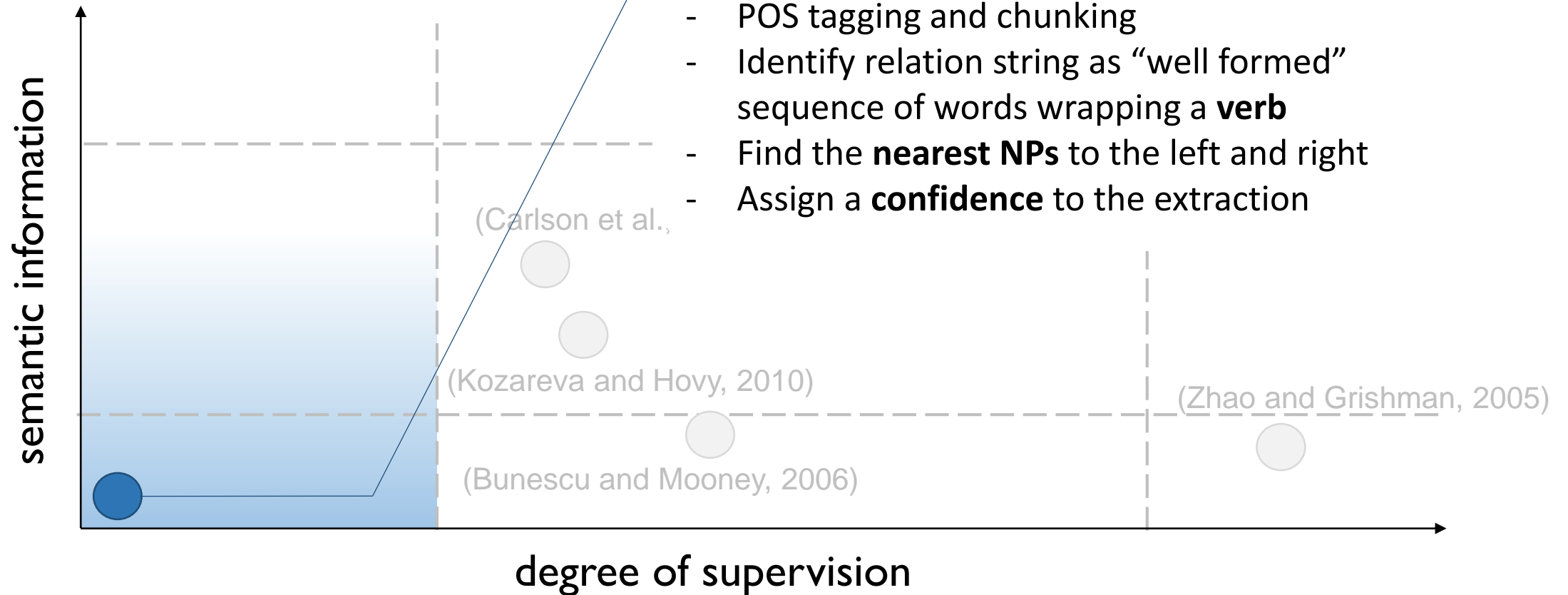
How?

ReVerb (Fader et al., 2011)



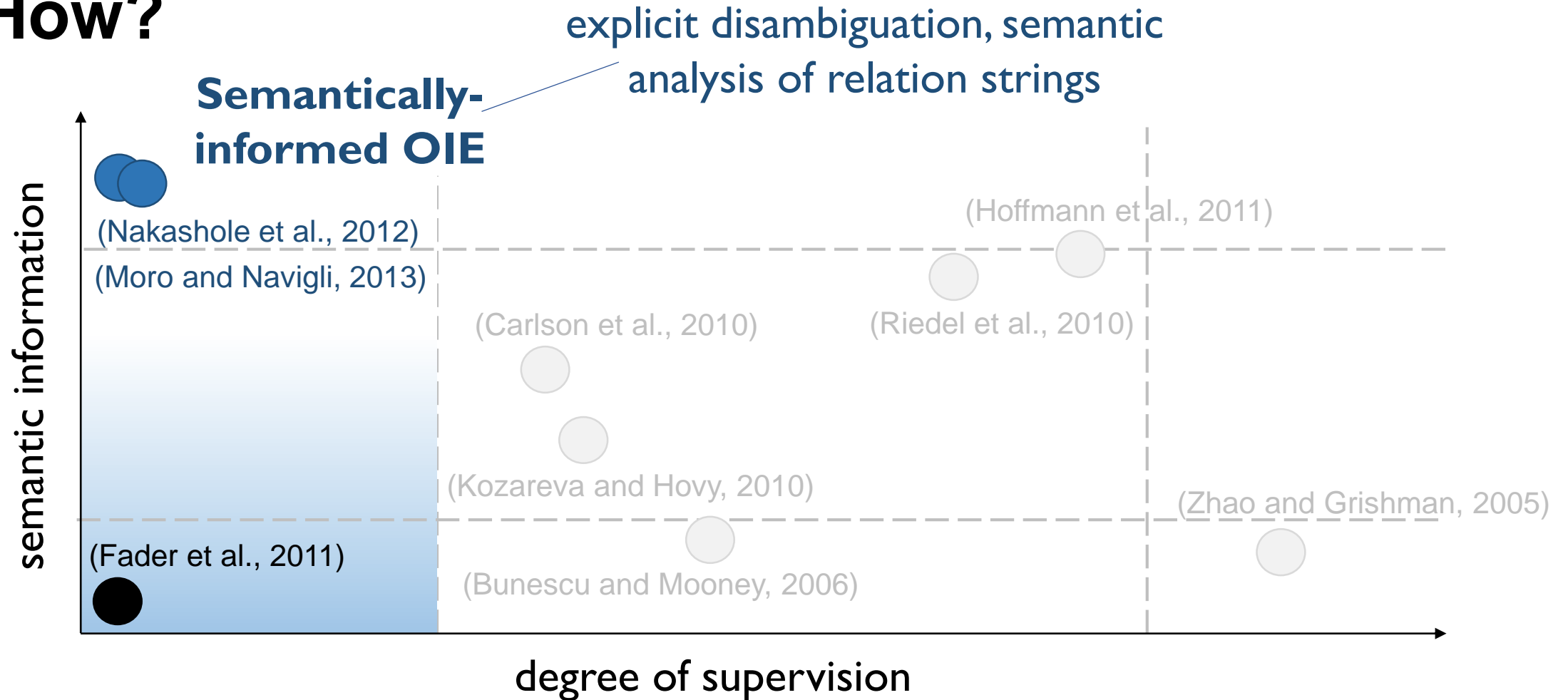
For each sentence in the corpus:

- POS tagging and chunking
- Identify relation string as “well formed” sequence of words wrapping a **verb**
- Find the **nearest NPs** to the left and right
- Assign a **confidence** to the extraction



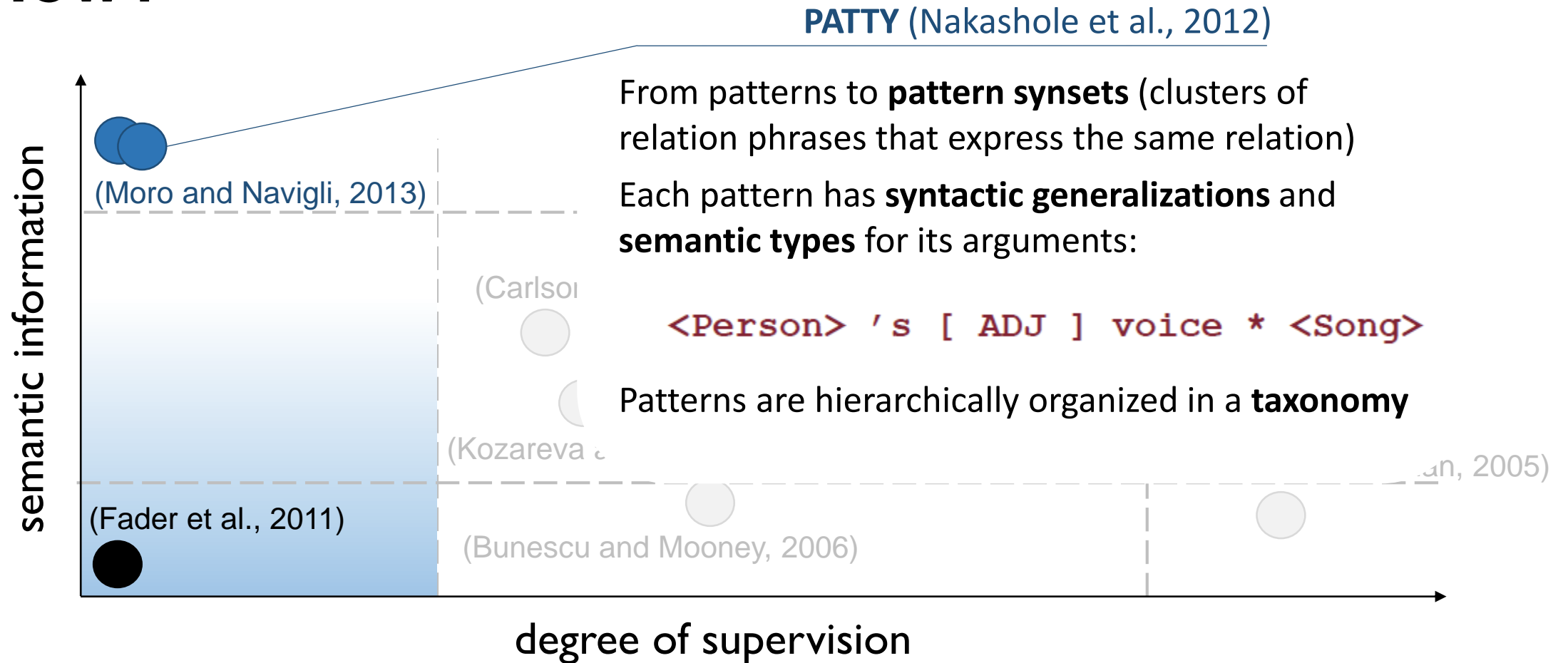
Information Extraction

How?



Information Extraction

How?



Information Extraction

How?

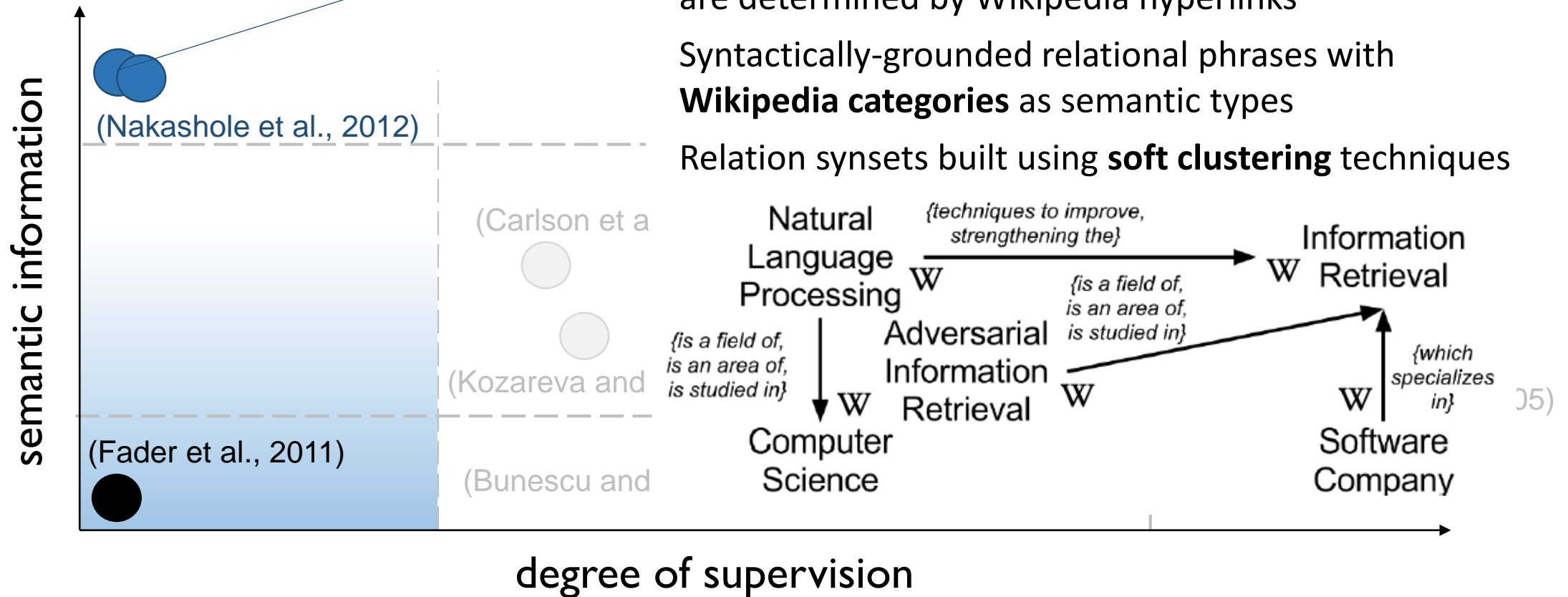
WiSeNet (Moro and Navigli, 2013)

Wikipedia-based Semantic Network: triples in the KB are determined by Wikipedia hyperlinks

Syntactically-grounded relational phrases with

Wikipedia categories as semantic types

Relation synsets built using **soft clustering** techniques



(Open) Information Extraction

OIE is great, but...

Sparsity: many relation phrases express the same relationship (e.g. synonyms, paraphrases)

Ambiguity: arguments (and relation phrases) are ambiguous!



Outline

IE and OIE: some background

DefIE: OIE from textual definitions

Claudio Delli Bovi, Luca Telesca and Roberto Navigli.

Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis.

Transactions of the Association for Computational Linguistics (TACL), 2015.



SAPIENZA
UNIVERSITÀ DI ROMA

KBUnify: KB disambiguation and unification

Delli Bovi, Espinosa-Anke, Navigli: **EMNLP 2015**



DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The idea:

instead of targeting massive and noisy corpora (like the web) and then trying to find a smart way to cope with the noise

target smaller but “denser” (and virtually noise-free) corpora of **definitional knowledge**.





DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The idea:

instead of targeting massive and noisy corpora (like the web) and then trying to find a smart way to cope with the noise

target smaller but “denser” (and virtually noise-free) corpora of **definitional knowledge**.

Apply OIE techniques to extract as much information as possible!





DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The tools:

- An underlying **inventory/knowledge base** (to which arguments and relation patterns will be connected)
- A **WSD/EL system** (to disambiguate concepts and entity mentions across the input text)
- A **syntactic parser** (to construct meaningful relation patterns and avoid sparsity)



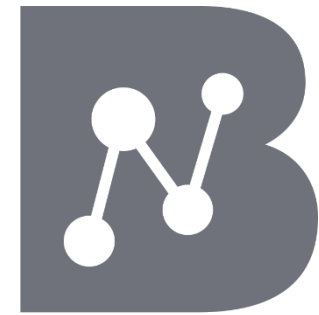
DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The tools:

- An underlying inventory/knowledge base (to which arguments and relation patterns will be connected)
- A **WSD/EL system** (to disambiguate concepts and entity mentions across the input text)
- A **syntactic parser** (to construct meaningful relation patterns and avoid sparsity)

<http://babelnet.org>



BabelNet

14 million entries
both **lexicographic**
and **encyclopedic**
knowledge



DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The tools:

- An underlying **inventory/knowledge base** (to which arguments and relation patterns will be connected)
- A **WSD/EL system** (to disambiguate concepts and entity mentions across the input text)
- A **syntactic parser** (to construct meaningful relation patterns and avoid sparsity)

<http://babelfy.org>



unified graph-based approach to **EL** and **WSD**

unsupervised, based on **BabelNet**



DefIE: OIE from textual definitions

 <http://lcl.uniroma1.it/defie>

The tools:

- An underlying **inventory/knowledge base** (to which arguments and relation patterns will be connected)
- A **WSD/EL system** (to disambiguate concepts and entity mentions across the input text)
- A **syntactic parser** (to construct meaningful relation patterns and avoid sparsity)

<http://svn.ask.it.usyd.edu.au/trac/candc>

C&C tools

log-linear parser and supertagger based on **CCG**

(theoretically) suited to **long-distance dependencies**



DefIE: How it works



<http://lcl.uniroma1.it/defie>

I. Extracting relation instances

“Atom Heart Mother is the fifth album by English band Pink Floyd.”

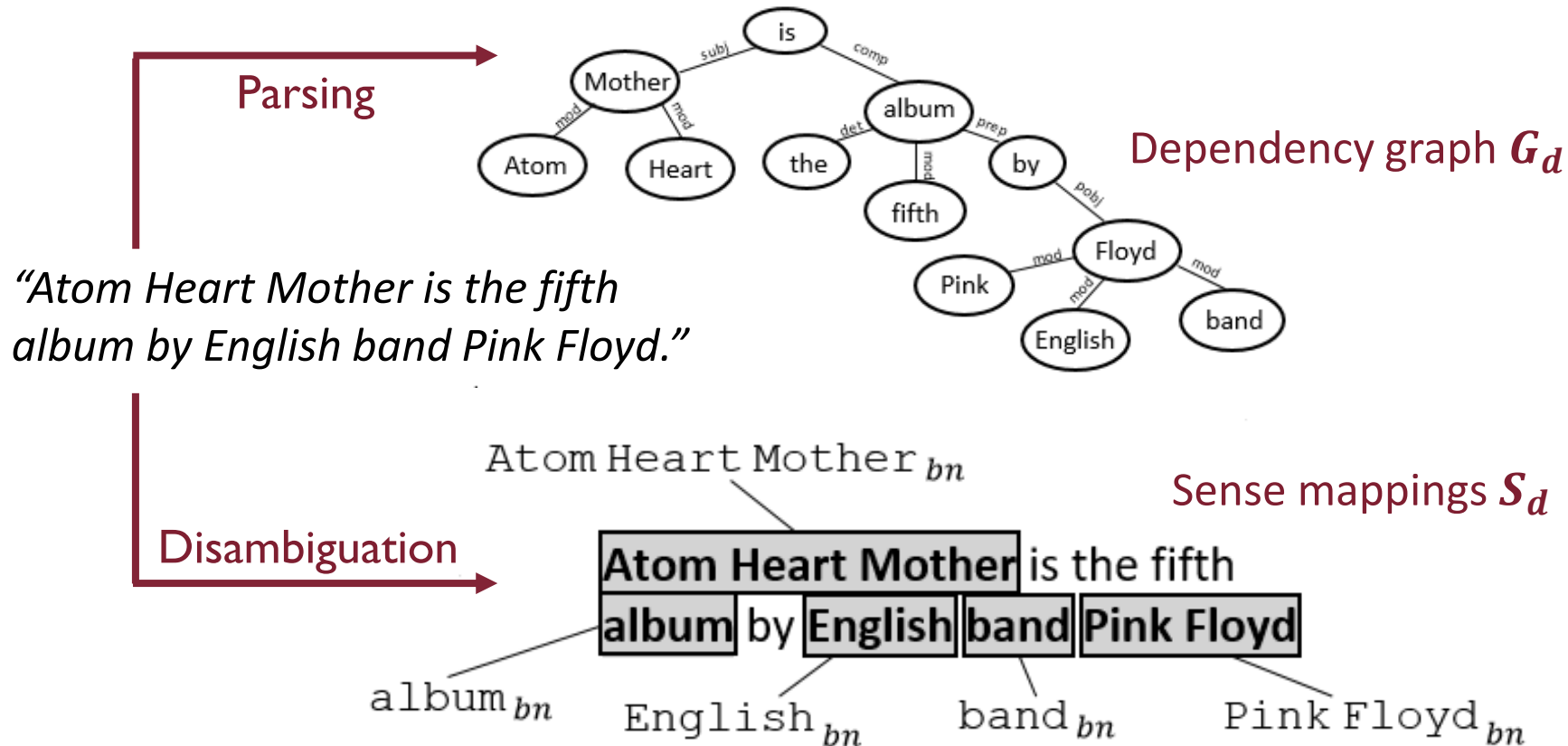
Textual definition *d*



DefIE: How it works

 <http://lcl.uniroma1.it/defie>

I. Extracting relation instances

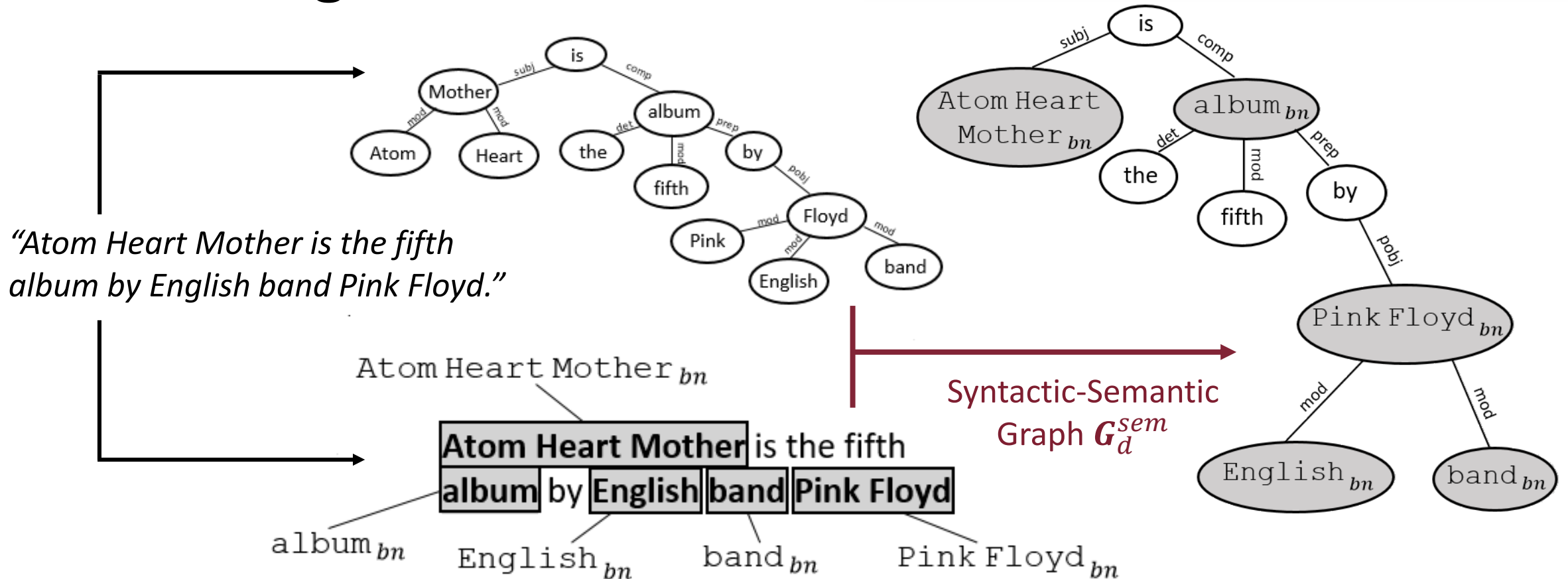




DefIE: How it works

<http://lcl.uniroma1.it/defie>

I. Extracting relation instances

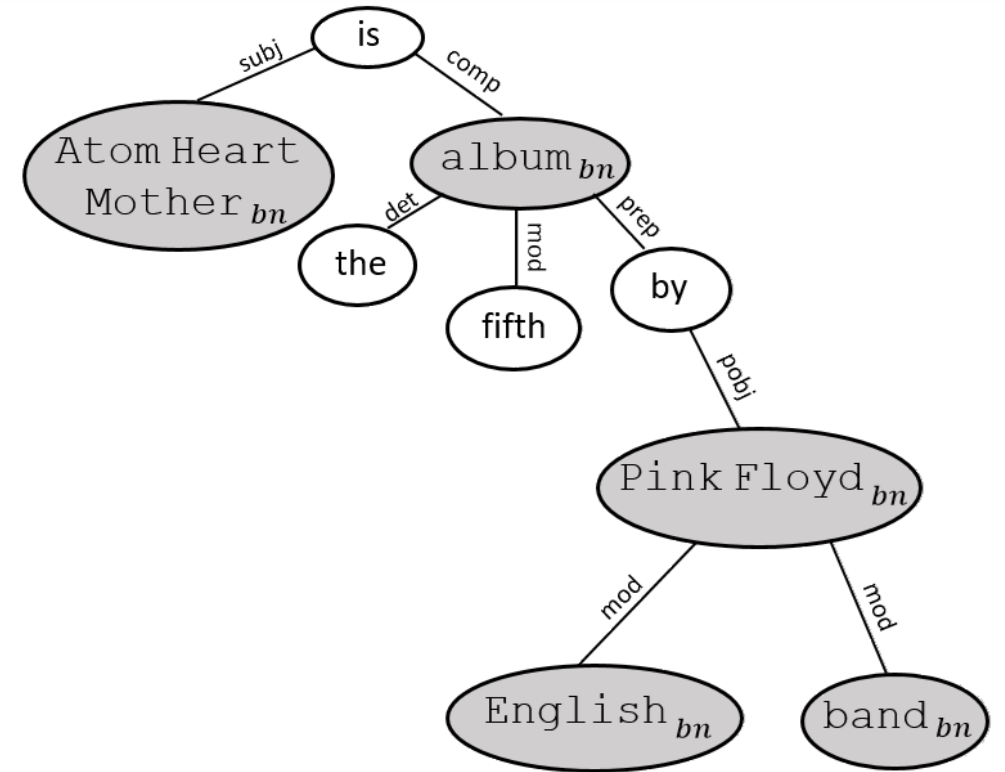




DefIE: How it works

 <http://lcl.uniroma1.it/defie>

I. Extracting relation instances





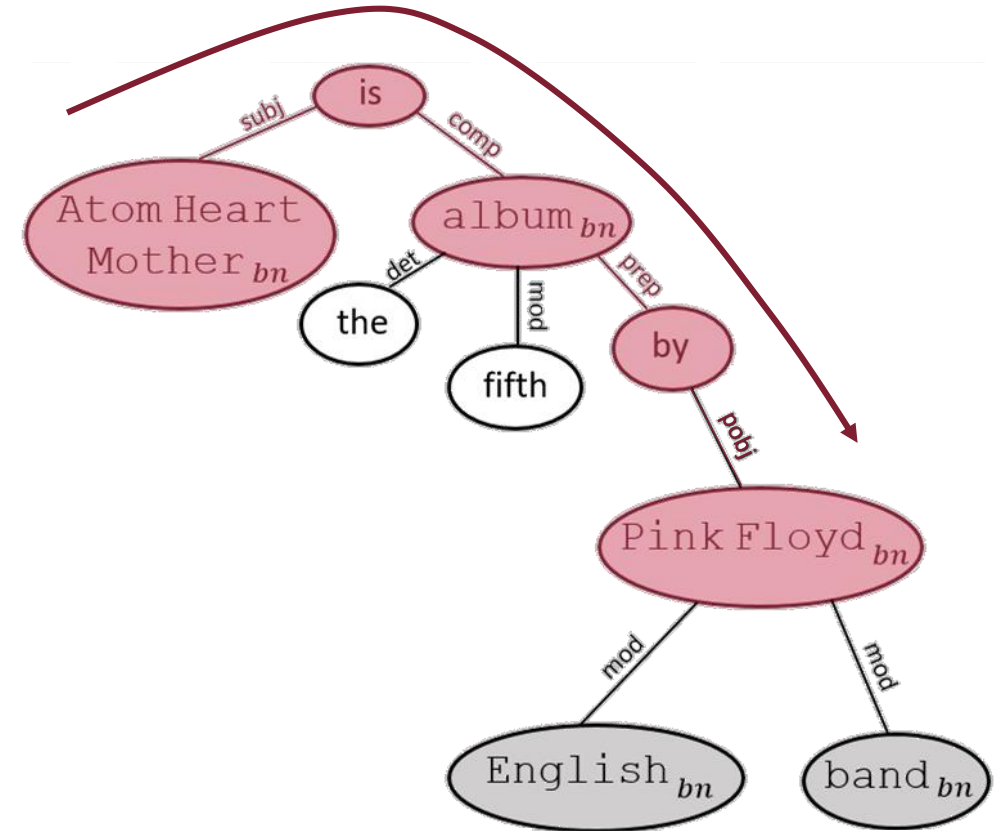
DefIE: How it works

 <http://lcl.uniroma1.it/defie>

I. Extracting relation instances

Extraction 1

$X \rightarrow \text{is} \rightarrow \text{album}_{bn}^1 \rightarrow \text{by} \rightarrow Y$
 $X = \text{Atom Heart Mother}_{bn}^1$
 $Y = \text{Pink Floyd}_{bn}^1$





DefIE: How it works

 <http://lcl.uniroma1.it/defie>

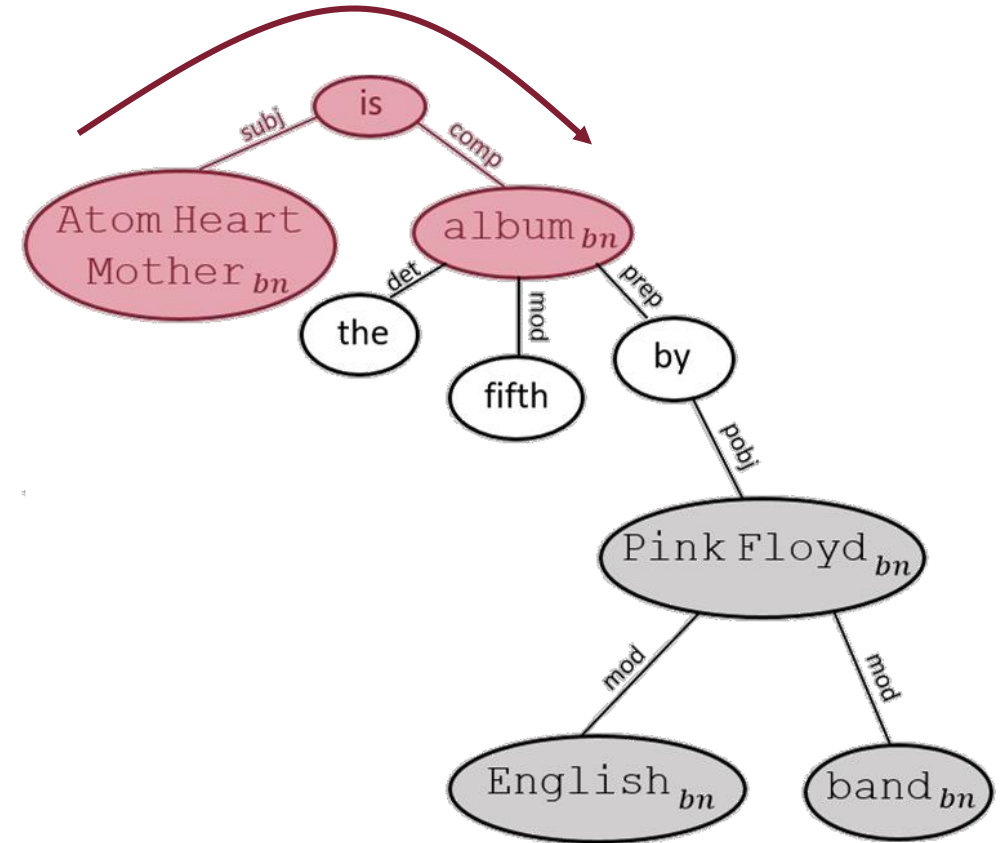
I. Extracting relation instances

Extraction 2

$$\left\{ \begin{array}{l} X \rightarrow is \rightarrow Y \\ X = \text{Atom Heart Mother}_{bn}^1 \\ Y = \text{album}_{bn}^1 \end{array} \right.$$

Extraction 1

$$\left\{ \begin{array}{l} X \rightarrow is \rightarrow \text{album}_{bn}^1 \rightarrow by \rightarrow Y \\ X = \text{Atom Heart Mother}_{bn}^1 \\ Y = \text{Pink Floyd}_{bn}^1 \end{array} \right.$$





DefIE: How it works



<http://lcl.uniroma1.it/defie>

2. Relation typing and scoring



DefIE: How it works



<http://lcl.uniroma1.it/defie>

2. Relation typing and scoring

For each relation R :

Substitute each domain and range argument with its **hypernym h** (using the BabelNet taxonomy) and generate a **probability distribution over semantic types** for the two sets



DefIE: How it works



<http://lcl.uniroma1.it/defie>

2. Relation typing and scoring

For each relation R :

Substitute each domain and range argument with its **hypernym** h (using the BabelNet taxonomy) and generate a **probability distribution over semantic types** for the two sets

Compute the **entropy** of R as
$$H_R = - \sum_{i=1}^n p(h_i) \log_2 p(h_i)$$



DefIE: How it works

 <http://lcl.uniroma1.it/defie>

2. Relation typing and scoring

For each relation R :

Compute the **score** of R as

$$\text{score}(R) = \frac{|S_R|}{(H_R + 1) \text{length}(r)}$$

Total number of
extracted instances
for R

Domain and range
entropy of R

Length of the
relation pattern of R



DefIE: How it works

 <http://lcl.uniroma1.it/defie>

2. Relation typing and scoring

Pattern	Score	Entropy
<i>X directed by Y</i>	4 025.80	1.74
<i>X known for Y</i>	2 590.70	3.65
<i>X is election district_{bn}¹ of Y</i>	110.49	0.83
<i>X is composer_{bn}¹ from Y</i>	39.92	2.08
<i>X is street_{bn}¹ named after Y</i>	1.91	2.24
<i>X is village_{bn}² founded in 1912 in Y</i>	0.91	0.18



DefIE: How it works



<http://lcl.uniroma1.it/defie>

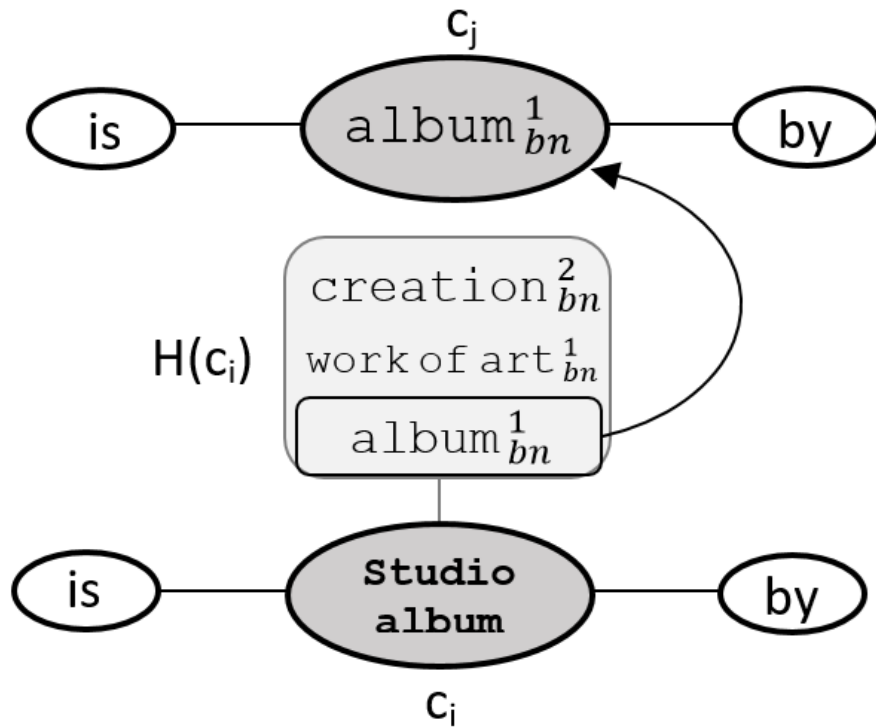
3. Relation taxonomization



DefIE: How it works

 <http://lcl.uniroma1.it/defie>

3. Relation taxonomization



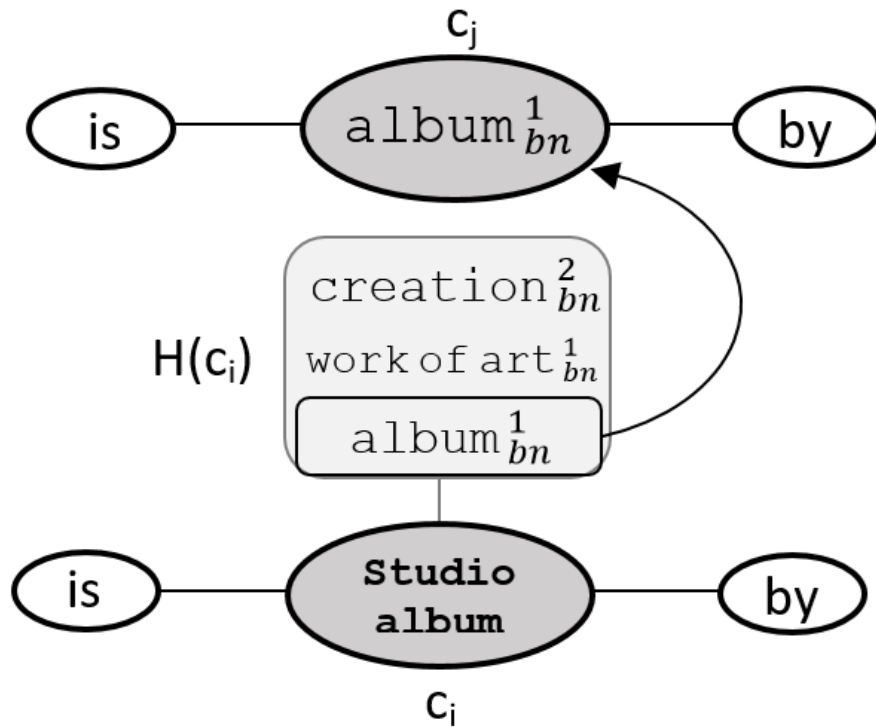
Hypernym Generalization



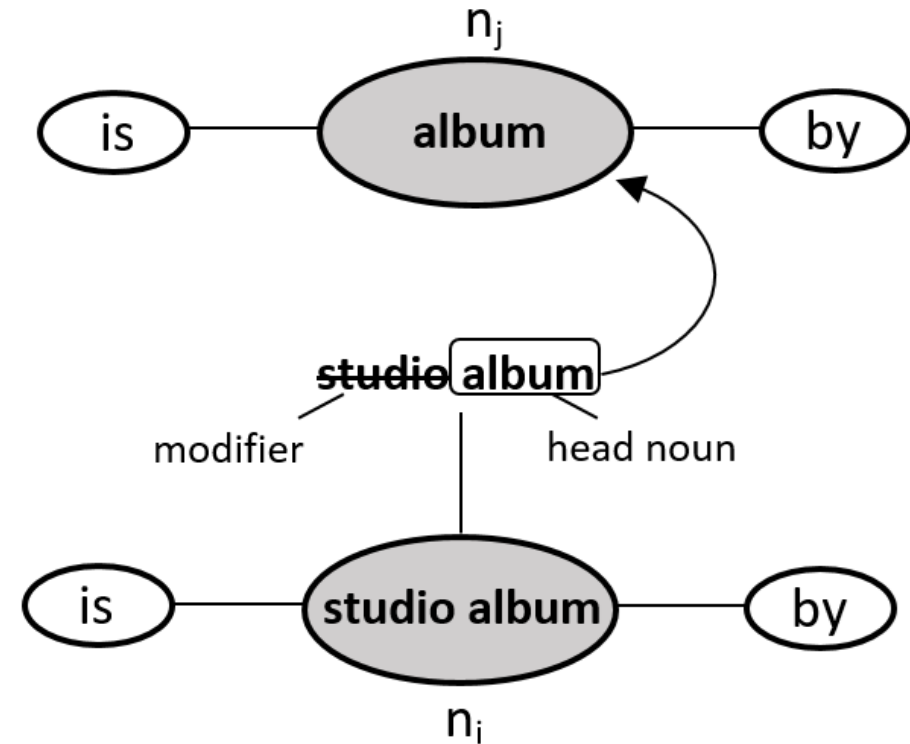
DefIE: How it works

 <http://lcl.uniroma1.it/defie>

3. Relation taxonomization



Hypernym Generalization



Substring Generalization



DeflE: Results

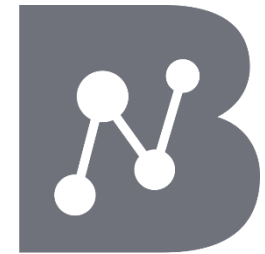


<http://lcl.uniroma1.it/defie>

Dataset:

whole set of English textual definitions in BabelNet 2.5

4 357 327 items from **5** different sources (Wikipedia, WordNet, Wikidata, Wiktionary, OmegaWiki)



BabelNet



DefIE: Results



<http://lcl.uniroma1.it/defie>

	DefIE	NELL	PATTY	ReVerb	WiSeNet
# Relations	255 881	298	1 631 531	664 746	245 935
Avg. extractions	81.68	7 013.03	9.68	22.16	9.24
# Extractions	20 352 903	2 089 883	15 802 946	14 728 268	2 271 807
# Entities	2 398 982	1 996 021	1 087 907	3 327 425	1 636 307
# Edges in the taxonomy	44 412	-	20 339	-	-



DefIE: Results



<http://lcl.uniroma1.it/defie>

Other evaluations:

- **Precision** and **coverage** of relations
- **Novelty** of information
- Quality of relation **taxonomization**
- Quality of **entity linking/disambiguation**
- **Impact** of definition sources

...



DefIE: Results



<http://lcl.uniroma1.it/defie>

Other evaluations:

- **Precision** and **coverage** of relations
- **Novelty** of information
- Quality of relation **taxonomization**
- Quality of **entity linking/disambiguation**
- **Impact** of definition sources

...

Data and output soon available for download on the website!



Outline

IE and OIE: some background

DefIE: OIE from textual definitions

Delli Bovi, Telesca, Navigli: **TACL** (to appear)

KBUnify: KB disambiguation and unification

Claudio Delli Bovi, Luis Espinosa-Anke and Roberto Navigli.

Knowledge Base Unification via Sense Embeddings and Disambiguation.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language

Processing (EMNLP), pages 726–736, Lisbon, Portugal, 17-21 September 2015.



SAPIENZA
UNIVERSITÀ DI ROMA



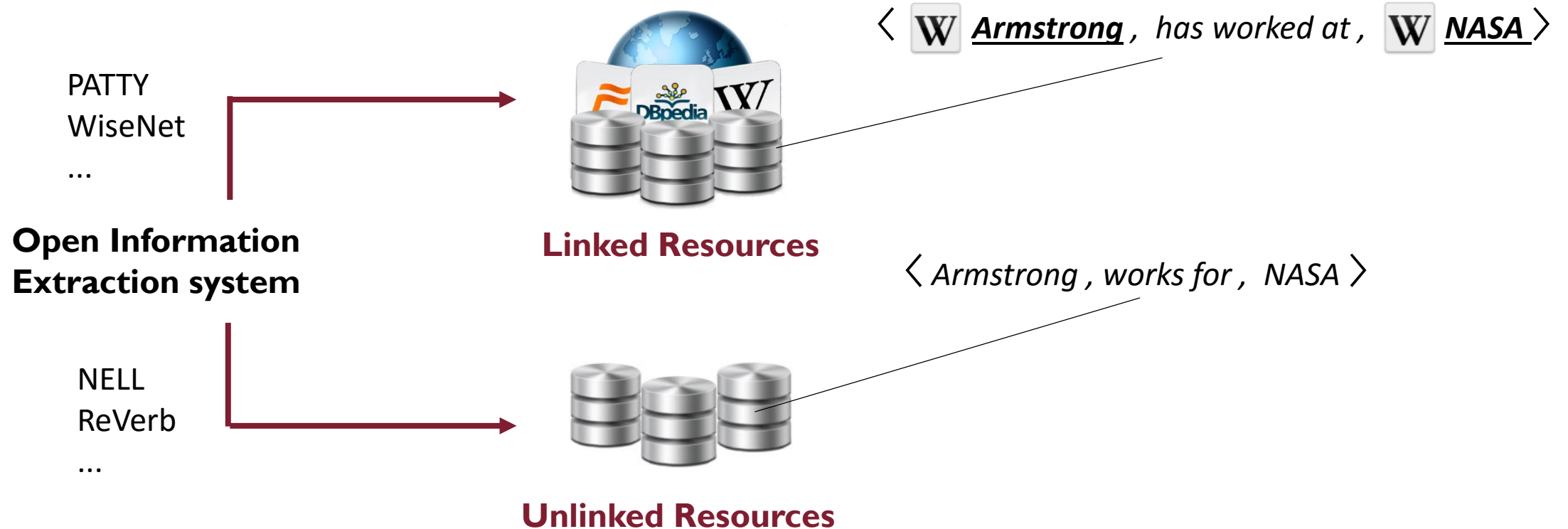
UNIVERSITAT
POMPEU FABRA



KB-Unify: Knowledge base unification via sense embeddings and disambiguation

 <http://lcl.uniroma1.it/kb-unify>

The idea:

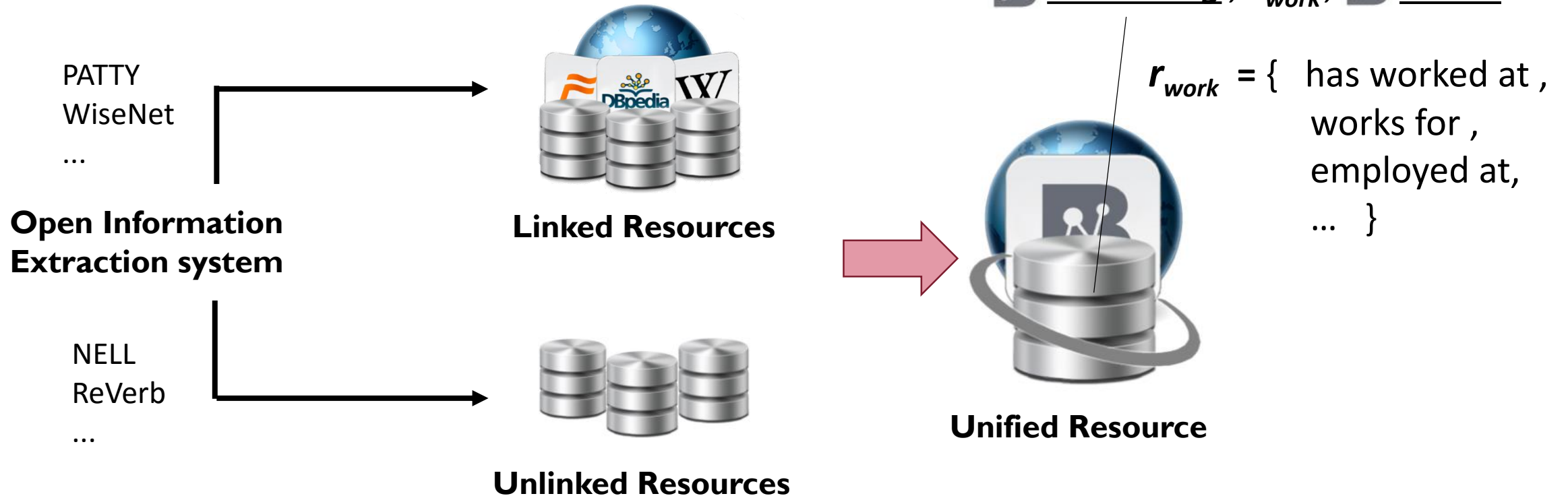




KB-Unify: Knowledge base unification via sense embeddings and disambiguation

 <http://lcl.uniroma1.it/kb-unify>

The idea:





KB-Unify: Knowledge base unification via sense embeddings and disambiguation

 <http://lcl.uniroma1.it/kb-unify>

The tools:

- A **WSD/EL system** (to disambiguate unlinked resources)
- A unified **sense inventory \mathbf{S}** (to make the various resources “speak to each other”)
- A unified **vector space \mathbf{V}_S** (to associate a vector with each item of \mathbf{S})



KB-Unify: Knowledge base unification via sense embeddings and disambiguation

 <http://lcl.uniroma1.it/kb-unify>

The tools:

- A WSD/EL system (to disambiguate unlinked resources)



Babelfy

- A unified sense inventory \mathbf{S} (to make the various resources “speak to each other”)



Babelnet

- A unified **vector space \mathbf{V}_S** (to associate a vector with each item of \mathbf{S})



KB-Unify: Knowledge base unification via sense embeddings and disambiguation

 <http://lcl.uniroma1.it/kb-unify>

The tools:

- A **WSD/EL system** (to disambiguate unlinked resources)
- A unified **sense inventory \mathbf{S}** (to make the various resources “speak to each other”)
- A unified **vector space \mathbf{V}_S** (to associate a vector with each item of \mathbf{S})

SensEmbed

(Iacobacci et al., 2015)

Sense-based embedding model

Popular word2vec architecture (**skip-gram**) trained on a **sense-annotated corpus**

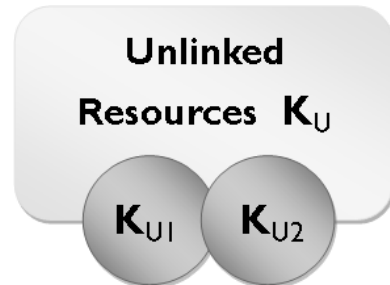
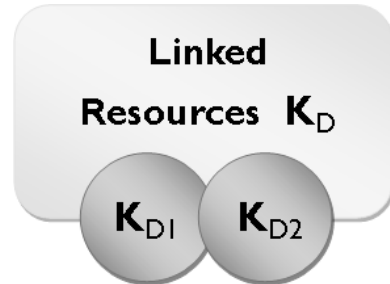


KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>

A bird's-eye view

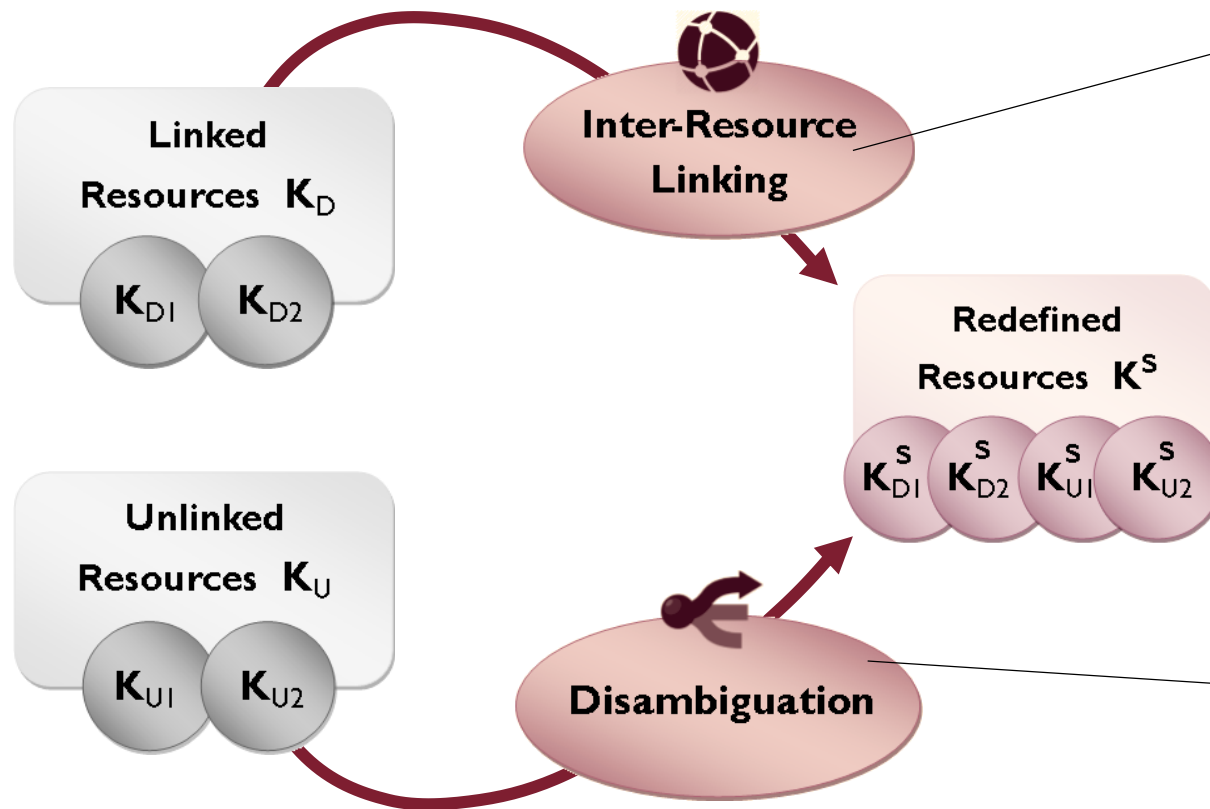




KB-Unify: How it works

 <http://lcl.uniroma1.it/kb-unify>

A bird's-eye view



use **BabelNet** mappings to
redefine each linked resource

disambiguate each unlinked
resource using Babelnet as
sense inventory (**more on this
later!**)



KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>



Disambiguation



KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>



Disambiguation

Two basic intuitions:

1. Among all triples in target knowledge base, some of them (even if ambiguous) will be **easier to disambiguate**;

e.g. < Armstrong , works for , NASA >



KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>



Disambiguation

Two basic intuitions:

1. Among all triples in target knowledge base, some of them (even if ambiguous) will be **easier to disambiguate**;

e.g. < Armstrong , works for , NASA >

2. In general, the disambiguation strategy should vary according to the **degree of specificity** of each relation.



KB-Unify: How it works

 <http://lcl.uniroma1.it/kb-unify>



Disambiguation

Group the set of unlinked triples by relation

For each relation r :

- Extract and disambiguate a subset of high-confidence **seed argument pairs** for r ;
- Estimate the **specificity** of r by looking at the distribution of its disambiguated seeds in the vector space \mathbf{V}_S ;
- Disambiguate the remaining argument pairs of r with Babelify either **triple-by-triple** (if r is general) or **all at once** (if r is specific).



KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>



Identifying seed argument pairs

< Armstrong ,

works for ,

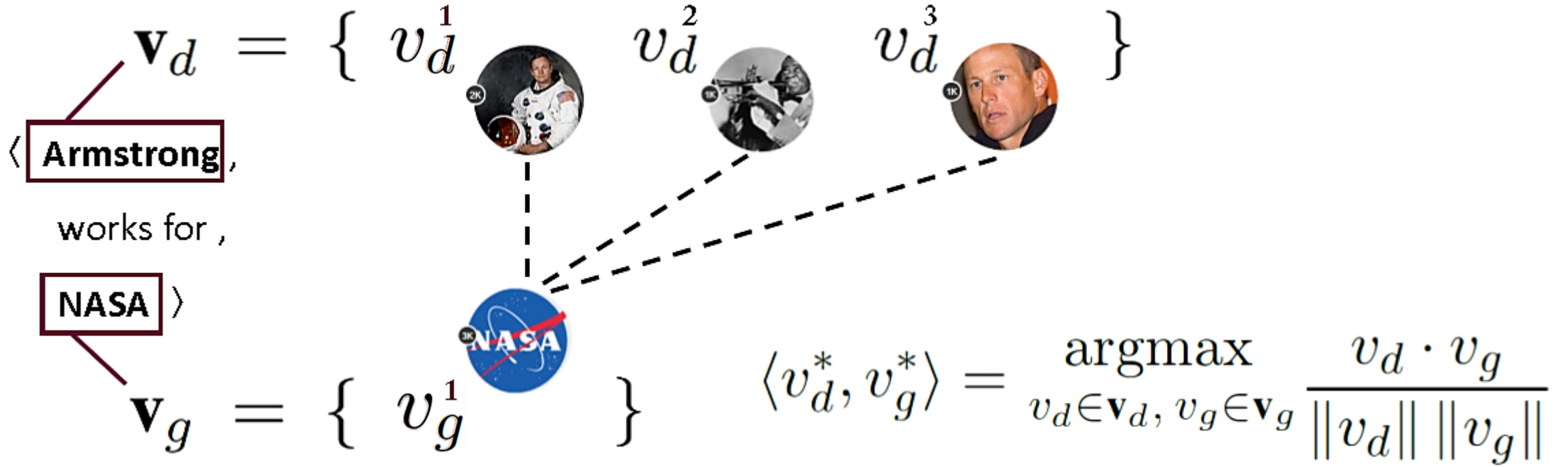
NASA >



KB-Unify: How it works

<http://lcl.uniroma1.it/kb-unify>

Identifying seed argument pairs

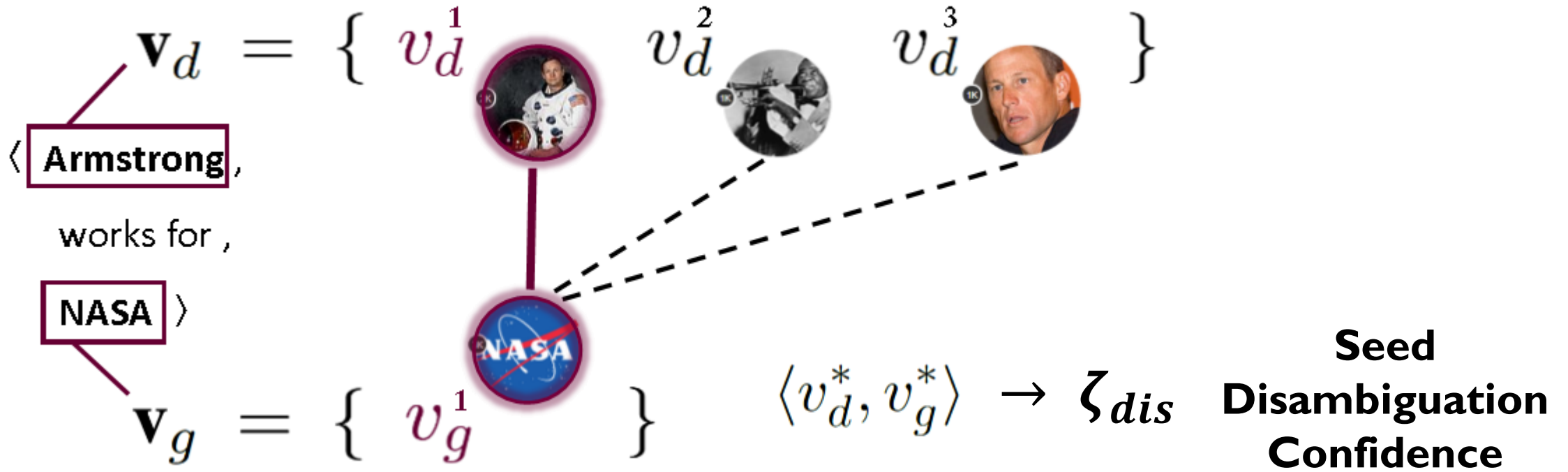




KB-Unify: How it works

<http://lcl.uniroma1.it/kb-unify>

Identifying seed argument pairs





KB-Unify: How it works



<http://lcl.uniroma1.it/kb-unify>



Ranking relations by specificity

$$\mu_k = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} \frac{v}{\|v\|}, \quad k \in \{D, G\}$$

Domain/Range Centroids

$$\sigma_k^2 = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} (1 - \cos(v, \mu_k))^2$$

Domain/Range Variances



KB-Unify: How it works

 <http://lcl.uniroma1.it/kb-unify>



Ranking relations by specificity

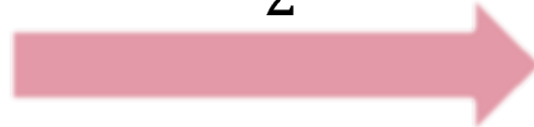
$$\mu_k = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} \frac{v}{\|v\|}, \quad k \in \{D, G\}$$

Domain/Range Centroids

$$\sigma_k^2 = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} (1 - \cos(v, \mu_k))^2$$

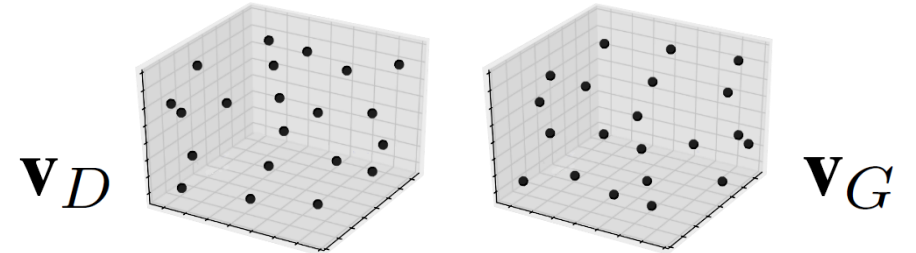
Domain/Range Variances

$$Gen(r) = \frac{\sigma_D^2 + \sigma_G^2}{2}$$

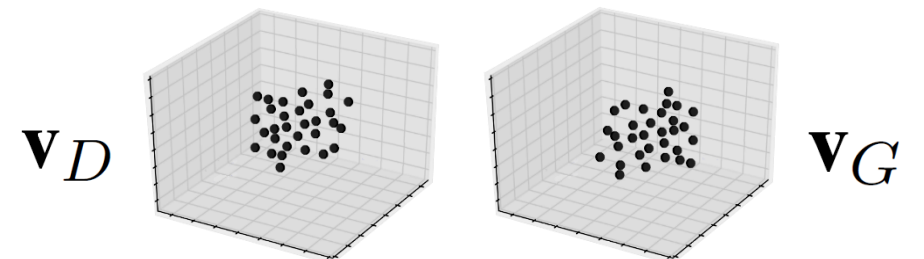


Specificity
threshold:
 δ_{spec}

High $Gen(r) (> \delta_{spec})$



Low $Gen(r) (\leq \delta_{spec})$

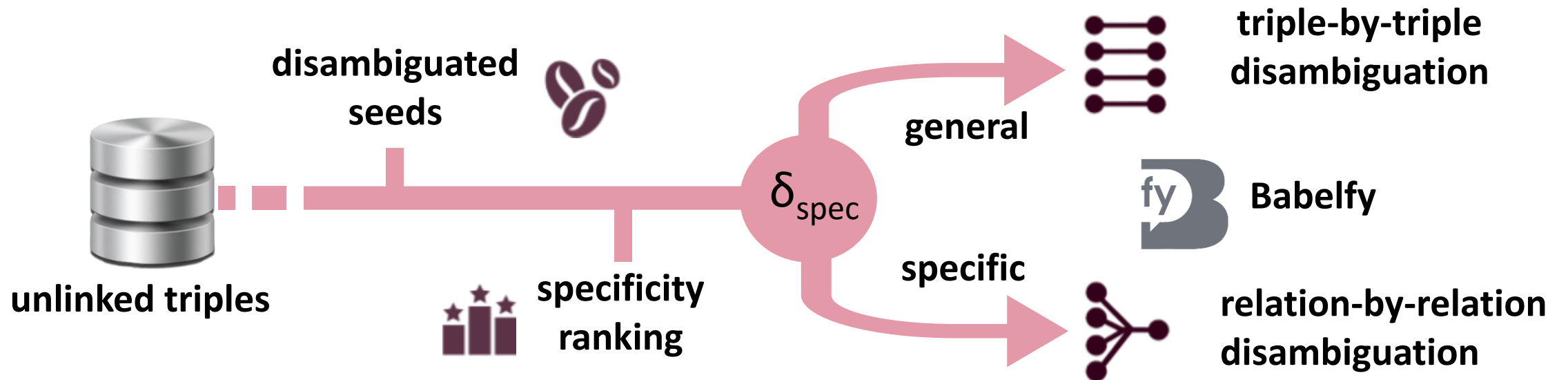




KB-Unify: How it works

 <http://lcl.uniroma1.it/kb-unify>

Disambiguation with Relation Context

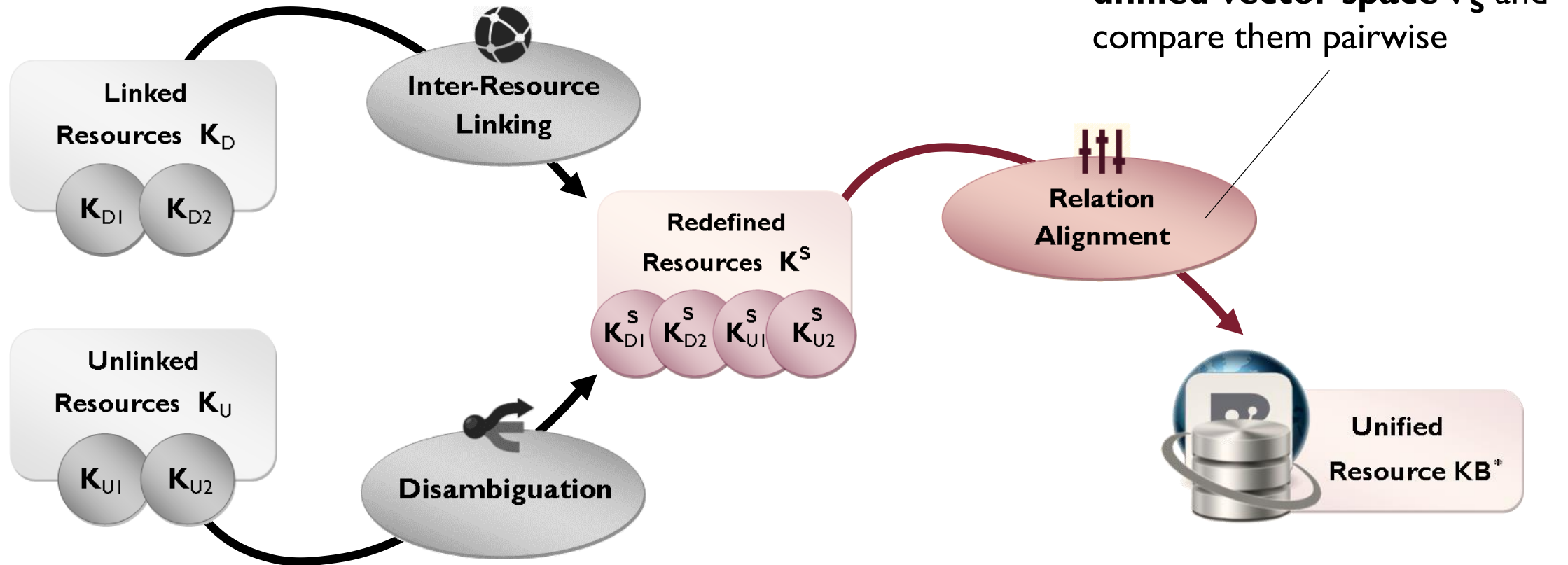




KB-Unify: How it works

<http://lcl.uniroma1.it/kb-unify>

A bird's-eye view





KB-Unify: How it works



<http://icl.uniroma1.it/kb-unify>



Relation alignment



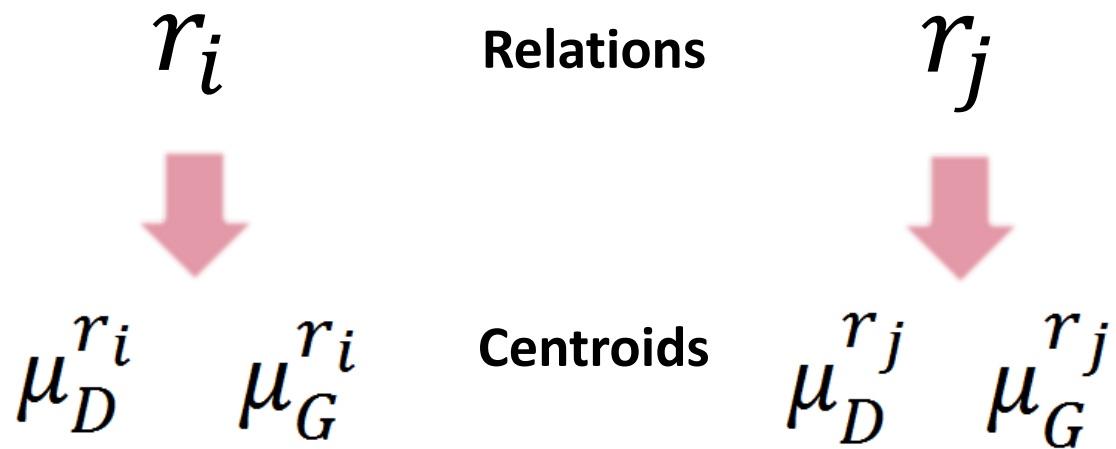
KB-Unify: How it works



<http://icl.uniroma1.it/kb-unify>

Relation alignment

For each relation pair $\langle r_i, r_j \rangle$:



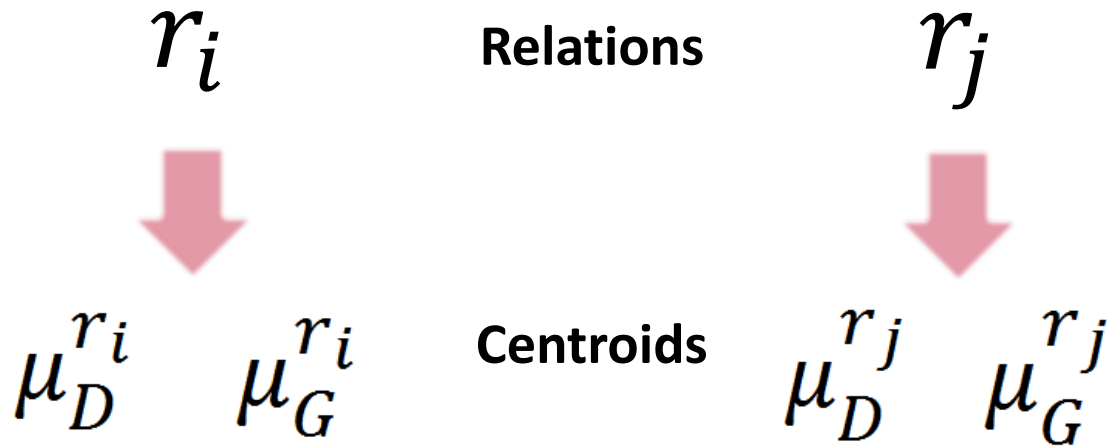


KB-Unify: How it works

<http://lcl.uniroma1.it/kb-unify>

Relation alignment

For each relation pair $\langle r_i, r_j \rangle$:



Compare domain and range centroids pairwise:

$$s_k = \frac{\mu_k^{r_i} \cdot \mu_k^{r_j}}{\|\mu_k^{r_i}\| \|\mu_k^{r_j}\|}$$

$k \in \{D, G\}$

Relation Centroid Similarity



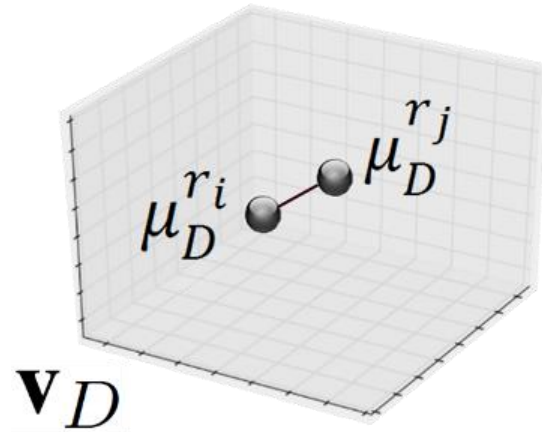
KB-Unify: How it works

 <http://icl.uniroma1.it/kb-unify>

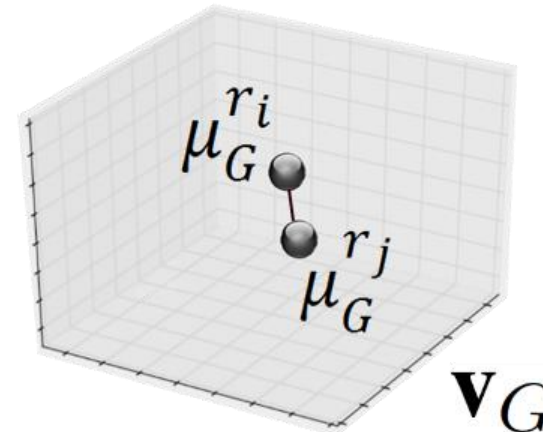
Relation alignment

Fix a similarity threshold δ_{align} :

Domain
Centroids



Range
Centroids



$\frac{1}{2} (s_D + s_G) \geq \delta_{align}$? Align r_i and r_j and merge them in the same cluster



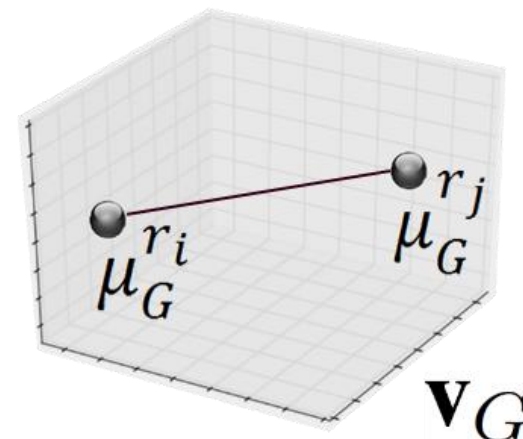
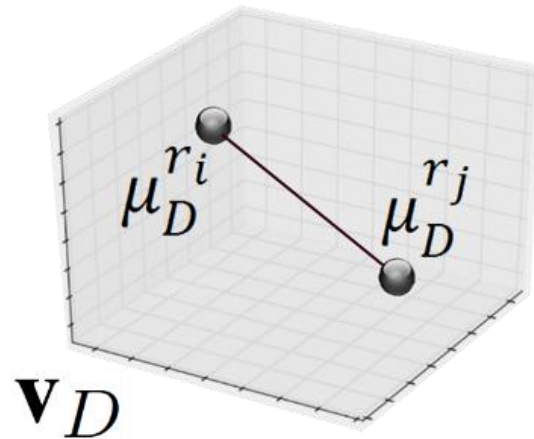
KB-Unify: How it works

 <http://icl.uniroma1.it/kb-unify>

Relation alignment

Fix a similarity threshold δ_{align} :

Domain
Centroids



Range
Centroids

$\frac{1}{2} (s_D + s_G) < \delta_{align}$? Leave r_i and r_j in separate clusters



KB-Unify: Experiments

 <http://lcl.uniroma1.it/kb-unify>

Evaluation

Experimental setup:

Linked Resources K_D :



1,631,531 relations
15,802,946 triples

245,935 relations
2,271,807 triples

Unlinked Resources K_U :



298 relations
2,245,050 triples

1,299,844 relations
14,728,268 triples



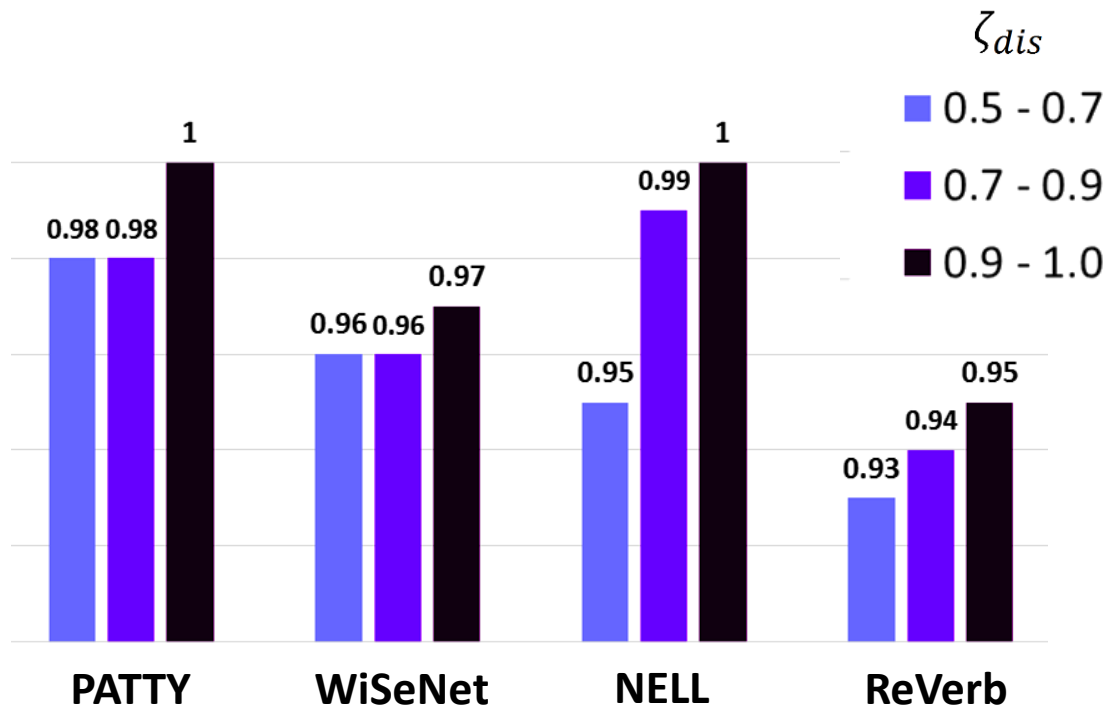
KB-Unify: Experiments

 <http://icl.uniroma1.it/kb-unify>

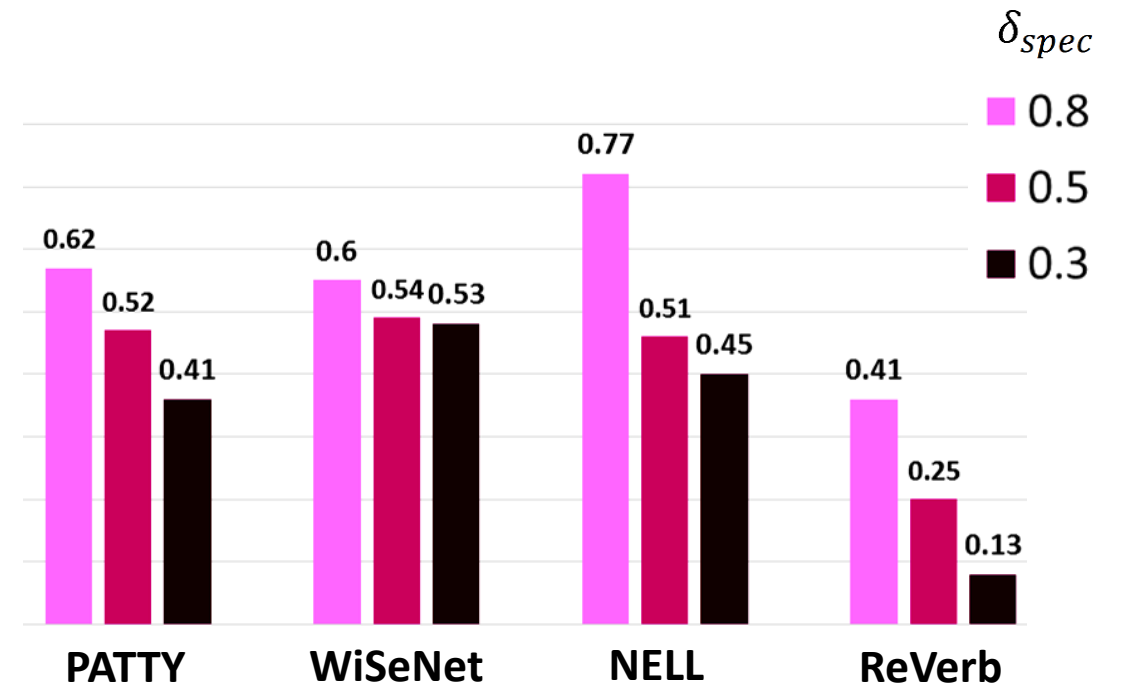


Disambiguation

Seed Precision:



Coverage:





KB-Unify: Experiments



<http://lcl.uniroma1.it/kb-unify>



Specificity ranking

For each ranked relation compute $Gen(r)$ against the **average argument similarity** \bar{s} :

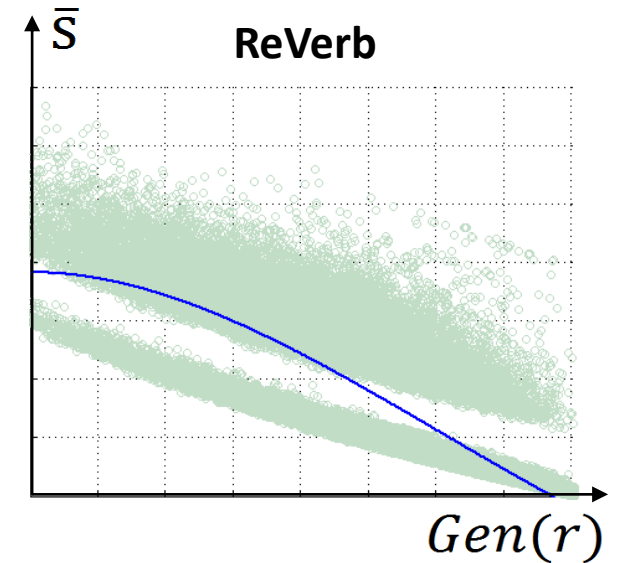
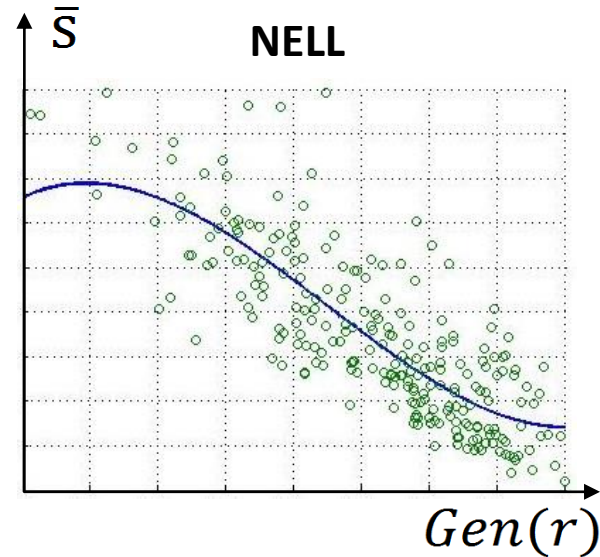
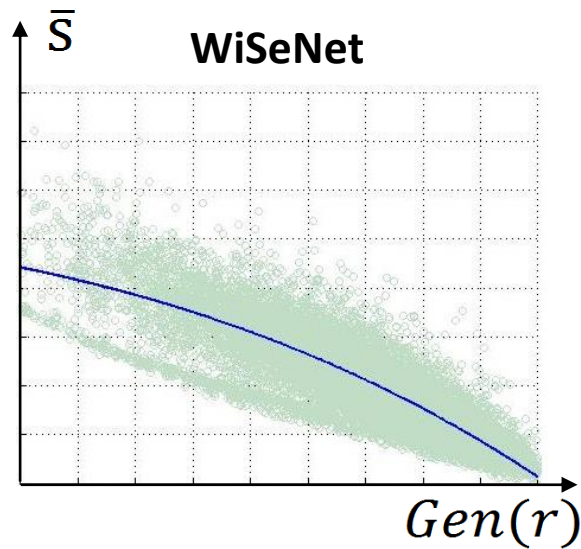
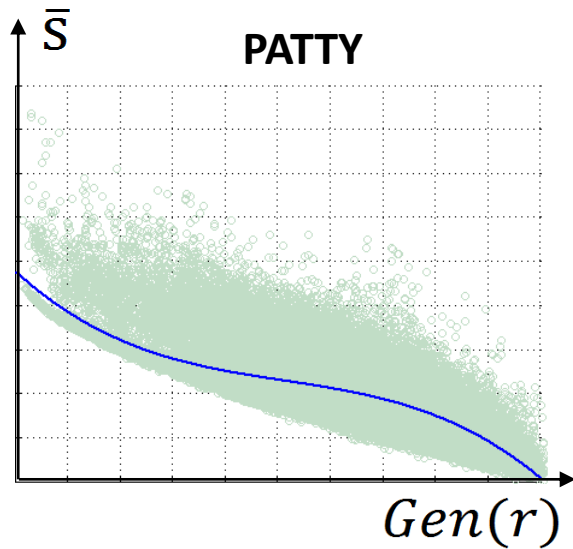


KB-Unify: Experiments

 <http://lcl.uniroma1.it/kb-unify>

Specificity ranking

For each ranked relation compute $Gen(r)$ against the **average argument similarity** \bar{s} :



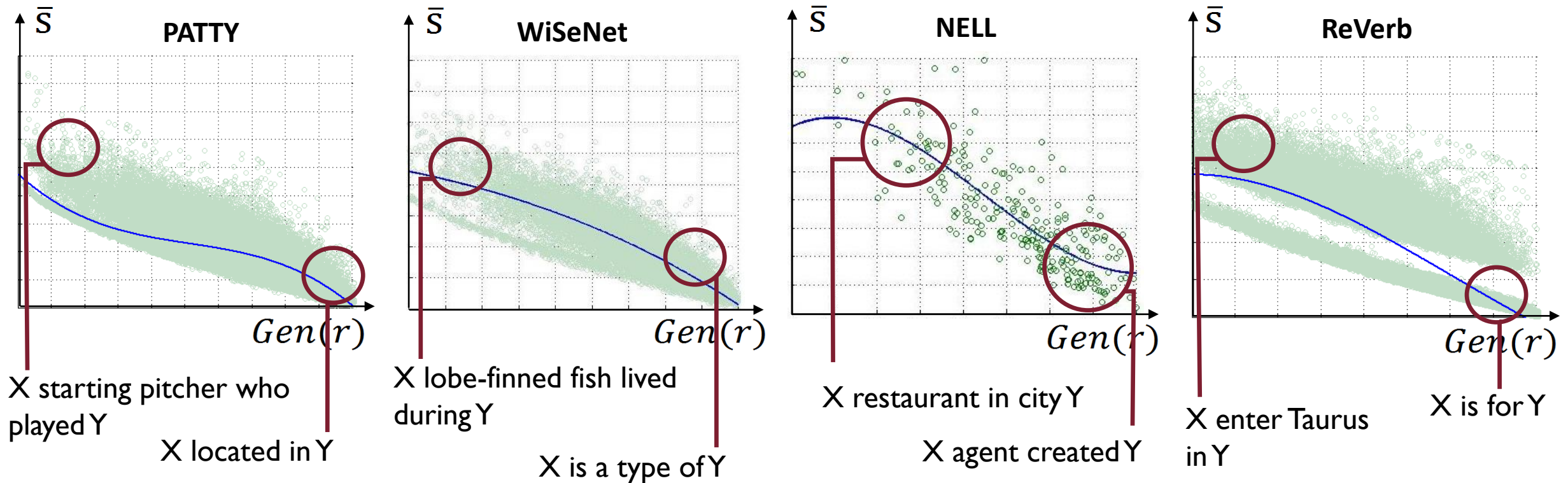


KB-Unify: Experiments

 <http://lcl.uniroma1.it/kb-unify>

Specificity ranking

For each ranked relation compute $Gen(r)$ against the **average argument similarity** \bar{s} :





KB-Unify: Experiments

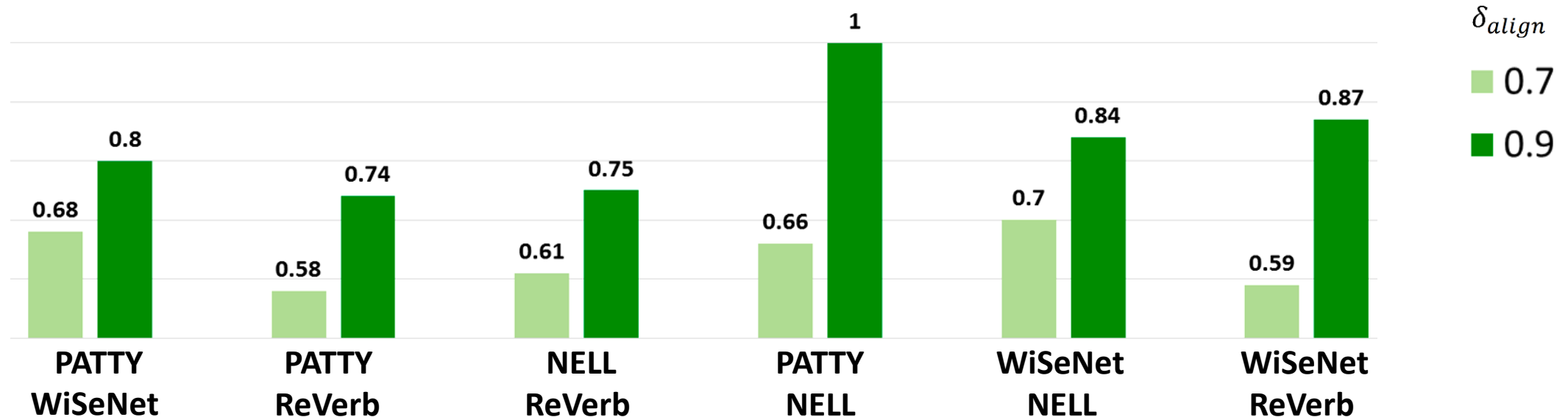


<http://lcl.uniroma1.it/kb-unify>



Cross-resource relation alignment

Samples of **150 candidate alignments** for different alignment thresholds δ_{align} manually evaluated (in terms of **paraphrasing**) by two human judges





KB-Unify: Experiments



<http://lcl.uniroma1.it/kb-unify>



Cross-resource relation alignment

Some examples:

PATTY-WISENET		ζ_{align}	NELL-PATTY		ζ_{align}
portrayed	's character	0.84	worksfor	was hired by	0.72
debuted in	first appeared in	0.86	riveremptiesintoriver	tributary of	0.89
PATTY-REVERB		ζ_{align}	NELL-WISENET		ζ_{align}
language in	is spoken in	0.81	animaleatfood	feeds on	0.72
mostly known for	plays the role of	0.70	teahomestadium	play their home games at	0.88
NELL-REVERB		ζ_{align}	REVERB-WISENET		ζ_{align}
bookwriter	is a novel by	0.88	has a selection of	offers	0.82
personleadscity	is the mayor of	0.60	had grown up in	was born and raised in	0.85

Wrap up and Conclusion

Wrap up and Conclusion



DefIE: A full-fledged OIE pipeline targeted to textual definitions, with explicit semantic characterization of both arguments and relation patterns

Wrap up and Conclusion



DefIE: A full-fledged OIE pipeline targeted to textual definitions, with explicit semantic characterization of both arguments and relation patterns



KB-Unify: An approach to knowledge base disambiguation and unification based on a shared sense inventory and a sense-based vector space model

Wrap up and Conclusion

Take-home message(s):

Web-scale OIE is absolutely great, but...

1. **Definitional knowledge is important:** sometimes it is worth it to just step back and analyze from where valuable information is extracted (**quality vs. quantity**)

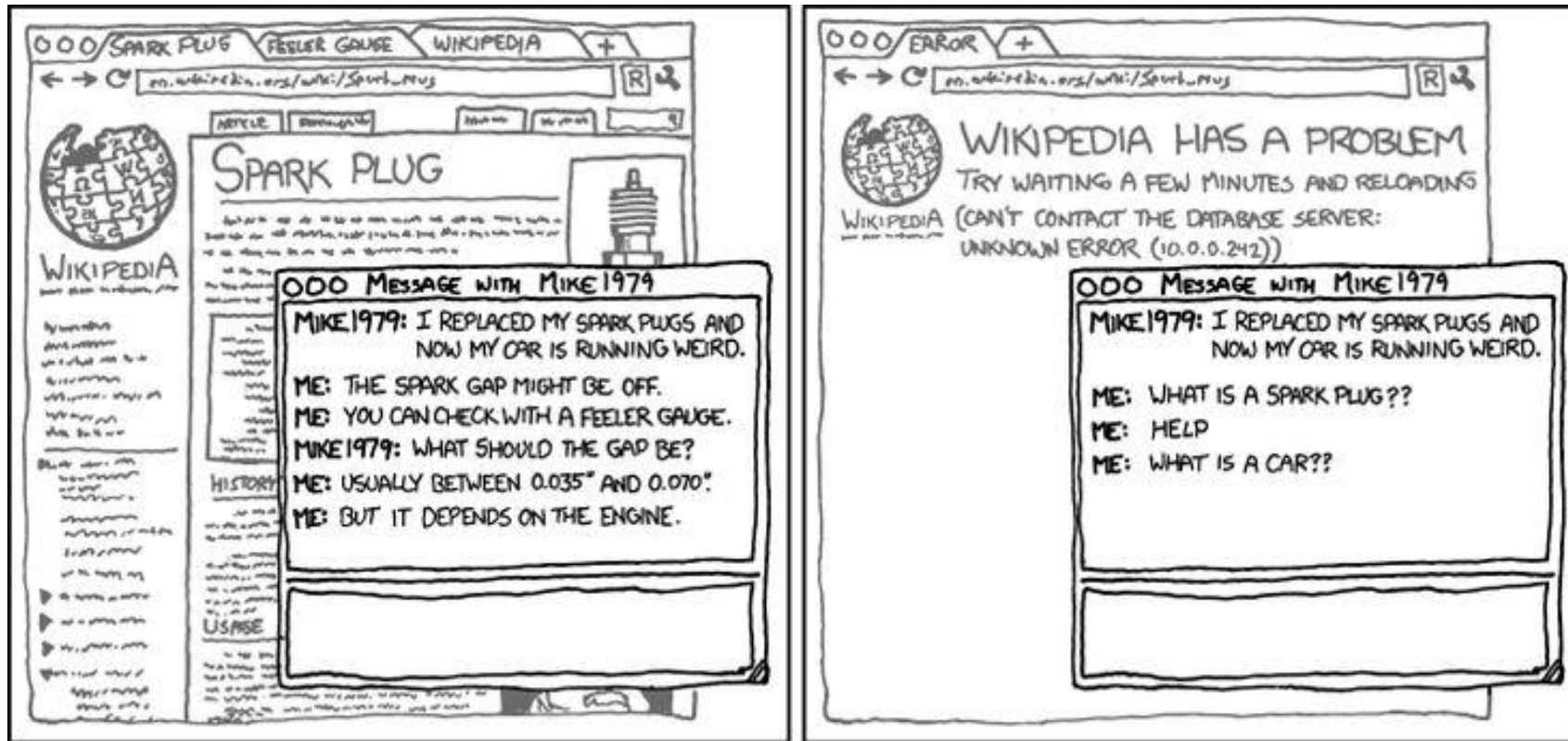
2. **Making sense of the output is important:** semantic analysis can be used to let different OIE outputs “speak to each other” and benefit from mutual enrichment

🇬🇧 Thanks!

🇪🇸 Gràcies!

🇪🇸 ¡Gracias!

🇮🇹 Grazie!



WHEN WIKIPEDIA HAS A SERVER OUTAGE, MY APPARENT IQ DROPS BY ABOUT 30 POINTS.