Karlsruhe Institute of Technology

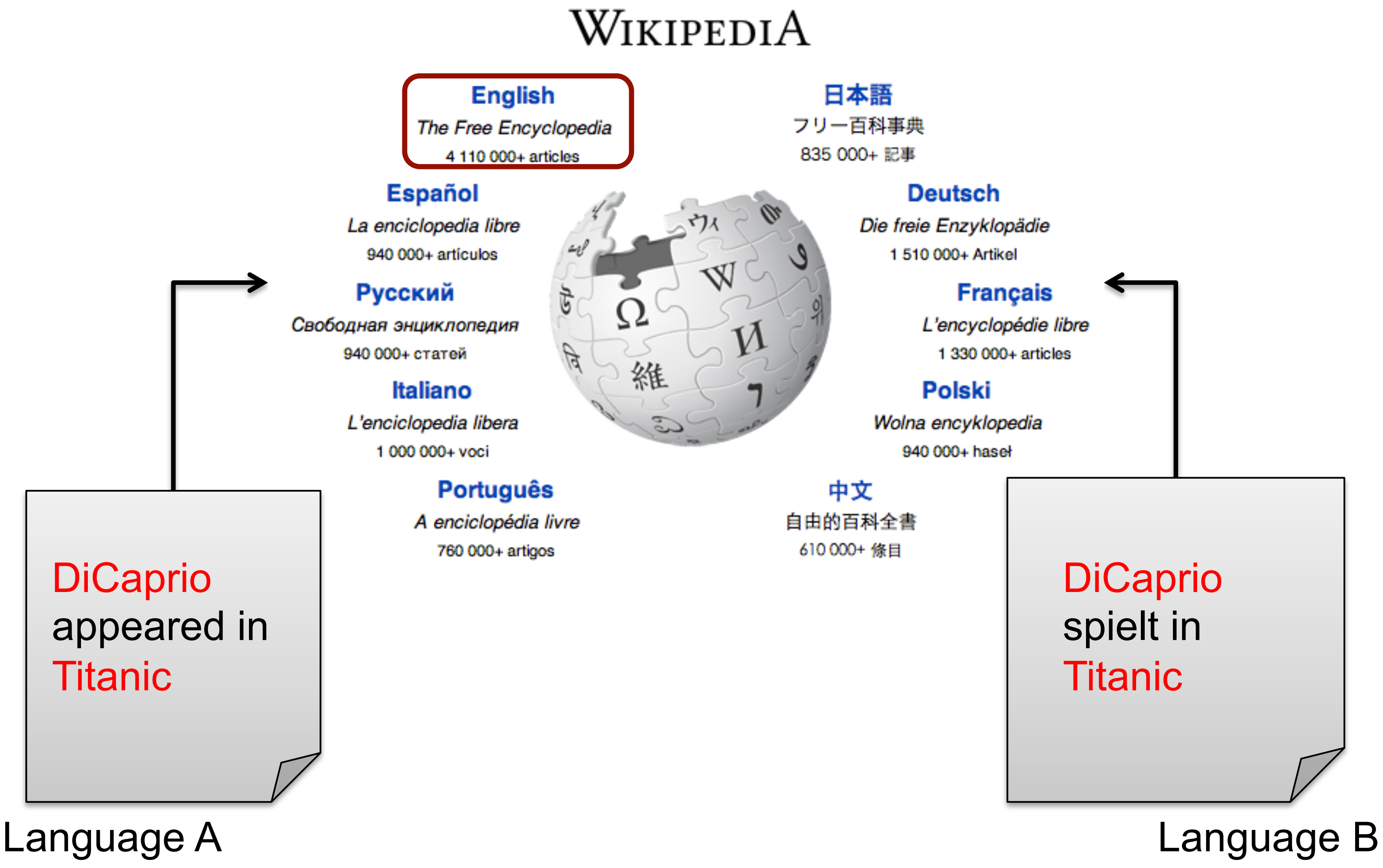
Learning a Cross-Lingual Semantic Representation of Relations Expressed in Text

Achim Rettinger, Artem Schumilin, Steffen Thoma, Basil Ell

{rettinger, steffen.thoma, basil.ell}@kit.edu, artem.schumilin@student.kit.edu

KNOWLEDGE MANAGEMENT GROUP
INSTITUTE OF APPLIED INFORMATICS AND FORMAL DESCRIPTION METHODS (AIFB)

Cross-lingual Entity Linking



WIKIPEDIA

English
The Free Encyclopedia
4 110 000+ articles

日本語
フリー百科事典
835 000+ 記事

Español
La enciclopedia libre
940 000+ artículos

Deutsch
Die freie Enzyklopädie
1 510 000+ Artikel

Русский
Свободная энциклопедия
940 000+ статей

Français
L'encyclopédie libre
1 330 000+ articles

Italiano
L'enciclopedia libera
1 000 000+ voci

Polski
Wolna encyklopedia
940 000+ haseł

Português
A enciclopédia livre
760 000+ artigos

中文
自由的百科全書
610 000+ 條目

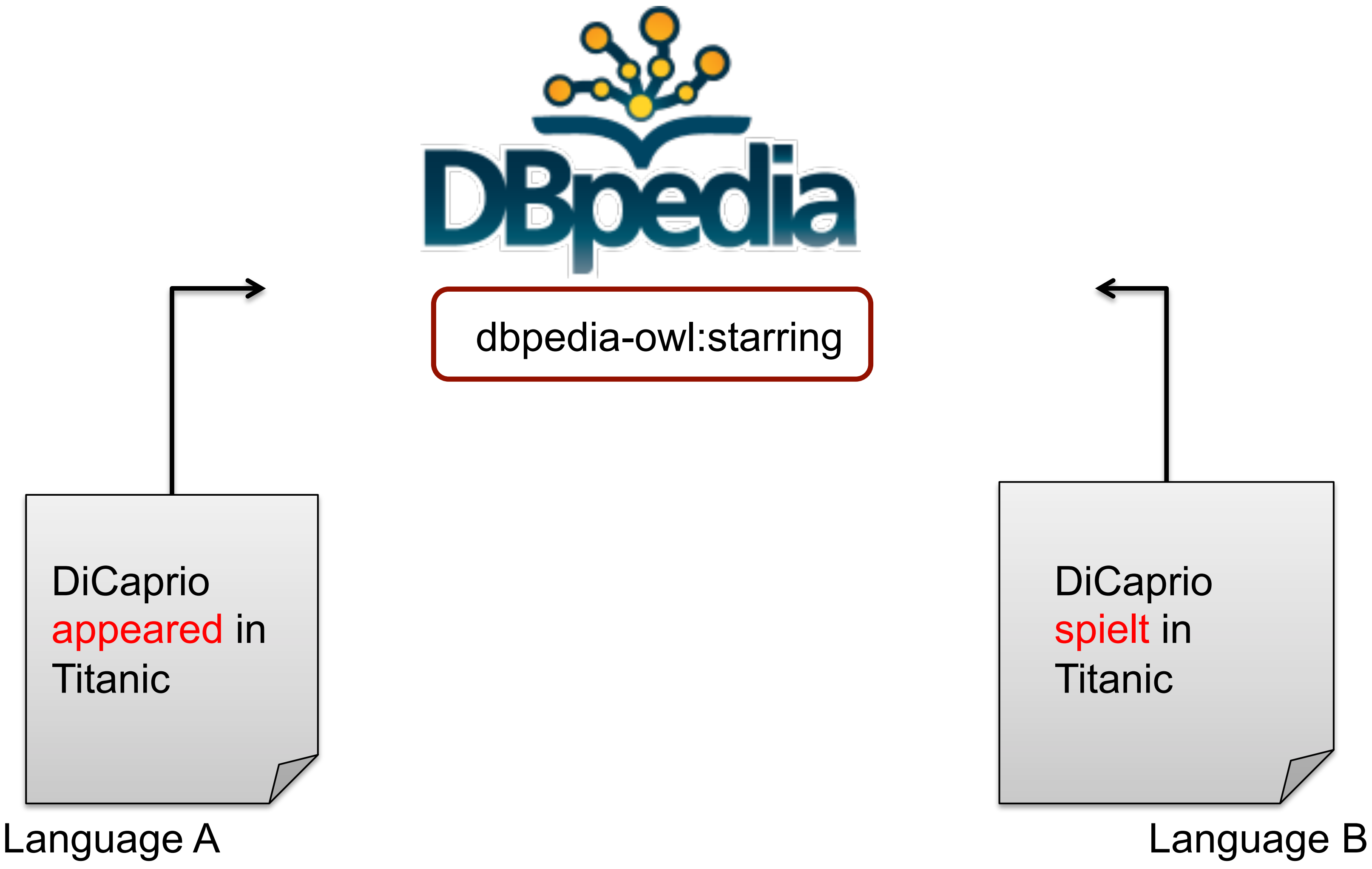
DiCaprio appeared in Titanic

Language A

DiCaprio spielt in Titanic

Language B

Cross-lingual Relation Linking



dbpedia-owl:starring

DiCaprio appeared in Titanic

Language A

DiCaprio spielt in Titanic

Language B

Entity vs. Relation Linking

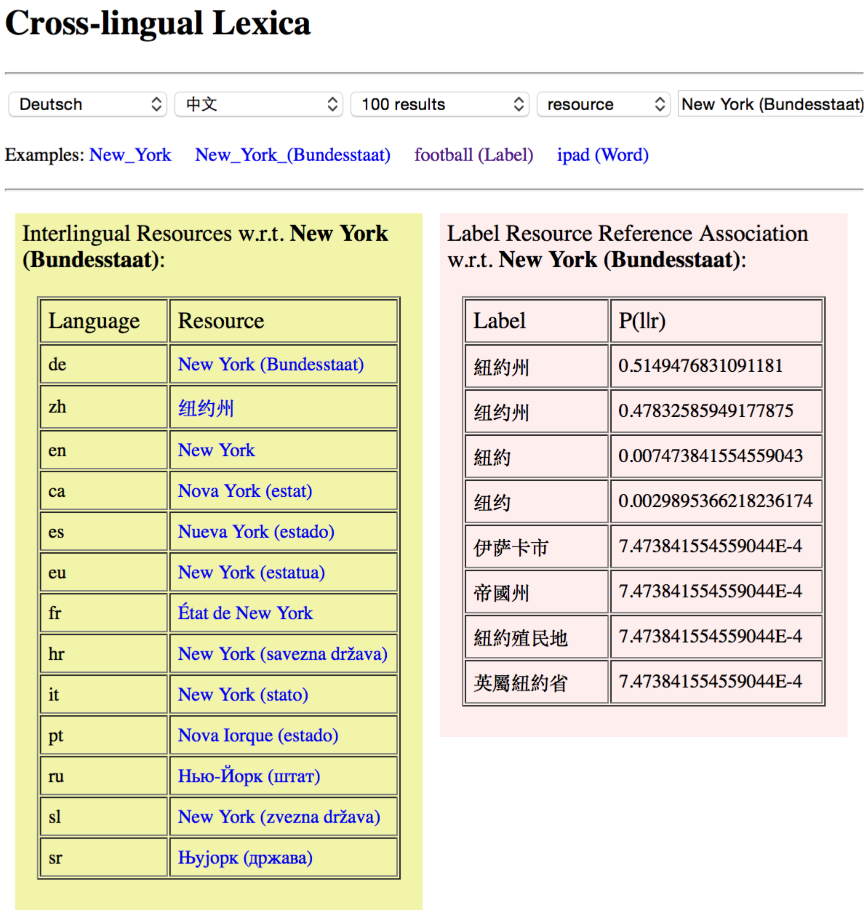
- Named Entities similar across languages
- Many non-English labels
- Lots of training data

➢ Hard

- High variability how relation can be expressed
- One English predicate label per language
- No training data

➢ Harder

Representations of Entities vs. Relations



Cross-lingual Lexica

Examples: New_York New_York_(Bundesstaat) football (Label) ipad (Word)

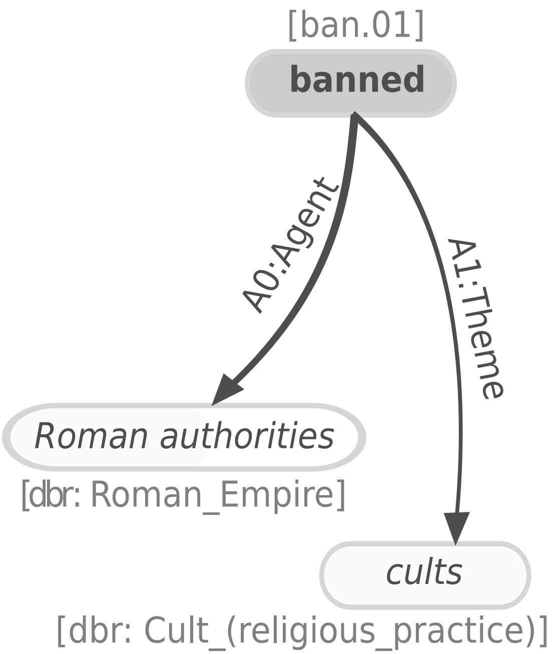
Interlingual Resources w.r.t. **New York (Bundesstaat)**:

Language	Resource
de	New York (Bundesstaat)
zh	纽约州
en	New York
ca	Nova York (estat)
es	Nueva York (estado)
eu	New York (estatua)
fr	État de New York
hr	New York (savezna država)
it	New York (stato)
pt	Nova Iorque (estado)
ru	Нью-Йорк (штат)
sl	New York (zvezna država)
sr	Њујорк (држава)

Label Resource Reference Association w.r.t. **New York (Bundesstaat)**:

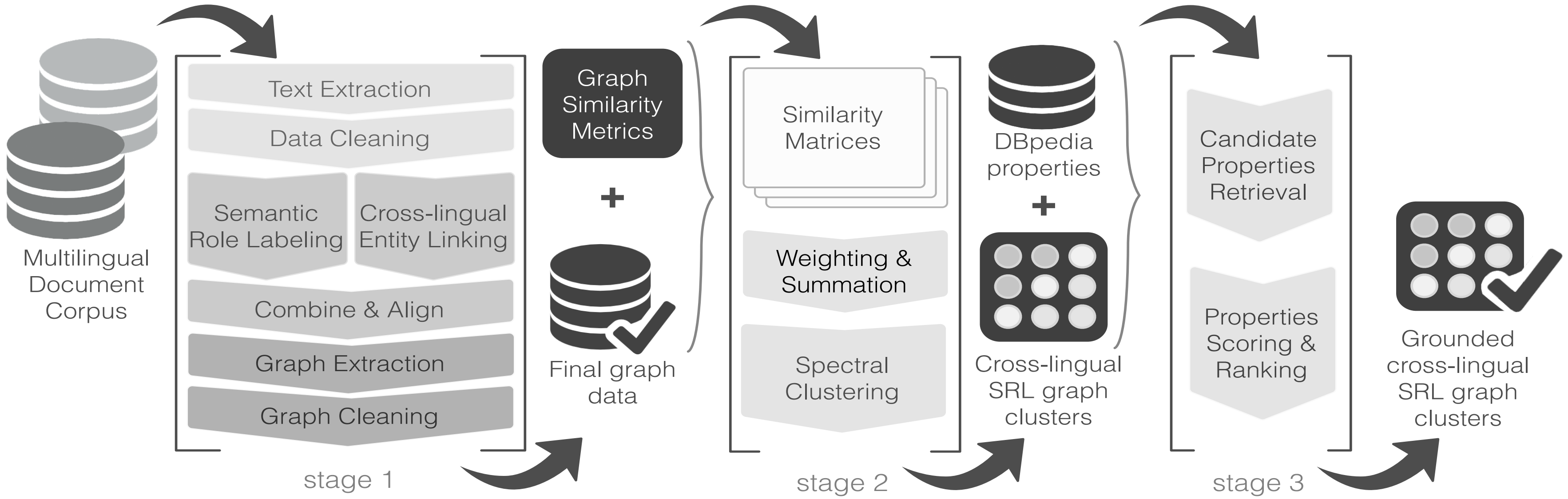
Label	P(l|r)
紐約州	0.5149476831091181
纽约州	0.47832585949177875
紐約	0.007473841554559043
纽约	0.0029895366218236174
伊萨卡市	7.473841554559044E-4
帝國州	7.473841554559044E-4
紐約殖民地	7.473841554559044E-4
英屬紐約省	7.473841554559044E-4

- Pattern formalism



- http://km.aifb.kit.edu/services/nlp-dbpedia/
- Output: List of graphs per relation

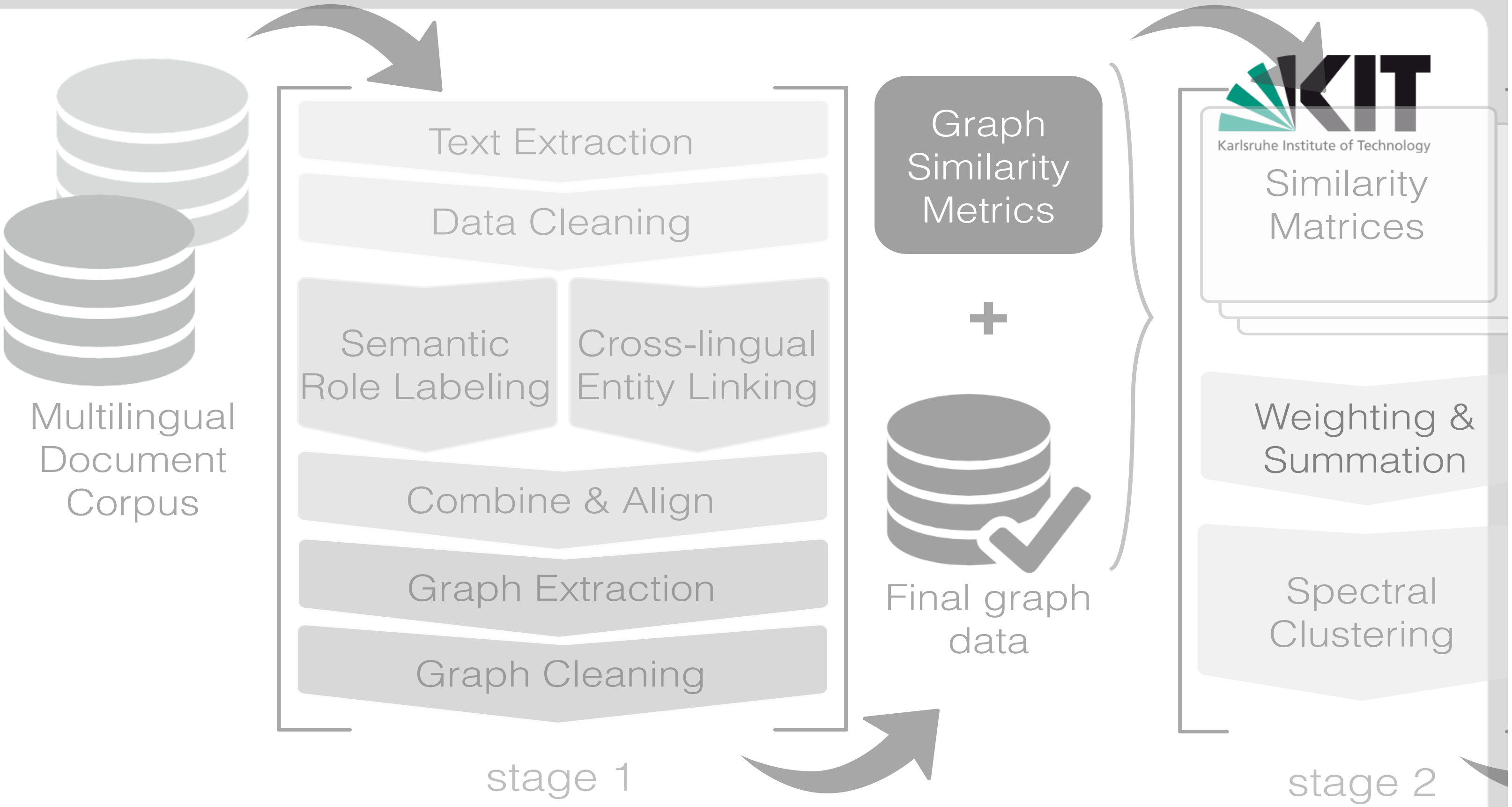
Cross-lingual Relation Clustering and Grounding - Pipeline



Stage 1. Extract cross-lingual semantic representation of predicates

Stage 2. Find similar graphs

Stage 3. Link clusters of similar graphs



Multilingual Document Corpus

Text Extraction

Data Cleaning

Semantic Role Labeling

Cross-lingual Entity Linking

Combine & Align

Graph Extraction

Graph Cleaning

stage 1

Graph Similarity Metrics

+

Final graph data

Similarity Matrices

Weighting & Summation

Spectral Clustering

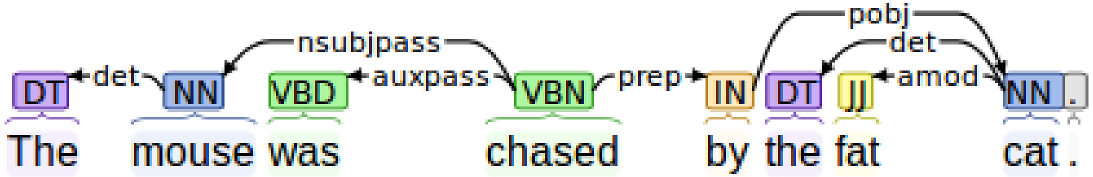
stage 2

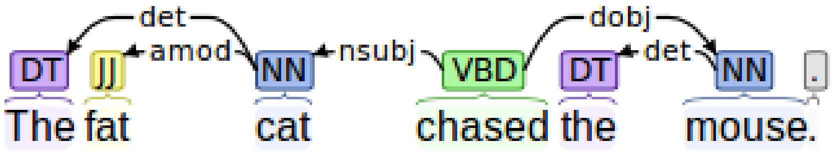
Stage 1

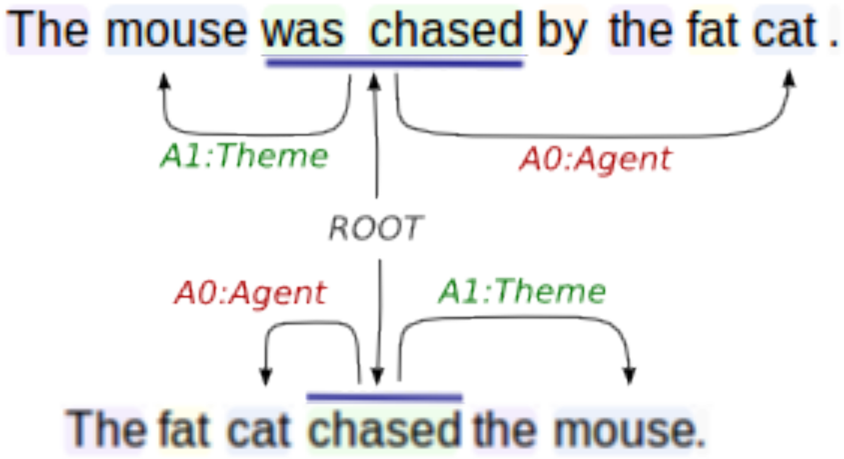
CROSS-LINGUAL SEMANTIC REPRESENTATION

Dependency Parsing vs Semantic Parsing

- Advantage of a semantic over shallow syntactic representation



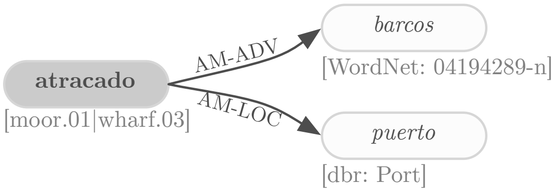



The mouse was chased by the fat cat .
A1:Theme
A0:Agent
ROOT
A0:Agent
A1:Theme
The fat cat chased the mouse.

Multi-lingual SRL Graphs

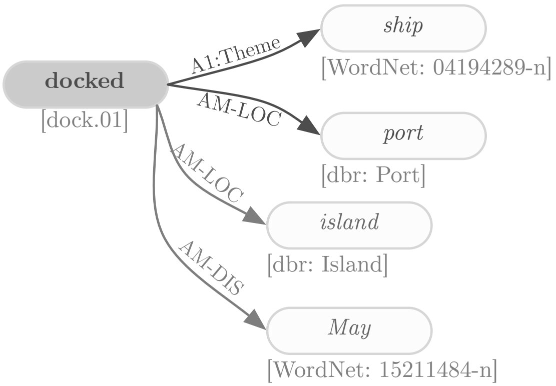
Spanish sentence 1:

En mayo de 1937 el Deutschland estaba **atracado** en el *puerto* de Palma, en Mallorca, junto con otros *barcos* de guerra neutrales de las armadas británica e italiana.



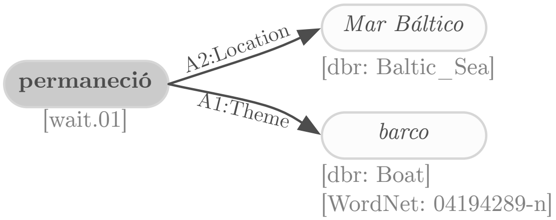
English sentence 2:

In *May* 1937, the *ship* was **docked** in the *port* of Palma on the *island* of Majorca, along with several other neutral warships, including vessels from the British and Italian navies.



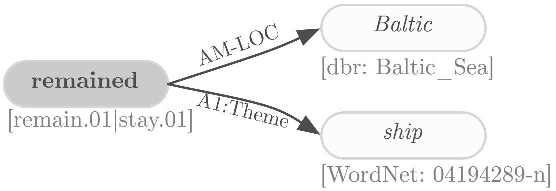
Spanish sentence 3:

Los problemas en sus motores obligaron a una serie de reparaciones que culminaron en una revisión completa a fines de 1943, tras lo que el *barco* **permaneció** en el *Mar Báltico*.



English sentence 4:

Engine problems forced a series of repairs culminating in a complete overhaul at the end of 1943, after which the *ship* **remained** in the *Baltic*.



Cross-lingual Meaning Representation

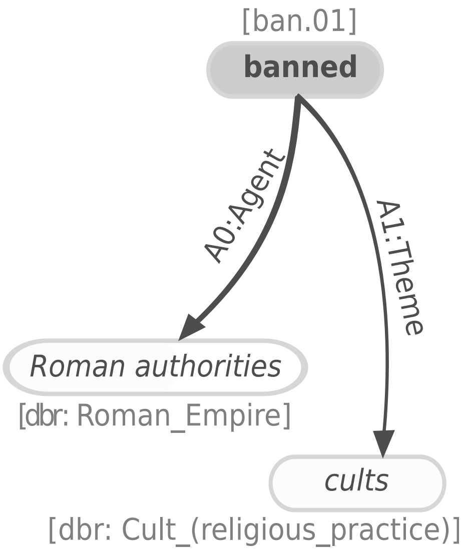
Only a few cults were banned by the Roman authorities...

SRL

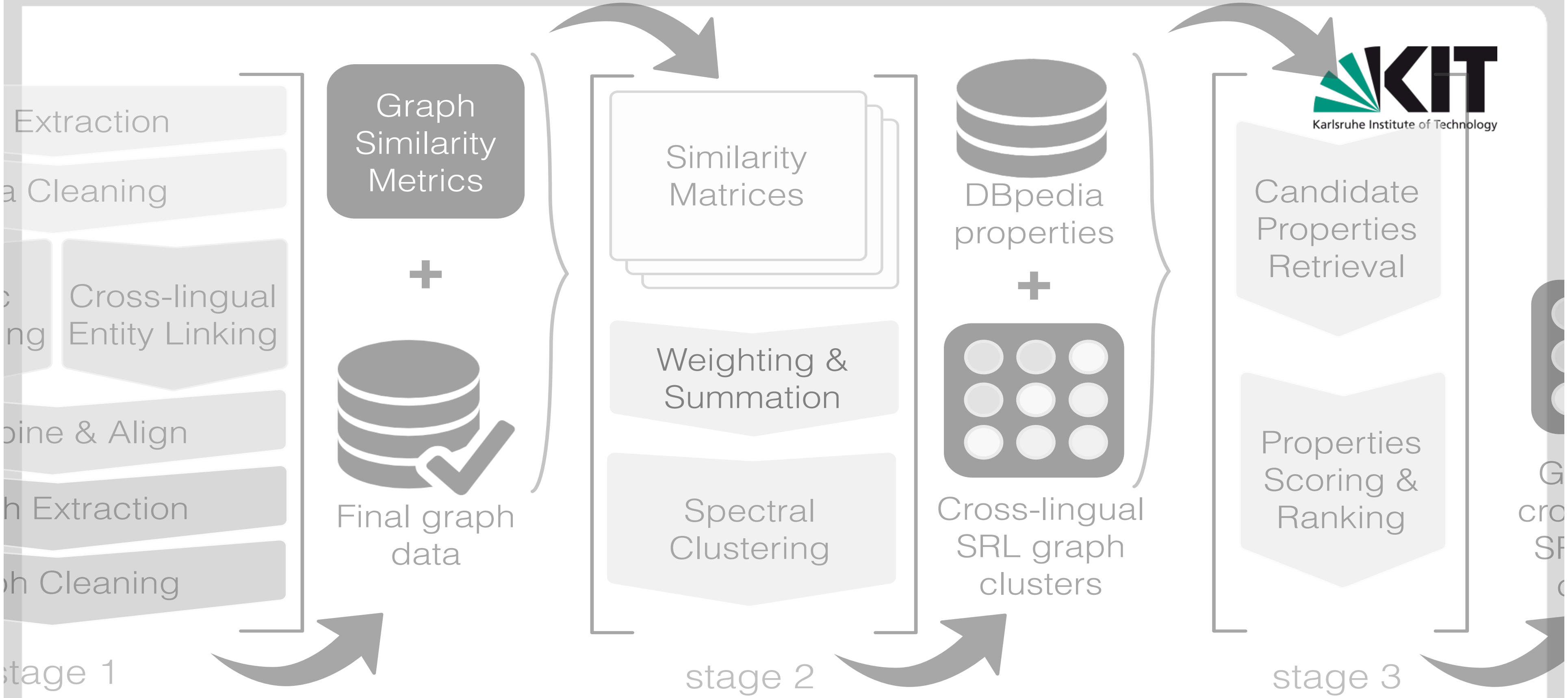
```
<frame displayName="ban.01" ID="F541" sentenceID="57" tokenID="57.6" >
    <argument displayName="cult" role="A1:Theme" id="W544" />
    <argument displayName="imperial_roman" role="A0:Agent" id="E1" />
    ...
    <descriptions>
        <description URI="00796392-v" displayName="ban" knowledgeBase="WordNet-3.0" />
    </descriptions>
</frame>
```

entity linking

```
<DetectedTopic URL="http://dbpedia.org/resource/Cult_(religious_practice)" mention="cults"
    displayName="Cult (religious practice)" from="7064" to="7069" weight="0.01" \>
<DetectedTopic URL="http://dbpedia.org/resource/Roman_Empire" mention="Roman authorities"
    displayName="Roman Empire" from="7089" to="7106" weight="0.393" \>
```



Entity Linking tool:
https://people.aifb.kit.edu/lzh/xlisa/



Stage 2

CLUSTERING CROSS-LINGUAL SRL GRAPHS

Similarity Metrics for SRL Graphs

- Three symmetric measures to quantify similarity of graph pairs
  - Compare root predicates *p(g)*

    $$m_1(g_i, g_j) := \begin{cases} 1 \ , p(g_i) = p(g_j) \\ 0 \ , else \end{cases}$$

  - Jaccard *similarity of the predicates‘* annotated *argument sets* $A(g)$

    $$m_2(g_i, g_j) := \frac{|A(g_i) \cap A(g_j)|}{|A(g_i) \cup A(g_j)|}$$

  - Jaccard similarity of the predicates‘ role label sets $B(g)$

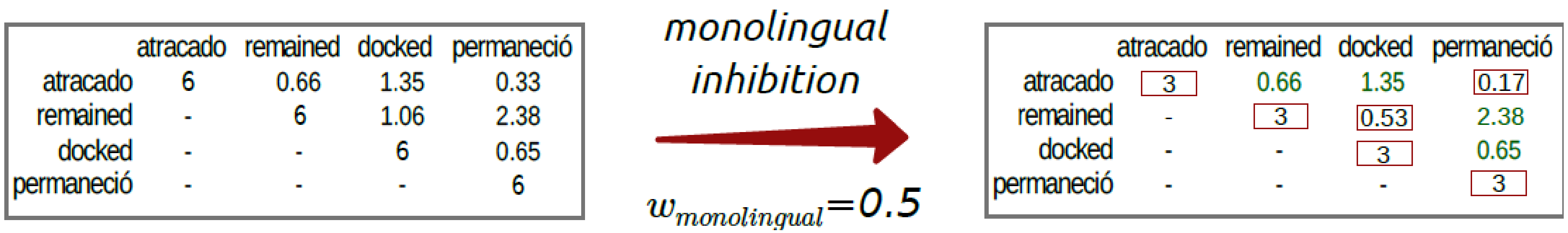
    $$m_3(g_i, g_j) := \frac{|B(g_i) \cap B(g_j)|}{|B(g_i) \cup B(g_j)|}$$

Similarity Matrix

- Extended Similarity Matrix $S↑*$
  - Linear combination of $m1, m2, m3$
  - Apply cross-lingual constraint
    - decrease weight of monolingual predicate graph pairs

$$W_{ij} = \begin{cases} w_{monolingual} & \text{if i and j are monolingual} \\ 1 & \text{if i and j are crosslingual} \end{cases}$$

$$S^*_{ij} = W_{ij} \cdot S_{ij}$$



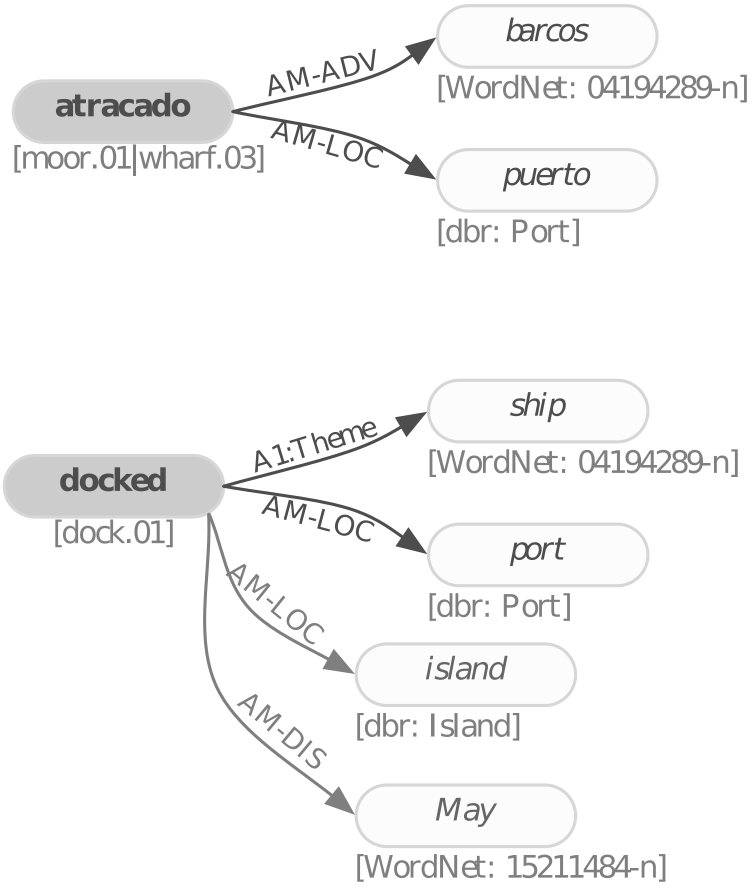
	atracado	remained	docked	permaneció
atracado	6	0.66	1.35	0.33
remained	-	6	1.06	2.38
docked	-	-	6	0.65
permaneció	-	-	-	6

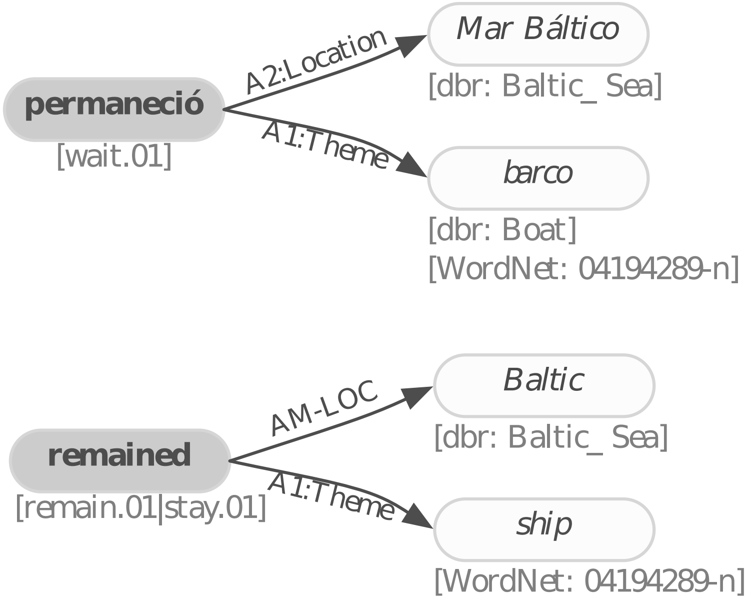
monolingual inhibition

$w_{monolingual}$=0.5

	atracado	remained	docked	permaneció
atracado	3	0.66	1.35	0.17
remained	-	3	0.53	2.38
docked	-	-	3	0.65
permaneció	-	-	-	3

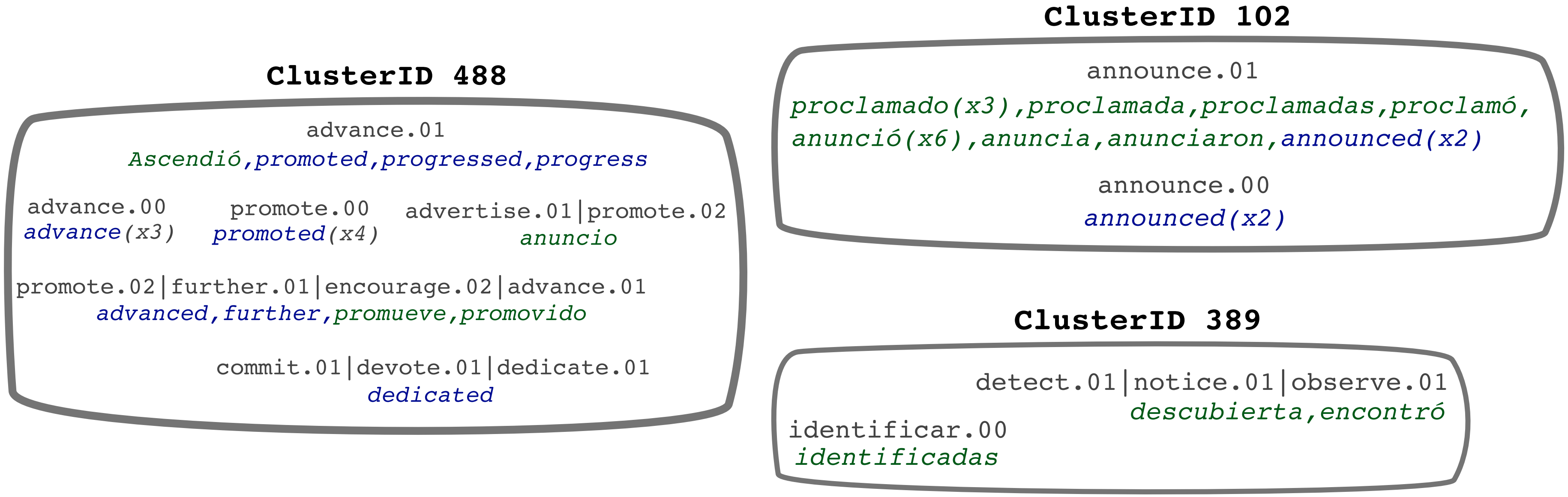
Resulting Similarity Matrix





	atracado	remained	docked	permaneció
atracado	6	0.66	1.35	0.33
remained	-	6	1.06	2.38
docked	-	-	6	0.65
permaneció	-	-	-	6

Result of Spectral Clustering



ClusterID 488

advance.01
Ascendió,promoted,progressed,progress

advance.00
advance(x3)

promote.00
promoted(x4)

advertise.01|promote.02
anuncio

promote.02|further.01|encourage.02|advance.01
advanced,further,promueve,promovido

commit.01|devote.01|dedicate.01
dedicated

ClusterID 102

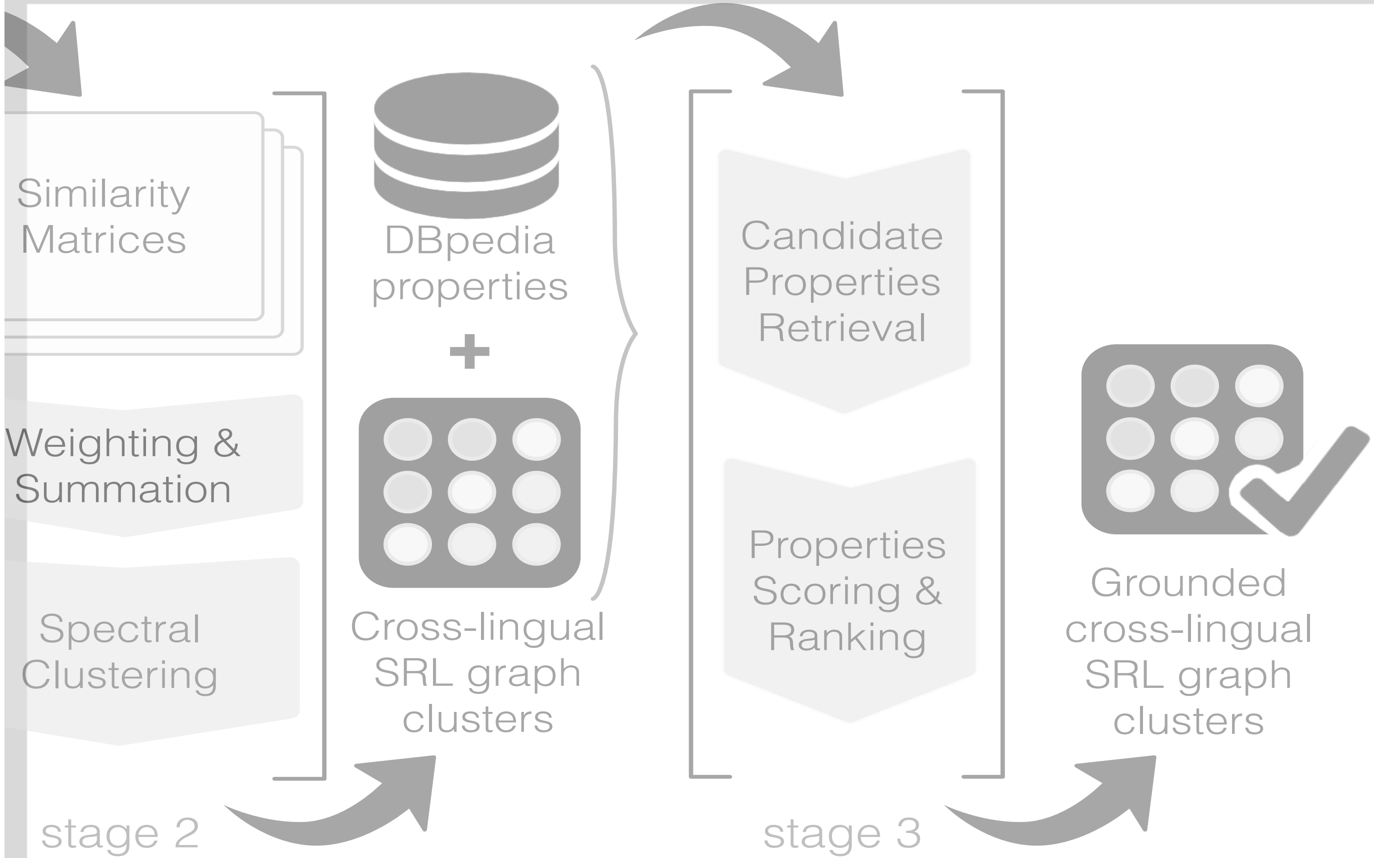
announce.01
proclamado(x3),proclamada,proclamadas,proclamó,
anunció(x6),anuncia,anunciaron,announced(x2)

announce.00
announced(x2)

ClusterID 389

detect.01|notice.01|observe.01
descubierta,encontró

identificar.00
identificadas

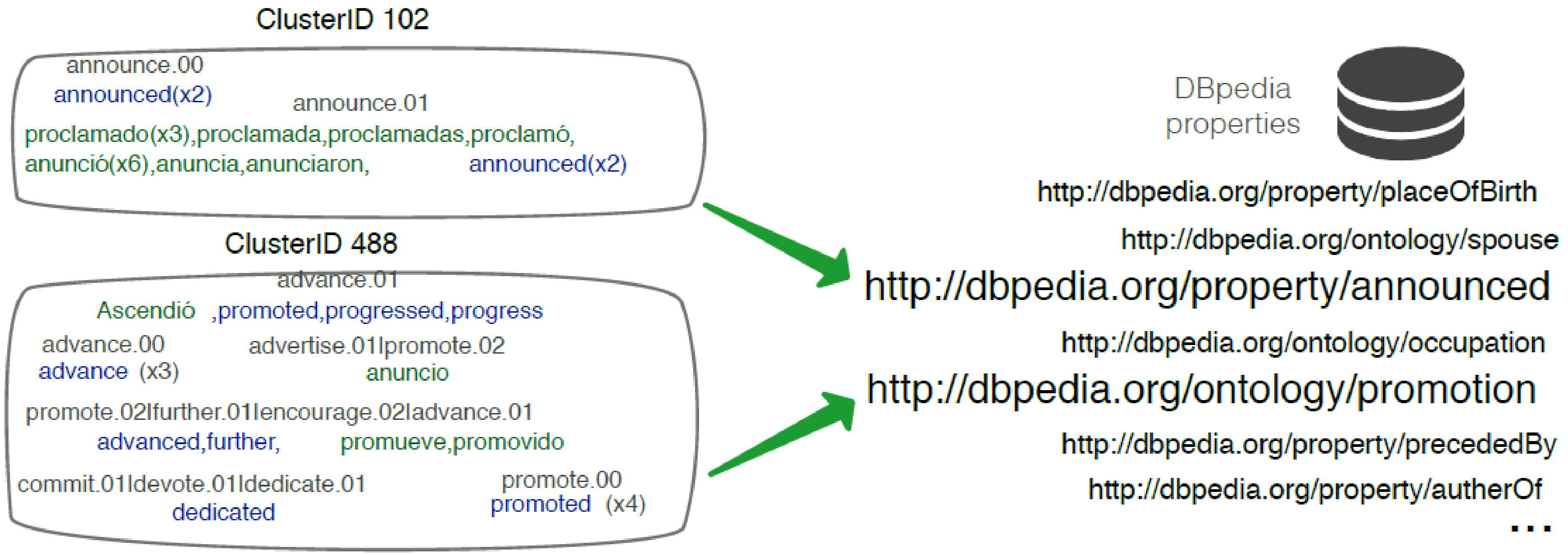


Stage 3

LINKING GRAPH CLUSTERS

Linking Predicate Clusters to DBpedia Properties

- For a given cluster of predicate graphs, generate a list of candidate DBpedia properties
  - collect the contained entities
  - query DBpedia for the set of associated in- and outbound properties
  - Rank the candidates by their absolute frequency



ClusterID 102

announce.00
announced(x2)
announce.01
proclamado(x3),proclamada,proclamadas,proclamó,
anunció(x6),anuncia,anunciaron,
announced(x2)

ClusterID 488

advance.01
Ascendió ,promoted,progressed,progress
advance.00
advance (x3)
advertise.01lpromote.02
anuncio
promote.02lfurther.01lencourage.02ladvance.01
advanced,further,
promueve,promovido
commit.01ldevote.01ldedicate.01
dedicated
promote.00
promoted (x4)

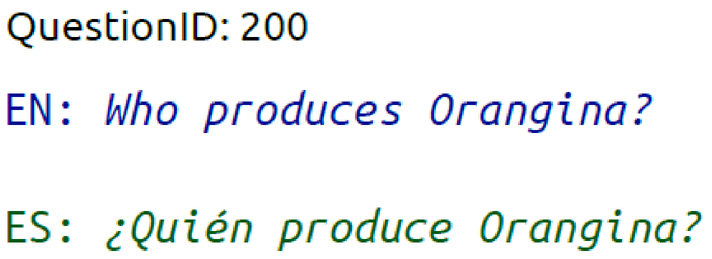
DBpedia properties

http://dbpedia.org/property/placeOfBirth
http://dbpedia.org/ontology/spouse
http://dbpedia.org/property/announced
http://dbpedia.org/ontology/occupation
http://dbpedia.org/ontology/promotion
http://dbpedia.org/property/precededBy
http://dbpedia.org/property/autherOf
...

EVALUATION QALD

QALD Challenge

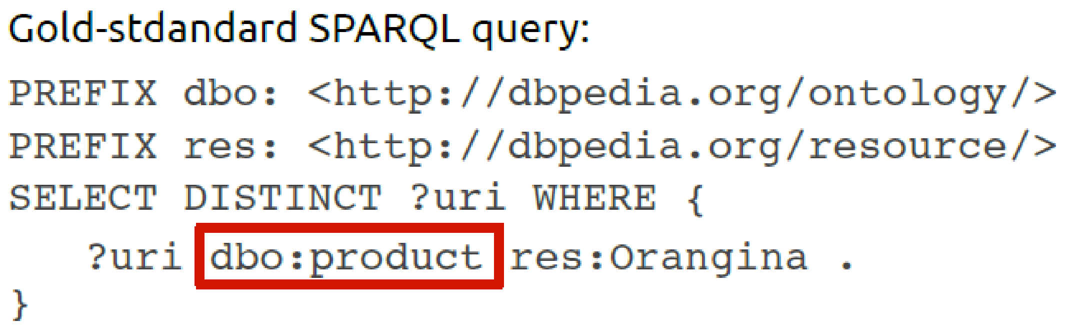
- Identify the formal properties contained in a natural-language question
- Multilingual Question Answering over Linked Data 4 (Task-1 of QALD4)
    - 188 in-scope questions in EN and ES



QuestionID: 200

EN: *Who produces Orangina?*

ES: *¿Quién produce Orangina?*



Gold-stdandard SPARQL query:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
SELECT DISTINCT ?uri WHERE {
   ?uri dbo:product res:Orangina .
}
```

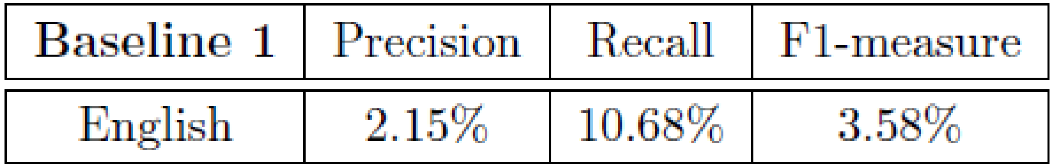
Baselines

- Baseline 1: String Similarity-based Property Linking
  - between the question tokens and all DBpedia property labels
  - Naive, mono-lingual due to labels mostly in EN

- Baseline 2: Entity-based Property Linking
  - Query properties of entities appearing in the given question
  - String similarity between property labels and question tokens
  - Two ways to extract entities from question:
    - WITHOUT SRL: Do plain entity linking
    - WITH SRL: Generate SRL graph and take only the annotated arguments

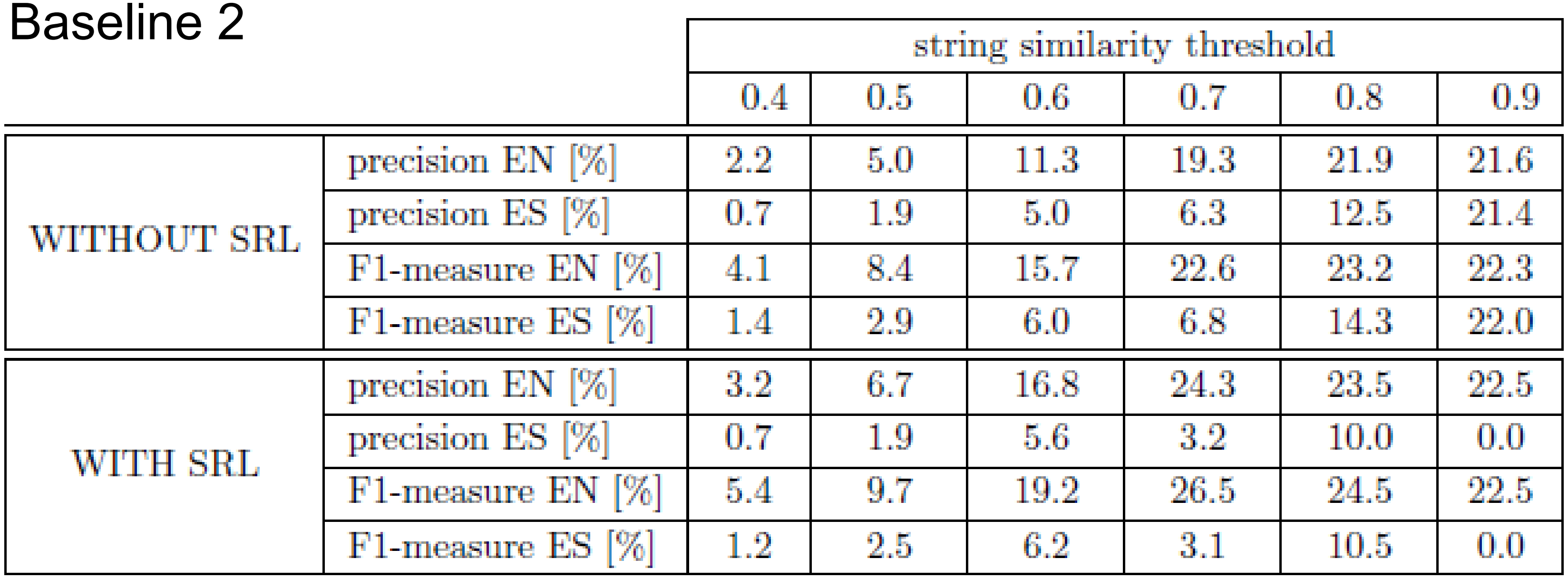
Results: Baseline Performance

- Baseline1



Baseline 1	Precision	Recall	F1-measure
English	2.15%	10.68%	3.58%

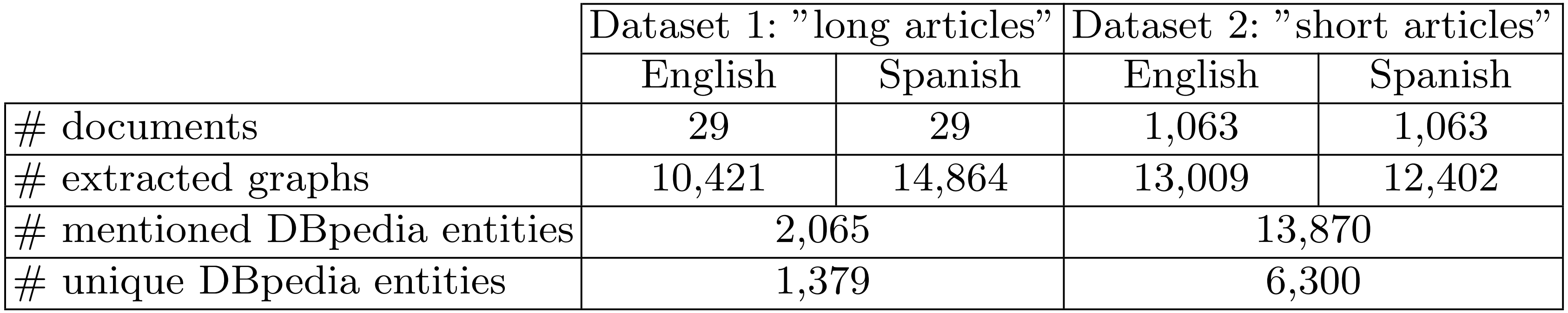
- Baseline 2



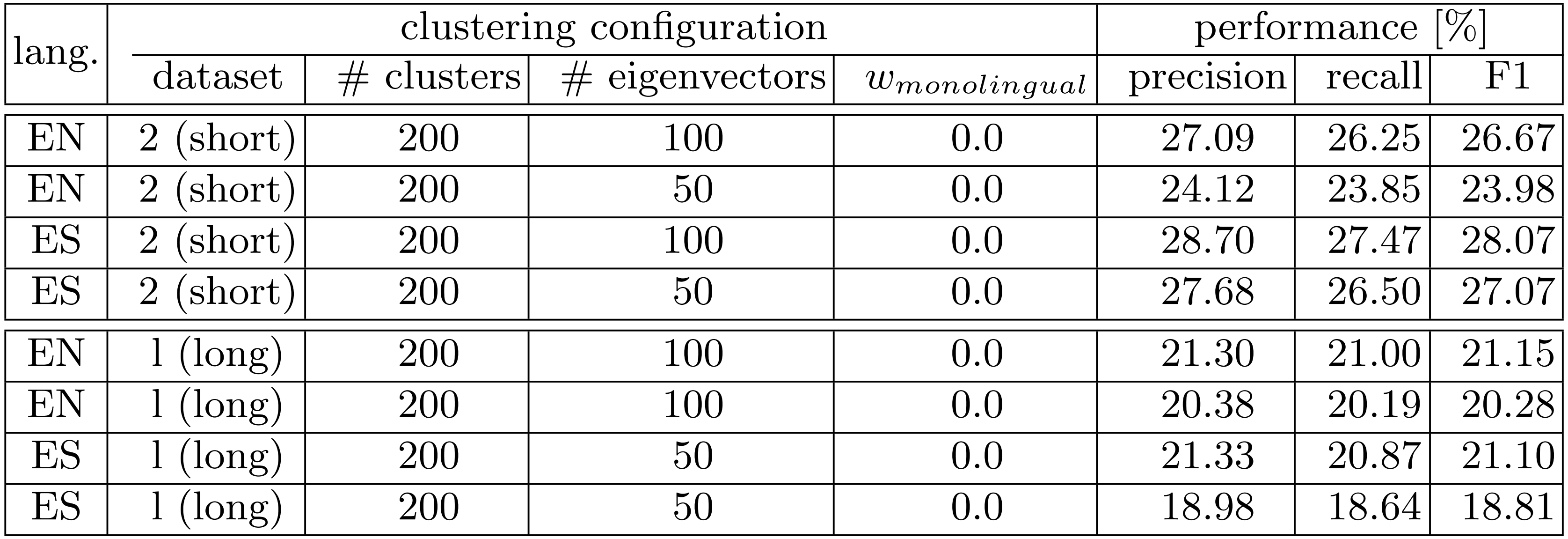
		string similarity threshold					
		0.4	0.5	0.6	0.7	0.8	0.9
WITHOUT SRL	precision EN [%]	2.2	5.0	11.3	19.3	21.9	21.6
	precision ES [%]	0.7	1.9	5.0	6.3	12.5	21.4
	F1-measure EN [%]	4.1	8.4	15.7	22.6	23.2	22.3
	F1-measure ES [%]	1.4	2.9	6.0	6.8	14.3	22.0
WITH SRL	precision EN [%]	3.2	6.7	16.8	24.3	23.5	22.5
	precision ES [%]	0.7	1.9	5.6	3.2	10.0	0.0
	F1-measure EN [%]	5.4	9.7	19.2	26.5	24.5	22.5
	F1-measure ES [%]	1.2	2.5	6.2	3.1	10.5	0.0

Table 5.2: Performance of Baseline 2 without and with SRL graph extraction for different string similarity threshold values.

Results: Effect of diversity of data



	Dataset 1: ”long articles”		Dataset 2: ”short articles”	
	English	Spanish	English	Spanish
# documents	29	29	1,063	1,063
# extracted graphs	10,421	14,864	13,009	12,402
# mentioned DBpedia entities	2,065		13,870	
# unique DBpedia entities	1,379		6,300	

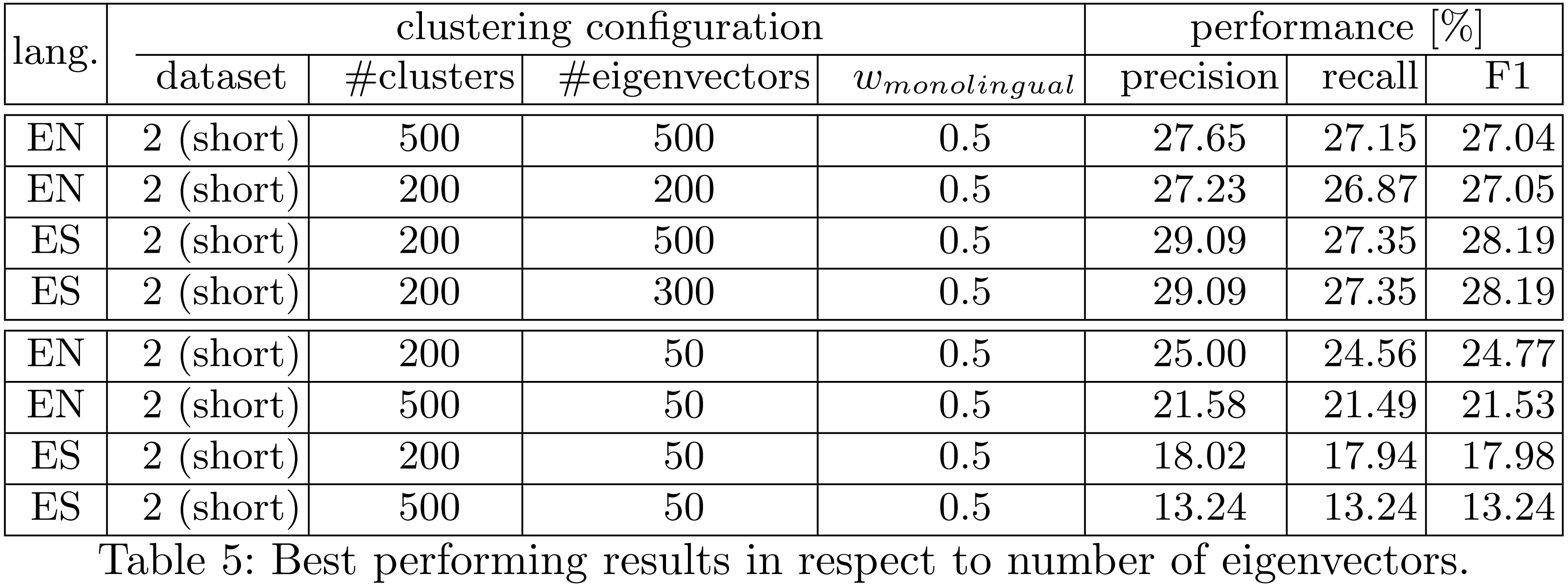


lang.	clustering configuration				performance [%]		
	dataset	# clusters	# eigenvectors	$w_{monolingual}$	precision	recall	F1
EN	2 (short)	200	100	0.0	27.09	26.25	26.67
EN	2 (short)	200	50	0.0	24.12	23.85	23.98
ES	2 (short)	200	100	0.0	28.70	27.47	28.07
ES	2 (short)	200	50	0.0	27.68	26.50	27.07
EN	1 (long)	200	100	0.0	21.30	21.00	21.15
EN	1 (long)	200	100	0.0	20.38	20.19	20.28
ES	1 (long)	200	50	0.0	21.33	20.87	21.10
ES	1 (long)	200	50	0.0	18.98	18.64	18.81

Table 4: Best performing results for ”short articles” vs ”long articles”.

Results: Effect of # of eigenvectors

- More eigenvectors increases performance
- Outperforms Baseline 2 by ca. 6%
- Robust in terms of input language



lang.	clustering configuration				performance [%]		
	dataset	#clusters	#eigenvectors	$w_{monolingual}$	precision	recall	F1
EN	2 (short)	500	500	0.5	27.65	27.15	27.04
EN	2 (short)	200	200	0.5	27.23	26.87	27.05
ES	2 (short)	200	500	0.5	29.09	27.35	28.19
ES	2 (short)	200	300	0.5	29.09	27.35	28.19
EN	2 (short)	200	50	0.5	25.00	24.56	24.77
EN	2 (short)	500	50	0.5	21.58	21.49	21.53
ES	2 (short)	200	50	0.5	18.02	17.94	17.98
ES	2 (short)	500	50	0.5	13.24	13.24	13.24

Table 5: Best performing results in respect to number of eigenvectors.

SUMMARY

Cross-lingual Semantic Clusters of Relations

- Extract **relation expressions** from different languages
- Extracted expressions are **embedded in a semantic graph**, describing the context this expression appears in.
- Semantically-related relation expressions and their associated context **are disambiguated and clustered across languages**.
- If existing, relation clusters are linked **to their corresponding property** in the English DBpedia

Thank you!

QUESTIONS?