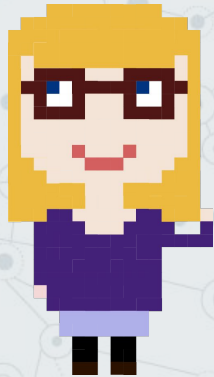




Multilinguality in Knowledge Graphs

Lucie-Aimée Kaffee

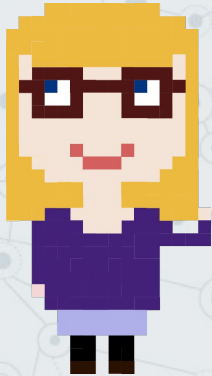
About Lucie



- Postdoc at University of Copenhagen (CopeNLU)
- PhD from University of Southampton (2020) as Marie-Curie ITN research fellow
- Research fellow at TIB Hannover (2019)
- Research internship at Bloomberg L.P. (2019/20)
- Software Developer at Wikimedia Deutschland in the Wikidata team (2014-2016)

About Lucie

Research Interests



- Multilingual Data
- Wikidata & Wikipedia
- Languages and Labels in Knowledge Graphs (this presentation)
- Natural Language Processing (NLP), with focus on application for KGs and integrating social science knowledge

Structure of the presentation

- Motivation
- Research Questions and respective studies/results
- Outlook on current work



Knowledge Graphs and Multilingual Data

Motivation

- A large number of applications serve multilingual users, e.g. question answering systems



Amazon Alexa, Image: CC-BY Pixabay



Google Home, Image: CC-BY-SA
[Wikimedia Commons](https://commons.wikimedia.org/wiki/File:Google_Home.jpg)

The screenshot displays a Bloomberg Terminal interface with multiple data panels. The top panel shows 'China' market data, including 'Onshore FI Overview' with columns for Yield, Last Yield, and Last. Below this is a 'CFETS Key Rates' table with columns for Current and Chg. The right side of the terminal shows 'Bond Futures' and 'CFETS Key Rates' with columns for Last and Chg. The bottom panel shows a 'Financial Analysis' table for 'GRAB' with columns for FY 2013, FY 2014, FY 2015, FY 2016, Current/LTM, FY 2017 Est., and FY 2018 Est. The table includes rows for Revenue, Earnings, and Cash Flow.

	FY 2013	FY 2014	FY 2015	FY 2016	Current/LTM	FY 2017 Est.	FY 2018 Est.
Revenue, Adj	7,393.0	7,736.5	7,715.0	7,553.7	7,232.4	7,355.5	7,626.8
Gross Profit, Adj	49.2	29.3	189.3	149.3	146.6		
Operating Profit, Adj	1,883.2	2,097.7	2,187.8	2,213.7	2,114.6	2,009.9	2,136.9
Net Income, Adj	590.3	636.2	688.4	695.2	600.0	565.9	633.2
EPS, Adj	-5.3	1.2	0.7	0.2	0.9	0.7	0.8
Free Cash Flow	18.6	386.2	267.3	156.8	18.3	68.8	105.0

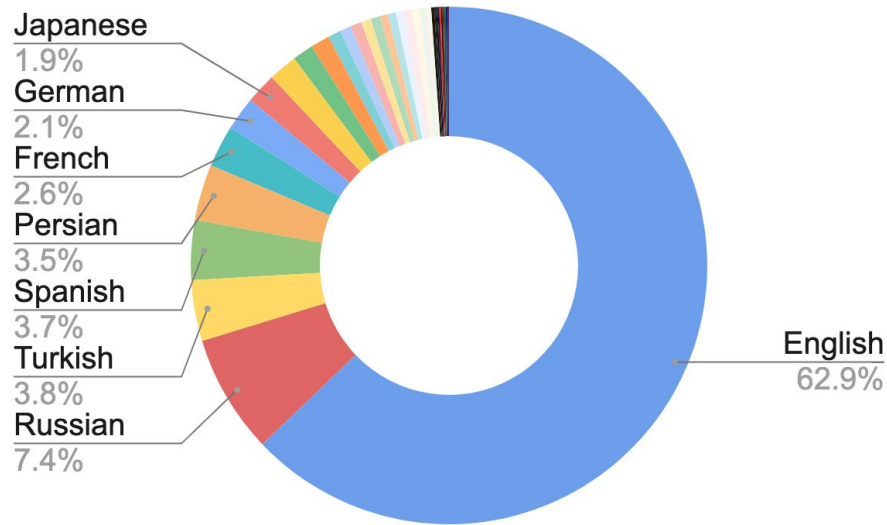
Bloomberg Terminal, Image Credits: [Bloomberg L.P.](https://www.bloomberg.com)

Motivation

- A large number of applications serve multilingual users, e.g. question answering systems
- However, a large number of users is not supported in their (native) language due to a lack of multilingual data

Motivation

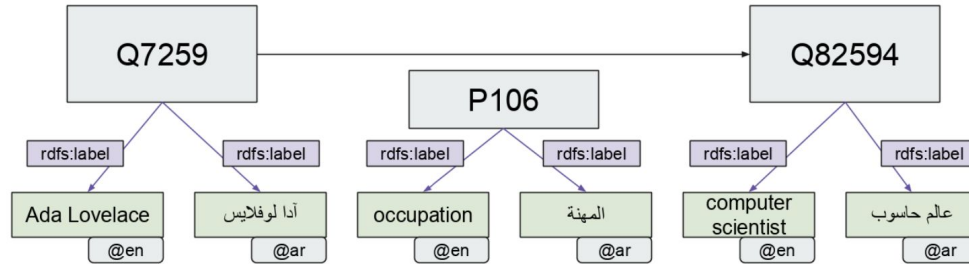
- Lack of language coverage on the web overall



Source: https://w3techs.com/technologies/overview/content_language

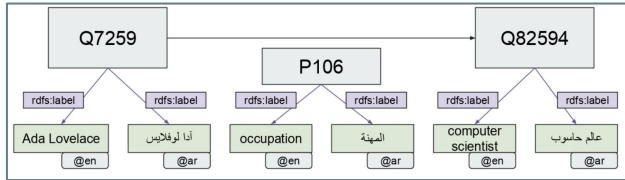
Motivation

- Knowledge graphs support a multilingual representation of concepts



Motivation

- Knowledge graphs could support a larger coverage of languages through, e.g., text generation



WIKIPEDIA The Free Encyclopedia

Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (*née* **Byron**; 10 December 1815 – 27 November 1852) was an English mathematician and writer, chiefly known for her work on Charles Babbage's proposed mechanical general-purpose computer, the *Analytical Engine*. She was the first to recognise that the machine had applications beyond pure calculation, and to have published the first *algorithm* intended to be carried out by such a machine. As a result, she is often regarded as the first computer programmer.^{[a][b]}

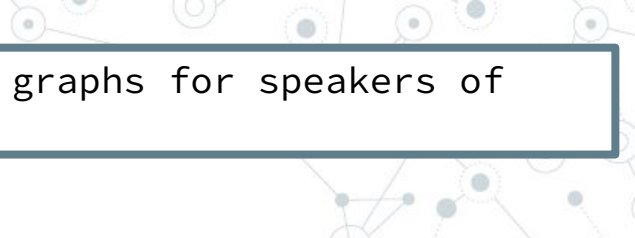
Ada Byron was the only child of poet Lord Byron and mathematician Lady Byron.^[a] All of Byron's other children were born out of wedlock to other women.^[a] Byron separated from his wife a month after Ada was born and left England forever four months later. He commemorated the parting in a poem that begins, "Is thy face like thy mother's my fair child! ADA! sole daughter of my house and heart?".^[7] He died in Greece when Ada was eight years old. Her mother remained bitter and promoted Ada's interest in mathematics and logic in an effort to prevent her from developing her father's perceived insanity. Despite this, Ada remained interested in him, naming her two sons Byron and Gordon. Upon her death, she was buried next to him at her request. Although often ill in her childhood, Ada pursued her studies assiduously. She married William King in 1835. King was made *Earl of Lovelace* in 1838, Ada thereby becoming *Countess of Lovelace*.

Her educational and social exploits brought her into contact with scientists such as Andrew Crosse, Charles Babbage, Sir David Brewster, Charles Wheatstone, Michael Faraday and the author Charles Dickens, contacts which she used to further her education. Ada described her approach as "poetical science"^[a] and herself as an "Analyst (& Metaphysician)".^[a]

When she was a teenager (18), her mathematical talents led her to a long working relationship and friendship with fellow British mathematician Charles Babbage, who is known as "the father of computers". She was in particular interested in Babbage's work on the Analytical Engine. Lovelace first met him in June 1833, through their mutual friend, and her private tutor, Mary Somerville.

Between 1842 and 1843, Ada translated an article by Italian military engineer Luigi Menabrea about the calculating engine, supplementing it with an elaborate set of notes, simply called "Notes". Lovelace's notes are important in the early history of computers, containing what many consider to be

The Right Honourable
The Countess of Lovelace
Daguerrotype by Antoine Claudet (c. 1843)^[1]
Born The Hon. Augusta Ada Byron
10 December 1815
London, England



How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

RQ2 Does knowledge about languages in a knowledge graph support ranking them for question answering?

RQ3 Can we automatically translate knowledge graph labels and aliases?

RQ4 Can we reuse Wikidata's multilingual data to generate Wikipedia summaries?



RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?



Pavlos Vougiouklis, Huawei
Alessandro Piscopo, BBC

RQ2 Does knowledge about languages in a knowledge graph support ranking them for question answering?

Kemele M. Endris, Amazon

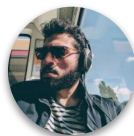


TIB
LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

RQ3 Can we automatically translate knowledge graph labels and aliases?

Bloomberg

RQ4 Can we reuse Wikidata's multilingual data to generate Wikipedia summaries?



Hady Elsahar, Meta
Pavlos Vougiouklis, Huawei

How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

Kaffee and Simperl: The Human Face of the Web of Data: A Cross-sectional Study of Labels [SEMANTiCS 2018]; Kaffee et al.: When Humans and Machines Collaborate: Cross-lingual Label Editing in Wikidata [OpenSym 2019]; Kaffee et al.: A glimpse into Babel: an analysis of multilinguality in Wikidata [OpenSym 2017]; Kaffee and Simperl: Analysis of Editors' Languages in Wikidata [OpenSym 2018]

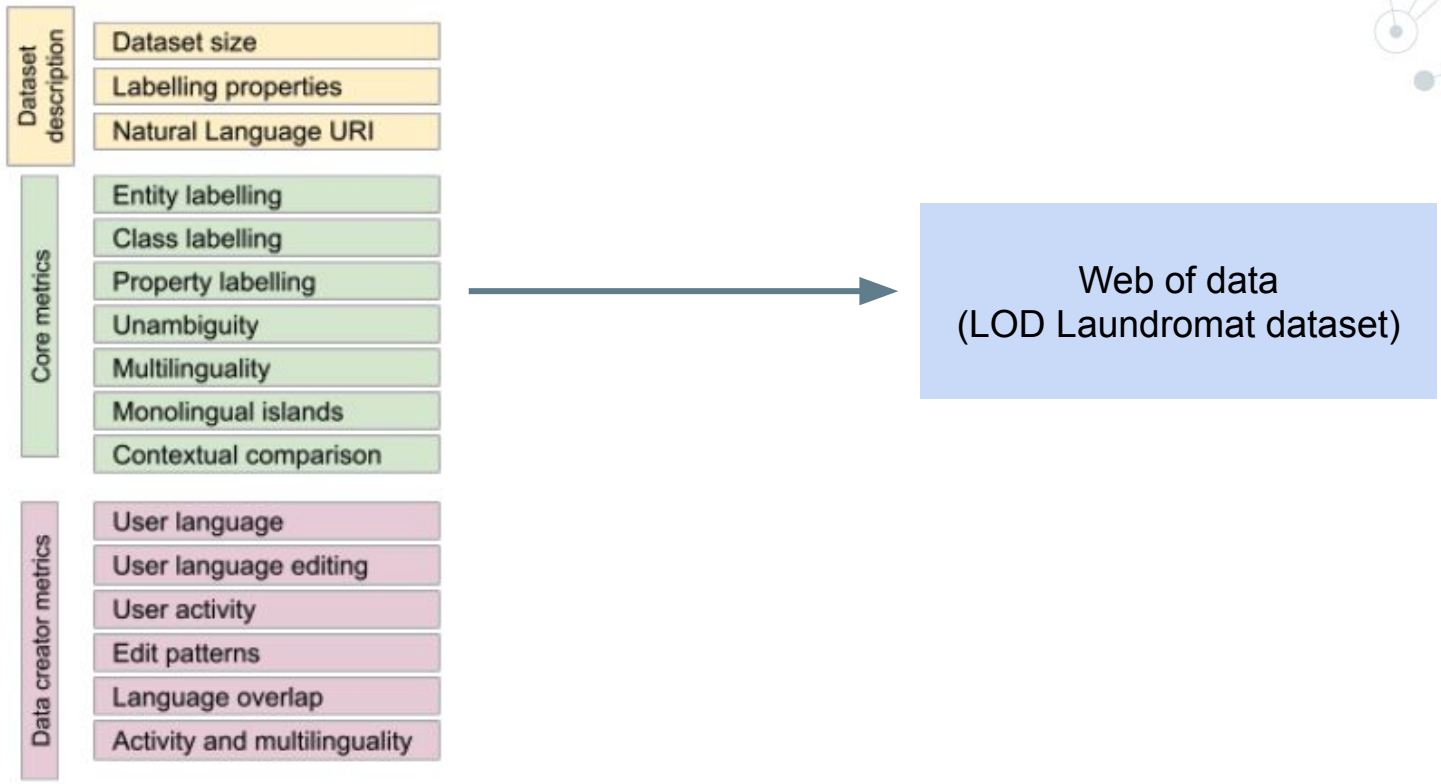
Framework to measure multilinguality in KGs

Dataset description
Dataset size
Natural language URI
Labelling properties

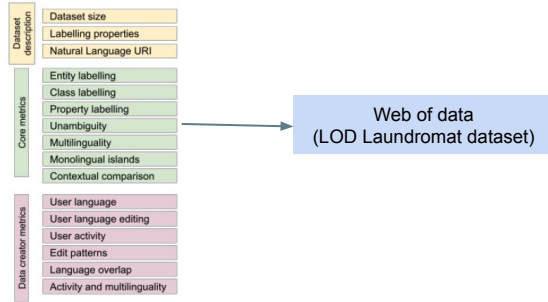
Core metrics
Entity label completeness
Class label completeness
Property label completeness
Unambiguity
Multilinguality
Monolingual Islands
Contextual comparison

Data creator metrics
User language
User language editing
User activity
Edit patterns
Language overlap
Activity and Multilinguality

Framework to measure multilinguality in KGs



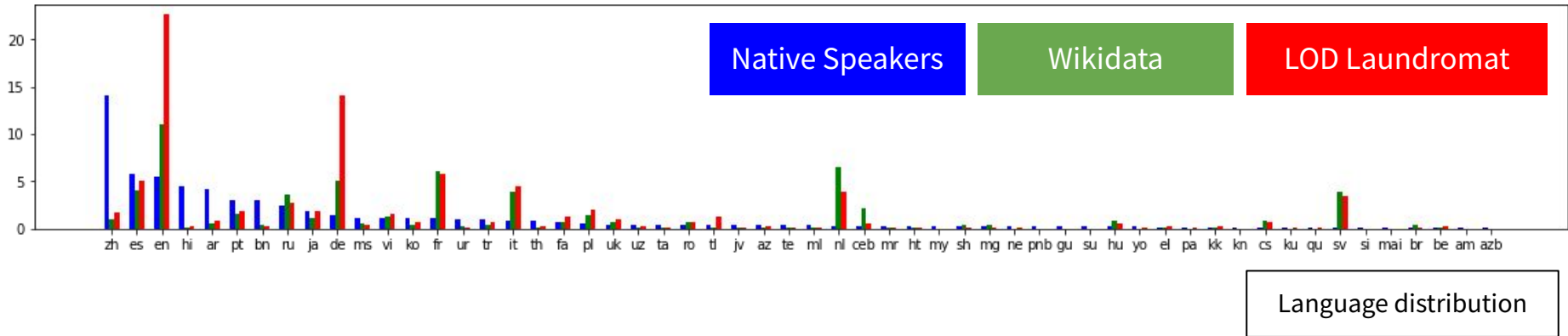
Framework to measure multilinguality in KGs

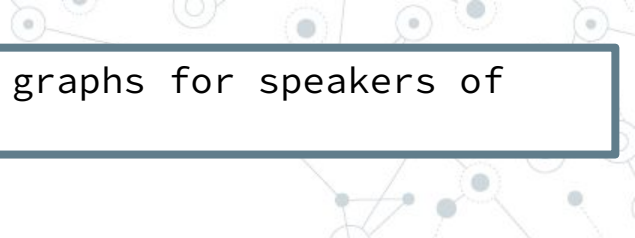


- **rdfs:label** is by far the most used labelling property
- Only **5.42%** of subjects are **labelled**, **80.9%** of properties are labelled
- The **five** most used **languages** cover **over 50%** of labels
- **83.2%** of entities are labelled in only **one language**

Multilinguality in Wikidata

- Compared to the Web of data at large, Wikidata's editors create a more diverse knowledge graph



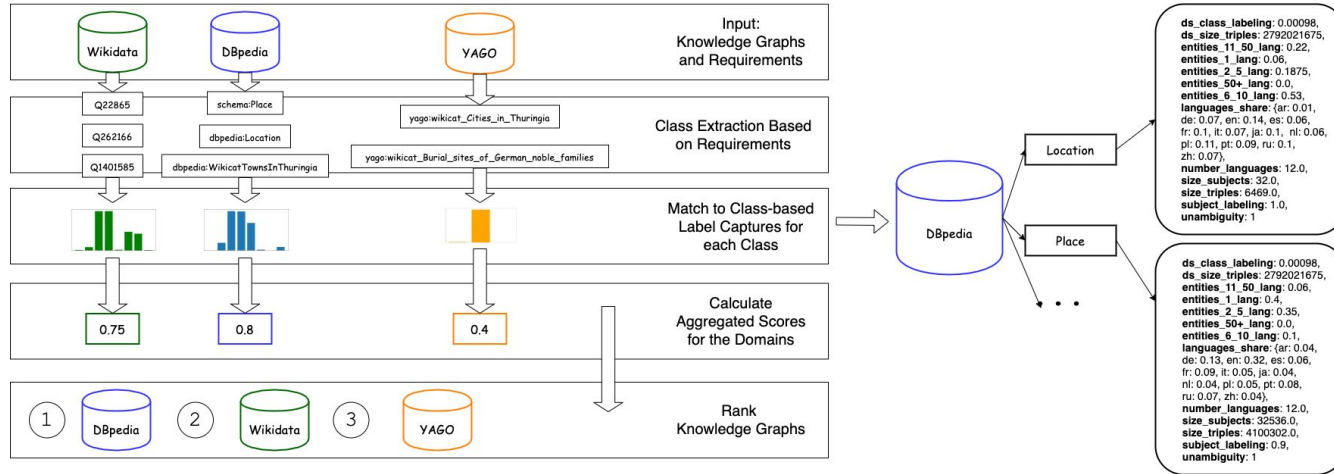


How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

RQ2 Does knowledge about languages in a knowledge graph support ranking them for question answering?

LINGVO, framework to rank KGs based on labels and languages



1

Where was Bach born?

English

2

db:Bach dbo:birthPlace db:Germany
db:Bach rdfs:label "Johann Sebastian Bach"@en
db:Germany rdfs:label "Germany"@en
db:Germany rdfs:label "BRD"@en
db:Germany rdfs:label "Deutschland"@de
db:birthPlace rdfs:label "birth place"@en

Knowledge Graph 1 (KG1)

Wo wurde Bach geboren?

German

wd:Q255 wd:P19 wd:Q183
wd:Q183 rdfs:label "Germany"@en
wd:Q183 rdfs:label "Deutschland"@de
wd:Q183 rdfs:label "Almanya"@tr
wd:Q255 rdfs:label "Johann Sebastian Bach"@en
wd:Q255 rdfs:label "Johann Sebastian Bach"@de
wd:P19 rdfs:label "Geburtsort"@de
wd:P19 rdfs:label "birth place"@en

Knowledge Graph 2 (KG2)

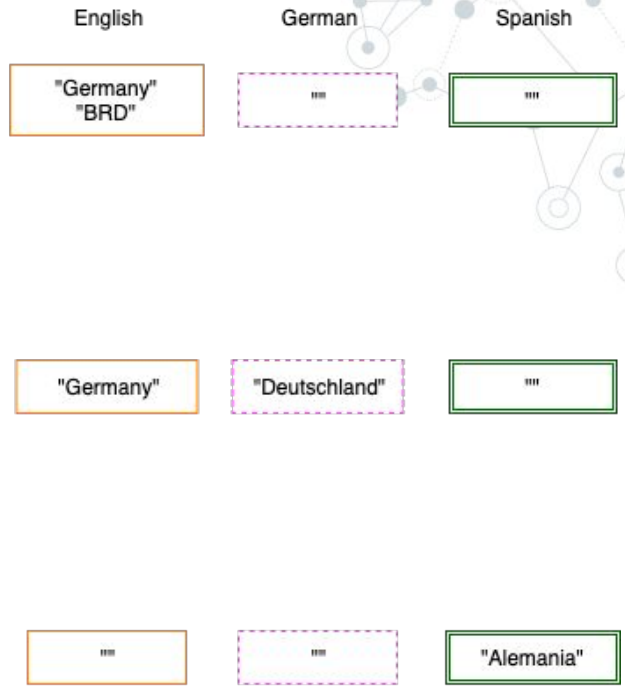
¿Dónde nació Bach?

Spanish

ex:12 ex:31 ex:13
ex:12 rdfs:label "Johann Sebastian Bach"@es
ex:31 rdfs:label "lugar de nacimiento"@es
ex:13 rdfs:label "Alemania"@es
ex:12 rdfs:label "Allemagne"

Knowledge Graph 3 (KG3)

3



1

Select a Knowledge Graph

3

Where was Bach born?

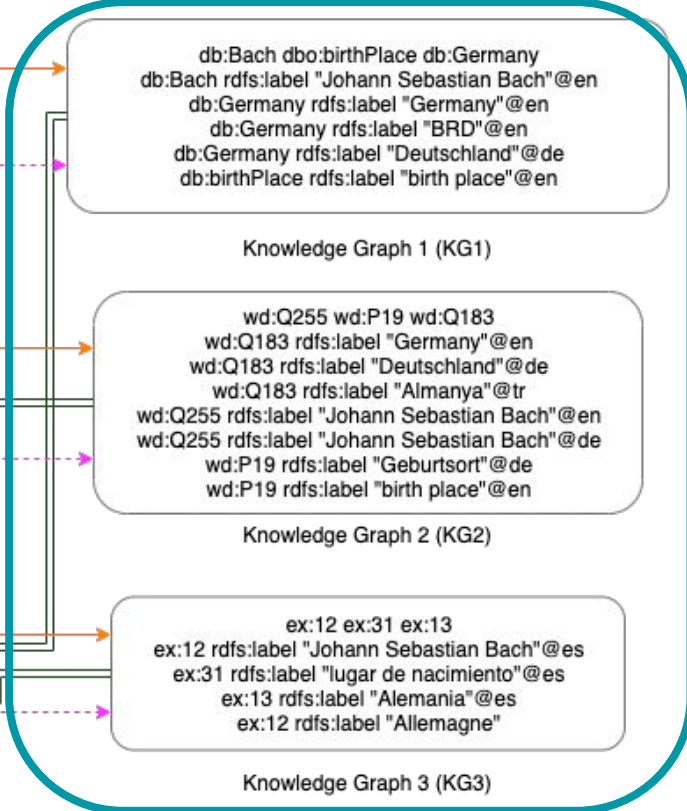
English

Wo wurde Bach geboren?

German

¿Dónde nació Bach?

Spanish



English
"Germany"
"BRD"

German
""

Spanish
""

"Germany"

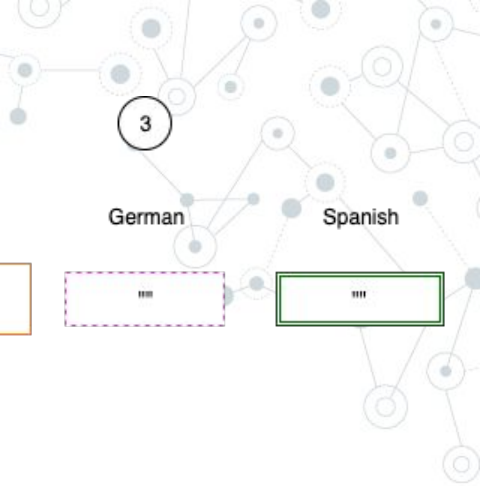
"Deutschland"

""

""

""

"Alemania"



1

That answers the question best

Where was Bach born?

English

Wo wurde Bach geboren?

German

¿Dónde nació Bach?

Spanish

```

db:Bach dbo:birthPlace db:Germany
db:Bach rdfs:label "Johann Sebastian Bach"@en
db:Germany rdfs:label "Germany"@en
db:Germany rdfs:label "BRD"@en
db:Germany rdfs:label "Deutschland"@de
db:birthPlace rdfs:label "birth place"@en

```

Knowledge Graph 1 (KG1)

```

wd:Q255 wd:P19 wd:Q183
wd:Q183 rdfs:label "Germany"@en
wd:Q183 rdfs:label "Deutschland"@de
wd:Q183 rdfs:label "Almanya"@tr
wd:Q255 rdfs:label "Johann Sebastian Bach"@en
wd:Q255 rdfs:label "Johann Sebastian Bach"@de
wd:P19 rdfs:label "Geburtsort"@de
wd:P19 rdfs:label "birth place"@en

```

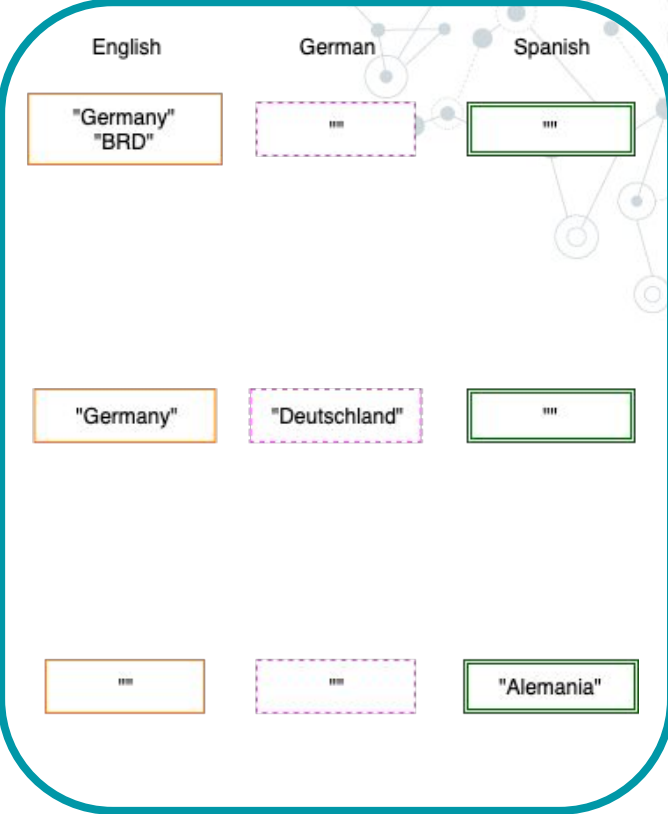
Knowledge Graph 2 (KG2)

```

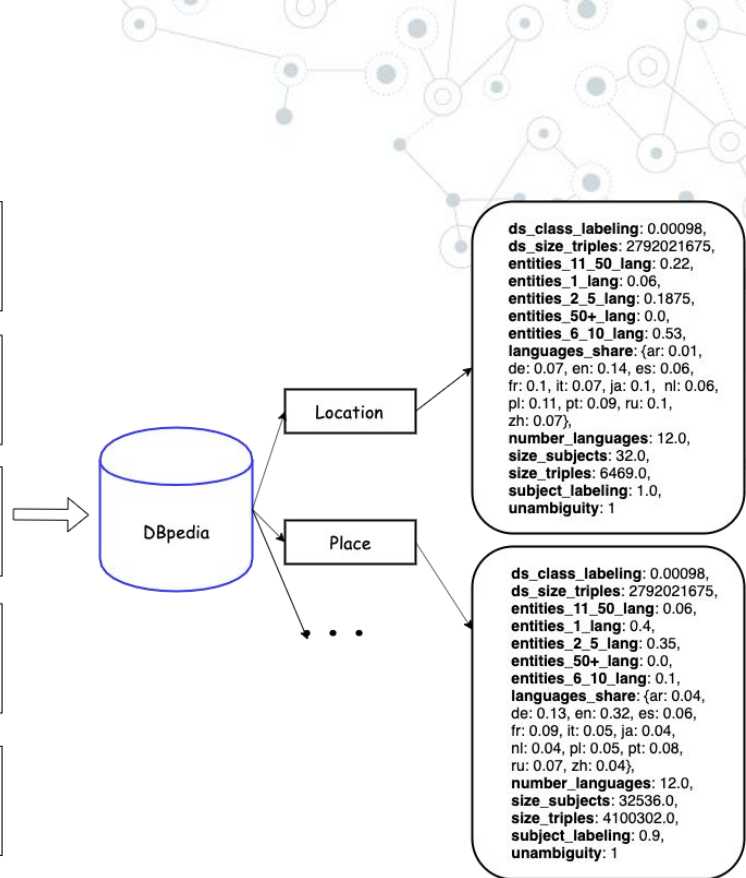
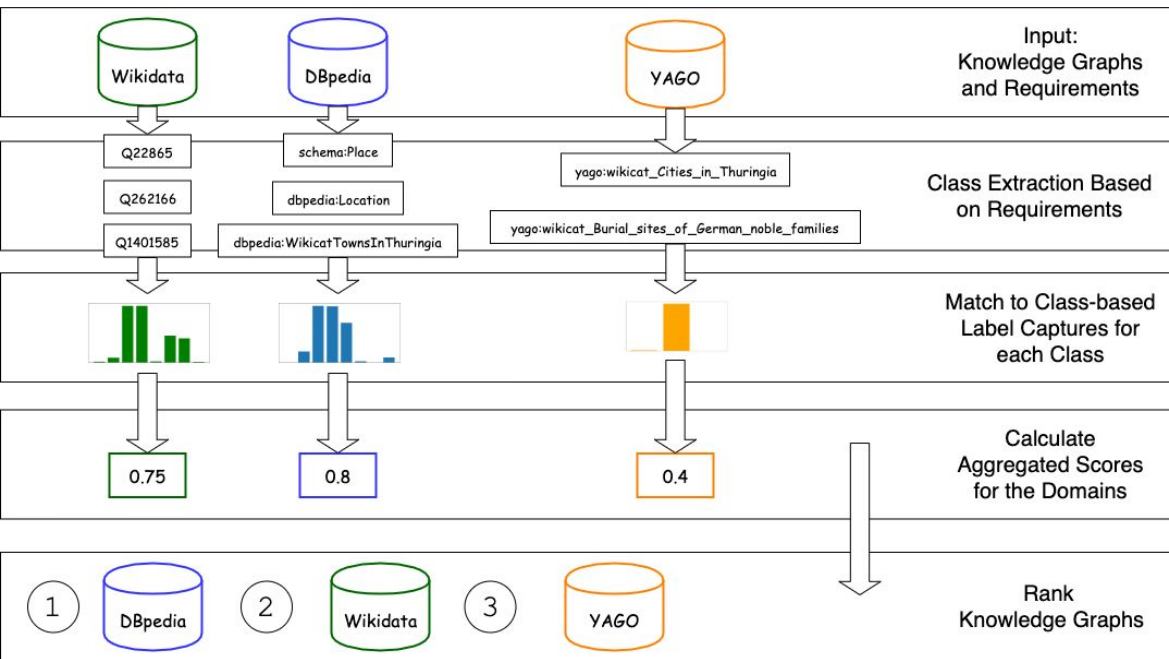
ex:12 ex:31 ex:13
ex:12 rdfs:label "Johann Sebastian Bach"@es
ex:31 rdfs:label "lugar de nacimiento"@es
ex:13 rdfs:label "Alemania"@es
ex:12 rdfs:label "Allemagne"

```

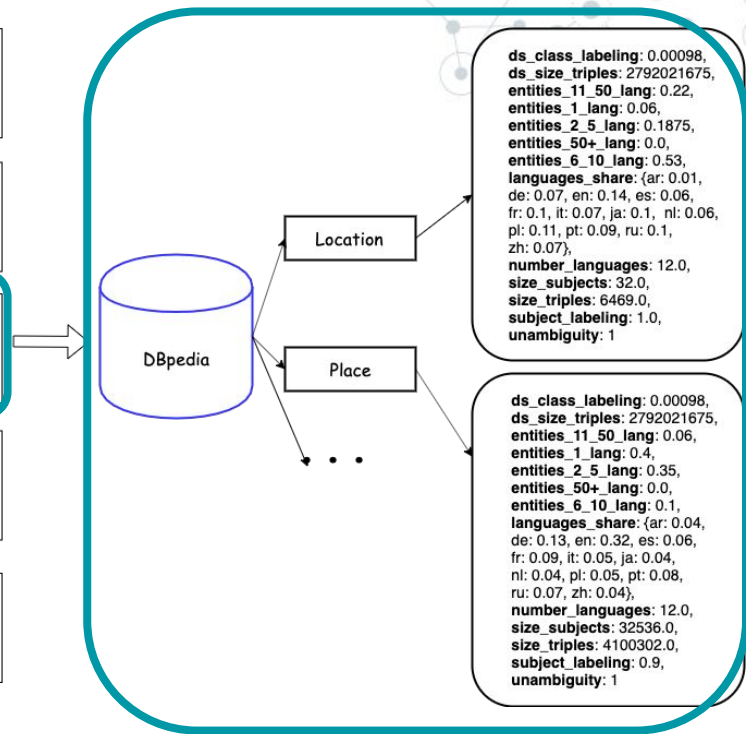
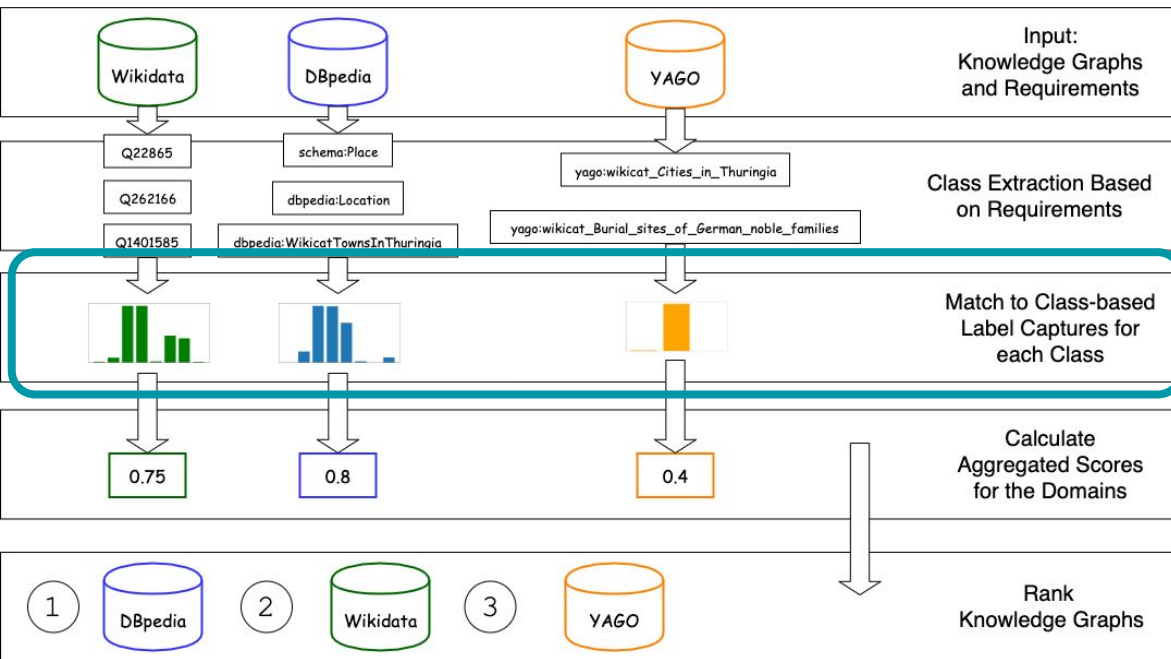
Knowledge Graph 3 (KG3)



Rank Knowledge Graphs

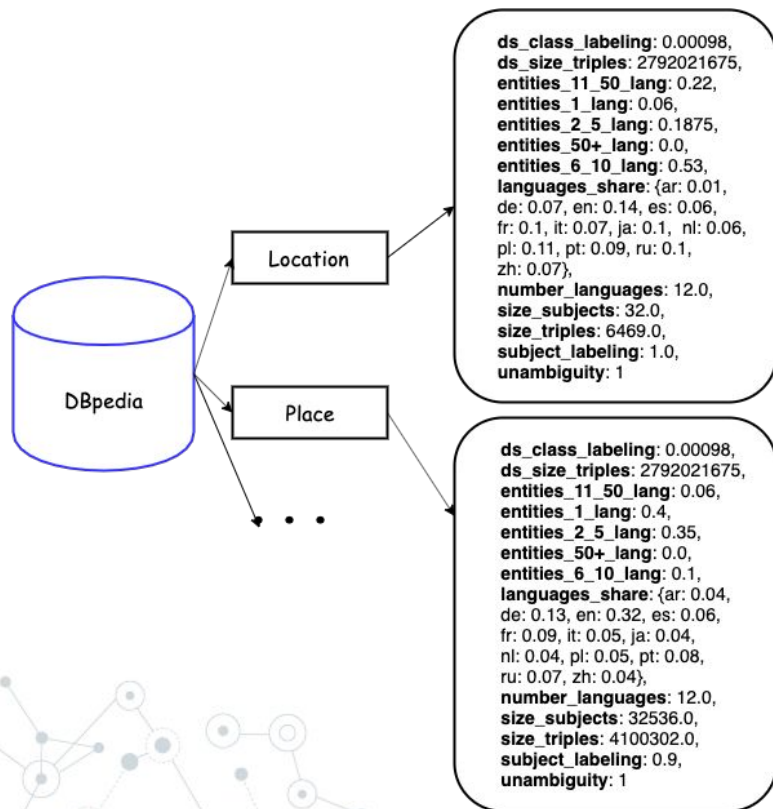


Rank Knowledge Graphs



Based on the framework

Class-based Label Capture (CLC)

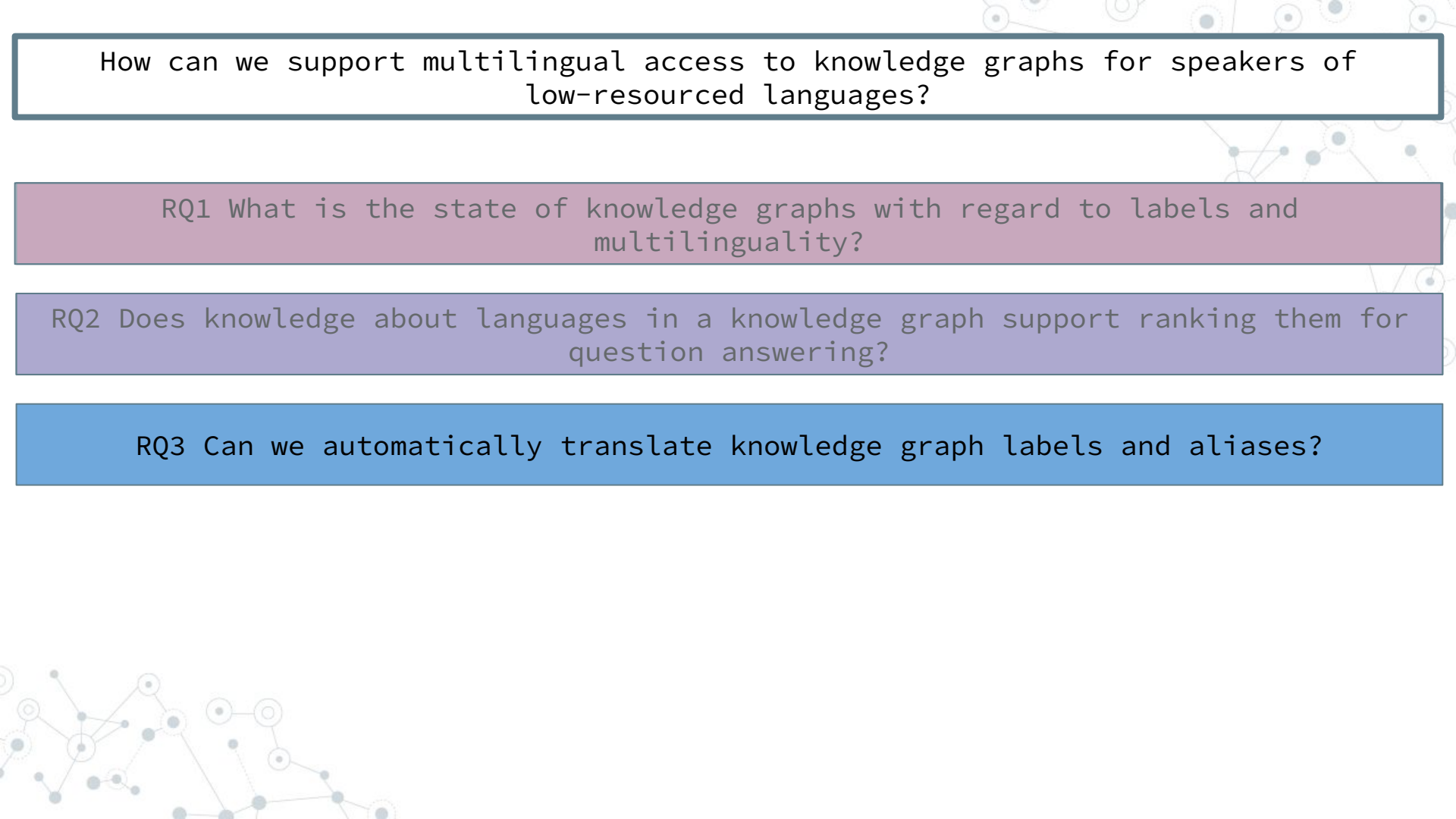


Capture knowledge about language and label coverage for each class in the knowledge graph

E.g. number of languages, label coverage

QALM dataset

- ◎ Based on QALD-9 (benchmark for multilingual QA over DBpedia, answers as entity URIs)
- ◎ Manual translation of SPARQL queries for five KGs
 - DBpedia, Wikidata, YAGO, MusicBrainz, LinkedMDB
- ◎ Answers in English, Spanish, Hindi
- ◎ Crowdsourced best answers
- ◎ Separated in domains, such as company, politics
- ◎ <https://github.com/luciekaffee/QALM>
(still has to be published somewhere properly)



How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

RQ2 Does knowledge about languages in a knowledge graph support ranking them for question answering?

RQ3 Can we automatically translate knowledge graph labels and aliases?

Translation and transliteration of KG labels



- Entities in the *company* domain
- Can be translated or transliterated

Translation and transliteration of KG labels

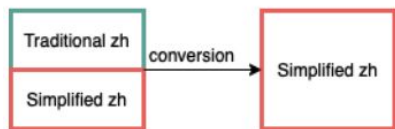


- Entities in the *company* domain
- Can be translated or transliterated

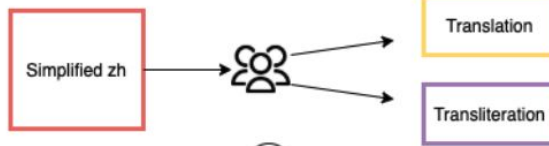
transfers the meaning of a word from one to another language

transfers a word from one script to another

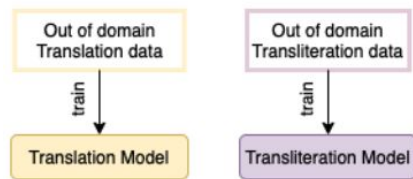
Translation and transliteration of KG labels



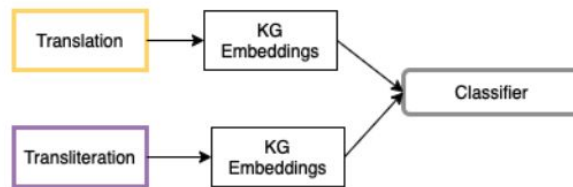
1



2

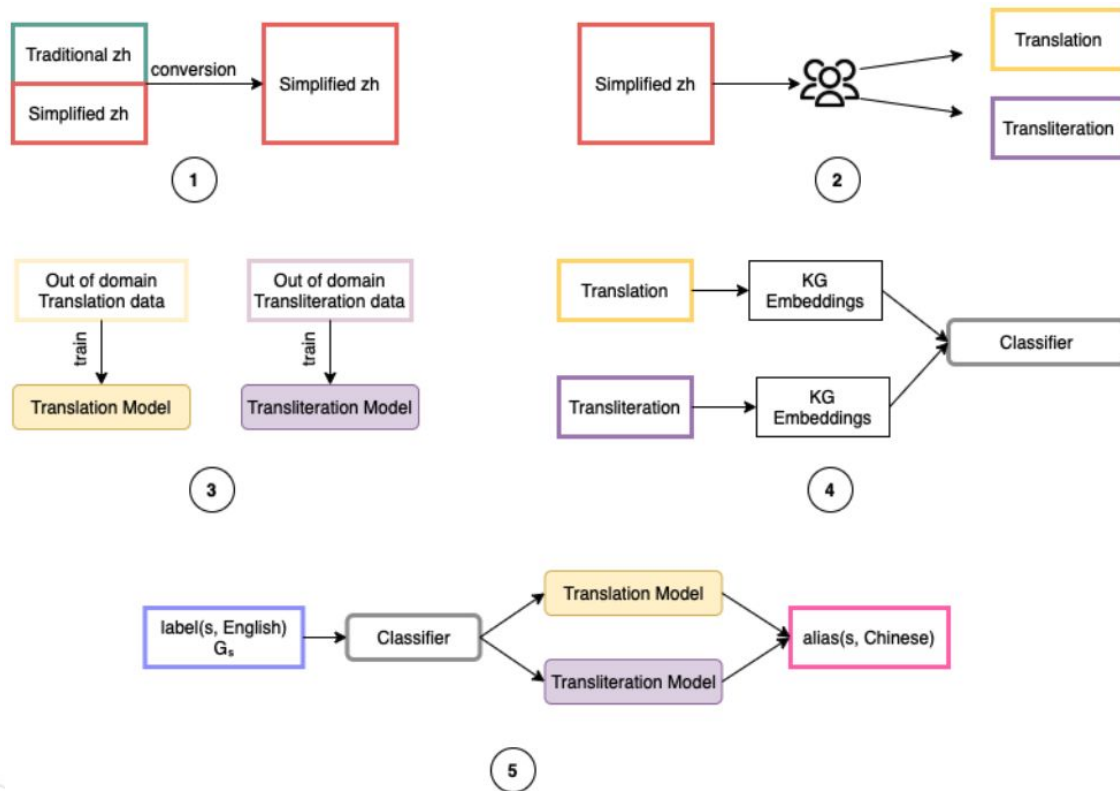


3



4

Translation and transliteration of KG labels



How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

RQ2 Does knowledge about languages in a knowledge graph support ranking them for question answering?

RQ3 Can we automatically translate knowledge graph labels and aliases?

RQ4 Can we reuse Wikidata's multilingual data to generate Wikipedia summaries?

Kaffee et al.: Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders [ESWC 2018]; Kaffee et al.: Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata [NAACL 2018]; Kaffee et al.: Using Natural Language Generation to Bootstrap Missing Wikipedia Articles: A Human-centric Perspective [Semantic Web Journal 2021]

Article Placeholder on Wikipedia

مراكش



54

مُراكش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

معرفةات

معرف مكتبة البرلمان الوطني (NDL)

00629288

معرف فري بيس

m/054rw/

معرف جيونيمز

2542997

بداية (تدشين)

1062

التسمية المحلية

مراكش (العربية)

□□□□□□ (الآمازيغية وسط الأطلس)

Triples from Wikidata

مراكش



54

مُراكش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

معرفةات

معرف مكتبة البرلمان الوطني (NDL)

00629288

معرف فري بيس

m/054rw/

معرف جيونيمز

2542997

بداية (تدشين)

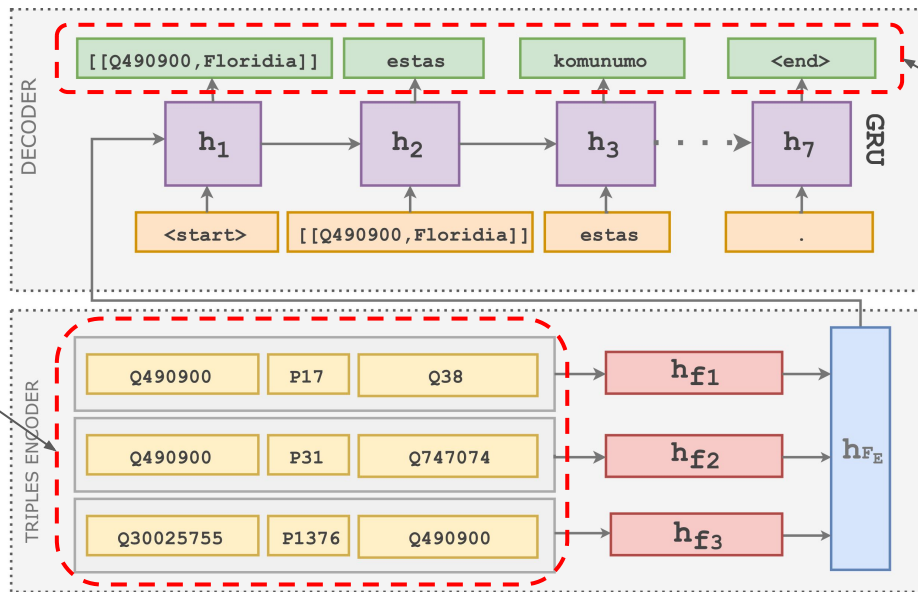
1062

التسمية المحلية

مراكش (العربية)

□□□□□□ (الآمازيغية وسط الأطلس)

Text generation from Wikidata for Wikipedia in low-resource languages



Encoded
Wikidata Triples

Generation of 1-2
sentences
descriptions of the
input triples.

Generated description of the topic

مراكش



51

مُراكش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

معارف

معرف مكتبة البرلمان الوطني (NDL)

00629288

معرف فري بيس

m/054rw/

معرف جيونيمز

2542997

بداية (تدشين)

1062

التسمية المحلية

مراكش (العربية)

□□□□□□ (الآمازيغية وسط الأطلس)

- Tested with Wikipedia readers in Arabic and Esperanto
- Results of the reader study:
We generate sentences of comparable fluency, that “feel” like Wikipedia sentences

		Fluency	Appropriateness	
		Mean	SD	Part of Wikipedia
Arabic	Ours	4.7	1.2	77%
	Wikipedia	4.6	0.9	74%
	News	5.3	0.4	35%
Esper.	Ours	4.5	1.5	69%
	Wikipedia	4.9	1.2	84%
	News	4.2	1.2	52%

Study with Wikipedia editors: We generate sentences that are highly reused by editors

	Category	Examples	%
Arabic	WD	<p>خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.</p> <p>خماسي كلوريد الزرنيخ هو مركب كيميائي له الصيغة (AtClu2085) ، ويكون على شكل بلورات بيضاء.</p>	45.45%
	PD	<p>بيتش باتوم (بالإنجليزية (كلمة ناقصة) Ohio) هي منطقة سكنية تقع في الولايات المتحدة (في (كلمة ناقصة).</p> <p>بيتش باتوم (بالإنجليزية: Beach Batom) هي قرية تقع في الولايات المتحدة الأمريكية في برووك كاونتي.</p>	33.33%
	ND	<p>دير علا هي بلدة تقع في جنوب غرب إيران.</p> <p>دير علا، أو بيثر، هي قرية أردنية</p>	21.21%
Esperanto	WD	<p>Zederik estas komunumo en la nederlanda provinco Zuid-Holland.</p> <p>Zederik estas komunumo en la nederlanda provinco Zuid-Hooland kaj estas ĉirkaŭata de la municipoj Lopik kaj Zederik.</p>	78.98%
	PD	<p>Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando, kiu havis (manka nombro) loĝantojn en (jaro).</p> <p>Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando.</p>	15.79%
	ND	<p>Ibiúna estas municipo de la brazila subŝtato San-Paŭlo, kiu taksis (manka nombro) enloĝantojn en (jaro).</p> <p>Ibiúna estas brazila [[municipo]] kiu troviĝas en la administra unuo [[San-Paŭlo]].</p>	5.26%

- Interview with Wikipedia editors across 6 languages
- Mistakes by the network are ignored
- Can lead to too much trust (no provenance of the text)
- “magical threshold” for the length of a generated text
- Good for new editors (as a starting help)

Language	Sentence displayed to participants	# Participants
Arabic (ar)	مَرَاكُش (بالأمازيغية: <rare>) هي مدينة مغربية تقع شمال البلاد.	4
Swedish (sv)	Marrakech (arabiska <rare>, Berberspråk <rare>) är en stad i sydvästra Marocko, vid foten av <rare>.	2
Hebrew (he)	מרקש (בערבית: <rare>) היא עיר מדברית בדרום מערב מרוקו למרגלות הרי <rare>.	1
Persian (fa)	شهر مراکش (به بربری: <rare>) یکی از شهرهای کشور مراکش و مرکز استان مراکش <rare> است.	1
Indonesian (id)	Marrakesh (Arab: <rare>) ialah kota di barat daya Maroko di kaki <rare>.	1
Ukrainian (uk)	Марракеш (араб. <rare>) — важливе імперське місто в Марокко, розташованого біля підніжжя гір <rare>.	1

A decorative network diagram in the top-left corner, consisting of various sized grey circles (nodes) connected by thin grey lines (edges). Some nodes are solid grey, while others are hollow with a grey outline. The network is dense and irregular, extending from the top-left towards the center.

Outlook on recent work

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It features a cluster of grey nodes of different sizes connected by thin lines. Some nodes are solid grey, and others are hollow with a grey outline. The network is dense and irregular, extending from the bottom-right towards the center.

Entity linking over varying KBs with time

- Time-stratified English Wikipedia snapshots from 2013 to 2022
- Entities that change over time and newly introduced entities for each year
- Decrease of existing entity linking approaches by 3.1% for existing, changing entities
- For new entities this accuracy drop is up to 17.9%

Probing PTLMs for cross-cultural differences in values

- LMs embed knowledge about cultures and values
- Probing how those values align with previous social-science studies of cross-cultural values
- PTLMs capture differences across cultures, i.e., there are differences in the values across different language PTLMs
- Those do not align with previous studies of cross-cultural differences in values

Papers

- Zaporojets et al.: TempEL: Linking Dynamically Evolving and Newly Emerging Entities [Neurips 2022]
- Arora et al.: Probing Pre-Trained Language Models for Cross-Cultural Differences in Values
- Kaffee et al.: Using Natural Language Generation to Bootstrap Missing Wikipedia Articles: A Human-centric Perspective [Semantic Web Journal 2022]
- Amaral et al.: Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach [Semantic Web Journal 2021]
- Kaffee et al.: Ranking Knowledge Graphs By Capturing Knowledge about Languages and Labels [KCAP 2019]
- Kaffee et al.: When Humans and Machines Collaborate: Cross-lingual Label Editing in Wikidata [OpenSym 2019]
- Kaffee and Simperl: The Human Face of the Web of Data: A Cross-sectional Study of Labels [SEMANTiCS 2018]
- Kaffee et al.: Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders [ESWC 2018]
- Kaffee et al.: Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata [NAACL 2018]
- Kaffee and Simperl: Analysis of Editors' Languages in Wikidata [OpenSym 2018]
- Pellissier Tanon and Kaffee: Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases [WikiWorkshop @ WWW 2018]
- Vougiouklis et al.: Neural Wikipedia: Generating Textual Summaries from Knowledge Base Triples [Journal of Web Semantics 2018]
- Kaffee et al.: A glimpse into Babel: an analysis of multilinguality in Wikidata [OpenSym 2017]
- Piscopo et al.: Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References [ISWC 2017]
- Piscopo et al.: What do Wikidata and Wikipedia Have in Common? An Analysis of their Use of External References [OpenSym 2017]

Grants & Awards

- 2022: SWSA Distinguished Dissertation Award
- 2020: Scribe's reference API, *WikiCred Project Grant*
- 2019: Scribe: Supporting Under-resourced Wikipedia Editors in Creating New Articles, *Wikimedia Project Grant*
- 2016: Marie Skłodowska-Curie ITN PhD Fellowship