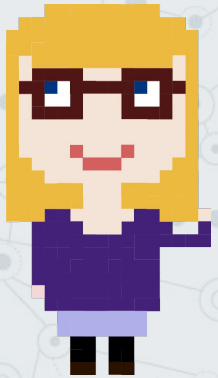
Multilinguality in Knowledge Graphs

Lucie-Aimée Kaffee

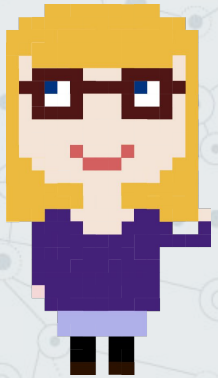
About Lucie



- 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

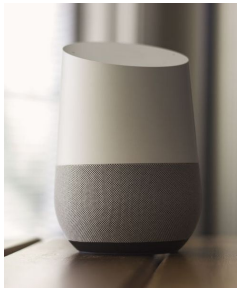
- Motivation
- Research Questions and respective studies/results
- Outlook on current work

Knowledge Graphs and Multilingual Data

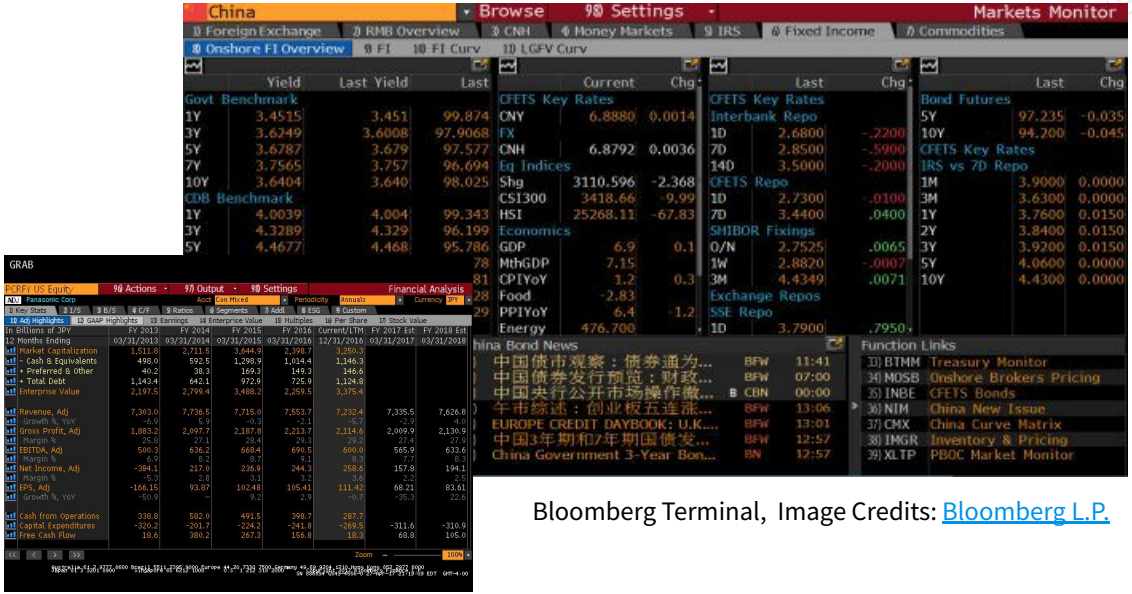
Motivation

- 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]



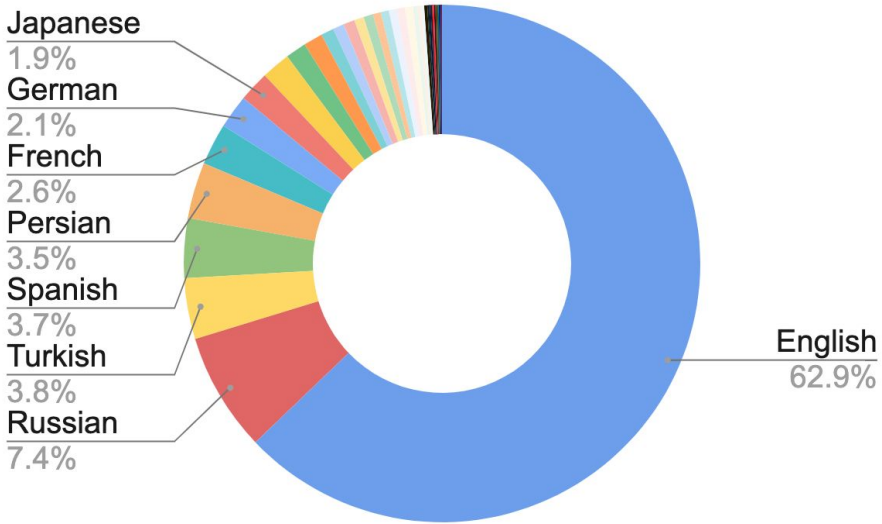
Bloomberg Terminal, Image Credits: [Bloomberg L.P.]

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

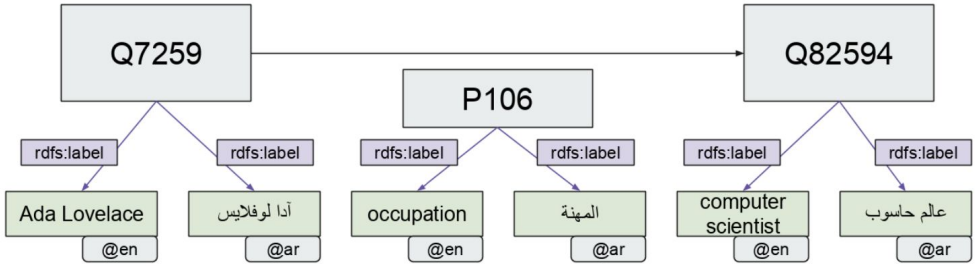
- Lack of language coverage on the web overall



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

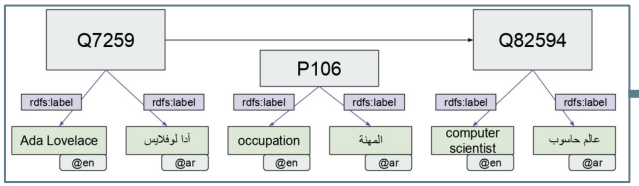
Motivation

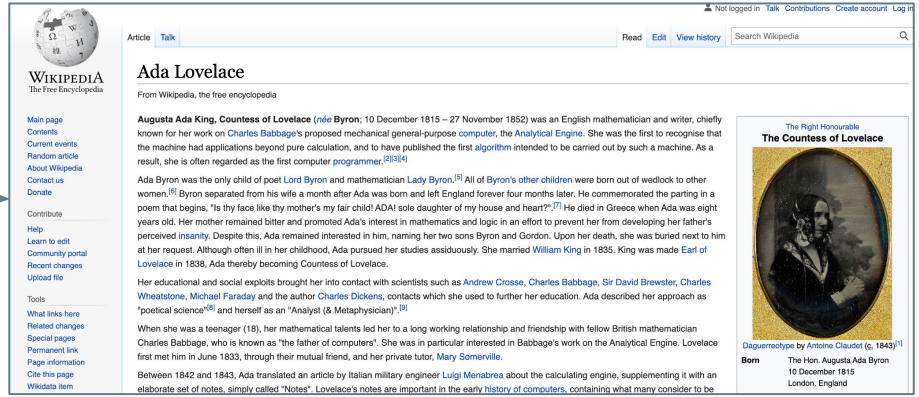
- Knowledge graphs support a multilingual representation of concepts



Motivation

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





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

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]

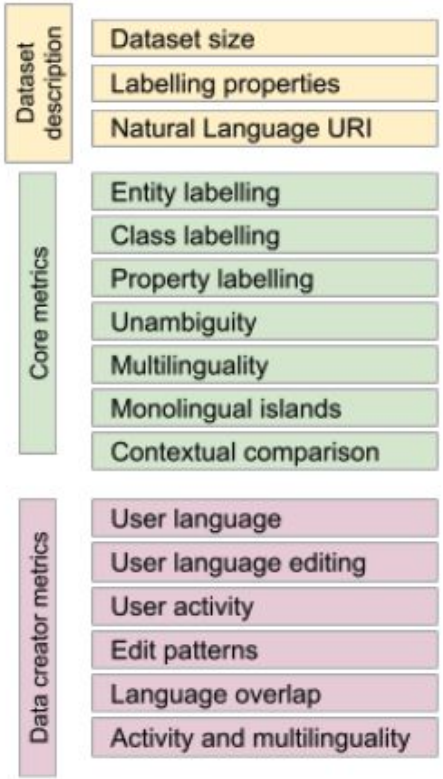
Framework to measure multilinguality in KGs

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

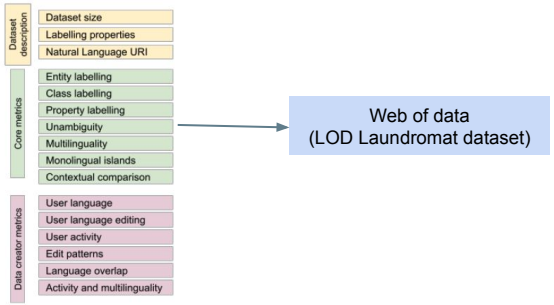
Data creator metrics
User language
User language editing
User activity
Edit patterns
Language overlap
Activity and Multilinguality

Framework to measure multilinguality in KGs



Web of data
(LOD Laundromat dataset)

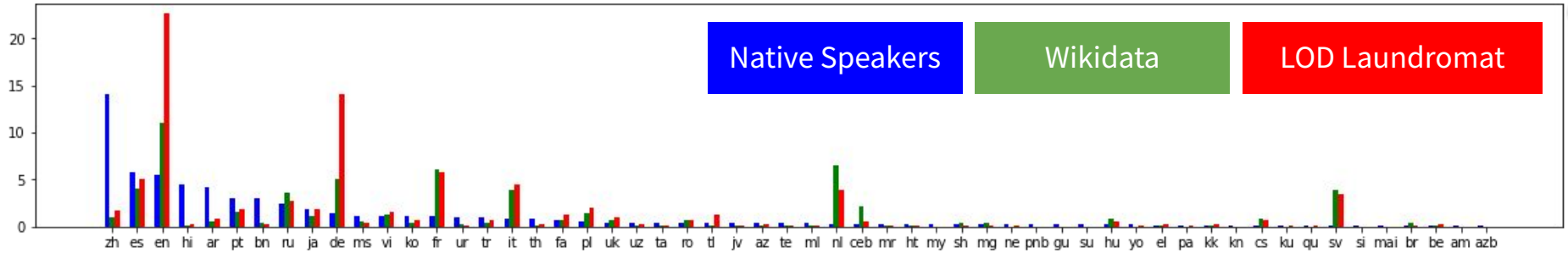
Framework to measure multilinguality in KGs



- **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



Language distribution

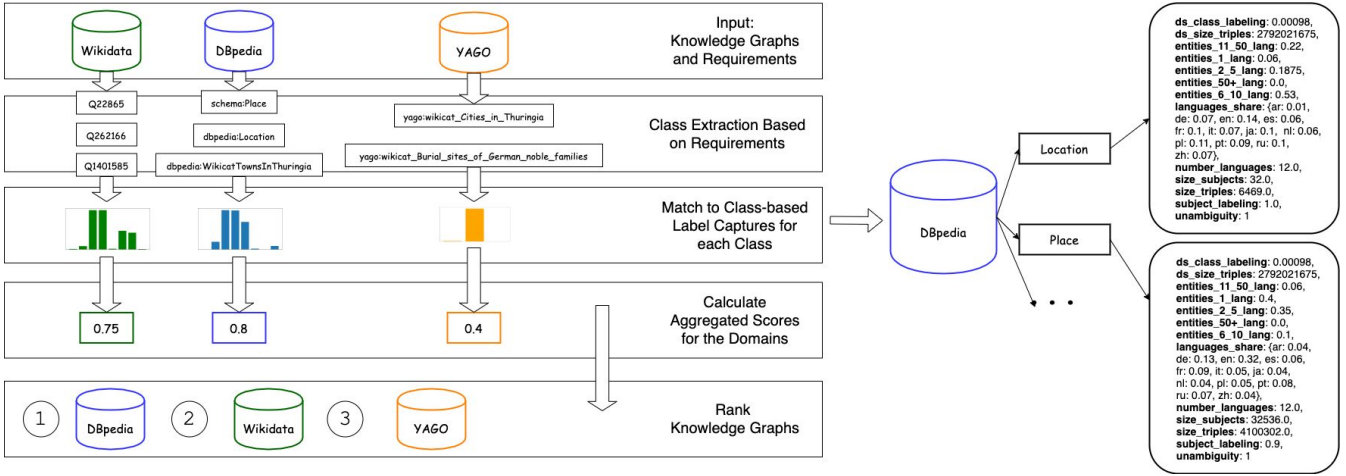
How can we support multilingual access to knowledge graphs for speakers of low-resourced languages?

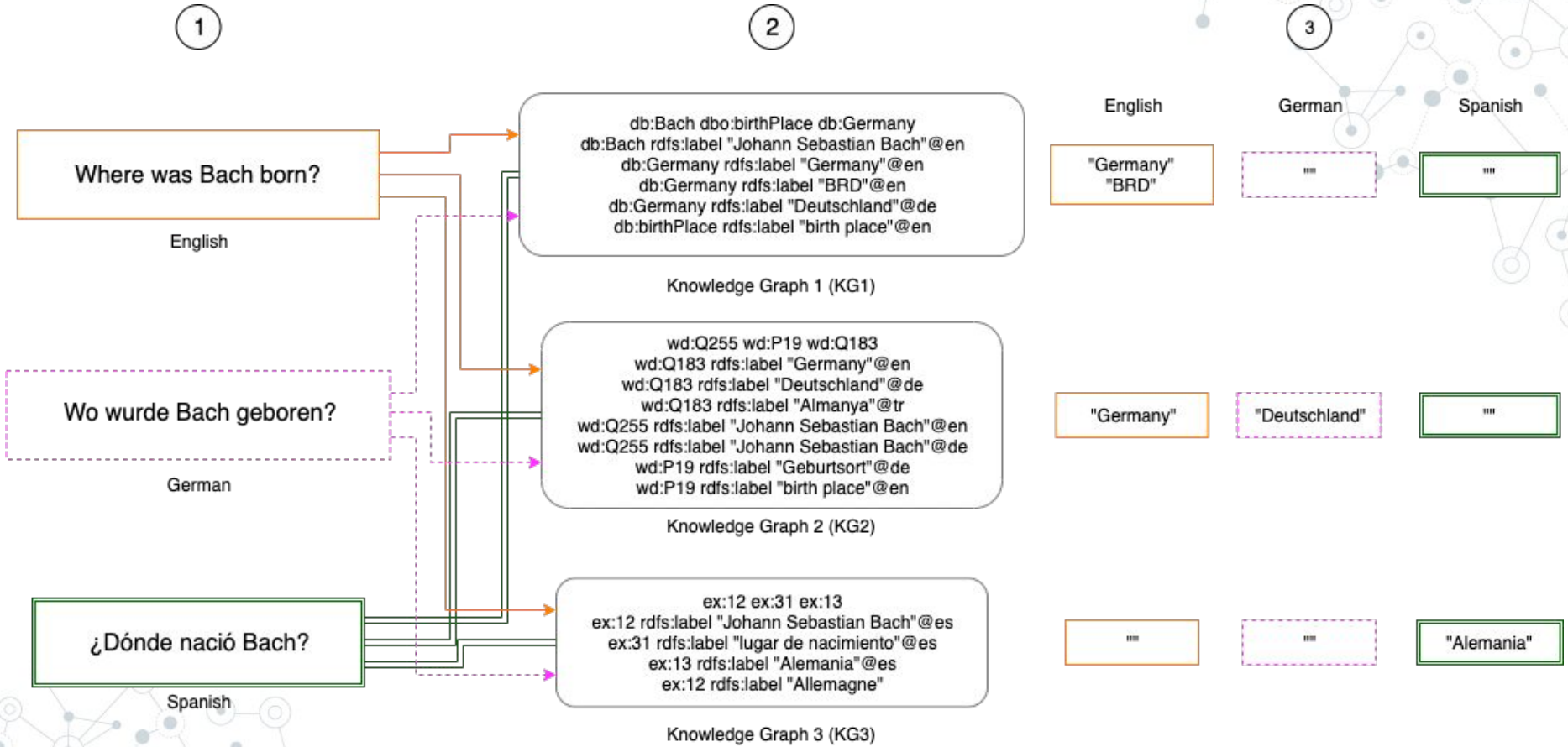
RQ1 What is the state of knowledge graphs with regard to labels and multilinguality?

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

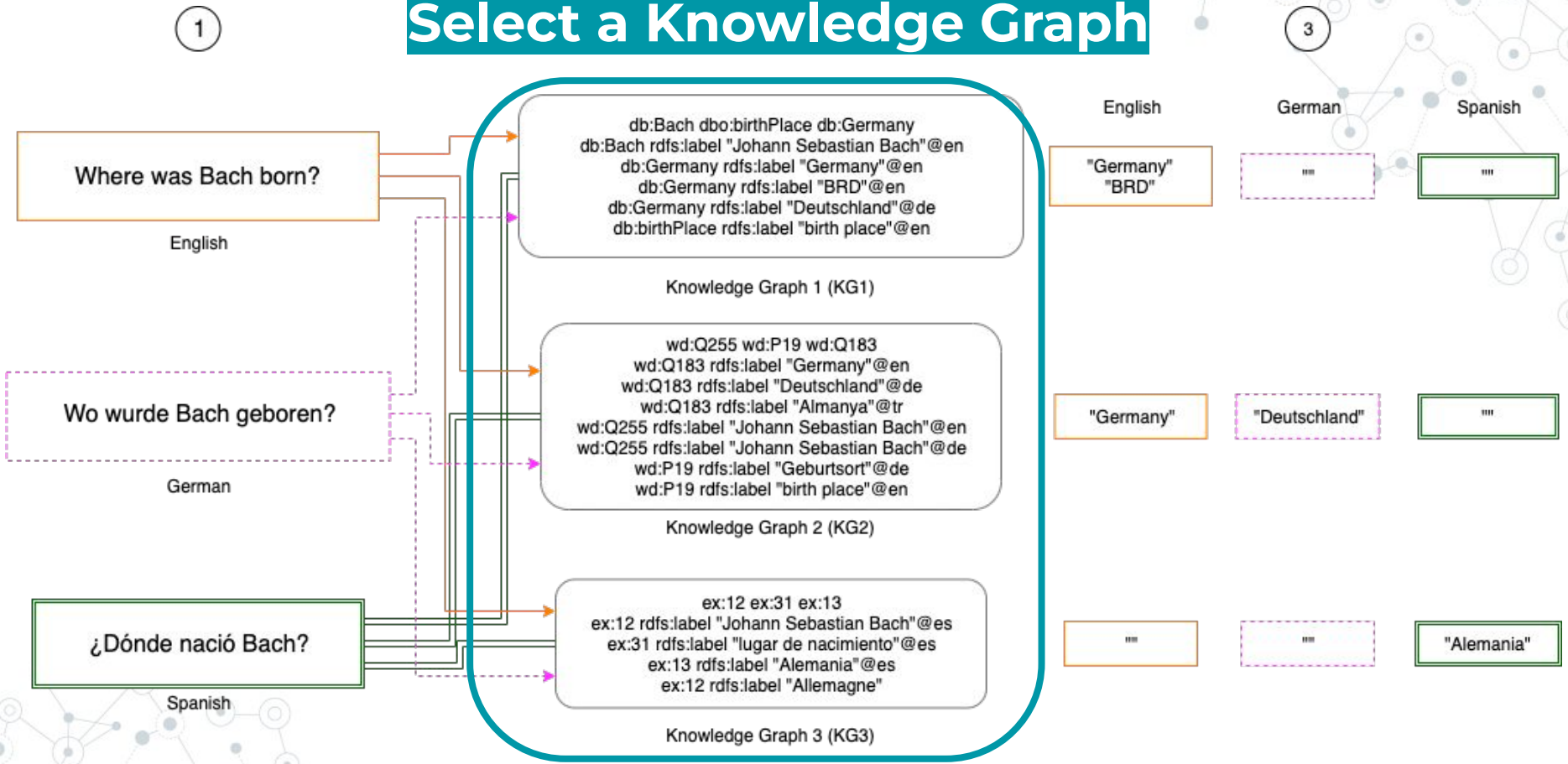
Kaffee et al.: Ranking Knowledge Graphs By Capturing Knowledge about Languages and Labels [KCAP 2019]

LINGVO, framework to rank KGs based on labels and languages

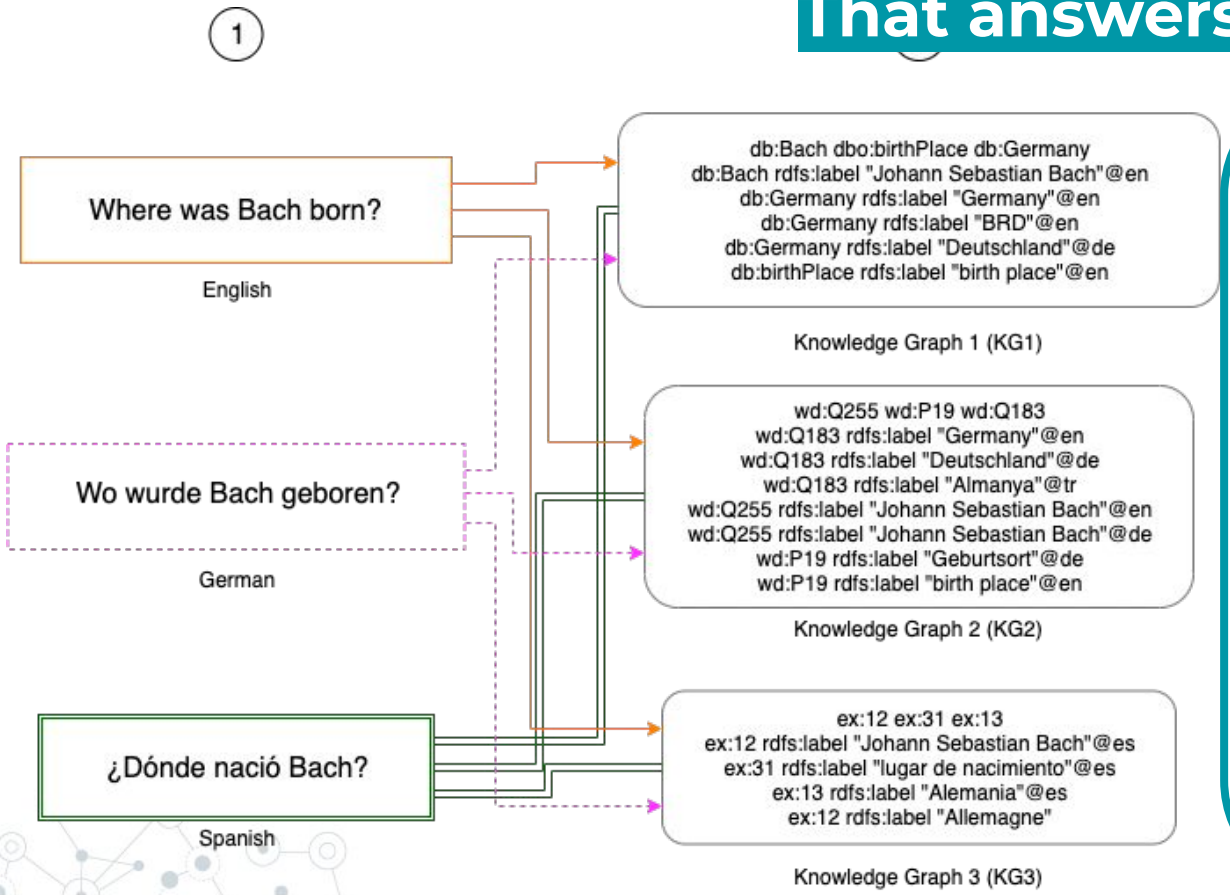



1
2
3
Where was Bach born?
English
Wo wurde Bach geboren?
German
¿Dónde nació Bach?
Spanish
db:Bach dbo:birthPlace db:Germany
db:Bach rdfs:label "Johann Sebastian Bach"@en
db:Germany rdfs:label "Germany"@en
db:Germany rdfs:label "BRD"@en
db:Germany rdfs:label "Deutschland"@de
db:birthPlace rdfs:label "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)
English
German
Spanish
"Germany"
"BRD"
""
""
"Germany"
"Deutschland"
""
""
""
"Alemania"

Select a Knowledge Graph

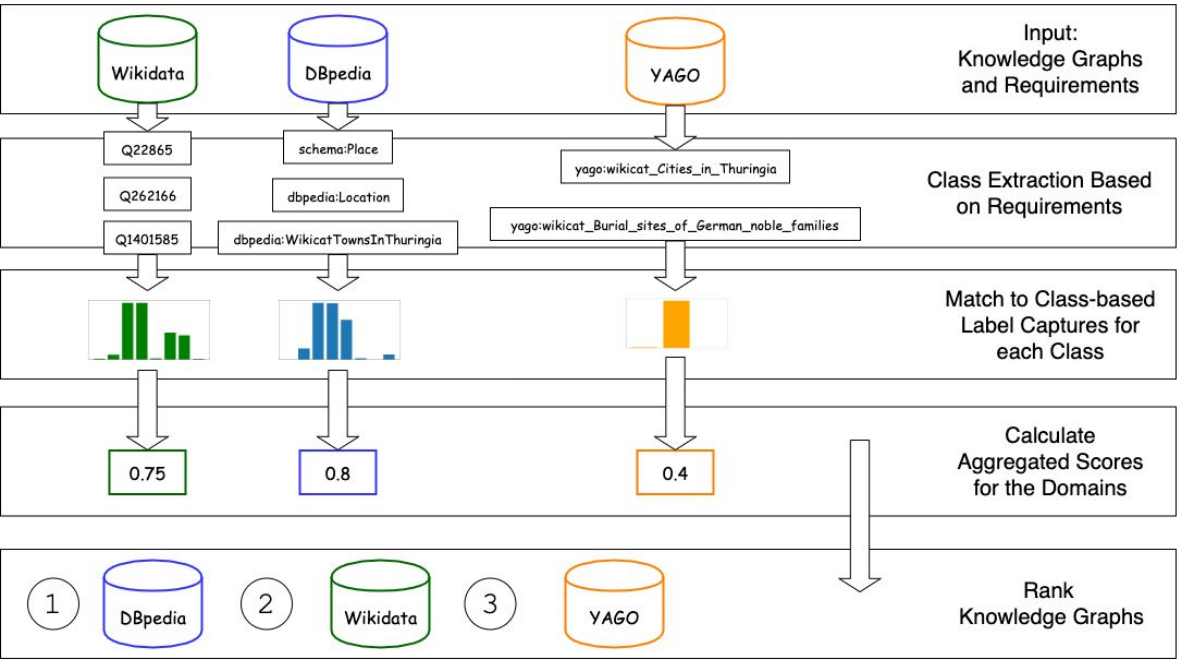


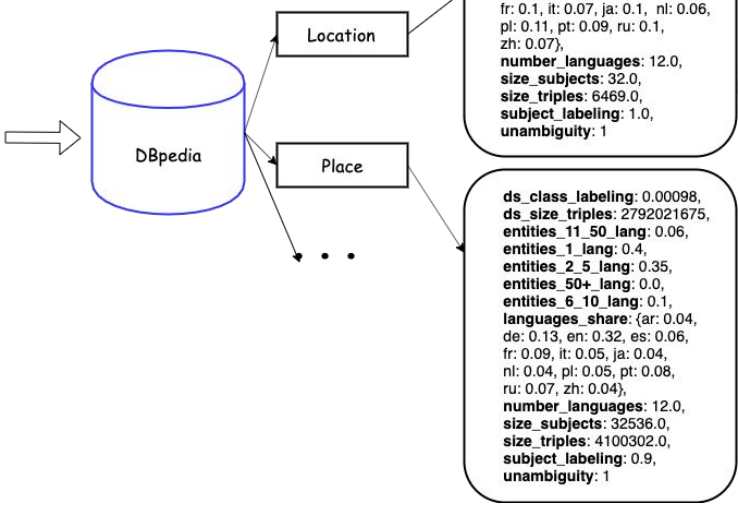
That answers the question best



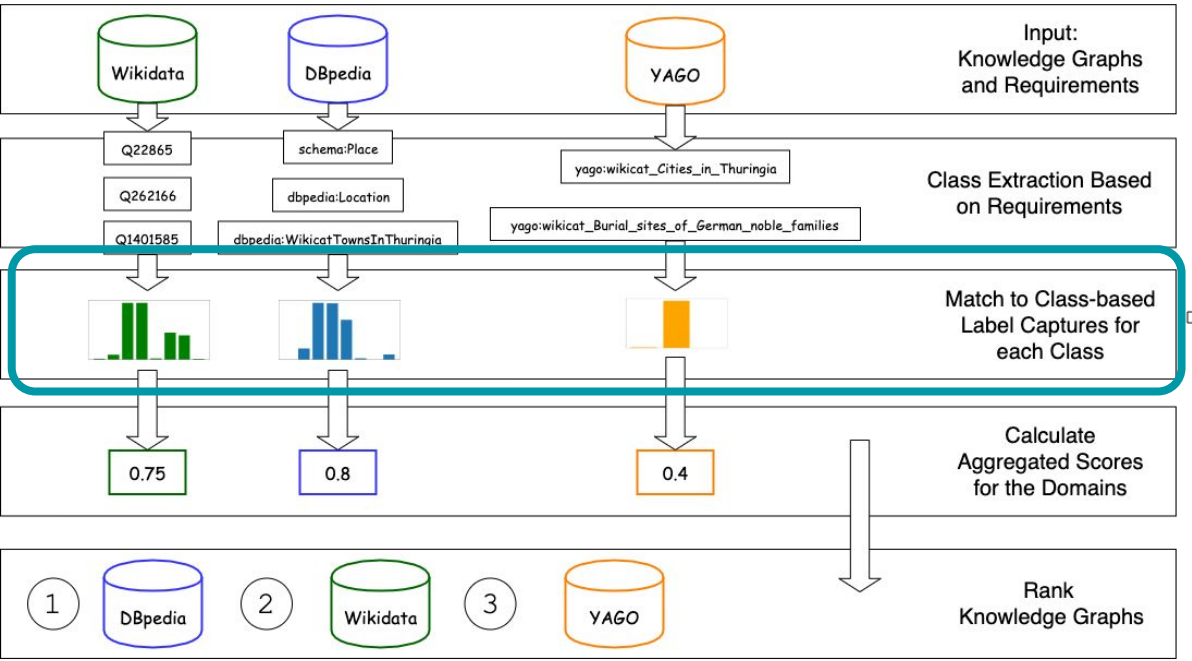
English	German	Spanish
"Germany" "BRD"	""	""
"Germany"	"Deutschland"	""
""	""	"Alemania"

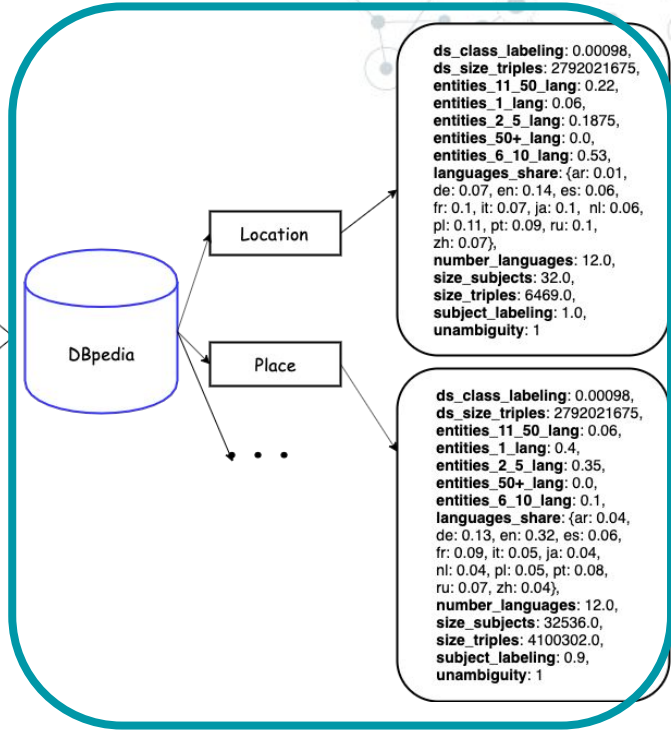
Rank Knowledge Graphs





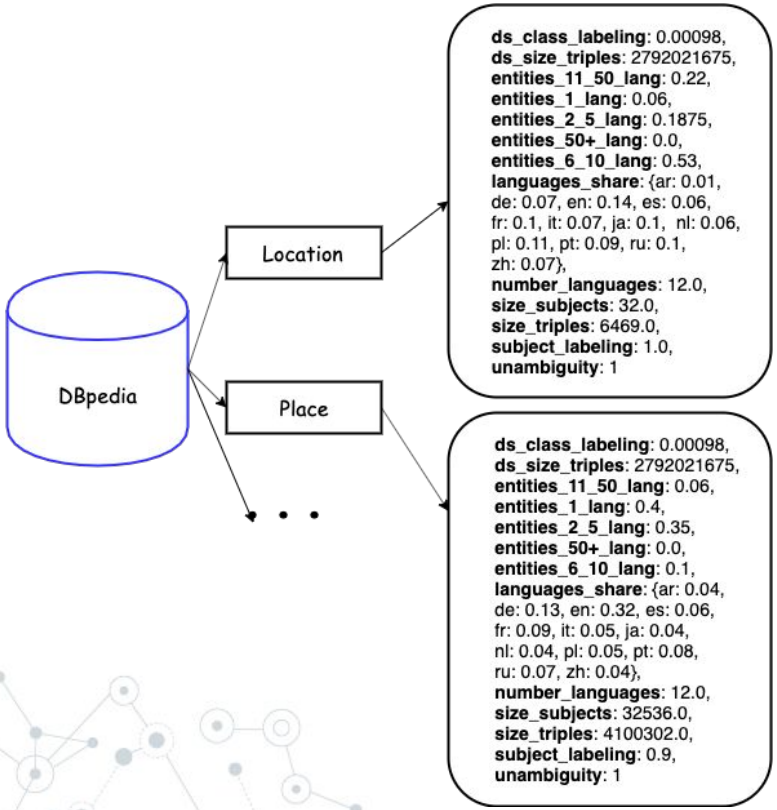
Rank Knowledge Graphs





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
- Seperated 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

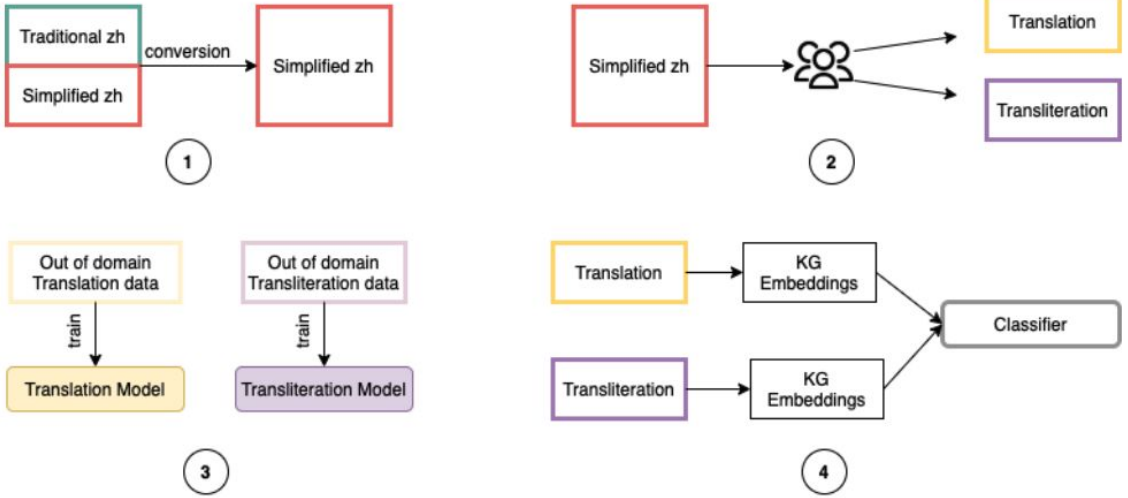
Translation and transliteration of KG labels

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

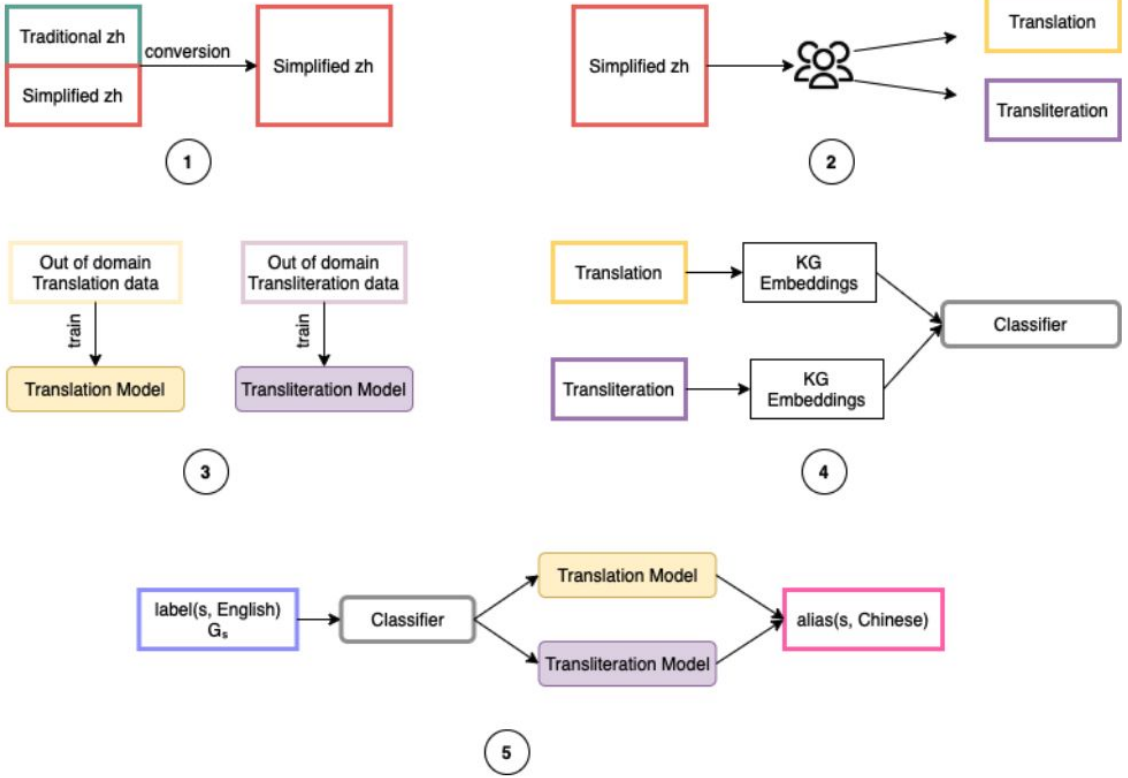
transfers the meaning of a word from one to another language

transfers a word from one script to another

Translation and transliteration of KG labels



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]

ArticlePlaceholder on Wikipedia

مراكش

مُرَاكُش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

معرفات

معرف مكتبة البرلمان الوطني (NDL)
00629288
معرف فري بيس
m/054rw/
معرف جيونيمز
2542997

بداية (تدشين)

1062

التسمية المحلية

مراكش (العربية)
□□□□□□ (الأمازيغية وسط الأطلس)

Triples from Wikidata

مراكش

مُرَّاكُش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

التسمية المحلية

مراكش (العربية)

□□□□□□ (الأمازيغية وسط الأطلس)

بداية (تدشين)

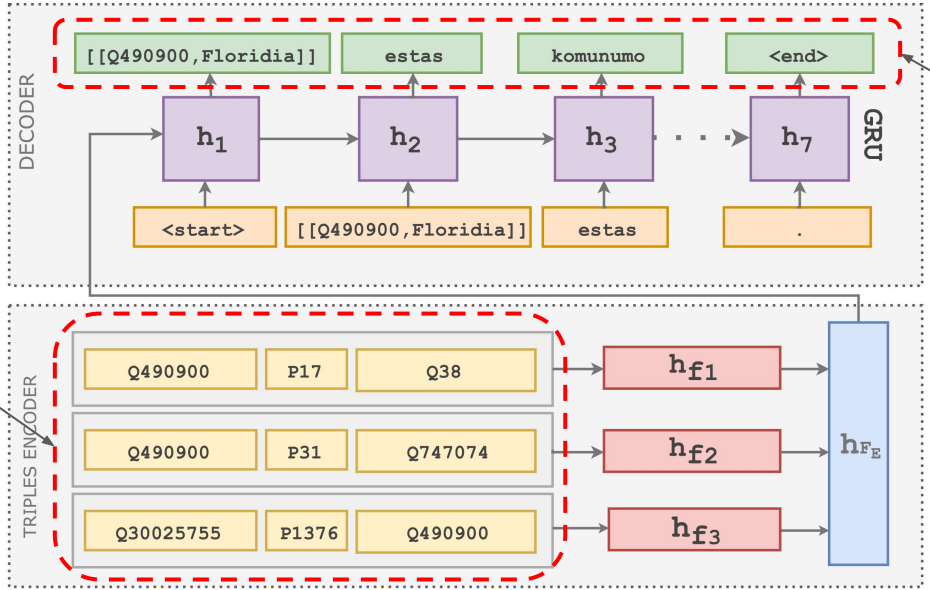
1062

معرفات

معرف مكتبة البرلمان الوطني (NDL)
00629288
معرف فري بيس
/m/054rw
معرف جيونيمز
2542997

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

مُرَّاكُش هي مدينة مغربية تقع شمال البلاد .

حالة خاصة من

مدينة

بداية (تدشين)

1062

التسمية المحلية

مراكش (العربية)

□□□□□□ (الأمازيغية وسط الأطلس)

معرفات

معرف مكتبة البرلمان الوطني (NDL)

00629288

معرف فري بيس

m/054rw/

معرف جيونيمز

2542997

- 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	خماسي كلوريد الزرنيخ هو مُركب كميائي له الصيغة(AtClu2085)، ويكون على شكل بلورات بيضاء.	45.45%
	PD	بيتش باتوم أوهايو (بالإنجليزية (كلمة ناقصة) Ohio)هي منطقة سكنية تقع في الولايات المتحدة في(كلمة ناقصة.)	33.33%
	ND	دير علا هي بلدة تقع في جنوب غرب إيران.	21.21%
Esperanto	WD	Zederik estas komunumo en la nederlanda provinco Zuid-Holland.	78.98%
	PD	Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando, kiu havis (manka nombro) loĝantojn en (jaro).	15.79%
	ND	Ibiúna estas municipo de la brazila subŝtato San-Paŭlio, kiu taksis (manka nombro) enloĝantojn en (jaro).	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)	مُرَاكُش (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

Outlook on recent work

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%

Klim Zaporojets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, Isabelle Augenstein
TempEL: Linking Dynamically Evolving and Newly Emerging Entities, NeurIPS 2022

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

Arnav Arora, Lucie-Aimée Kaffee, Isabelle Augenstein
Probing Pre-Trained Language Models for Cross-Cultural Differences in Values (under review)

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 Wikipedian: 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