Hellina Hailu Nigatu, Min Li, Maartje Ter Hoeve, Saloni Potdar, Sarah Chasins

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Completion for Low-Resourced Languages

Berkeley
UNIVERSITY OF CALIFORNIA

April 16, 2025

Work Currently Under Review!!

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Construction for Low-Resourced Languages.

Anonymous ACL submission

Abstract

Knowledge Graphs represent real-world entities and the relationships between them. Multilingual Knowledge Graph Construction (mKGC) refers to the task of automatically constructing or predicting missing entities and links for knowledge graphs in a multilingual setting. In this work, we reformulate the mKGC task as a Question Answering (QA) task and introduce mRAKL: a Retrieval-Augmented Generation (RAG) based system to perform mKGC. We achieve this by using the head entity and linking relation in a question, and having our model predict the tail entity as an answer. Our experiments focus primarily on two low-resourced languages: Tigrinya and Amharic. We experiment with using higher-resourced languages Arabic and English for cross-lingual transfer. With a BM25 retriever, we find that the RAG-based approach improves performance over a no-context setting. Further, our ablation studies show that with an idealized retrieval system, mRAKL improves accuracy by 4.92 and 8.79 percentage points for Tigrinya and Amharic respectively.

KGs is expensive (Paulheim, 2018). Recent work has investigated the use of pre-trained Language Models (LMs) for KG Construction (e.g. Saxena et al., 2022a; Yao et al., 2019). However, most of the work is focused on English, for which LMs have good performance (Zhou et al., 2022).

Multilingual Knowledge Graph Construction (mKGC) research allows us to (1) extend the downstream benefits of KGs to multiple languages, and (2) capture culturally nuanced and relevant information across languages. However, the challenges of mKGC are exacerbated for languages with limited data available. Prior work using LMs for mKGC relies on pre-training LMs with large amounts of structured data (e.g., Zhou et al. (2022) train on a KG with 52M triples). However, languages on the long tail do not have such datasets available (Joshi et al., 2020). Based on official statistics, only 0.2% of the total entities in Wikidata (Vrandečić and Krötzsch, 2014) have labels in the low-resourced language Amharic.[1] Additionally, most pre-trained LMs do not have good performance for low-resourced languages (Ojo et al., 2024).

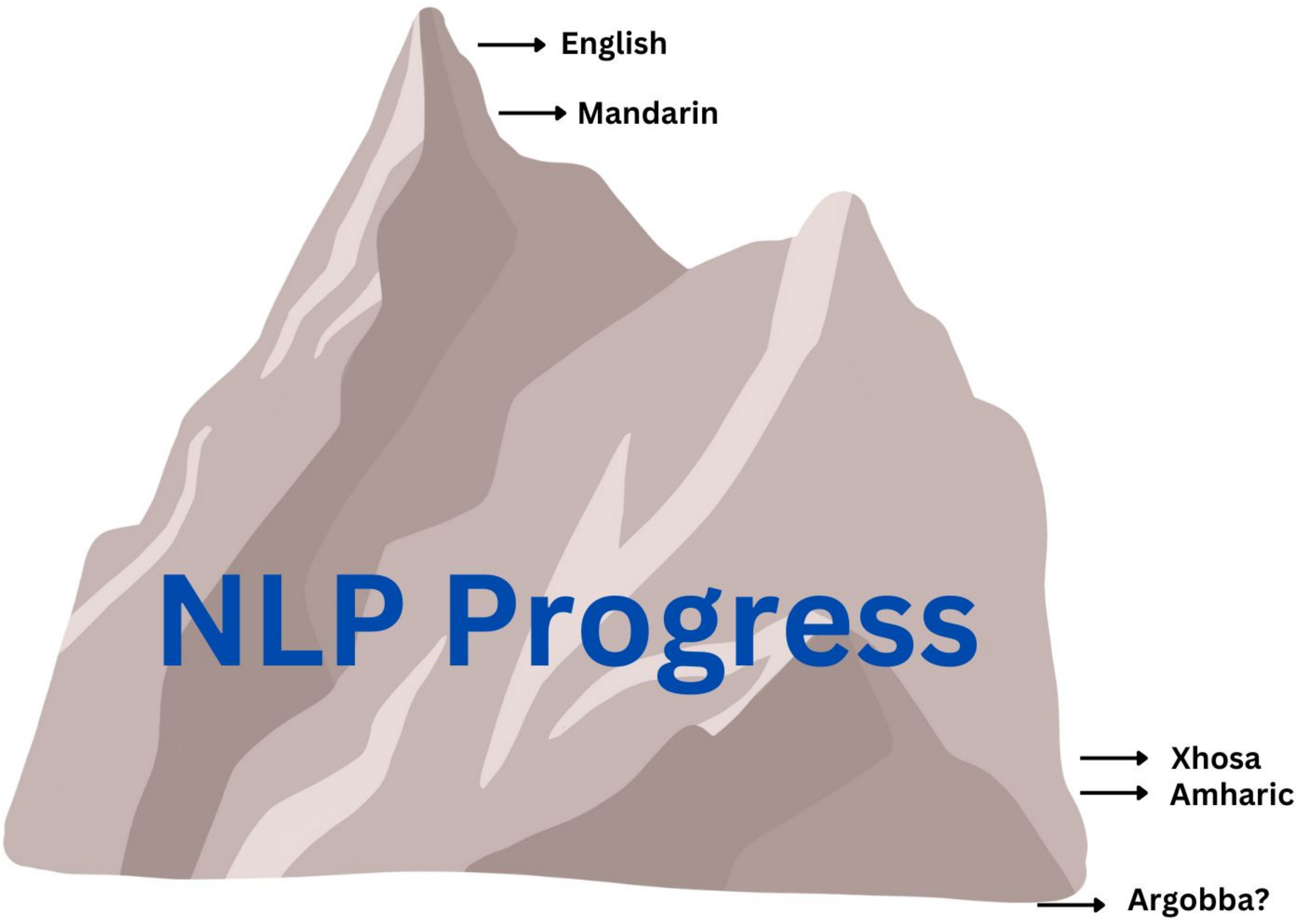
Content

- Motivation
- Method
- Experiments & Results
- Future Work

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Completion for Low-Resourced Languages

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Completion for Low-Resourced Languages

NLP progress is limited to a few languages.



NLP History

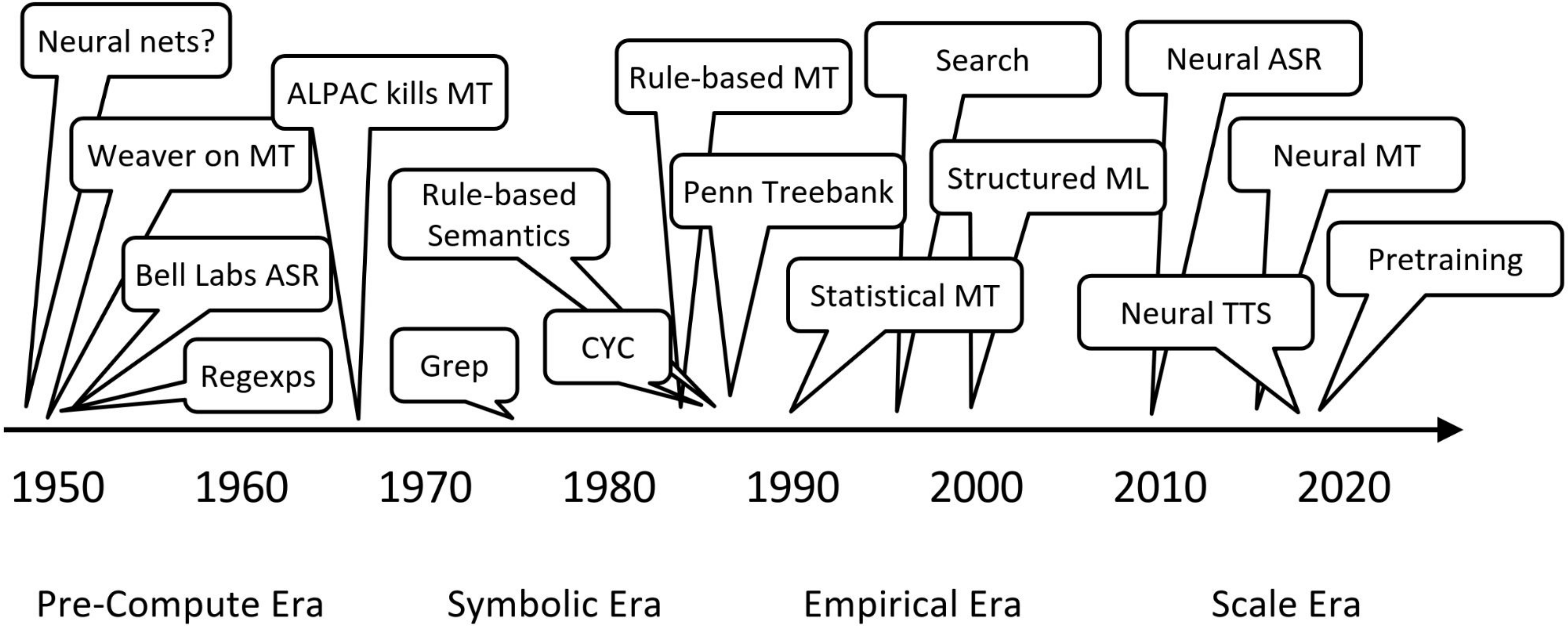


Figure from CS 288: Natural Language Processing by Dan Klein

NLP History

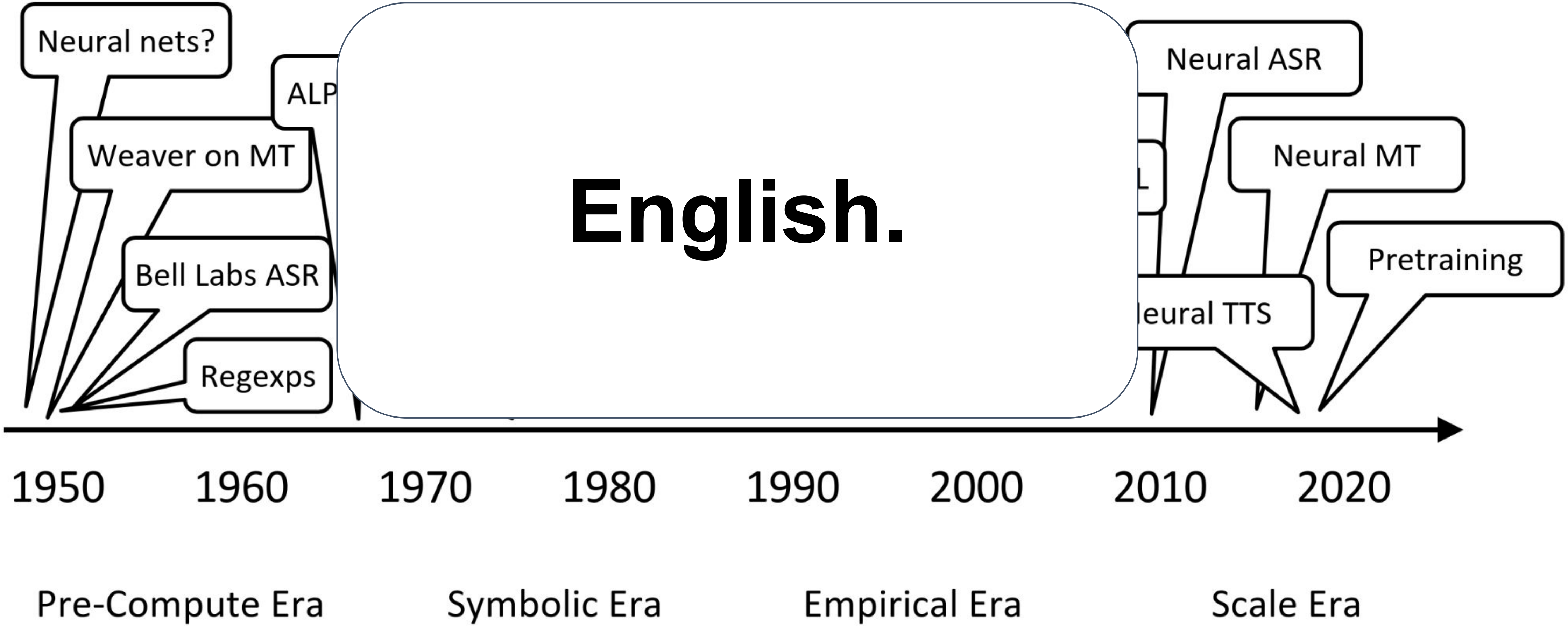
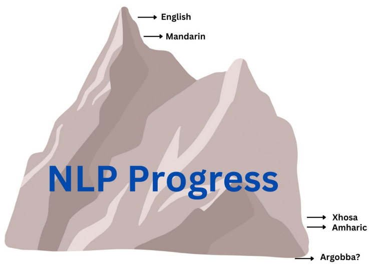


Figure from CS 288: Natural Language Processing by Dan Klein



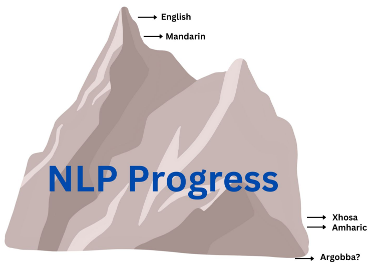
The #BenderRule: On Naming the Languages We Study and Why It Matters

14.SEP.2019 . 15 MIN READ

Emily M. Bender

High Resource Languages vs Low Resource Languages

Progress in the field of Natural Language Processing (NLP) depends on the existence of language resources: digitized collections of written, spoken or signed language, often with gold standard labels or annotations



The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Pratik Joshi* **Sebastin Santy*** **Amar Budhiraja***
Kalika Bali **Monojit Choudhury**
Microsoft Research, India
{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com

Abstract

Language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications. In this paper we look at the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. Our quantitative inves-

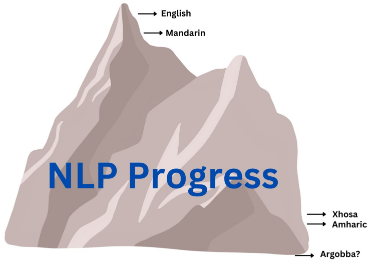
(a) ACL + NAACL + EACL + EMNLP (b) LREC + WS

Figure 1: Number of papers with mentions of **X** and **Y** language for two sets of conferences.

High Resource Lang
Languages

Progress in the field of Natural Language

>88% of the world's languages “are still ignored in the aspect of language technologies.”



High Resource Lang
Languages

Progress in the field of Natural Language
digitized collections of written, spoken o

f Linguistic Diversity and Inclusion in the NLP World

ratik Joshi* Sebastin Santy* Amar Budhiraja*
Kalika Bali Monojit Choudhury
Microsoft Research, India
{t-prjos, t-se ar.budhiraja, kalikab, monojitc}@microsoft.com

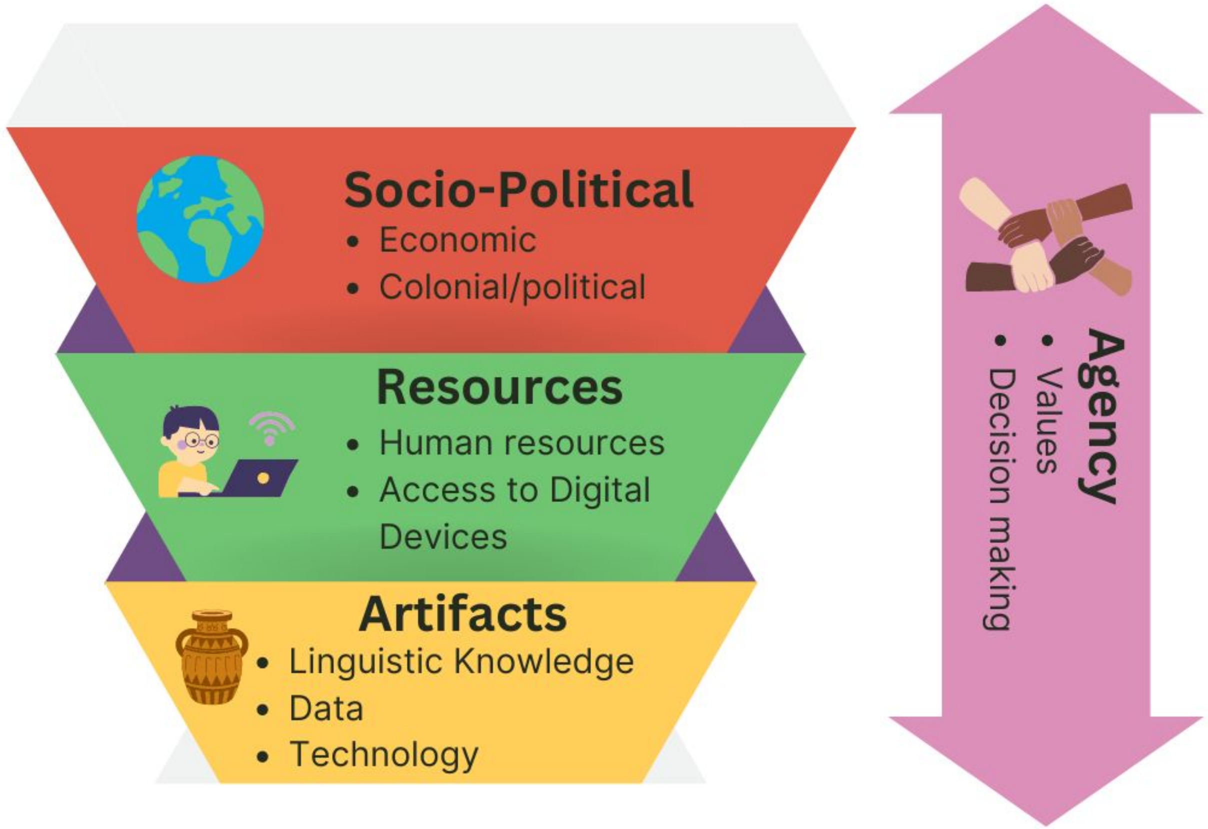
Abstract

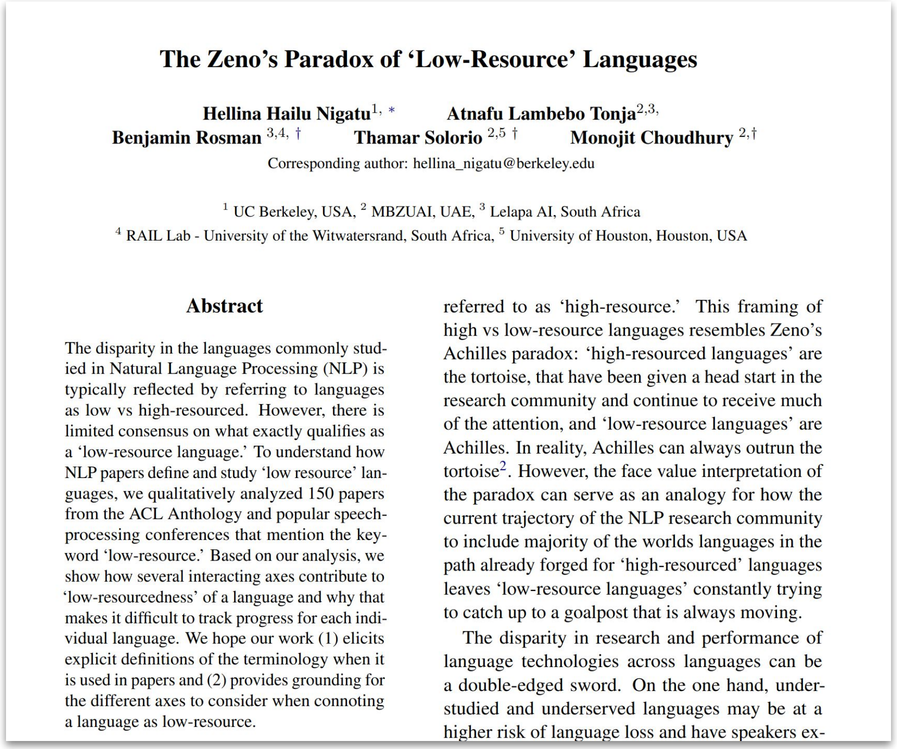
Language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications. In this paper we look at the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. Our quantitative inves-

(a) ACL + NAACL + EACL + EMNLP (b) LREC + WS

Figure 1: Number of papers with mentions of **X** and **Y** language for two sets of conferences.

What is a low-resourced language?





The Zeno's Paradox of 'Low-Resource' Languages

Hellina Hailu Nigatu[1, *] Atnafu Lambebo Tonja[2,3,]
Benjamin Rosman [3,4, †] Thamar Solorio [2,5 †] Monojit Choudhury [2,†]
Corresponding author: hellina_nigatu@berkeley.edu

[1] UC Berkeley, USA, [2] MBZUAI, UAE, [3] Lelapa AI, South Africa
[4] RAIL Lab - University of the Witwatersrand, South Africa, [5] University of Houston, Houston, USA

Abstract

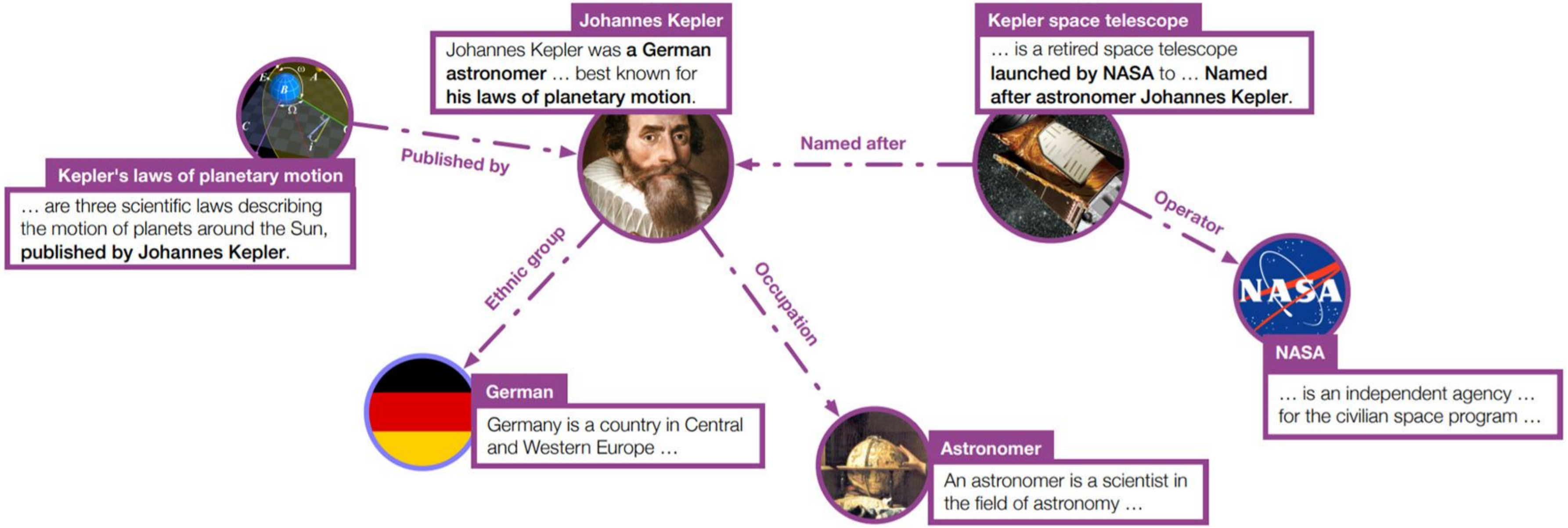
The disparity in the languages commonly studied in Natural Language Processing (NLP) is typically reflected by referring to languages as low vs high-resourced. However, there is limited consensus on what exactly qualifies as a 'low-resource language.' To understand how NLP papers define and study 'low resource' languages, we qualitatively analyzed 150 papers from the ACL Anthology and popular speech-processing conferences that mention the keyword 'low-resource.' Based on our analysis, we show how several interacting axes contribute to 'low-resourcedness' of a language and why that makes it difficult to track progress for each individual language. We hope our work (1) elicits explicit definitions of the terminology when it is used in papers and (2) provides grounding for the different axes to consider when connoting a language as low-resource.

referred to as 'high-resource.' This framing of high vs low-resource languages resembles Zeno's Achilles paradox: 'high-resourced languages' are the tortoise, that have been given a head start in the research community and continue to receive much of the attention, and 'low-resource languages' are Achilles. In reality, Achilles can always outrun the tortoise[2]. However, the face value interpretation of the paradox can serve as an analogy for how the current trajectory of the NLP research community to include majority of the worlds languages in the path already forged for 'high-resourced' languages leaves 'low-resource languages' constantly trying to catch up to a goalpost that is always moving.

The disparity in research and performance of language technologies across languages can be a double-edged sword. On the one hand, understudied and underserved languages may be at a higher risk of language loss and have speakers ex-

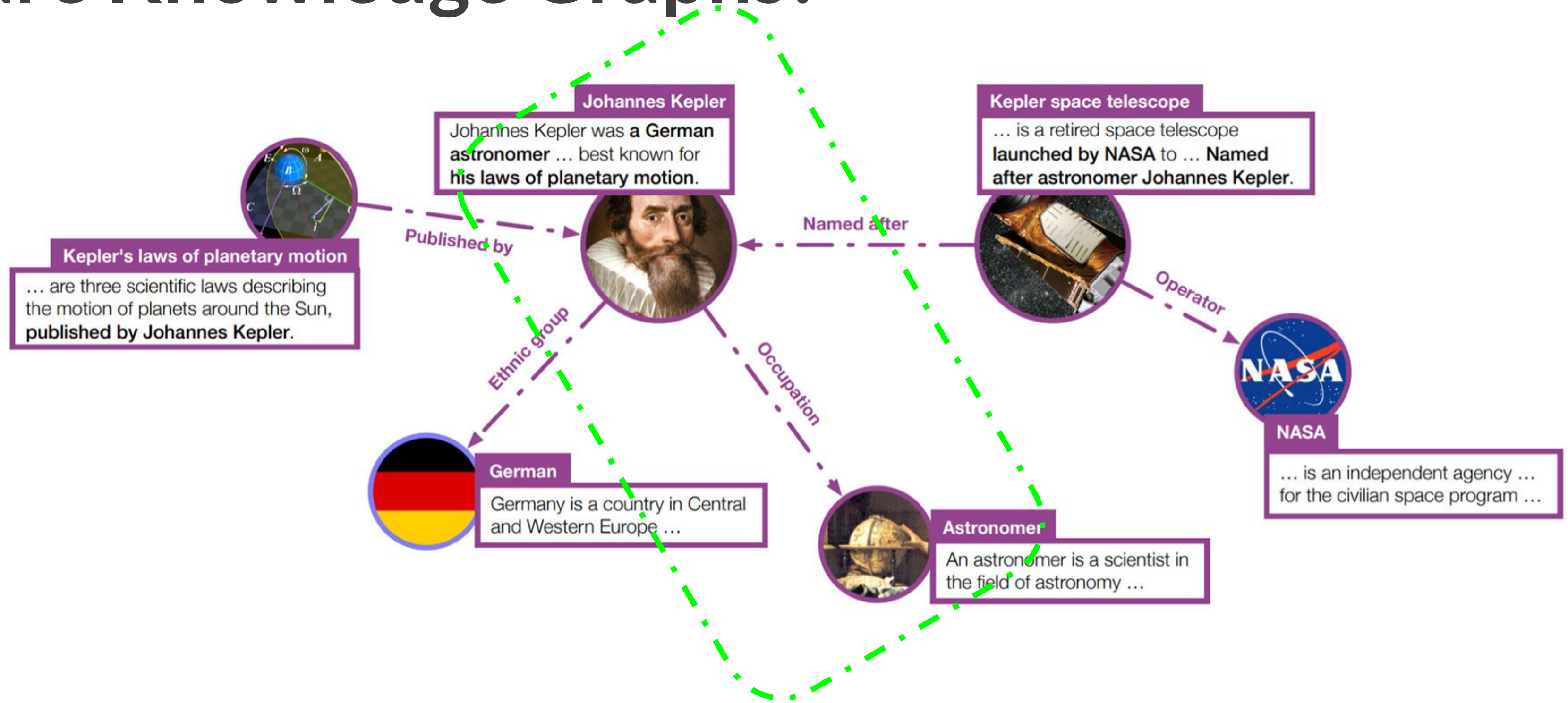
mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Completion for Low-Resourced Languages

What are Knowledge Graphs?



https://deepgraphlearning.github.io/project/wikidata5m

What are Knowledge Graphs?

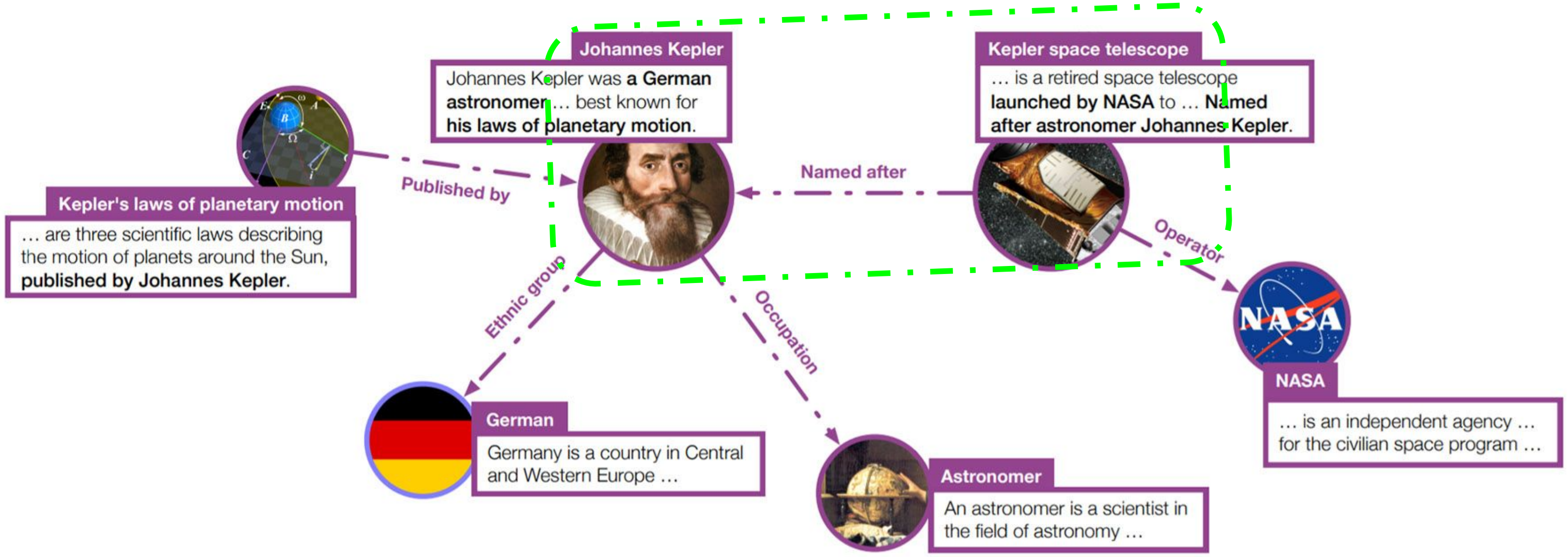


KG: <head, relation, tail>

<Johannes Kepler, Occupation, Astronomer>

https://deepgraphlearning.github.io/project/wikidata5m

What are Knowledge Graphs?



KG: <head, relation, tail>

<Johannes Kepler, Occupation, Astronomer>
<Kepler Space Telescope, Named after, Johannes Kepler>
…

https://deepgraphlearning.github.io/project/wikidata5m

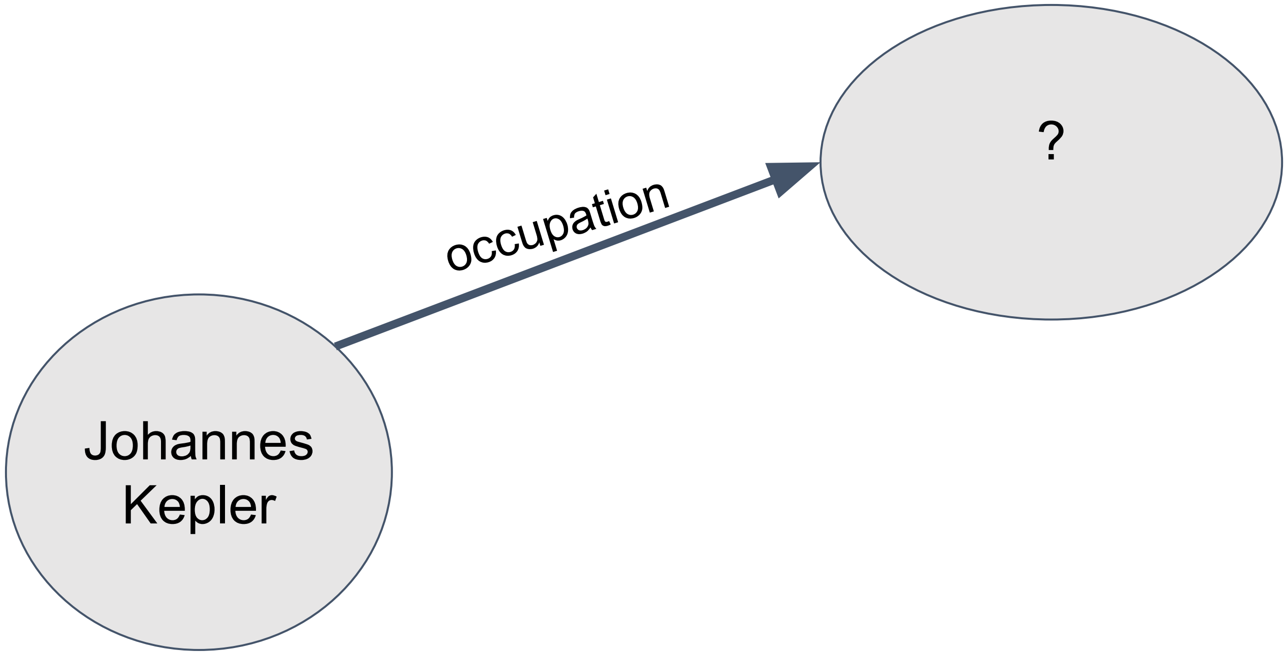
What are Knowledge Graphs?

Have been used in:

- Question Answering
- Dialogue Systems
- Recommendation Systems

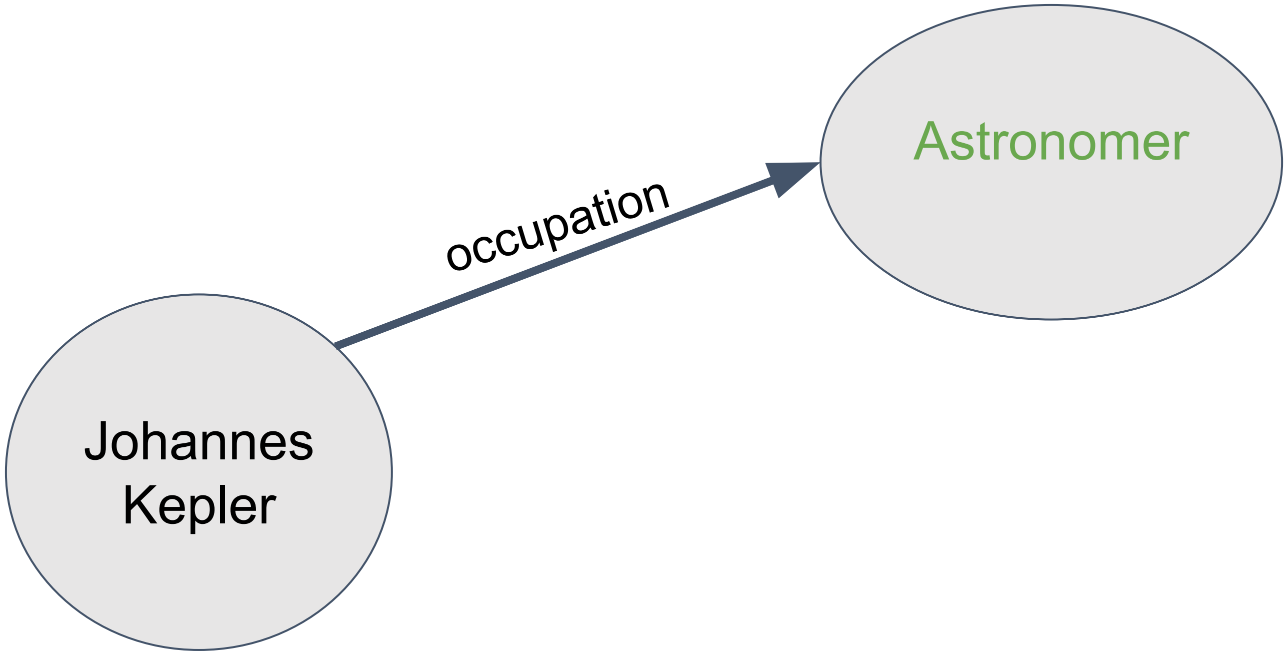
What is Knowledge Graph Completion?

KG: <head, relation, ?>



What is Knowledge Graph Completion?

KG: <head, relation, **tail**>



Knowledge Graph Completion

Option 1: Manual Construction

How much is a Triple?

Estimating the Cost of Knowledge Graph Creation

Heiko Paulheim

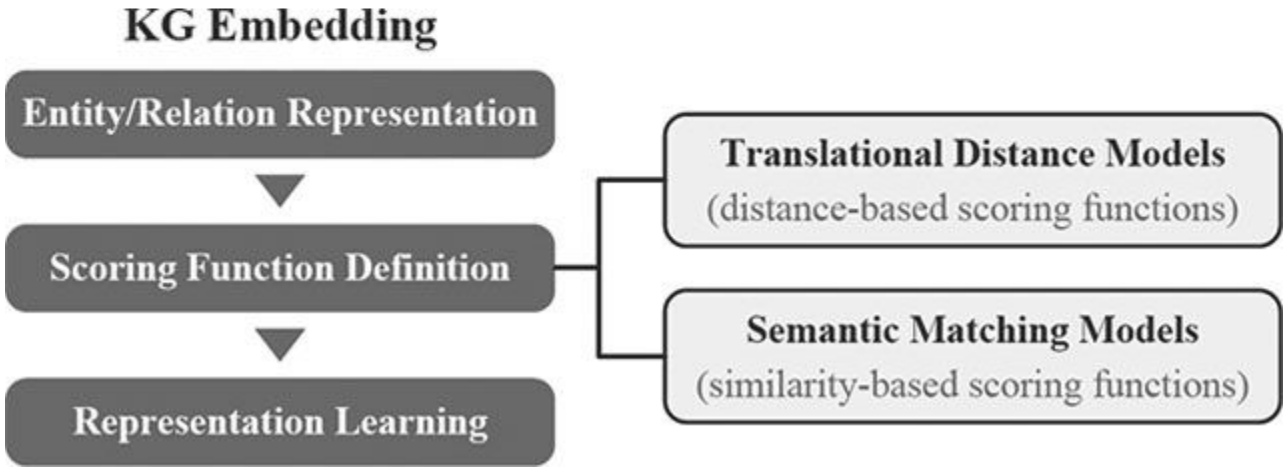
Data and Web Science Group, University of Mannheim, Germany
heiko@informatik.uni-mannheim.de

Abstract. Knowledge graphs are used in various applications and have been widely analyzed. A question that is not very well researched is: what is the price of their production? In this paper, we propose ways to estimate the cost of those knowledge graphs. We show that the cost of manually curating a triple is between \$2 and \$6, and that the cost for automatically created knowledge graphs is by a factor of 15 to 250 cheaper (i.e., 1¢ to 15¢ per statement). Furthermore, we advocate for taking cost into account as an evaluation metric, showing the correspondence between cost per triple and semantic validity as an example.

Keywords: Knowledge Graphs, Cost Estimation, Automation

Knowledge Graph Completion

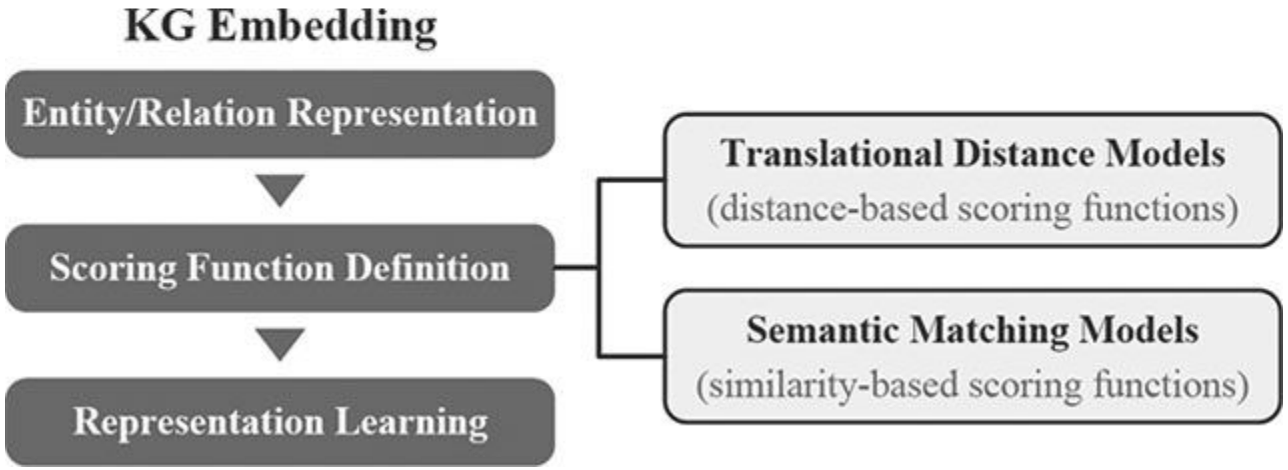
Option 2: KG Embedding Models



Q. Wang, Z. Mao, B. Wang and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724-2743, 1 Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

Knowledge Graph Completion

Option 2: KG Embedding Models

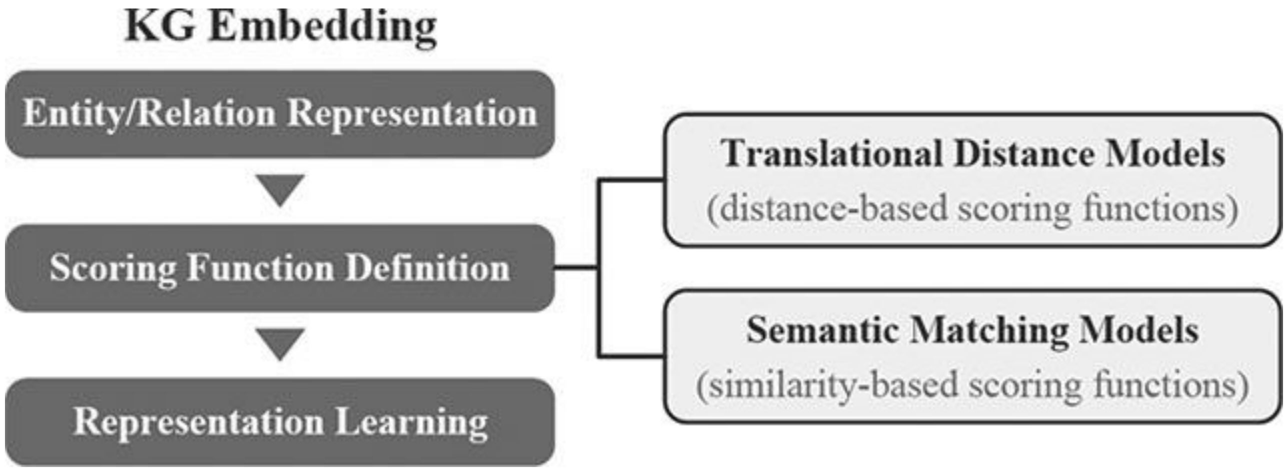


- High quality and performance

Q. Wang, Z. Mao, B. Wang and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724-2743, 1 Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

Knowledge Graph Completion

Option 2: KG Embedding Models



✓

- High quality and performance

✗

- Model size increase with KG size
- Separate model for downstream tasks

Q. Wang, Z. Mao, B. Wang and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724-2743, 1 Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

Knowledge Graph Construction

Option 3: Transformer Based KGE Models

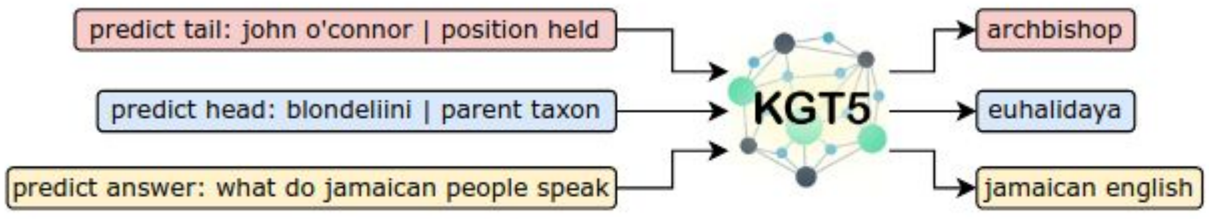


Figure 1: Overview of our method KGT5. KGT5 is first trained on the link prediction task (predicting head/tail entities, given tail/head and relation). For question answering, the same model is further finetuned using QA pairs.

[Sequence-to-Sequence Knowledge Graph Completion and Question Answering]() (Saxena et al., ACL 2022)

Knowledge Graph Construction

Option 3: Transformer Based KGE Models

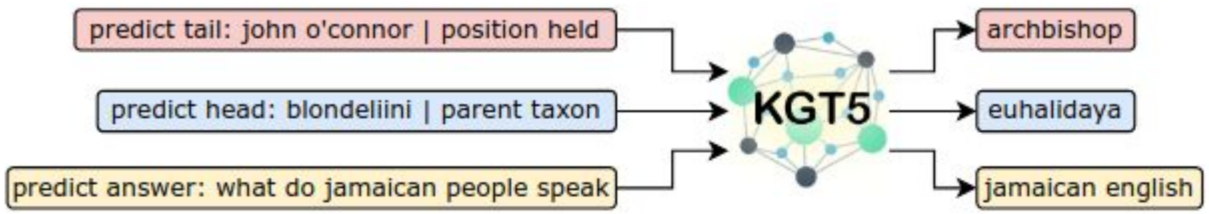


Figure 1: Overview of our method KGT5. KGT5 is first trained on the link prediction task (predicting head/tail entities, given tail/head and relation). For question answering, the same model is further finetuned using QA pairs.

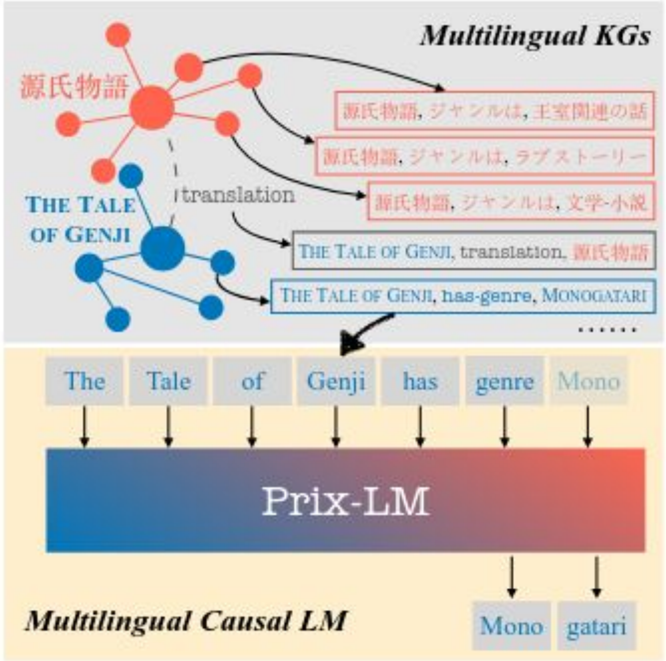


Figure 1: An illustration of the main idea supporting Prix-LM: it infuses complementary multilingual knowledge from KGs into a multilingual causal LM; e.g., Japanese KG stores more comprehensive genre information of THE TALE OF GENJI than KGs in other languages. Through cross-lingual links (translations), such knowledge is then propagated across languages.

[Sequence-to-Sequence Knowledge Graph Completion and Question Answering]() (Saxena et al., ACL 2022)

[Prix-LM: Pretraining for Multilingual Knowledge Base Construction]() (Zhou et al., ACL 2022)

Knowledge Graph Construction

Option 3 + Context: Transformer Based KGE Models with Context

Figure 1: Overview of KGT5-context (at bottom) and comparison to KGT5 (on top); real example from Wikidata5M, best viewed in color. KGT5-context differs from KGT5 in that it appends the neighboring relations and entities of *Yambaó* (a drama movie) to the verbalized query. Both models then apply T5, sample predictions from the decoder, map the samples to entities, and rank by sample logit scores.

[Friendly Neighbors: Contextualized Sequence-to-Sequence Link Prediction](#) (Kochsiek et al., RepL4NLP 2023)

Knowledge Graph Construction

Option 3: Transformer Based KGE Models with/without Context

✓

- Good quality and performance
- Model size independent of KG size
- Can use knowledge in pre-training
- End-to-end trainable for downstream tasks

✗

- Mostly for mid- to high resourced languages
- require structured data for context

[Friendly Neighbors: Contextualized Sequence-to-Sequence Link Prediction]() (Kochsiek et al., RepL4NLP 2023)

How do we make this work for Low-resourced languages?

⚠️ Current methods rely on large, structured data for training!

How do we make this work for Low-resourced languages?

⚠️ Current methods rely on large, structured data for training!

We do not currently have large structured datasets for low-resourced languages

How do we make this work for Low-resourced languages?

⚠️ Current methods rely on large, structured data for training!

We do not currently have large structured datasets for low-resourced languages

Unstructured data is more easily accessible for low-resourced languages

How can we use unstructured data which is more easily accessible to perform mKGC for low-resourced languages?

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Construction for Low-Resourced Languages

Method

Step 1: Reformulating KGC as a Question answering task

<Johannes Kepler, Occupation, ?> → <What is Johannes Kepler’s Occupation?>

Method

Step 1: Reformulating KGC as a Question answering task

Step 2: Have a generative model predict the answer given the question

What is Johannes Kepler's Occupation?

generative LM

Astronomy

Method

Step 1: Reformulating KGC as a Question answering task

Step 2: Have a generative model predict the answer given the question

Step 3: Retrieve context from unstructured data to aid generation

retriever Model

[... Kepler … fathers of modern astronomy]
What is Johannes Kepler's Occupation?

generative LM

Astronomy

Method

Step 1: Reformulating KGC as a Question answering task

Step 2: Have a generative model predict the answer given the question

Step 3: Retrieve context from unstructured data to aid generation

retriever Model

[... Kepler … fathers of modern astronomy]
What is Johannes Kepler's Occupation?

generative LM

Astronomy

Step 4: Do this multilingually!

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

Step 1: Extract Triples from Wikidata

Triple

(Q106368583, P19, Q115)

Step 2: Get labels for the head, relation, and tail in each language.

Language	Labels
Tigrinya	(ሱራፌኤል ዳግናቸው , ቦታ ልደት, ኢትዮጵያ)
English	(Surafel Dagnachew, place of birth, Ethiopia)
Amharic	(ሱራፌል ዳኛቸው , የትውልድ ቦታ, ኢትዮጵያ)
Arabic	(سورافيل داجناشيو, مكان الولادة, إثيوبيا)

Step 3: Verbalize the triple as a question-answer pair.

Language	Question	Answer
Tigrinya	ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?	ኢትዮጵያ
English	What is Surafel Dagnachew's place of birth?	Ethiopia
Amharic	የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?	ኢትዮጵያ
Arabic	ما هو مكان ولادة سورافيل داجناشيو ؟	إثيوبيا

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

Step 1: Extract Triples from Wikidata

Triple

(Q106368583, P19, Q115)

Step 2: Get labels for the head, relation, and tail in each language.

Language	Labels
Tigrinya	(ሱራፊኤል ዳግናቸው , ቦታ ልደት, ኢትዮጵያ)
English	(Surafel Dagnachew, place of birth, Ethiopia)
Amharic	(ሱራፌል ዳኛቸው , የትውልድ ቦታ, ኢትዮጵያ)
Arabic	(سورافيل داجناشيو, مكان الولادة, إثيوبيا)

Step 3: Verbalize the triple as a question-answer pair.

Language	Question	Answer
Tigrinya	ናይ ሱራፊኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?	ኢትዮጵያ
English	What is Surafel Dagnachew's place of birth?	Ethiopia
Amharic	የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?	ኢትዮጵያ
Arabic	ما هو مكان ولادة سورافيل داجناشيو ؟	إثيوبيا

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

Step 1: Extract Triples from Wikidata

Triple

(Q106368583, P19, Q115)

Step 2: Get labels for the head, relation, and tail in each language.

Language	Labels
Tigrinya	(ሱራፊኤል ዳግናቸው , ቦታ ልደት, ኢትዮጵያ)
English	(Surafel Dagnachew, place of birth, Ethiopia)
Amharic	(ሱራፌል ዳኛቸው , የትውልድ ቦታ, ኢትዮጵያ)
Arabic	(سورافيل داجناشيو, مكان الولادة, إثيوبيا)

Step 3: Verbalize the triple as a question-answer pair.

Language	Question	Answer
Tigrinya	ናይ ሱራፊኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?	ኢትዮጵያ
English	What is Surafel Dagnachew's place of birth?	Ethiopia
Amharic	የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?	ኢትዮጵያ
Arabic	ما هو مكان ولادة سورافيل داجناشيو ؟	إثيوبيا

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

KG	Triples	Head	Tail
Tigrinya	3.5k	244	170
Amharic	34k	8568	5058

Table 1: Details on size of KGs in the two target languages.

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

KG	Triples	Head	Tail
Tigrinya	3.5k	244	170
Amharic	34k	8568	5058

Table 1: Details on size of KGs in the two target languages.

		Tigrinya KG		Amharic KG	
Language	Wiki	Head	Tail	Head	Tail
Amharic	14.04K	79.50	86.47	100	100
Arabic	1.23M	95.49	99.41	79.56	94.36
English	6.84M	100	100	90.40	98.39
Tigrinya	506	100	100	3.60	4.03

Table 2: Percentage of the head and tail entities in each of the target language KGs with textual representations in each of the transfer languages.

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

KG	Triples	Head	Tail
Tigrinya	3.5k	244	170
Amharic	34k	8568	5058

Table 1: Details on size of KGs in the two target languages.

		Tigrinya KG		Amharic KG	
Language	Wiki	Head	Tail	Head	Tail
Amharic	14.04K	79.50	86.47	100	100
Arabic	1.23M	95.49	99.41	79.56	94.36
English	6.84M	100	100	90.40	98.39
Tigrinya	506	100	100	3.60	4.03

Table 2: Percentage of the head and tail entities in each of the target language KGs with textual representations in each of the transfer languages.

Method

Target Languages: Amharic and Tigrinya

Transfer Languages: Arabic and English

Preparing our dataset

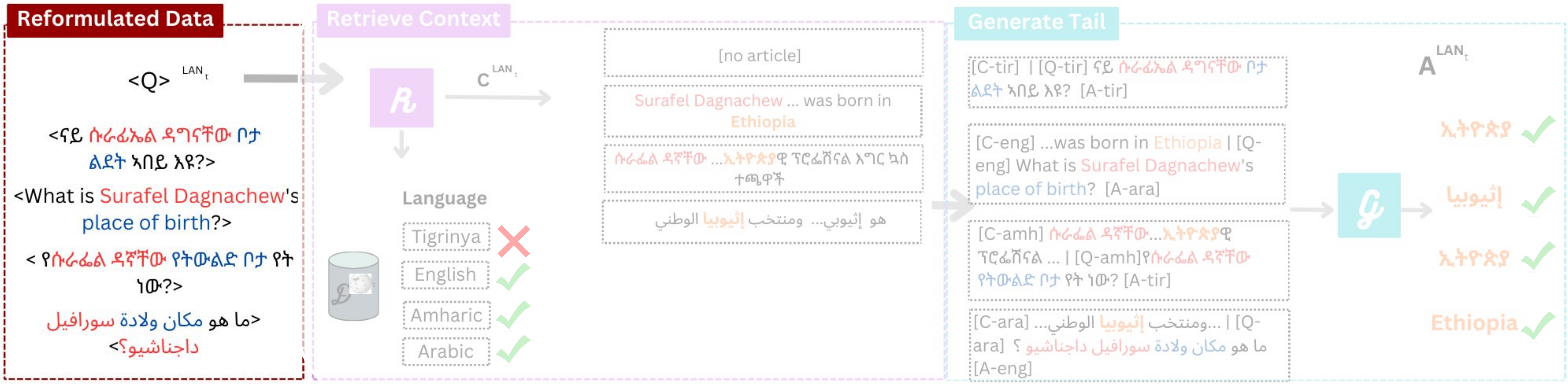
KG	Triples	Head	Tail
Tigrinya	3.5k	244	170
Amharic	34k	8568	5058

Table 1: Details on size of KGs in the two target languages.

		Tigrinya KG		Amharic KG	
Language	Wiki	Head	Tail	Head	Tail
Amharic	14.04K	79.50	86.47	100	100
Arabic	1.23M	95.49	99.41	79.56	94.36
English	6.84M	100	100	90.40	98.39
Tigrinya	506	100	100	3.60	4.03

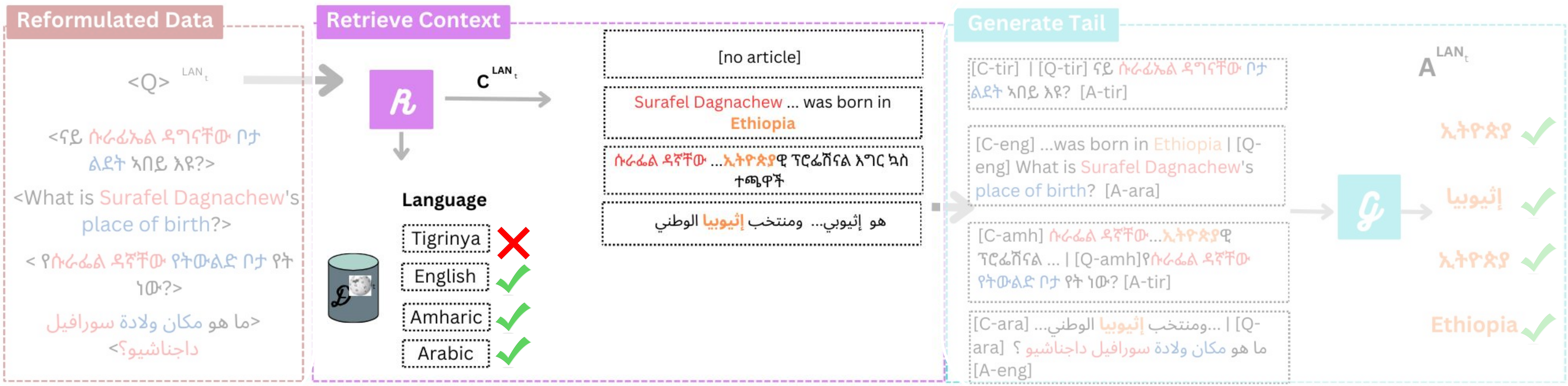
Table 2: Percentage of the head and tail entities in each of the target language KGs with textual representations in each of the transfer languages.

Methods



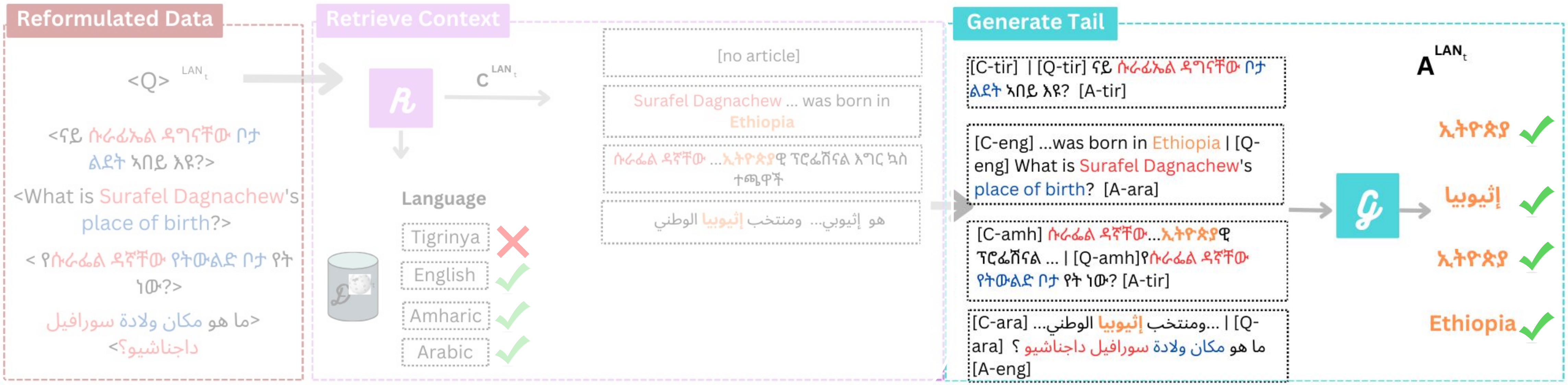
Reformulated Data
<Q> LAN t
<ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?>
<What is Surafel Dagnachew's place of birth?>
< የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?>
<ما هو مكان ولادة سورافيل داجناشيو؟>
Retrieve Context
C LAN t
Language
Tigrinya
English
Amharic
Arabic
[no article]
Surafel Dagnachew ... was born in Ethiopia
Generate Tail
[C-tir] | [Q-tir] ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ? [A-tir]
[C-eng] ...was born in Ethiopia | [Q-eng] What is Surafel Dagnachew's place of birth? [A-ara]
[C-amh] ... | [Q-amh] ... [A-tir]
[C-ara] ... | [Q-ara] ... [A-eng]
A LAN t
ኢትዮጵያ
إثيوبيا
ኢትዮጵያ
Ethiopia

Methods



Reformulated Data
<Q> LAN t
<ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?>
<What is Surafel Dagnachew's place of birth?>
< የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?>
<ما هو مكان ولادة سورافيل داجناشيو؟>
Retrieve Context
C LAN t
Language
Tigrinya
English
Amharic
Arabic
[no article]
Surafel Dagnachew ... was born in Ethiopia
ሱራፌል ዳኛቸው ...ኢትዮጵያዊ ፕሮፌሽናል እግር ኳስ ተጫዋች
هو إثيوبي... ومنتخب إثيوبيا الوطني
Generate Tail
[C-tir] | [Q-tir] ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ? [A-tir]
[C-eng] ...was born in Ethiopia | [Q-eng] What is Surafel Dagnachew's place of birth? [A-ara]
[C-amh] ሱራፌል ዳኛቸው...ኢትዮጵያዊ ፕሮፌሽናል ... | [Q-amh]የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው? [A-tir]
[C-ara] ...ومنتخب إثيوبيا الوطني | [Q-ara] ما هو مكان ولادة سورافيل داجناشيو ؟ [A-eng]
A LAN t
ኢትዮጵያ
إثيوبيا
ኢትዮጵያ
Ethiopia

Methods



Reformulated Data
<Q> LAN
<ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ?>
<What is Surafel Dagnachew's place of birth?>
< የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው?>
<ما هو مكان ولادة سورافيل داجناشيو؟>
Retrieve Context
C LAN
[no article]
Surafel Dagnachew ... was born in Ethiopia
ሱራፌል ዳኛቸው ...ኢትዮጵያዊ ፕሮፌሽናል እግር ኳስ ተጫዋች
هو إثيوبي... ومنتخب إثيوبيا الوطني
Language
Tigrinya
English
Amharic
Arabic
Generate Tail
[C-tir] | [Q-tir] ናይ ሱራፌኤል ዳግናቸው ቦታ ልደት ኣበይ እዩ? [A-tir]
[C-eng] ...was born in Ethiopia | [Q-eng] What is Surafel Dagnachew's place of birth? [A-ara]
[C-amh] ሱራፌል ዳኛቸው...ኢትዮጵያዊ ፕሮፌሽናል ... | [Q-amh]የሱራፌል ዳኛቸው የትውልድ ቦታ የት ነው? [A-tir]
[C-ara] ...ومنتخب إثيوبيا الوطني | [Q-ara] ما هو مكان ولادة سورافيل داجناشيو ؟ [A-eng]
A LAN
ኢትዮጵያ
إثيوبيا
ኢትዮጵያ
Ethiopia

Experiments & Results

Language Models have **very little parametric knowledge** in low-resourced languages.

Language →		Tigrinya	Amharic
Zero-Shot	mT5*	-	0.49
	AfriTeVa †	0.22	0.61
	Aya*	0.67	1.52
	GPT-4	**2.23**	**5.83**
Finetuned	mT5	2.01	23.32
	AfriTeva	**5.13**	**29.15**

Experiments & Results

Language Models trained on **smaller number of related languages perform better***.

Language →		Tigrinya	Amharic
Zero-Shot	mT5*	-	0.49
	AfriTeVa †	0.22	0.61
	Aya*	0.67	1.52
	GPT-4	**2.23**	**5.83**
Finetuned	mT5	2.01	23.32
	AfriTeva	**5.13**	**29.15**

The Less the Merrier? Investigating Language Representation in Multilingual Models (Nigatu et al., Findings 2023)

Experiments & Results

Language Models trained on **smaller number of related languages perform better***.

Language →		Tigrinya	Amharic
Zero-Shot	mT5*	-	0.49
	AfriTeVa †	0.22	0.61
	Aya*	0.67	1.52
	GPT-4	**2.23**	**5.83**
Finetuned	mT5	2.01	23.32
	AfriTeva	**5.13**	**29.15**

*in line with prior work [The Less the Merrier? Investigating Language Representation in Multilingual Models] (Nigatu et al., Findings 2023)

Experiments & Results

mRAKL outperforms prior work for low-resourced language context.

	Tigrinya KG		Amharic KG	
	H@1	H@10	H@1	H@10
KGT5-No-Context	6.91	28.57	32.58	52.57
KGT5-Description	5.8	23.44	32.91	43.32
KGT5-One-Hop	4.46	24.33	28.83	48.17
(ours) No-Context	5.13	26.11	29.15	54.81
(ours) Self-Context	**11.83**	**34.59**	**41.37**	**61.87**

Experiments & Results

mRAKL outperforms prior work for low-resourced language context.

	Tigrinya KG		Amharic KG	
	H@1	H@10	H@1	H@10
KGT5-No-Context	6.91	28.57	32.58	52.57
KGT5-Description	5.8	23.44	32.91	43.32
KGT5-One-Hop	4.46	24.33	28.83	48.17
(ours) No-Context	5.13	26.11	29.15	54.81
(ours) Self-Context	**11.83**	**34.59**	**41.37**	**61.87**

	Tigrinya		Amharic	
	% con.	% tail	% con.	% tail
KGT5-Description	49.77	1.78	6.3	0.71
KGT5-One-Hop	48.83	0.89	25.77	1.65

Experiments & Results

mRAKL outperforms prior work for low-resourced language context.

	Tigrinya KG		Amharic KG	
	H@1	H@10	H@1	H@10
KGT5-No-Context	6.91	28.57	32.58	52.57
KGT5-Description	5.8	23.44	32.91	43.32
KGT5-One-Hop	4.46	24.33	28.83	48.17
(ours) No-Context	5.13	26.11	29.15	54.81
(ours) Self-Context	**11.83**	**34.59**	**41.37**	**61.87**

	Tigrinya		Amharic	
	% con.	% tail	% con.	% tail
KGT5-Description	49.77	1.78	6.3	0.71
KGT5-One-Hop	48.83	0.89	25.77	1.65

Experiments & Results

mRAKL allows for implicit cross-lingual link prediction, where **transfer language context helps improve generation performance**.

Query: ሰማያዊ የምን አይነት ነው?
(What is Blue an instance of?)

Tail: ቀለም
(Color)

Context Language	Context	Prediction
Amharic	እና እኔም አልሰራ ካሉ ይቅር እንጂ የምን ክስ የምን ጣጣ ነው ብየ ተከራከርኩ በርካታ መጽሔቶችም የምን ጊዜም ታላቁ አርቲስት በማለት ይገልጹታል	ቀለም ✓
Arabic	انتفض القيصر واقفا للاحتجاج وصاح القيصر في عجب ماذا ماذا هيكل نموذج اللون أزرق أحمر أخضر أمر استخدام علم جمهورية أذربيجان ينص على أن لون علم الدولة دقيق	لون ✓
English	That is what happened in this instance For instance he hosted a dinner party where he dyed all the food **blue** because he claimed there weren t enough blue foods	color ✓

Experiments & Results

mRAKL allows for implicit cross-lingual link prediction, where **transfer language context helps improve generation performance**.

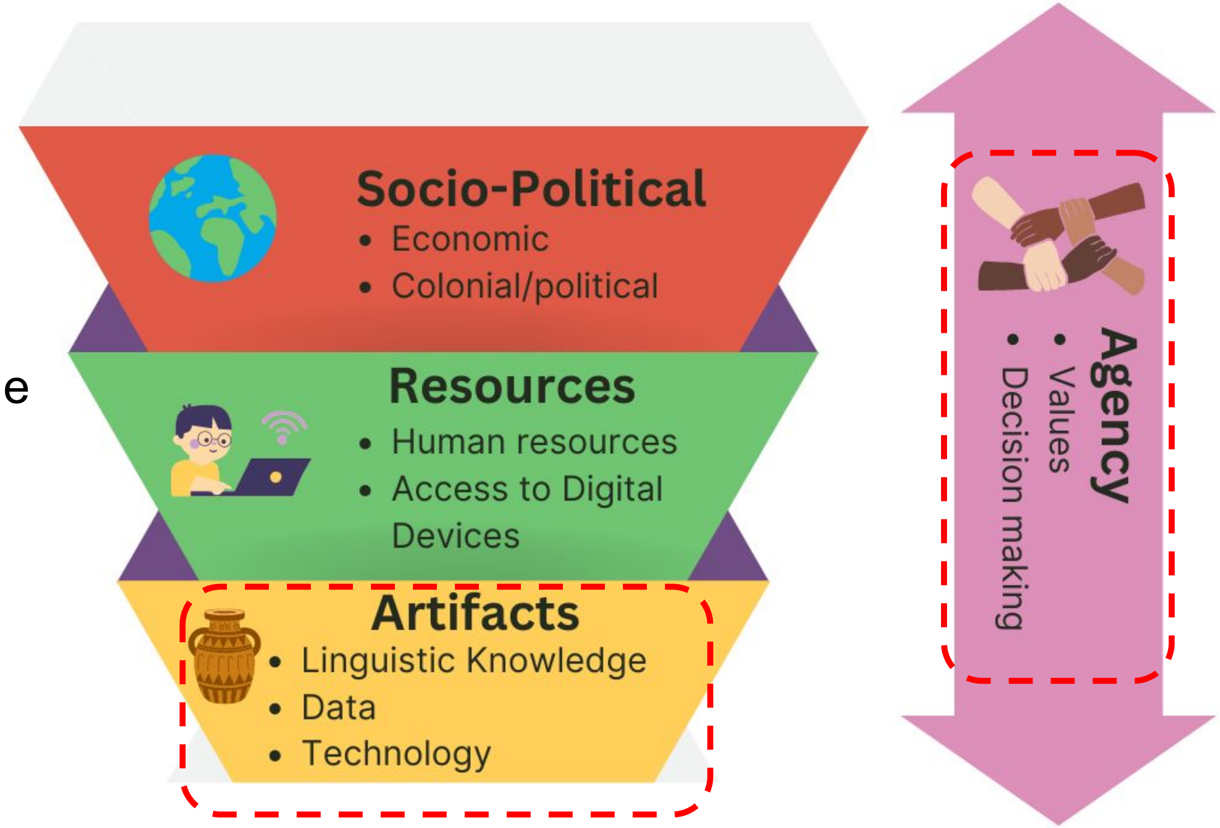
Target lang. →		Tigrinya					Amharic				
Context lang. →		Amh	Ara	Eng	Tir	Avg.	Amh	Ara	Eng	Tir	Avg.
H@1	No-Context	11.64	12.08	14.06	14.06	12.97	30.26	25.34	31.32	8.79	27.81
	LaBSE	12.10	10.29	13.17	13.62	12.30	29.15	24.36	30.69	10.49	27.07
	BM25	**13.70**	**12.53**	**15.84**	**16.51**	**14.65**	**32.68**	**27.75**	**32.48**	**10.99**	**29.82**
H@3	No-Context	**22.60**	**21.48**	22.77	22.32	22.29	38.79	33.81	40.11	**17.68**	36.43
	LaBSE	21.19	18.12	20.76	20.38	22.53	38.35	32.97	39.25	17.26	35.74
	BM25	21.23	21.25	**23.88**	**23.88**	**22.57**	**40.49**	**35.44**	**45.58**	17.09	**37.57**
H@10	No-Context	**39.72**	36.91	38.83	38.16	38.40	48.69	43.38	**50.36**	29.78	45.41
	LaBSE	39.50	36.02	36.38	37.95	37.45	46.91	42.02	47.88	**30.79**	44.77
	BM25	37.21	**37.58**	**39.73**	**39.73**	**38.57**	**48.88**	**44.06**	49.20	29.95	**46.38**

Our Contributions

- **Dataset**: QA and KG
  - We started with what is available in the target languages.
- **Models**: Retriever, Generator
- **Technology**: mRAKL

Our Contributions

- **Dataset**: QA and KG
  - We started with what is available in the target languages.
- **Models**: Retriever, Generator
- **Technology**: mRAKL



Future Work

- Applying this to QA
- Adding more transfer languages
- Adding more unstructured data

Past and Future, Future Work

(Past) Why don’t we have data on Wikipedia for low-resourced languages?

- Hellina Hailu Nigatu, John Canny, Sarah Chasins. (2023). "A Need Finding Study with Low-Resourced Language Content Creators." Proceedings of 4th ACM African Human-Computer Interaction Conference (AfriCHI 2023)
- Hellina Hailu Nigatu, John Canny, Sarah Chasins. (2024). "Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study" *Proceedings of ACM Conference on Human Factors in Computing Systems (ACM CHI).*

WIKIPEDIA
The Free Encyclopedia

(Future, Future) How can we build Dialect-Aware Language Technologies to support low-resourced Wikipedia article creation?

- Machine Translation
- Speech Recognition
- Audio Archives

Visit my Website!!

hellina_nigatu@berkeley.edu