On the Inter-Linguistic Disparity of Knowledge Graphs: Bridging the Gap between English and non-English Languages

Simone Conia · *UniDive - 2nd General Meeting | WG-{1,3,4}* · Apple & Sapienza University of Rome

UniDive · COST European Cooperation in Science & Technology · Funded by the European Union · FAIR Future Artificial Intelligence Research · Sapienza Università di Roma

Abstract

Recent work in Natural Language Processing and Computer Vision has been using textual information – e.g., entity names and descriptions – available in **knowledge graphs to ground neural models to high-quality structured data**. However, when it comes to **non-English languages**, the quantity and quality of textual information are comparatively scarce.

To address this issue, we introduce the novel task of automatic **Knowledge Graph Completion (KGE)** and perform a thorough investigation on bridging the gap in both the quantity and quality of textual information between English and non-English languages. More specifically, we: i) bring to light the problem of increasing multilingual coverage and precision of entity names and descriptions in Wikidata; ii) demonstrate that state-of-the-art methods, namely, Machine Translation (MT), Web Search (WS), and Large Language Models (LLMs), struggle with this task; iii) present **M-NTA, a novel unsupervised approach that combines MT, WS, and LLMs to generate high-quality textual information**; and, iv) study the impact of increasing multilingual coverage and precision of non-English textual information in Entity Linking, Knowledge Graph Completion, and Question Answering.

As part of our effort towards better multilingual knowledge graphs, we also introduce **WikiKGE-10, the first human-curated benchmark to evaluate KGE approaches in 10 languages** across 7 language families.

Multilingual Knowledge Graphs | Overview

Knowledge graphs (KGs) encode our collective understanding of the world in a structured representation.

Before

Wikidata

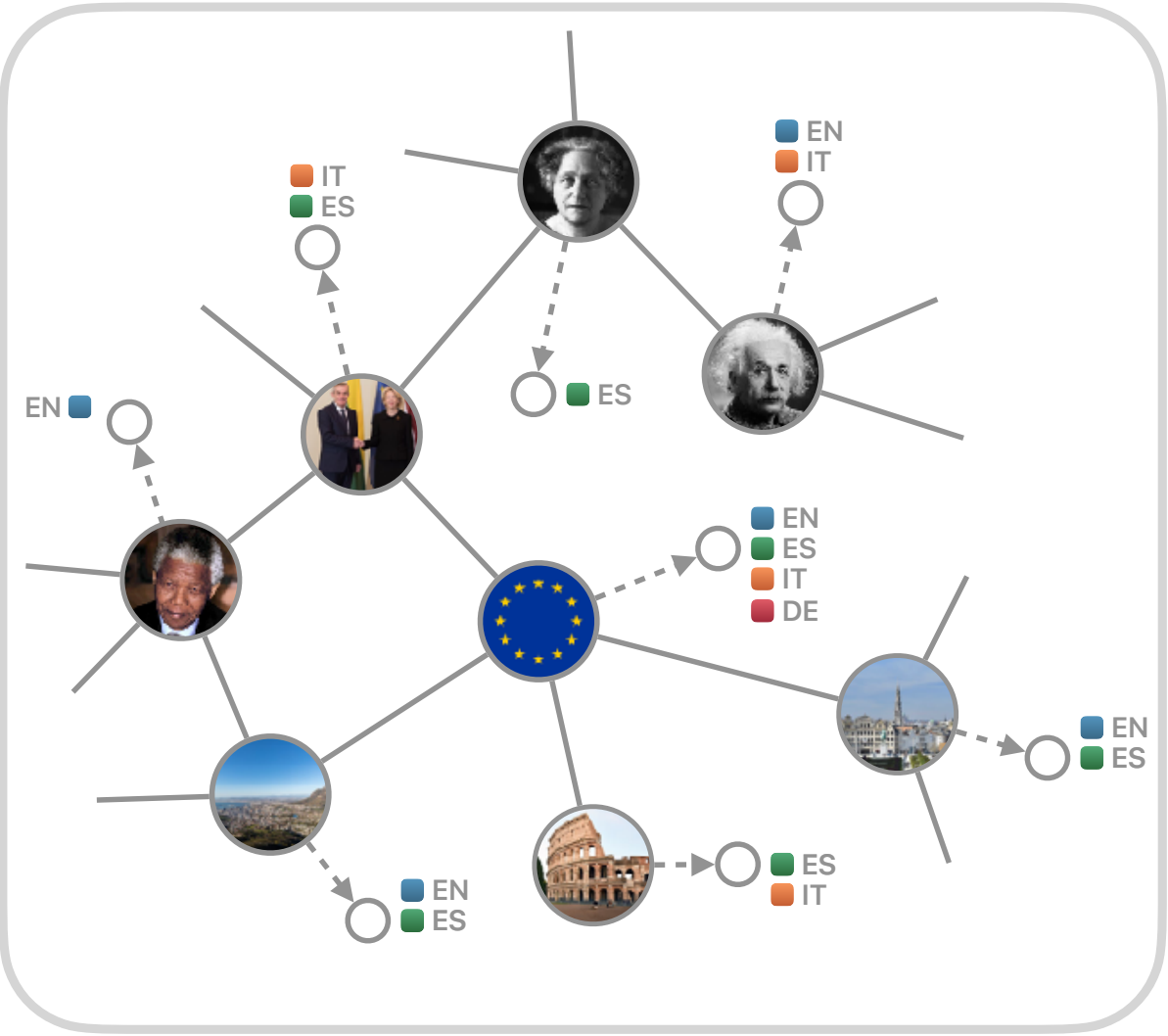


Image of Elsa Einstein: © Bundesarchiv Bild 1929
Image of politicians: © Ernests Dinka 2016
Image of Nelson Mandela: © John Mathew Smith 2001
Image of Albert Einstein: © Orren Jack Turner 1947

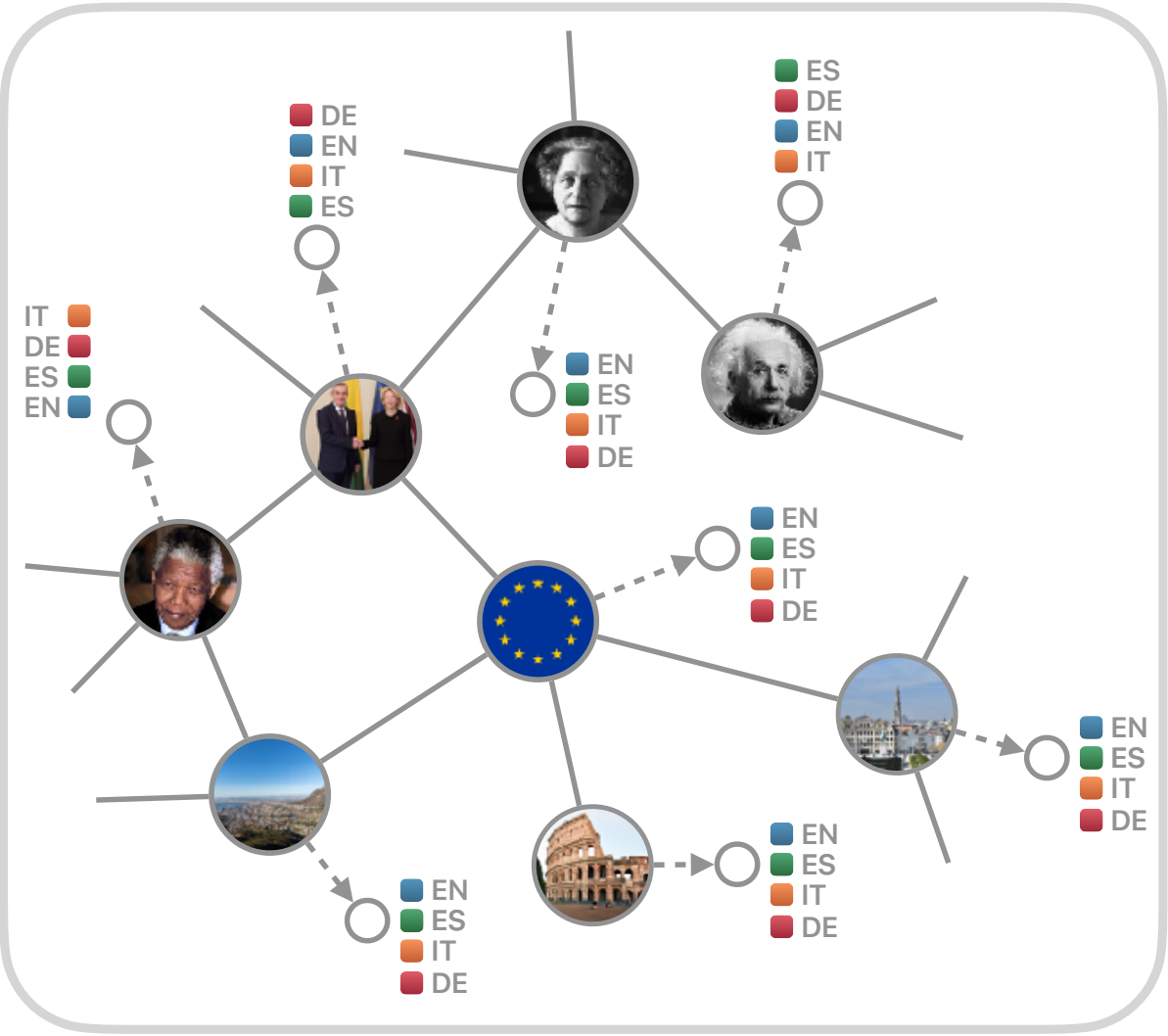
How close can we get to language parity?

→

Knowledge Graph Enhancement

After

Multilingually-enriched Wikidata



Textual information in **multilingual** KGs lacks

Coverage | Precision

↓

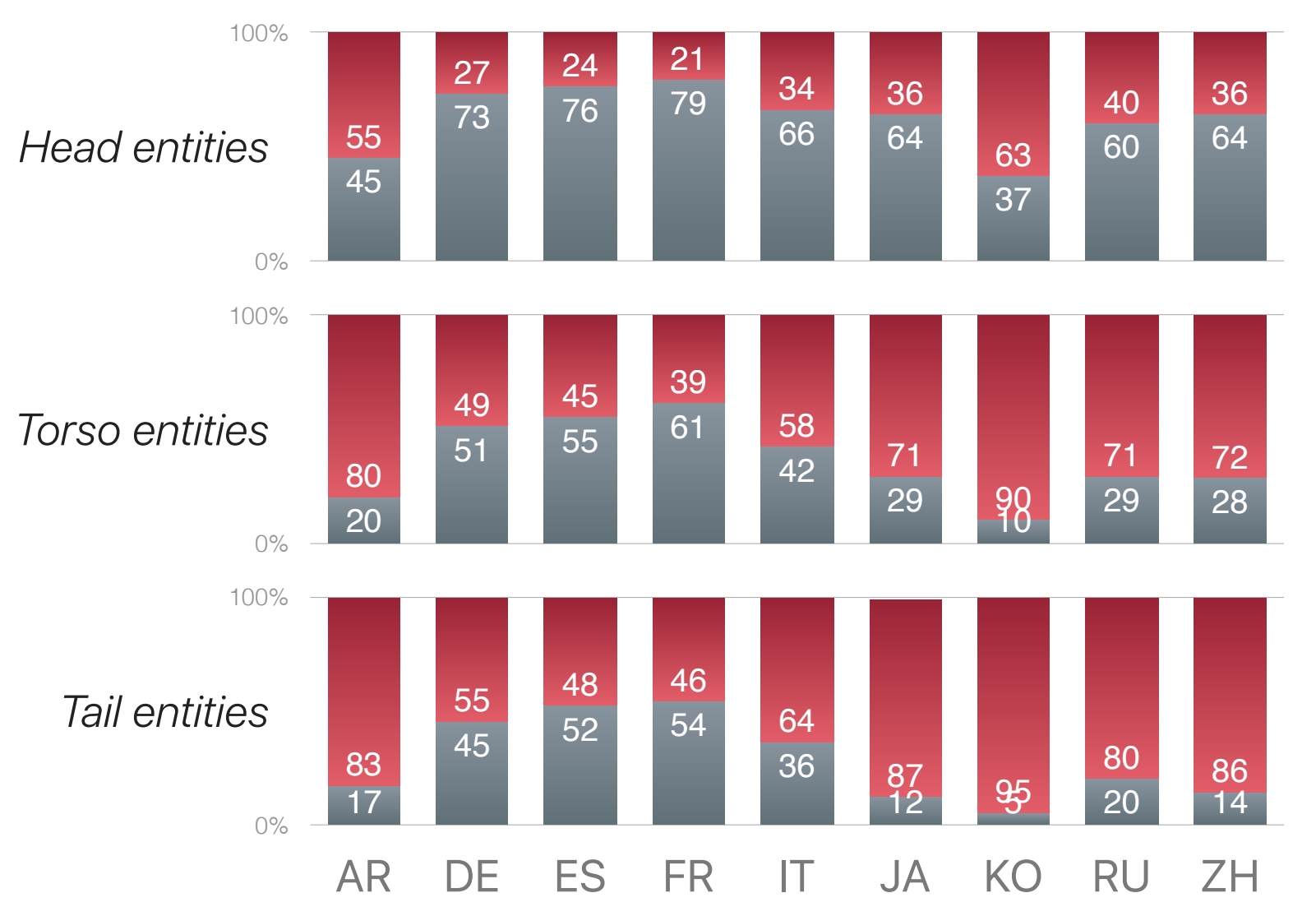
Limited **multilingual** applications

- Language Modeling
- Recommenders
- Question Answering
- Entity Linking
- Image Classification
- Information Retrieval
- Sense Disambiguation
- + Other tasks

Textual Information in Multilingual Knowledge Graphs | Current Limitations

Even for entity names, multilingual coverage is far from 100%.

Wikidata: **Coverage** of non-English entity names vs. English



Many models use KGs out of the box but textual information in KGs is not always precise.

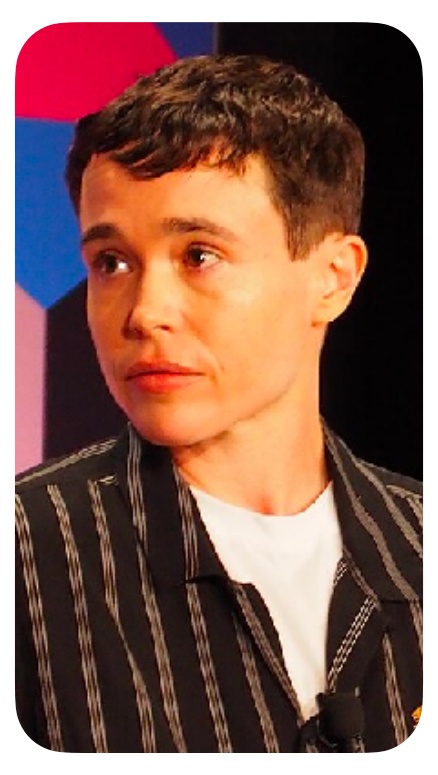
KGs contains human errors

Language	Entity name
English	Olivier Giroud
French	Olivier Giroud
Spanish	**Oliver** Giroud
Japanese	オリヴィエ・ジルー
Chinese	奥利维耶·吉鲁
...	...

Image of Olivier Giroud: © Антон Зайцев 2018

A spelling error in the primary name of a popular entity in Wikidata

KGs contains stale entries



Language	Entity name
English	Elliot Page
Trad. Chinese	艾略特·佩吉 (Elliot Page)
Simp. Chinese	艾莲·佩奇 (Ellen Page)
...	...

Image of Elliot Page: © Frypie 2023

This entity name has not been updated to reflect changes in the real world.

KGs contains under-specific information

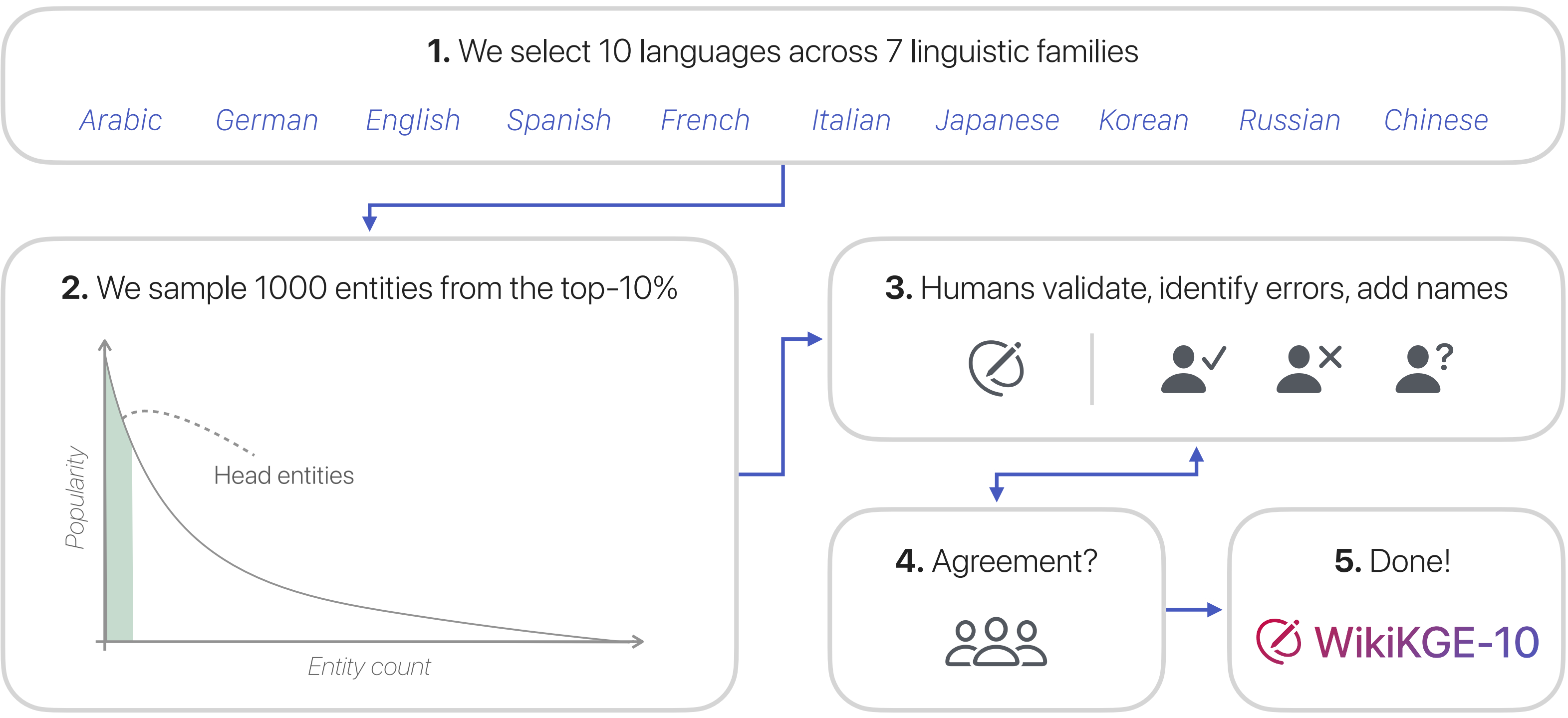
Lang.	Entity description
EN	Japanese composer (1952–2023)
ES	músico japonés
FR	musicien, compositeur, producteur et acteur japonais
...	...

Image of Ryuichi Sakamoto: © Joi Ito 2007

For this entity, different languages have descriptions with different information

WikiKGE-10 | A human-graded benchmark for evaluating KGE systems in 10 languages

How did we create WikiKGE-10?



1. We select 10 languages across 7 linguistic families

Arabic *German* *English* *Spanish* *French* *Italian* *Japanese* *Korean* *Russian* *Chinese*

2. We sample 1000 entities from the top-10%

3. Humans validate, identify errors, add names

4. Agreement?

5. Done!

WikiKGE-10

WikiKGE-10 contains over 35k manually-graded entity names across 10 languages.

	AR	DE	EN	ES	FR	IT	JA	KO	RU	ZH	All
Entities	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	**10,000**
Entity names in WikiKGE-10	4,213	3,498	2,837	4,320	3,548	3,156	2,999	3,874	3,901	4,088	**36,434**
- **Valid names** in Wikidata	2,521	2,336	2,090	2,732	2,330	1,840	2,235	2,136	2,706	2,569	**23,495**
- **Invalid names** in Wikidata	320	491	219	571	530	236	486	329	507	830	**4,663**

Strong agreement
Krippendorf's alpha ≅ **0.94**

Wikidata is incomplete
+35-40% names in WikiKGE-10

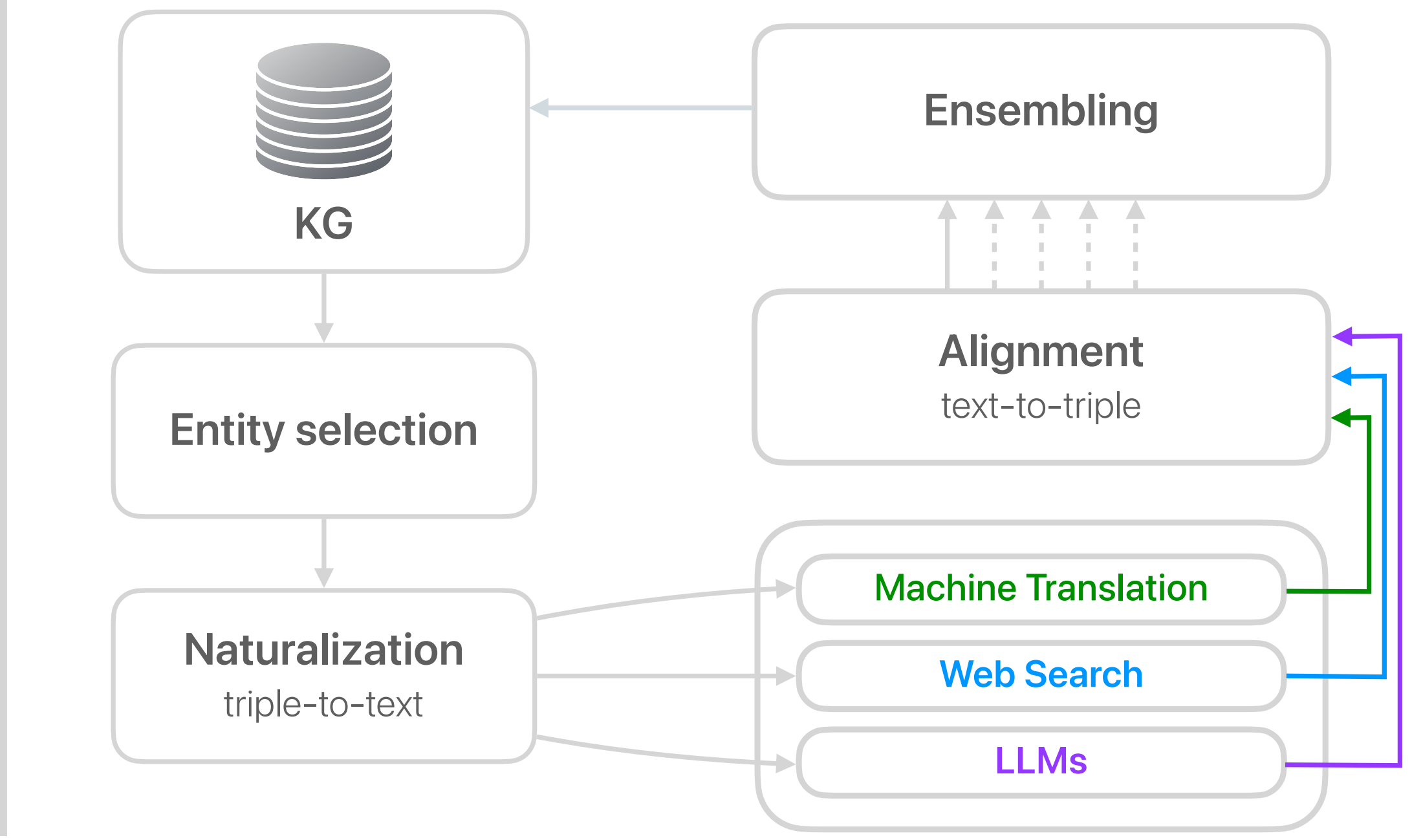
Wikidata is imprecise
20% names are incorrect

M-NTA | Combining MT, WS, and LLMs for KGE

M-NTA leverages the complementary knowledge across locales and tools to provide accurate predictions.

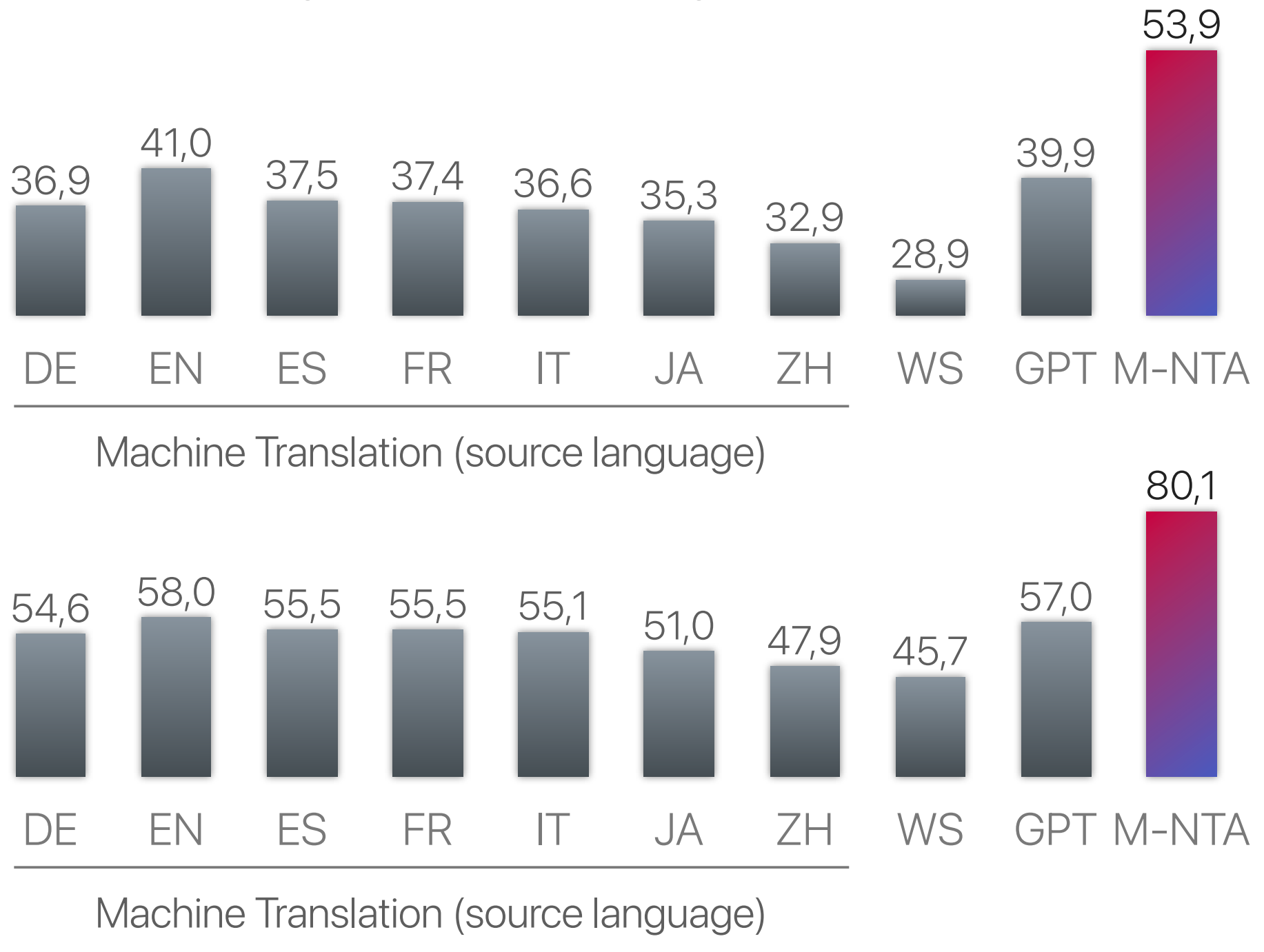
M-NTA

For entity names
& entity descriptions



M-NTA can successfully combine information from multiple tools, sources, and languages.

Avg. F1 scores on coverage and precision



Main takeaway

Generating a fact from **multiple sources and languages may offer complementary pieces of information**, which provide varying views on our world knowledge.

KGE | Impact on downstream tasks

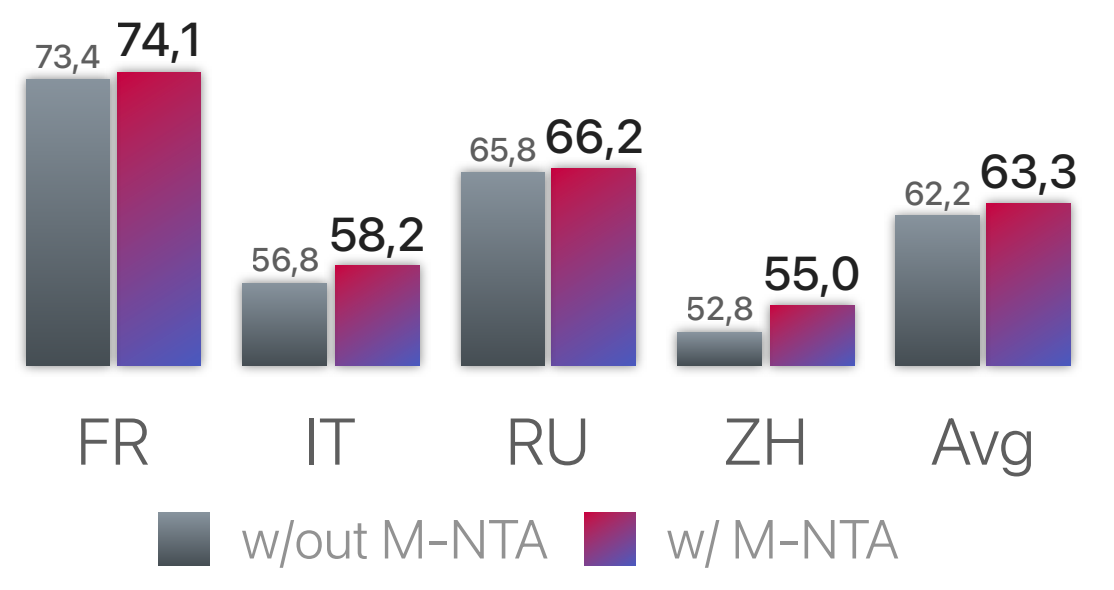
Multilingual **Entity Linking**

"El comandante **Armstrong** fue el primer ser humano que pisó la superficie del satélite terrestre el 21 de julio de 1969 a las 2:56 (hora internacional UTC) […]."

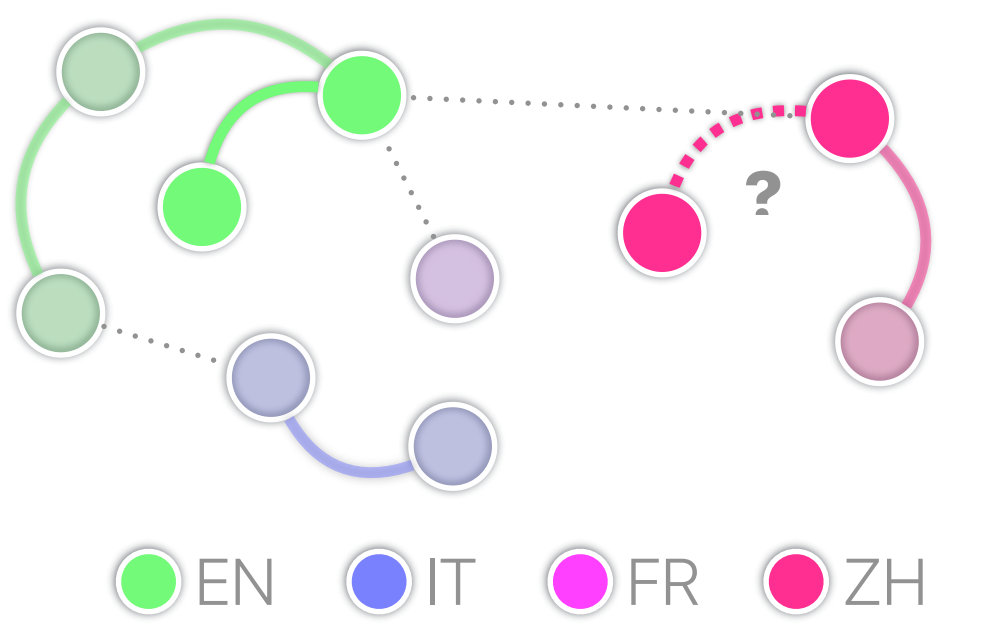
Neil Armstrong
American astronaut
Image of Neil Armstrong: © NASA 1969

Edwin Howard Armstrong
American electrical engineer
Image of Edwin Howard Armstrong: © Florian Bachrach 1950

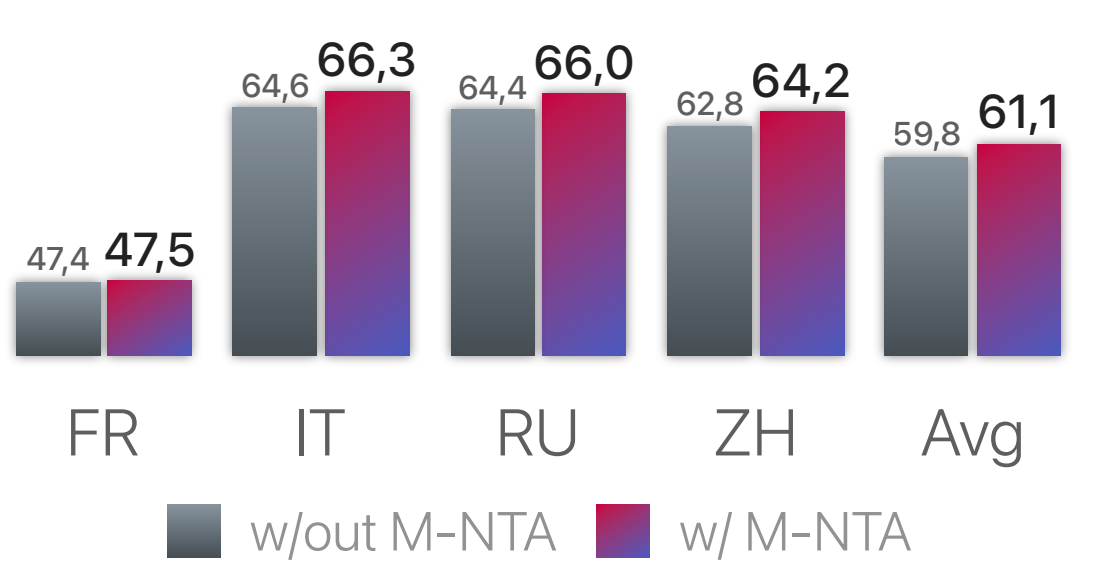
mGENRE on Wikinews-7



Multilingual **KG Completion**



AlignKGC on DBP-5L

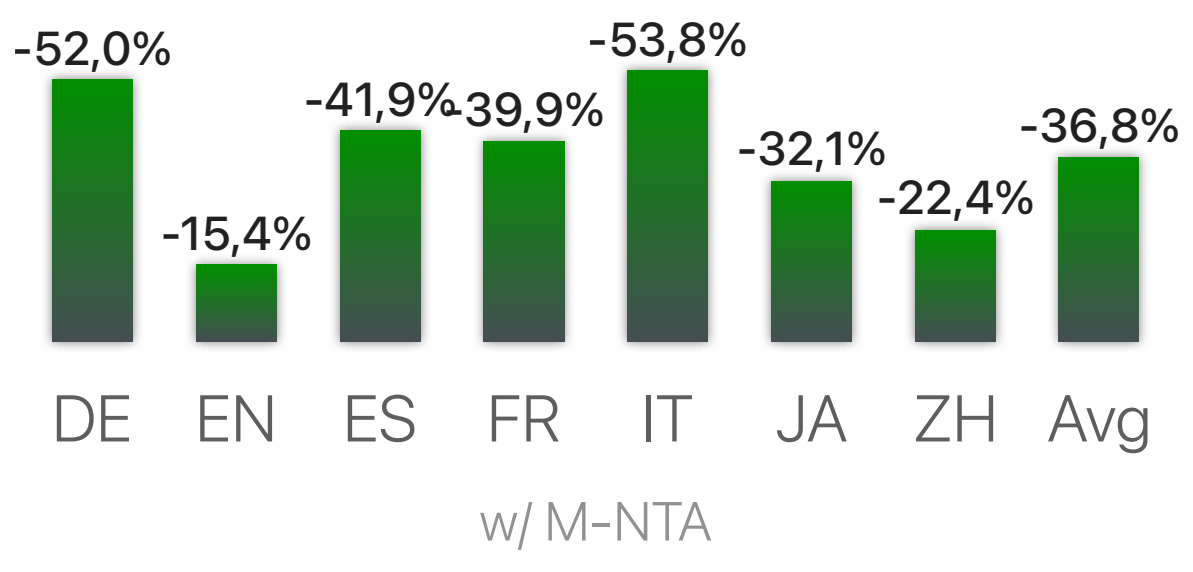


Multilingual **KGQA**

Q: *Quanto è alto Barack Obama?*
A: *1,87 metri*

Q: 영화 식스티 세컨즈의 배우는
A: 안젤리나 졸리, 윌리엄 리 스콧, ...

Unanswerable queries in MKQA



Conclusion

WikiKGE-10 is available at:

https://github.com/apple/ml-kge

Simone Conia
simone.conia@uniroma1.it

Special thanks to

Min Li
min_li6@apple.com

Daniel Lee
daniel_lee1@ucalgary.ca

Umar Farooq Minhas
ufminhas@apple.com

Ihab Ilyas
iilyas@apple.com

Yunyao Li
yunyaol@adobe.com