Cross-Lingual Entity Linking for Low-Resource Languages

Shruti Rijhwani

SIGTYP Online Lecture Series

June 25, 2021

Carnegie Mellon University
Language Technologies Institute

What is Entity Linking?

Associating named entities from natural language with entries in a structured knowledge base.

What is Entity Linking?

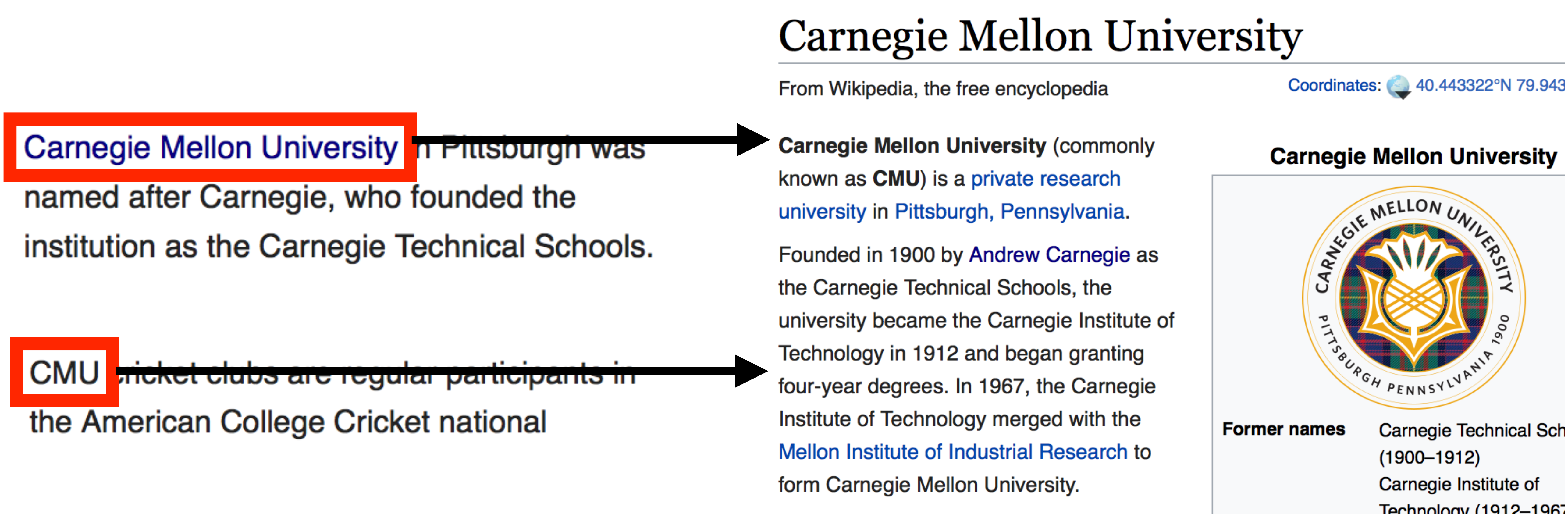
Associating named entities from natural language with entries in a structured knowledge base.

Carnegie Mellon University in Pittsburgh was named after Carnegie, who founded the institution as the Carnegie Technical Schools.

CMU cricket clubs are regular participants in the American College Cricket national

What is Entity Linking?

Associating named entities from natural language with entries in a structured knowledge base.



Why is Entity Linking Useful?

"CMU"

Carnegie Mellon University

Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988)
Type	Private research university
Established	1900 by Andrew Carnegie 1967 (merger with Mellon Institute)
Academic affiliations	AAU, Space-grant
Endowment	$2.67 billion (2020)[1]
President	Farnam Jahanian
Provost	James Garrett
Academic staff	1,483 (March 2020)[2]
Students	14,799 (Fall 2019)[3]

Knowledge bases contain structured information about entities.

Cross-lingual Entity Linking

Azerbaijani

1990-2000-ci illərdə Karnegi Mellon Universiteti ABŞ-ın elit universitetlərindən ibarət olan "Top 25" reytinqində əsas yerləri tutur. 1997-ci ilin

Telugu

మరియు వెయాన్ హాల్ నిర్మాణం ముగింపు మధ్య కాలంలో విద్యాలయం కార్నెగీ ఇన్స్టిట్యూట్ ఆఫ్ టెక్నాలజీ నుండి కార్నెగీ మెల్లన్ విశ్వవిద్యాలయంగా మారింది. కృత్రిమ మేధస్సు, వాణిజ్యం, రోబోటిక్స్

Portuguese

graduação em ciência da computação da Universidade Carnegie Mellon em primeiro

Cross-lingual Entity Linking

Azerbaijani

1990-2000-ci illərdə Karnegi Mellon Universiteti ABŞ-ın elit universitetlərindən ibarət olan "Top 25" reytinqində əsas yerləri tutur. 1997-ci ilin

Telugu

మరియు వెయాన్ హాల్ నిర్మాణం ముగింపు మధ్య కాలంలో విద్యాలయం కార్నెగీ ఇన్‌స్టిట్యూట్ ఆఫ్ టెక్నాలజీ నుండి కార్నెగీ మెల్లన్ విశ్వవిద్యాలయంగా మారింది. కృత్రిమ మేధస్సు, వాణిజ్యం, రోబోటిక్స్

Portuguese

graduação em ciência da computação da Universidade Carnegie Mellon em primeiro

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.94

Carnegie Mellon University (commonly known as **CMU**) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

Carnegie Mellon University

Former names	Carnegie Technical Sch (1900–1912) Carnegie Institute of Technology (1912–196

Cross-lingual Entity Linking

Azerbaijani

1990-2000-ci illərdə Karnegi Mellon Universiteti ABŞ-ın elit universitetlərindən ibarət olan "Top 25" reytinqində əsas yerləri tutur. 1997-ci ilin

Telugu

మరియు వెయాన్ హాల్ నిర్మాణం ముగింపు మధ్య కాలంలో విద్యాలయం కార్నెగీ ఇన్‌స్టిట్యూట్ ఆఫ్ టెక్నాలజీ నుండి కార్నెగీ మెల్లన్ విశ్వవిద్యాలయంగా మారింది. కృత్రిమ మేధస్సు, వాణిజ్యం, రోబోటిక్స్

Portuguese

graduação em ciência da computação da Universidade Carnegie Mellon em primeiro

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943

Carnegie Mellon University (commonly known as **CMU**) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

Carnegie Mellon University

Former names	Carnegie Technical Sch (1900–1912) Carnegie Institute of Technology (1912–196

Most prior work uses English knowledge bases.

Entity Linking Pipeline

CMU

Entity Linking Pipeline

$\text{CMU} \xrightarrow{\text{Candidate Retrieval}}$

Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
…

Entity Linking Pipeline

CMU $\xrightarrow{\text{Candidate Retrieval}}$

Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
...

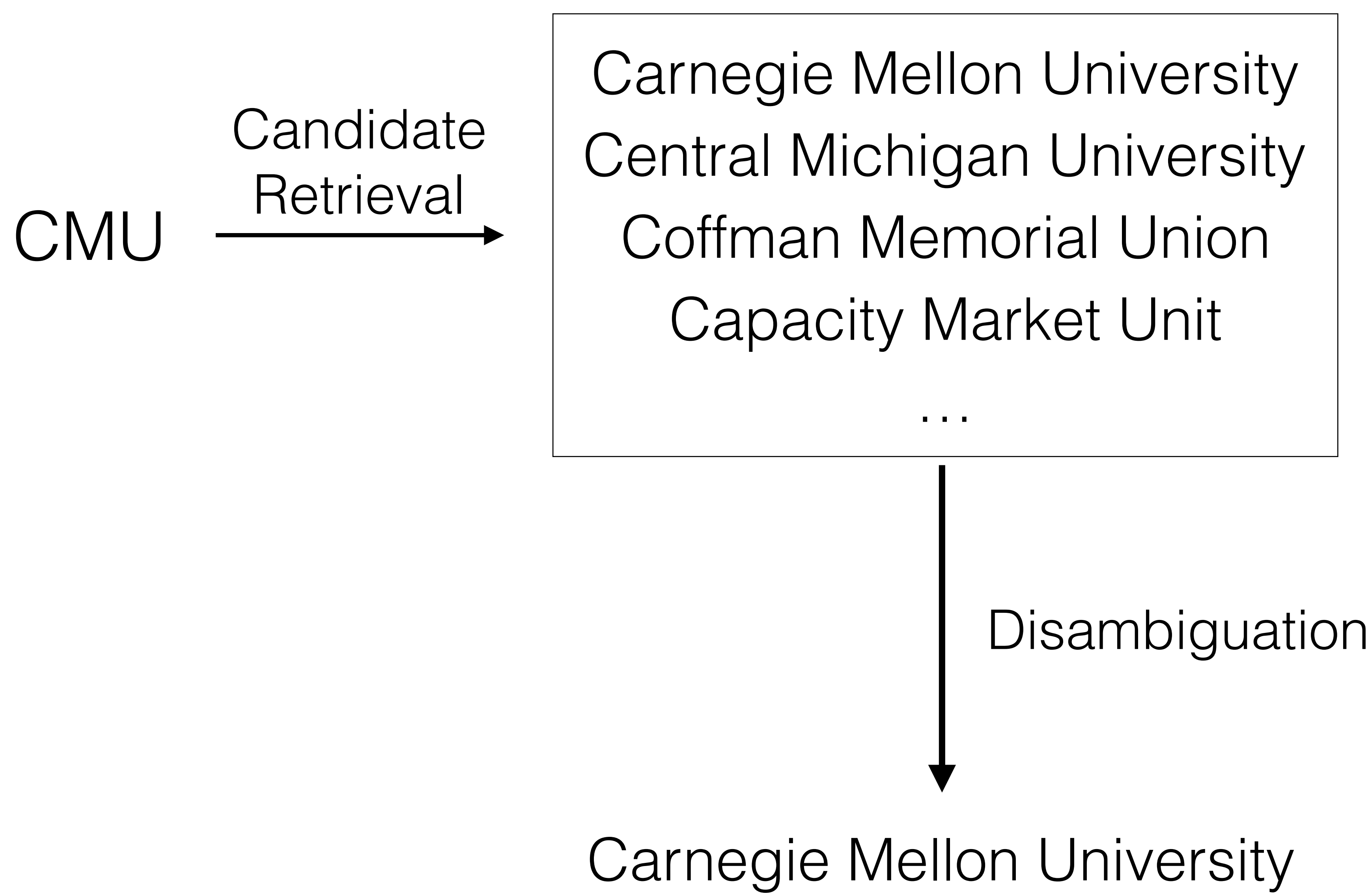
A fast algorithm that **shortlists candidates** from the millions of entries in the KB.

Entity Linking Pipeline

CMU $\xrightarrow{\text{Candidate Retrieval}}$

Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
…

Entity Linking Pipeline



Entity Linking Pipeline

CMU $\xrightarrow{\text{Candidate Retrieval}}$

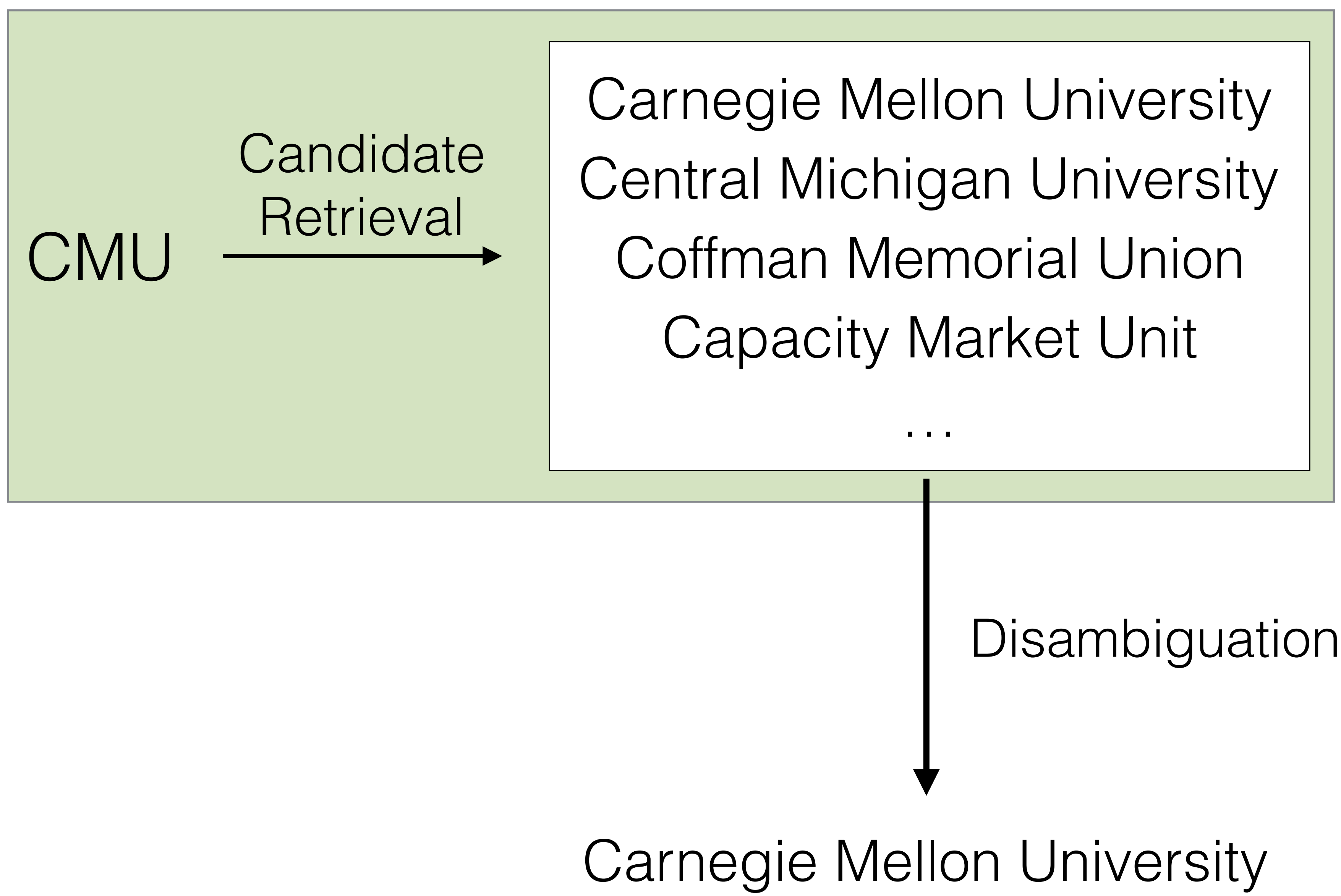
Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
…

Disambiguation

Carnegie Mellon University

A more complex method that **selects the best candidate** based on input context.

Entity Linking Pipeline



Candidate Retrieval

CMU

Candidate Retrieval

CMU

Input named
entity mention
from text

Candidate Retrieval

CMU $\xrightarrow{\text{Candidate Retrieval}}$

Method that **scores each KB entry** based on the input mention

Candidate Retrieval

CMU $\xrightarrow{\text{Candidate Retrieval}}$

Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
…

Top scoring KB entries as selected as candidates

Candidate Retrieval Methods

Candidate Retrieval Methods

- **Most methods use Wikipedia language links** for translating input entities to the KB language.

Candidate Retrieval Methods

- **Most methods use Wikipedia language links** for translating input entities to the KB language.

- Wikipedia exists in many languages and entities are linked between them.

Candidate Retrieval Methods

- **Most methods use Wikipedia language links** for translating input entities to the KB language.

- Wikipedia exists in many languages and entities are linked between them.

Amérika Sarékat → United States

source language (Javanese) *target language (English)*

Candidate Retrieval Methods

- **Most methods use Wikipedia language links** for translating input entities to the KB language.
- Wikipedia exists in many languages and entities are linked between them.

Amérika Sarékat	→	United States
source language (Javanese)		*target language (English)*

- These links are **used to create a dictionary** for translation between the source and target languages.

Low-resource Challenges

Low-resource Challenges

- ~300 languages have Wikipedia

Low-resource Challenges

- ~300 languages have Wikipedia
- Some sets of language links are **very small**

Low-resource Challenges

- ~300 languages have Wikipedia
- Some sets of language links are **very small**
  - Swedish, German, French, Dutch — 1M+ links

Low-resource Challenges

- ~300 languages have Wikipedia
- Some sets of language links are **very small**
  - Swedish, German, French, Dutch — 1M+ links
  - Tigrinya, Sango, Lao — fewer than 1000 links

Low-resource Challenges

- ~300 languages have Wikipedia
- Some sets of language links are **very small**
  - Swedish, German, French, Dutch — 1M+ links
  - Tigrinya, Sango, Lao — fewer than 1000 links
- There are **~7000 living languages** in the world!

Low-resource Challenges

- ~300 languages have Wikipedia
- Some sets of language links are **very small**
    - Swedish, German, French, Dutch — 1M+ links
    - Tigrinya, Sango, Lao — fewer than 1000 links
- There are **~7000 living languages** in the world!
- **How do we retrieve candidates without a high-coverage bilingual dictionary?**

Candidate Retrieval for Low-Resource Languages

Candidate Retrieval for Low-Resource Languages

- A method that **uses no bilingual resources in the source language.**

Candidate Retrieval for Low-Resource Languages

- A method that **uses no bilingual resources in the source language.**
  - Models are trained on **data from high-resource languages**.

Candidate Retrieval for Low-Resource Languages

- A method that **uses no bilingual resources in the source language.**
  - Models are trained on **data from high-resource languages**.
  - **Transferred to low-resource language entities**, with no fine-tuning or joint-training.

Candidate Retrieval for Low-Resource Languages

- A method that **uses no bilingual resources in the source language.**
  - Models are trained on **data from high-resource languages**.
  - **Transferred to low-resource language entities**, with no fine-tuning or joint-training.
- The method can also be used to **improve low-resource named entity recognition.**

Entity Linking Task

Entity Linking Task

CMU

Input named entity **mention** from text

Entity Linking Task

Knowledge base with
millions of **entries**

WIKIPEDIA
The Free Encyclopedia

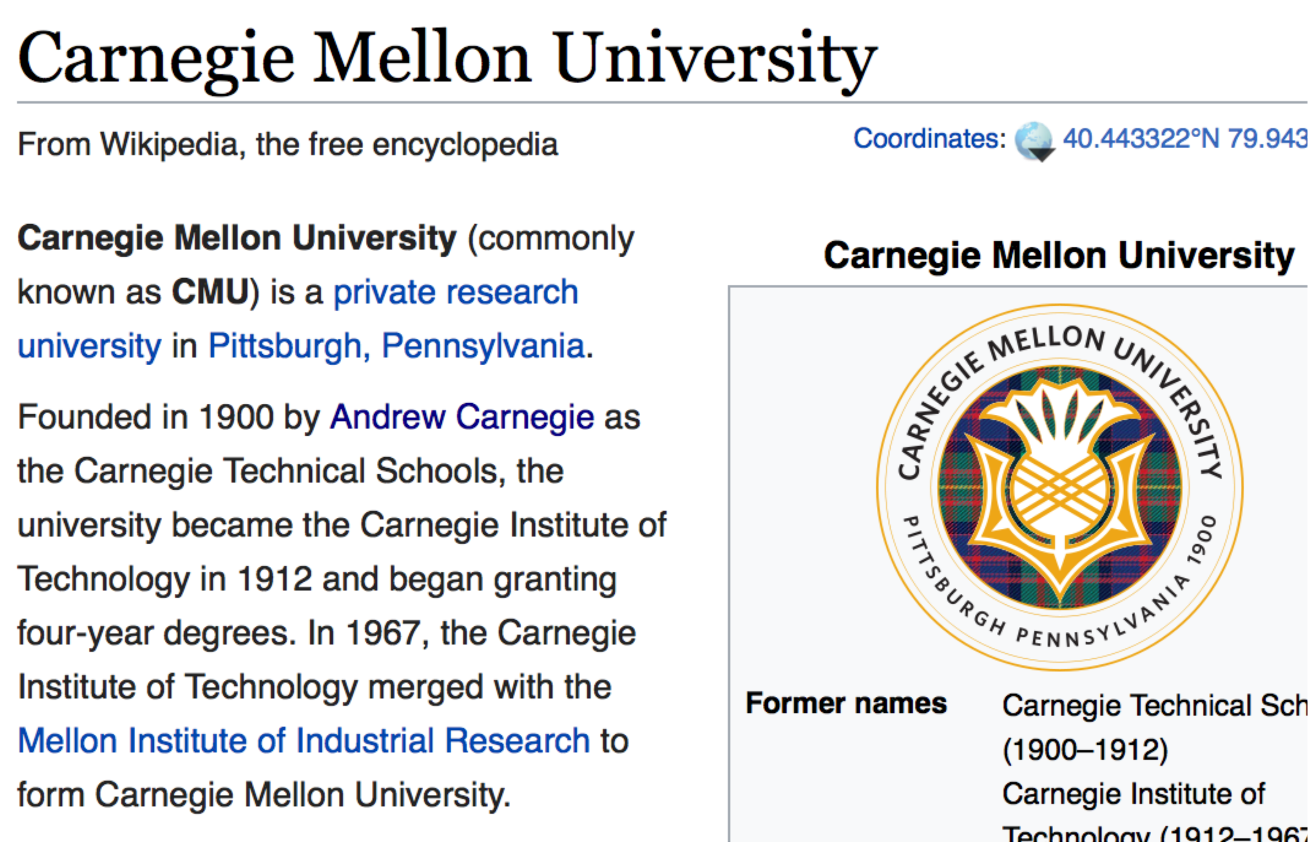
CMU

Input named
entity **mention**
from text

Entity Linking Task

Entity linking **scores the input with respect to each KB entry** to select the most appropriate one

WIKIPEDIA
The Free Encyclopedia


Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943

Carnegie Mellon University (commonly known as **CMU**) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

Carnegie Mellon University

Former names	Carnegie Technical Sch (1900–1912) Carnegie Institute of Technology (1912–196

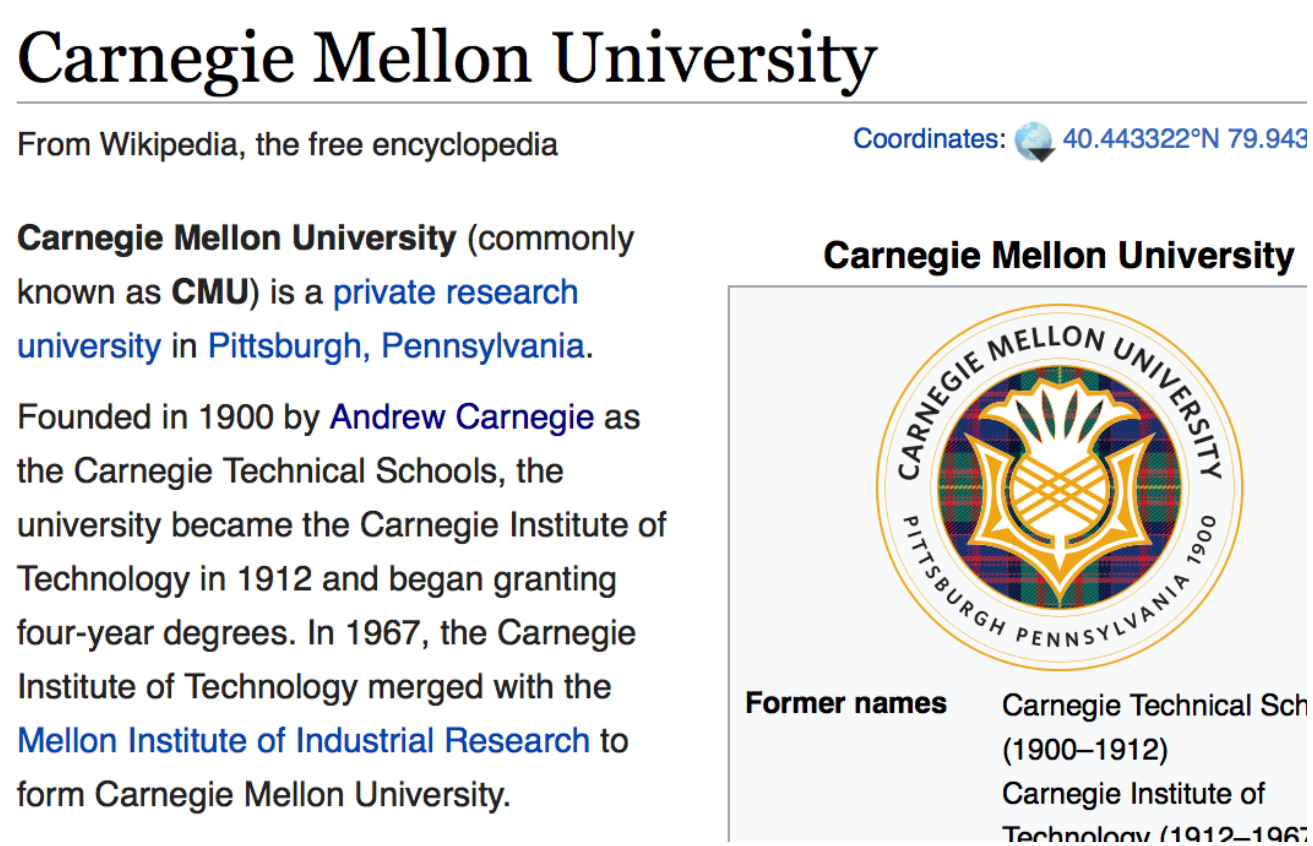
CMU

Input named entity **mention** from text

Entity Linking Task

Entity linking **scores the input with respect to each KB entry** to select the most appropriate one

Wikipedia
The Free Encyclopedia


Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943

Carnegie Mellon University (commonly known as **CMU**) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

Carnegie Mellon University

Former names Carnegie Technical Sch (1900–1912)
Carnegie Institute of

CMU

We focus on cases where the input is from a **low-resource language**

Pivot-Based Entity Linking

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to the low-resource language

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to the low-resource language

Pivoting

Link to closely-related “pivot” language, instead of English

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to the low-resource language

Pivoting

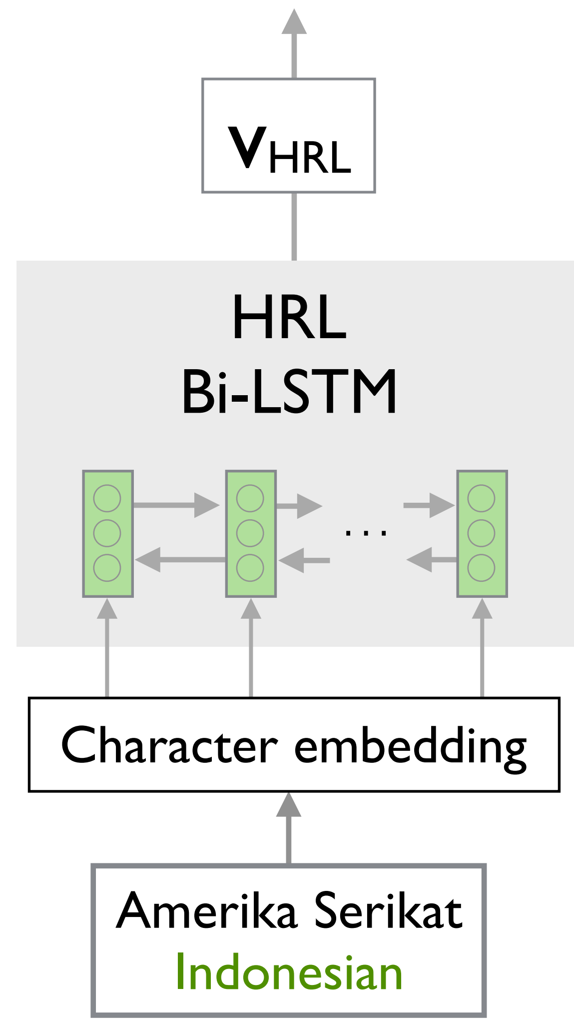
Link to closely-related "pivot" language, instead of English

Amérika Sarékat → Amerika Serikat ⋯⋯ United States

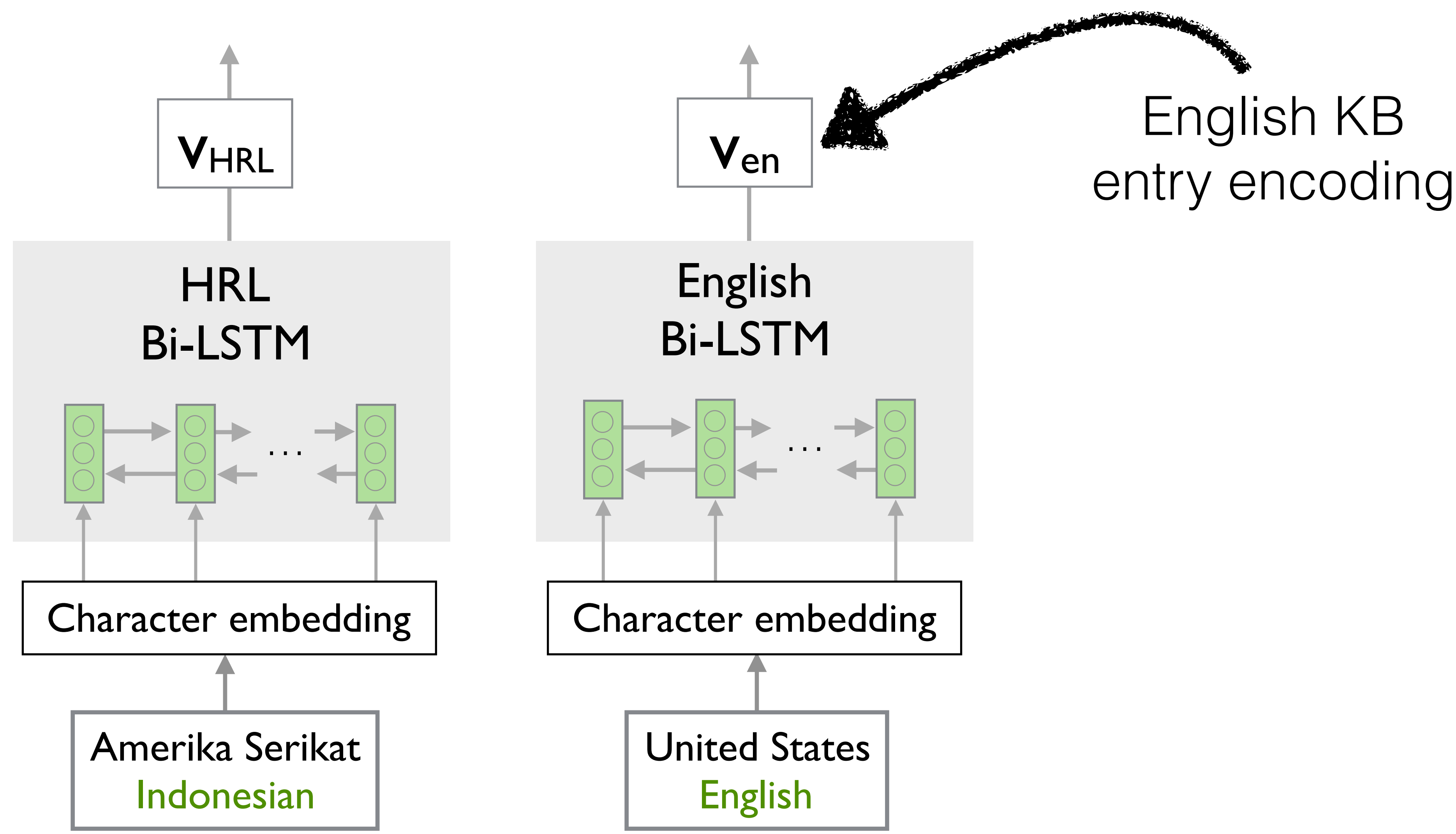
Javanese *Indonesian*

Learning an Entity Linking Model

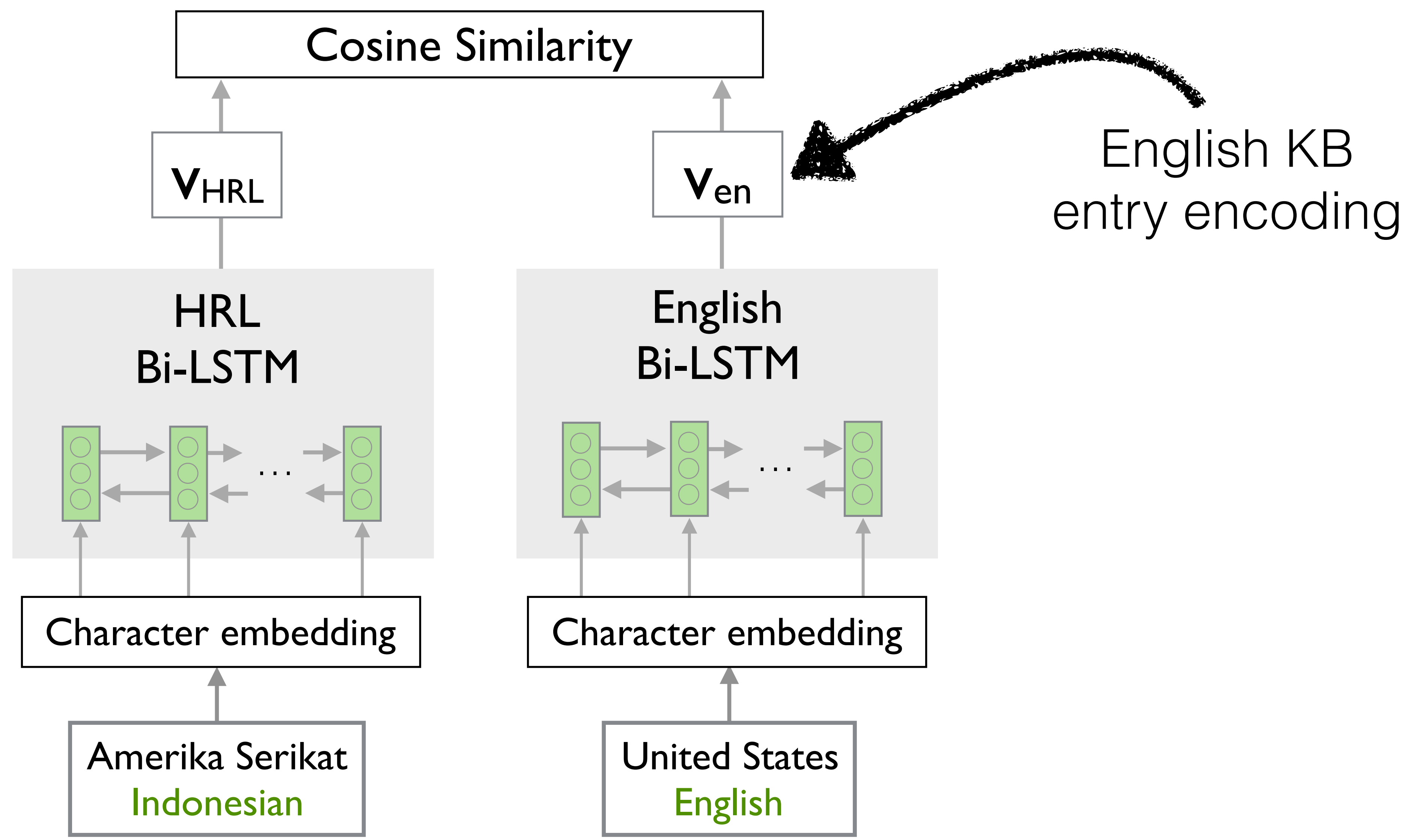
Learning an Entity Linking Model



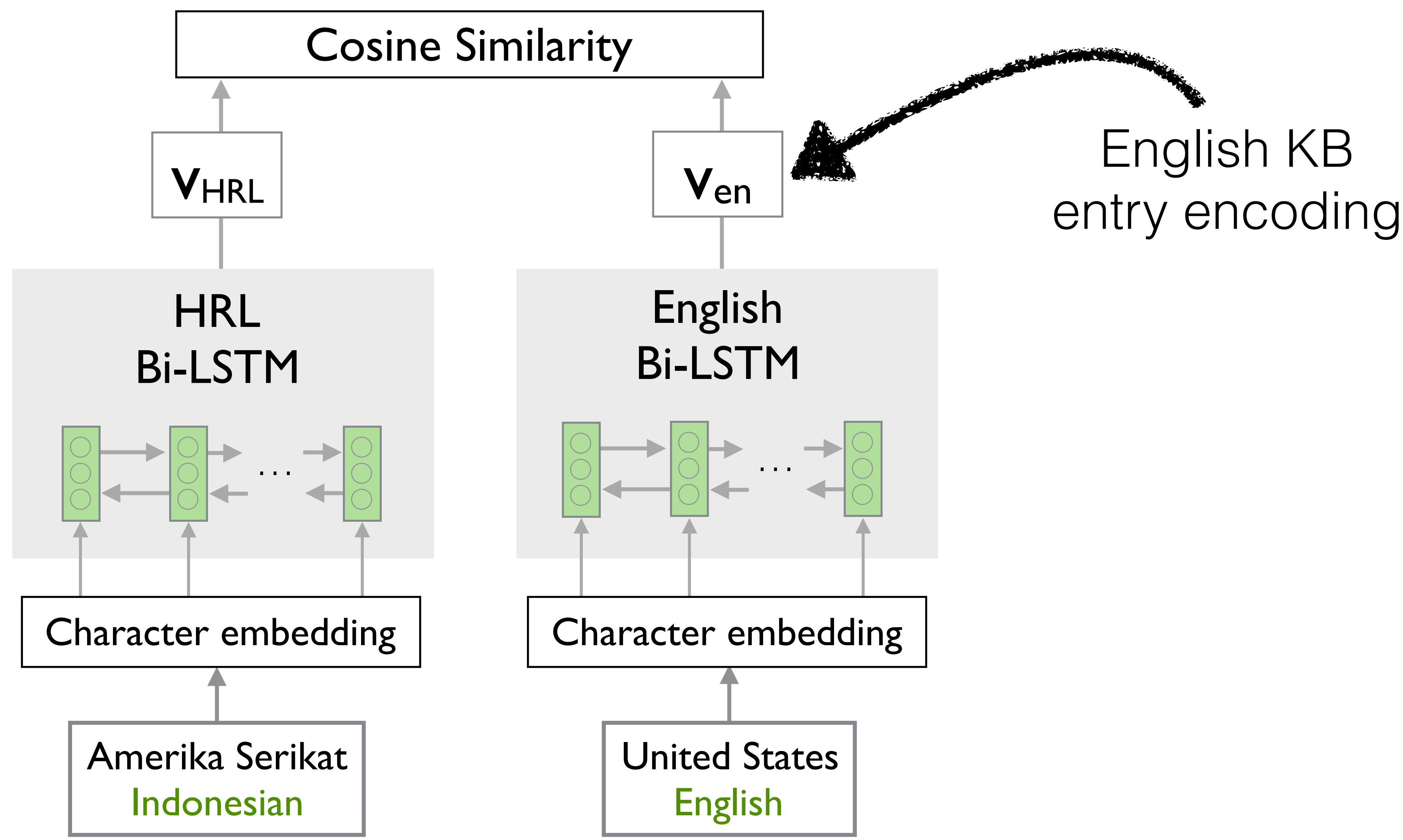
Learning an Entity Linking Model



Learning an Entity Linking Model



Learning an Entity Linking Model



Max-margin training objective forces the difference between scores of a positive training pair and a negative training pair to be at least margin **λ** apart.

$$\max(0, \text{sim}(e_{HRL}, e_{en}) - \text{sim}(e_{HRL}, e^*_{en}) + \lambda)$$

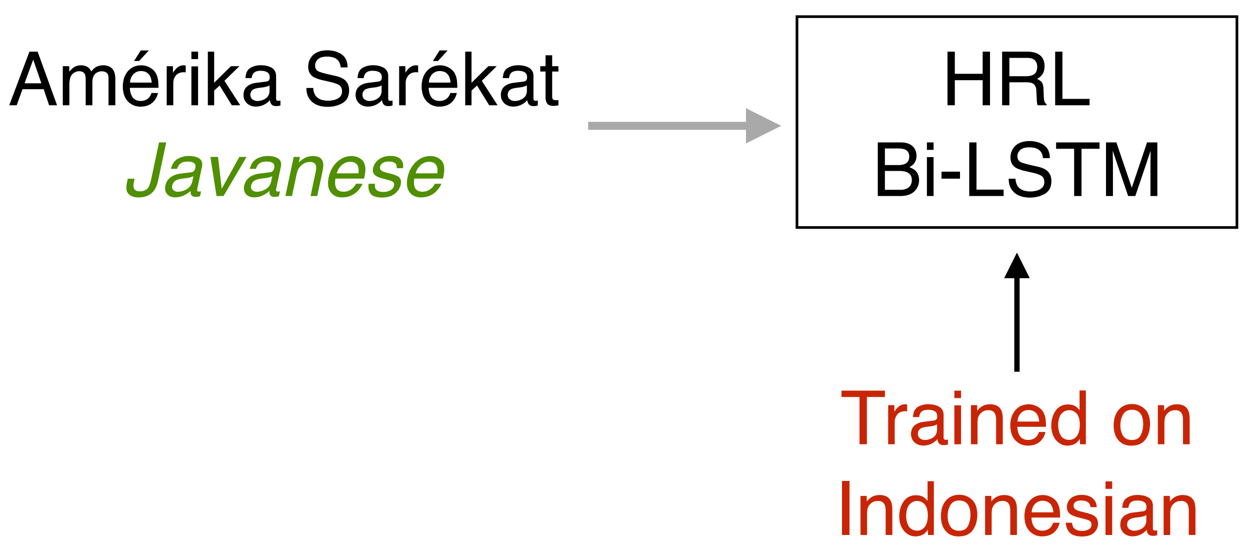
Zero-shot Transfer

Amérika Sarékat
Javanese

Zero-shot Transfer

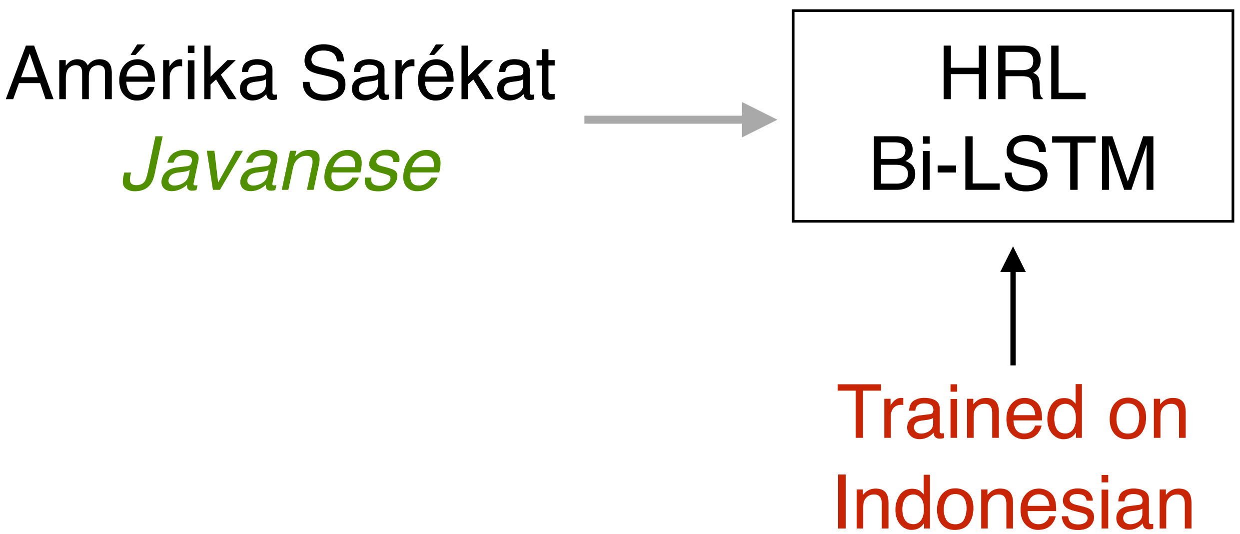
- Encode low-resource entity with character-LSTM **trained on a high-resource language**

Zero-shot Transfer



- Encode low-resource entity with character-LSTM **trained on a high-resource language**

Zero-shot Transfer



- Encode low-resource entity with character-LSTM **trained on a high-resource language**
- If HRL and LRL are **closely-related**, transfer can succeed

Zero-shot Transfer

Amérika Sarékat
Javanese

HRL
Bi-LSTM

Trained on
Indonesian

- Encode low-resource entity with character-LSTM **trained on a high-resource language**
- If HRL and LRL are **closely-related**, transfer can succeed

  Hindi and Marathi

  Thai and Lao

  Indonesian and Javanese

Pivoting with High-Resource

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Indonesian Lang. Link ············ *English Wikipedia*

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Indonesian Lang. Link ············ *English Wikipedia*

Language links between the high-resource language and English (usually a large set)

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Javanese input → *Indonesian Lang. Link* ········ *English Wikipedia*

Cosine similarity low-resource
input and high-resource entity.

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Javanese input → *Indonesian Lang. Link* ┈┈┈ *English Wikipedia*

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Javanese input → *Indonesian Lang. Link* ··········· *English Wikipedia*

Amérika Sarékat
Javanese

m_{LRL} → HRL Bi-LSTM

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Javanese input ⟶ *Indonesian Lang. Link* ········ *English Wikipedia*

Amerika Serikat
Indonesian

e_{HRL} → HRL Bi-LSTM

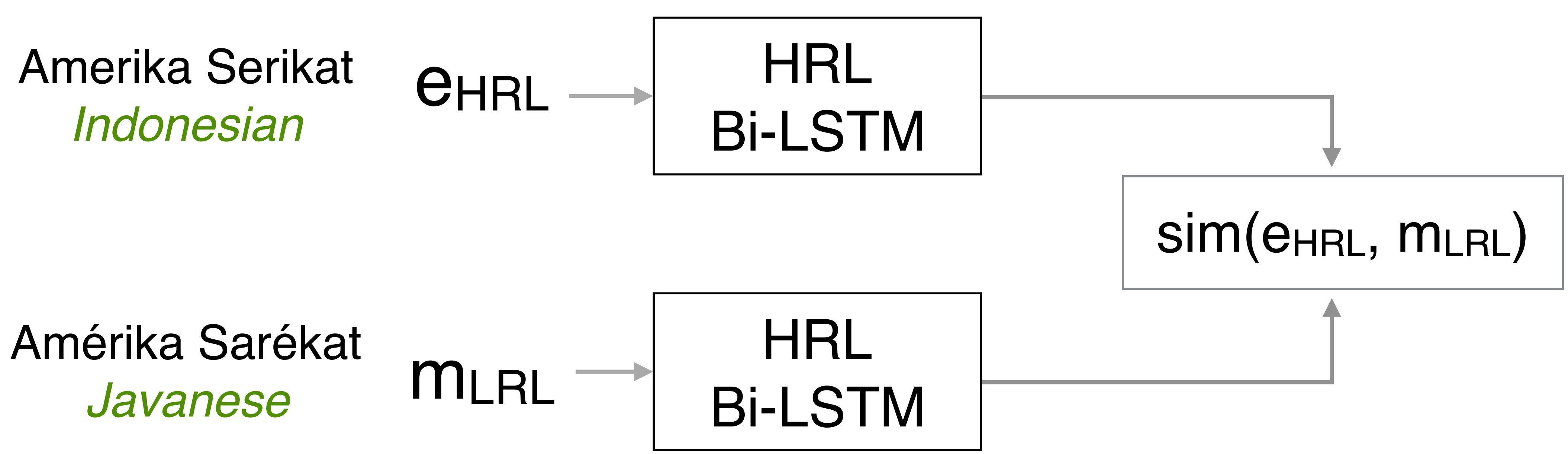
Amérika Sarékat
Javanese

m_{LRL} → HRL Bi-LSTM

Pivoting with High-Resource

- Low-resource entities are likely **more similar to those in a closely-related high-resource language** than to English.
- Calculate **scores with closely-related language entities** instead of directly with English.

Javanese input → *Indonesian Lang. Link* ············ *English Wikipedia*



Pivot-Based Linking Model

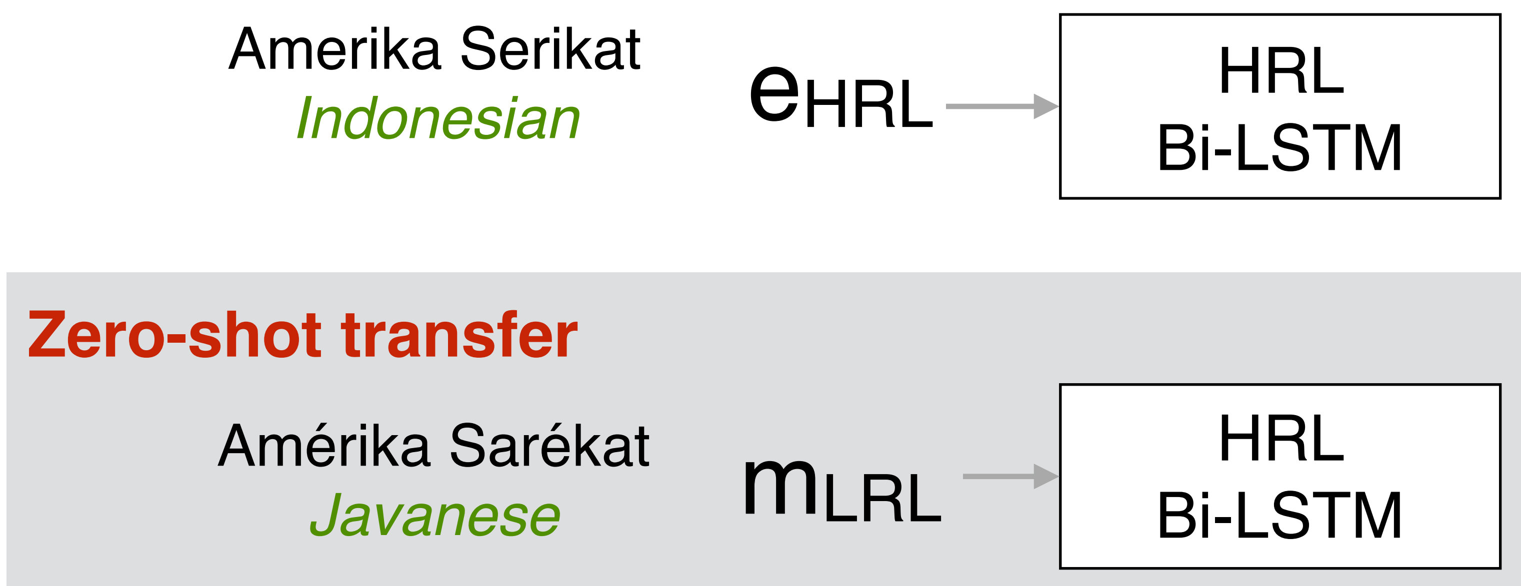
Pivot-Based Linking Model

Amérika Sarékat
Javanese

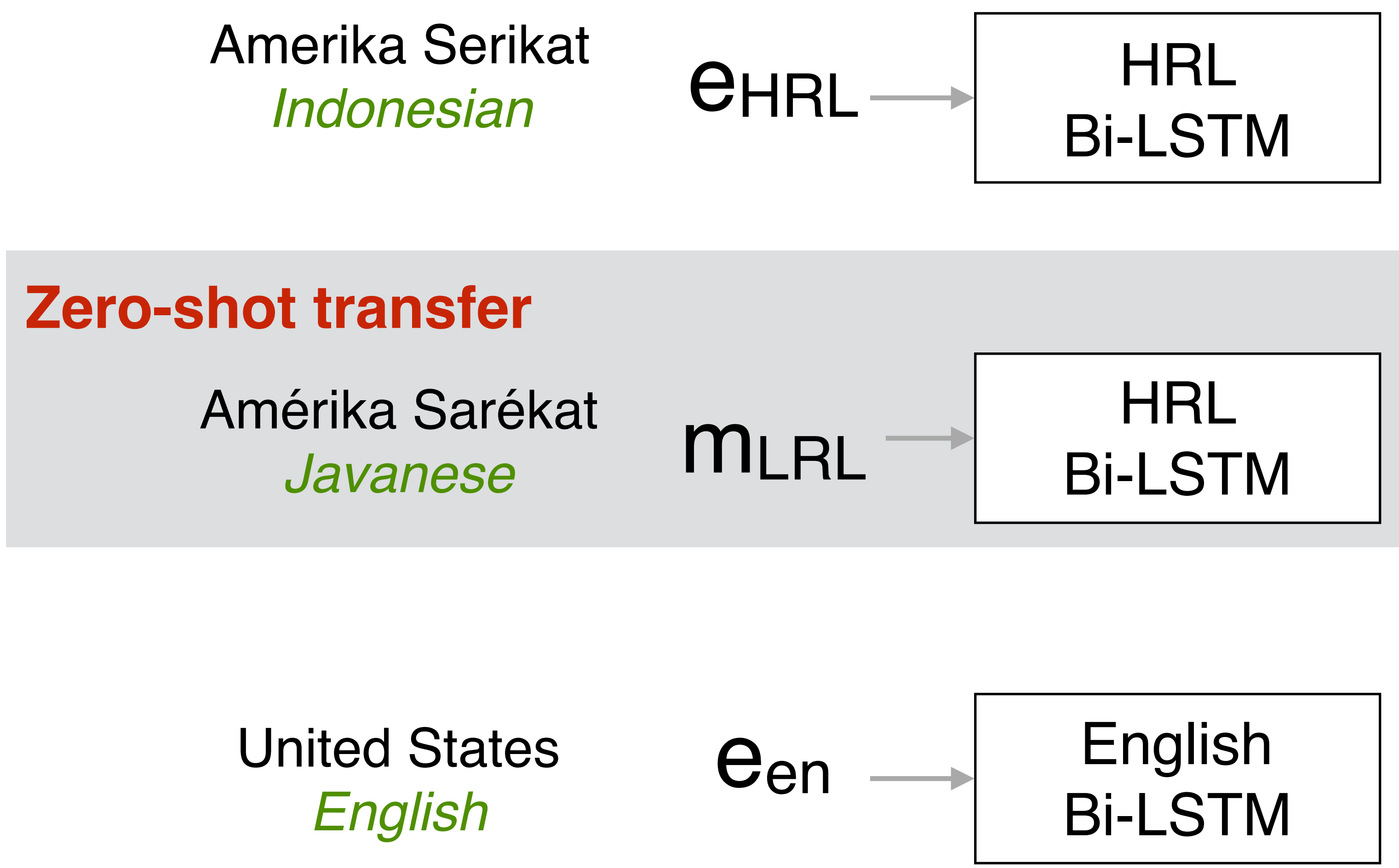
m_{LRL}

Pivot-Based Linking Model

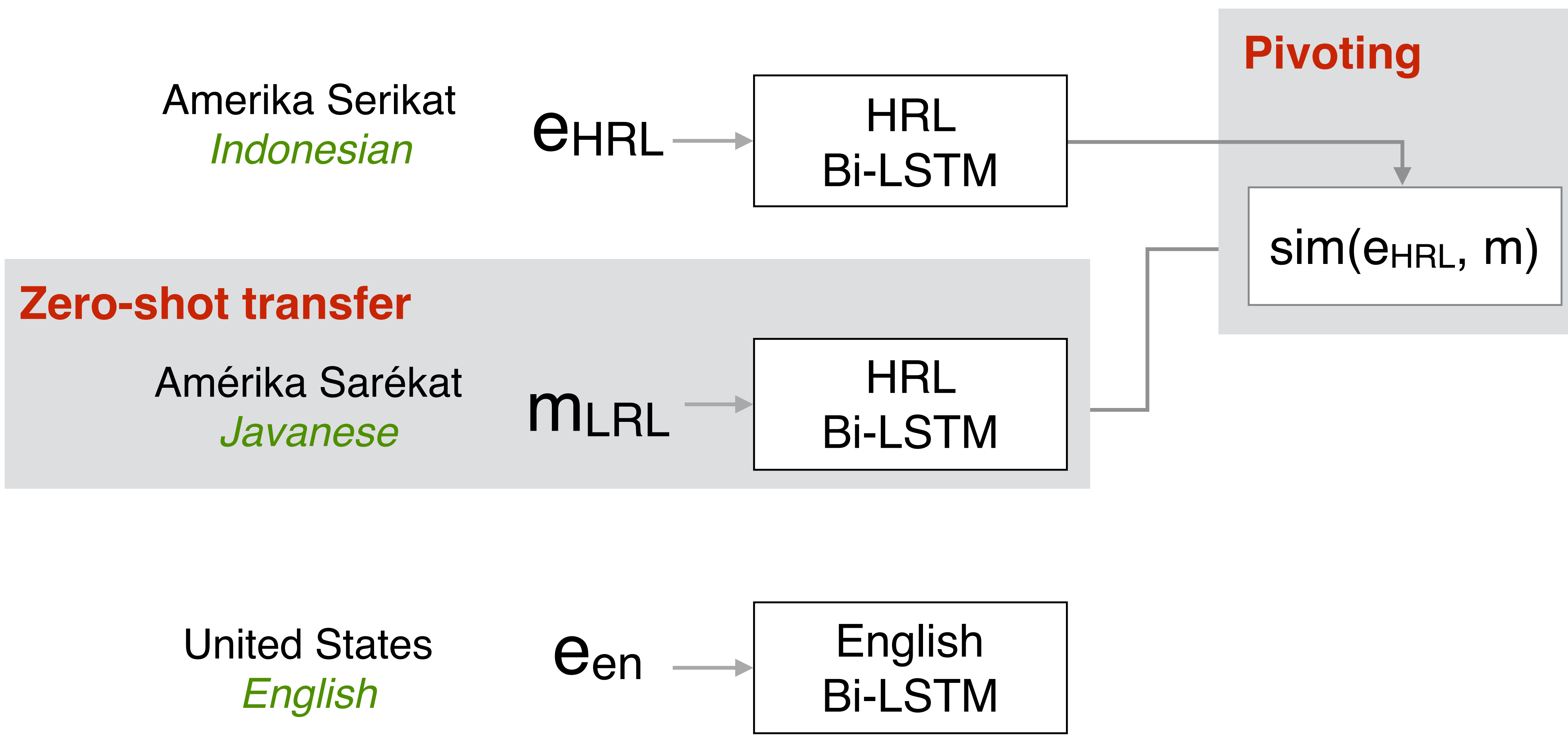
Pivot-Based Linking Model



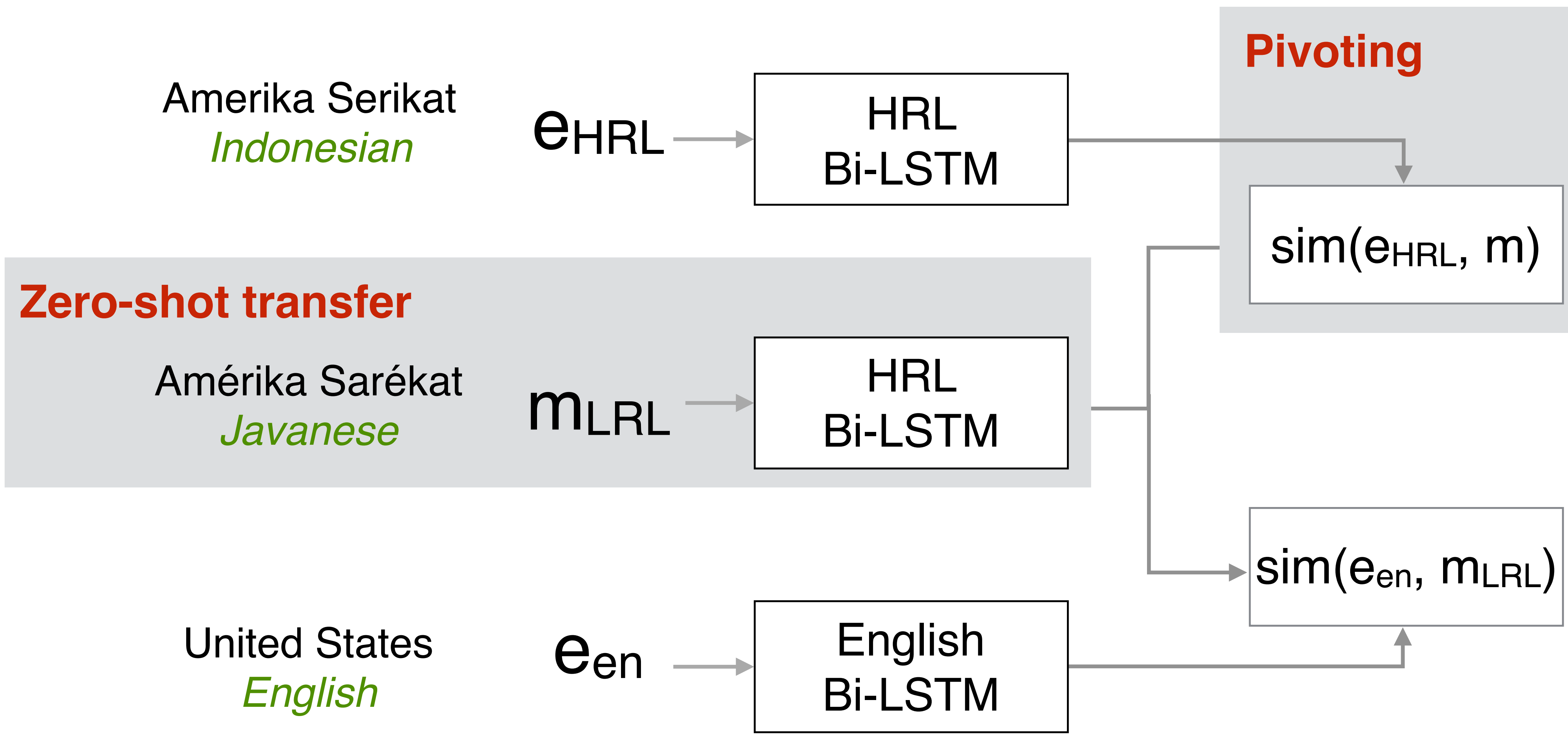
Pivot-Based Linking Model



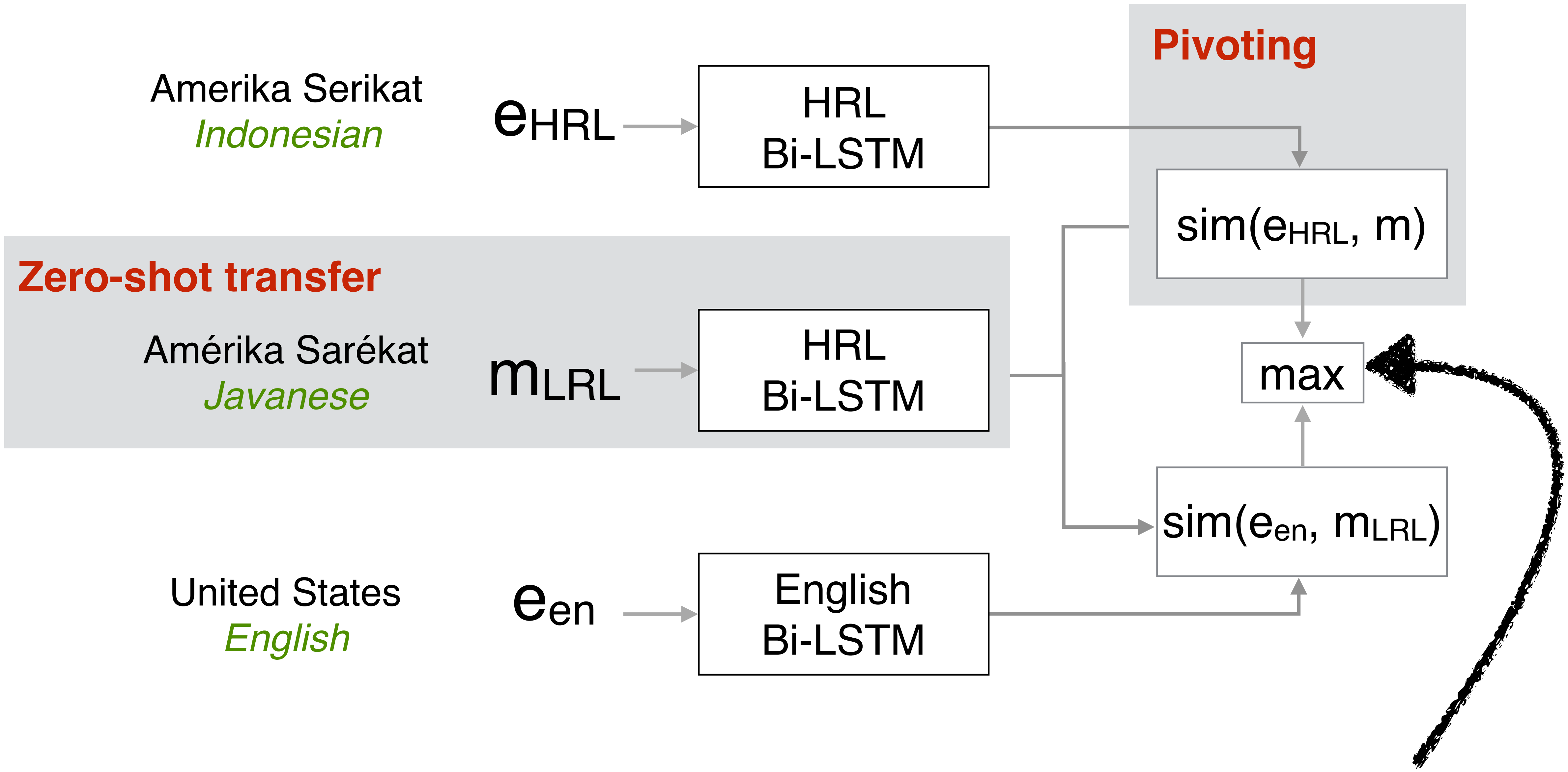
Pivot-Based Linking Model



Pivot-Based Linking Model



Pivot-Based Linking Model



Final score for the KB entry is the **maximum similarity score**

Phonological Transfer

Phonological Transfer

- What if the HRL and LRL use different scripts? Character-level **LSTM won’t transfer!**

Phonological Transfer

- What if the HRL and LRL use different scripts? Character-level **LSTM won’t transfer!**

- Use **language-universal International Phonetic Alphabet** (IPA) instead.

Phonological Transfer

- What if the HRL and LRL use different scripts? Character-level **LSTM won't transfer!**
- Use **language-universal International Phonetic Alphabet** (IPA) instead.

Lao		*Thai*	*English*
ຄາລາບາວ	x	คาราบาว	Carabao

Phonological Transfer

- What if the HRL and LRL use different scripts? Character-level **LSTM won't transfer!**

- Use **language-universal International Phonetic Alphabet** (IPA) instead.

	Lao		*Thai*	*English*
	ຄາລາບາວ	x	คาราบาว	Carabao
IPA	k^{h}aːlaːbaːw	→	k^{h}aːraːbaːw	

Experiments: Cross-lingual KB Title Linking

- 53 high-resource "pivot" languages
- **9 low-resource test languages**

Bengali	বাংলা
Javanese	Javanese
Lao	ພາສາລາວ
Marathi	मराठी
Punjabi	ਪੰਜਾਬੀ
Telugu	తెలుగు
Tigrinya	ትግርኛ
Ukrainian	українська мова
Uyghur	ئۇيغۇر تىلى

- English Wikipedia as KB

Results: Cross-lingual KB Title Linking

Results: Cross-lingual KB Title Linking

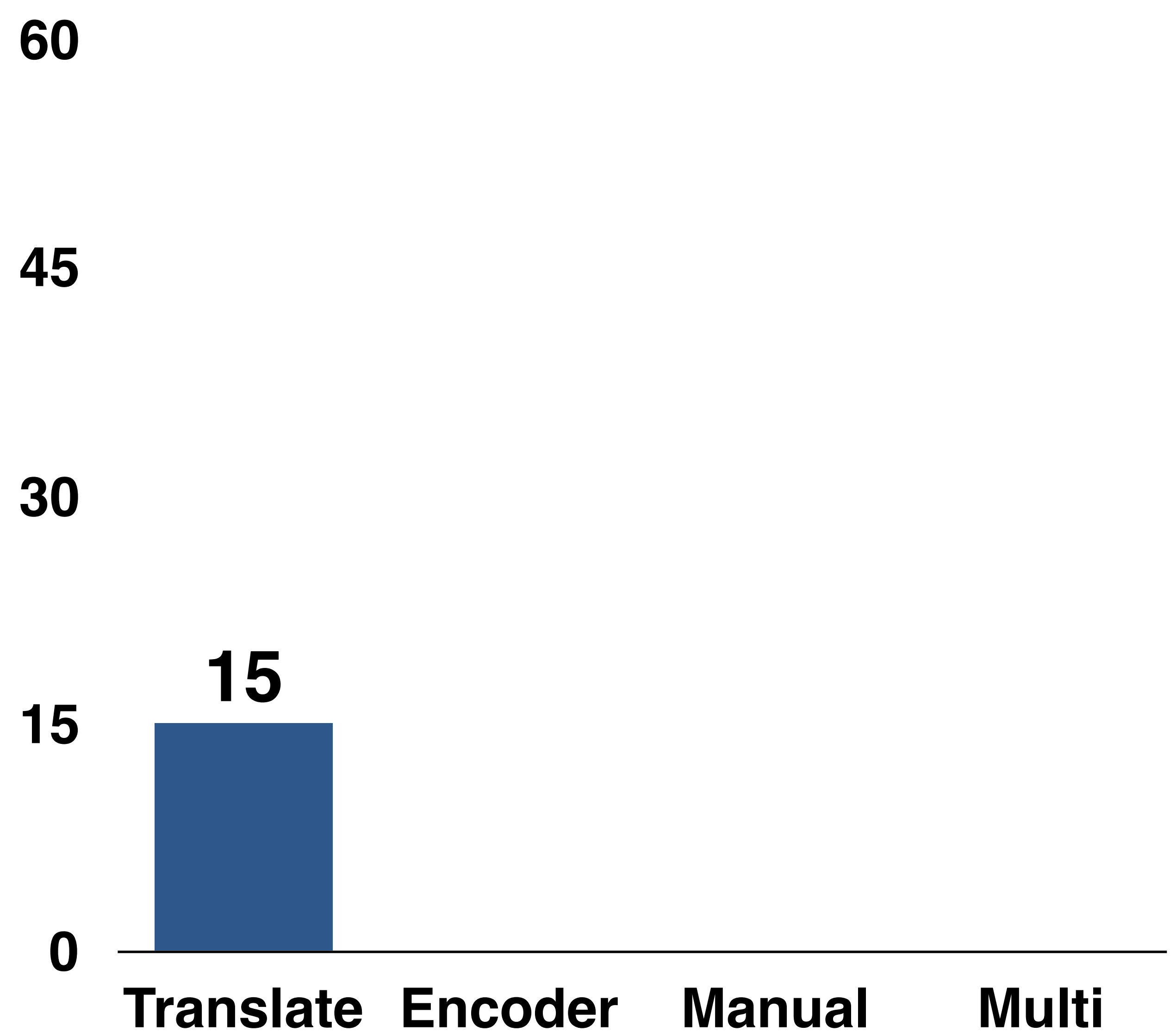
Zero-shot transfer: all models trained on high-resource language

Results: Cross-lingual KB Title Linking

Zero-shot transfer: all models trained on high-resource language

Baselines

- **Translate** using bilingual lexicon

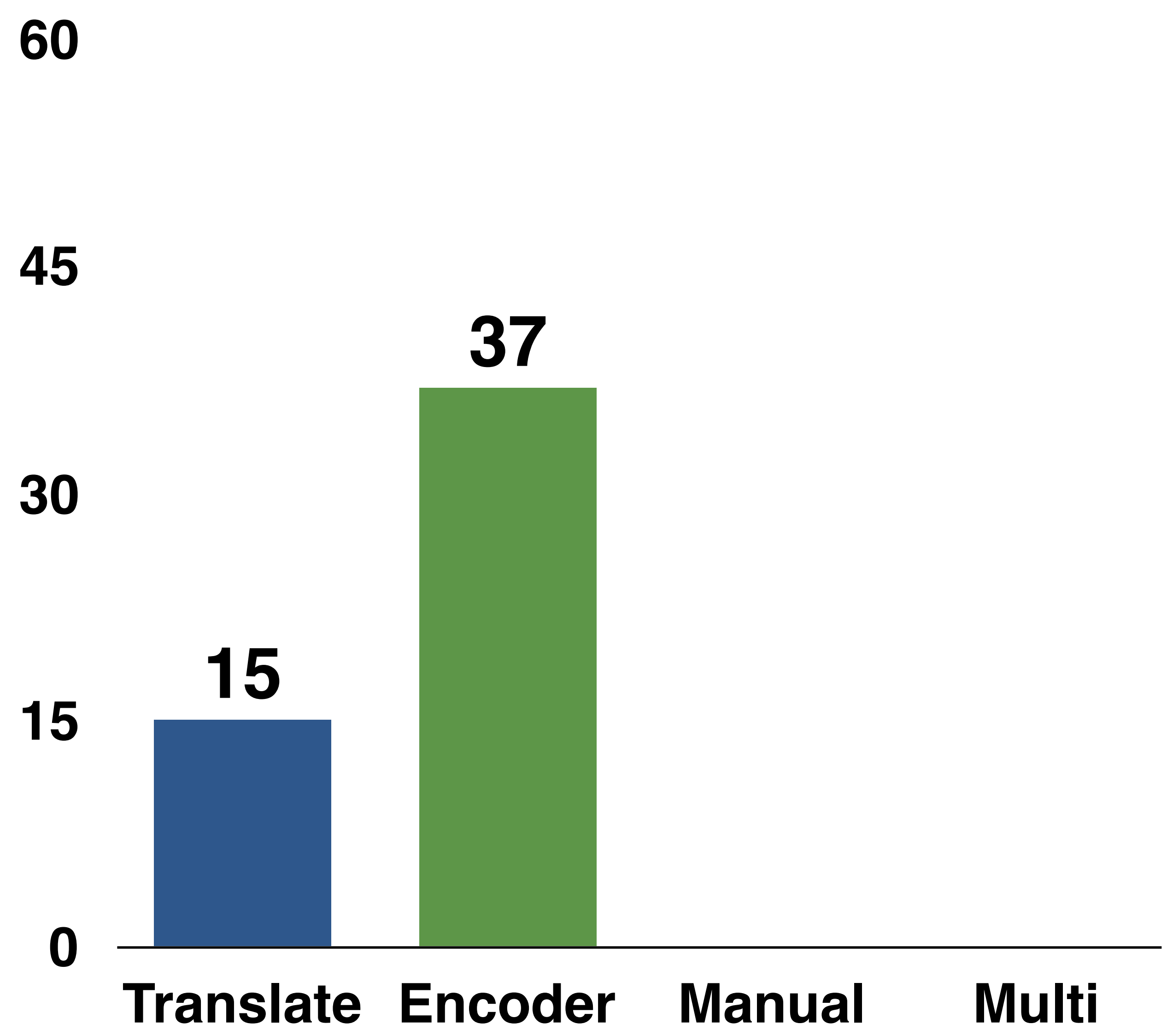


Results: Cross-lingual KB Title Linking

Zero-shot transfer: all models trained on high-resource language

Baselines

- **Translate** using bilingual lexicon
- Neural similarity **encoder**



Results: Cross-lingual KB Title Linking

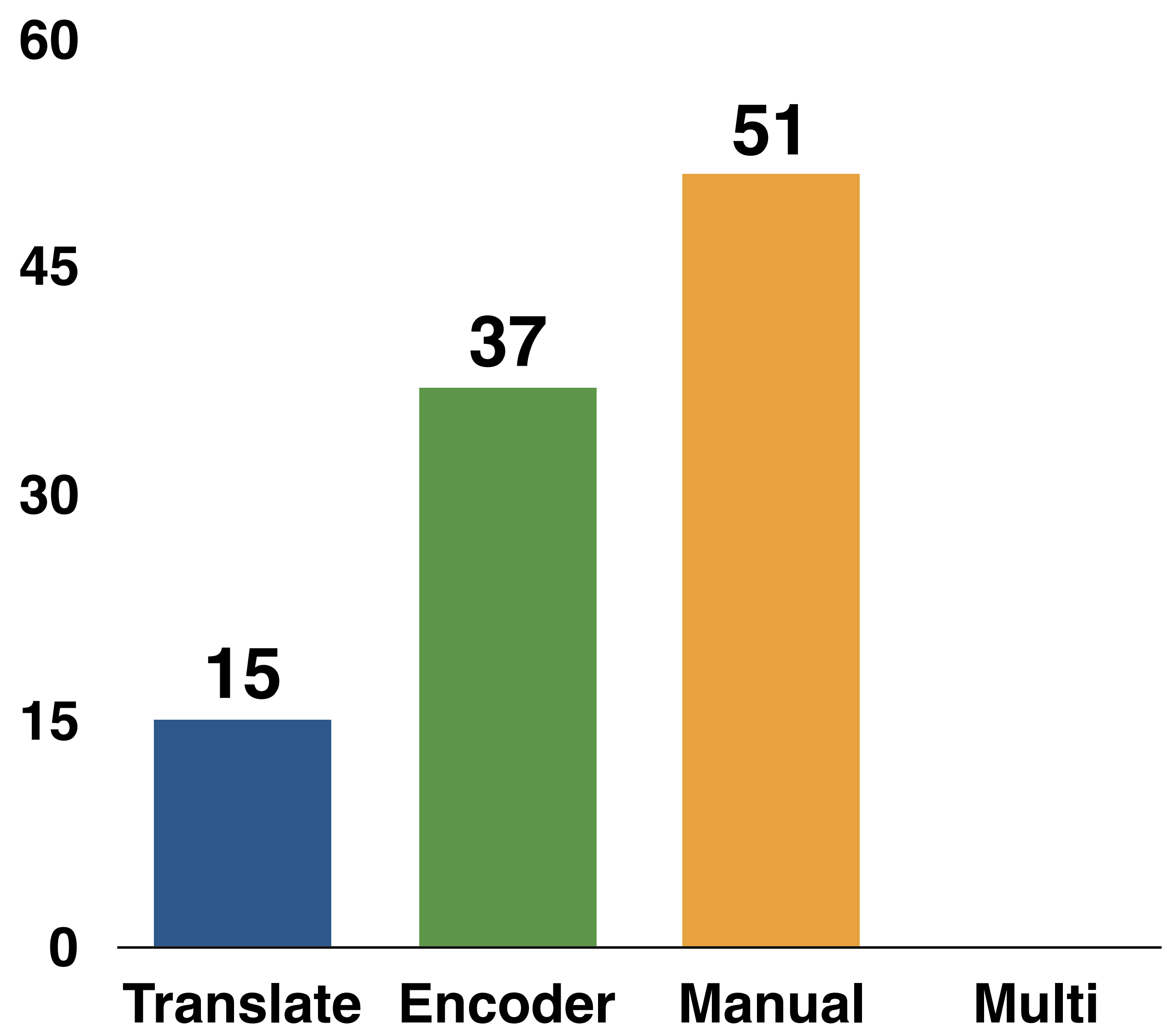
Zero-shot transfer: all models trained on high-resource language

Baselines

- **Translate** using bilingual lexicon
- Neural similarity **encoder**

Pivoting

- **Manual**ly-selected pivot language



Results: Cross-lingual KB Title Linking

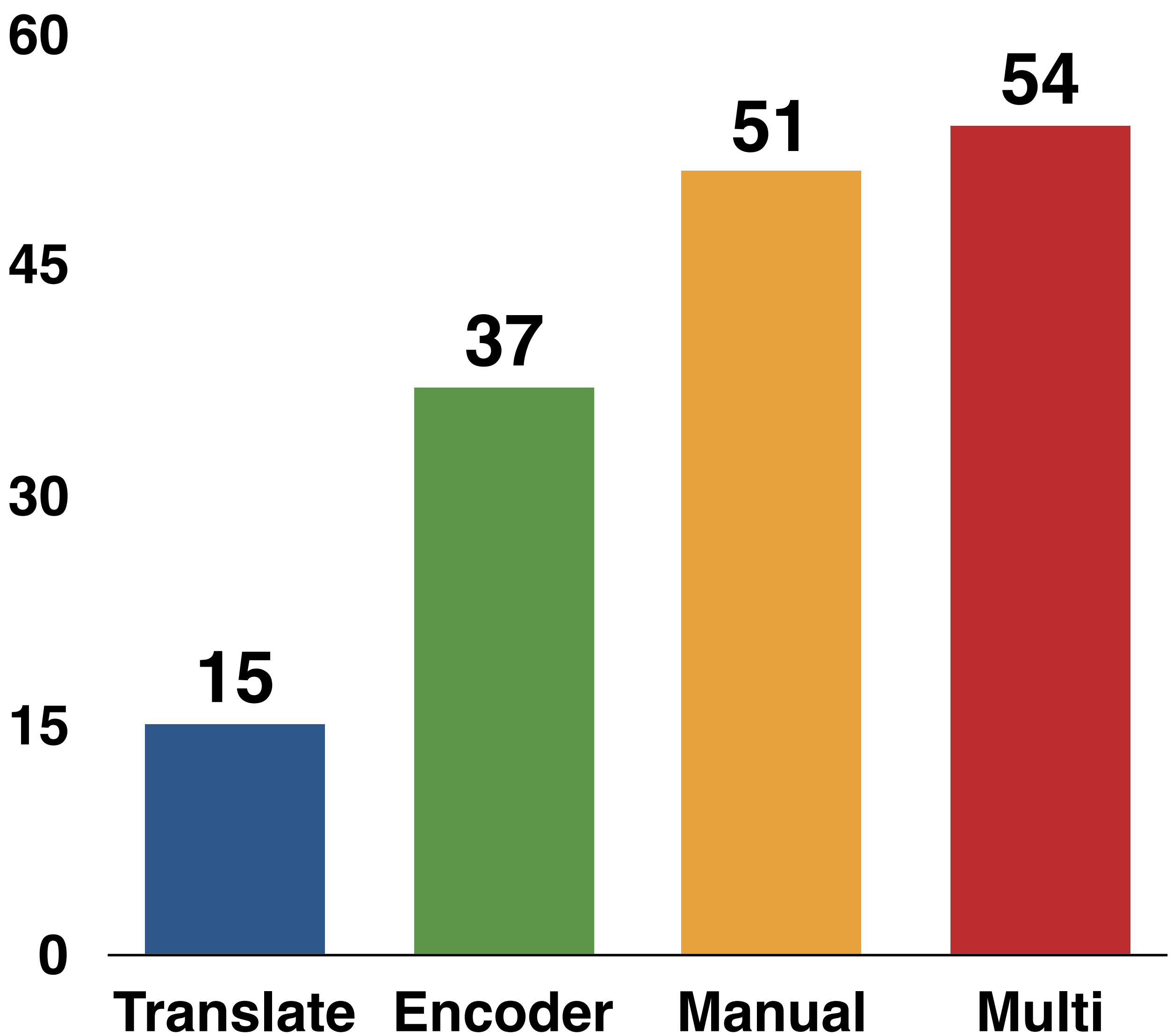
Zero-shot transfer: all models trained on high-resource language

Baselines

- **Translate** using bilingual lexicon
- Neural similarity **encoder**

Pivoting

- **Manual**ly-selected pivot language
- **Multi**ple pivots



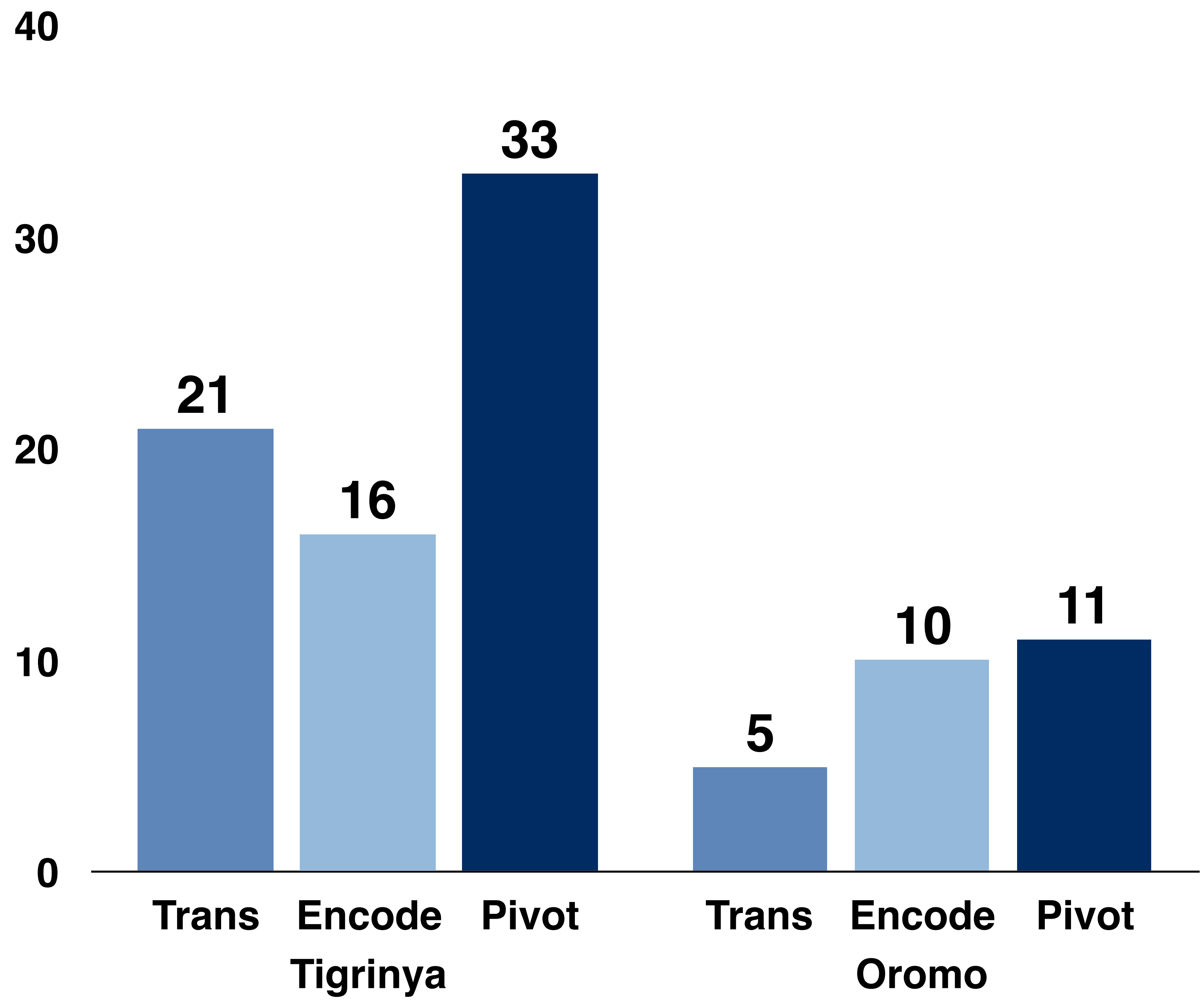
Experiments: Full Cross-lingual Entity Linking

Experiments: Full Cross-lingual Entity Linking

Pivots: Amharic for Tigrinya
and Somali for Oromo

Experiments: Full Cross-lingual Entity Linking

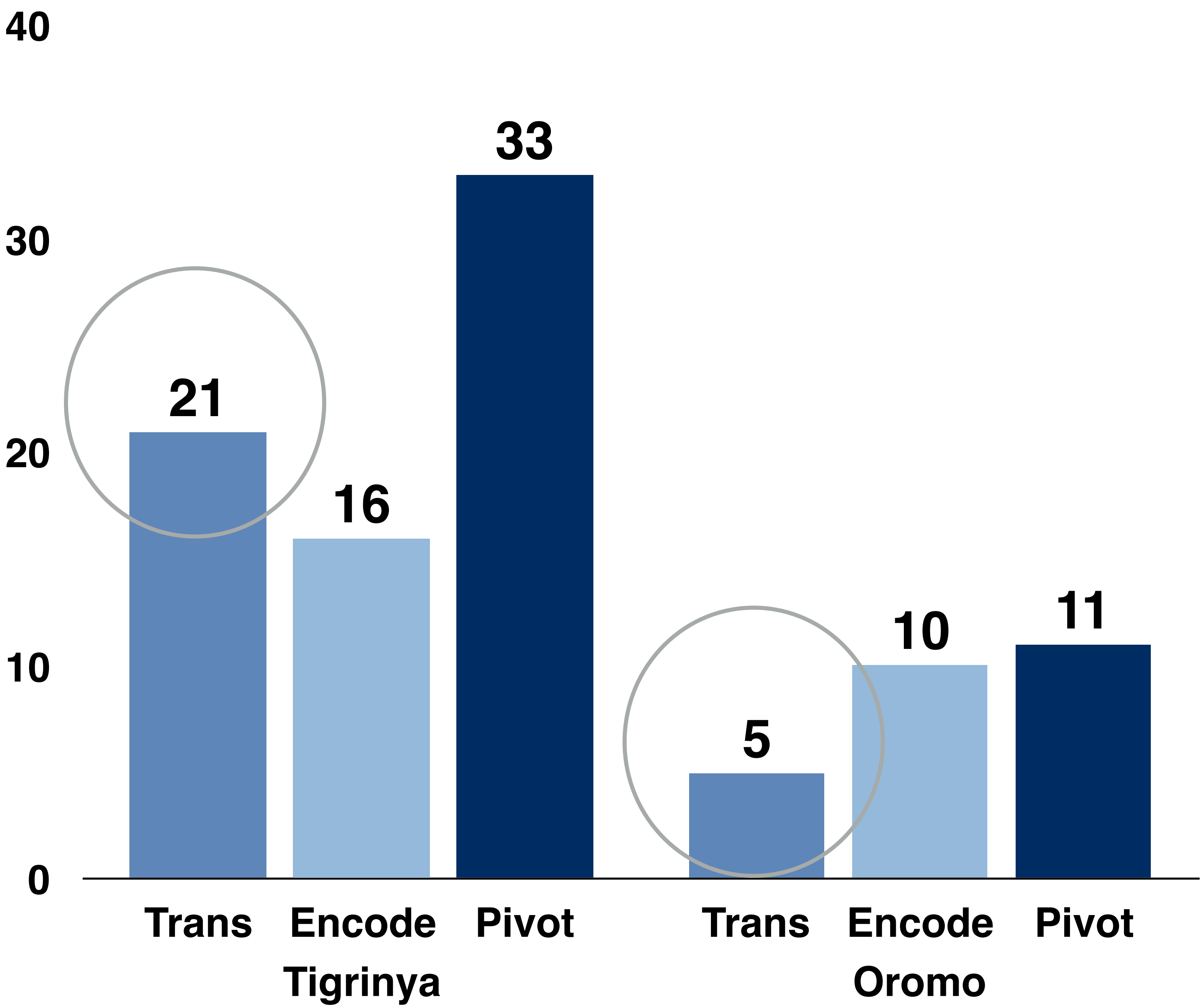
Pivots: Amharic for Tigrinya and Somali for Oromo



Experiments: Full Cross-lingual Entity Linking

Pivots: Amharic for Tigrinya and Somali for Oromo

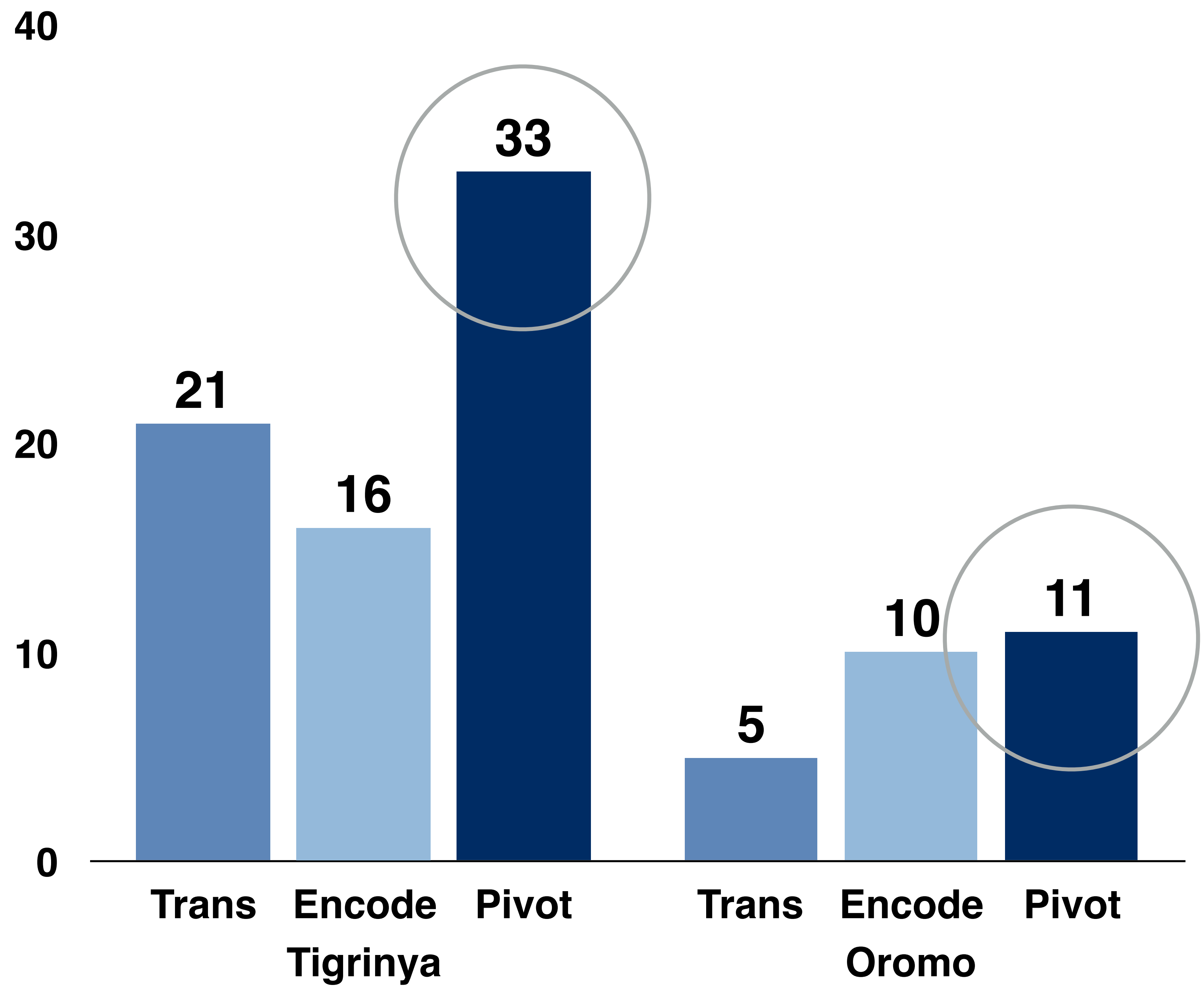
- Supervised translate does not work well because of small Wikipedia size



Experiments: Full Cross-lingual Entity Linking

Pivots: Amharic for Tigrinya and Somali for Oromo

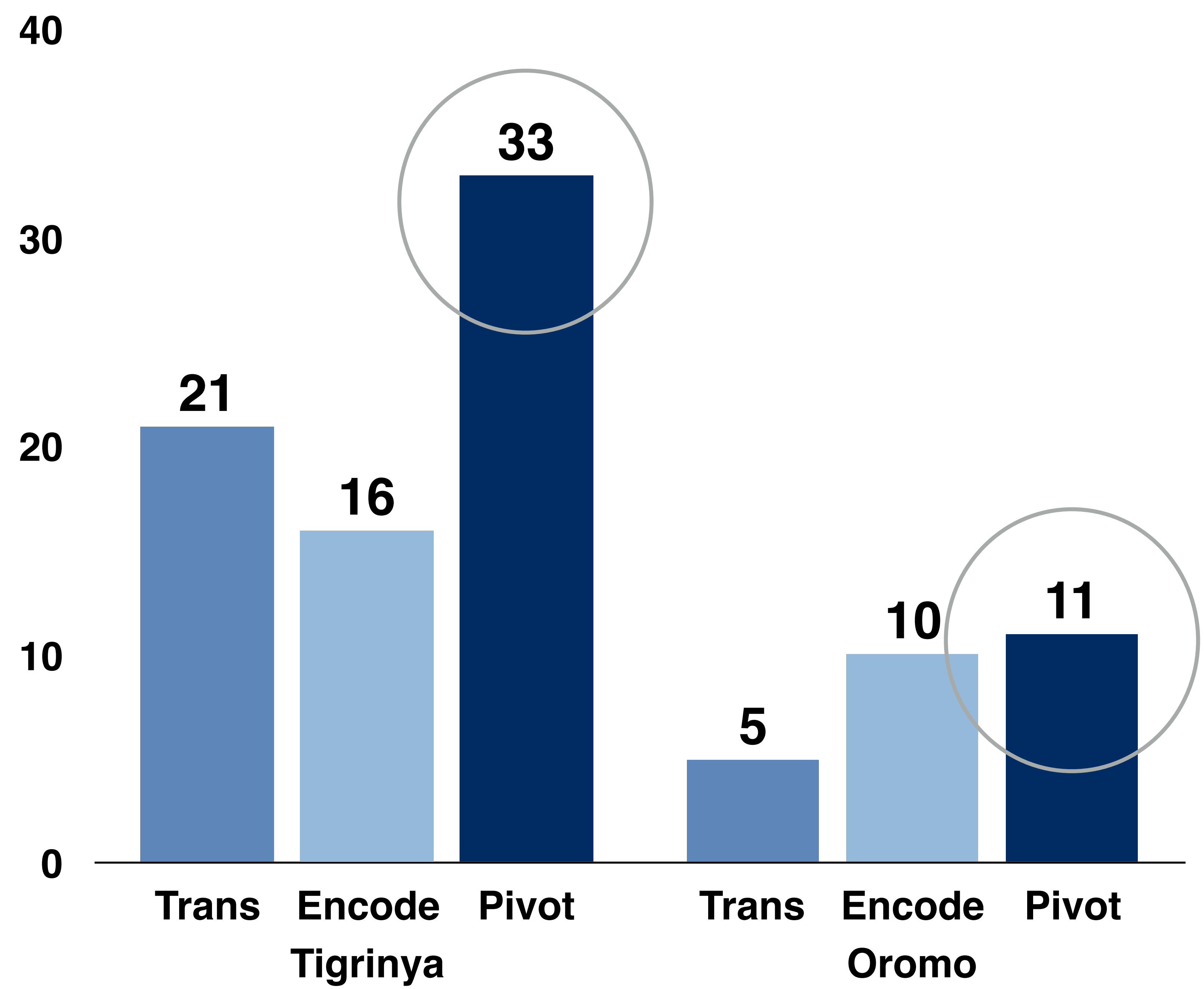
- Supervised translate does not work well because of small Wikipedia size
- Pivoting works better for Tigrinya than Oromo



Experiments: Full Cross-lingual Entity Linking

Pivots: Amharic for Tigrinya and Somali for Oromo

- Supervised translate does not work well because of small Wikipedia size
- Pivoting works better for Tigrinya than Oromo
- **Closely-related pivot language is necessary**: Amharic and Tigrinya are more similar than Somali and Oromo



Analysis: Phonological Transfer

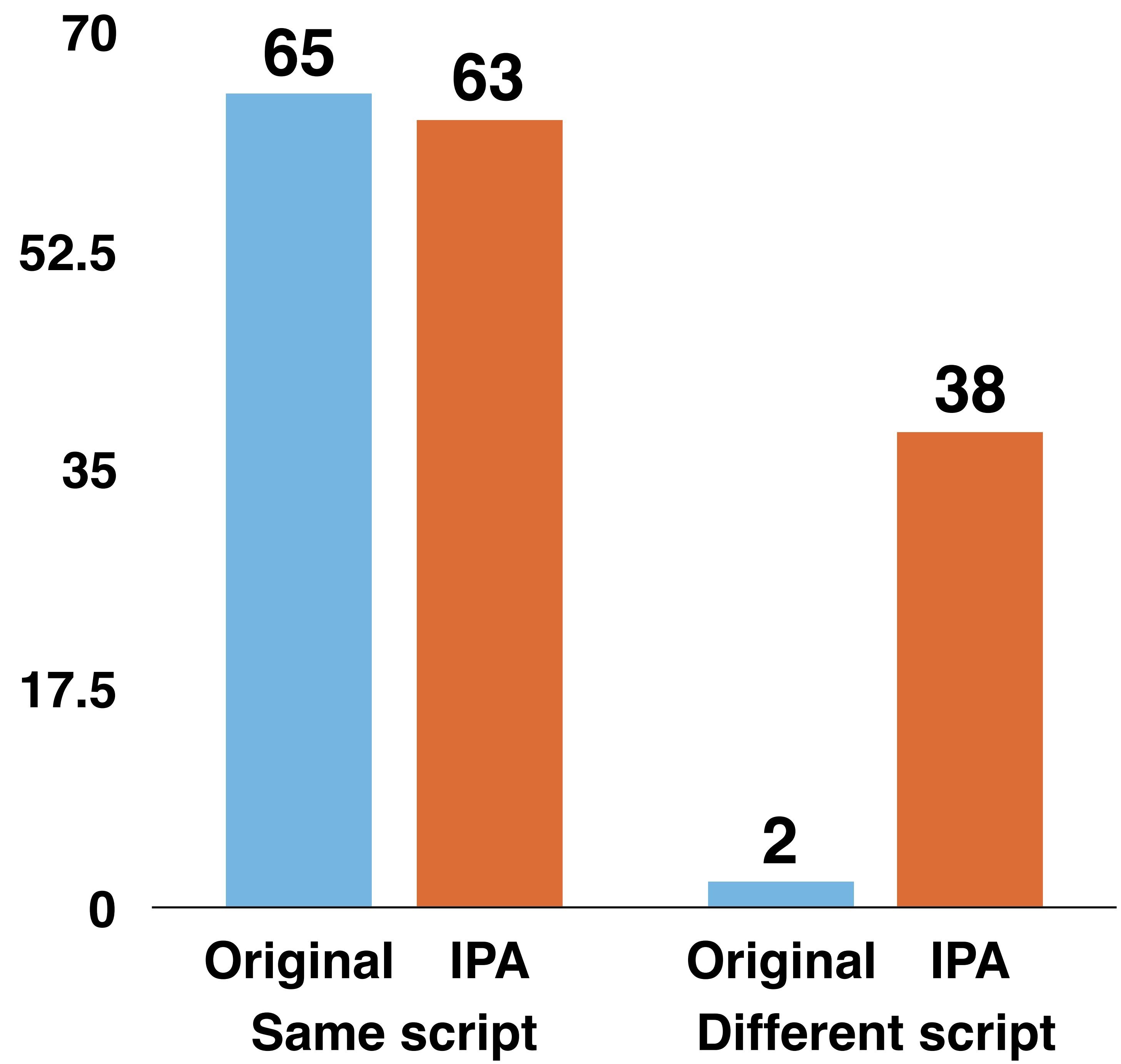
What if pivot and source languages use different scripts?

Use IPA to transfer!

Analysis: Phonological Transfer

What if pivot and source languages use different scripts?

Use IPA to transfer!

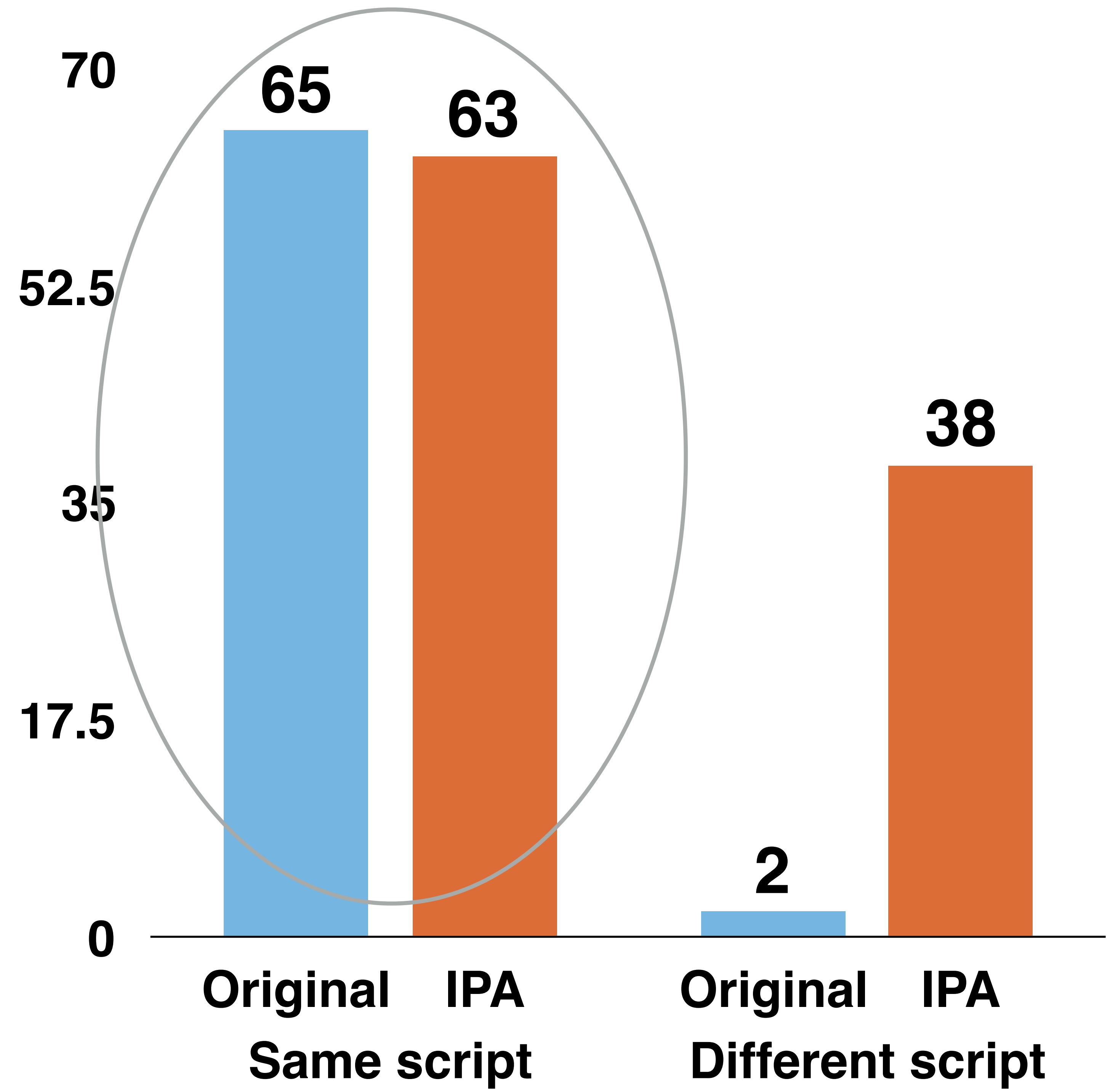


Analysis: Phonological Transfer

What if pivot and source languages use different scripts?

Use IPA to transfer!

- Original script transfer works better if the script is same — but not by too much!

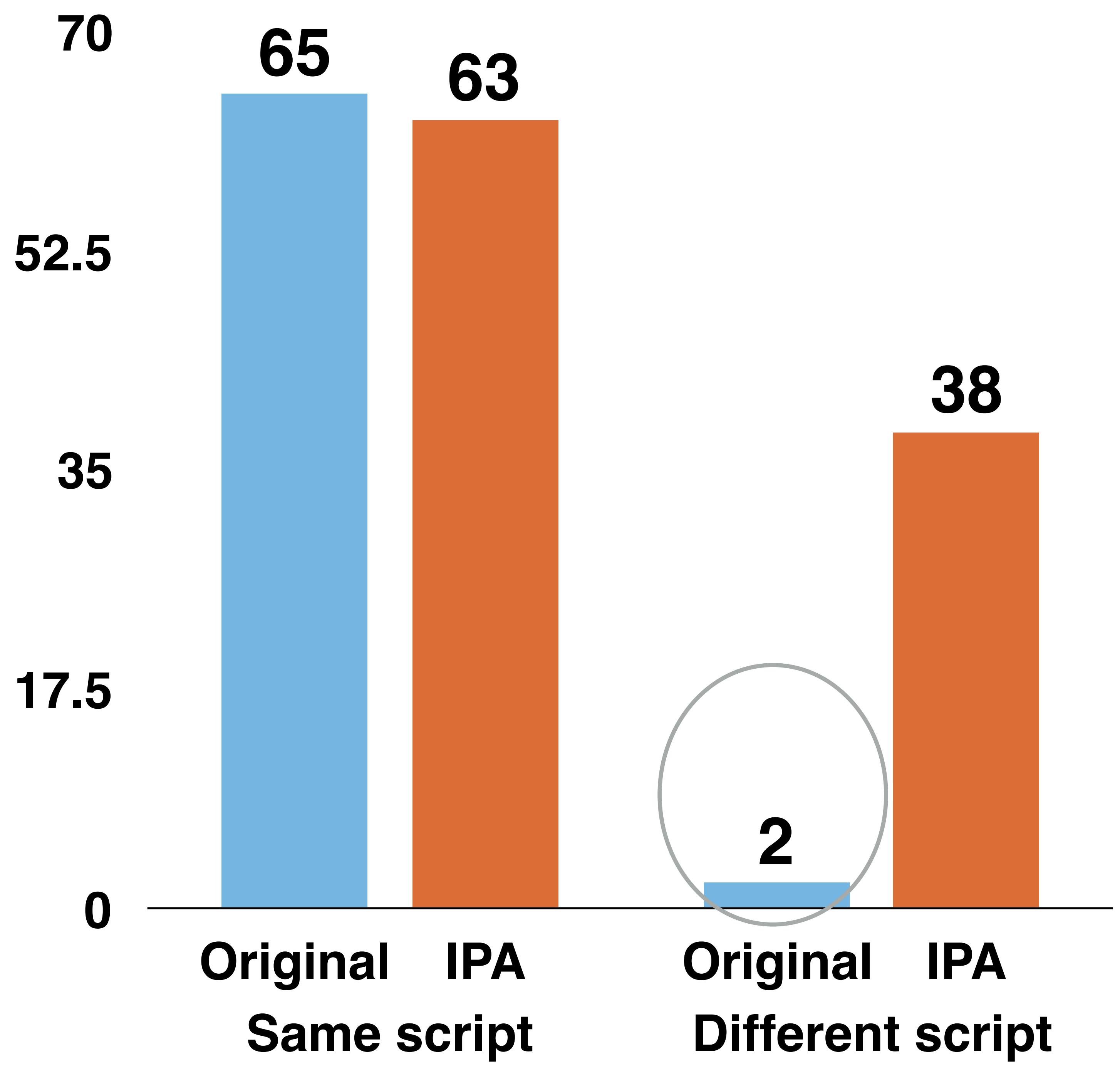


Analysis: Phonological Transfer

What if pivot and source languages use different scripts?

Use IPA to transfer!

- Original script transfer works better if the script is same — but not by too much!
- Original script transfer completely fails when scripts are different

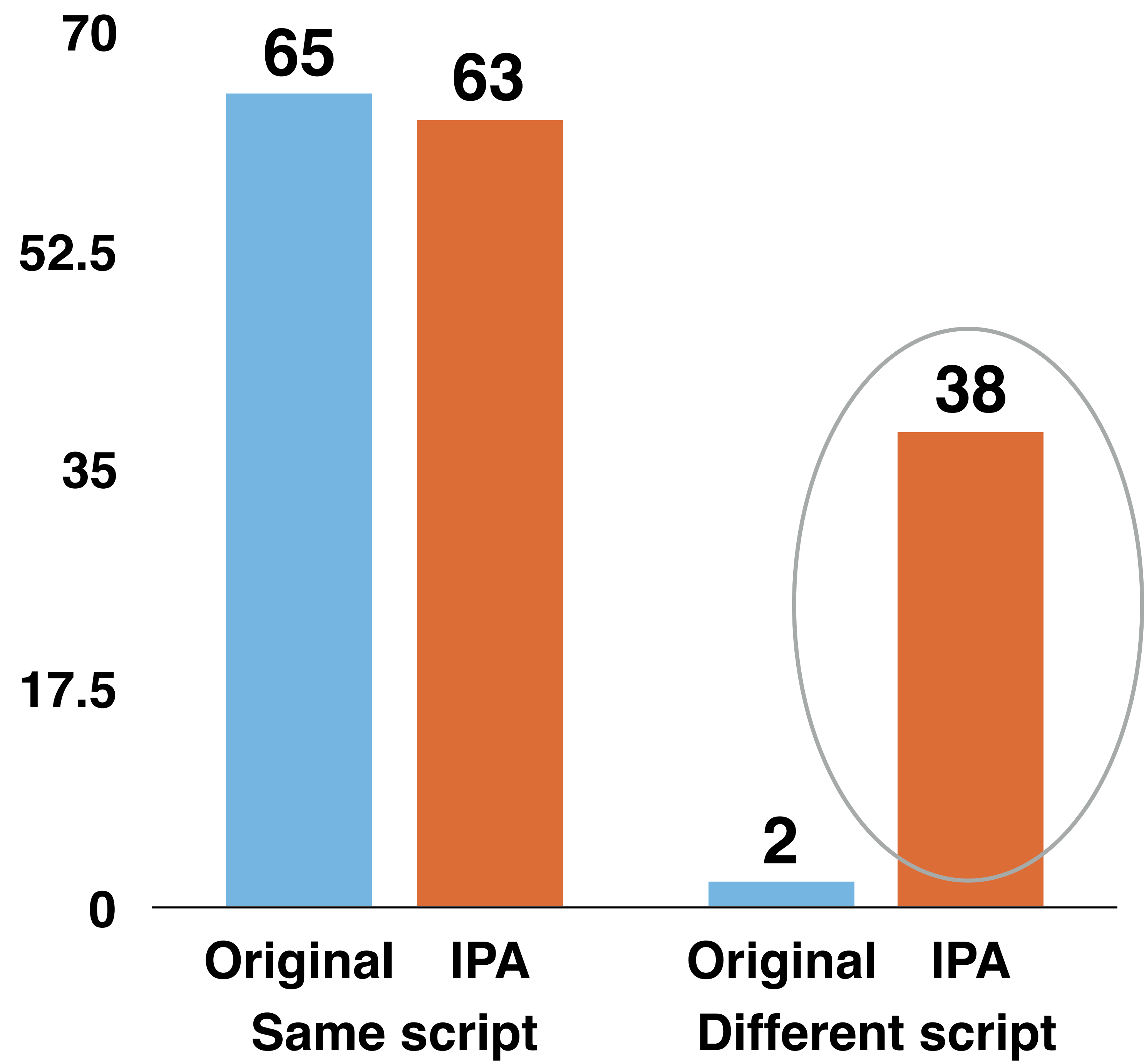


Analysis: Phonological Transfer

What if pivot and source languages use different scripts?

Use IPA to transfer!

- Original script transfer works better if the script is same — but not by too much!
- Original script transfer completely fails when scripts are different
- **IPA representation leads to much better zero-shot transfer**



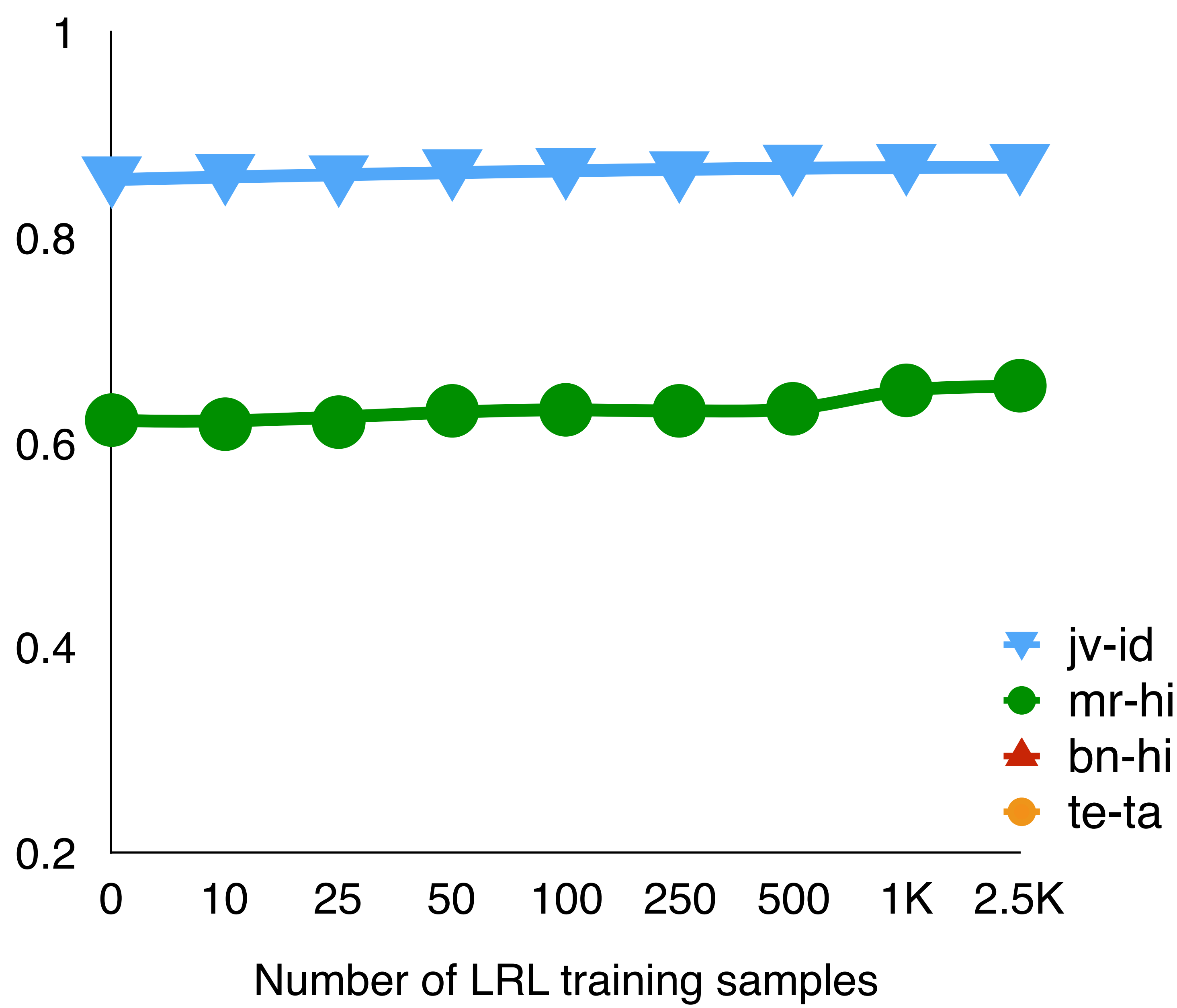
Analysis: Joint Training with Low-resource Language

Add low-resource language samples to training data

Analysis: Joint Training with Low-resource Language

Add low-resource language samples to training data

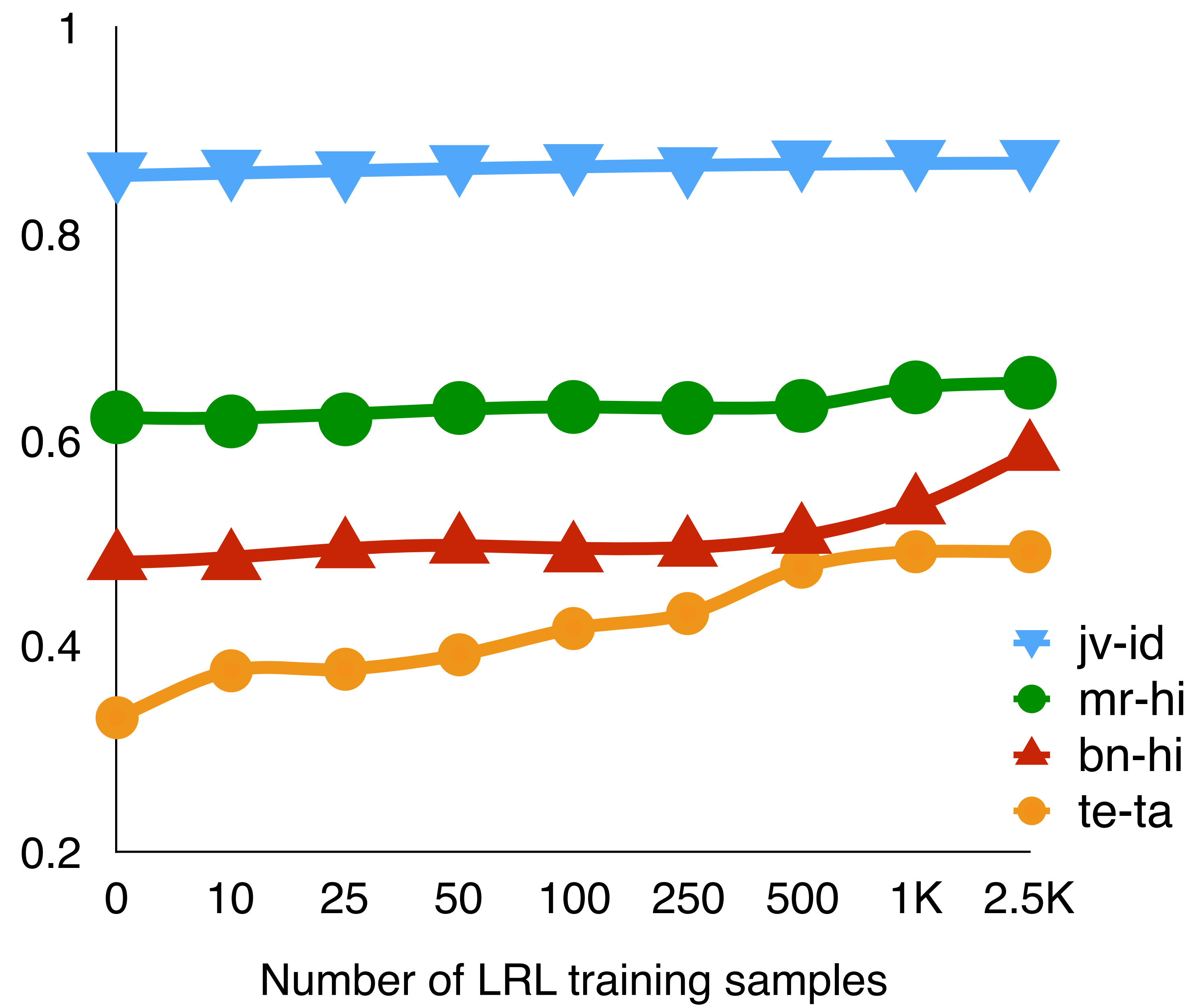
- Languages that are similar (*jv-id*, *hi-mr*) show small improvement



Analysis: Joint Training with Low-resource Language

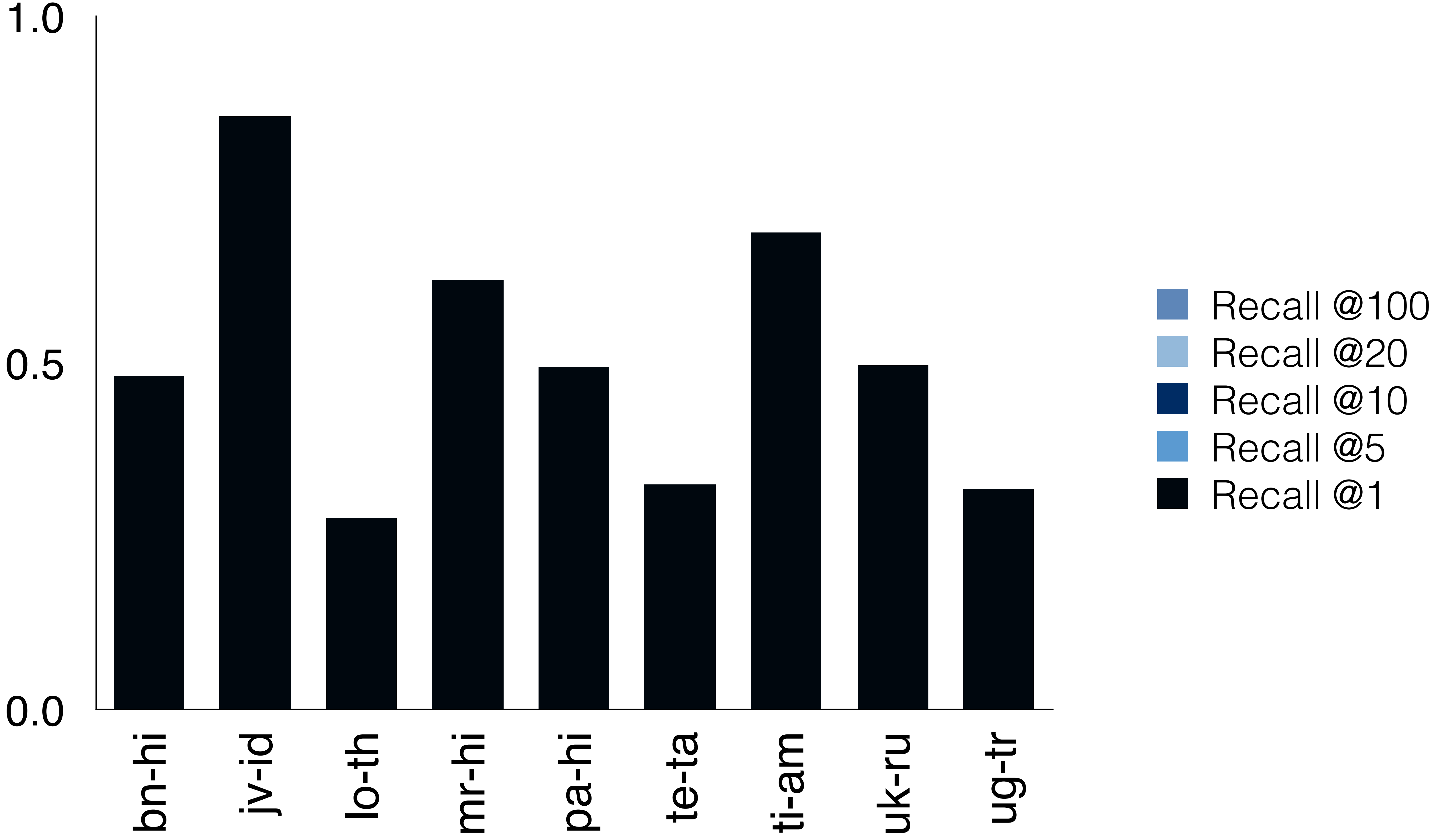
Add low-resource language samples to training data

- Languages that are similar (*jv-id*, *hi-mr*) show small improvement
- IPA transfer models show much more improvement (*te-ta*, *bn-hi*)

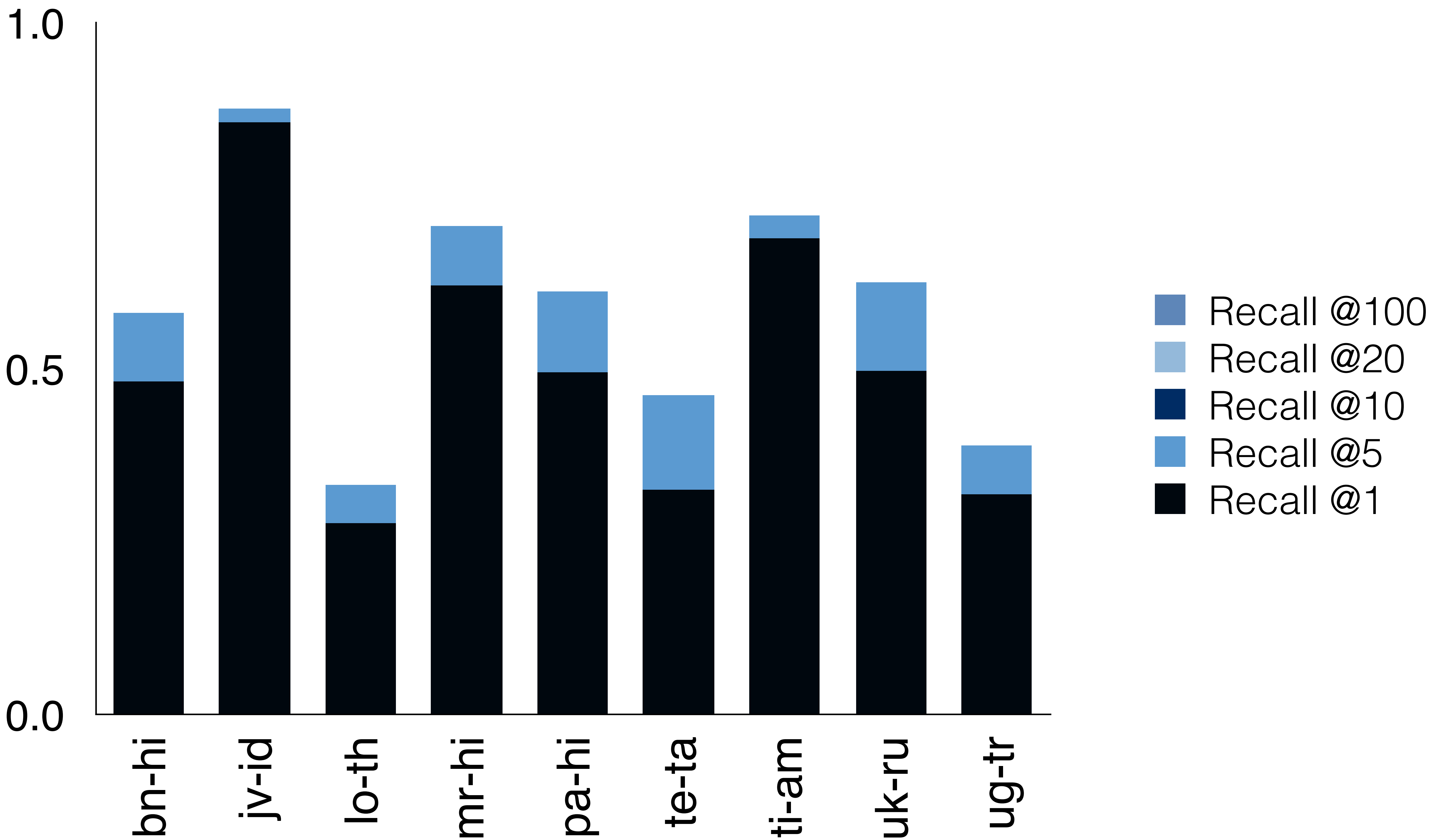


Analysis: Candidate Retrieval

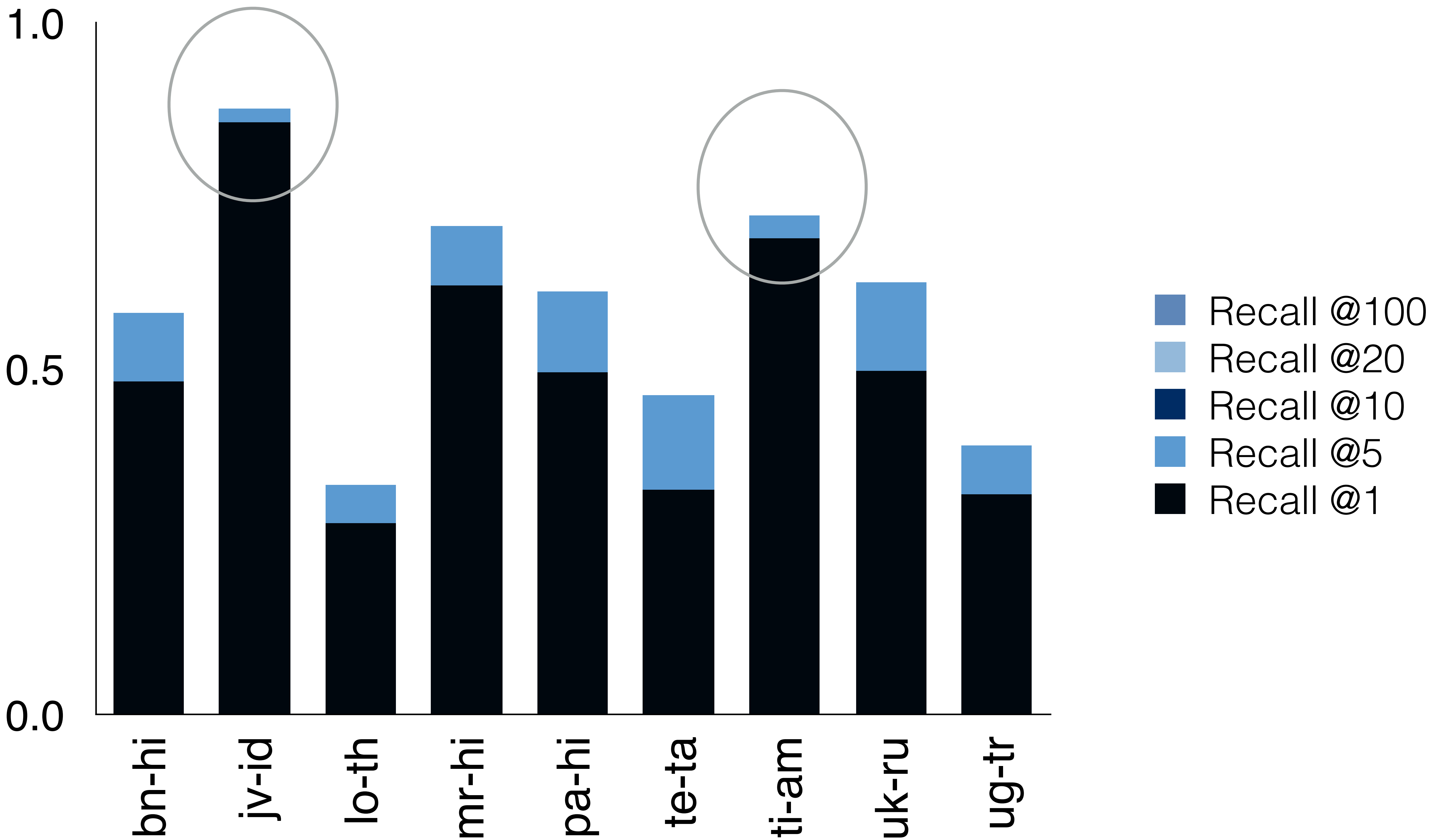
Analysis: Candidate Retrieval



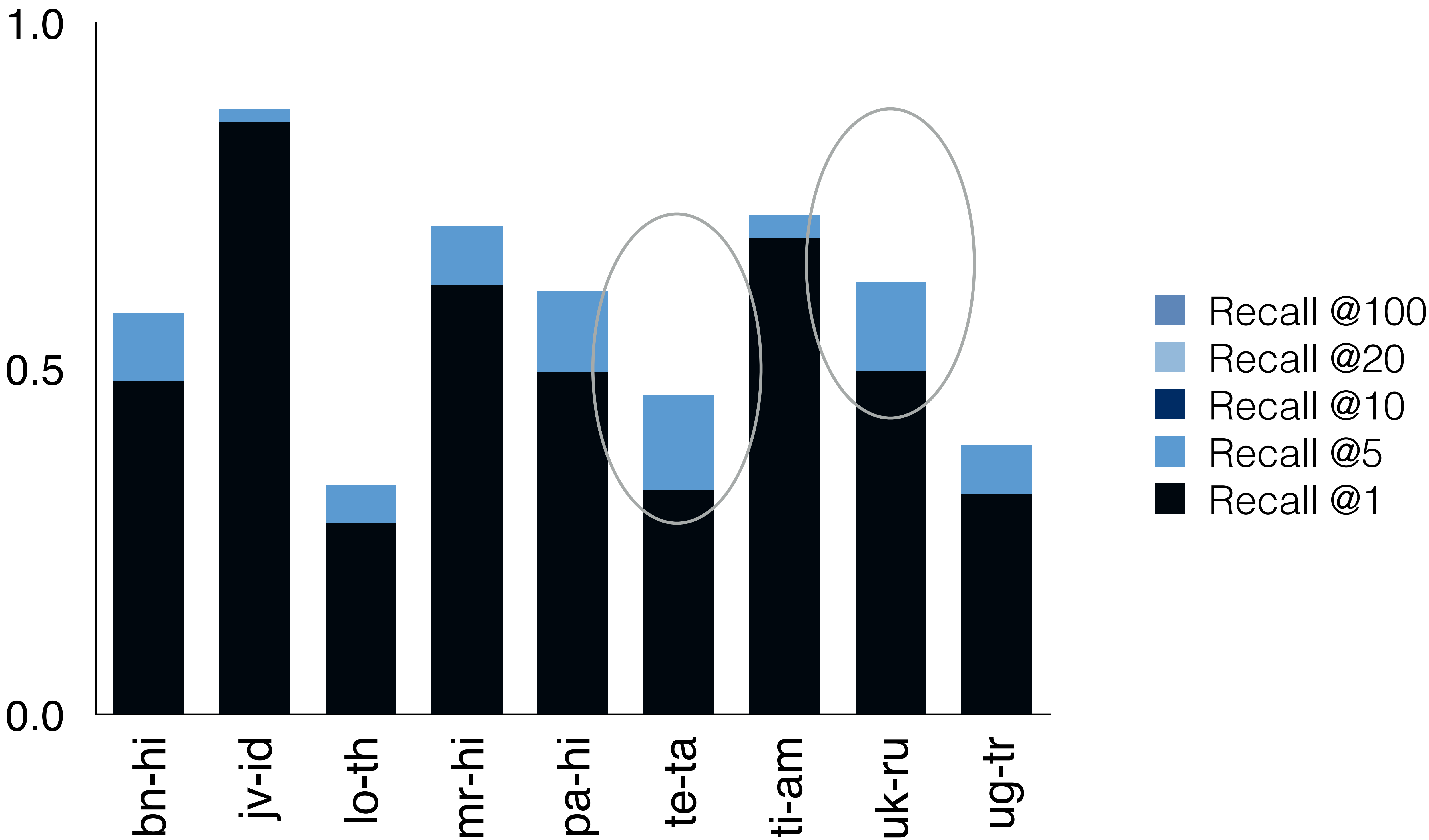
Analysis: Candidate Retrieval



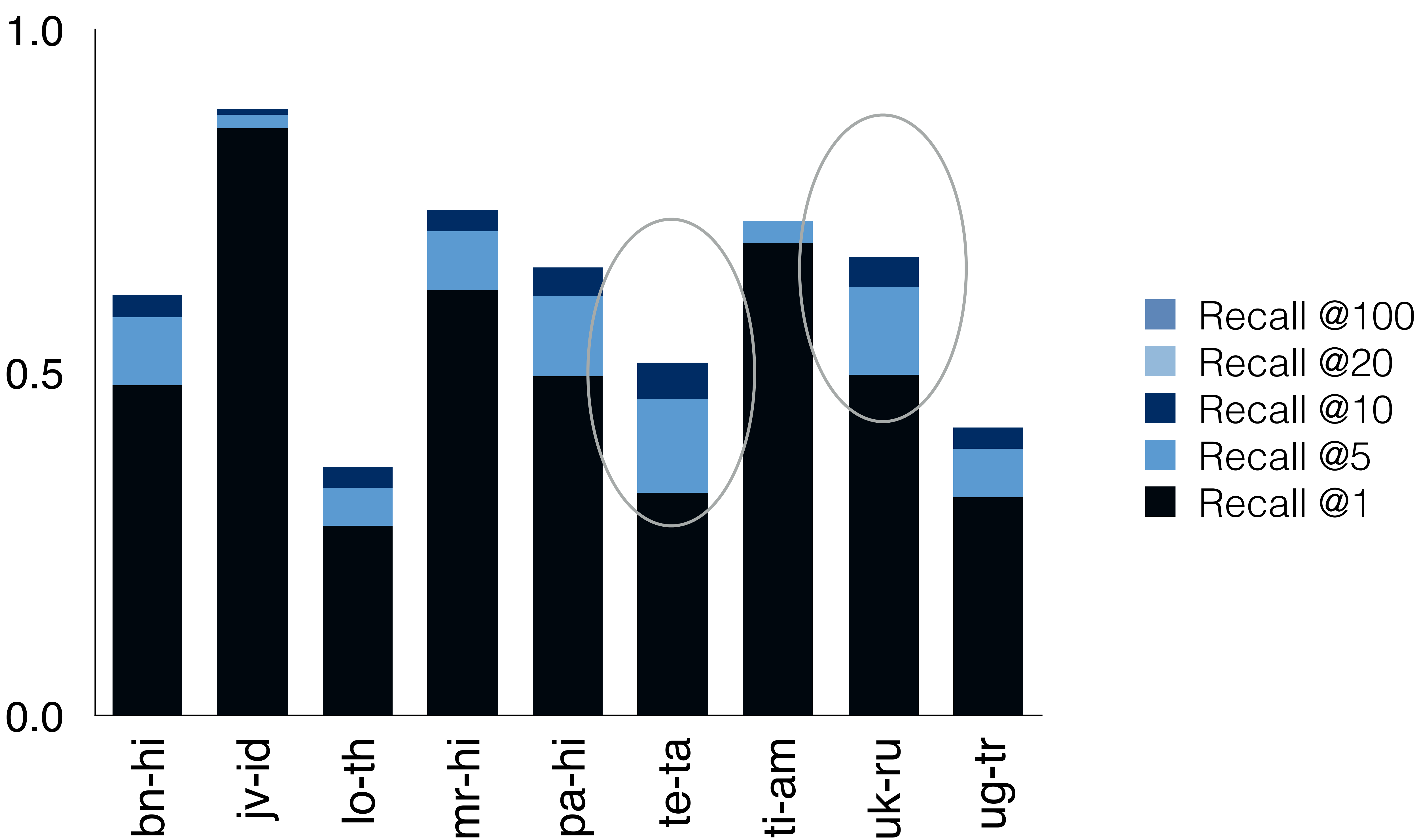
Analysis: Candidate Retrieval



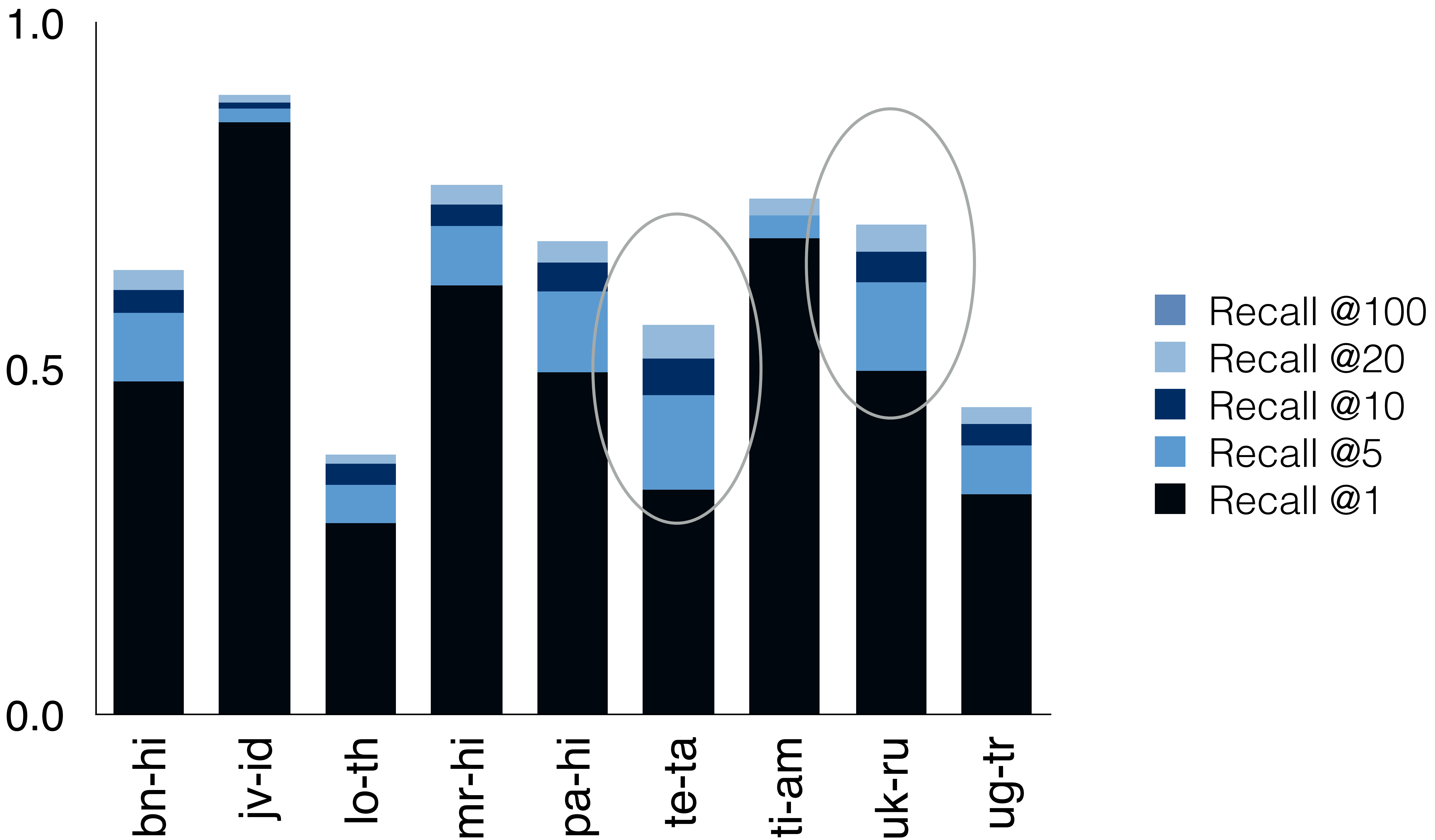
Analysis: Candidate Retrieval



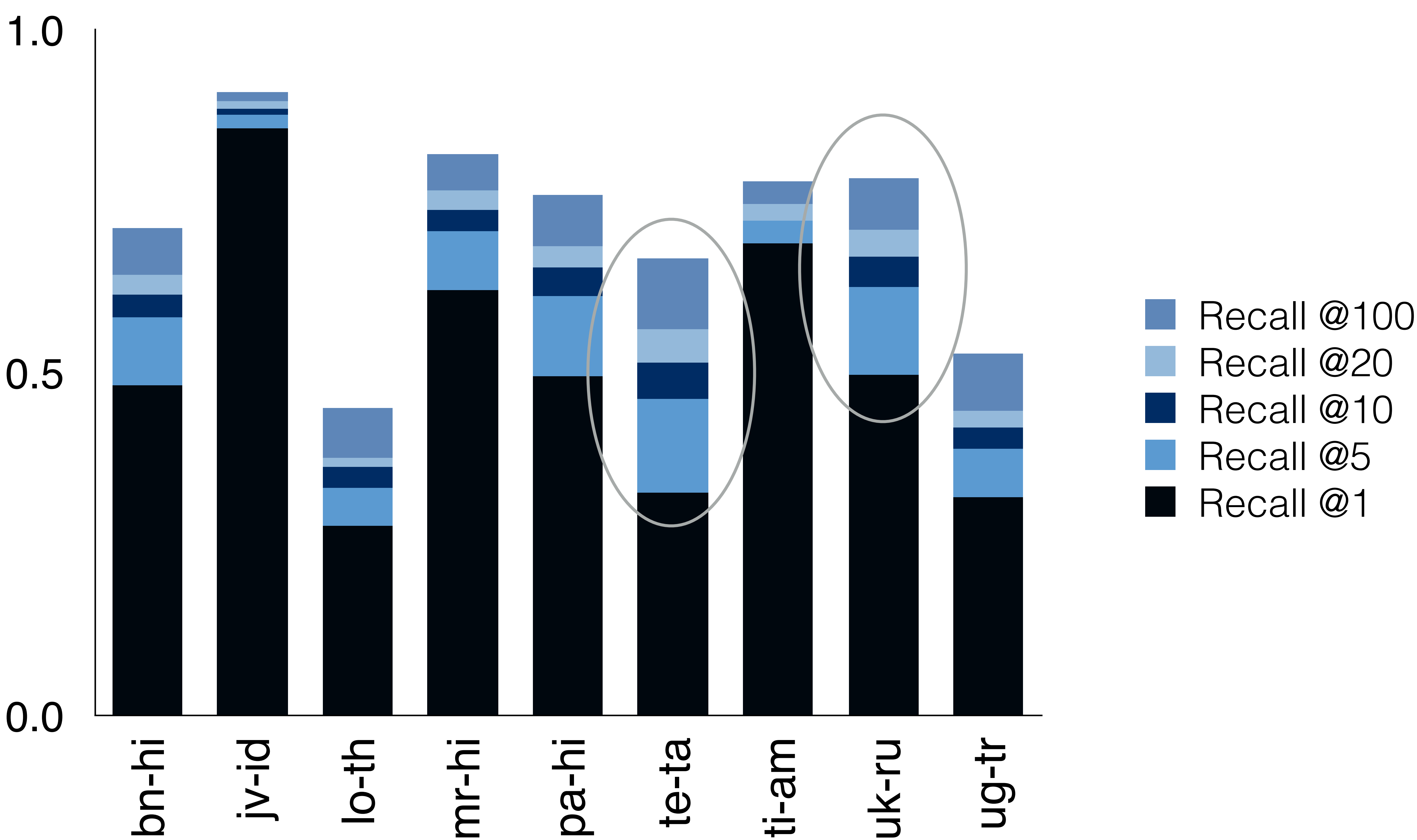
Analysis: Candidate Retrieval



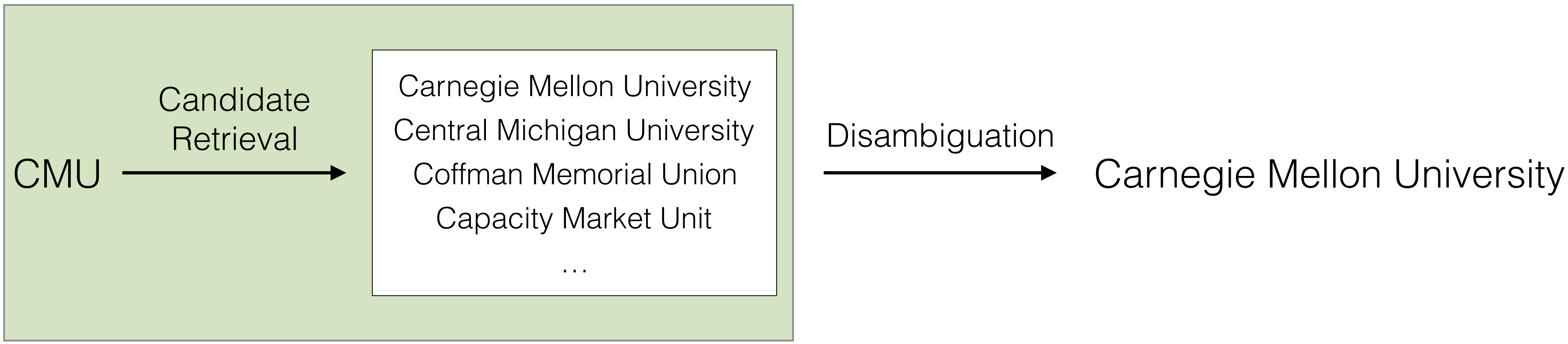
Analysis: Candidate Retrieval



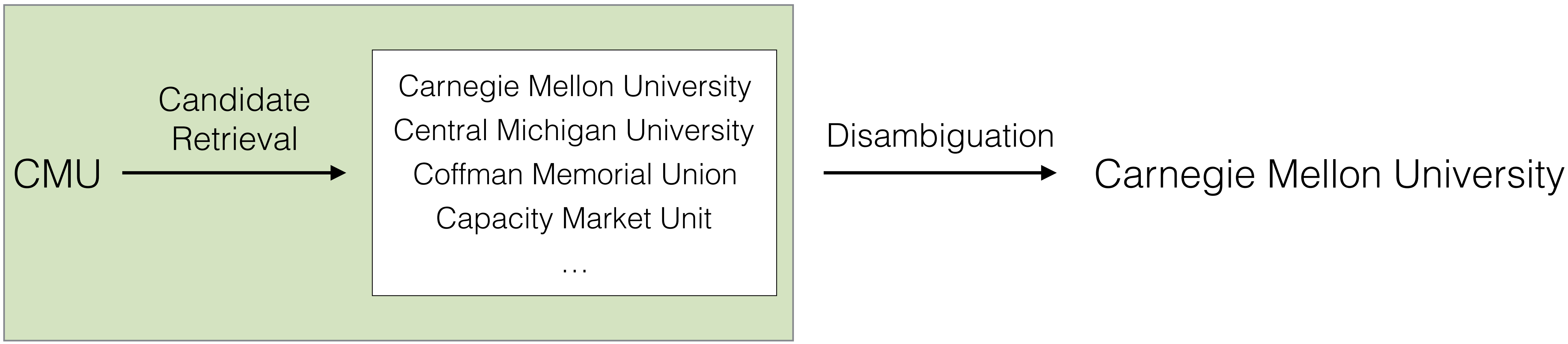
Analysis: Candidate Retrieval



Conclusion

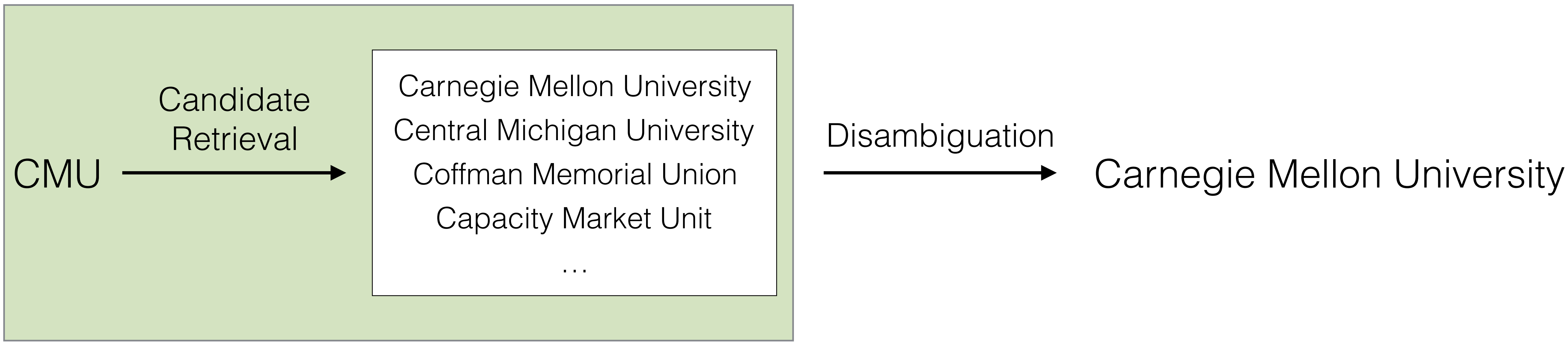

CMU
Candidate Retrieval
Carnegie Mellon University
Central Michigan University
Coffman Memorial Union
Capacity Market Unit
...
Disambiguation
Carnegie Mellon University

Conclusion



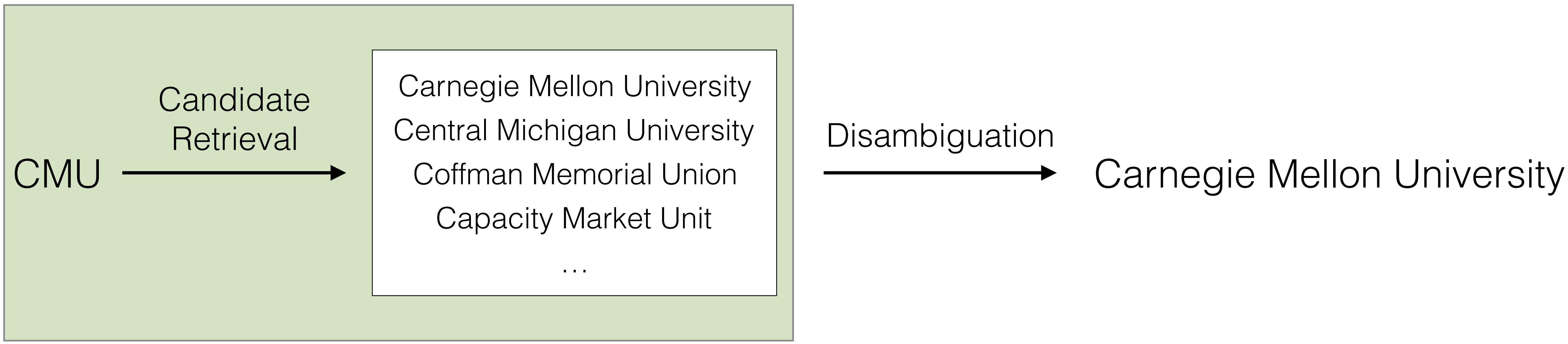
- PBEL is an entity linking method that leverages high-resource languages for zero-shot transfer.

Conclusion



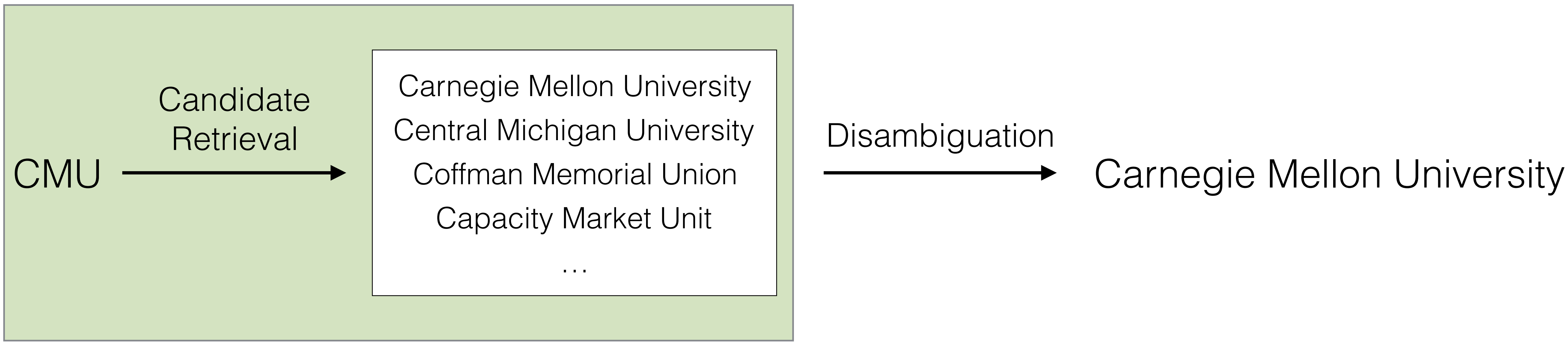
- PBEL is an entity linking method that leverages high-resource languages for zero-shot transfer.
    - Average **improvement of 17%** over baseline systems

Conclusion



- PBEL is an entity linking method that leverages high-resource languages for zero-shot transfer.
  - Average **improvement of 17%** over baseline systems
  - **IPA transfer improves pivoting by 36%** on average for dissimilar scripts

Conclusion



- PBEL is an entity linking method that leverages high-resource languages for zero-shot transfer.
  - Average **improvement of 17%** over baseline systems
  - **IPA transfer improves pivoting by 36%** on average for dissimilar scripts
- Can we do better?

Pivot-Based Entity Linking

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to the low-resource language

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to the low-resource language

Pivoting

Link to closely-related "pivot" language, instead of English

Amérika Sarékat → Amerika Serikat ········ United States

Javanese *Indonesian*

Pivot-Based Entity Linking

A method to score input entities that **uses no bilingual resources in the source language.**

Zero-shot transfer

Train the entity linking model on a high-resource language and transfer to th

Can we do better?

Pivoting

Link to closely-related "pivot" language, instead of English

Amérika Sarékat → Amerika Serikat ······· United States

Javanese *Indonesian*

Error Analysis

Error Analysis

- PBEL performs better on low-resource languages than translation-based methods.
    - But, significant room for improvement remains.

Error Analysis

- PBEL performs better on low-resource languages than translation-based methods.
  - But, significant room for improvement remains.
- **Systematic analysis of errors** made by PBEL
  - 100 errors from four low-resource languages
  - Manual inspection to **classify types of errors**

Examples of Errors

	Mention in Marathi	**English KB Entry**
DIRECT	कोबी स्मल्डर्स *Cobie Smulders*	Cobie Smulders

Mention has a **word-by-word mapping** with the target KB entry

Examples of Errors

	Mention in Marathi	English KB Entry
DIRECT	कोबी स्मल्डर्स *Cobie Smulders*	Cobie Smulders
ALIAS	जॅकोबा फ्रांसिस्का मरिया स्मल्डर्स *Jacoba Francisca Maria Smulders*	Cobie Smulders

Mention is an **alternate name** of the target KB entry

Examples of Errors

	Mention in Marathi	**English KB Entry**
DIRECT	कोबी स्मल्डर्स *Cobie Smulders*	Cobie Smulders
ALIAS	जॅकोबा फ्रांसिस्का मरिया स्मल्डर्स *Jacoba Francisca Maria Smulders*	Cobie Smulders
TRANS	कार्नेगी मेलॉन विद्यापीठ *Carnegie Mellon Vidyaapeeth*	Carnegie Mellon University

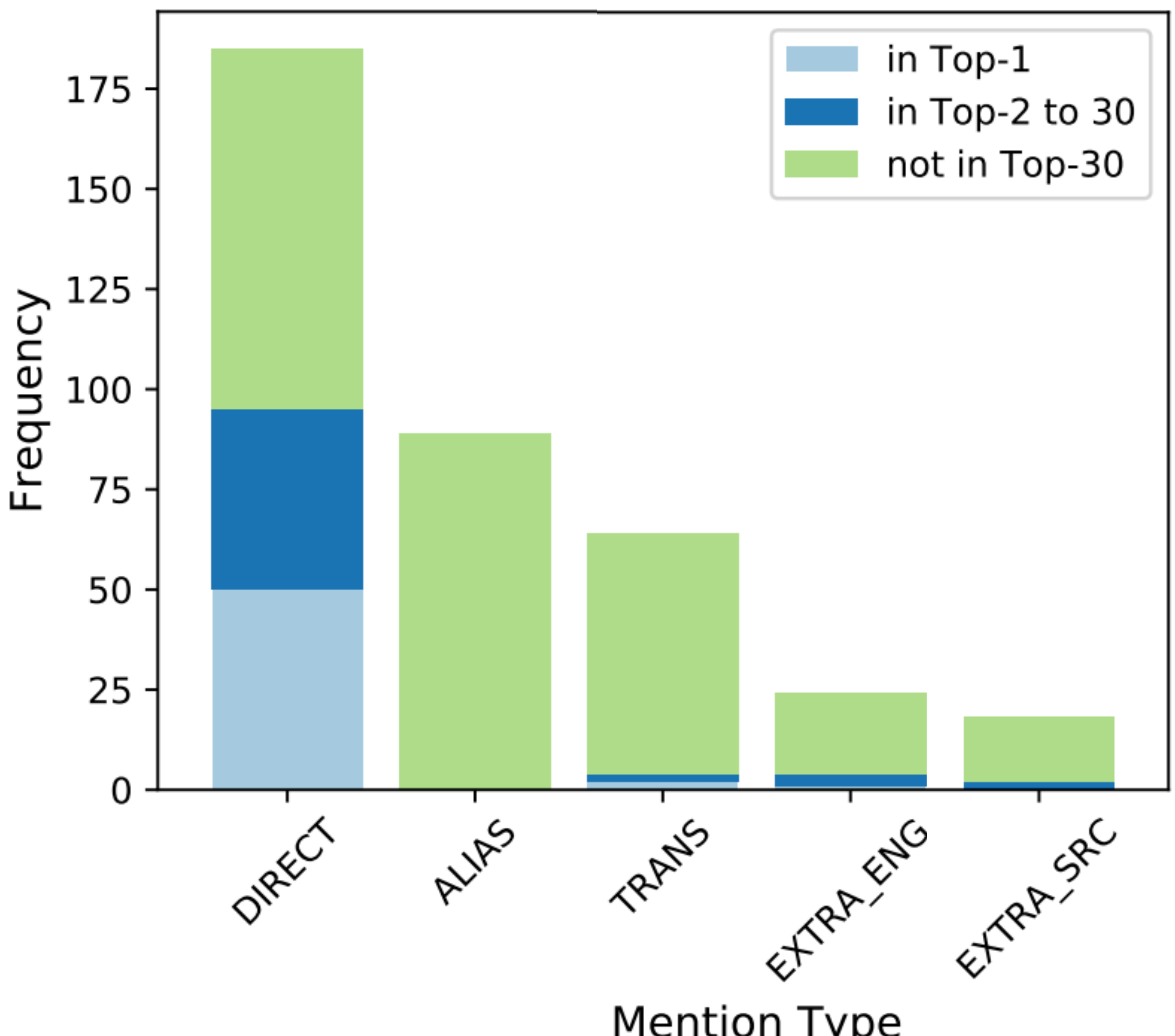
Common words in named entities like "university" are often **translated**

Examples of Errors

	Mention in Marathi	English KB Entry
DIRECT	कोबी स्मल्डर्स *Cobie Smulders*	Cobie Smulders
ALIAS	जॅकोबा फ्रांसिस्का मरिया स्मल्डर्स *Jacoba Francisca Maria Smulders*	Cobie Smulders
TRANS	कार्नेगी मेलॉन विद्यापीठ *Carnegie Mellon Vidyaapeeth*	Carnegie Mellon University
EXTRA	श्री गजानन महाराज *Shri Gajanan Maharaj*	Gajanan Maharaj

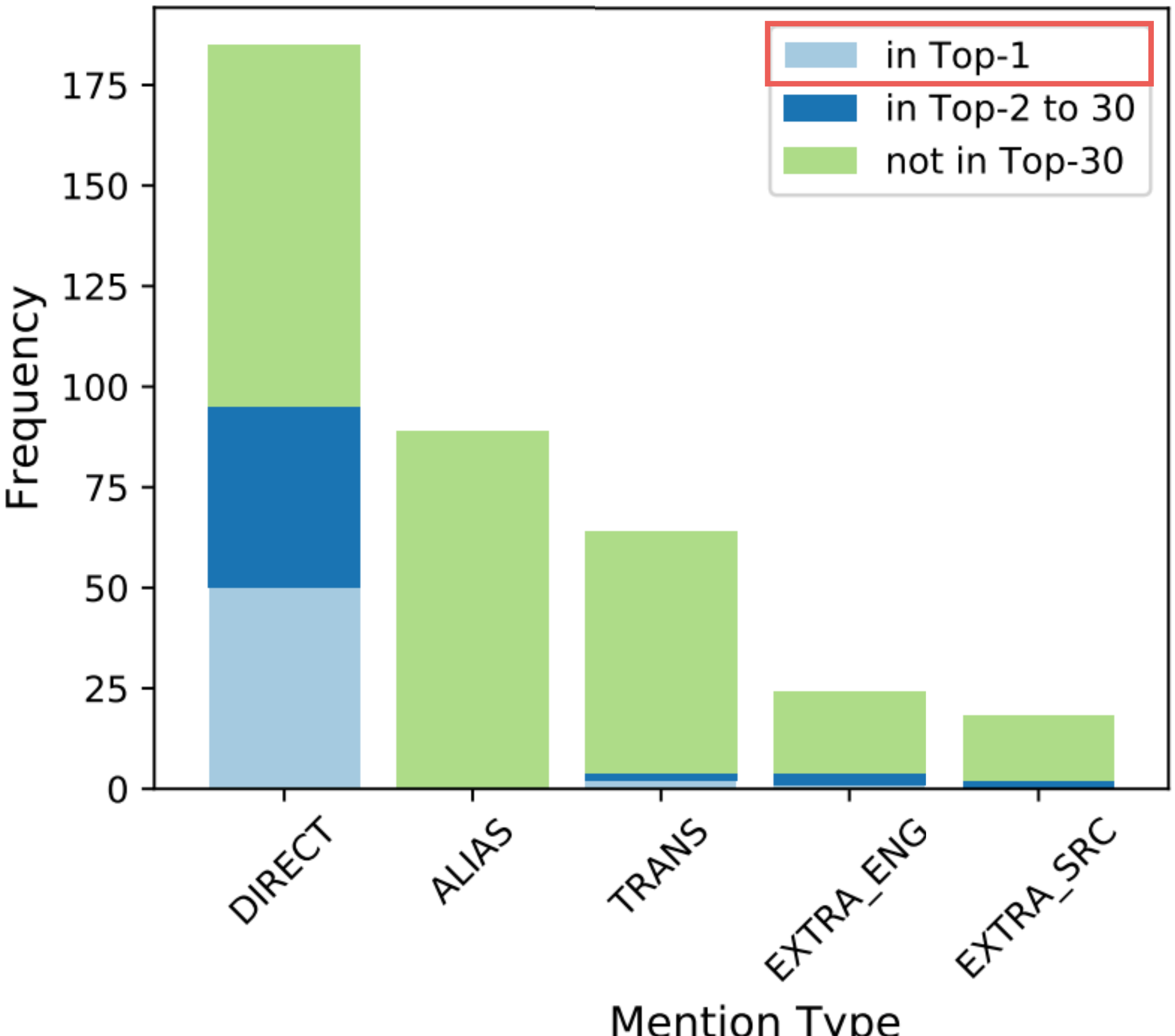
Mentions with **extra words** like honorifics

PBEL Error Distribution



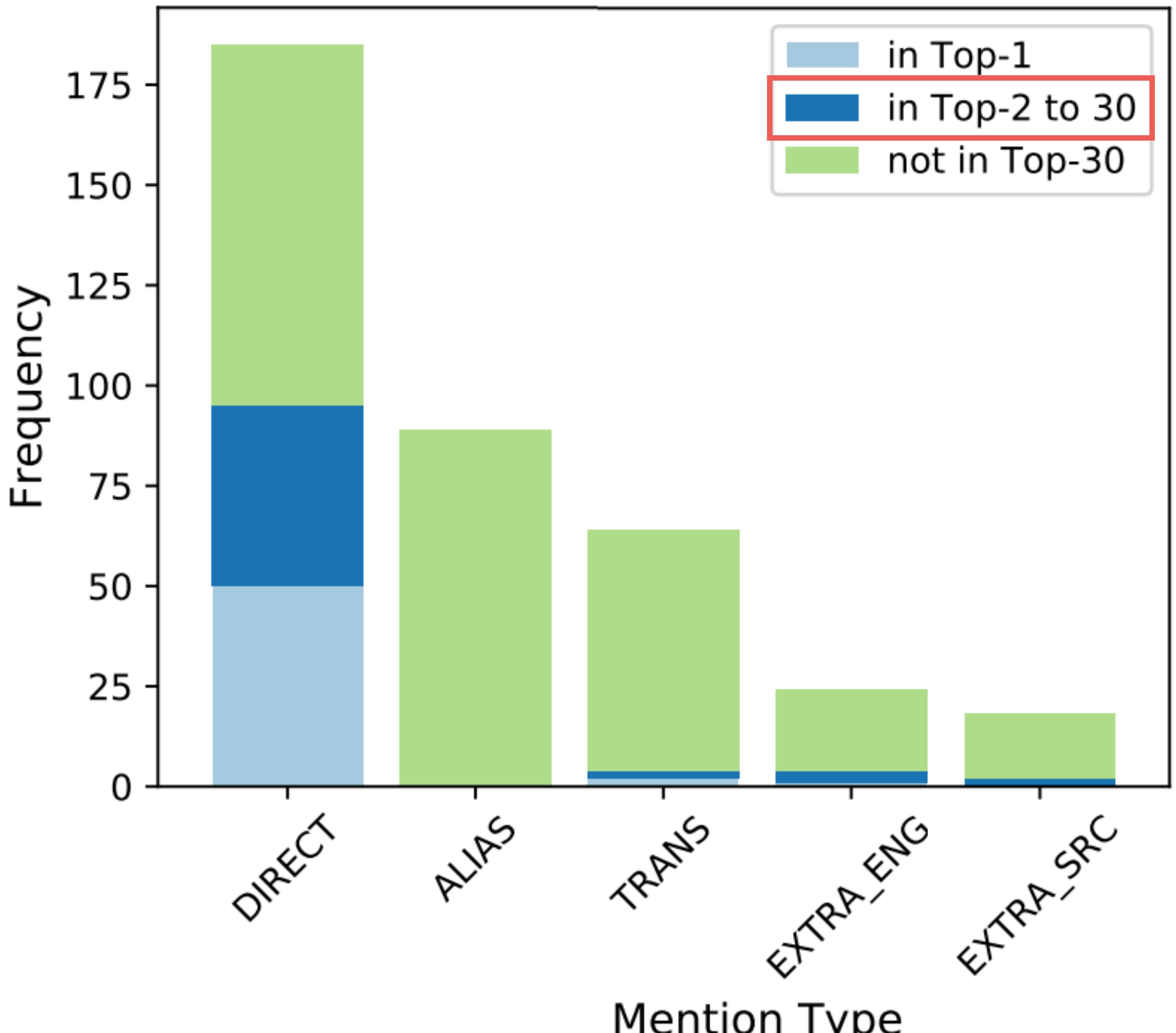
Position of the correct KB entry in candidate list

PBEL Error Distribution



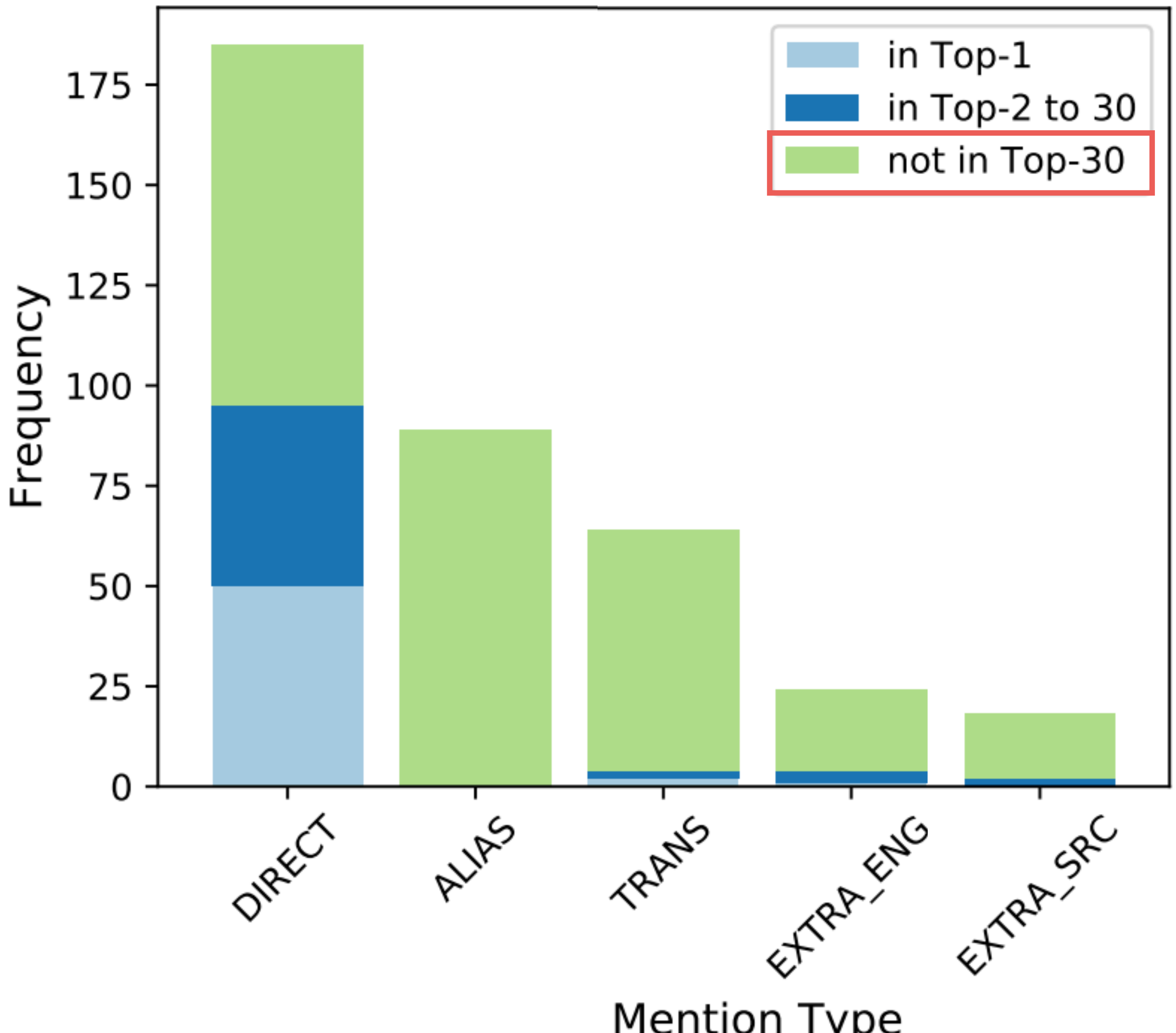
Position of the correct KB entry in candidate list

PBEL Error Distribution



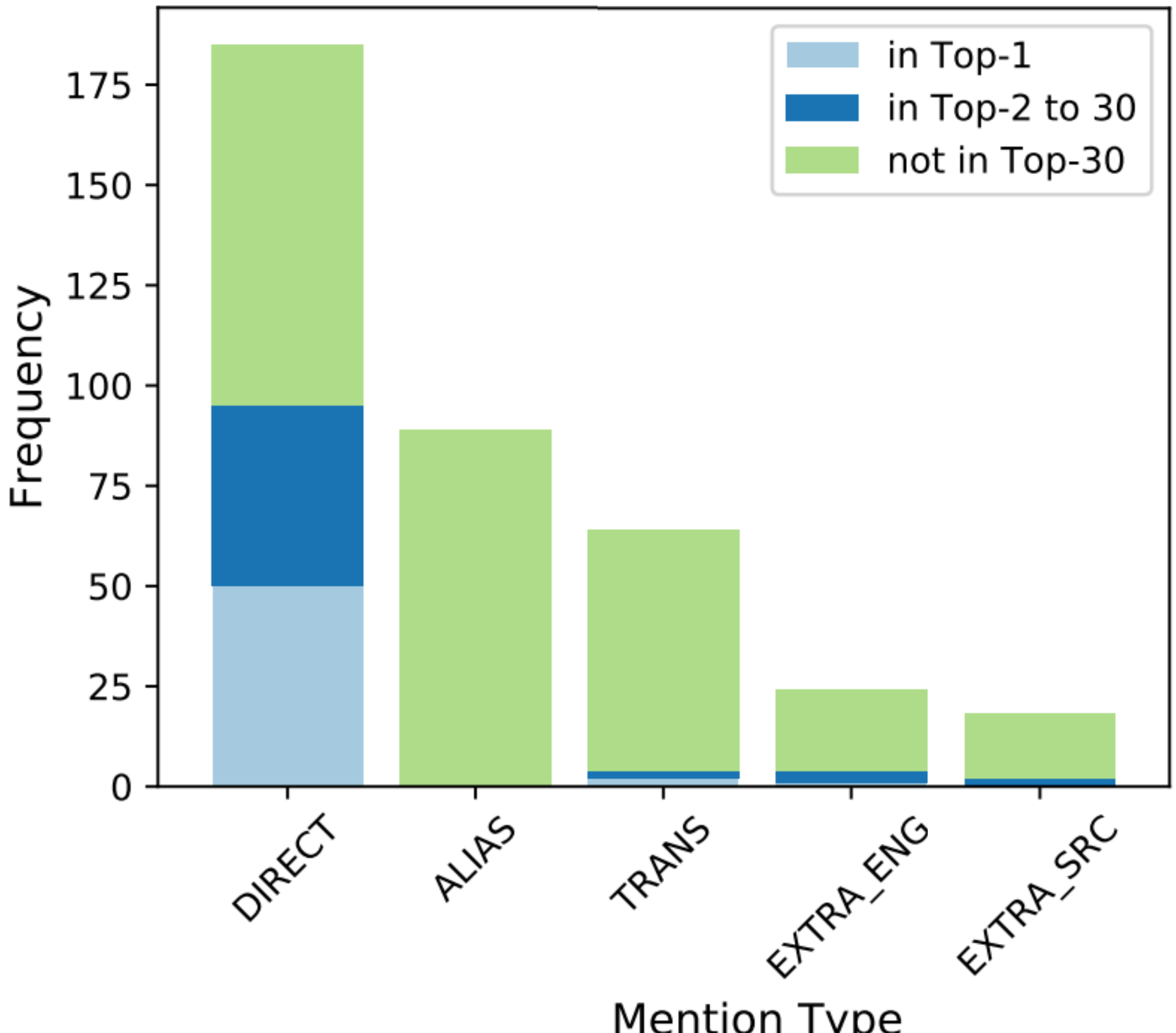
Position of the correct KB entry in candidate list

PBEL Error Distribution



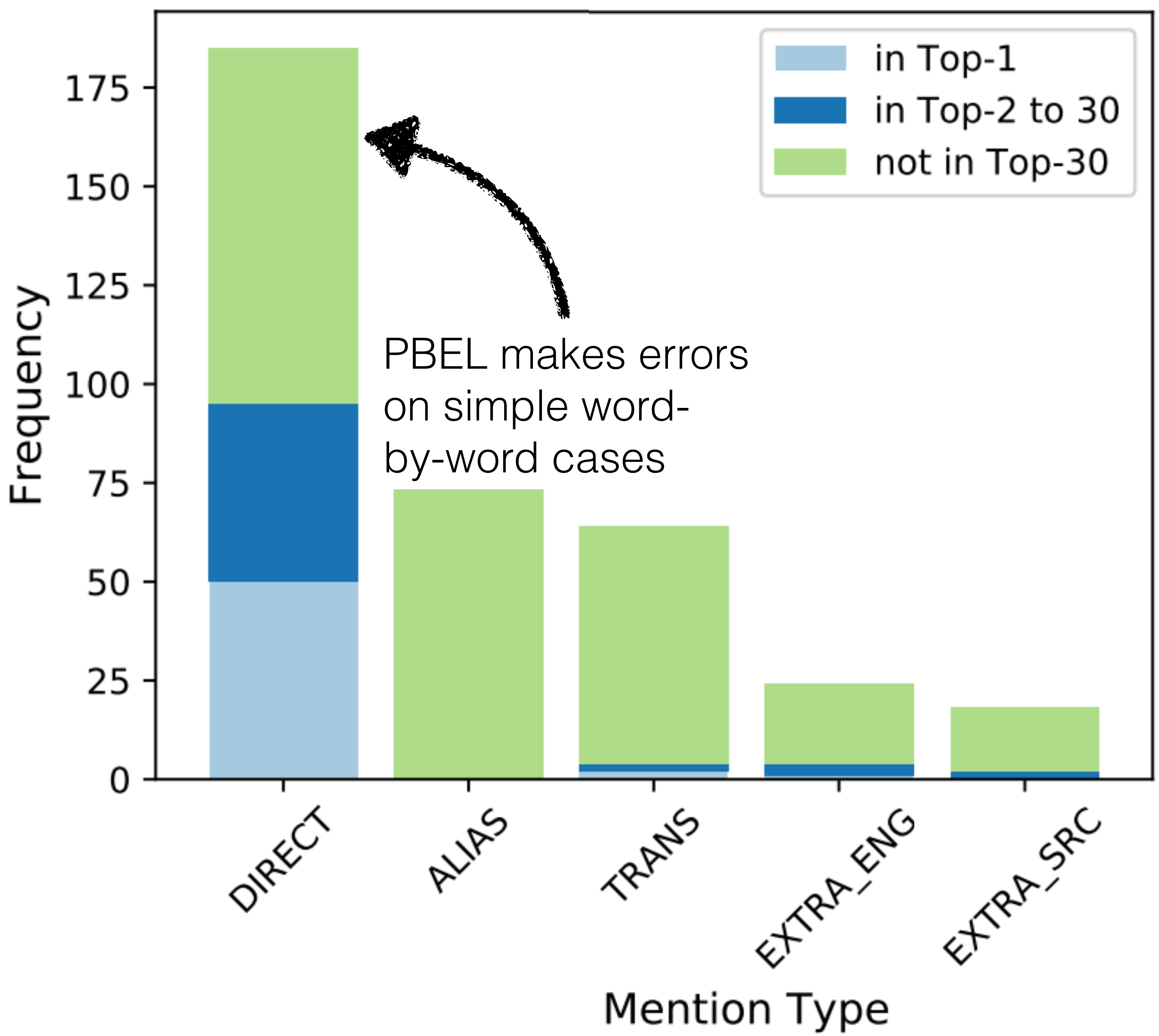
Position of the
correct KB entry
in candidate list

PBEL Error Distribution



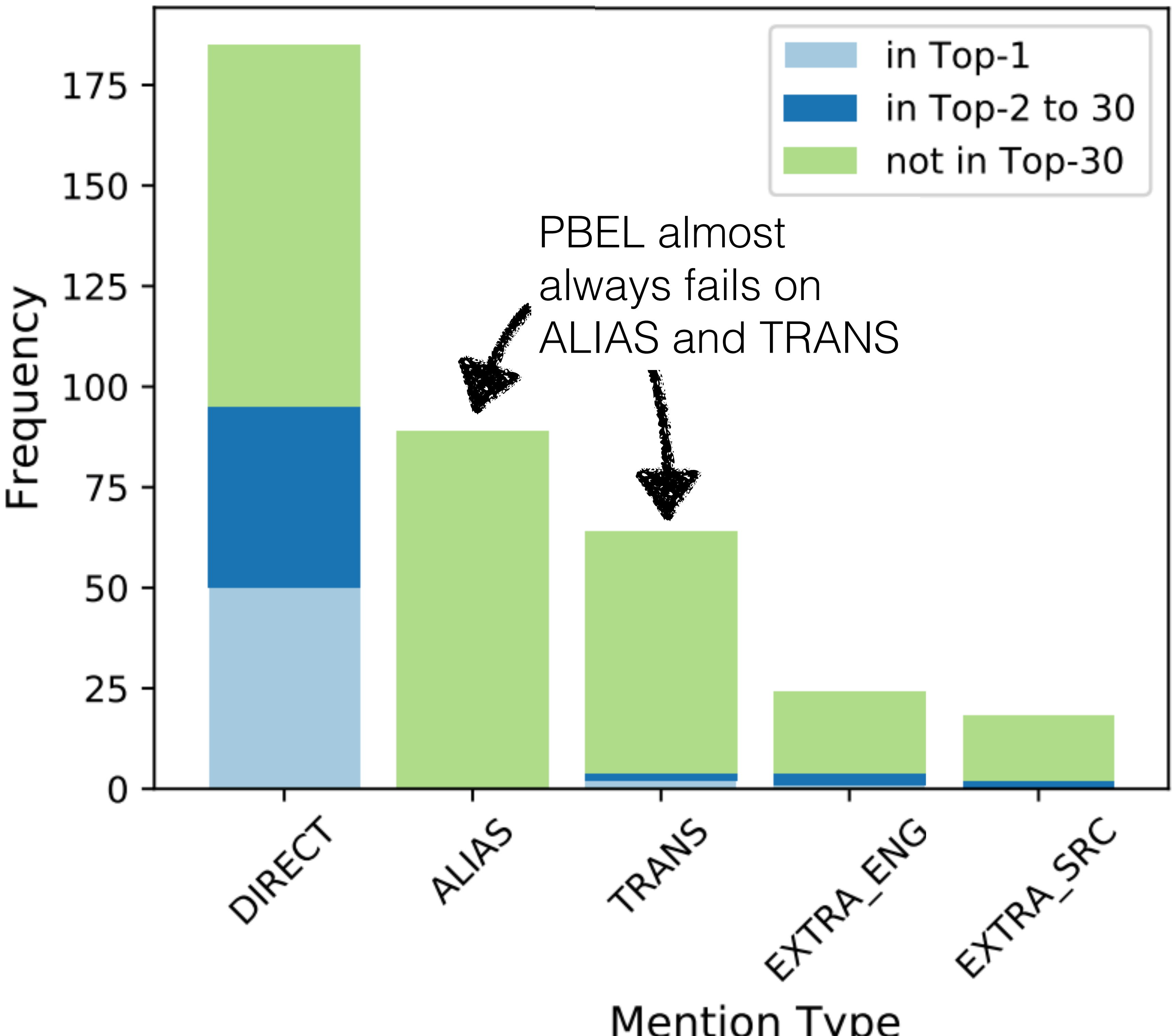
Position of the correct KB entry in candidate list

PBEL Error Distribution



Position of the correct KB entry in candidate list

PBEL Error Distribution



Position of the correct KB entry in candidate list

Eliminating Train-Test Discrepancy

Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**

Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**
    - PBEL training data has entity-entity pairs, usually word-by-word mappings

Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**
  - PBEL training data has entity-entity pairs, usually word-by-word mappings
  - Other types of mentions are not covered

Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**
  - PBEL training data has entity-entity pairs, usually word-by-word mappings
  - Other types of mentions are not covered
  - Example: Last name of a person entity

Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**
  - PBEL training data has entity-entity pairs, usually word-by-word mappings
  - Other types of mentions are not covered
  - Example: Last name of a person entity
- **Solution:** Add mention-entity pairs to the data

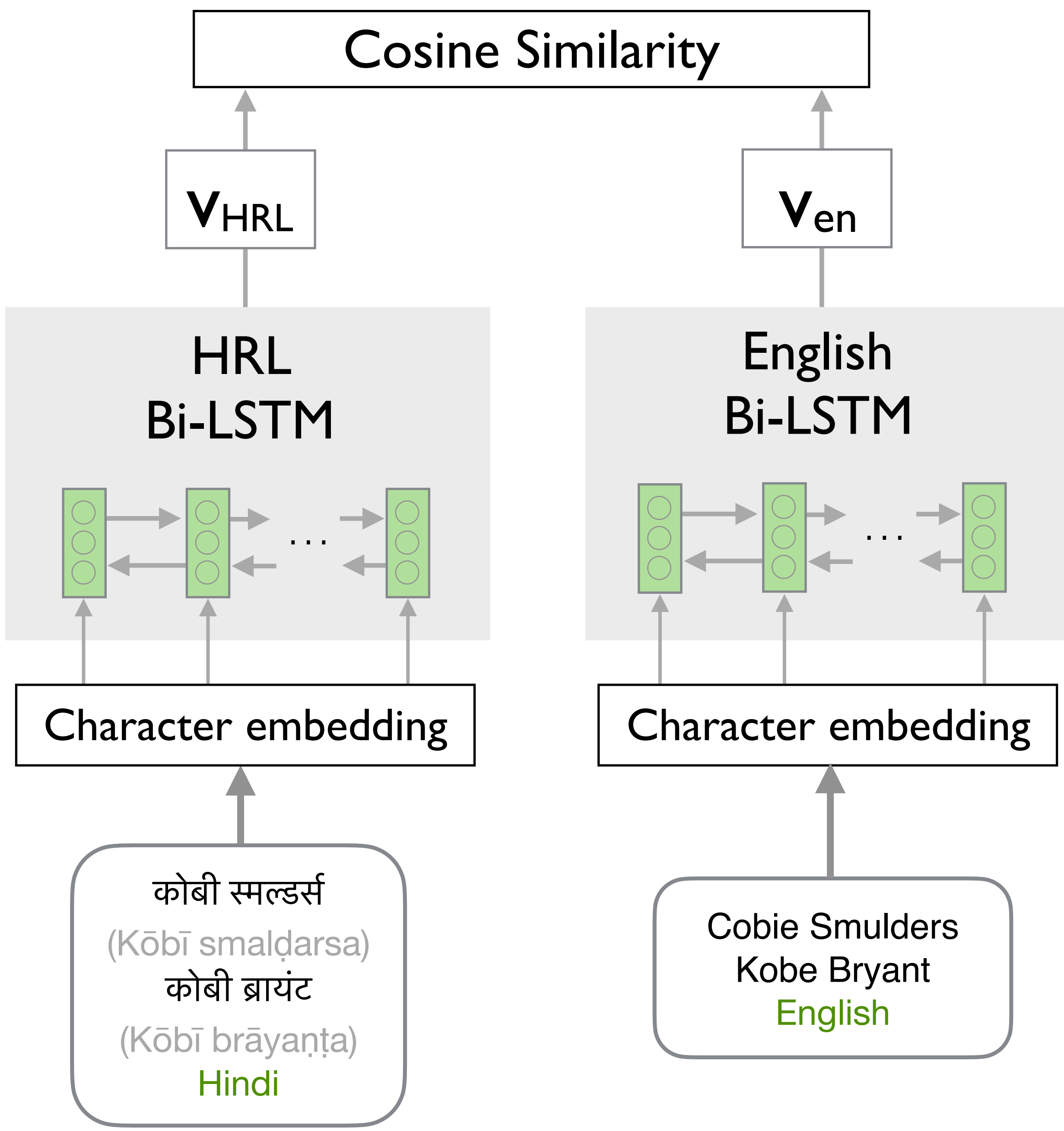
Eliminating Train-Test Discrepancy

- **Test mentions types are different from train**
  - PBEL training data has entity-entity pairs, usually word-by-word mappings
  - Other types of mentions are not covered
  - Example: Last name of a person entity
- **Solution:** Add mention-entity pairs to the data
  - From the high-resource language Wikipedia

Eliminating Train-Test Discrepancy

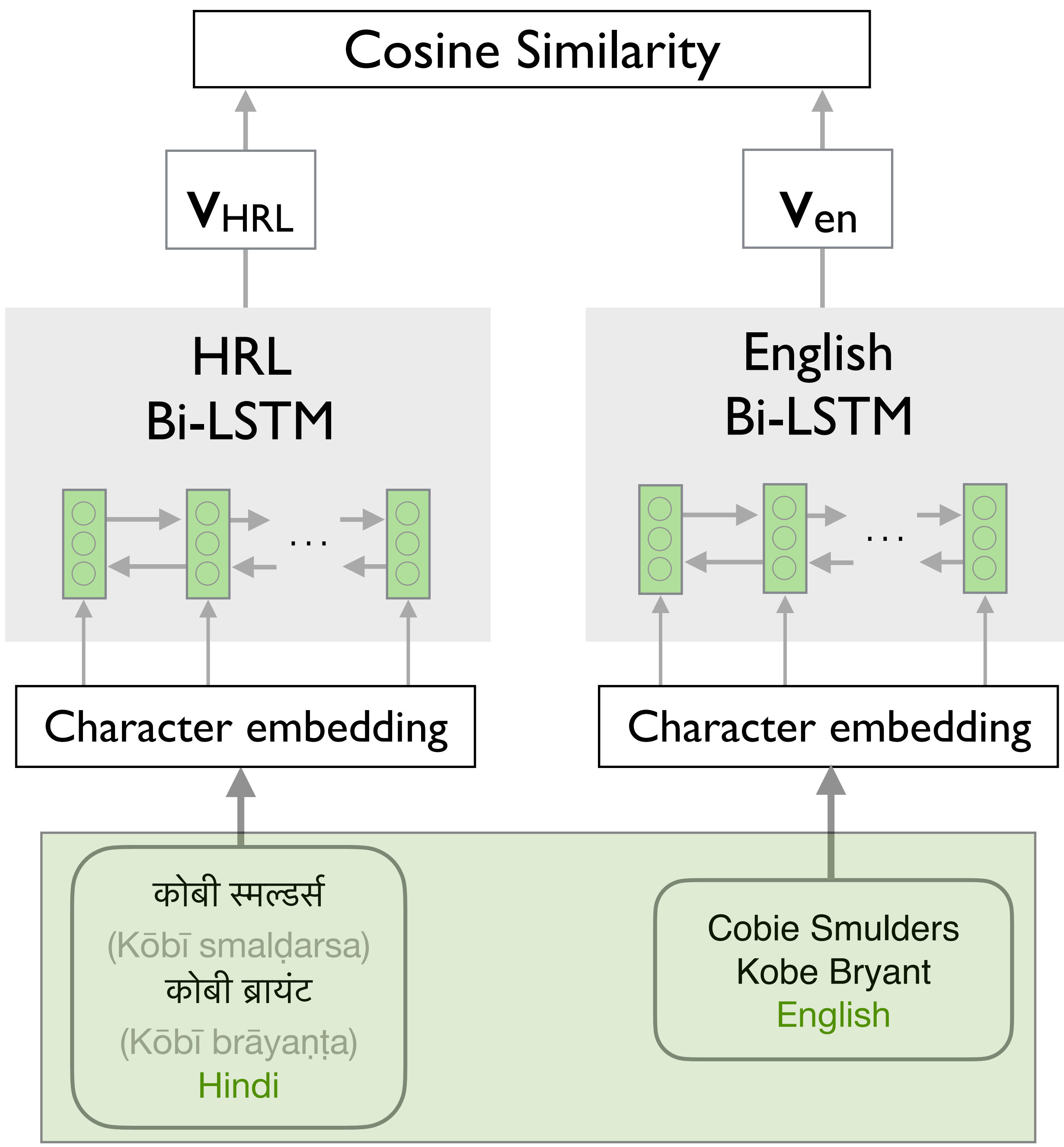
Train language: Hindi
Test language: Marathi

Eliminating Train-Test Discrepancy



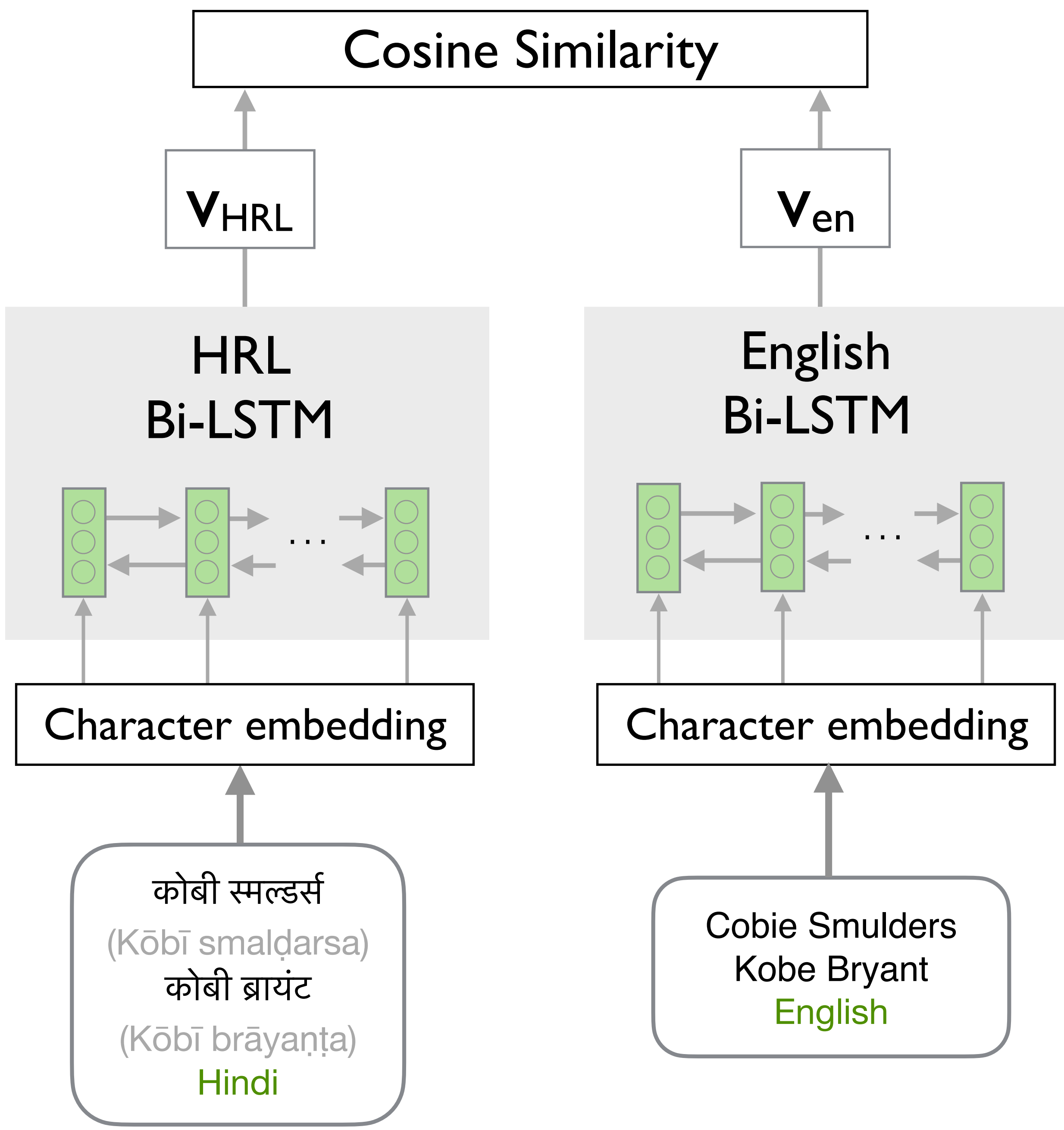
Train language: Hindi
Test language: Marathi

Eliminating Train-Test Discrepancy

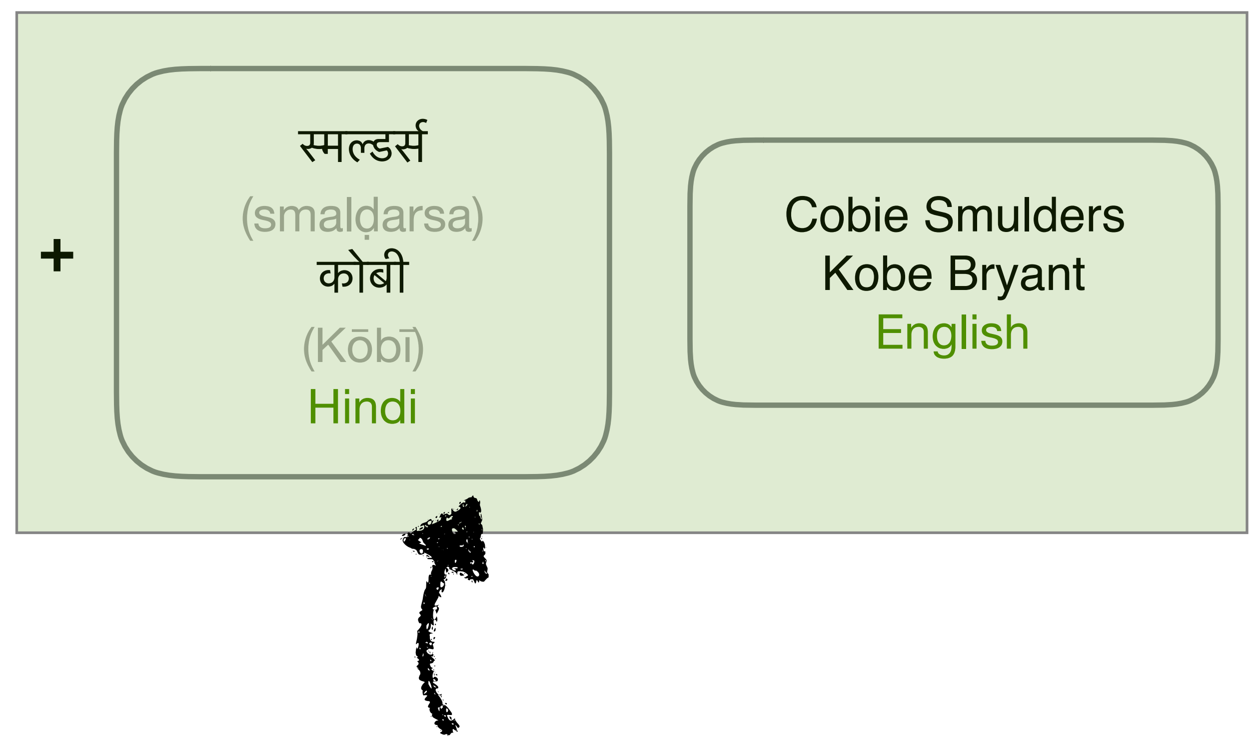


Train language: Hindi
Test language: Marathi

Eliminating Train-Test Discrepancy



Train language: Hindi
Test language: Marathi



Add **mention—entity** pairs to the training data

Utilizing English Entity Aliases

Utilizing English Entity Aliases

- **PBEL cannot match aliases** because it is only trained on the most common name for an entity.

Utilizing English Entity Aliases

- **PBEL cannot match aliases** because it is only trained on the most common name for an entity.
- **Solution:** add English aliases from Wikidata

Utilizing English Entity Aliases

- **PBEL cannot match aliases** because it is only trained on the most common name for an entity.
- **Solution:** add English aliases from Wikidata

Cobie Smulders **(Q200566)**

Also known as

Jacoba Francisca Maria Sm…
Jacoba Francisca Maria Sm…
Smulders, Cobie

Jacoba Francisca Maria Sm…
Jacoba Francisca Maria Sm…

WIKIDATA

More Explicit String Encoder

More Explicit String Encoder

- PBEL makes **errors on simple DIRECT** mentions

More Explicit String Encoder

- PBEL makes **errors on simple DIRECT** mentions
    - BiLSTM representations are likely not optimal

More Explicit String Encoder

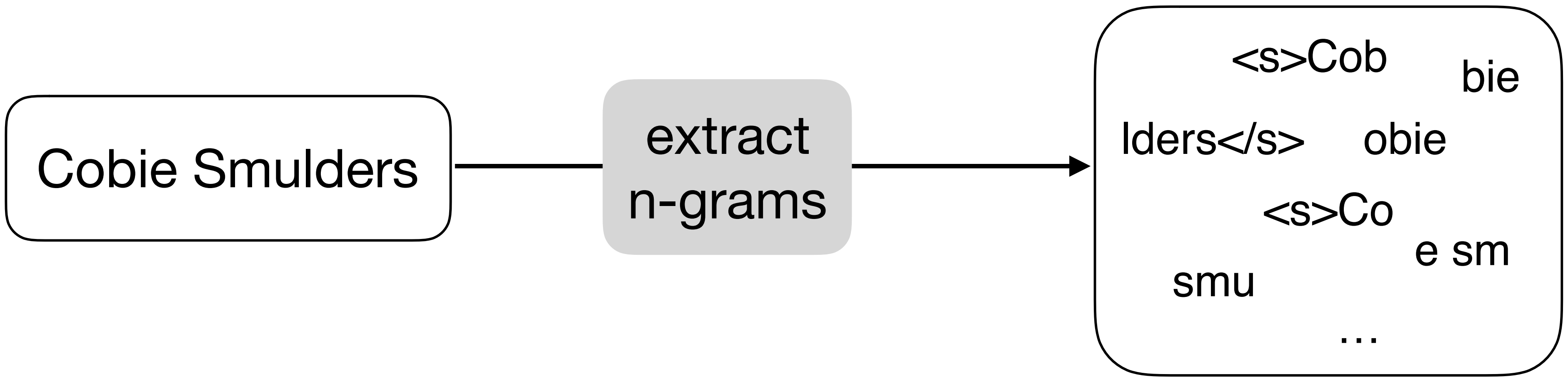
- PBEL makes **errors on simple DIRECT** mentions
  - BiLSTM representations are likely not optimal
- **Solution:** Replace the BiLSTM with character n-gram encoder (charagram)

More Explicit String Encoder

- **Solution:** Replace the BiLSTM with character n-gram encoder (charagram)

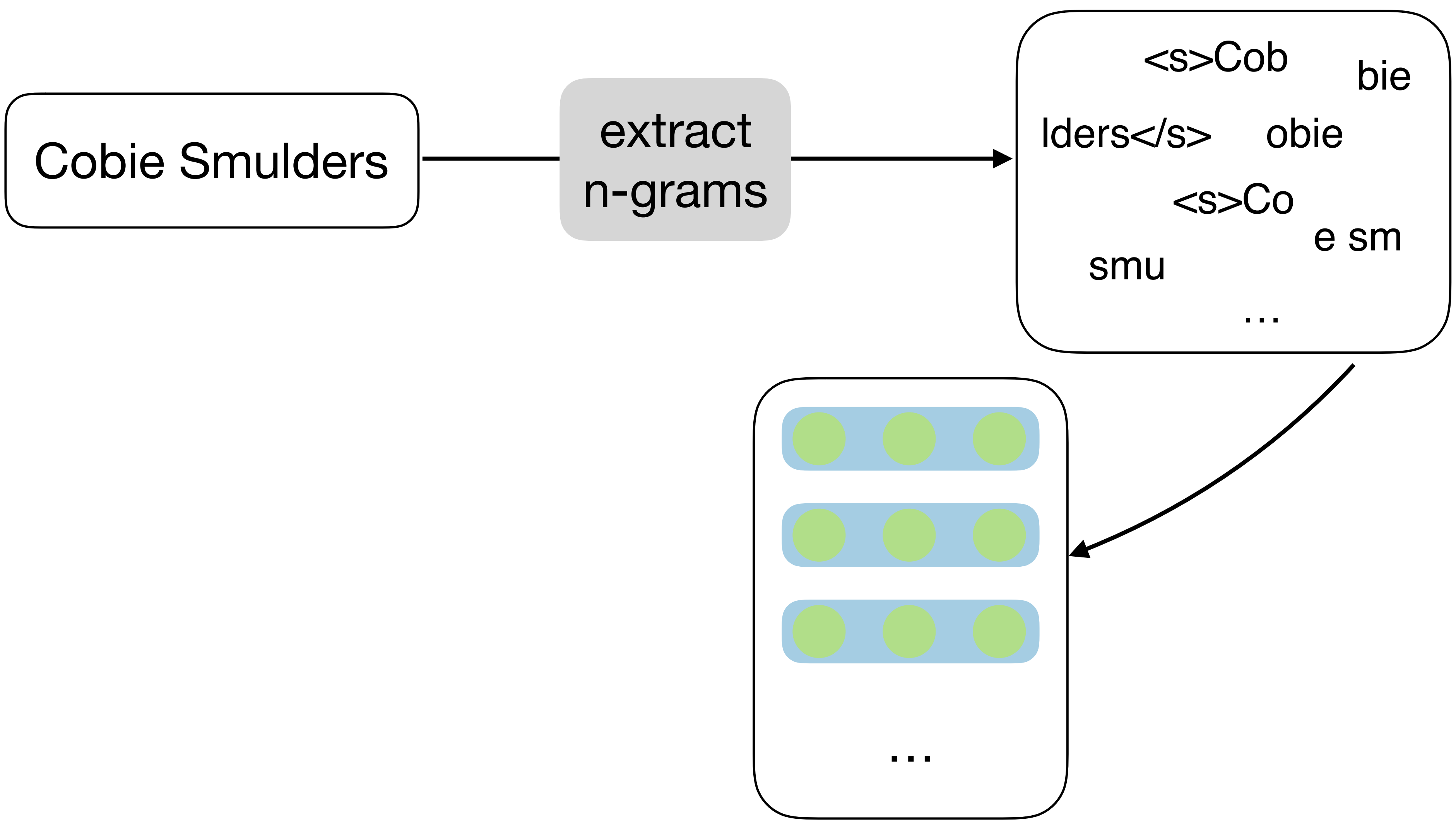
More Explicit String Encoder

- **Solution:** Replace the BiLSTM with character n-gram encoder (charagram)



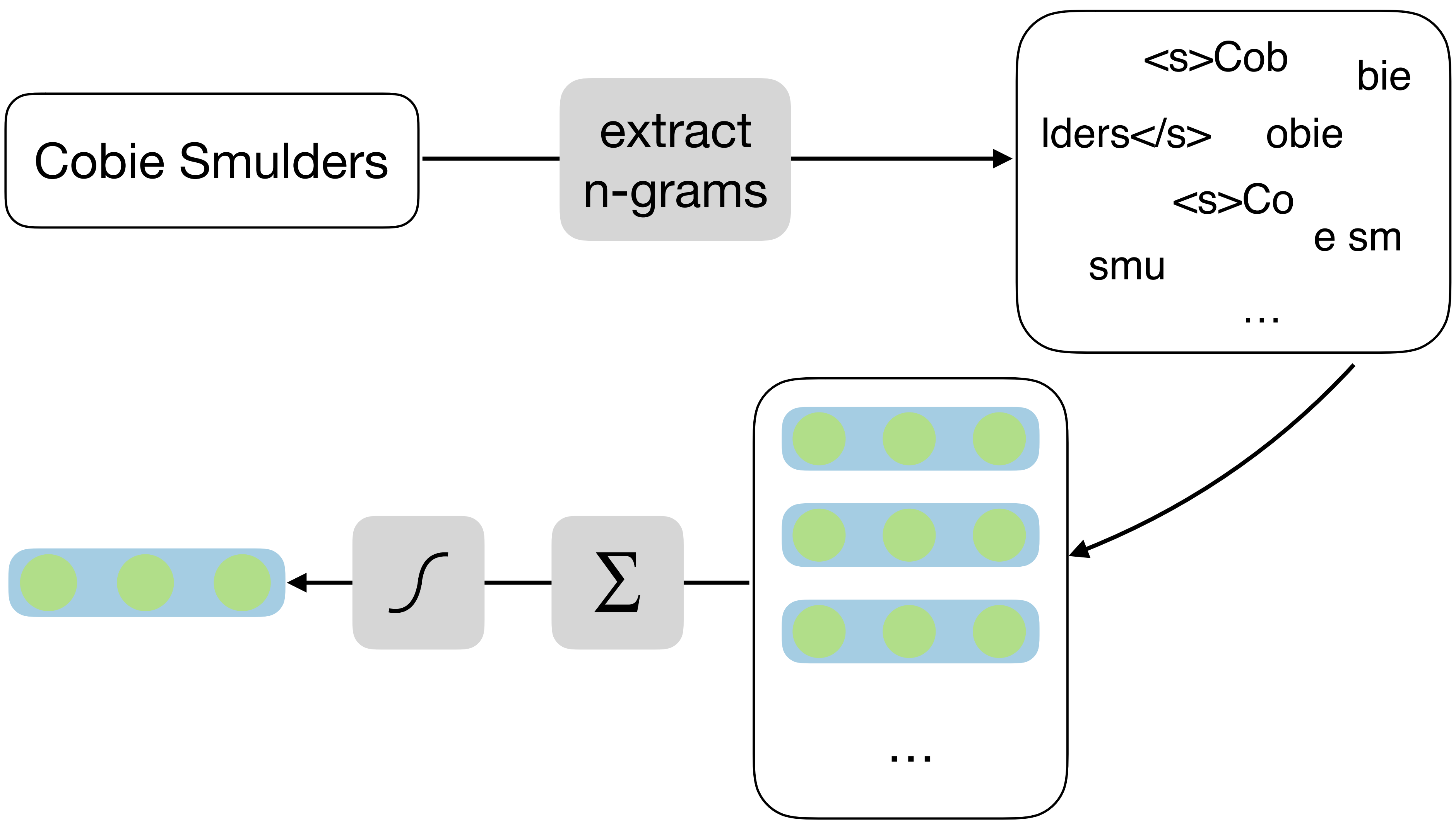
More Explicit String Encoder

- **Solution:** Replace the BiLSTM with character n-gram encoder (charagram)



More Explicit String Encoder

- **Solution:** Replace the BiLSTM with character n-gram encoder (charagram)



Candidate Retrieval with PBEL+

Candidate Retrieval with PBEL+

- With the core idea of the pivoting method, we add modifications based on the error analysis

Candidate Retrieval with PBEL+

- With the core idea of the pivoting method, we add modifications based on the error analysis
  - Eliminating the train-test discrepancy by **adding mention—entity pairs** to the data

Candidate Retrieval with PBEL+

- With the core idea of the pivoting method, we add modifications based on the error analysis
  - Eliminating the train-test discrepancy by **adding mention—entity pairs** to the data
  - Using **Wikidata aliases**

Candidate Retrieval with PBEL+

- With the core idea of the pivoting method, we add modifications based on the error analysis
  - Eliminating the train-test discrepancy by **adding mention—entity pairs** to the data
  - Using **Wikidata aliases**
  - **Replacing the LSTMs with charagram**, a more explicit string encoder.

Candidate Retrieval with PBEL+

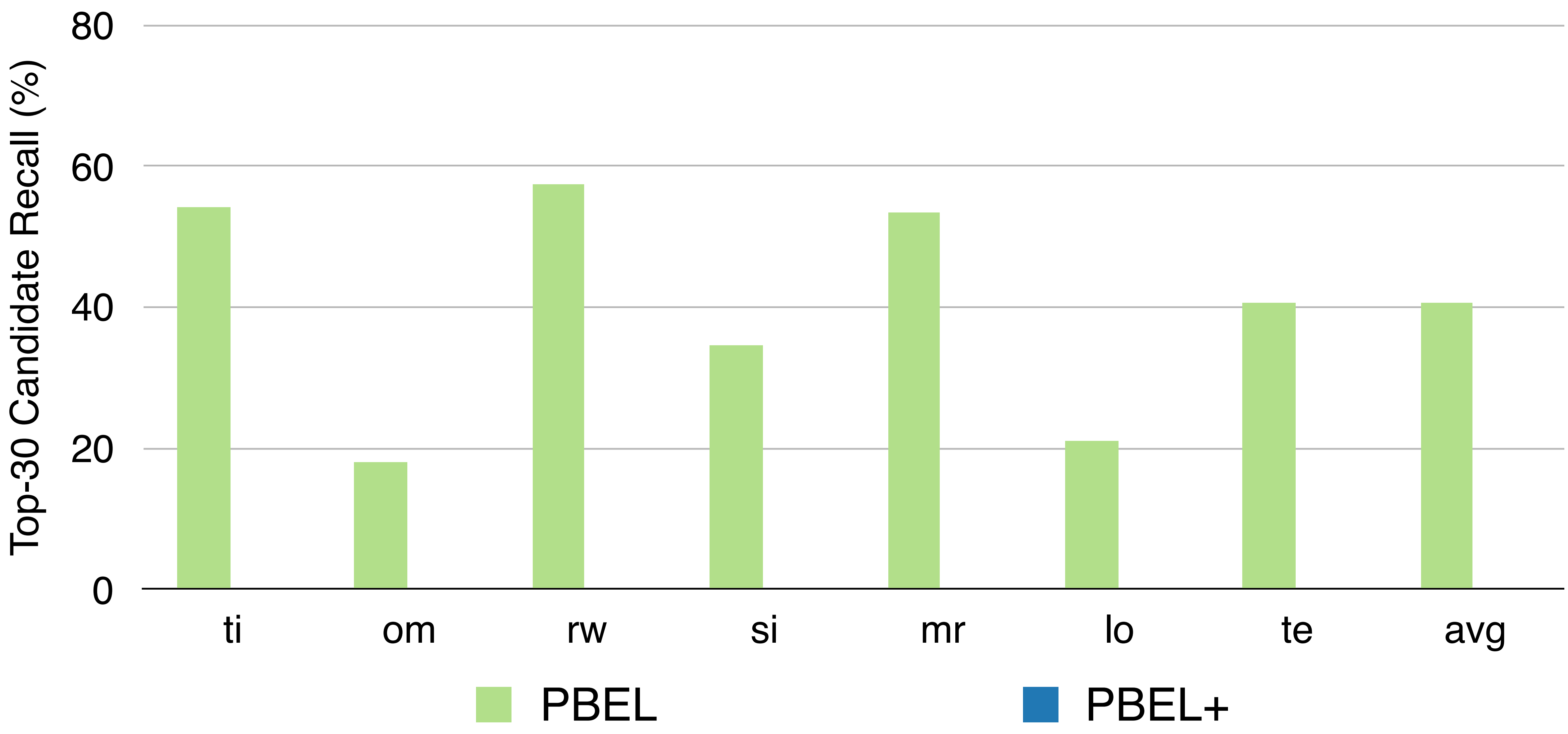
- With the core idea of the pivoting method, we add modifications based on the error analysis

Trained only on high-resource language data and the **zero-shot transfer and pivoting techniques** are still applied!

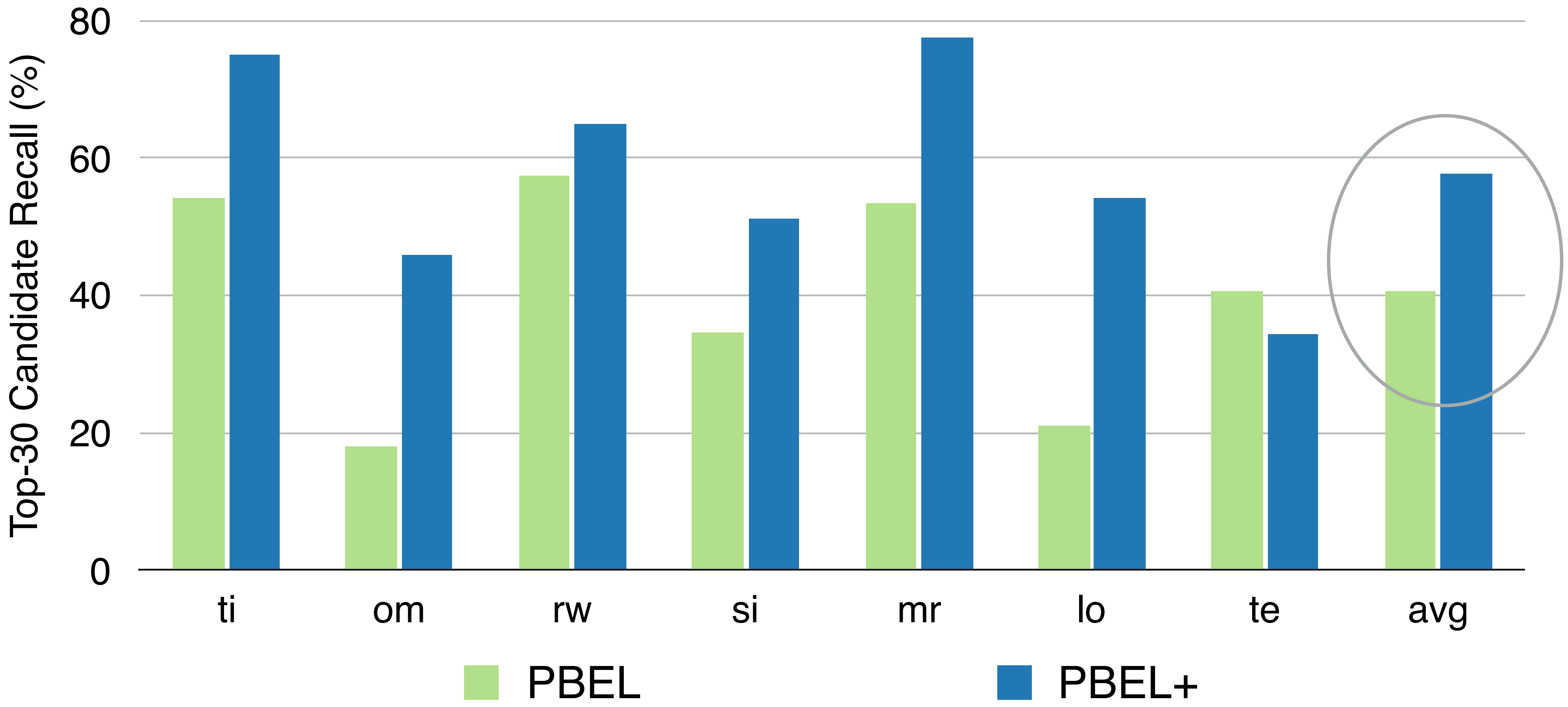
- **Replacing the LSTMs with charagram**, a more explicit string encoder.

Experiments: Top-30 Candidate Recall

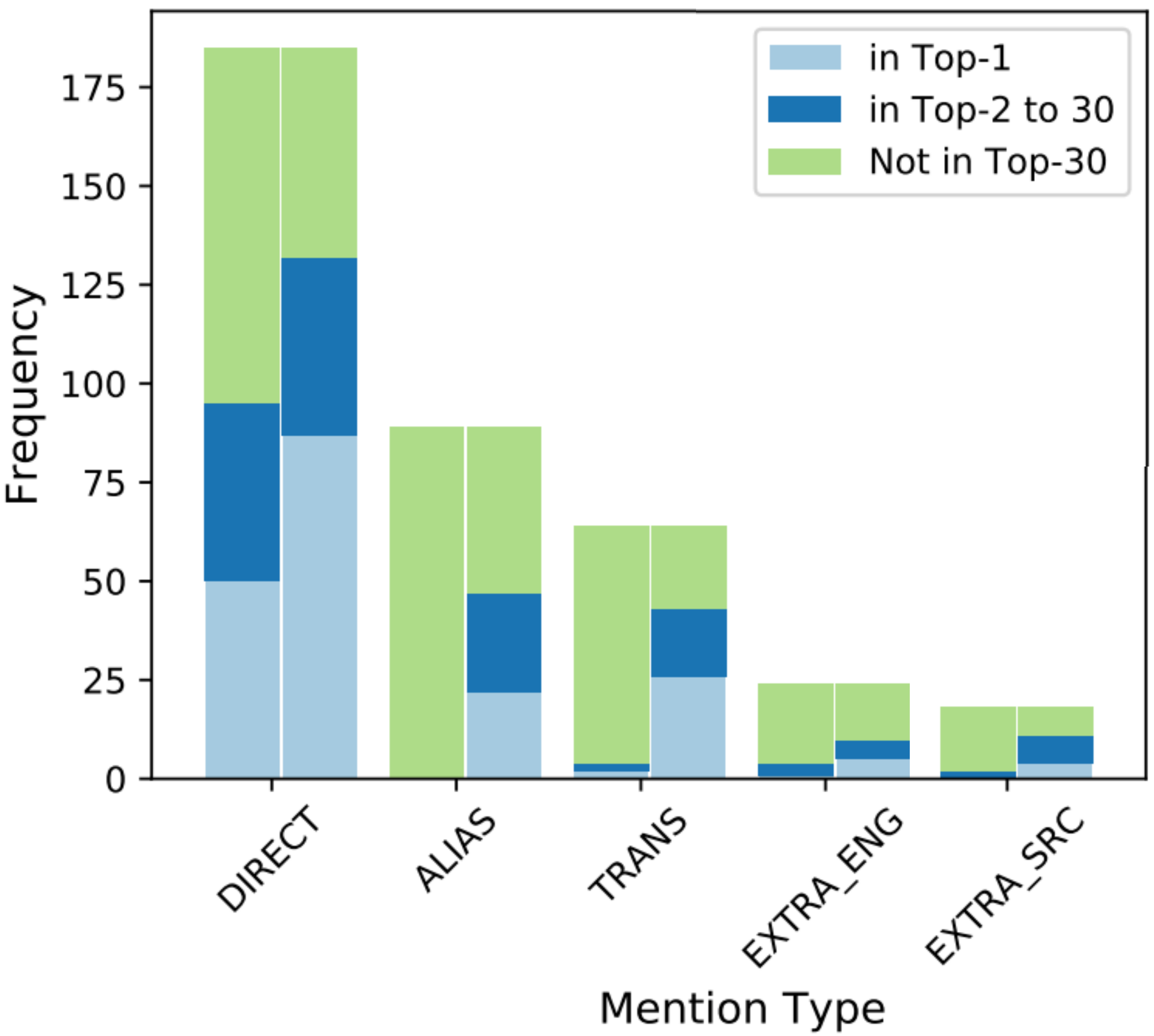
Experiments: Top-30 Candidate Recall



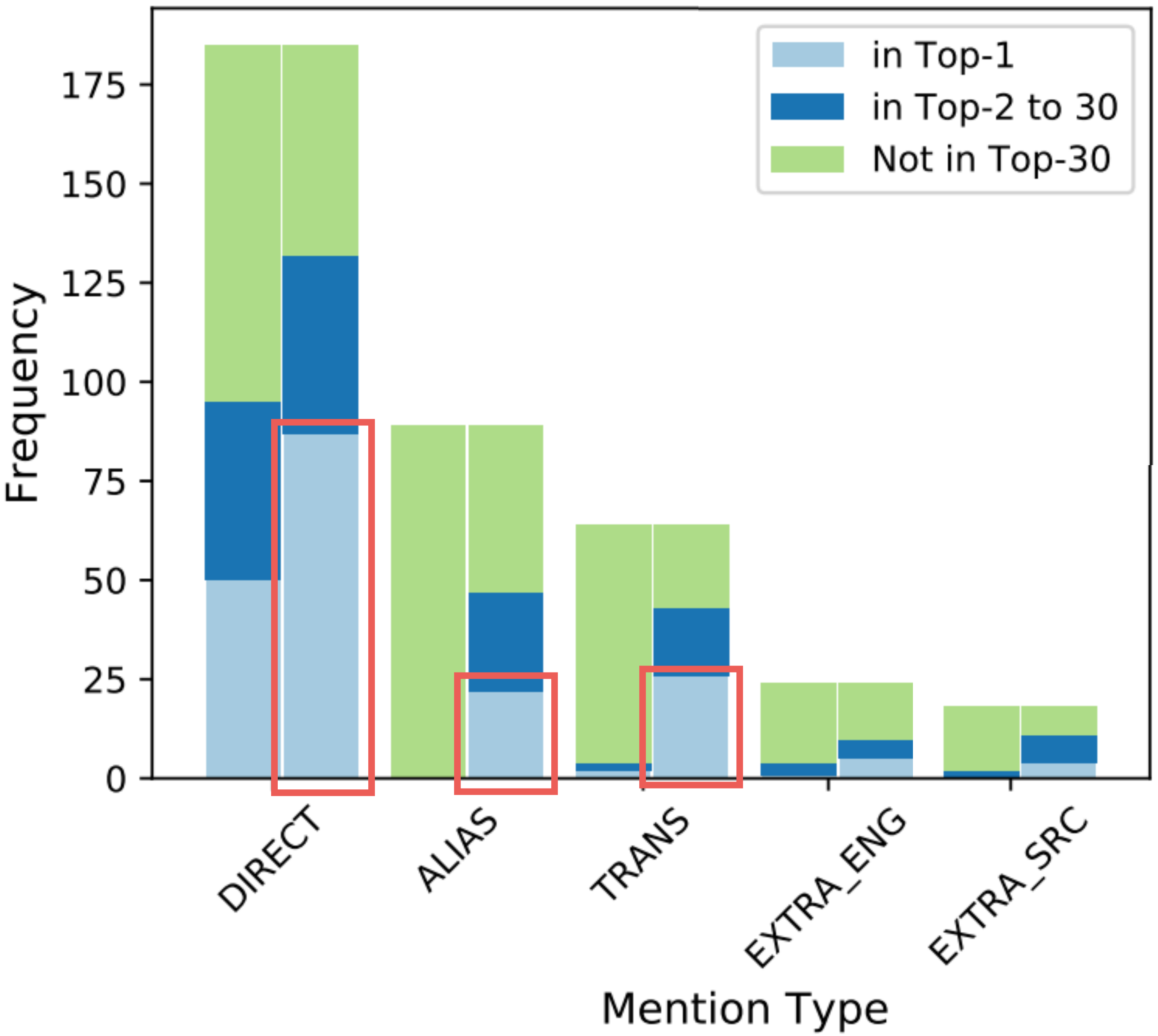
Experiments: Top-30 Candidate Recall



New Error Distribution

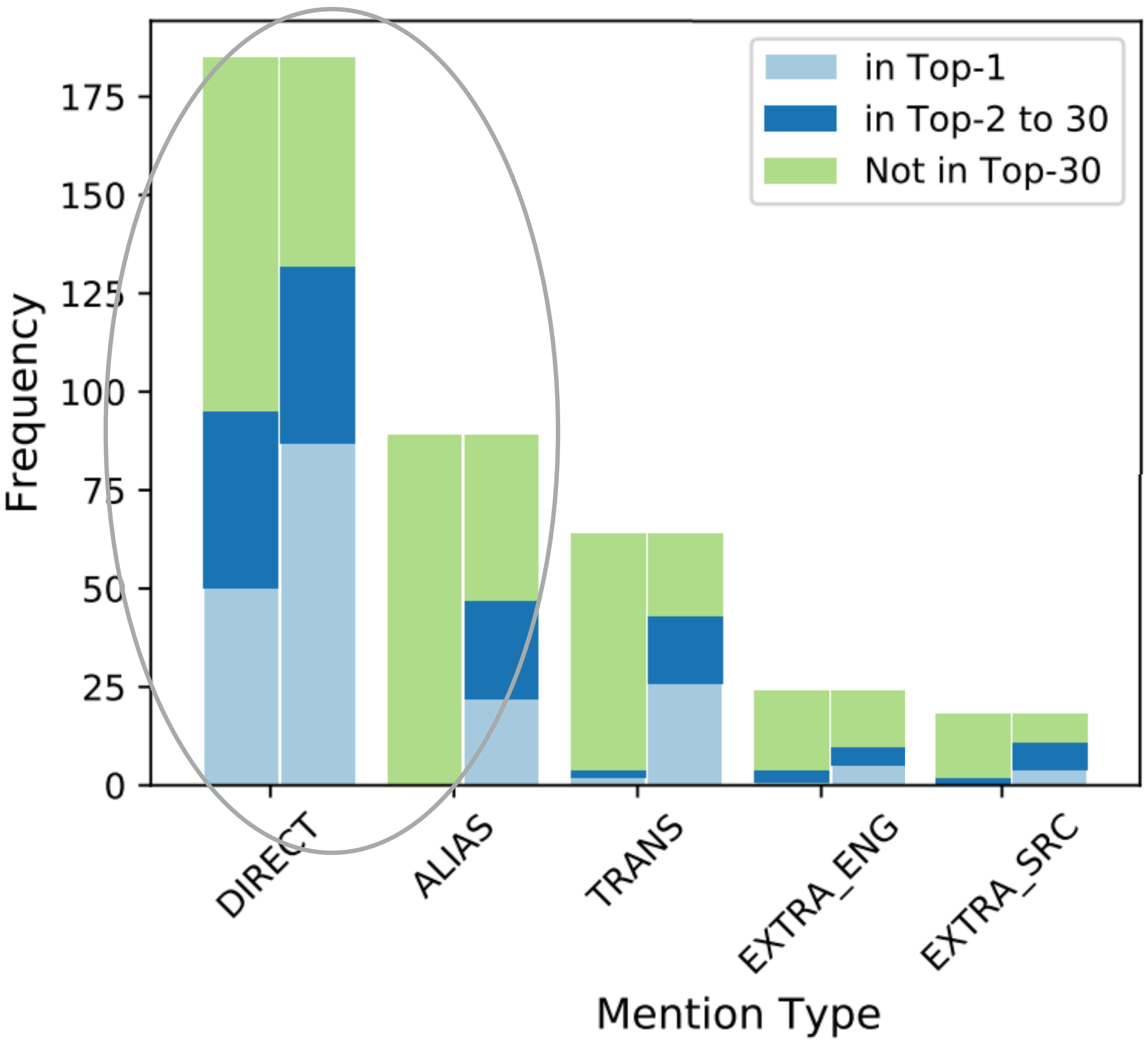


New Error Distribution



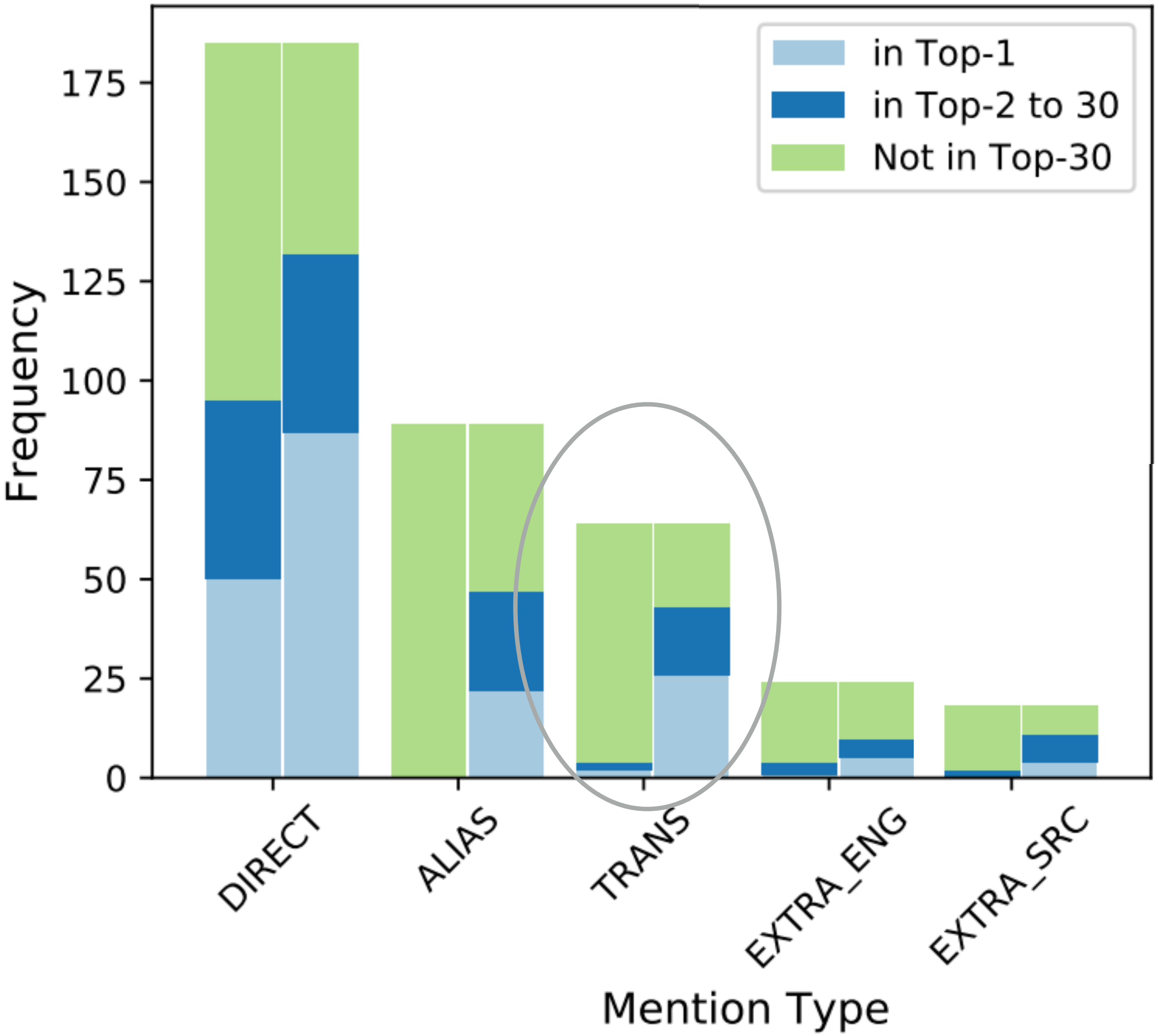
- Eliminates a large number of errors by ranking the correct entries the highest

New Error Distribution



- Eliminates a large number of errors by ranking the correct entries the highest
- Significant improvement on DIRECT and ALIAS
- Provides the downstream disambiguation model with **larger improvement headroom**

New Error Distribution



- Eliminates a large number of errors by ranking the correct entries the highest
- Significant improvement on DIRECT and ALIAS
- Provides the downstream disambiguation model with **larger improvement headroom**
- A number of TRANS errors are resolved as well

Candidate Retrieval with PBEL+

Candidate Retrieval with PBEL+

- A **systematic error analysis** over the zero-shot candidate retrieval model PBEL.

Candidate Retrieval with PBEL+

- A **systematic error analysis** over the zero-shot candidate retrieval model PBEL.
- Modifications to **handle the diverse realizations** of a mention and the unique identification of an entity.

Candidate Retrieval with PBEL+

- A **systematic error analysis** over the zero-shot candidate retrieval model PBEL.
- Modifications to **handle the diverse realizations** of a mention and the unique identification of an entity.
- A **better modeling strategy for strings** (the character n-gram model charagram).

Candidate Retrieval with PBEL+

- A **systematic error analysis** over the zero-shot candidate retrieval model PBEL.
- Modifications to **handle the diverse realizations** of a mention and the unique identification of an entity.
- A **better modeling strategy for strings** (the character n-gram model charagram).
- **Improves candidate recall by nearly 20 points** on average over seven low-resource languages.

Named Entity Recognition

Named Entity Recognition

Identify **named entities** and their **types**.

Named Entity Recognition

Identify **named entities** and their **types**.

Mark Watney visited Mars.

PER LOC

Named Entity Recognition

Identify **named entities** and their **types**.

Mark Watney visited Mars.

PER LOC

- F1 on English > 93%. What about low-resource languages?

Named Entity Recognition

Identify **named entities** and their **types**.

Mark Watney visited Mars.

PER LOC

- F1 on English > 93%. What about low-resource languages?
- Number of labeled sentences in low-resource datasets is less than 10% of CoNLL 2003 (English).

Named Entity Recognition

Identify **named entities** and their **types**.

Mark Watney visited Mars.

PER LOC

- F1 on English > 93%. What about low-resource languages?
- Number of labeled sentences in low-resource datasets is less than 10% of CoNLL 2003 (English).
- **Can we use data from English knowledge bases as supplemental information to improve NER models?**

Gazetteer Features for NER

Gazetteer Features for NER

- Before neural networks, named entity recognition systems used **linguistic features to improve performance.**

Gazetteer Features for NER

- Before neural networks, named entity recognition systems used **linguistic features to improve performance.**

- Integrating **handcrafted features with neural models is useful** for NER on English text (Wu et al., 2018).

Gazetteer Features for NER

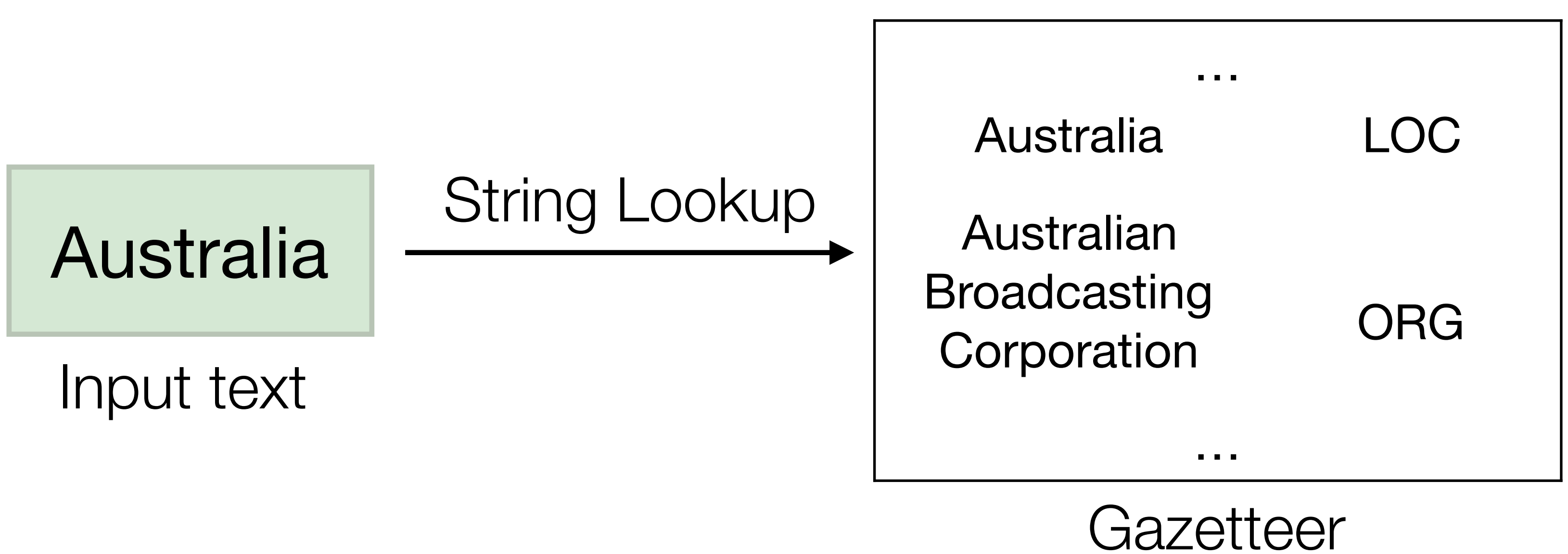
- Before neural networks, named entity recognition systems used **linguistic features to improve performance.**

- Integrating **handcrafted features with neural models is useful** for NER on English text (Wu et al., 2018).

- **Gazetteer features:** constructed with a string lookup in a list of entities called a gazetteer.

Gazetteer Features for NER

- Before neural networks, named entity recognition systems used **linguistic features to improve performance.**
- Integrating **handcrafted features with neural models is useful** for NER on English text (Wu et al., 2018).
- **Gazetteer features:** constructed with a string lookup in a list of entities called a gazetteer.

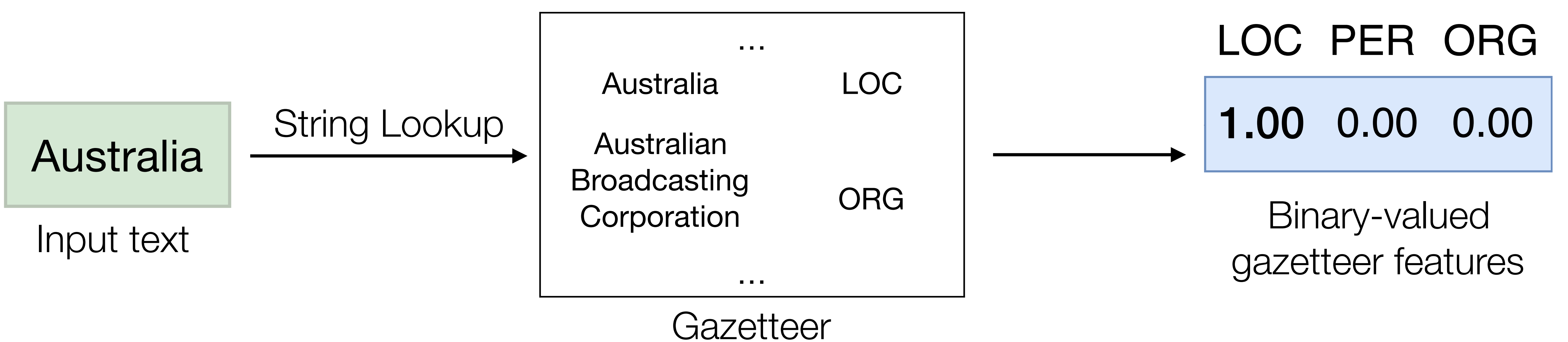


Gazetteer Features for NER

- Before neural networks, named entity recognition systems used **linguistic features to improve performance.**

- Integrating **handcrafted features with neural models is useful** for NER on English text (Wu et al., 2018).

- **Gazetteer features:** constructed with a string lookup in a list of entities called a gazetteer.



Low-Resource Challenges

Low-Resource Challenges

- Gazetteers are **limited in low-resource languages.**

Low-Resource Challenges

- Gazetteers are **limited in low-resource languages.**

Lang.	Entities in Wikipedia
English	2 million

Low-Resource Challenges

- Gazetteers are **limited in low-resource languages.**

Lang.	**Entities in Wikipedia**
English	2 million
Kinyarwanda	912
Oromo	313
Tigrinya	92

Low resource languages have much smaller gazetteers!

Low-Resource Challenges

- Gazetteers are **limited in low-resource languages.**

Lang.	**Entities in Wikipedia**
English	2 million
Kinyarwanda	912
Oromo	313
Tigrinya	92

Low resource languages have much smaller gazetteers!

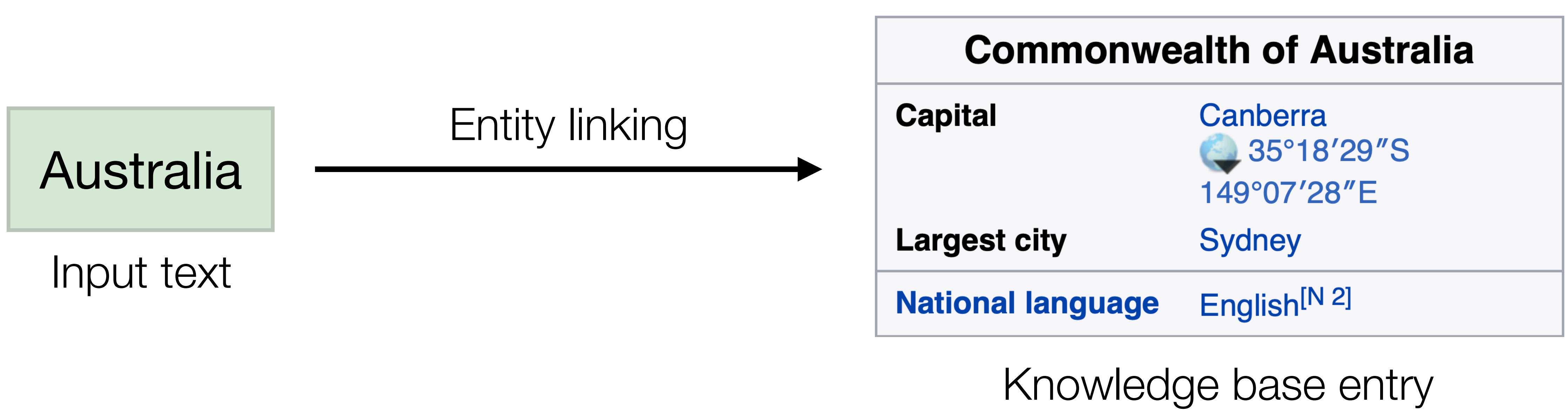
- Expanding gazetteers is **time-consuming and expensive**.
- **Finding annotators is difficult** for low-resource languages.

Soft Gazetteer Features

- Soft gazetteer features:
  - **Do not rely on large entity lists** in the target language.
  - Use readily available **English knowledge bases via entity linking** instead.

Soft Gazetteer Features

- Soft gazetteer features:
  - **Do not rely on large entity lists** in the target language.
  - Use readily available **English knowledge bases via entity linking** instead.



Soft Gazetteers: Method

Soft Gazetteers: Method

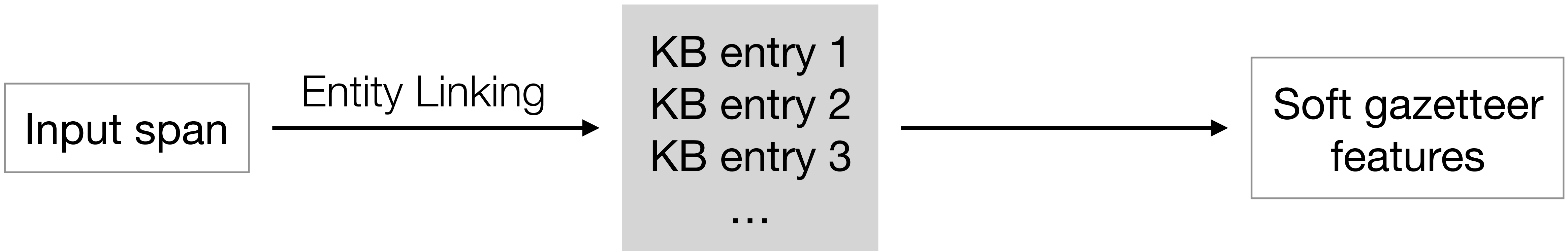
- Given an input sentence…

Soft Gazetteers: Method

- Given an input sentence…
- Candidate KB entries are retrieved for **each span in the sentence using pivot-based entity linking.**

Soft Gazetteers: Method

- Given an input sentence…
- Candidate KB entries are retrieved for **each span in the sentence using pivot-based entity linking.**
- The scores are used to create soft gazetteer features and applied to each word in the span.



Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

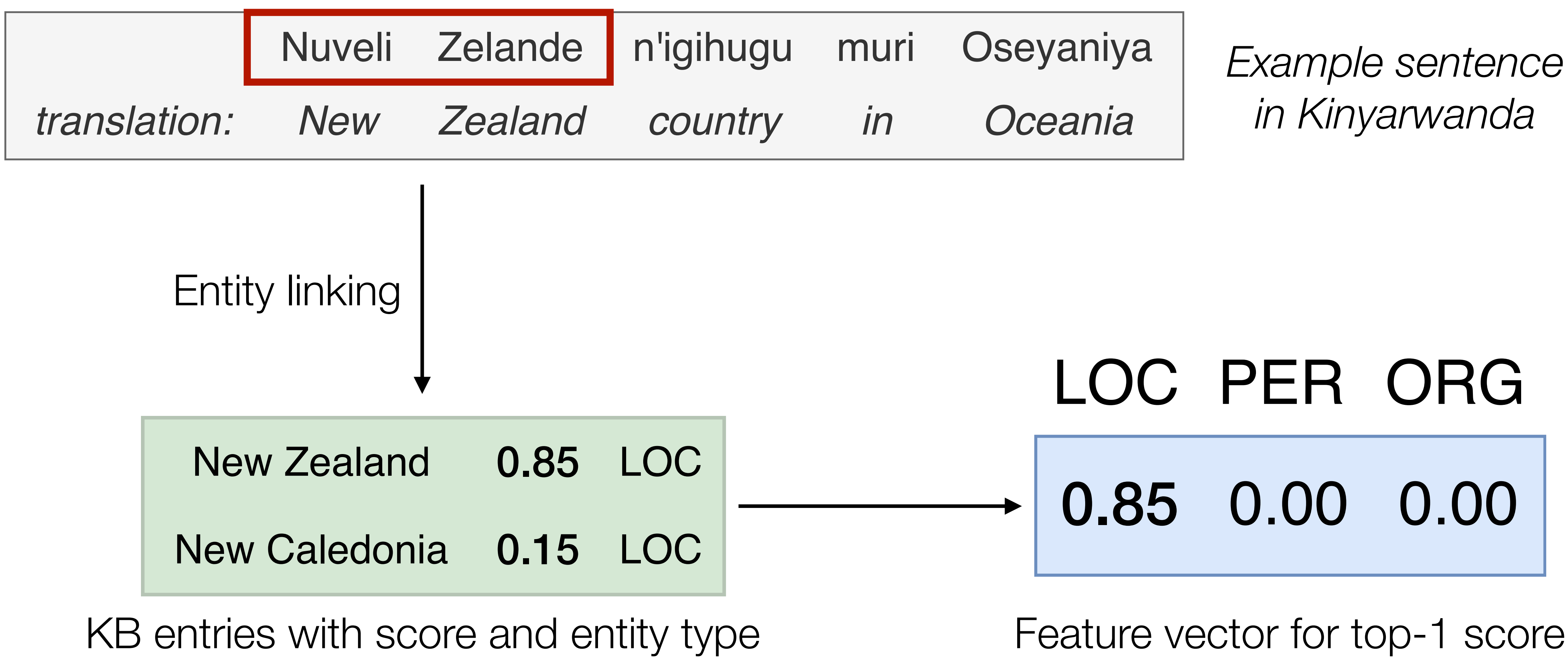
Entity linking

New Zealand	**0.85**	LOC
New Caledonia	**0.15**	LOC

KB entries with score and entity type

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**



Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

LOC	PER	ORG
0.85	0.00	0.00

Feature vector for top-1 score

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	$\boxed{\text{Nuveli}}$	$\boxed{\text{Zelande}}$	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

LOC	PER	ORG
0.85	0.00	0.00

Feature vector for top-1 score

Application to each word in the span

w_i ="Nuveli"

	LOC	PER	ORG
B-	**0.85**	**0.0**	**0.0**
I-	0.0	0.0	0.0

w_i ="Zelande"

	LOC	PER	ORG
B-	0.0	0.0	0.0
I-	**0.85**	**0.0**	**0.0**

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

LOC	PER	ORG
0.85	0.00	0.00

Feature vector for top-1 score

Application to each word in the span

w_i ="Nuveli"

	LOC	PER	ORG
$\boxed{\text{B-}}$	**0.85**	**0.0**	**0.0**
I-	0.0	0.0	0.0

w_i ="Zelande"

	LOC	PER	ORG
B-	0.0	0.0	0.0
I-	**0.85**	**0.0**	**0.0**

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

LOC	PER	ORG
0.85	0.00	0.00

Feature vector for top-1 score

→

Application to each word in the span

w_i ="Nuveli"

	LOC	PER	ORG
B-	**0.85**	**0.0**	**0.0**
I-	0.0	0.0	0.0

w_i ="Zelande"

	LOC	PER	ORG
B-	0.0	0.0	0.0
I-	**0.85**	**0.0**	**0.0**

Soft Gazetteers: An Example

Consider a feature that represents **the top scoring KB entry**

	Nuveli	Zelande	n'igihugu	muri	Oseyaniya
translation:	*New*	*Zealand*	*country*	*in*	*Oceania*

Example sentence in Kinyarwanda

Similarly:

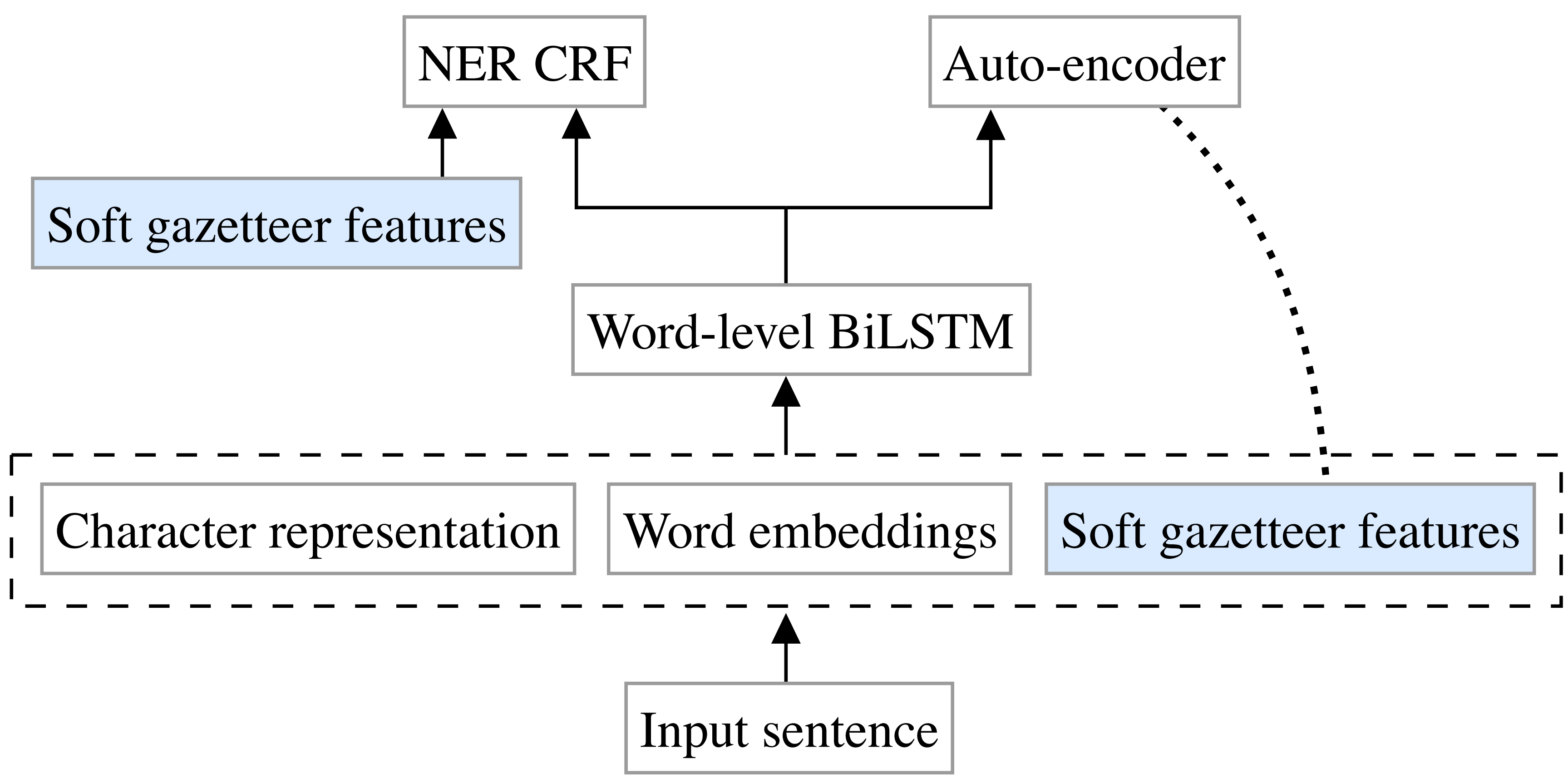
- top-3 candidate scores
- top-3 type-wise counts
- top-30 type-wise counts
- margins between top-4 candidates

Named Entity Recognition Model

- NER Model Architecture:
    - **Bi-LSTM** to encode the input
    - **CRF** to make a globally normalized prediction over the sequence

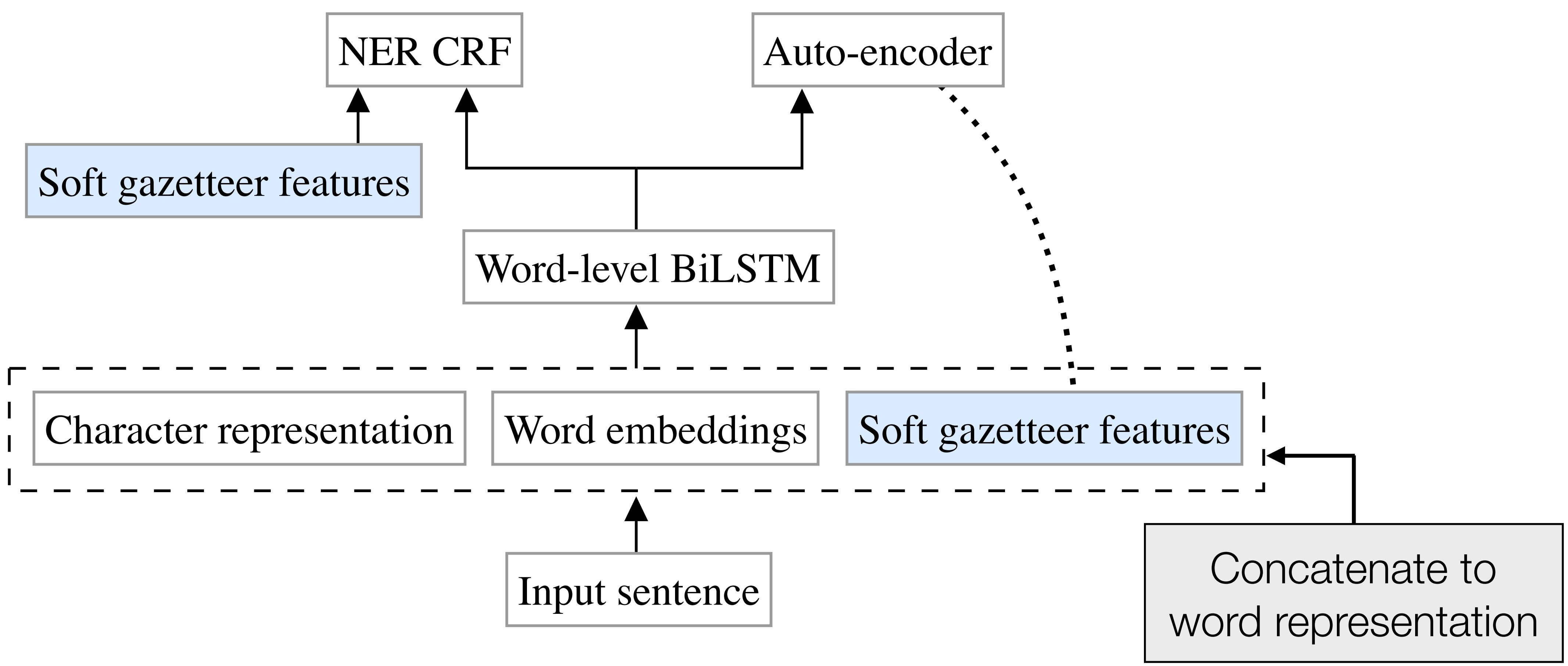
Named Entity Recognition Model

- NER Model Architecture:
  - **Bi-LSTM** to encode the input
  - **CRF** to make a globally normalized prediction over the sequence



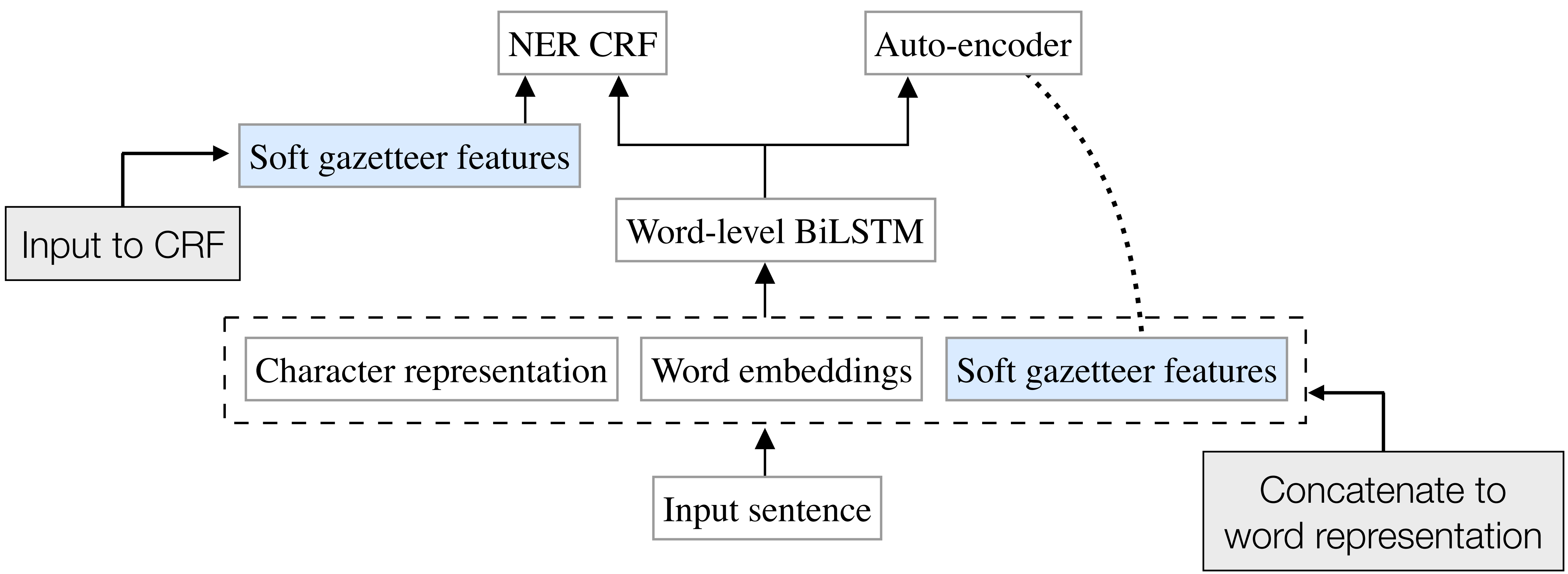
Named Entity Recognition Model

- NER Model Architecture:
  - **Bi-LSTM** to encode the input
  - **CRF** to make a globally normalized prediction over the sequence



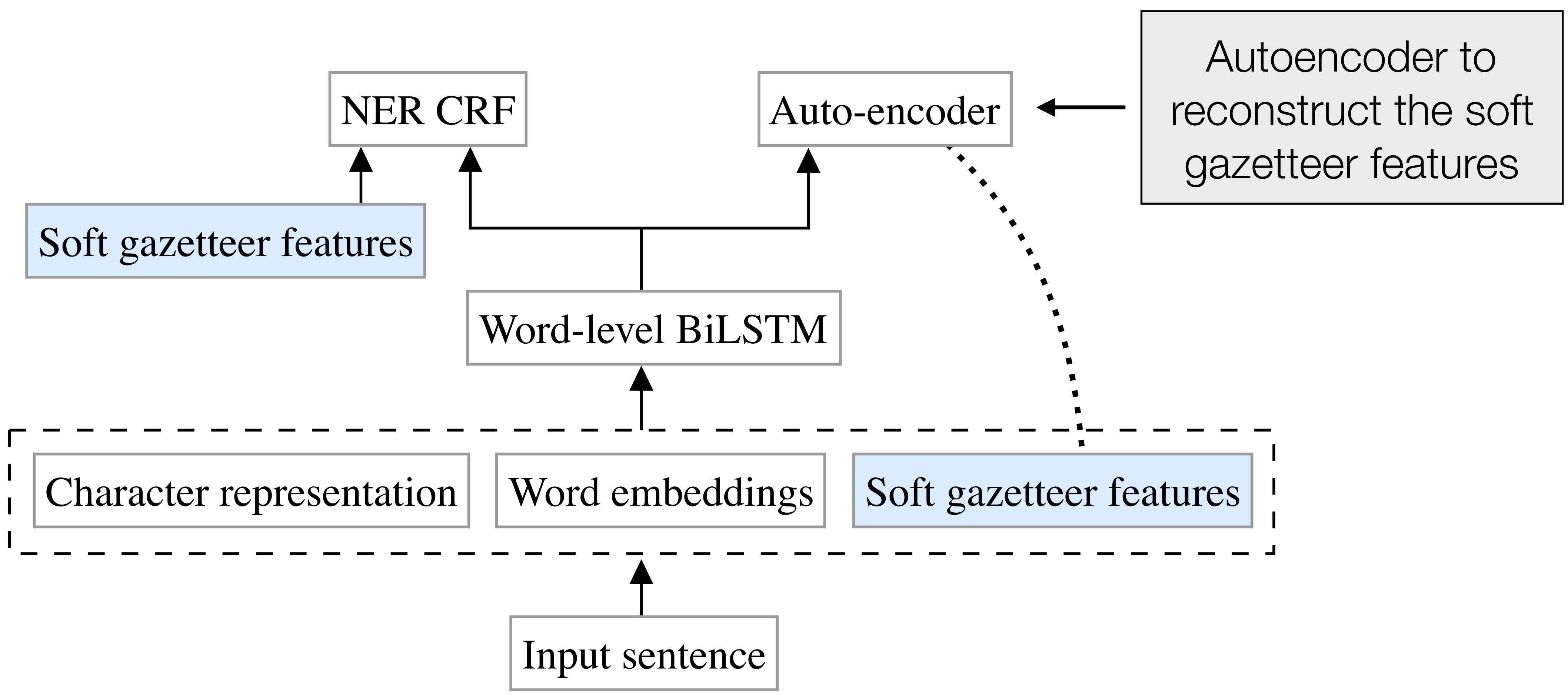
Named Entity Recognition Model

- NER Model Architecture:
  - **Bi-LSTM** to encode the input
  - **CRF** to make a globally normalized prediction over the sequence



Named Entity Recognition Model

- NER Model Architecture:
  - **Bi-LSTM** to encode the input
  - **CRF** to make a globally normalized prediction over the sequence



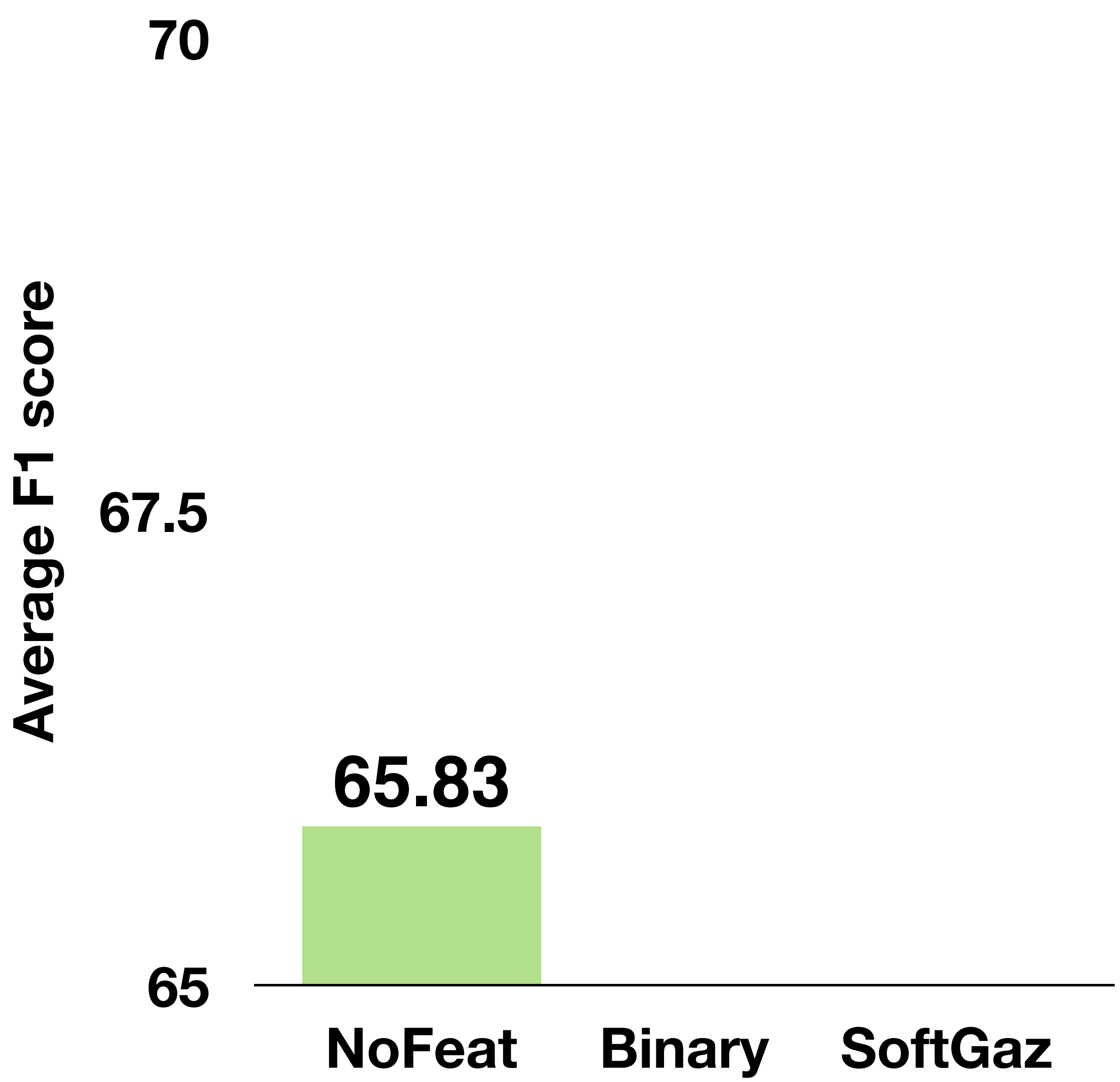
Experiments

Experiments

- **Four low-resource languages**: Kinyarwanda, Oromo, Sinhala, Tigrinya.

Experiments

- **Four low-resource languages**: Kinyarwanda, Oromo, Sinhala, Tigrinya.
- Baseline NER model with **no features**.

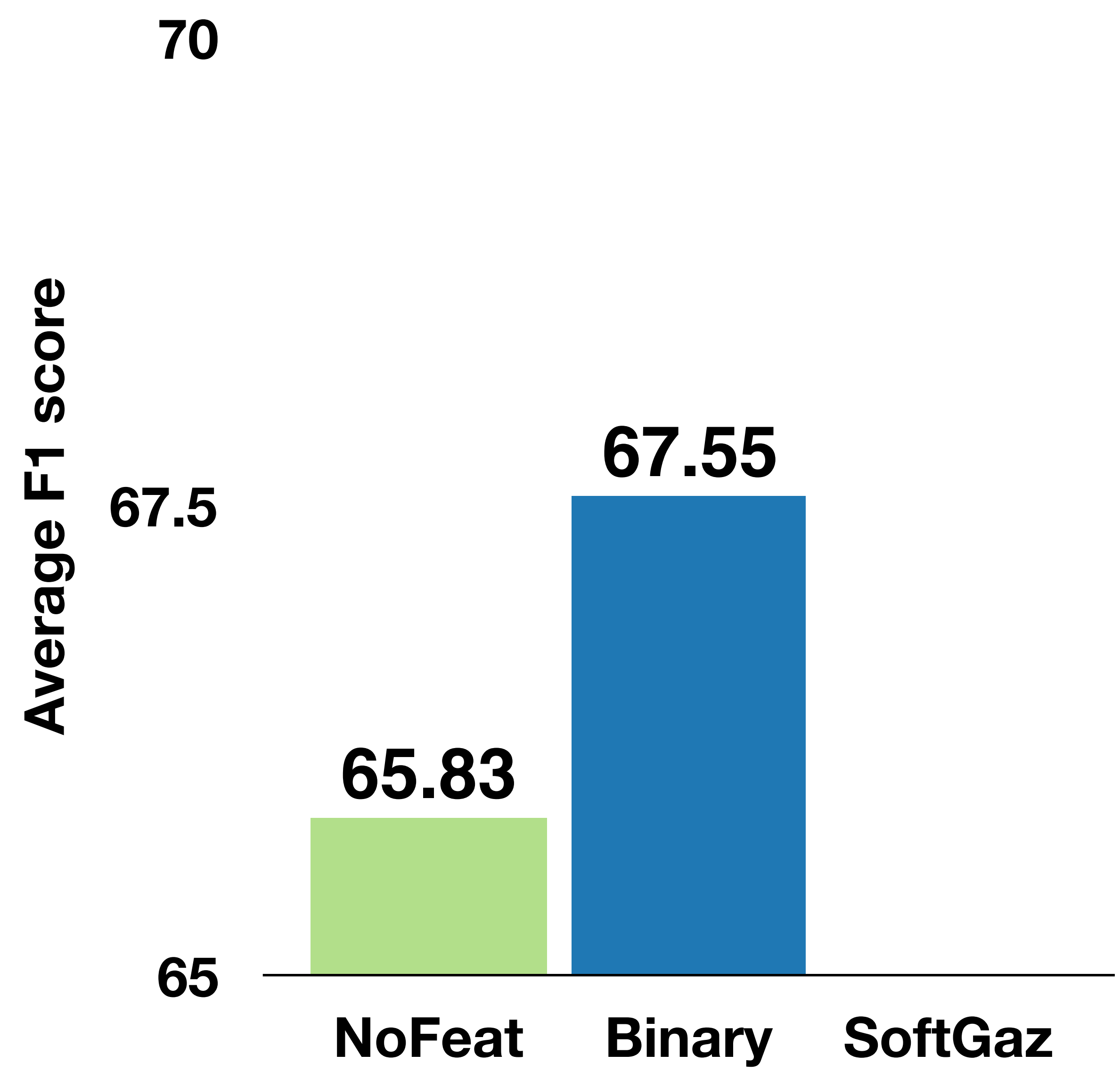


Experiments

- **Four low-resource languages**: Kinyarwanda, Oromo, Sinhala, Tigrinya.

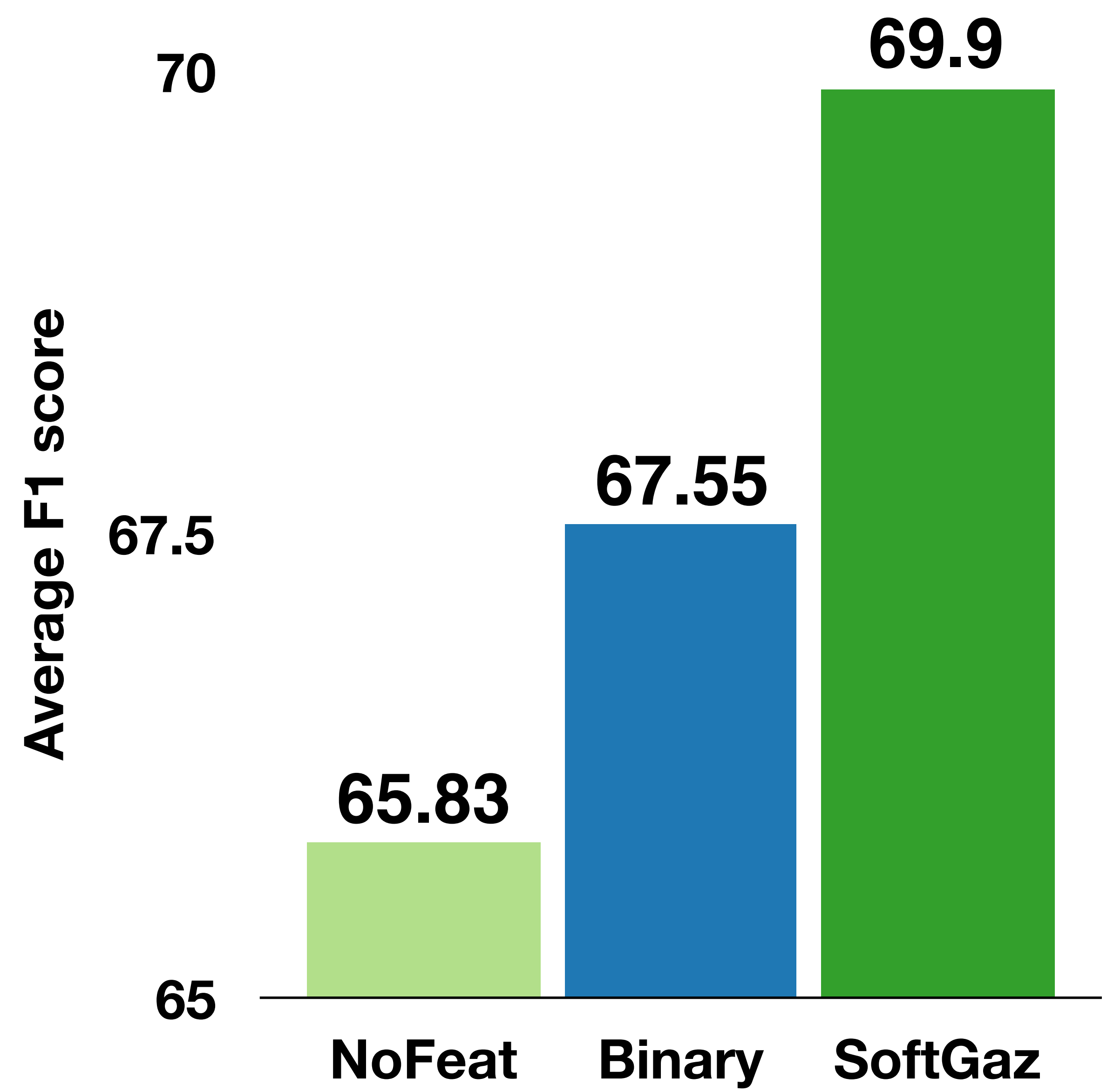
- Baseline NER model with **no features**.

- **Binary valued** gazetteer features.



Experiments

- **Four low-resource languages**: Kinyarwanda, Oromo, Sinhala, Tigrinya.
- Baseline NER model with **no features**.
- **Binary valued** gazetteer features.
- **Soft gazetteer** features from pivot-based entity linking.



Analysis

Analysis

- **What types of named entities benefit** from soft gazetteer features?

Analysis

- **What types of named entities benefit** from soft gazetteer features?
- Linkable entities that have a corresponding entry in the KB.

Analysis

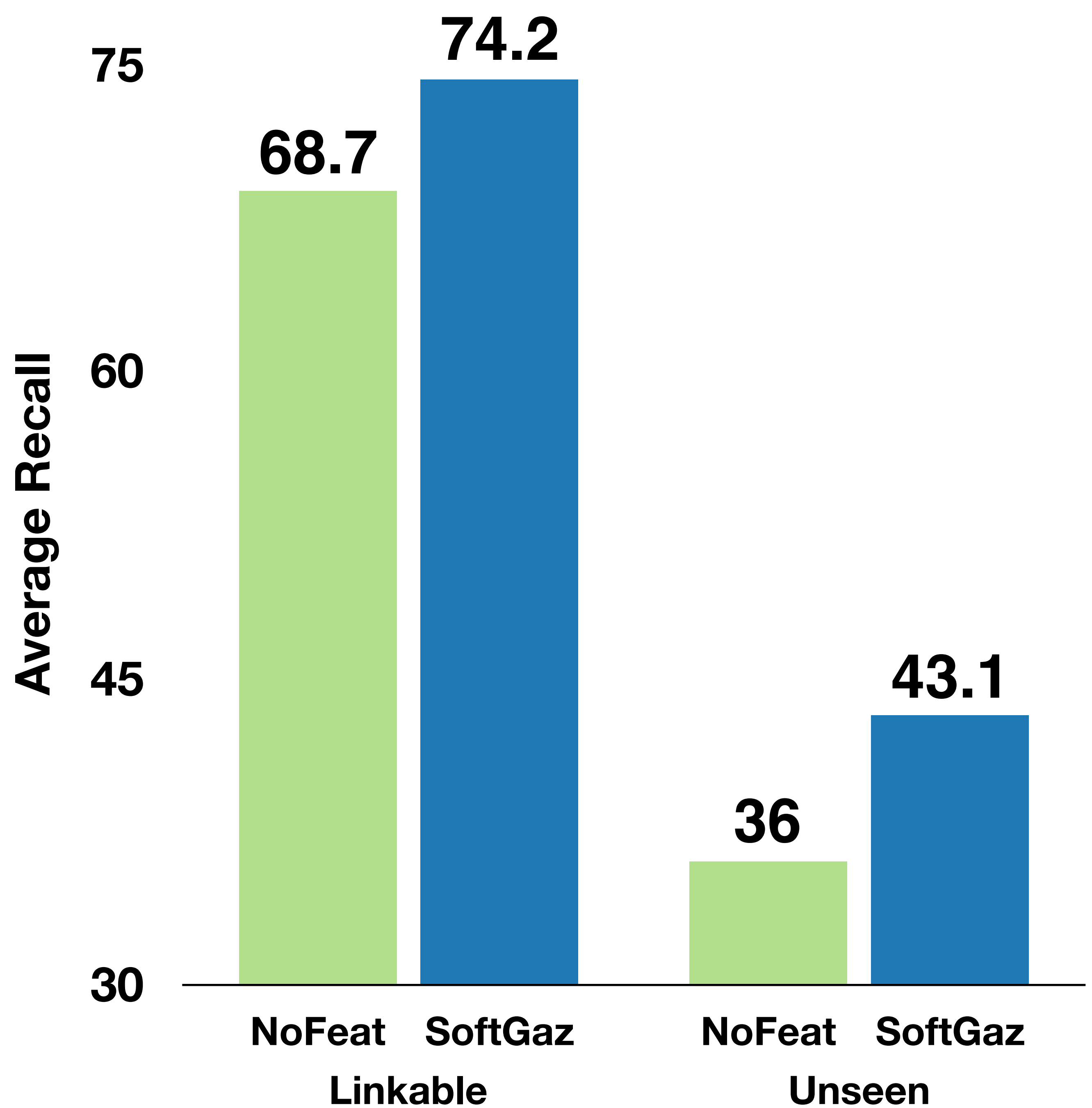
- **What types of named entities benefit** from soft gazetteer features?

- Linkable entities that have a corresponding entry in the KB.

- Unseen entities that are present in the test data but unseen during training.

Analysis

- **What types of named entities benefit** from soft gazetteer features?
- Linkable entities that have a corresponding entry in the KB.
- Unseen entities that are present in the test data but unseen during training.



Soft Gazetteers for NER

Soft Gazetteers for NER

- The soft gazetteer method creates features for NER that **does not rely on high-coverage entity lists**.

Soft Gazetteers for NER

- The soft gazetteer method creates features for NER that **does not rely on high-coverage entity lists**.
- Soft gazetteer features **improve NER over the baselines** in experiments on four low-resource languages.

Soft Gazetteers for NER

- The soft gazetteer method creates features for NER that **does not rely on high-coverage entity lists**.
- Soft gazetteer features **improve NER over the baselines** in experiments on four low-resource languages.
- **Future directions** include more sophisticated feature design and combinations of candidate lists from different entity linking methods.

Summary

Summary

- We presented a **method for candidate retrieval** in the cross-lingual entity linking setting.
    - Requires **no bilingual resources** in the source language.
- We presented **data and modeling improvements** to increase the accuracy of the candidate retrieval.
- We used the improved candidate retrieval method to **supplement low-resource NER models**.

More experiments and analysis for the methods are in the papers!

Zero-Shot Neural Transfer for Cross-Lingual Entity Linking
Shruti Rijhwani, Jiateng Xie, Graham Neubig, Jaime Carbonell
AAAI 2019

Improving Candidate Generation for Low-resource Cross-lingual Entity Linking
Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, Graham Neubig
TACL 2020

Soft Gazetteers for Low-Resource Named Entity Recognition
Shruti Rijhwani, Shuyan Zhou, Graham Neubig, Jaime Carbonell
ACL 2020

Thank you!