Leveraging supplemental representations for sequential transduction

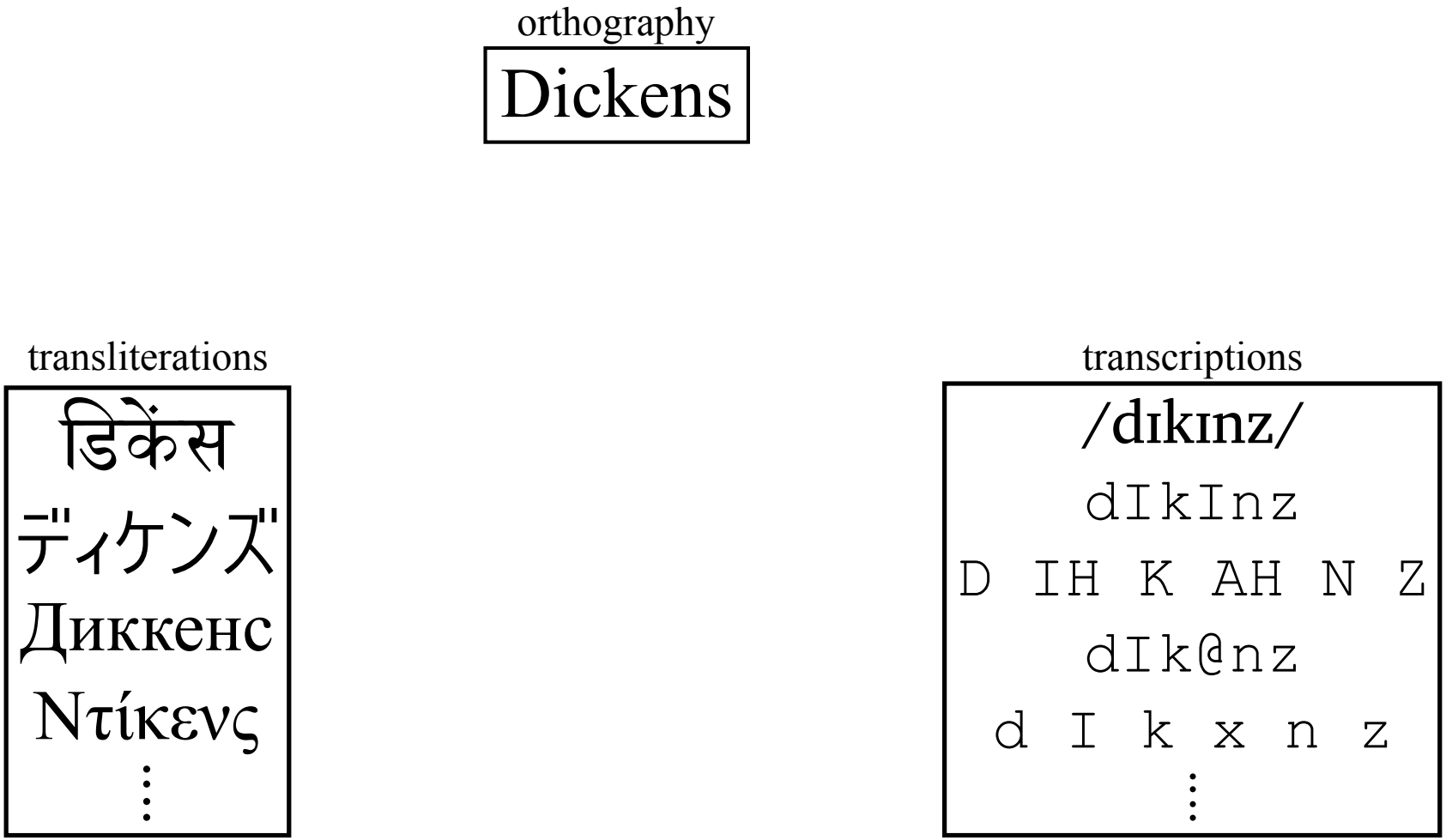
Aditya Bhargava[1] Grzegorz Kondrak[2]

[1]Department of Computer Science
University of Toronto

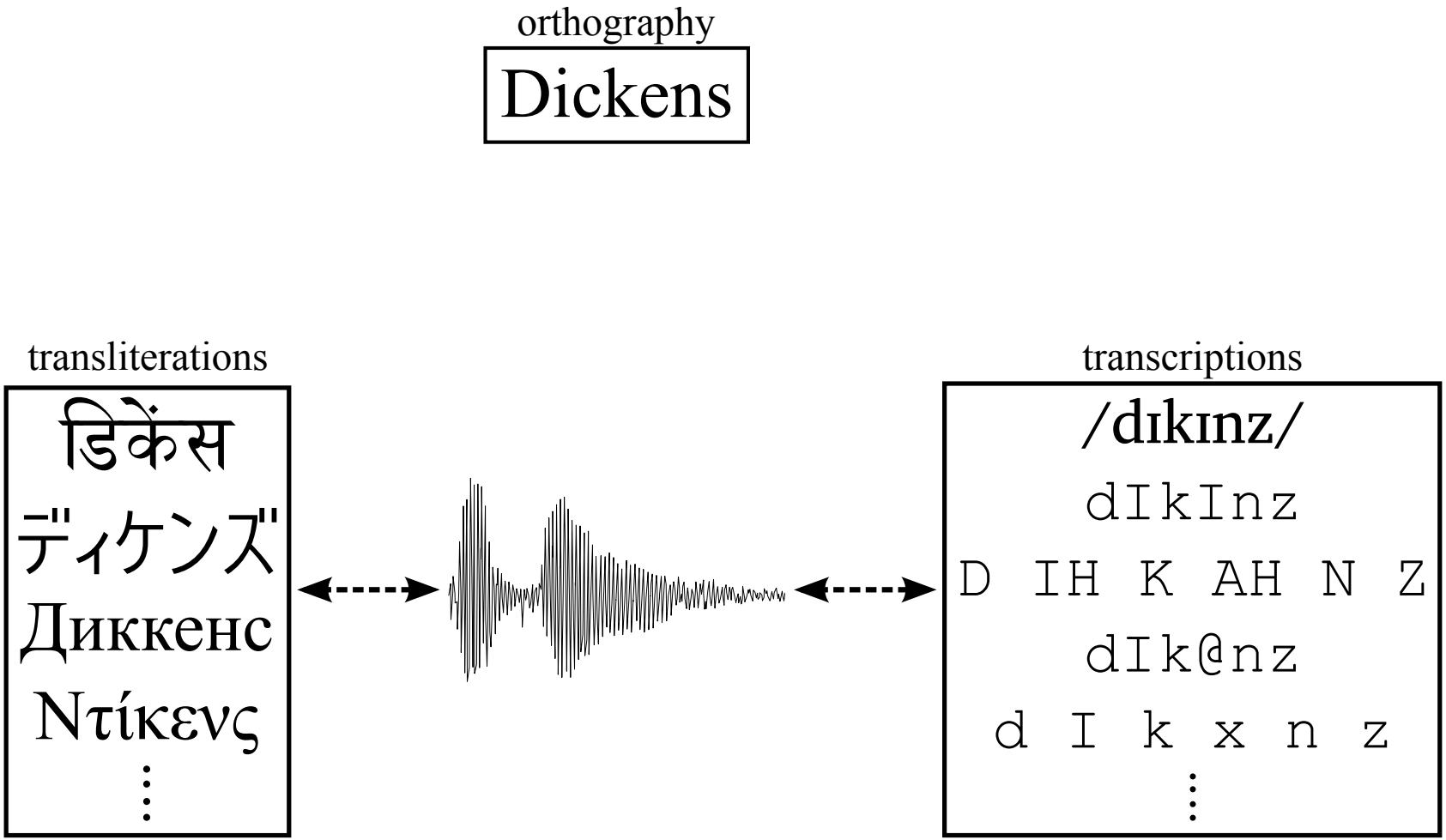
[2]Department of Computing Science
University of Alberta

NAACL-HLT 2012

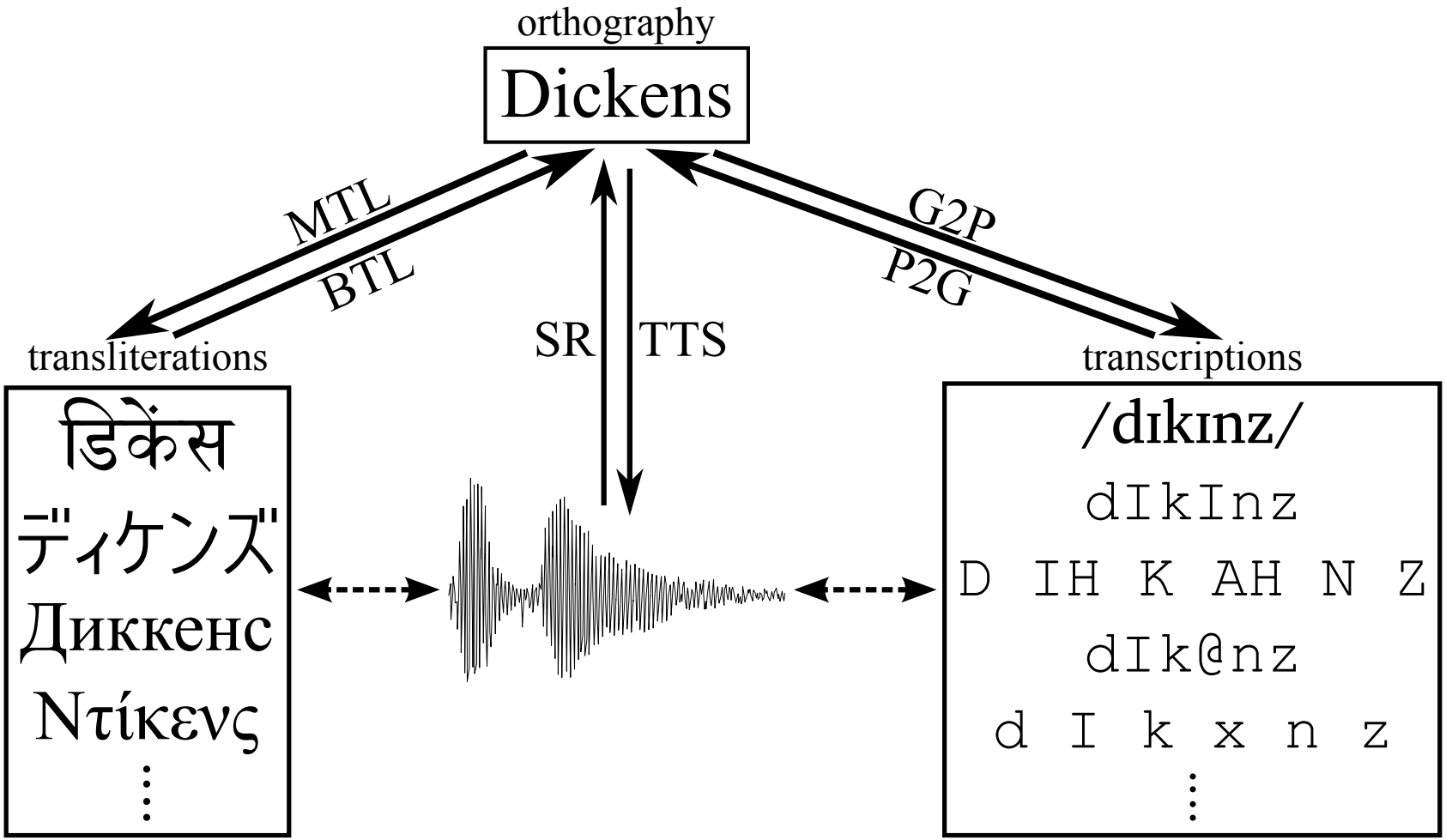
Pronunciation-based tasks



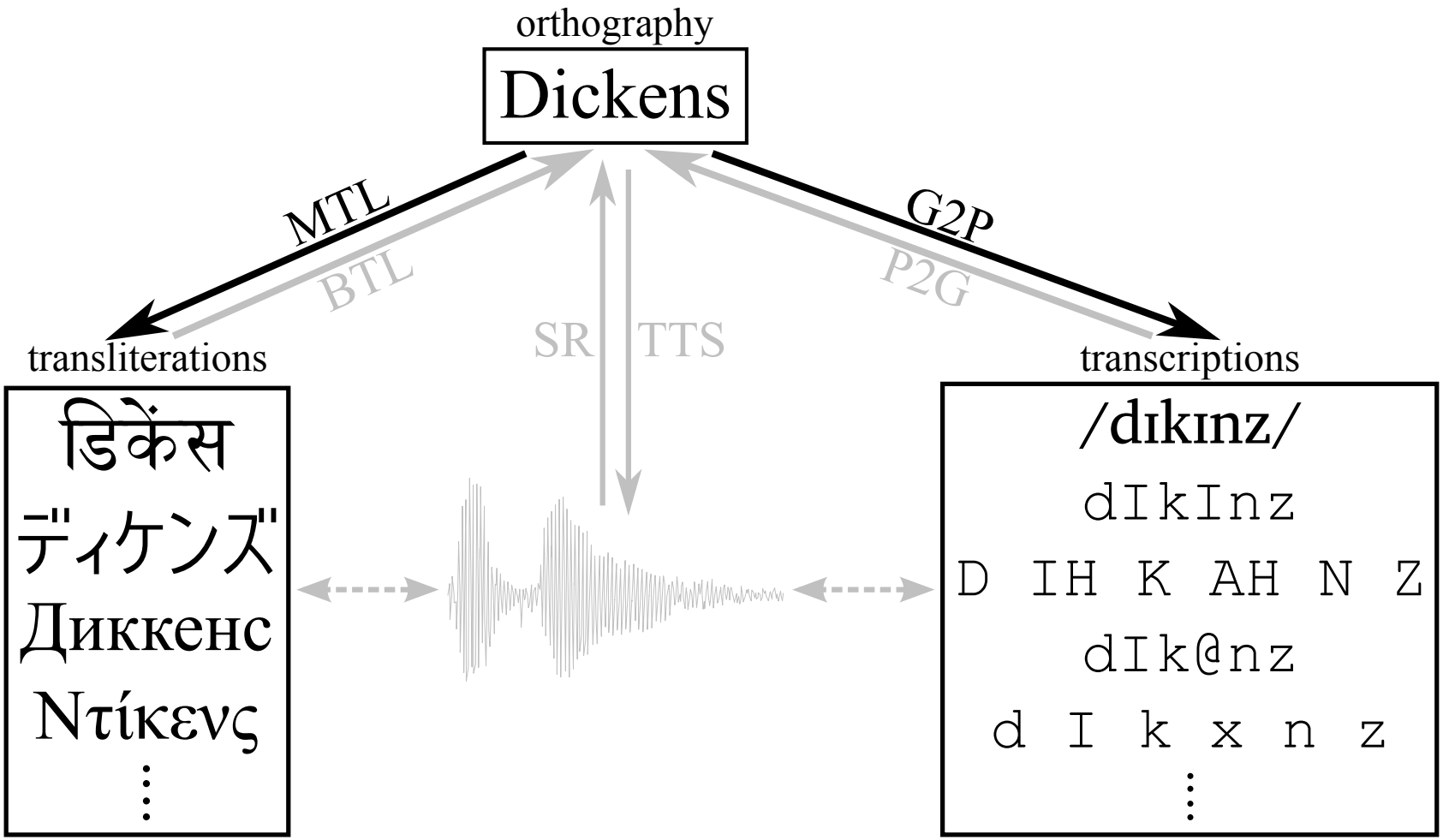
Pronunciation-based tasks



Pronunciation-based tasks



Pronunciation-based tasks



Overview

- x supplemental data for y
  - $x \in \{\text{transcription, transliteration}\}$
  - $y \in \{\text{G2P, MTL}\}$
- Rerank outputs from existing system
  - Features similar to base system, but applied to supplemental data
  - n-grams, alignment/similarity scores
- Same approach for system combination
  - Use another G2P/MTL system's outputs as supplemental data

Overview

- Excellent results
  - Up to 8.7% error reduction for system combination
  - MTL sees error reduction up to 14% from transliterations and 18% from transcriptions
  - G2P sees error reduction up to 43% from transcriptions

Overview

- Excellent results
  - Up to 8.7% error reduction for system combination
  - MTL sees error reduction up to 14% from transliterations and 18% from transcriptions
  - G2P sees error reduction up to 43% from transcriptions
  - But transliterations help G2P for names only

Overview

- Excellent results (mostly)
  - Up to 8.7% error reduction for system combination
  - MTL sees error reduction up to 14% from transliterations and 18% from transcriptions
  - G2P sees error reduction up to 43% from transcriptions
  - But transliterations help G2P for names only

Reranking method

- From ACL 2011
- Looks specifically at transliterations as supplemental data for G2P of names
  - Names are hard(er)
  - Transliteration is generally applied to named entities
  - Encodes relevant pronunciation information
- Using supplemental data, rerank *n*-best output list of G2P base system
- Additional findings:
  - Simple similarity-based methods don't work
  - Multiple languages are helpful

Reranking method

- Here, we experiment with:
  1. Transcriptions as supplemental data for both G2P and MTL
  2. Transcriptions and transliterations simultaneously
  3. G2P in general, rather than names only
  4. System combination as supplemental data

Reranking method

- Here, we experiment with:
  1. Transcriptions as supplemental data for both G2P and MTL
  2. Transcriptions and transliterations simultaneously
  3. G2P in general, rather than names only
  4. System combination as supplemental data

Reranking method

- Here, we experiment with:
  1. Transcriptions as supplemental data for both G2P and MTL
  2. Transcriptions and transliterations simultaneously
  3. G2P in general, rather than names only
  4. System combination as supplemental data

Reranking method

- Here, we experiment with:
  1. Transcriptions as supplemental data for both G2P and MTL
  2. Transcriptions and transliterations simultaneously
  3. G2P in general, rather than names only
  4. System combination as supplemental data

Reranking method

- Here, we experiment with:
  1. Transcriptions as supplemental data for both G2P and MTL
  2. Transcriptions and transliterations simultaneously
  3. G2P in general, rather than names only
  4. System combination as supplemental data

Related work

- G2P systems
  - Neural networks, instance-based learning, . . .
  - . . ., joint *n*-gram models (Sequitur), online discriminative learning (DirecTL+)
- MTL systems
  - Similarly many approaches
  - Lately Sequitur and DirecTL+ have performed quite well at NEWS

Related work

- Using heterogeneous data
  - Pivot through a third language for transliteration
  - Mostly useful for low-resource environments
  - Hard to incorporate more languages
  - Linear combination of system scores

Method

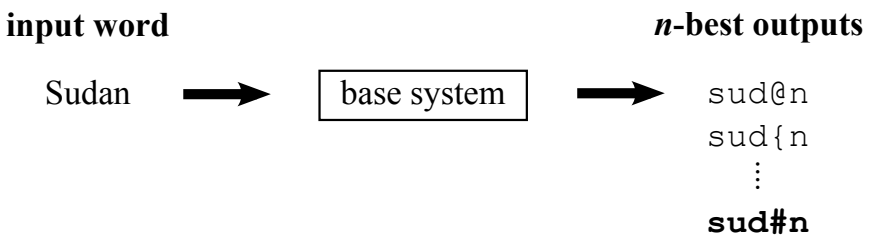
input word

Sudan

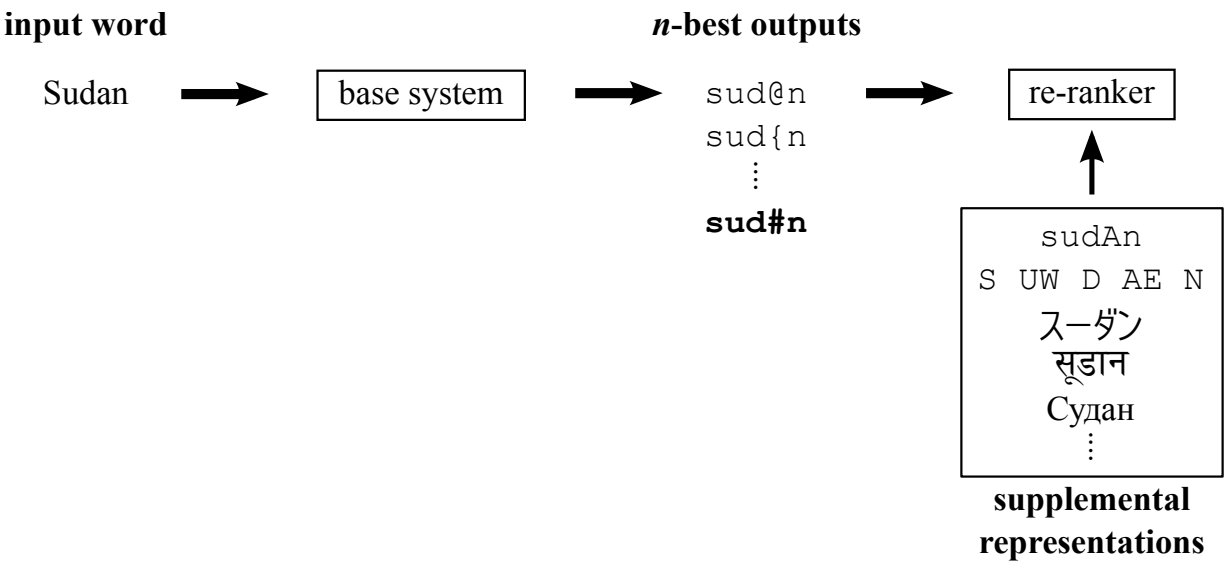
Method

input word

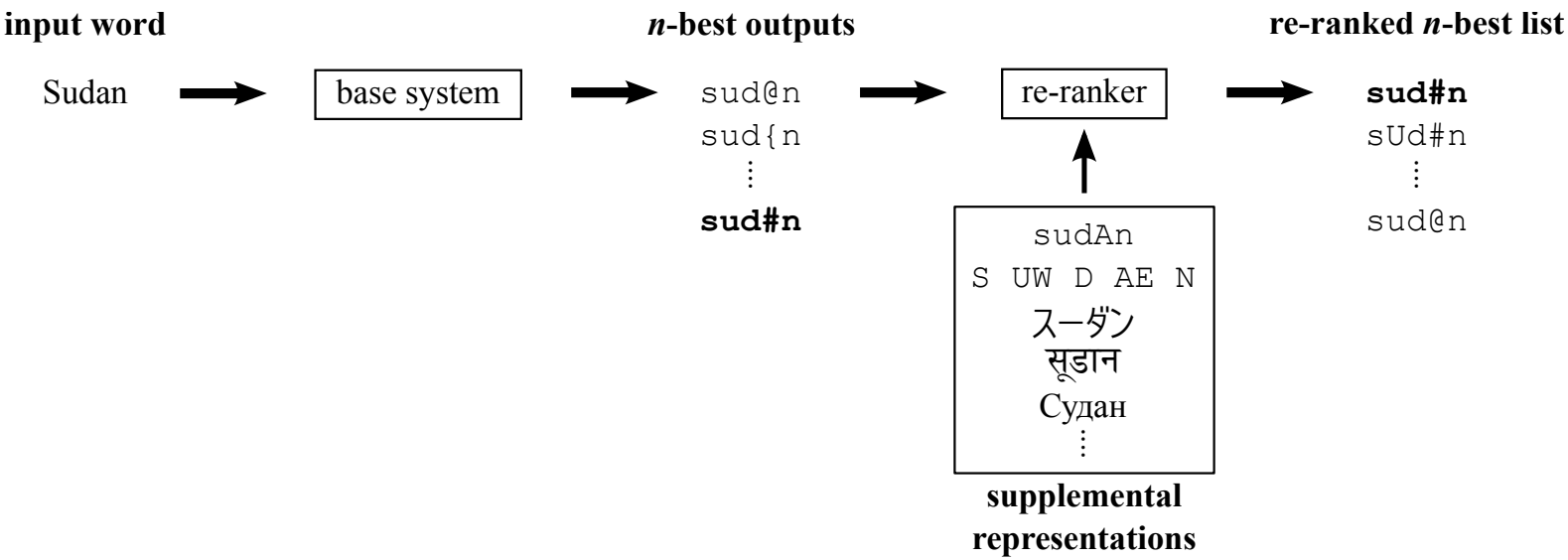
Method



Method


input word
Sudan
base system
n-best outputs
sud@n
sud{n
sud#n
re-ranker
sudAn
S UW D AE N
スーダン
सूडान
Судан
supplemental representations

Method


input word
Sudan
base system
n-best outputs
sud@n
sud{n
sud#n
re-ranker
sudAn
S UW D AE N
スーダン
सूडान
Судан
supplemental representations
re-ranked n-best list
sud#n
sUd#n
sud@n

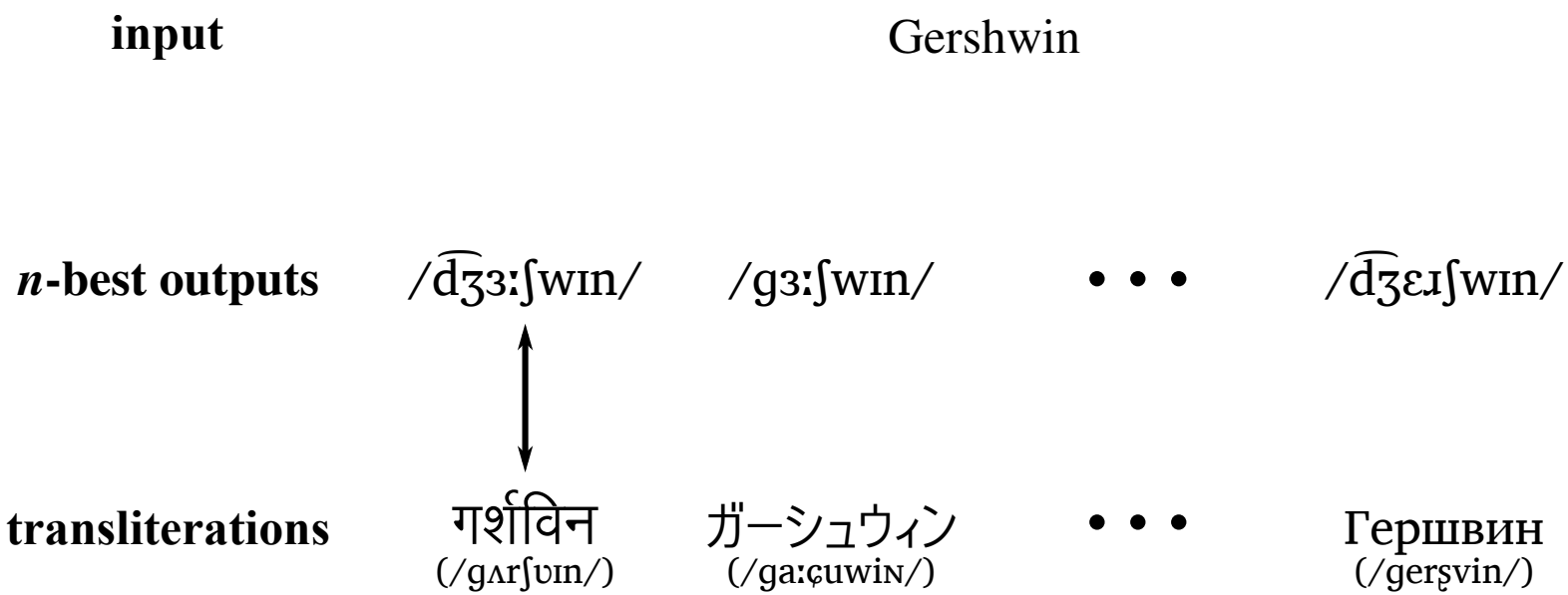
Method

input	Gershwin			
***n*-best outputs**	/d͡ʒɜːʃwɪn/	/gɜːʃwɪn/	• • •	/d͡ʒɛɹʃwɪn/

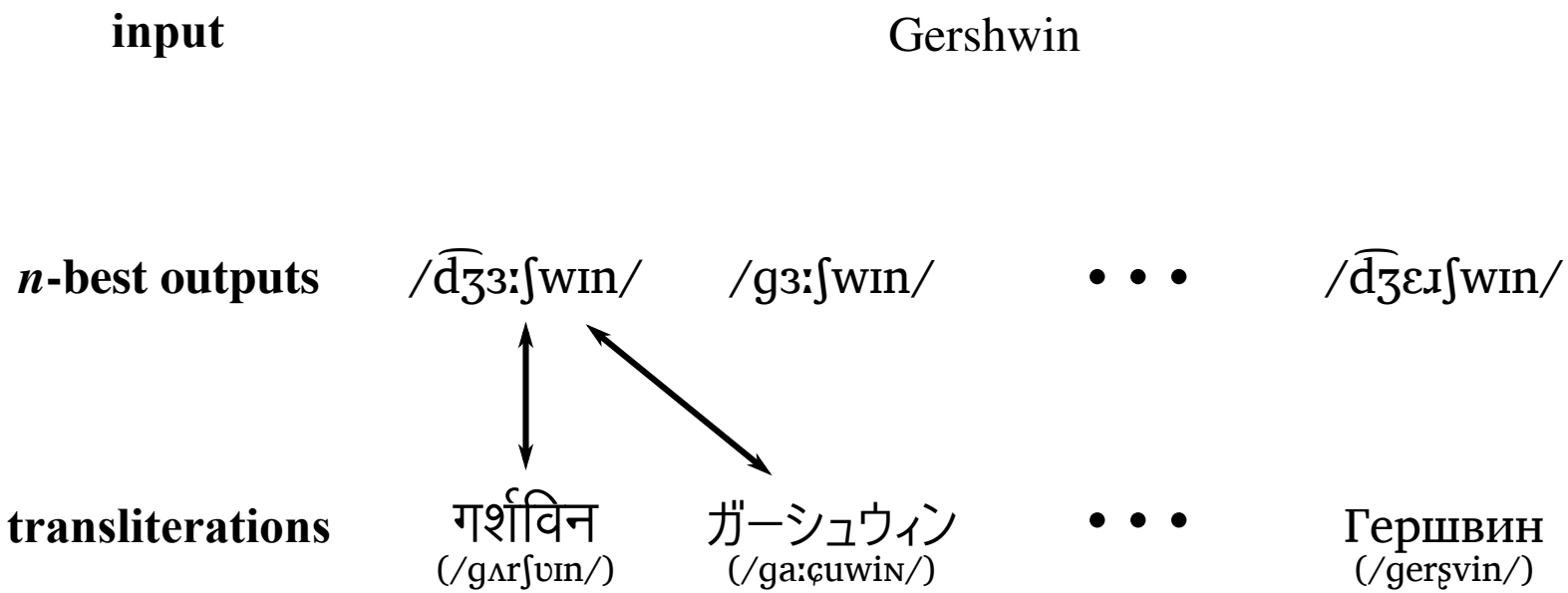
Method

input	Gershwin			
***n*-best outputs**	/d͡ʒɜːʃwɪn/	/gɜːʃwɪn/	• • •	/d͡ʒɛɹʃwɪn/
transliterations	गर्शविन (/gʌrʃʊɪn/)	ガーシュウィン (/gaːɕuwiɴ/)	• • •	Гершвин (/gerʂvin/)

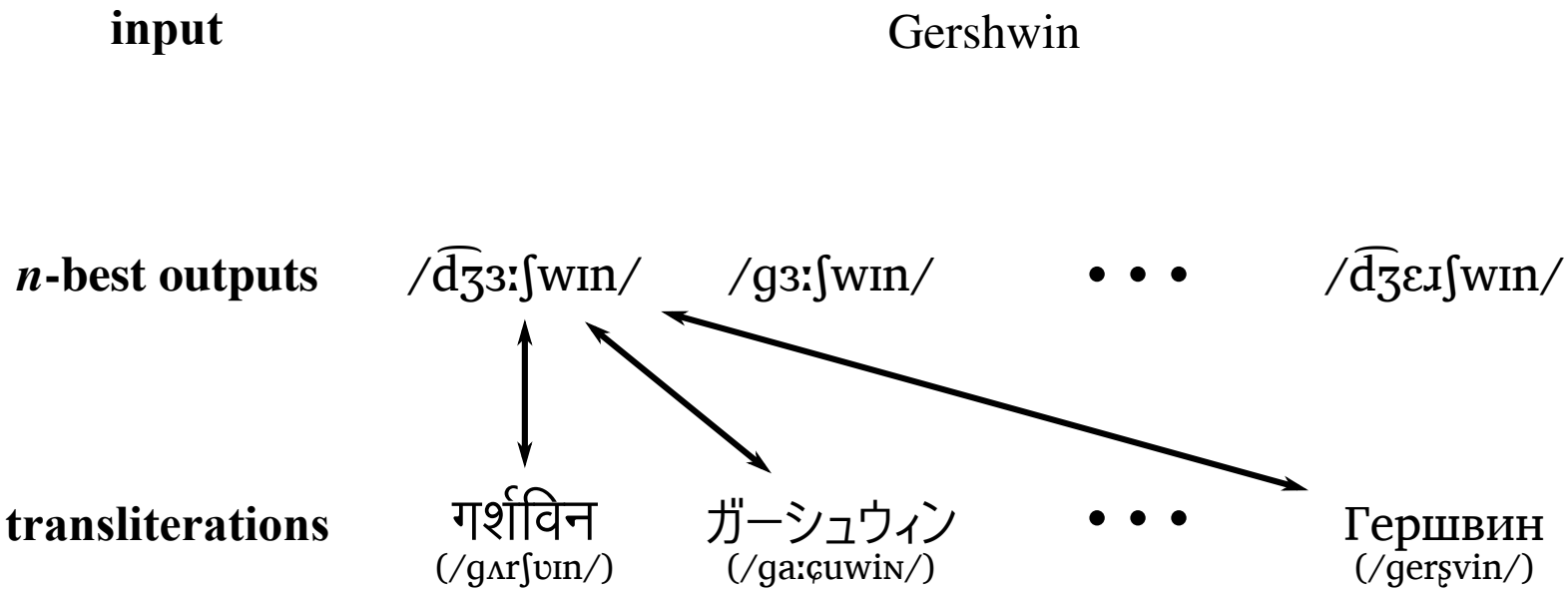
Method


input
Gershwin
n-best outputs
/d͡ʒɜːʃwɪn/
/gɜːʃwɪn/
• • •
/d͡ʒɛɹʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

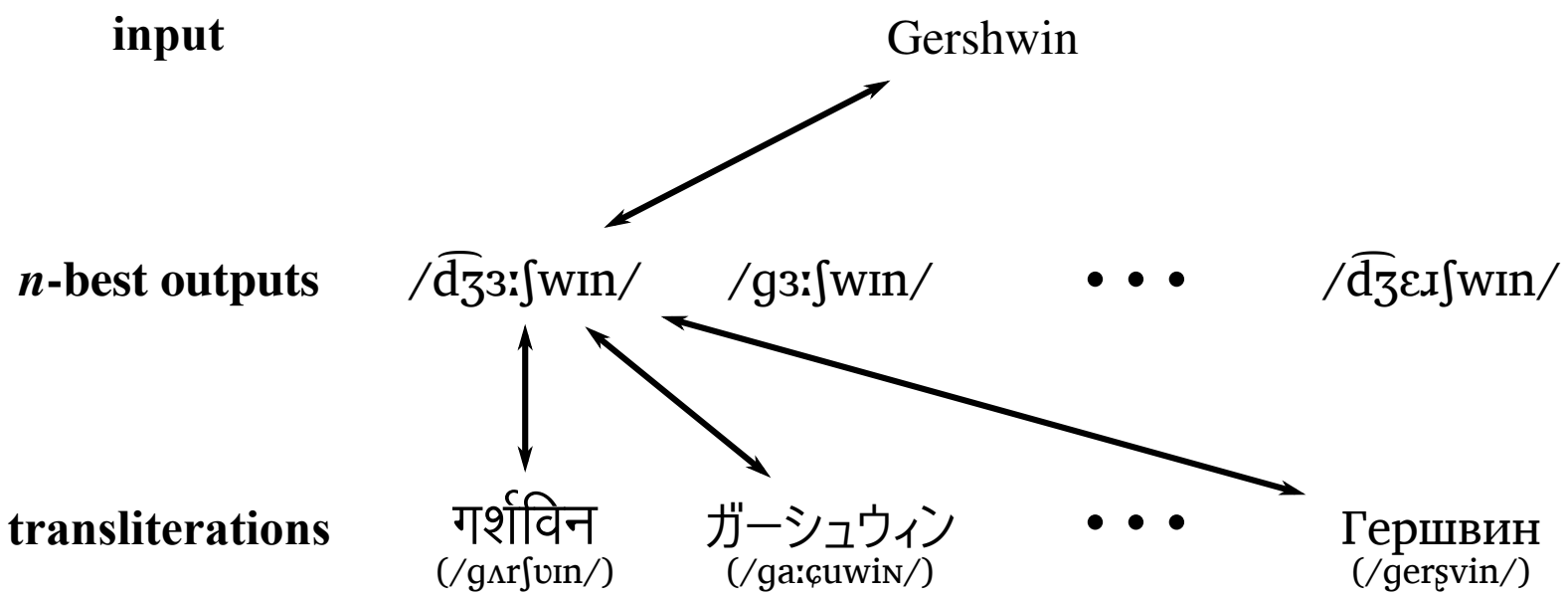
Method


input
Gershwin
n-best outputs
/d͡ʒɜːʃwɪn/
/gɜːʃwɪn/
• • •
/d͡ʒɛɹʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

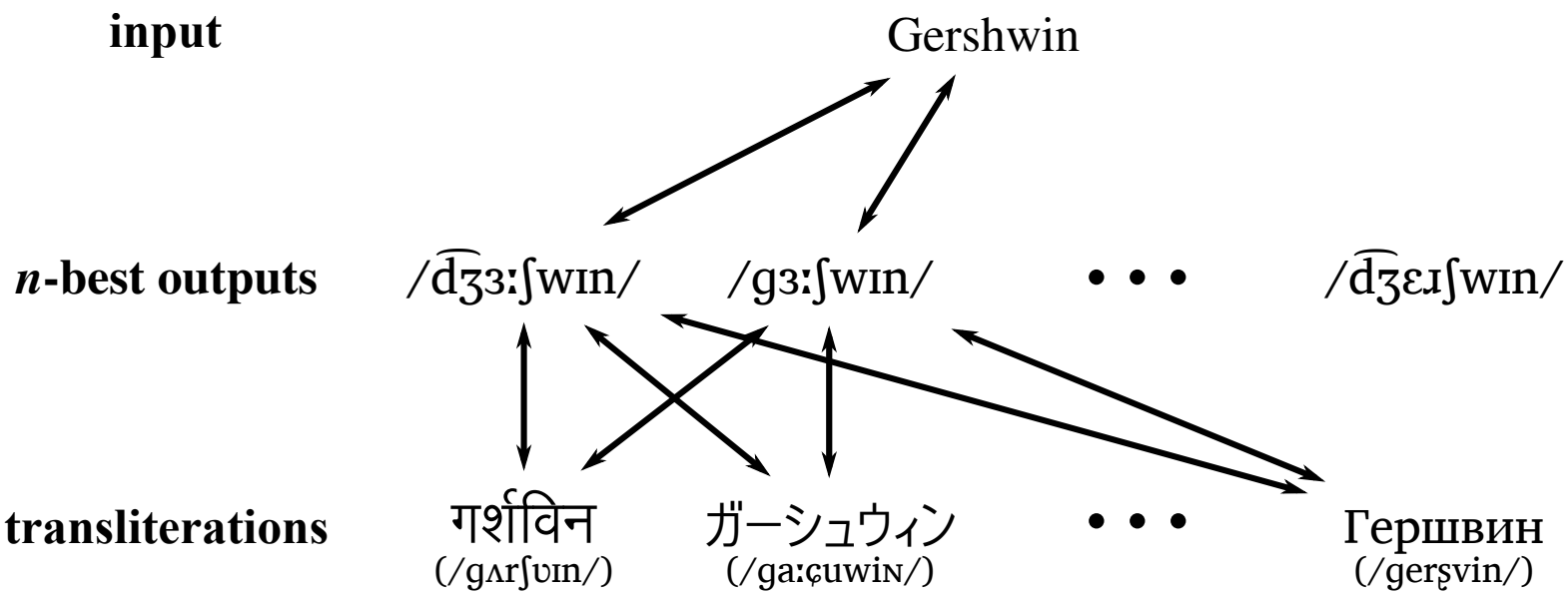
Method


input
Gershwin
n-best outputs
/d͡ʒɜːʃwɪn/
/gɜːʃwɪn/
• • •
/d͡ʒɛɹʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

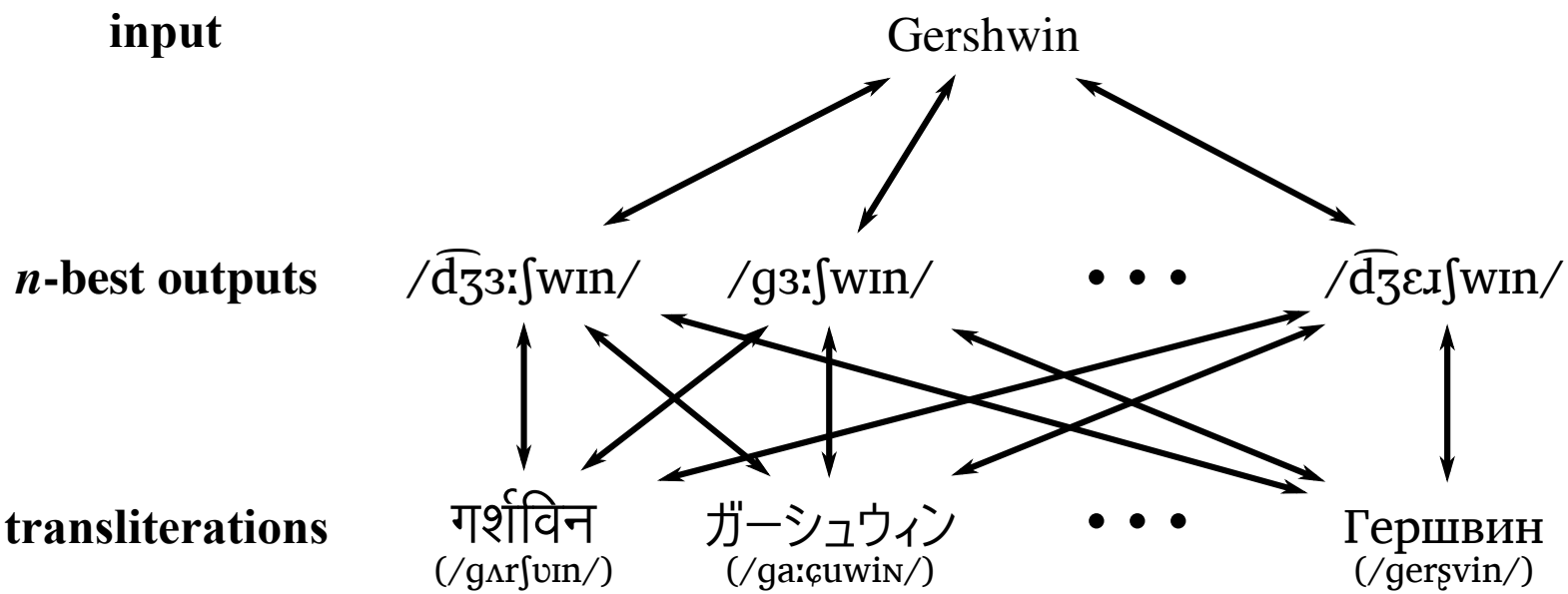
Method


input
Gershwin
n-best outputs
/ʤɜːʃwɪn/
/gɜːʃwɪn/
• • •
/ʤɛɹʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

Method


input
Gershwin
n-best outputs
/d͡ʒɜːʃwɪn/
/gɜːʃwɪn/
• • •
/d͡ʒɛɪʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

Method


input
Gershwin
n-best outputs
/ʤɜːʃwɪn/
/gɜːʃwɪn/
• • •
/ʤɛɪʃwɪn/
transliterations
गर्शविन
(/gʌrʃʊɪn/)
ガーシュウィン
(/gaːɕuwiɴ/)
• • •
Гершвин
(/gerʂvin/)

Data and base systems

- Transcriptions from Combilex and CELEX
- Transliterations from NEWS 2011
  - Experiment on English-to-Japanese transliteration
- 80/10/10 train/dev/test split
- Sequitur and DirecTL+ as base systems

G2P experiments

Supplemental transliterations

input	**candidate outputs**
McGee	m@kJi
	m@gi
	...
	m@CJi

G2P experiments

Supplemental transliterations

input	**candidate outputs**
McGee	m@kJi
	m@gi
	...
	m@CJi

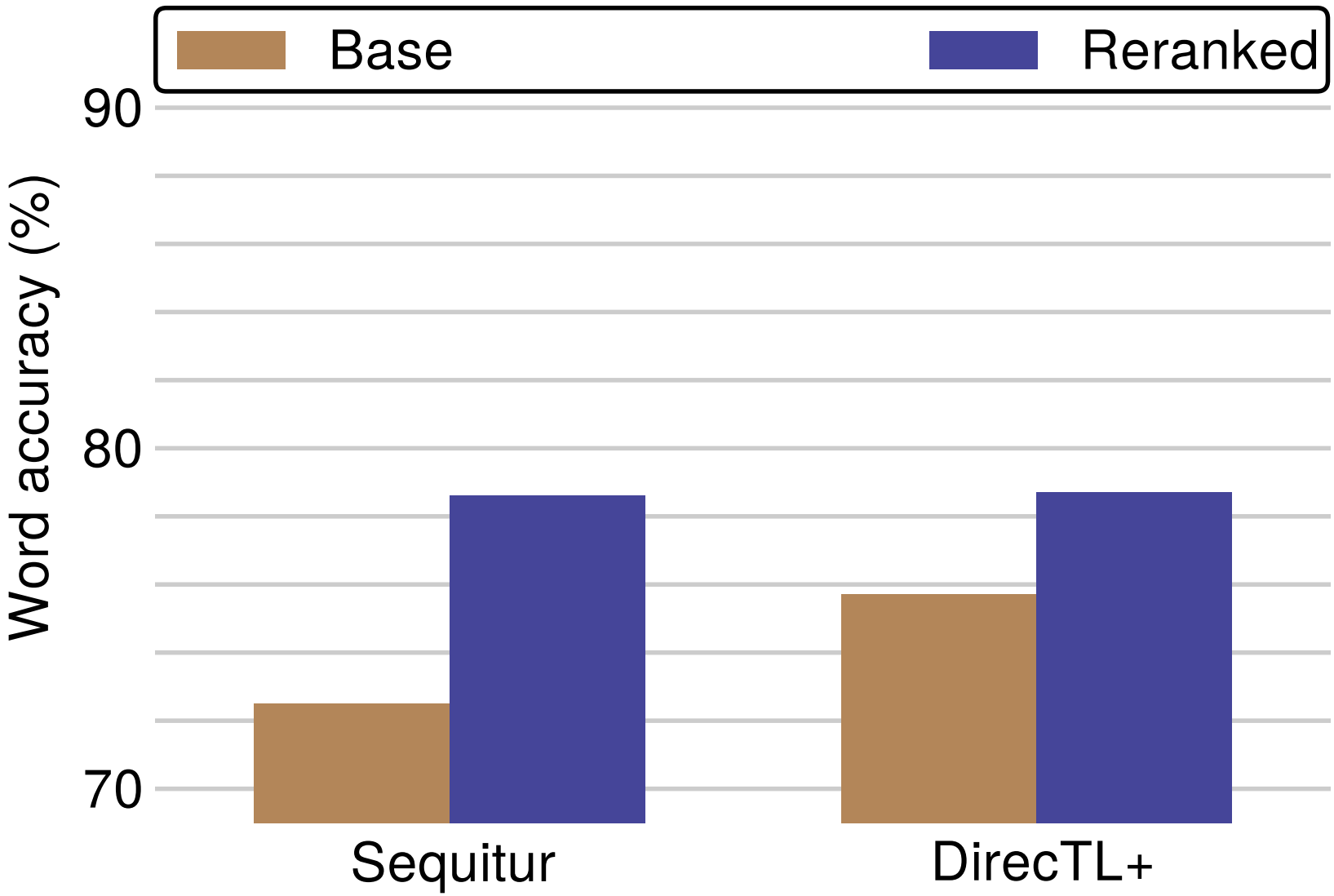
G2P experiments

Supplemental transliterations

input	**candidate outputs**	**supplemental**
McGee	m@kJi	मगी
	m@gi	マギー
	...	Макги
	m@CJi	

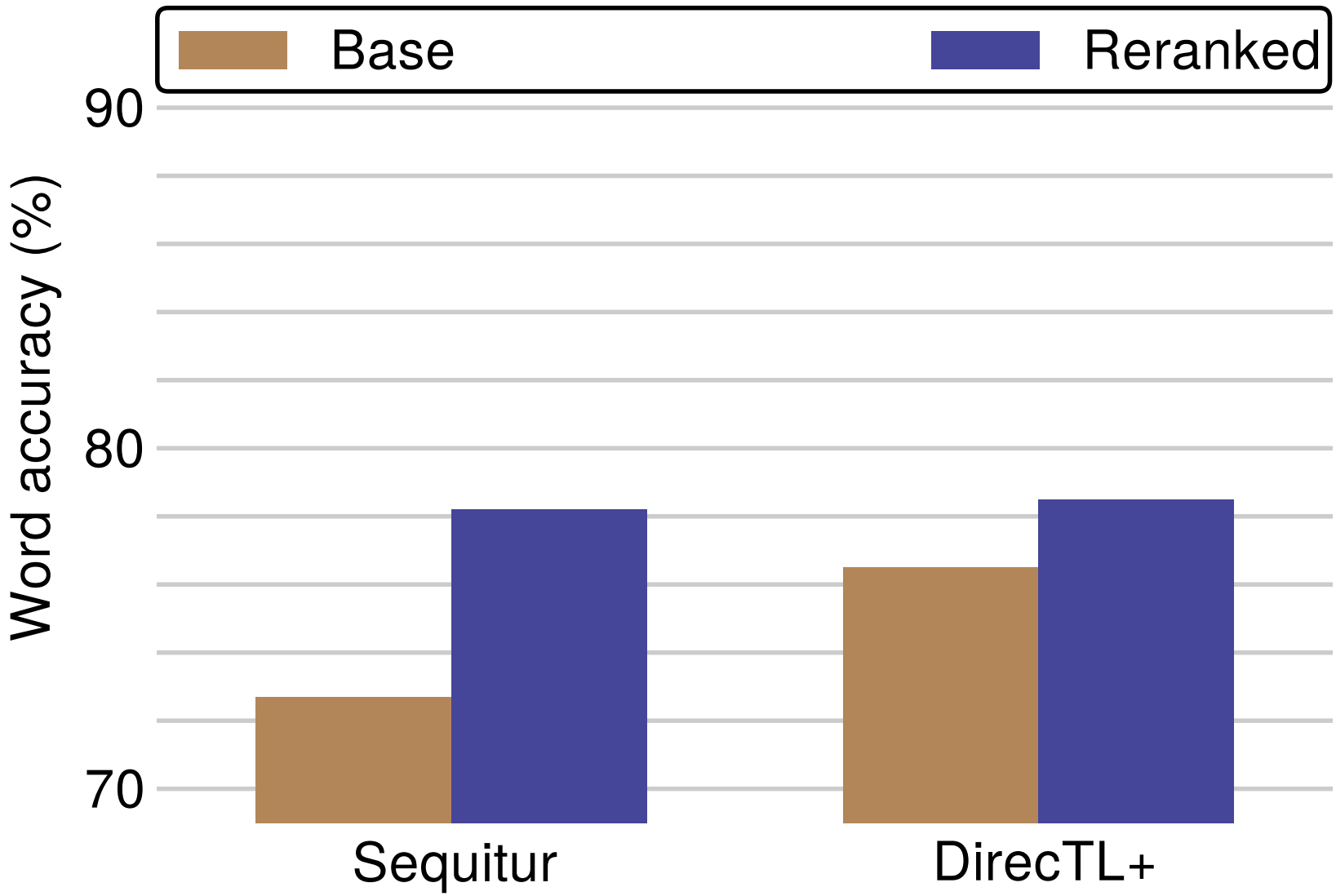
G2P experiments: names

Supplemental transliterations



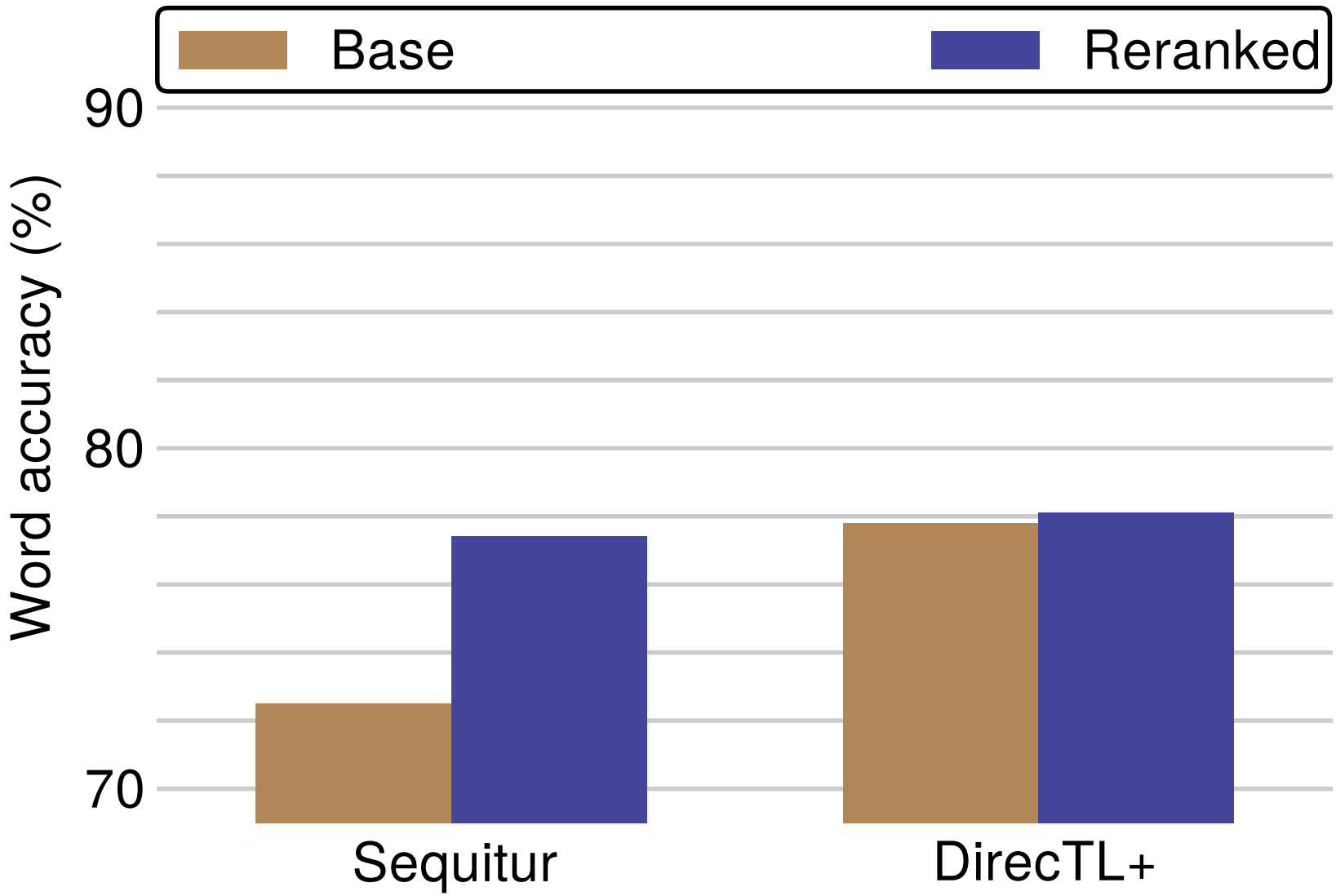
G2P experiments: full set

Supplemental transliterations



G2P experiments: core vocab

Supplemental transliterations



G2P experiments

Supplemental transcriptions

(word/name)	(CELEX)
input	**candidate outputs**
Sudan	sud@n
	sud{n
	...
	sud#n

G2P experiments

Supplemental transcriptions

(word/name)	(CELEX)
input	**candidate outputs**
Sudan	sud@n
	sud{n
	...
	sud#n

G2P experiments

Supplemental transcriptions

(word/name)	(CELEX)	(Combilex)
input	**candidate outputs**	**supplemental**
Sudan	sud@n	sudAn
	sud{n	
	...	
	sud#n	

G2P experiments: baselines

Supplemental transcriptions

- MERGE
  1. Convert Combilex to CELEX
  2. Merge with CELEX
  3. Train on combined set

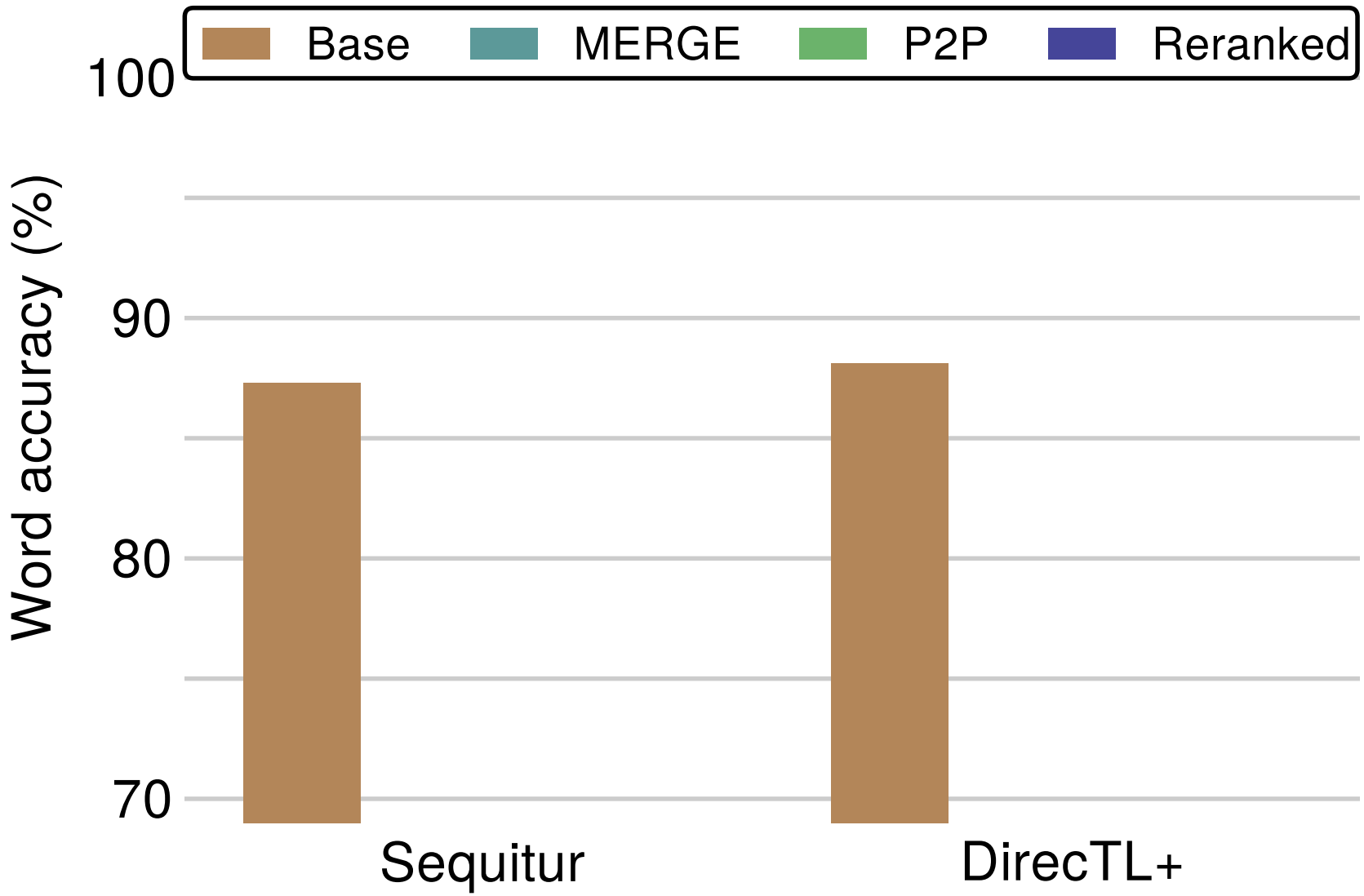
G2P experiments: baselines

Supplemental transcriptions

- MERGE
  1. Convert Combilex to CELEX
  2. Merge with CELEX
  3. Train on combined set
- P2P: phoneme-to-phoneme converter
  1. Intersect Combilex and CELEX
  2. Train a transduction system to convert Combilex to CELEX
  3. If a test word appears in Combilex, grab it from there and convert it to CELEX format

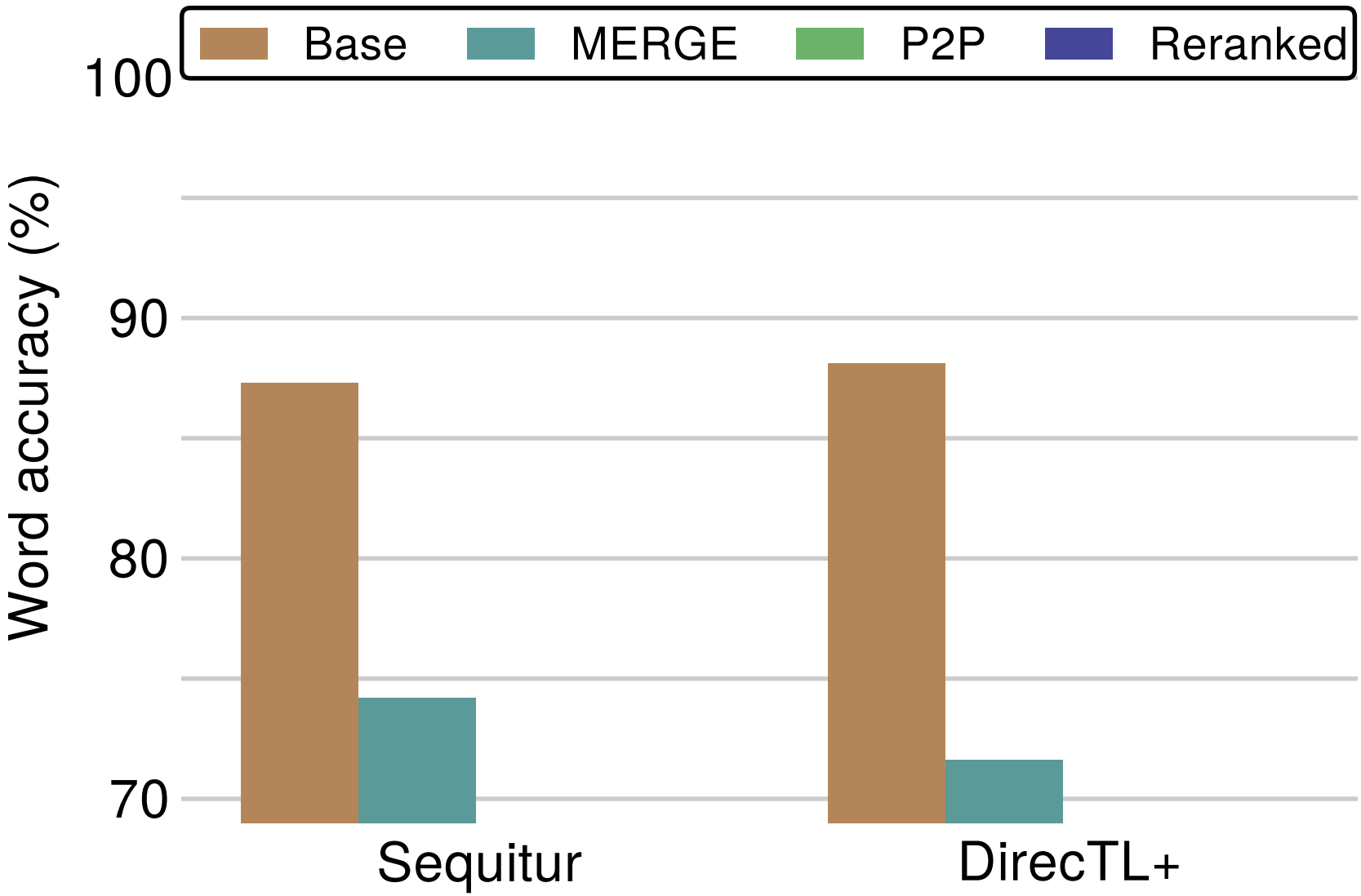
G2P experiments

Supplemental transcriptions: results



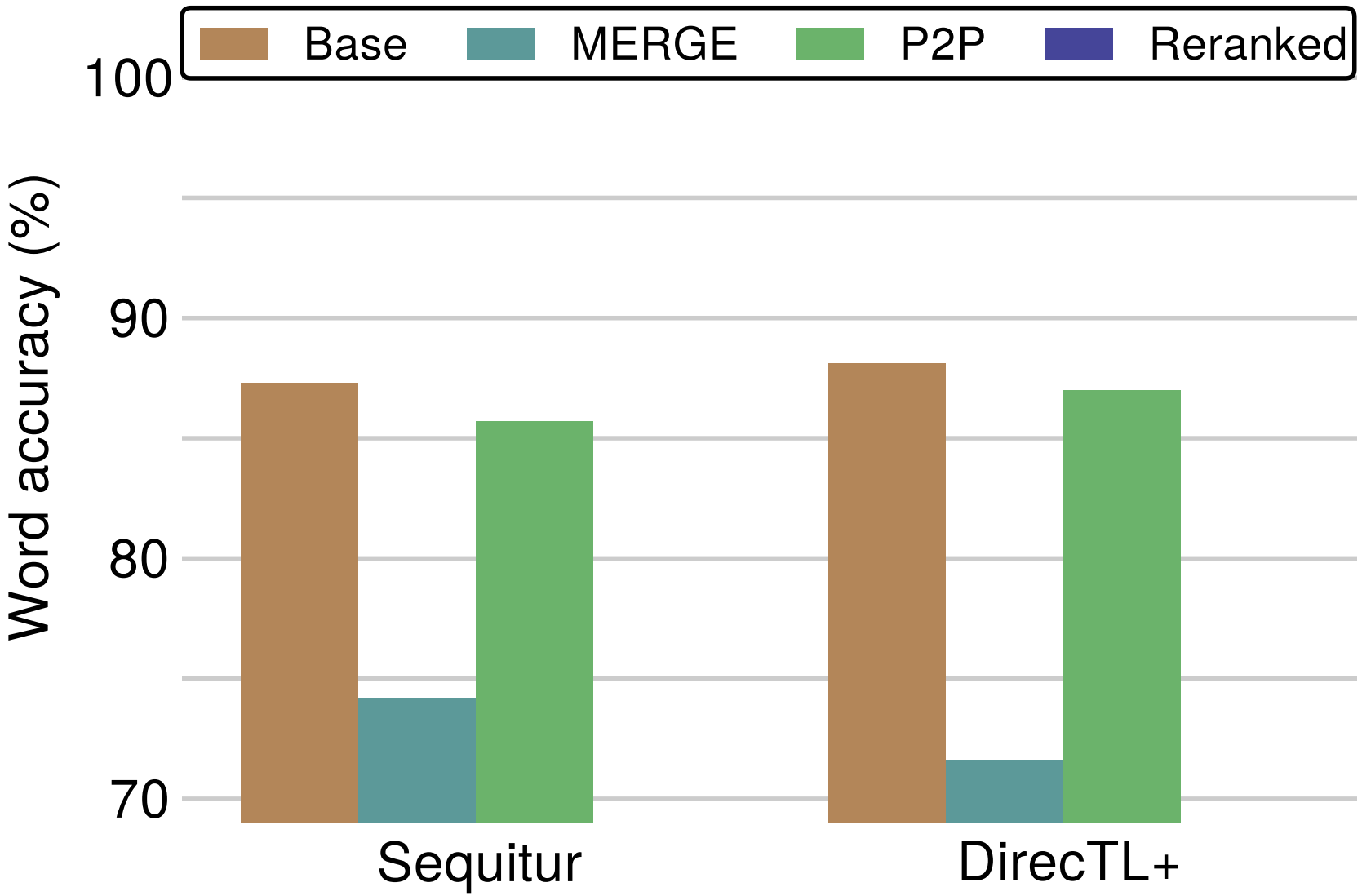
G2P experiments

Supplemental transcriptions: results



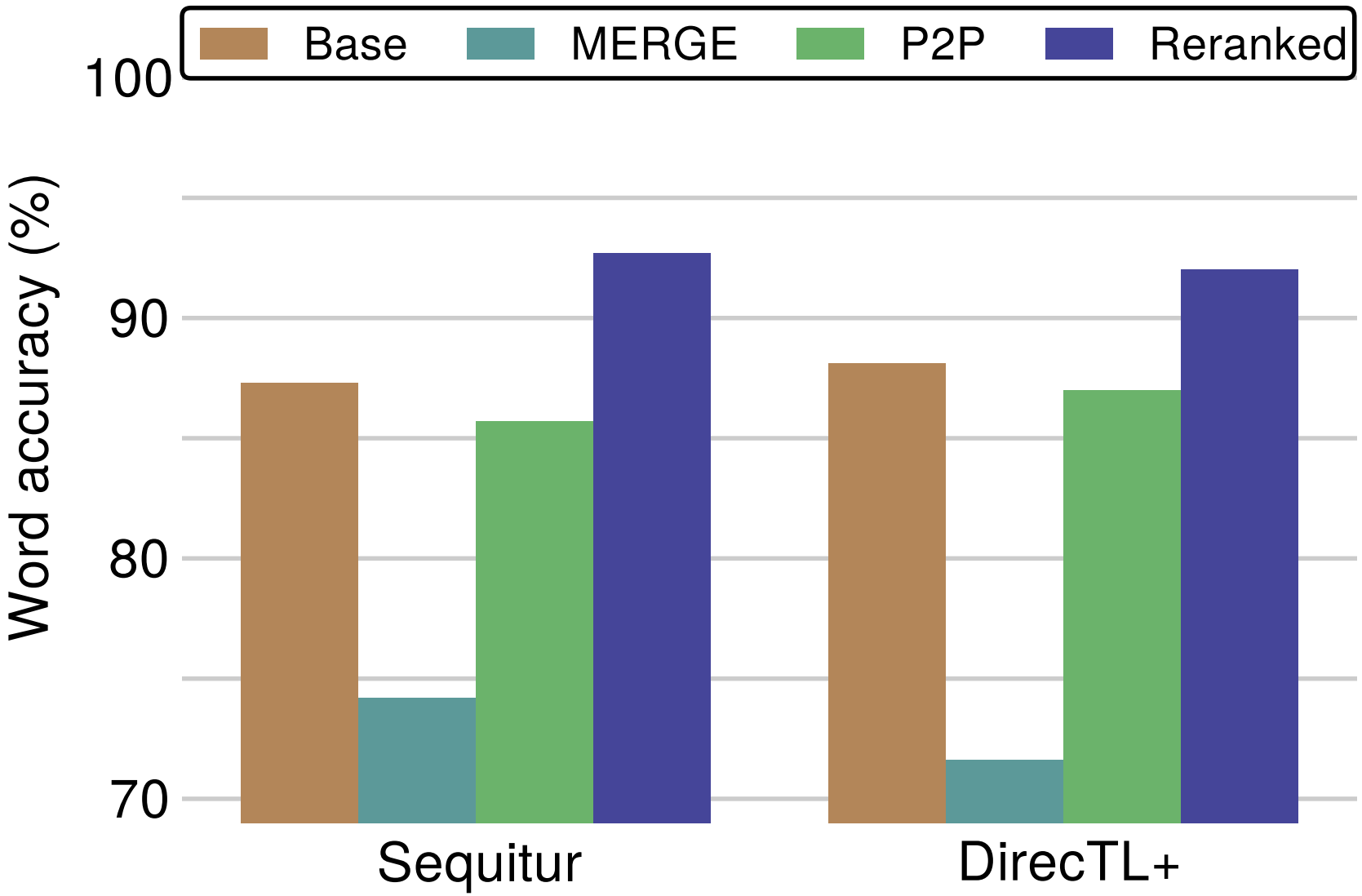
G2P experiments

Supplemental transcriptions: results



G2P experiments

Supplemental transcriptions: results



MTL experiments

Supplemental transliterations

input	**candidate outputs**
John Petrucci	जॉन पटरूसी
	जॉन पटरूची
	...
	जॉन पटरूक्सी

MTL experiments

Supplemental transliterations

input	**candidate outputs**
John Petrucci	जॉन पटरूसी
	जॉन पटरूची
	...
	जॉन पटरूक्सी

MTL experiments

Supplemental transliterations

input	**candidate outputs**	**supplemental**
John Petrucci	जॉन पटरूसी	ジョン ペトルーシ
	जॉन पटरूची	Джон Петруччи
	...	
	जॉन पटरूक्सी	

MTL experiments

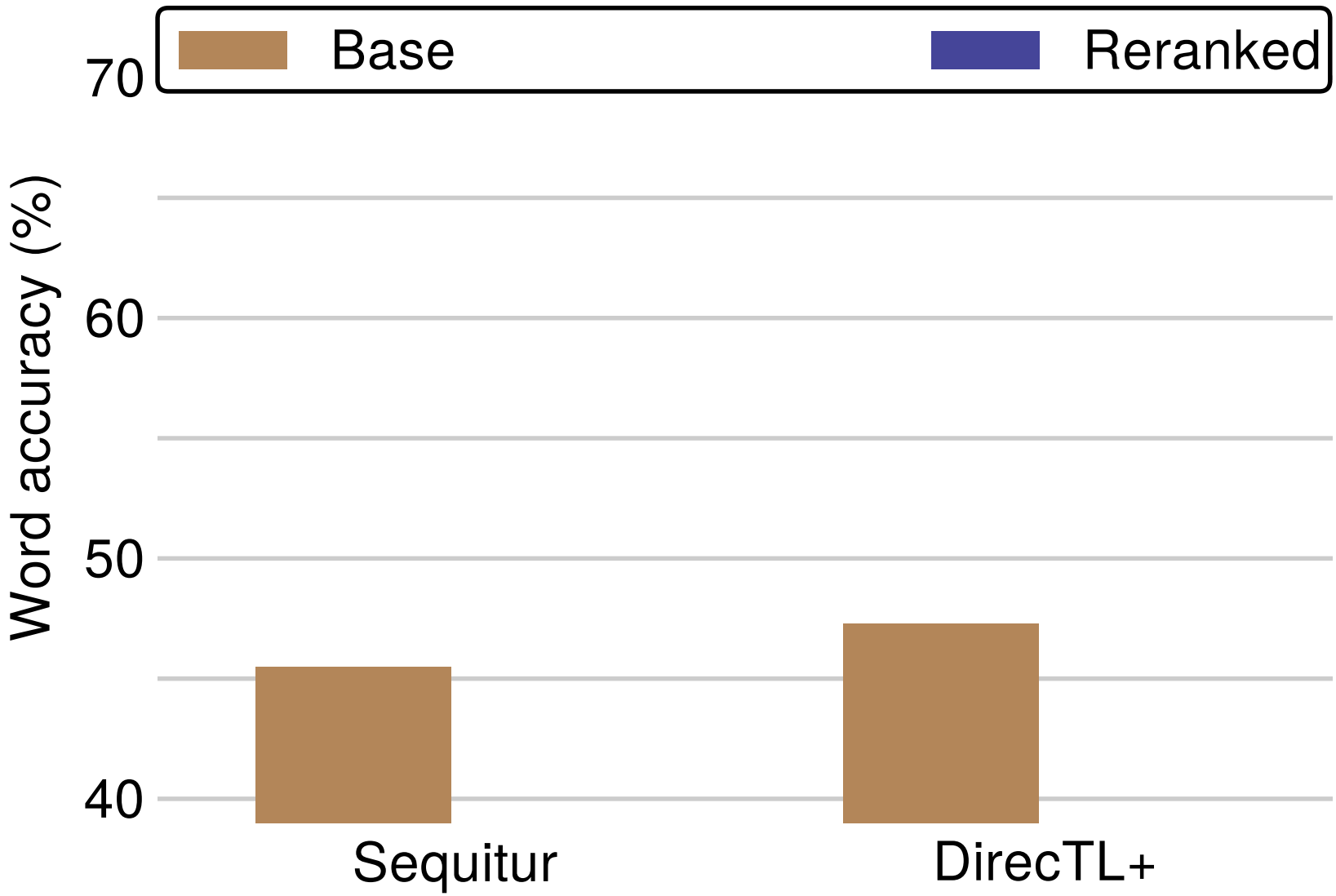
Supplemental transliterations

Wikipedia example

- John Petrucci article exists in English & Japanese, but not Hindi
- Want to automatically generate stub article in Hindi
  - Need transliteration of name
- Start from English, use Japanese (etc.) TLs to help generate Hindi TL

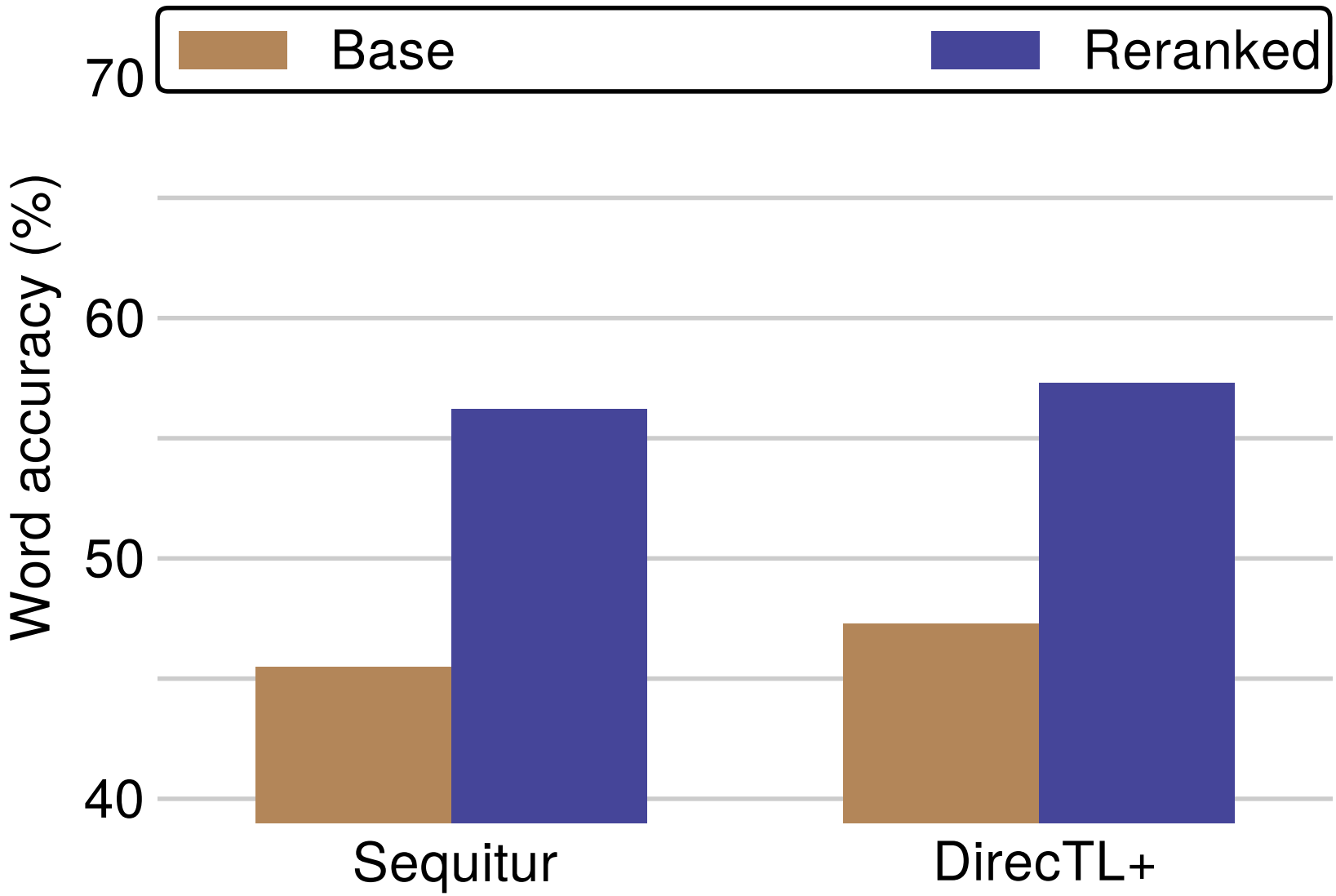
MTL experiments

Supplemental transliterations: results



MTL experiments

Supplemental transliterations: results



MTL experiments

Supplemental transcriptions

input	**candidate outputs**
Sudan	ズーダン
	スーダン
	...
	スユーダン

MTL experiments

Supplemental transcriptions

input	**candidate outputs**
Sudan	ズーダン
	スーダン
	...
	スユーダン

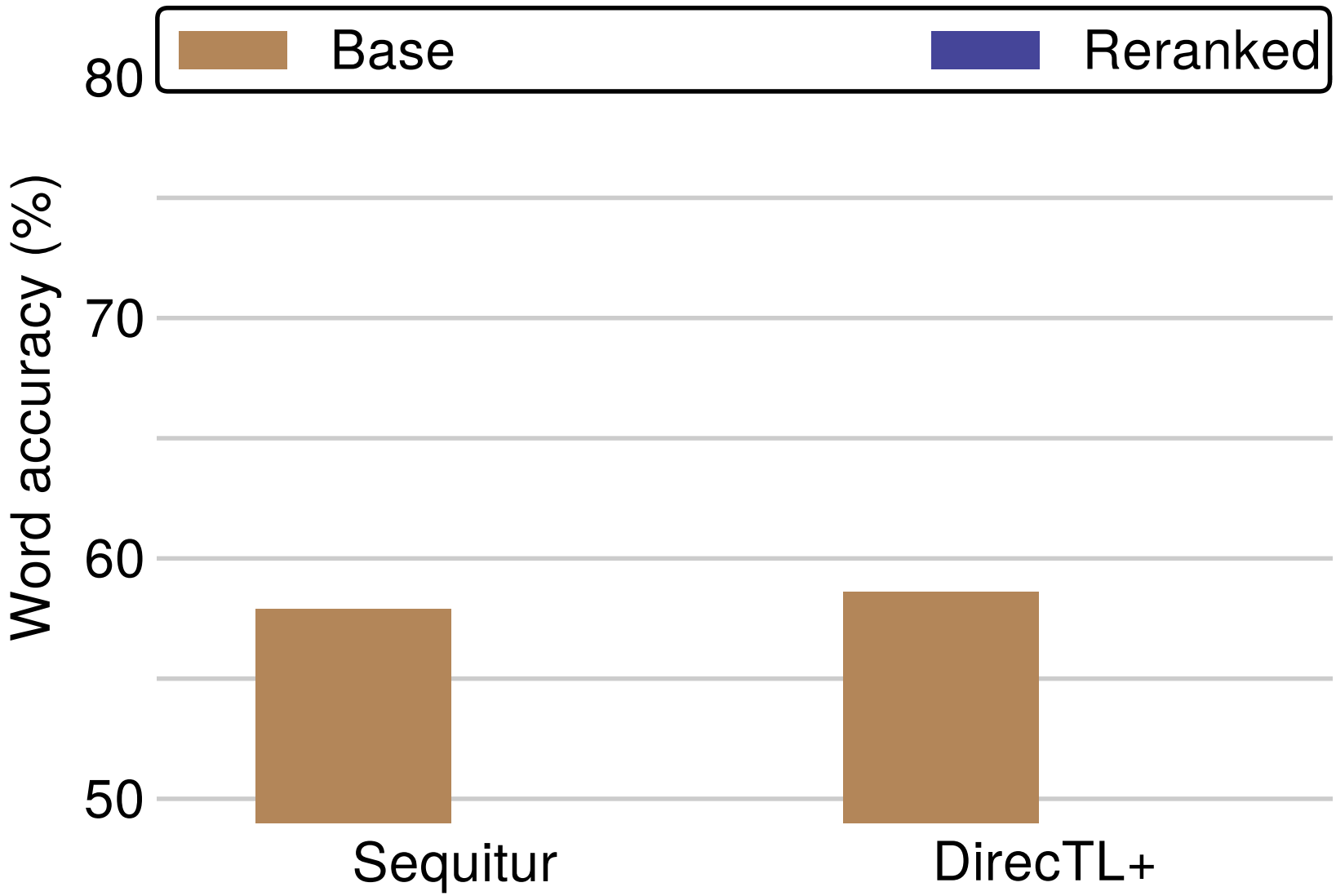
MTL experiments

Supplemental transcriptions

input	**candidate outputs**	**supplemental**
Sudan	ズーダン	`sud#n` **(CELEX)**
	スーダン	`sudAn` **(Combilex)**
	...	
	スユーダン	

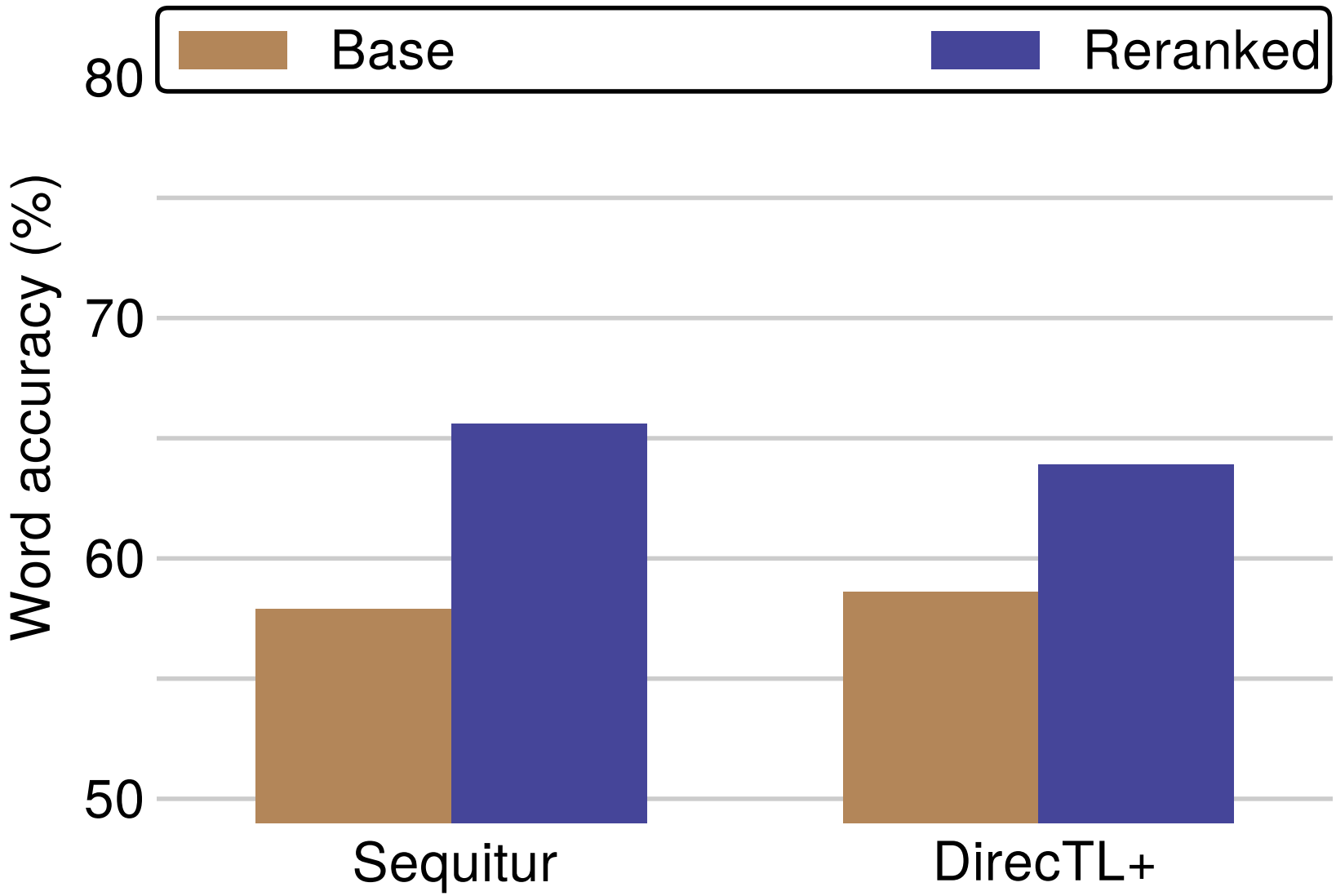
MTL experiments

Supplemental transcriptions: results



MTL experiments

Supplemental transcriptions: results



Analysis

- Method works across base systems, but magnitude of improvement varies
- Sequitur sees higher improvements
  1. Lower base score
  2. Higher oracle reranker score
  3. Reranking features are similar to those used in DirecTL+

Analysis

- Feature similarity indicates DirecTL+'s improvement comes from the supplemental representations, not new features

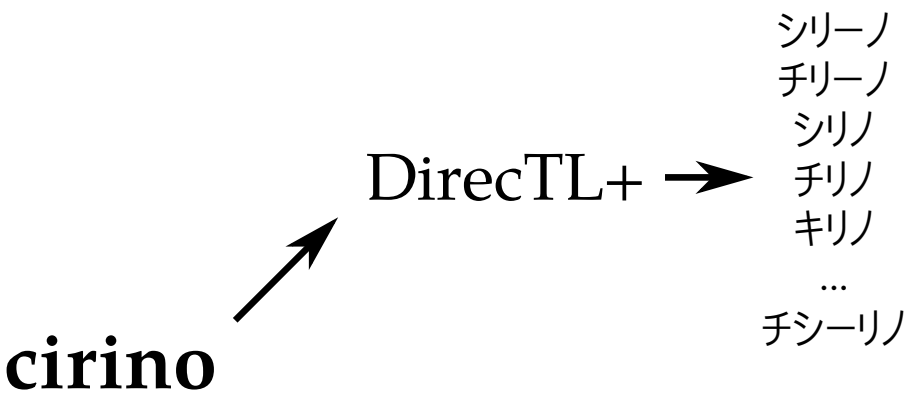
Analysis

- Using transcriptions and transliterations *simultaneously* doesn't provide any additional benefit

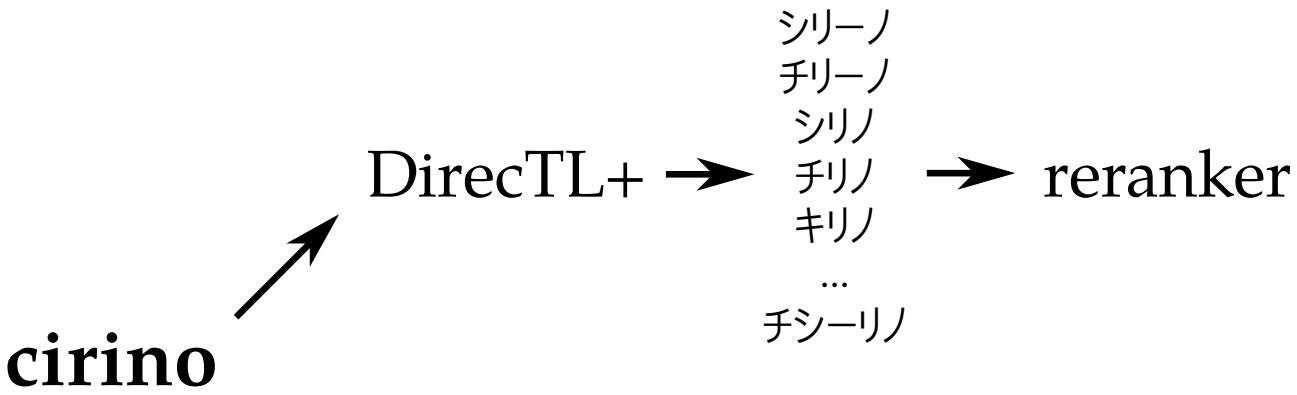
System combination

cirino

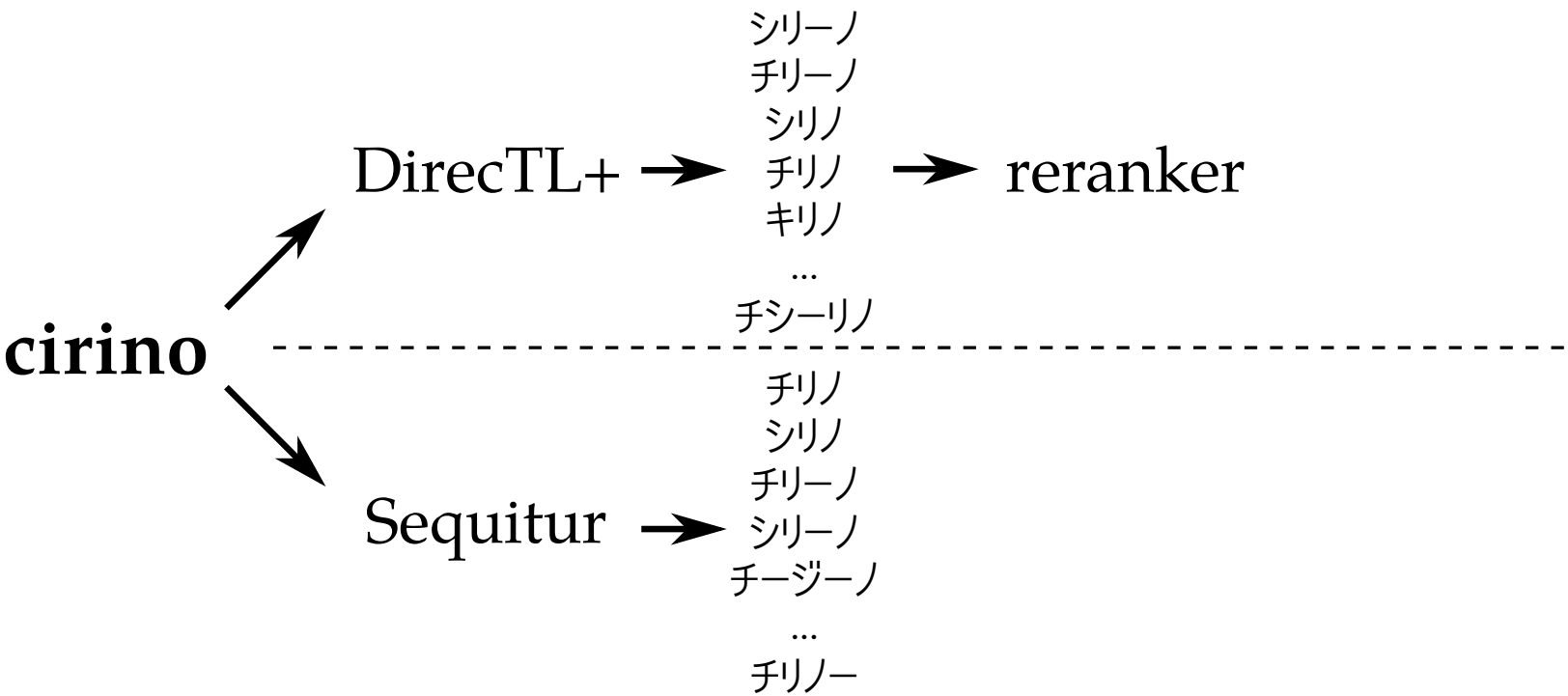
System combination



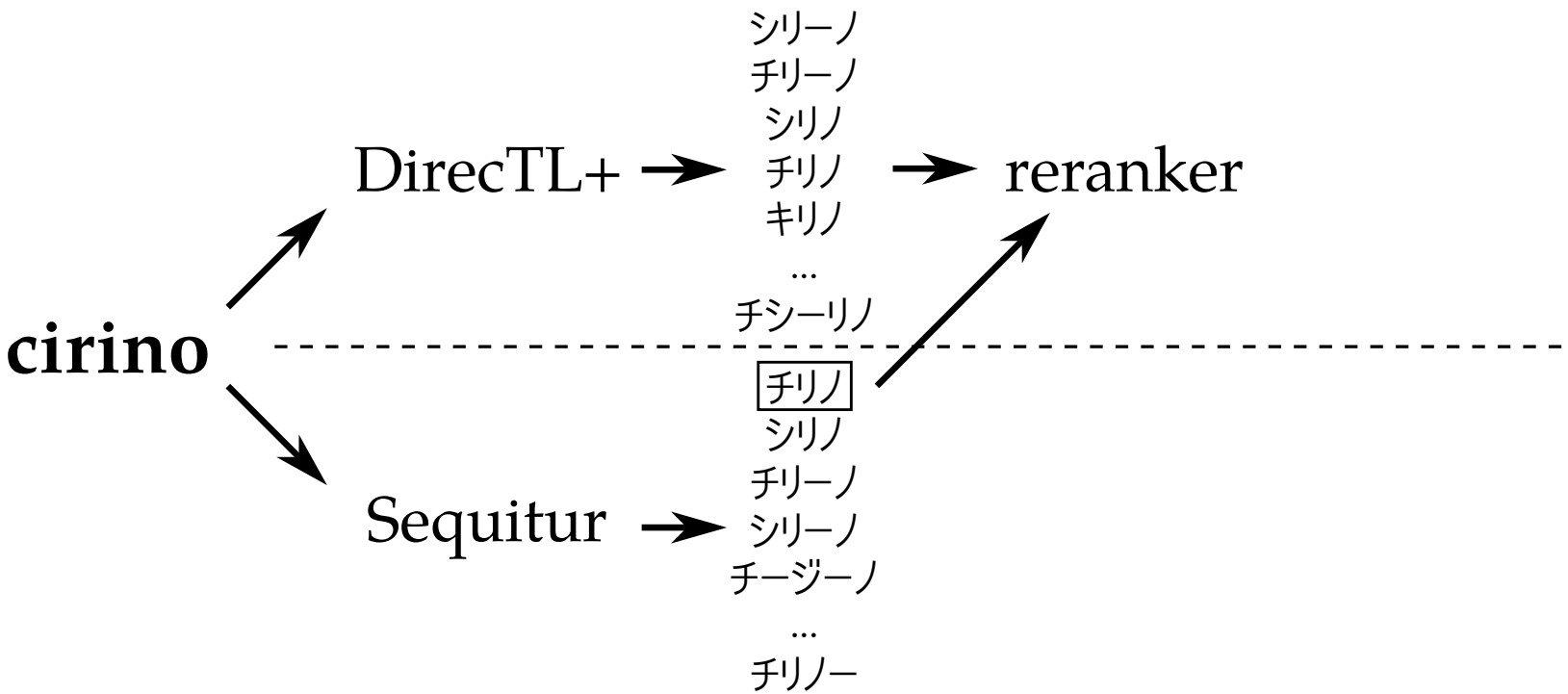
System combination



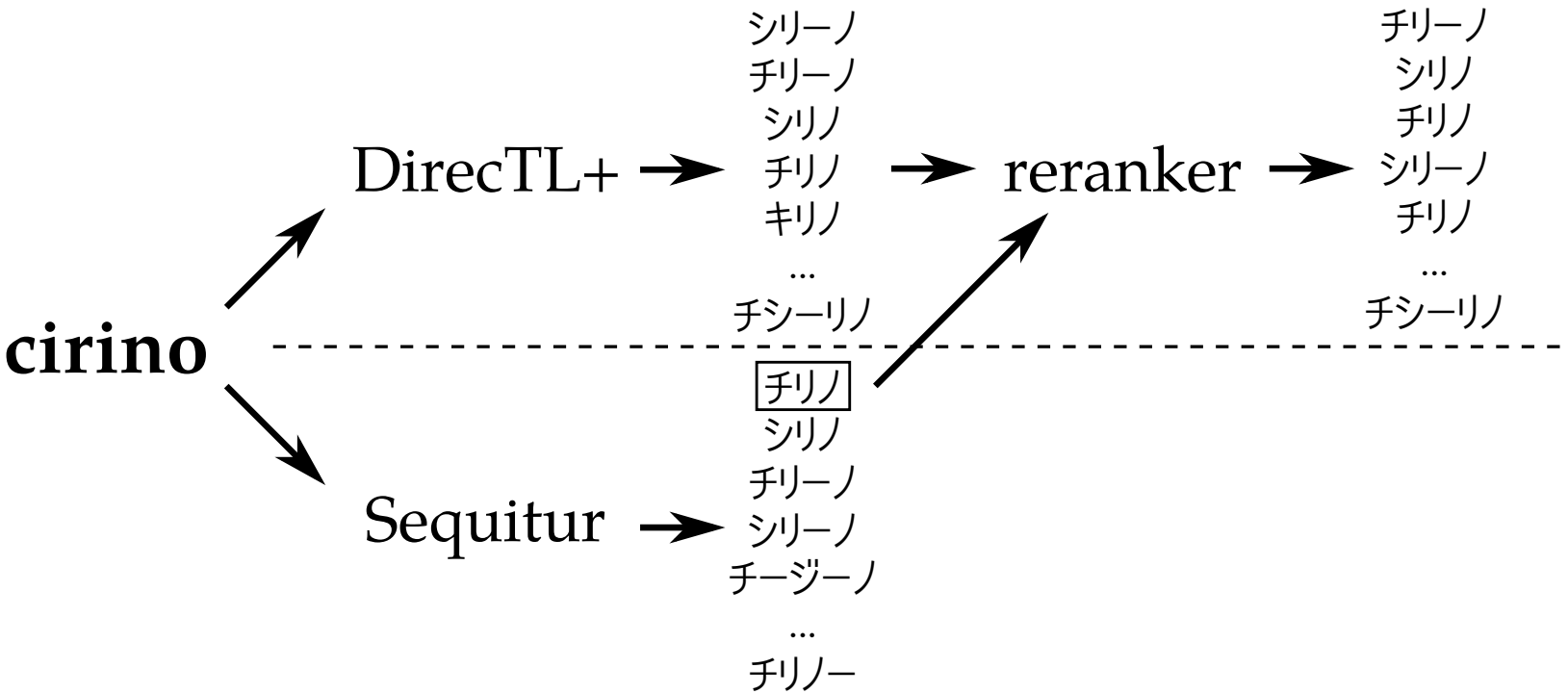
System combination



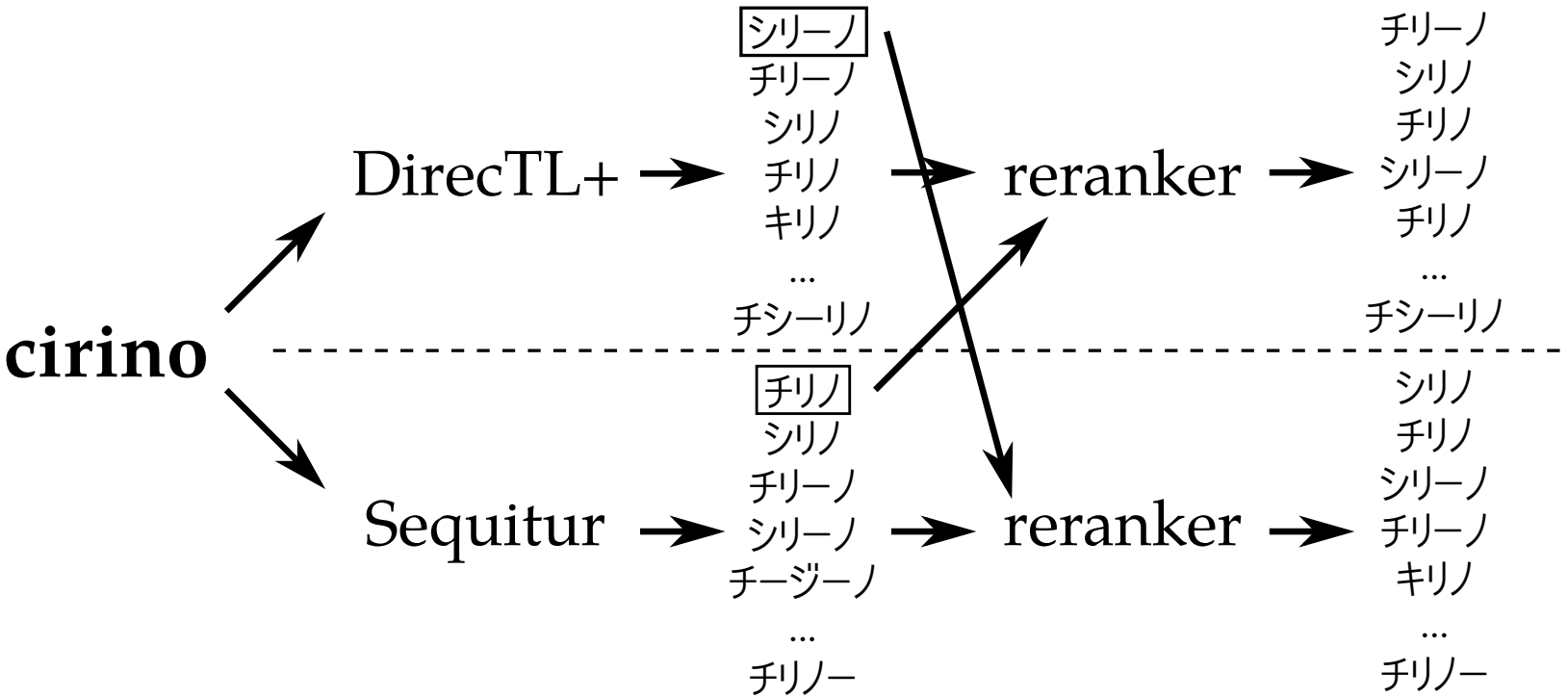
System combination



System combination



System combination



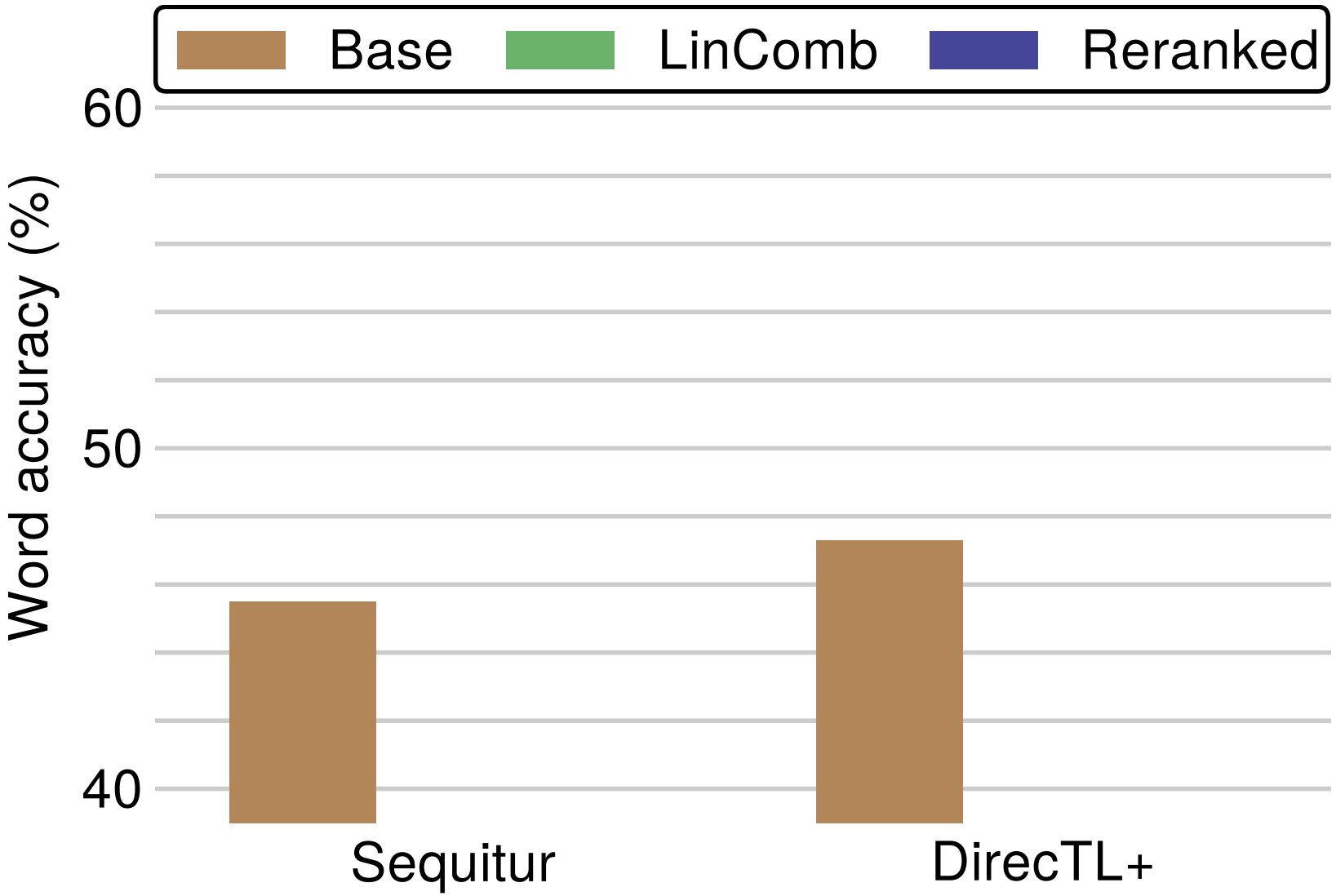
System combination

Baseline

- Linear combination baseline
  - Merge the base system lists
  - Linearly combine system scores
  - Manually tune linear parameter on training data

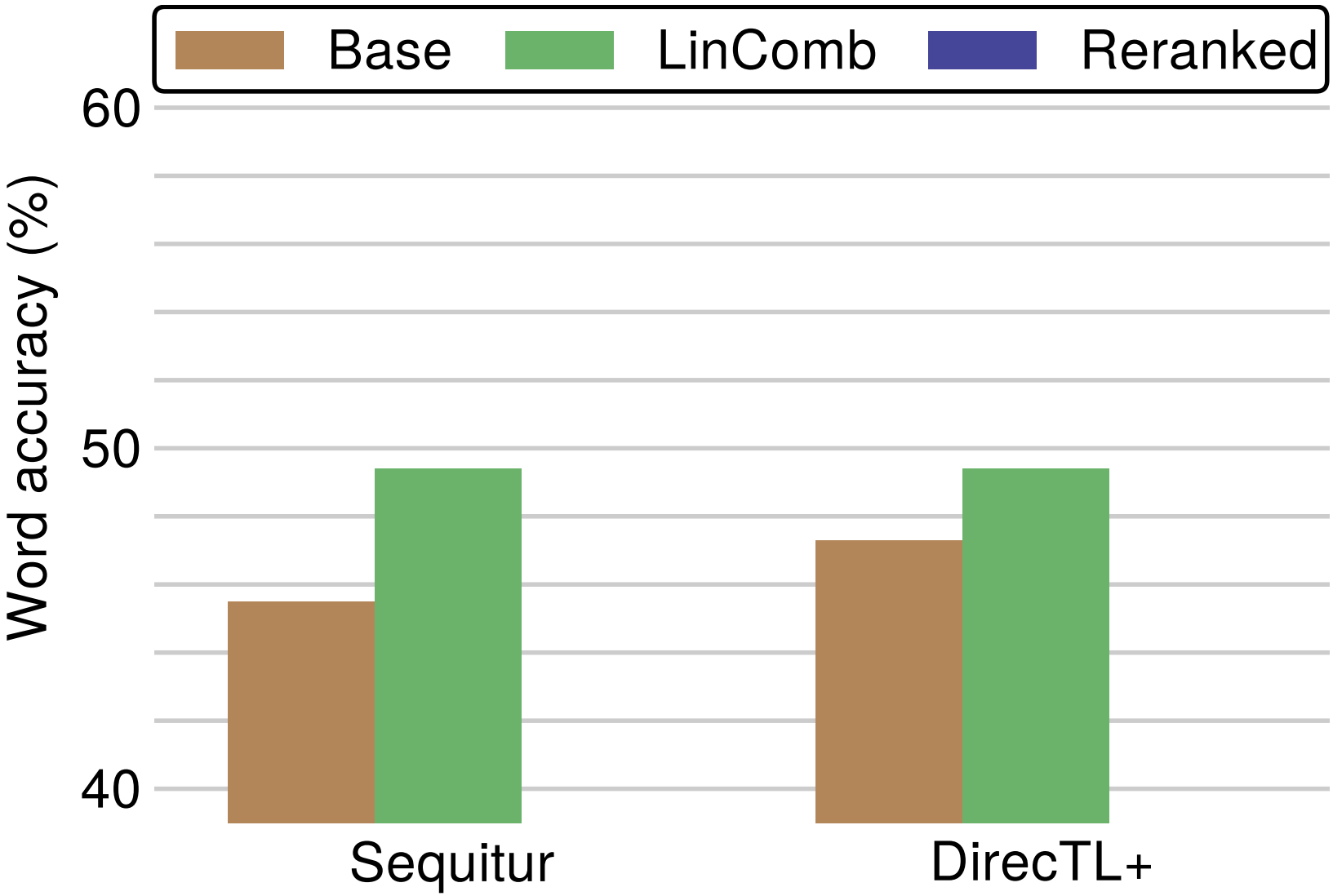
System combination

Results



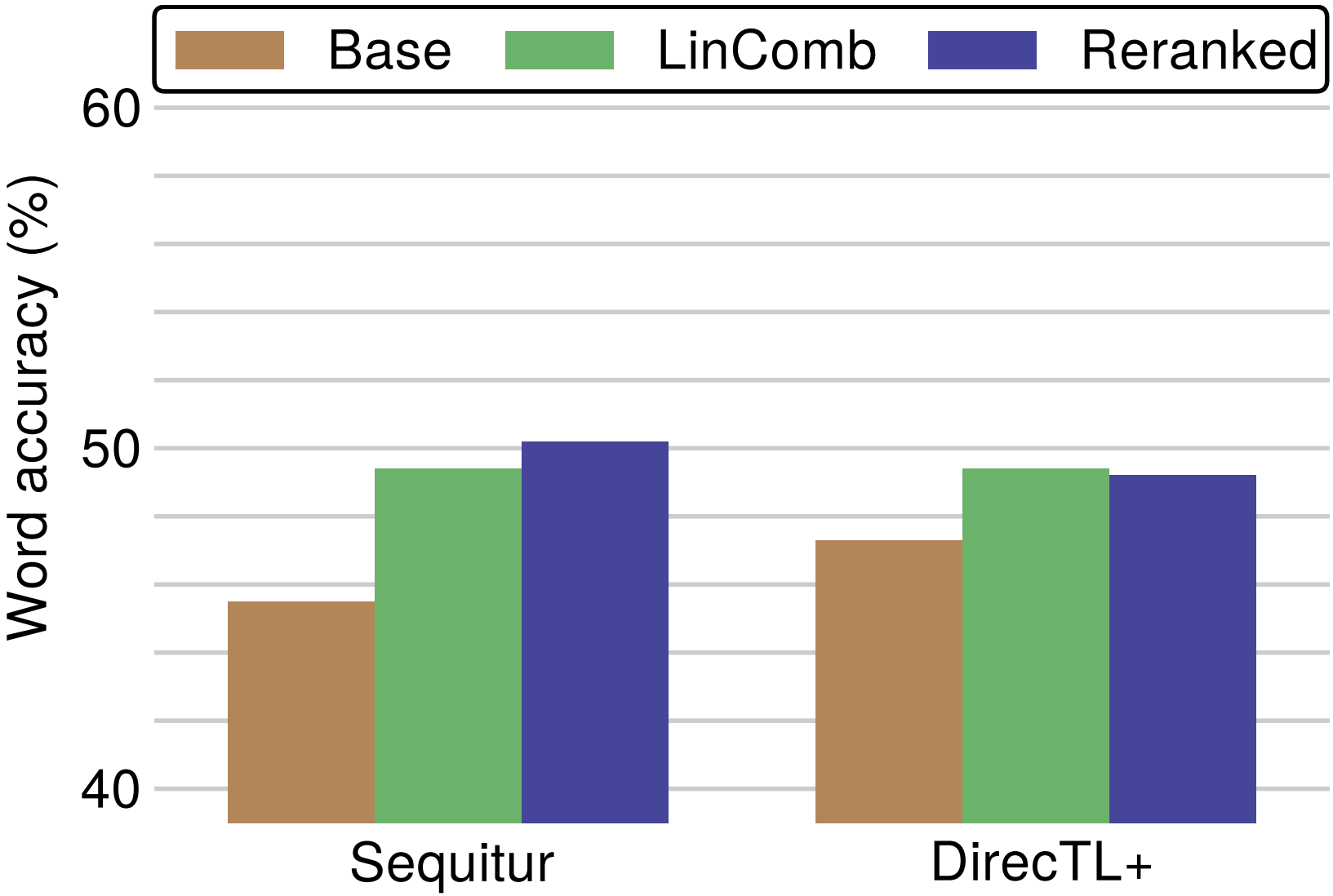
System combination

Results



System combination

Results



Summary

- Reranking approach effectively leverages supplemental transcriptions and transliterations for G2P and MTL
- Improvements across two base systems demonstrates that there is inherently useful information in the supplemental representations
- Treating another system's output as supplemental data works, but so does a linear combination

Future work

- Reranking is *post hoc*; direct integration might be more effective
- Incorporate supplemental *information* rather than *data*
- Other (noisy?) supplemental sources
  - Wikipedia IPA transcriptions
  - *Ad hoc* approximately-phonetic re-spellings