Target Conditioned Sampling:

Optimizing Data Selection for Multilingual NMT

Xinyi Wang, Graham Neubig

Language Technologies Institute
Carnegie Mellon University

NEULAB

Language Technologies Institute

Multilingual NMT

glg: A mañá que eu nunca vou

spa: Una mañana que nunca olvidaré .

por: Uma manhã que nunca vou esquecer .

ita: Una mattina che non dimenticherò mai .

jpn:その日の朝のことは 決して忘れることはないでしょう

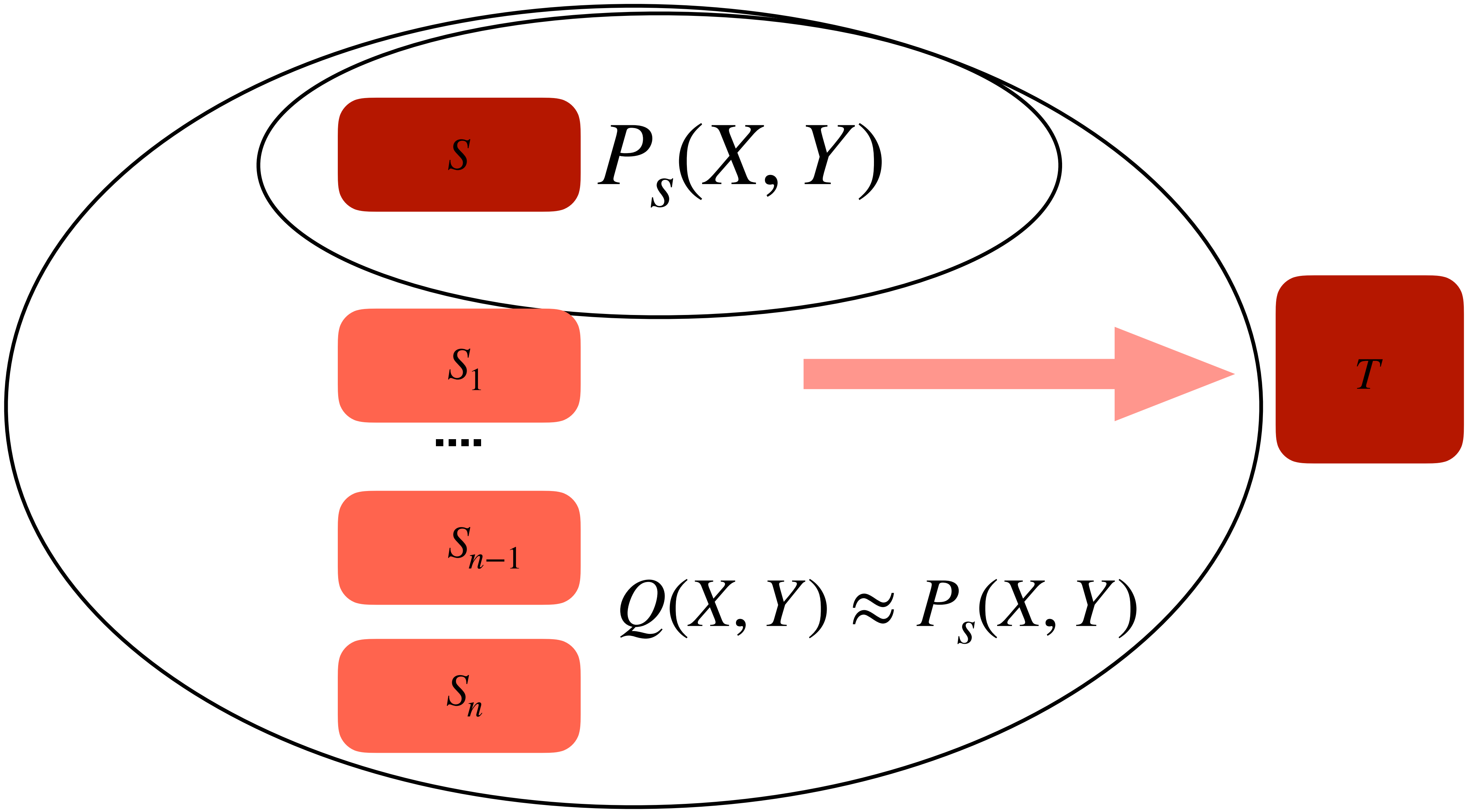
A morning that I will never forget .

- Particularly useful for low-resource languages (LRLs), such as Galician (glg)

Multilingual Training Paradigms

- Multi-lingual training (Dong et al. 2015, Firat et al. 2016)
- Train on related high-resource language, tune towards LRL (Zoph et al. 2016)
- Train on multilingual data, tune towards LRL (Neubig and Hu 2018, Gu et al. 2018)
- **Our proposal:** can we more intelligently select data in a less heuristic way?

Multilingual Objective for LRL NMT



- How to construct the $Q(X, Y)$?

Target Conditioned Sampling

union of targets

A morning that I will never forget.
....
When I was 11, I usually stay with

$Q(Y)$

A morning that I will never forget.

spa: Una mañana....
por: Uma manhã ..
ita: Una mattina ...
jpn:その日の朝...

$Q(X|y)$

Sampled Data

por: Uma manhã ..
A morning that I will never forget.

Choosing the Distributions

- $Q(Y)$
  - assume each language data comes from same domain
  - **uniform sample** from all target y can match $P_s(Y)$
- $Q(X|y)$
  - $P_s(X = x|y)$ measures how likely x is in language s
  - Approximate using **heuristic similarity measure** $sim(x, s)$, normalize over all multilingual x_i for a given target y

Estimating $sim(x, s)$

	Vocab Overlap	Language Model
Language Level	character n-gram between *S* and each language	score document of each language
Sentence Level	character n-gram between *S* and each sentence	use LM on *S* to score each sentence

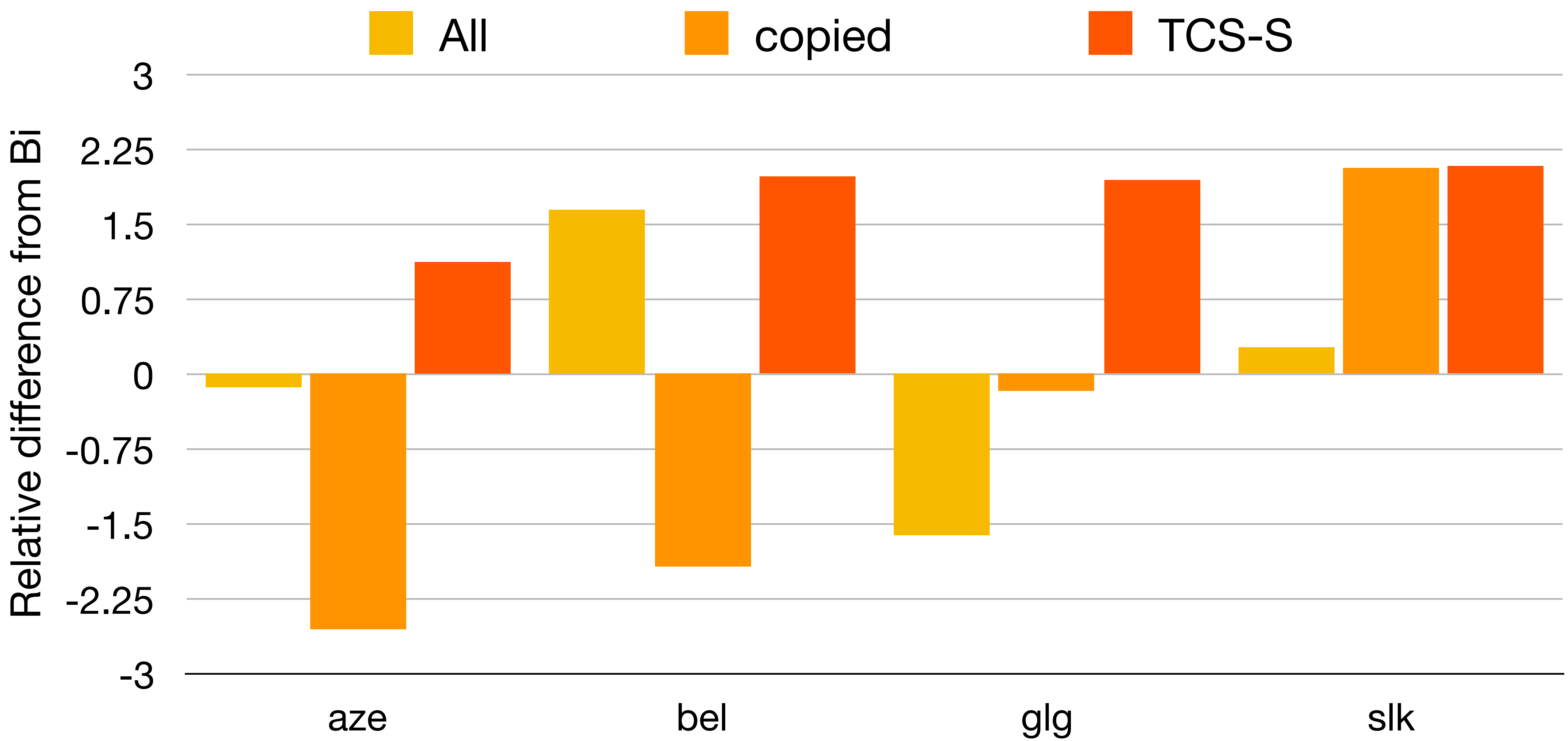
Algorithms

- First sample y based on $Q(Y)$, then sample (x_i, y) based on $Q(X \mid y)$
- **Stochastic (TCS-S):**
  - dynamically sample each mini batch
- **Deterministic (TCS-D):**
  - select $x' = \text{argmax}_x Q(x \mid y)$, fixed during training

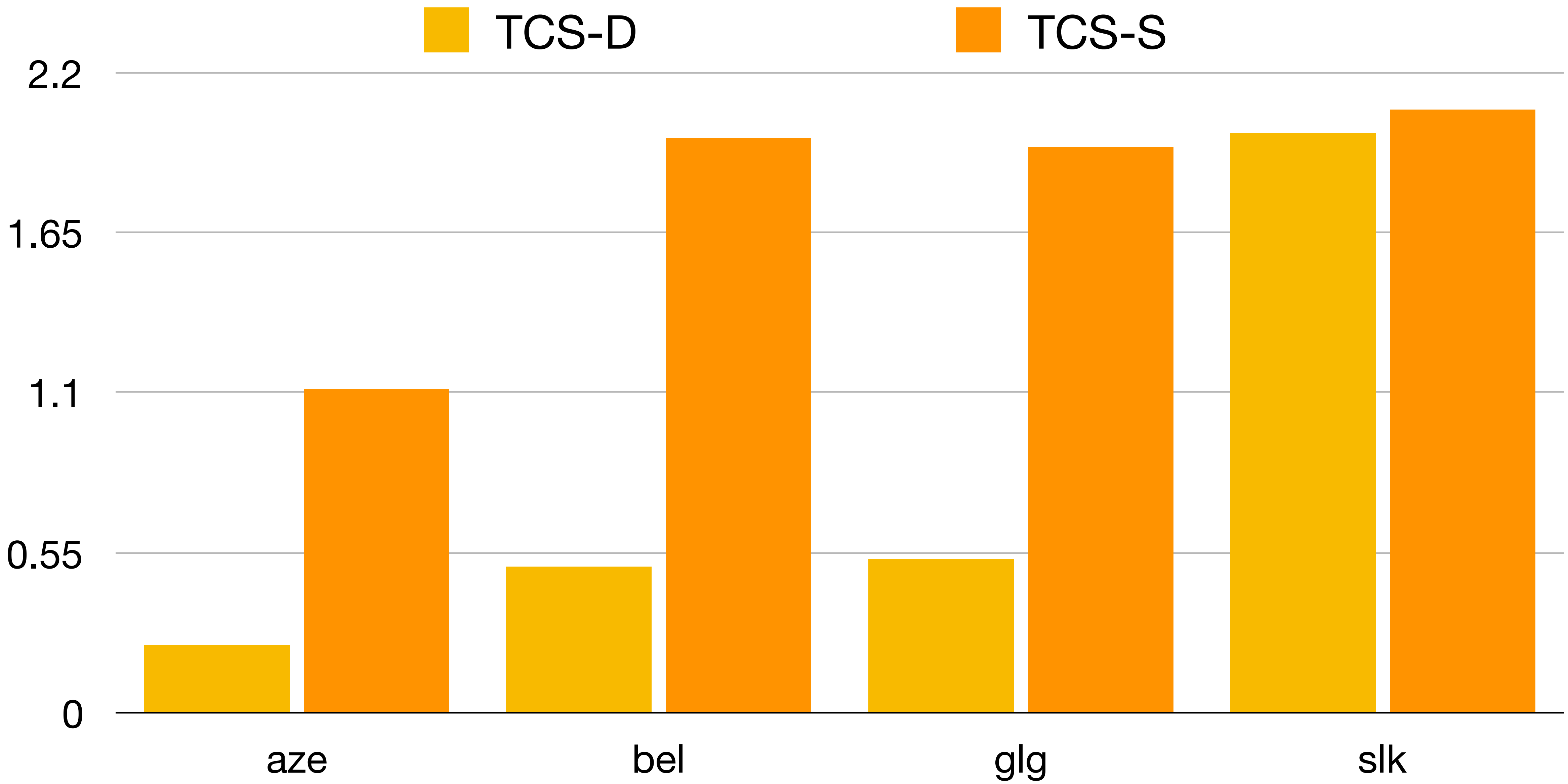
Experiment

- **Dataset**
  - 58-language-to-English TED dataset (Qi et al., 2018)
  - 4 test languages: Azerbaijani (aze), Belarusian (bel), Galician (glg), Slovak (slk)
- **Baselines**
  - **Bi:** each LRL paired with **one** related HRL (Neubig & Hu 2018)
  - **All:** train on all 59 languages
  - **Copied:** use union of English sentences as monolingual data by copying them to the source (Currey et al. 2017)

TCS vs. Baselines

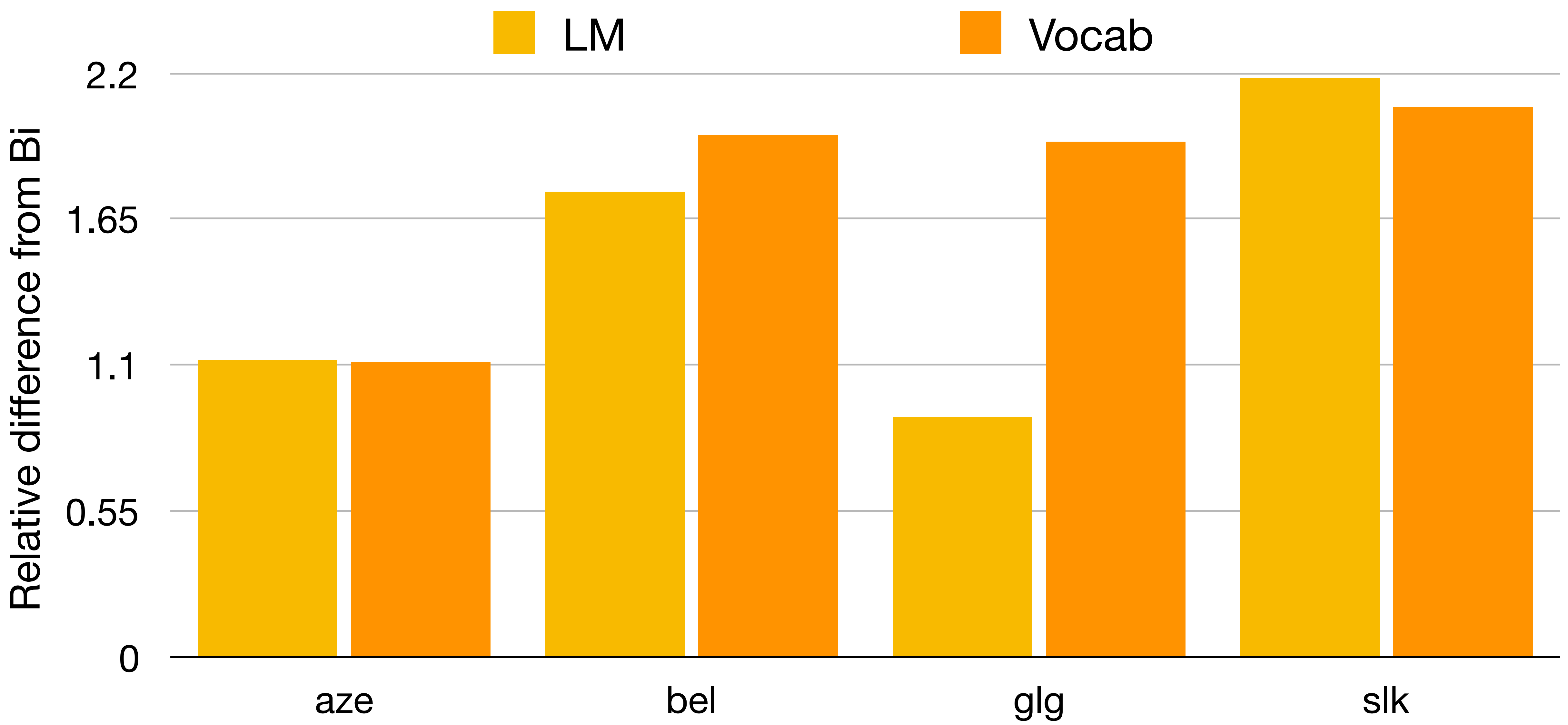


TCS-D vs. TCS-S



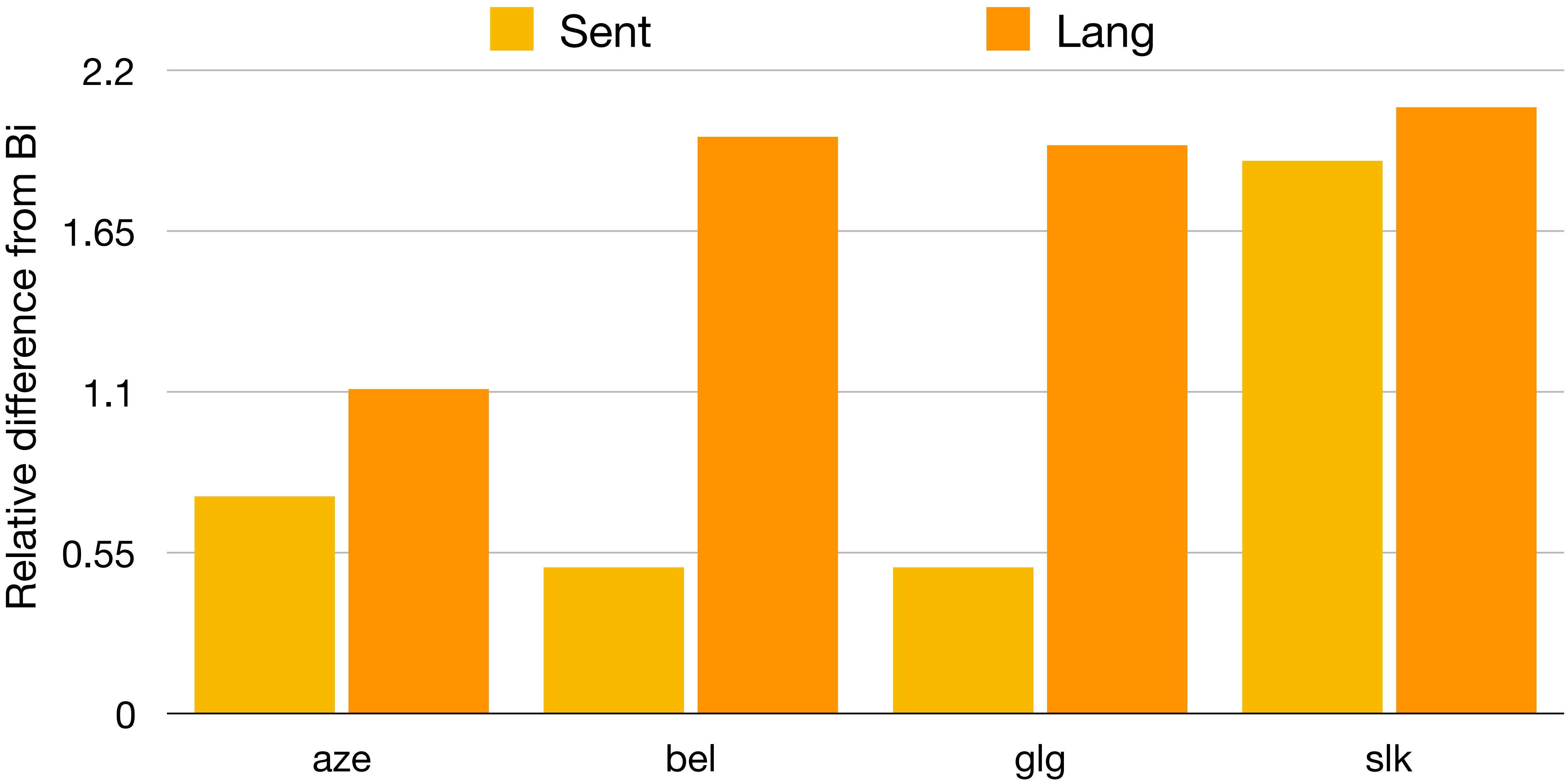
- TCS-D already brings gains, TCS-S generally performs better

LM vs. Vocab



- Simple vocab overlap heuristic is already competitive
- LM performs better for slk, with highest amount of data

Sent vs. Lang



- Language level heuristic is in general better

Conclusion

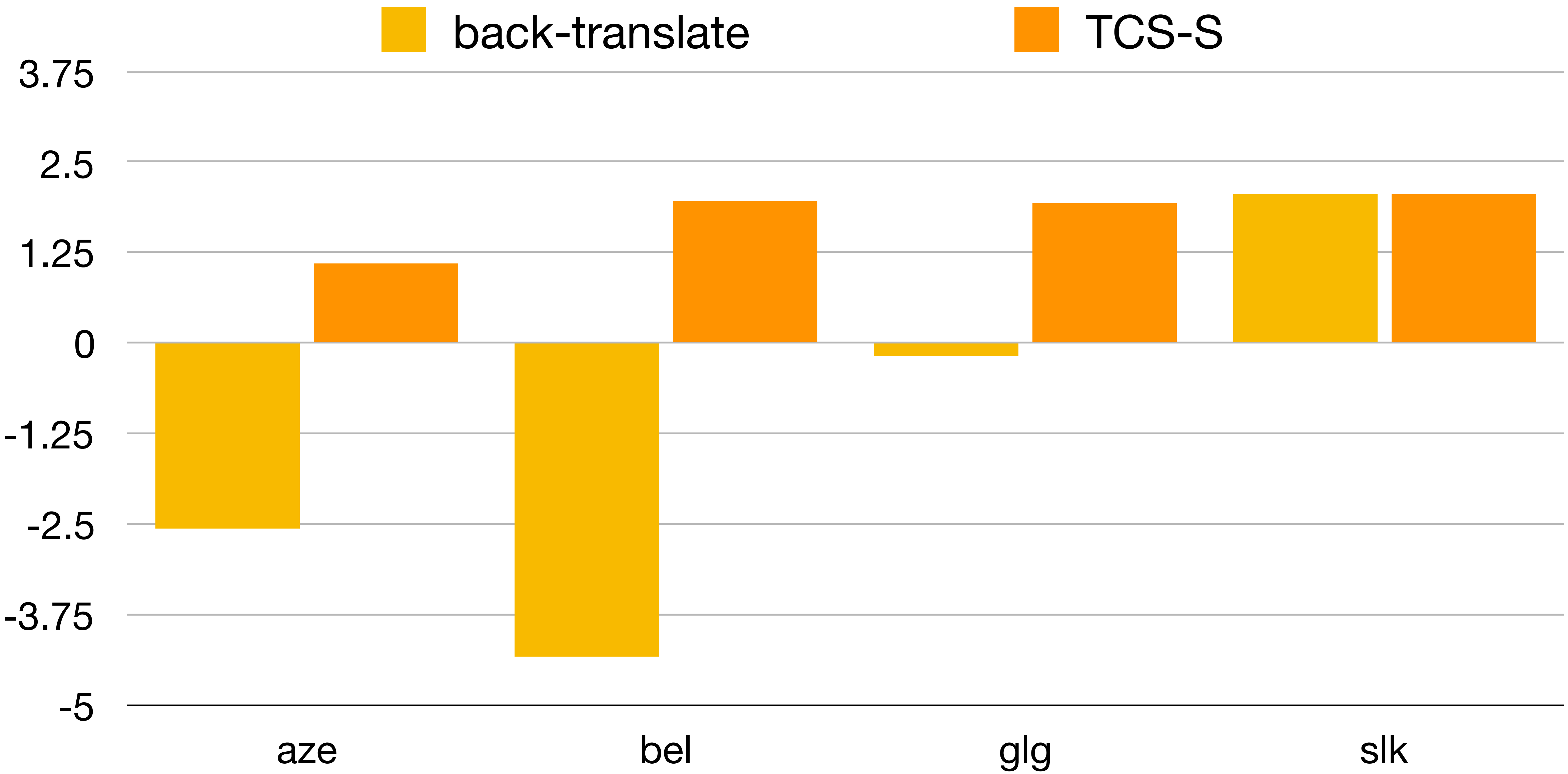
- TCS is a simple method for better multi-lingual data selection
- Brings significant improvements with little training overhead
- Simple heuristics work well for LRLs to estimate language similarity

https://github.com/cindyxinyiwang/TCS

Thank You! Questions?

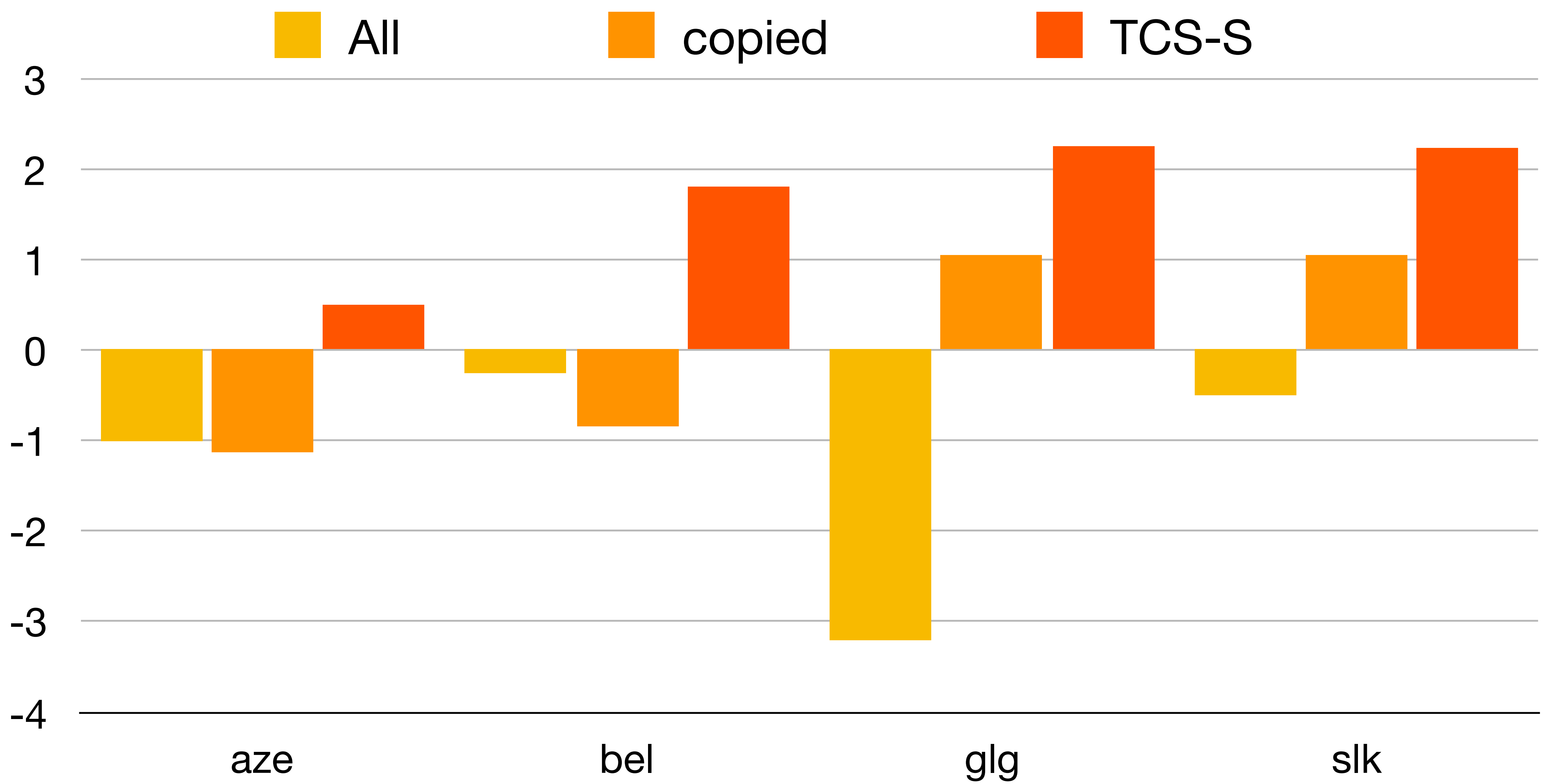
Extra Slides

Relationship with Back-Translation



- TCS approximates back-translate probability $P_s(X \mid y)$
- For LRL, heuristics performs better than back-translate model

Effect on SDE



- SDE: a better word encoding designed for multilingual data (Wang et. al. 2019)
- TCS still brings significant gains on top of SDE