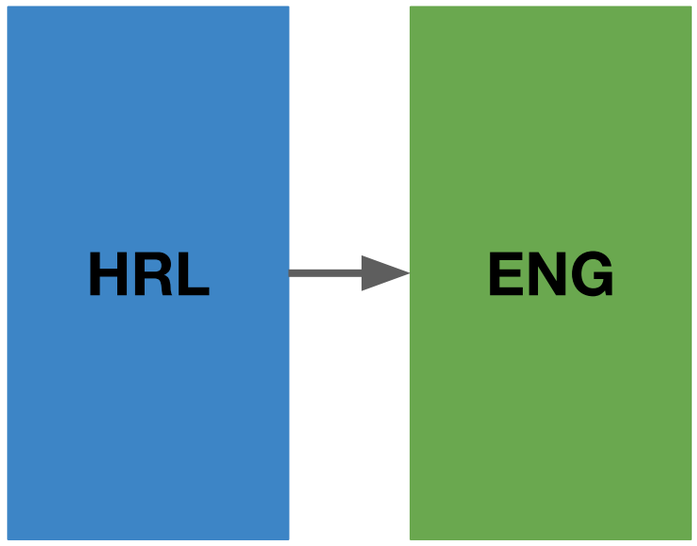
Generalized Data Augmentation for Low-Resource Translation

Mengzhou Xia, Xiang Kong,
Antonios Anastasopoulos, Graham Neubig

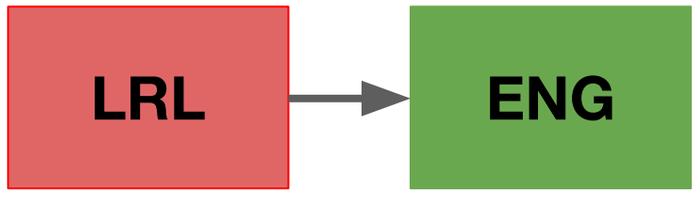
Language Technologies Institute
Carnegie Mellon University

Challenges in Low-resource MT

- MT of high-resource languages (HRLs) with large parallel corpora → good translations



- MT of low-resource languages (LRLs) with small parallel corpora → nonsense!



A Concrete Example

A system that is trained with **5000** sentence pairs on Azerbaijani and English ?

source - Atam balaca boz radiosunda BBC Xəbərlərinə qulaq asırdı.

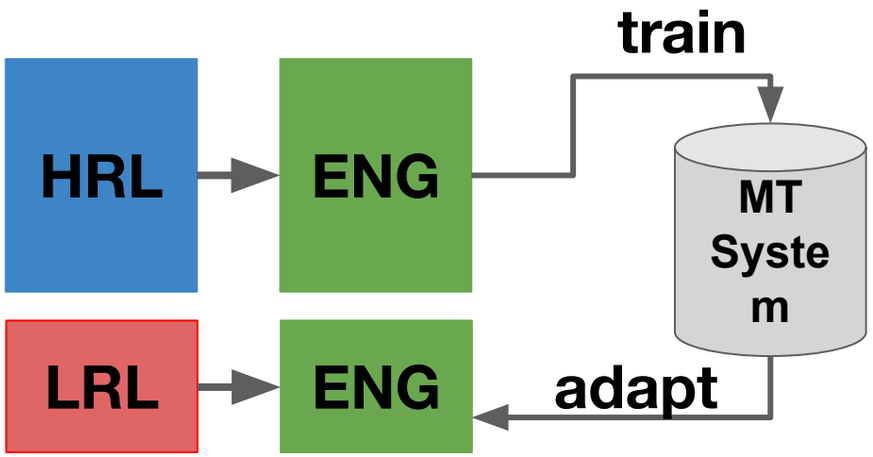
translation - So I’m going to became a lot of people.

reference - My father was listening to BBC News on his small , gray radio.

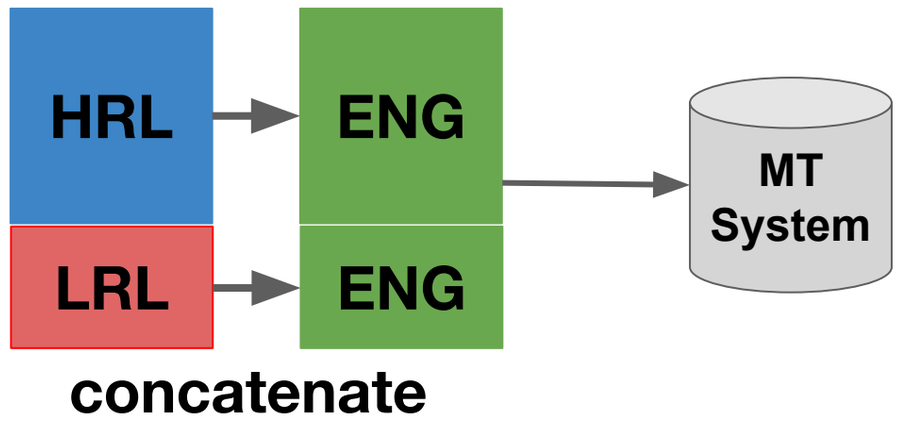
Does not convey the correct meaning at all.

Standard Approaches (1)

- Transfer HRL to LRL
  (Zoph et al., 2016; Nguyen and Chiang, 2017)



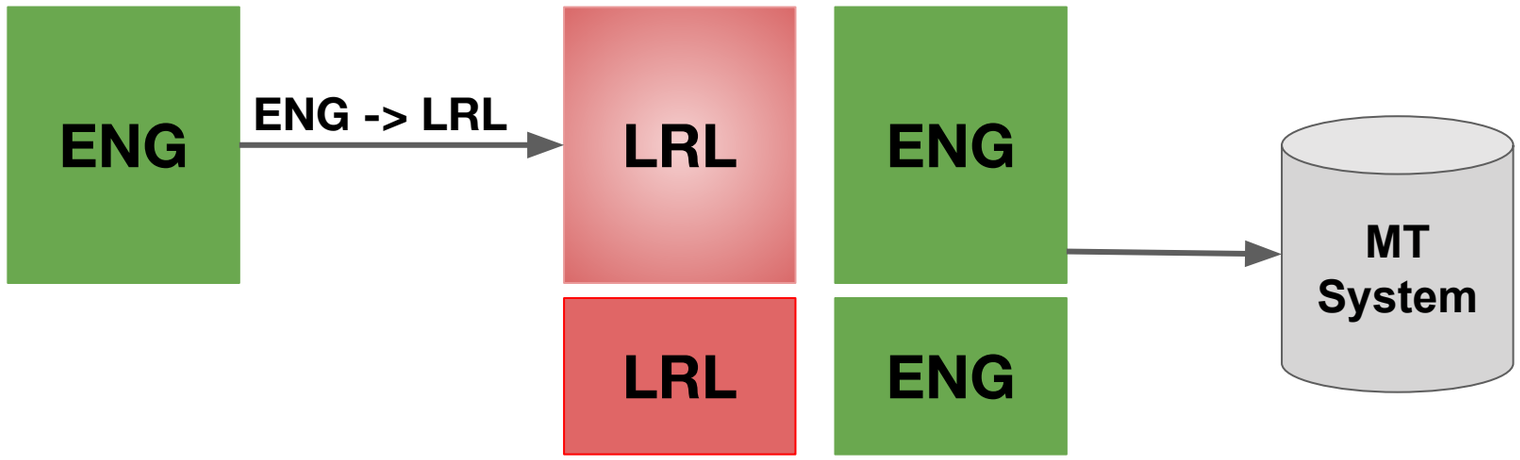
- Joint training with LRL and HRL parallel data
  (Johnson et al., 2017; Neubig and Hu, 2018)



- **Problems:** Suboptimal lexical/syntactic sharing.
  - Azerbaijani (LRL) word - zəfərin
  - Turkish (HRL) word - zaferin

Standard Approaches (2)

- Back translation (Sennrich et al. 2016)



- **Problems:** Poor-quality ENG->LRL system results in poor data.

This Work

- **Question:** Is there a better way of performing data augmentation for low-resource MT?

- **Contributions:**
  - A *generalized framework* for utilizing training data in low-resource MT.
  - New methods for *pivoting through related HRLs* to generate pseudo-parallel data.
  - An *extensive empirical study* comparing these methods, with gains of up to 1.5-8 BLEU.

Available Resources

ENG-M

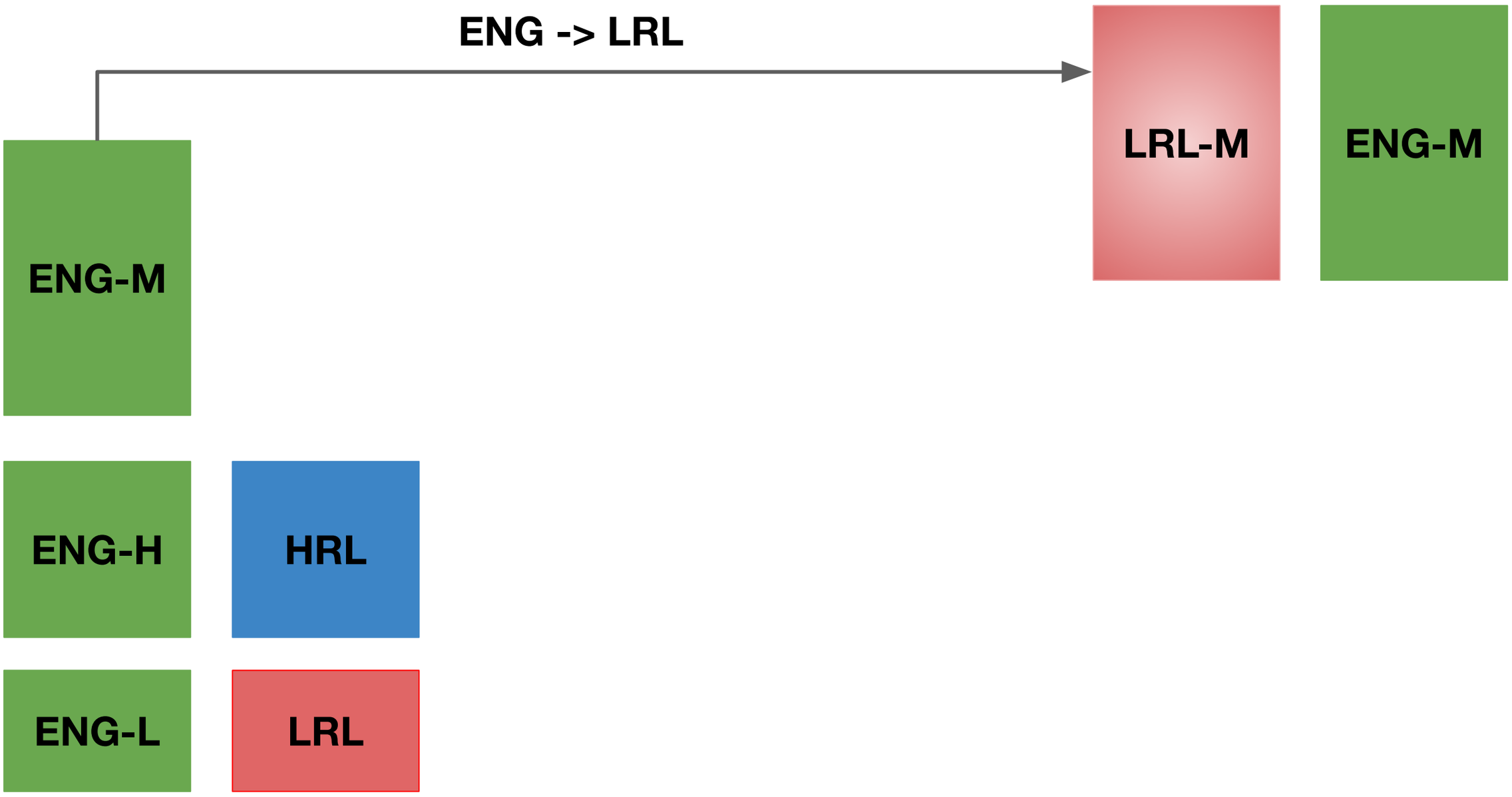
ENG-H

HRL

ENG-L

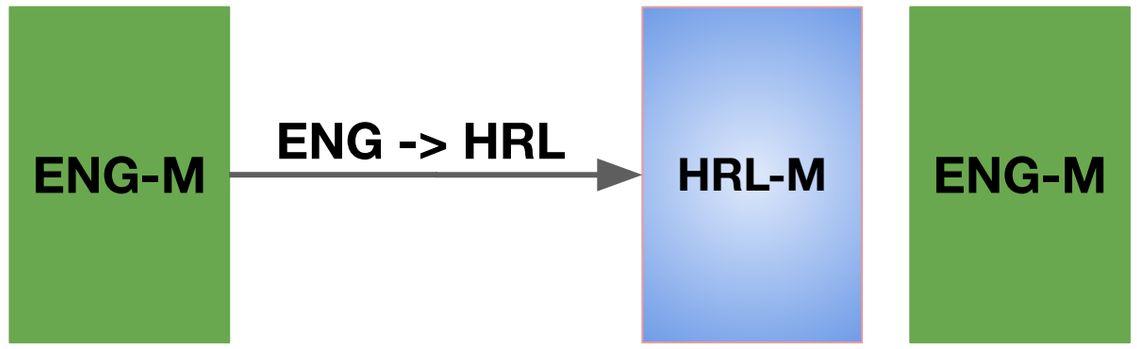
LRL

Available Resources + ENG-LRL Back-translation



Proposal 1: English -> HRL Augmentation

- **Problem:** ENG-LRL back-translation might be low quality



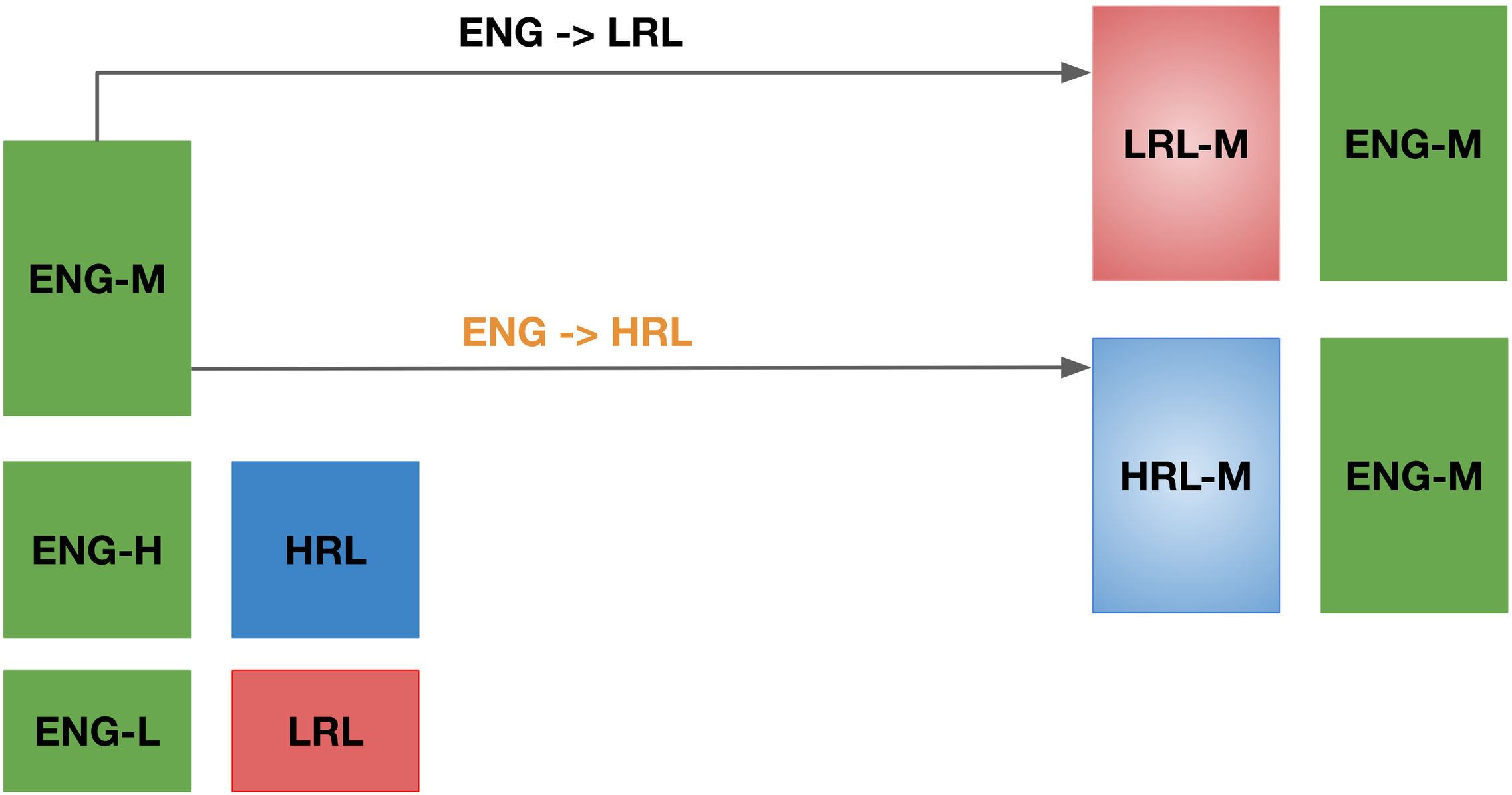
- **Idea: also back-translate into HRL**
  - more sentence pairs
  - vocabulary sharing of source-side
  - syntactic similarity of source-side
  - improves target-side LM

ENG: Thank you very much.

~~**AZE**: Ho Ho Ho.~~

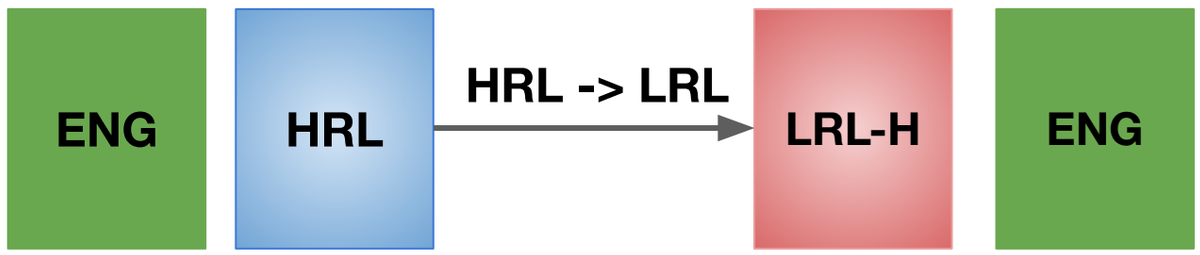
TUR: Çok teşekkür ederim.

Available Resources + ENG-LRL and ENG-HRL Back-translation



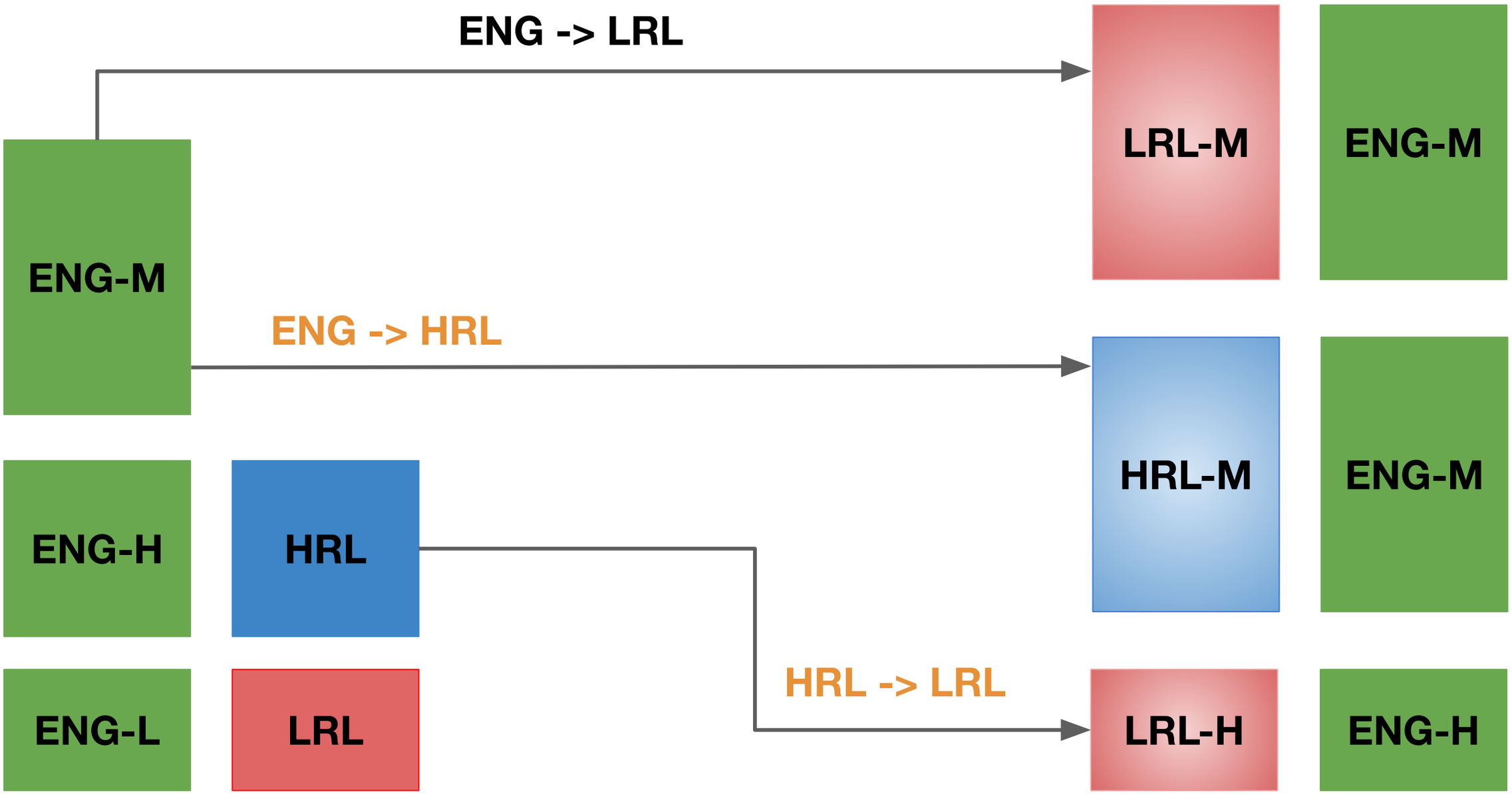
Proposal 2: Augmentation via Pivoting

- **Problem:** HRL-ENG data might suffer from lack of lexical/syntactic overlap
- **Idea: Translate existing HRL-ENG data**
  - Translate from HRL to LRL



TUR: Çok teşekkür ederim. → **AZE**: Çox sağ olun.
ENG: Thank you so much. **ENG**: Thank you so much.

Available Resources + ENG-LRL and ENG-HRL
Back-translation + Pivoting



Proposal 3: Back-Translation by Pivoting

- **Problem:** ENG-HRL back-translated data also suffers from lexical or syntactic mismatch

ENG: Thank you so much.

↓

TUR: Çok teşekkür ederim.

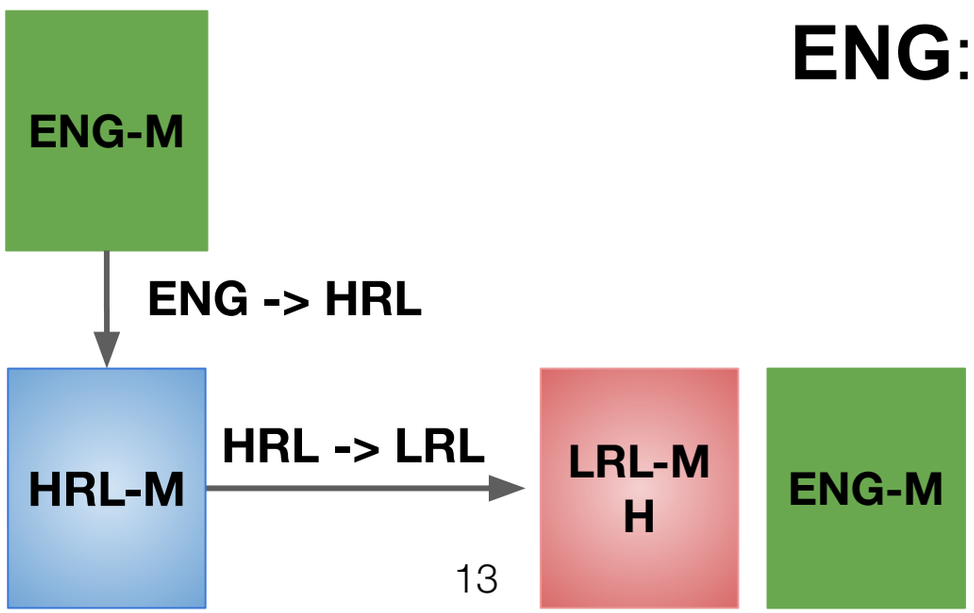
ENG: Thank you so much.

↓

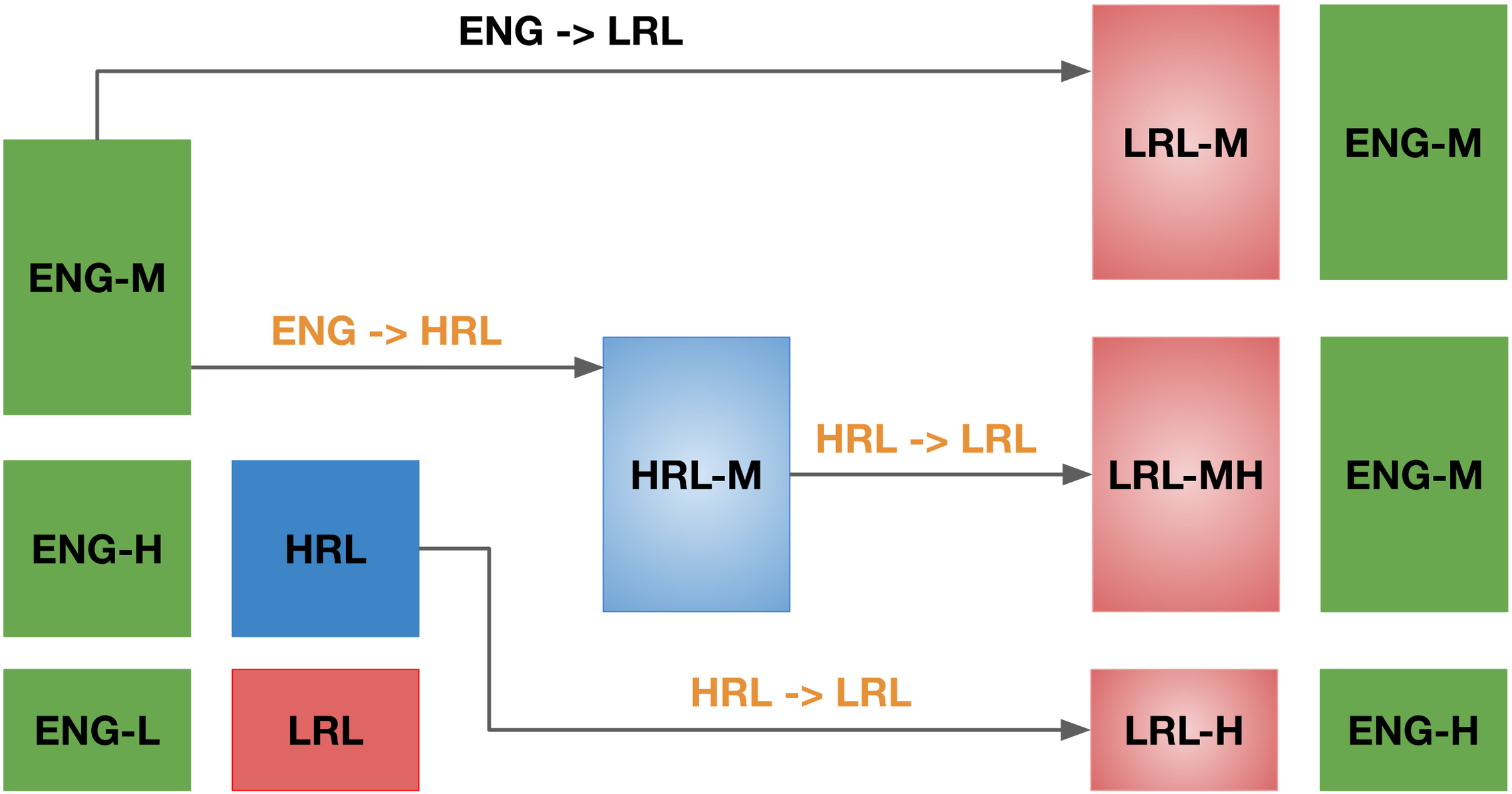
AZE: Çox sağ olun.

ENG: Thank you so much.

- **Proposal 3: ENG-HRL-LRL**
  - Large amount of English monolingual data can be utilized



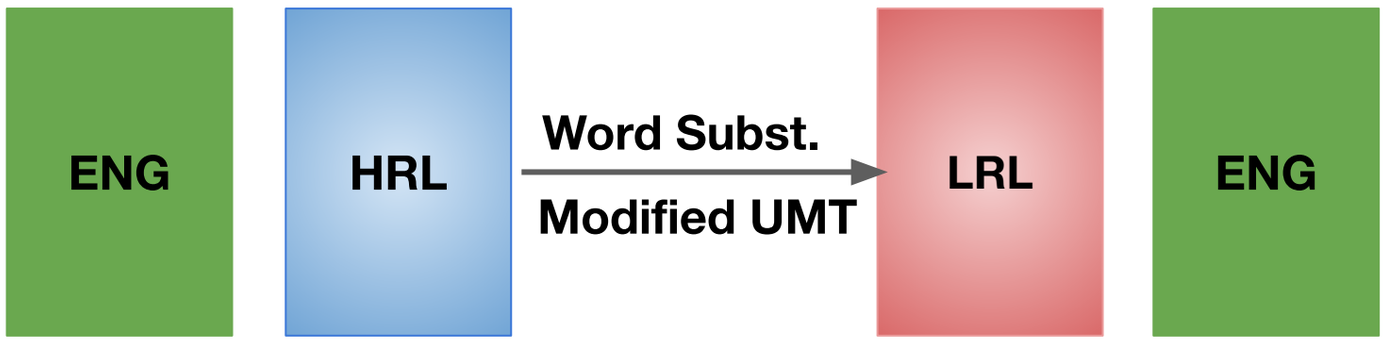
A **Generalized Framework** for Low-Resource Data Augmentation



HRL-LRL (Related Languages) Translation

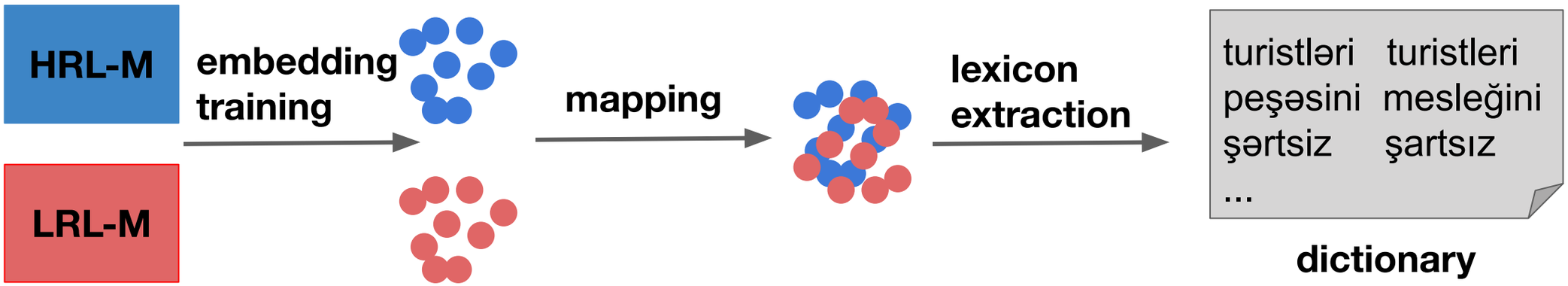
- Still a low resource setting! Standard supervised translation did not work well.

- We propose two simple techniques
  - Word substitution
  - Modified Unsupervised MT



HRL-LRL Translation - Word Substitution

- Lexicon Induction (e.g. Xing et al. 2015; Zhang et al. 2017; Lample et al. 2018)



- Corpus Construction
  - Replace HRL words with LRL ones to construct pseudo LRL-ENG corpus

bizim **ölkəmizin** utancını göstermek **üçün turistləri dəvət edir** .

bizim **ülkemizin** utancını göstermek **için turistleri davet eder** .

example

HRL-LRL Translation - Unsupervised MT

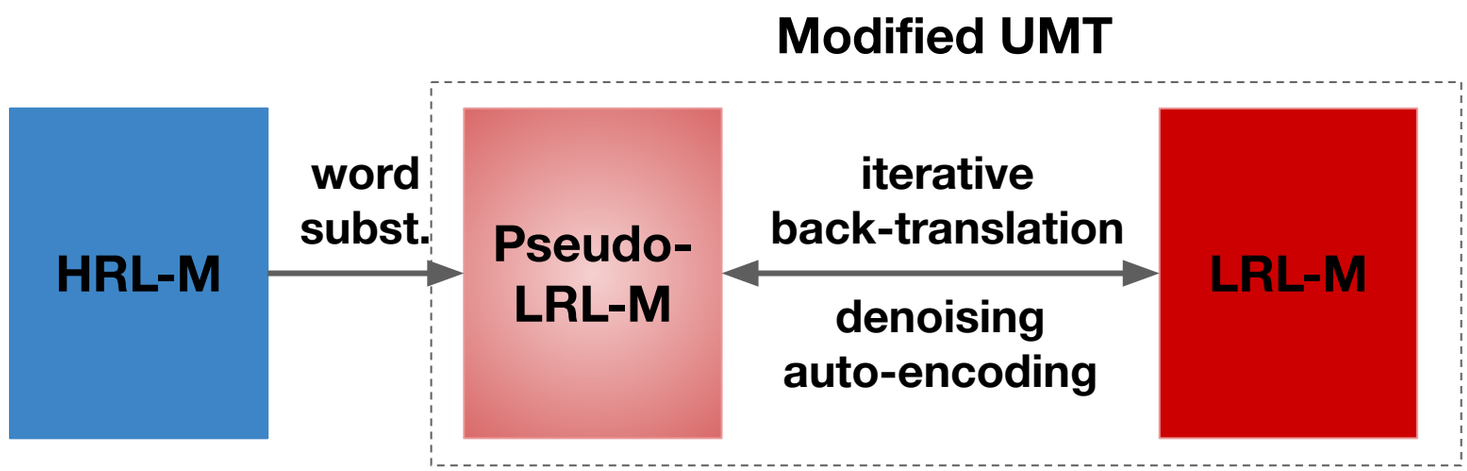
- **Word substitution still lacking:**
  - Is not context-dependent
  - Cannot handle reordering
  - Still have HRL words

- **An alternative:** unsupervised HRL-LRL MT!
  (Lample et al., 2018; Artetxe et al., 2018)

- **Problem:** direct unsupervised MT from HRL to LRL showed poor results.

HRL-LRL Translation - Modified Unsupervised MT

- Word substitution for HRL
- UMT over Pseudo-LRL and LRL corpus
- Jointly segmented => introduce more lexicon overlap
- Translate pseudo-LRL to LRL to construct LRL-ENG corpus

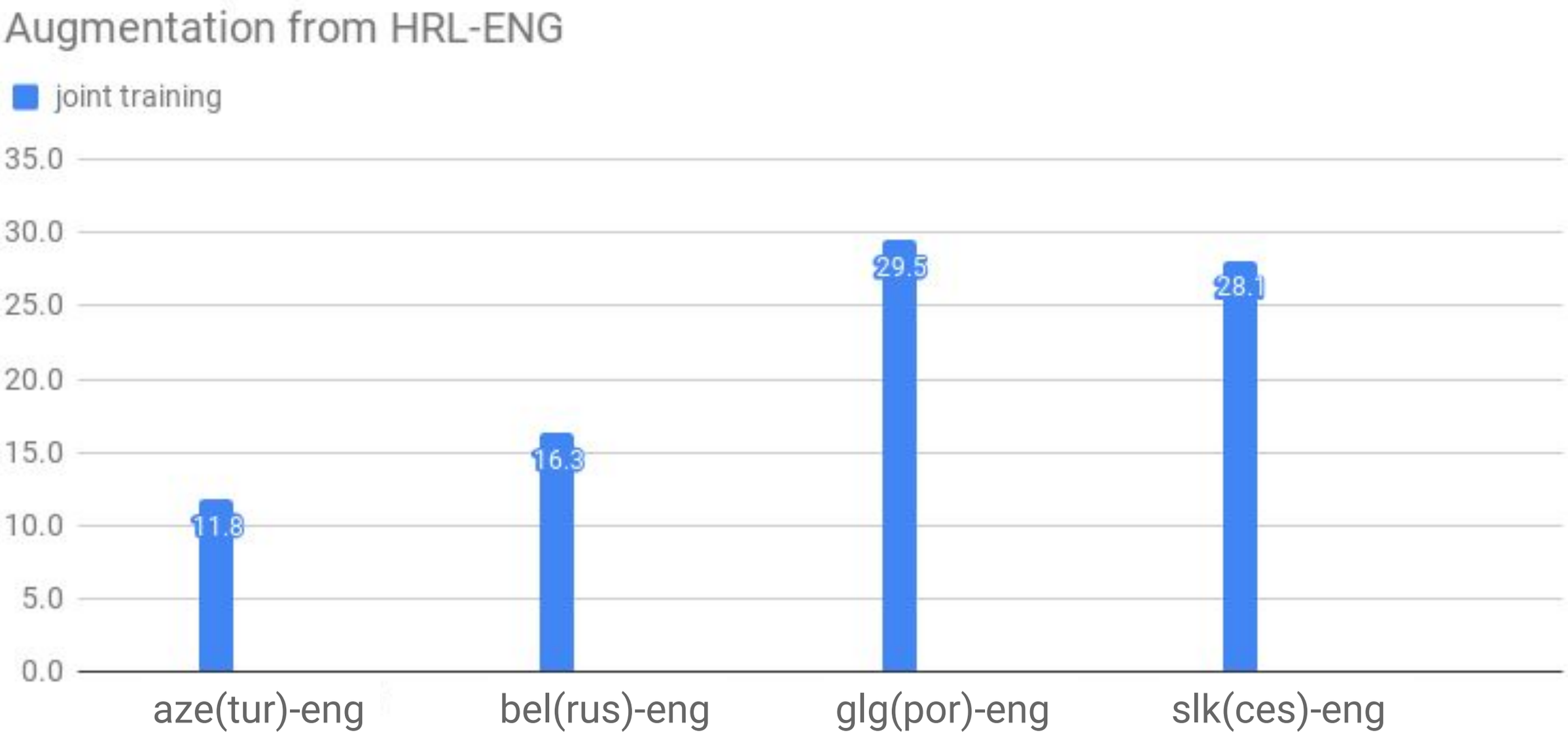


Experiments

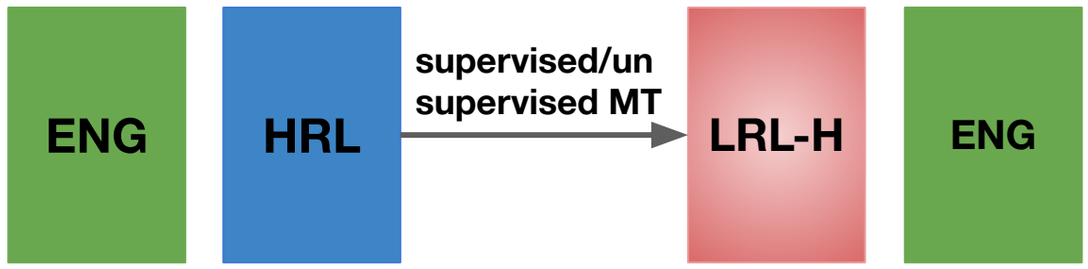
Experiment Setting - Dataset

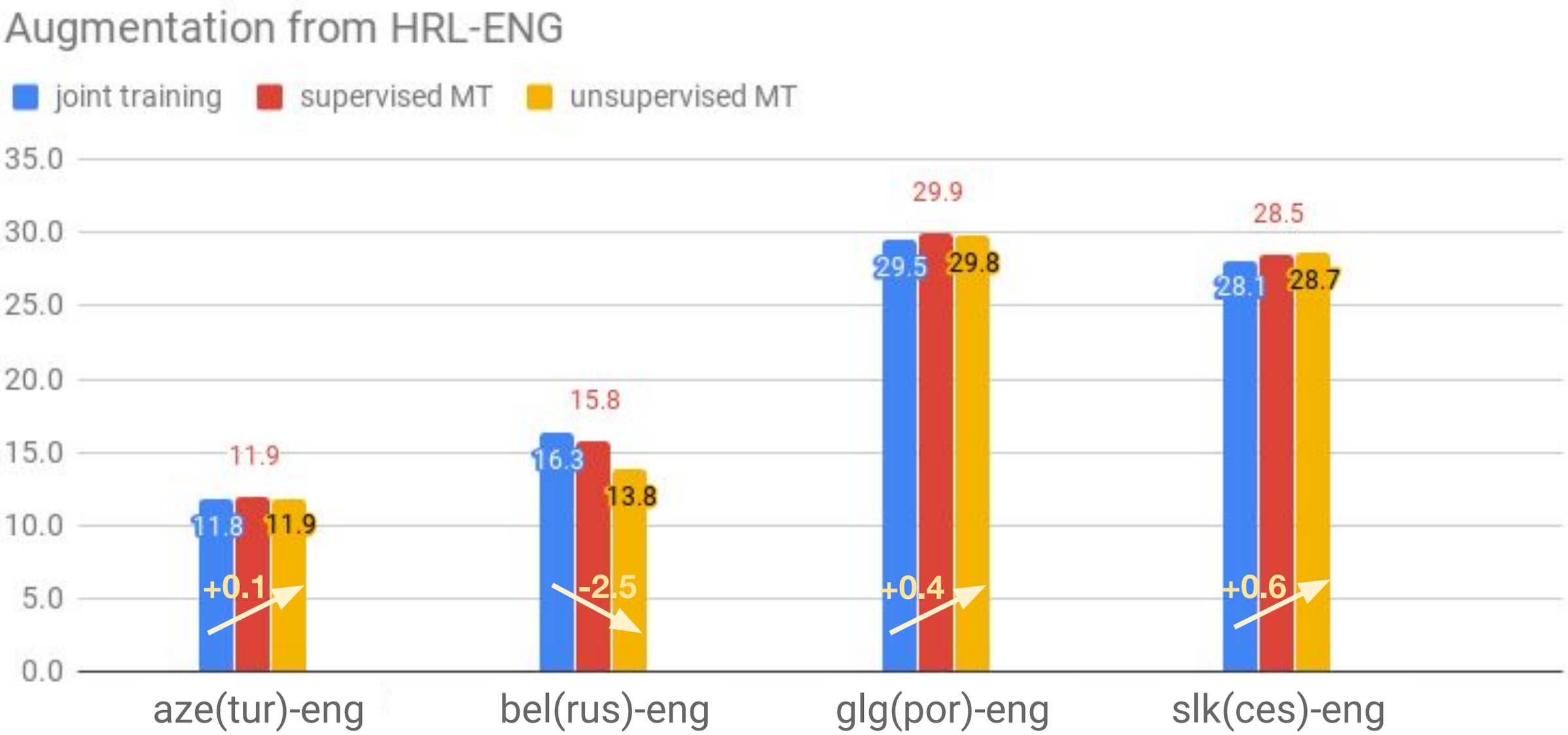
- **Parallel Data**: Ted Corpus (Qi et al., 2018)
  - LRL-ENG: 5.9-61K sentences
  - HRL-LRL: 5.7-44K sentences
  - HRL-ENG: 103-208K sentences
- **Monolingual Data**: Wiki Dumps
  - HRL, LRL, ENG: 2M sentences
- Sentence pieced 8k

Experiment Results



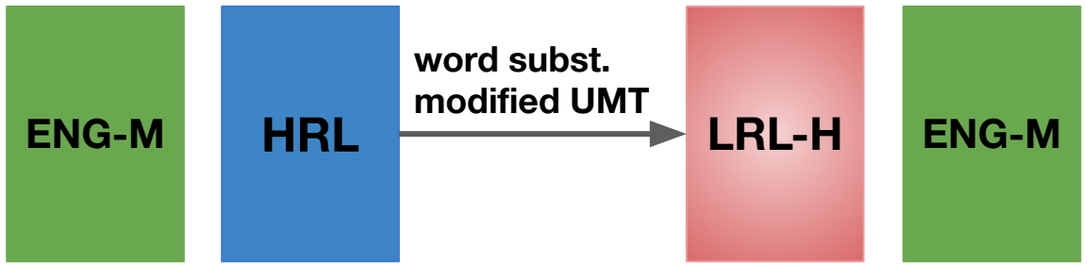
Experiment Results

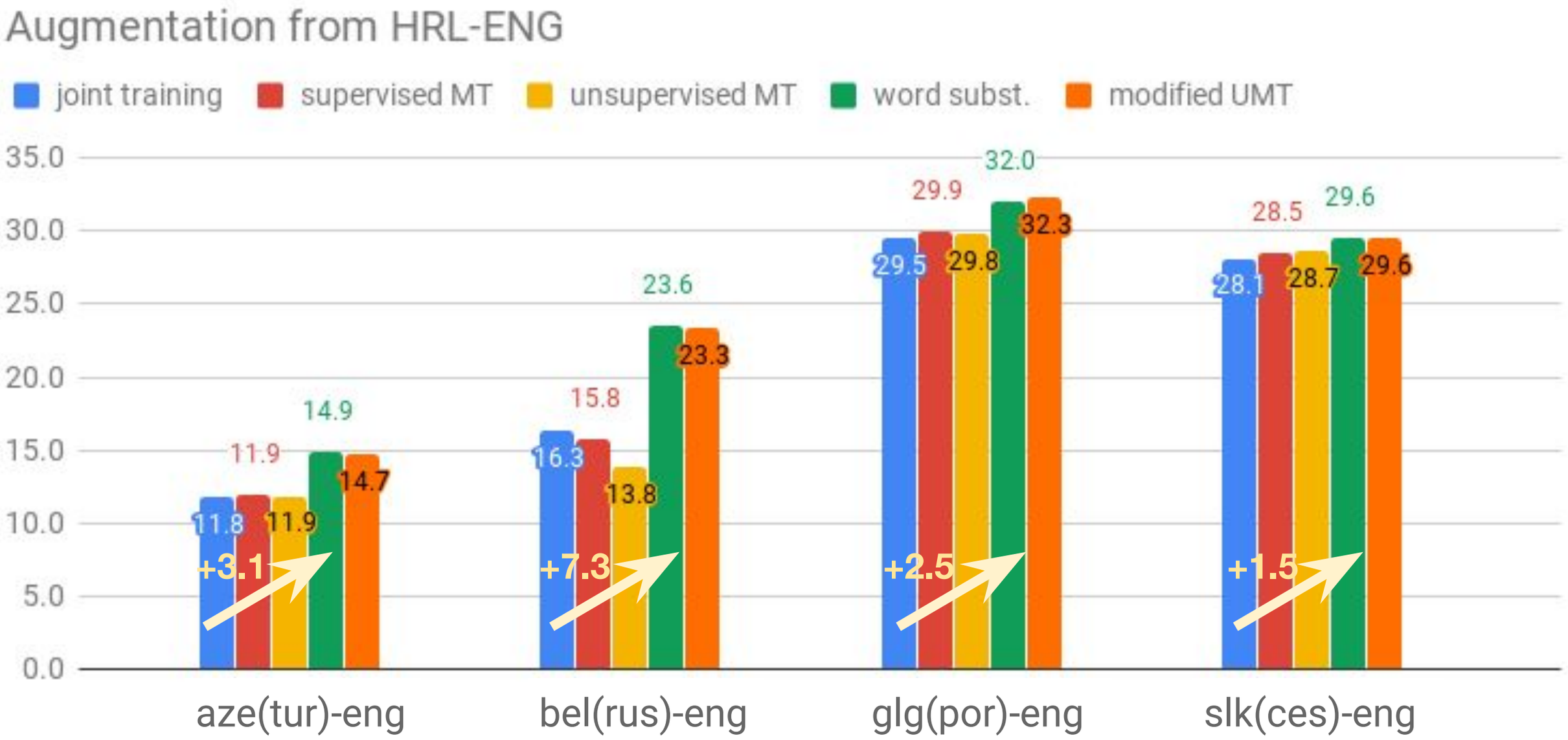




Low-resource supervised and vanilla unsupervised HRL-LRL translation do not lead to significant improvements.

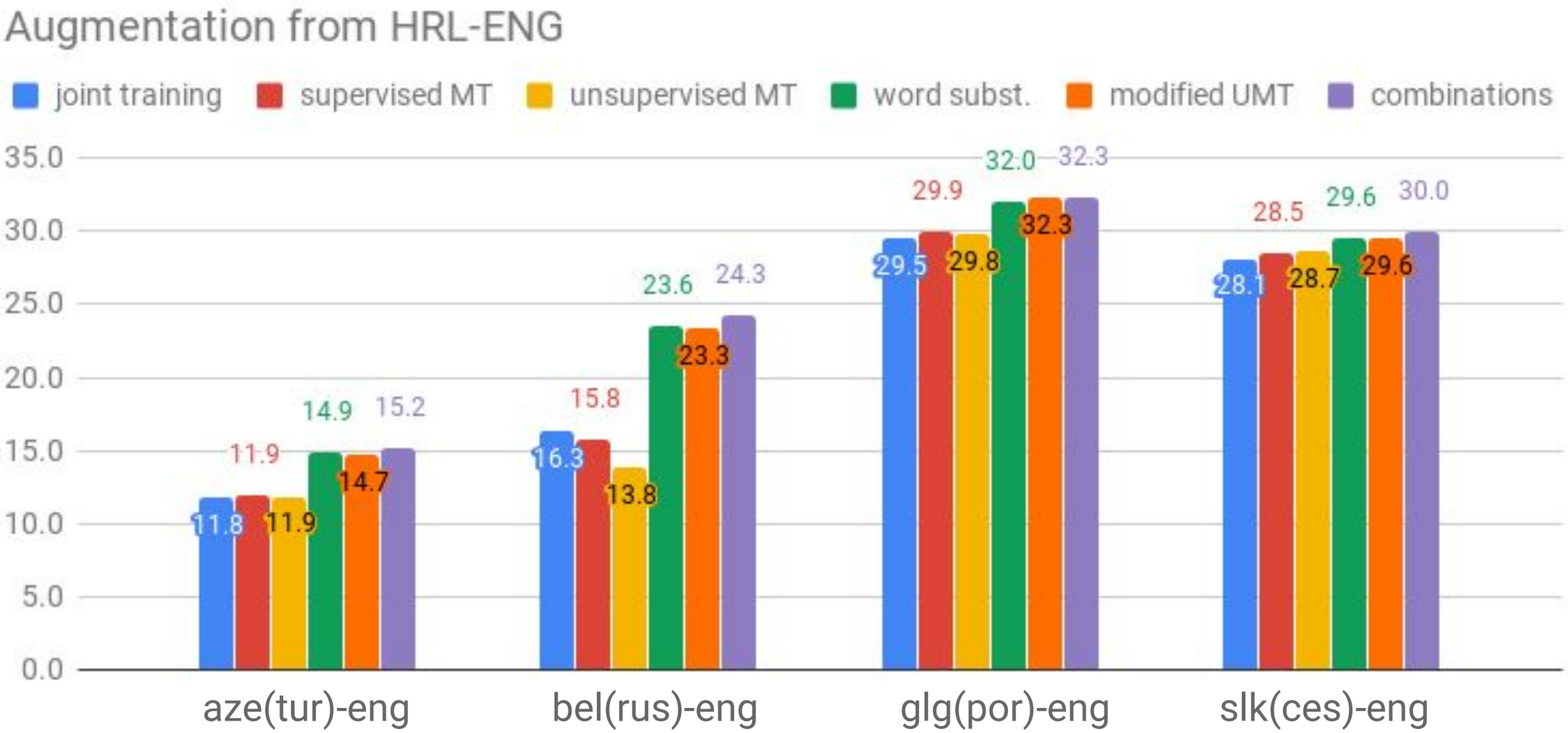
Experiment Results





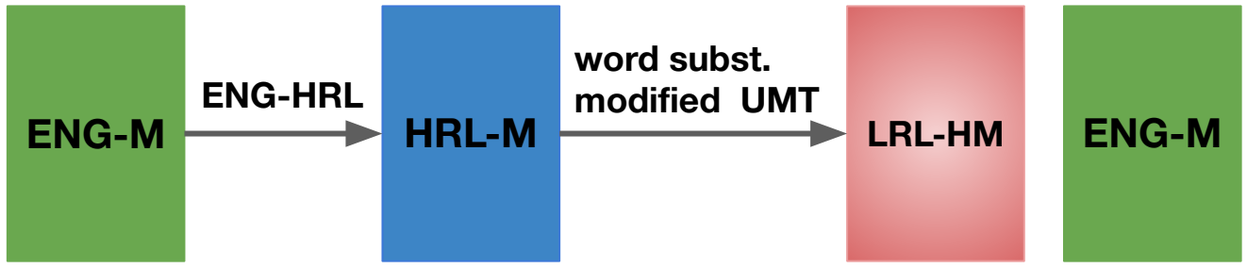
Our methods improve the performance by 1.5 - 7.3 BLEU points.

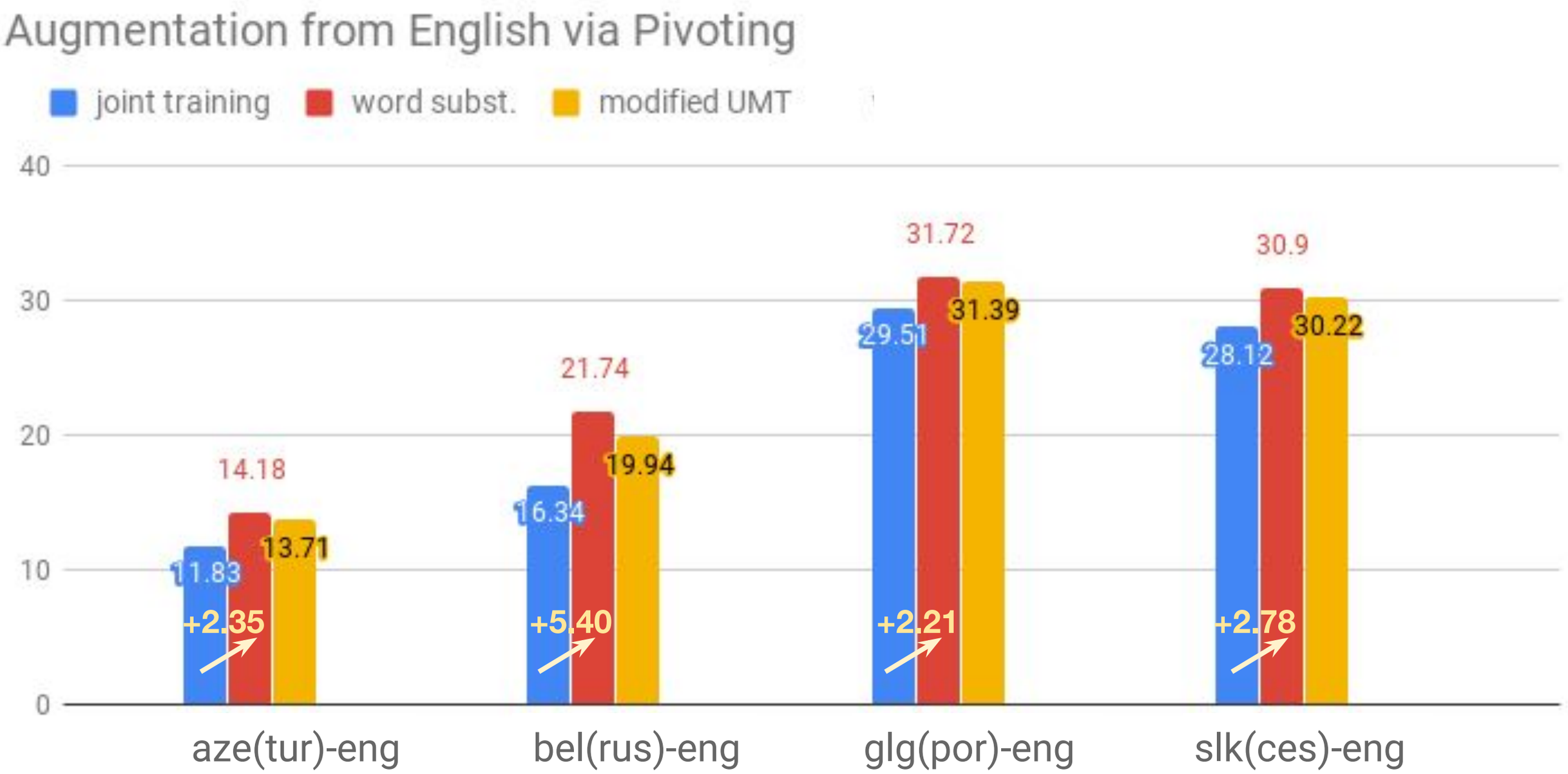
Experiment Results



The combination of the two methods give further improvements.

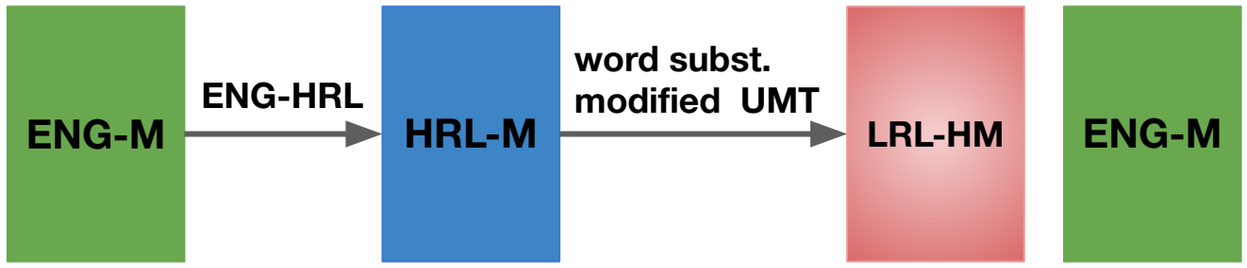
Experiment Results

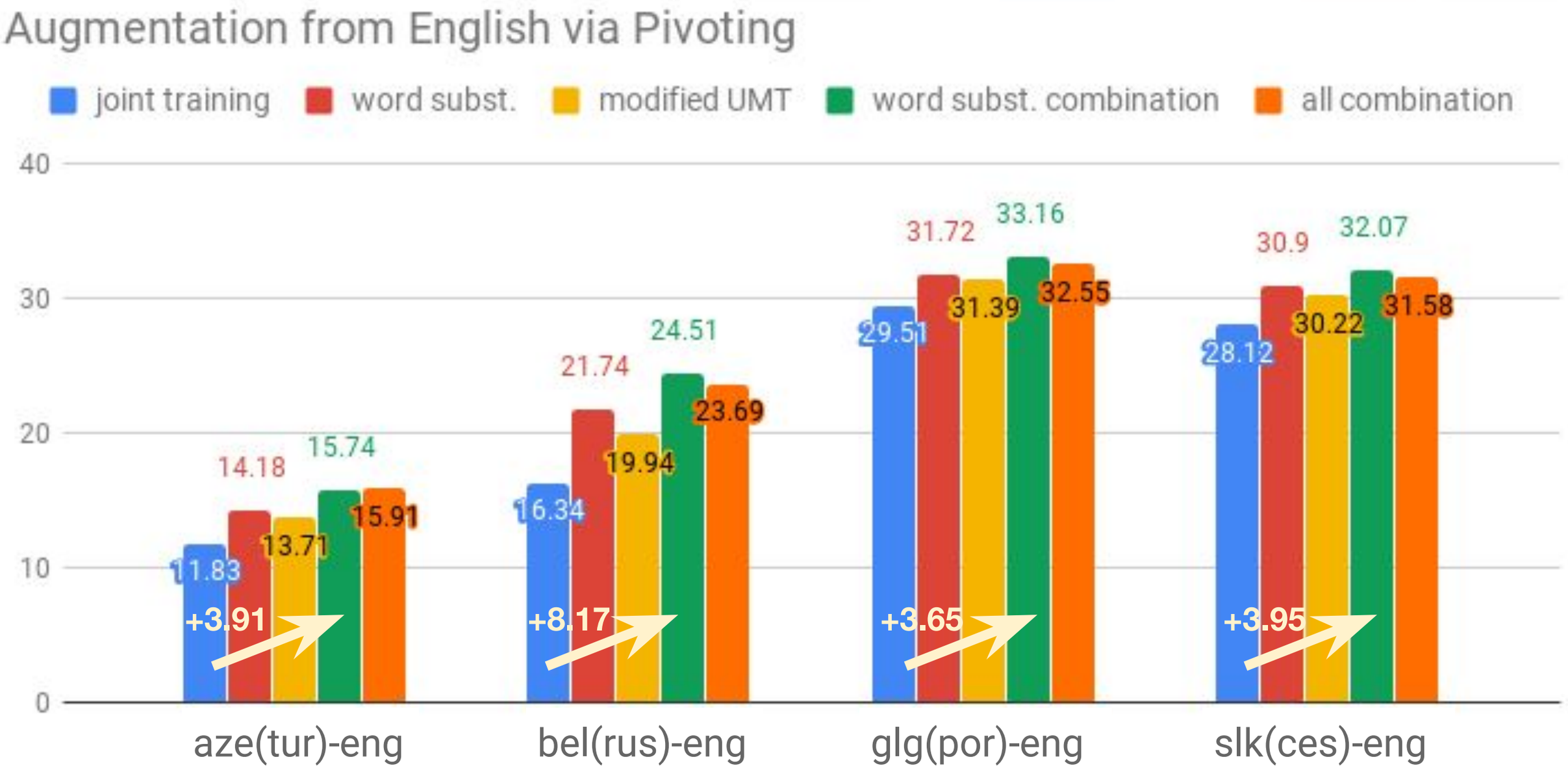




Augmentation from English with 200k sentences brings 2-5 BLEU improvements.

Experiment Results

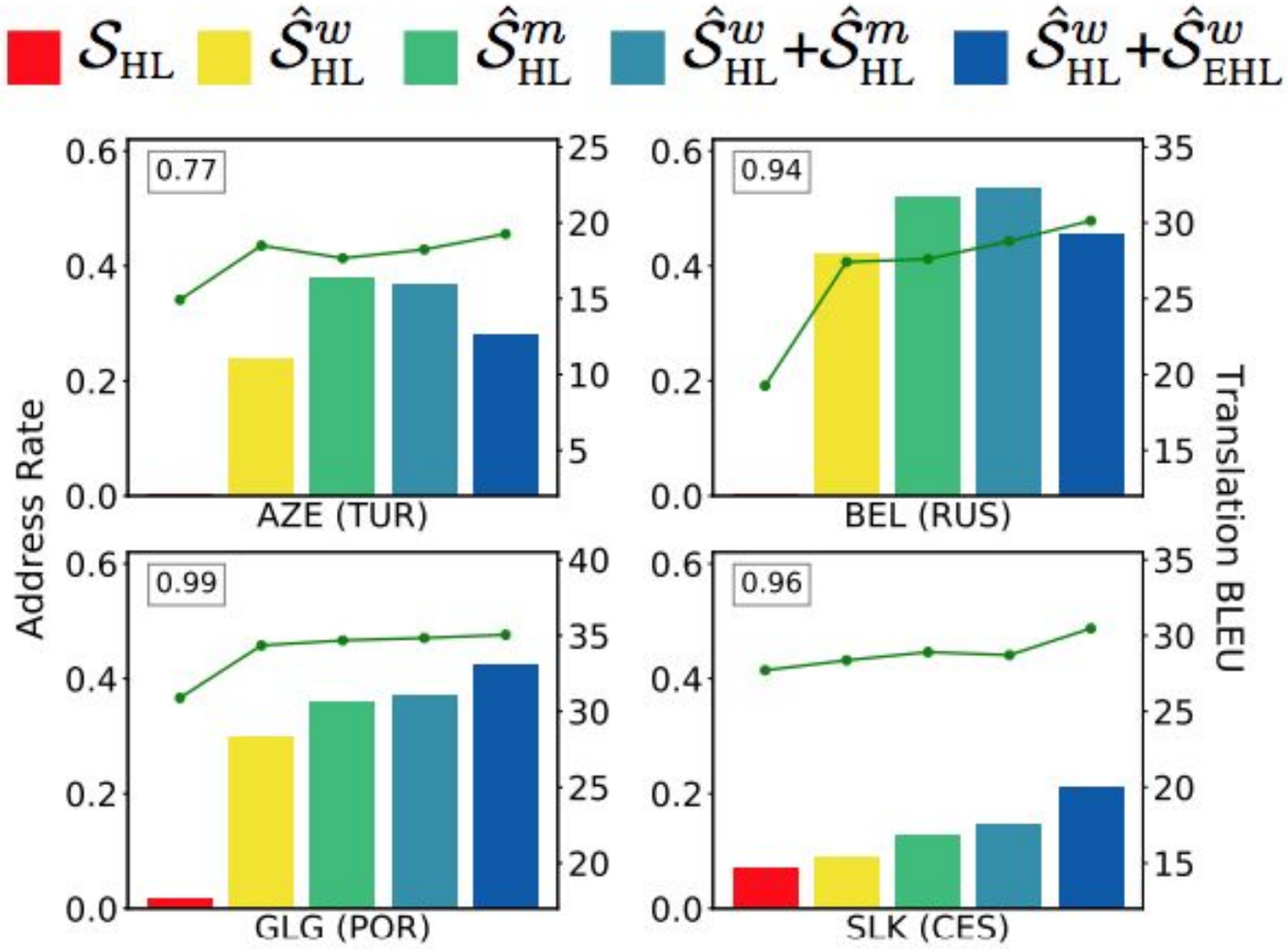




Combining the two methods give further improvements, ~4-8 BLEU in total.

Why does our methods do better?

Rare word address Rate - The percentage of rare words that become frequent after data augmentation



Rare word address rate (bars) **correlates with** LRL-ENG BLEU Scores (line plot).

A Concrete Example (Cont.)

source -
Atam balaca boz radiosunda BBC Xəbərlərinə qulaq asırdı.

translation output before data augmentation -
So I’m going to became a lot of people.

translation output after data augmentation-
My dad used to listen to BBC News on a little radio.

reference -
My father was listening to BBC News on his small , gray radio.

Conclusion

- Propose a generalized data augmentation framework
- Translating between related languages can improve LRL MT
- It’s important to make the best use of existing data for LRL MT

Thank you! Question?