CS11-737 Multilingual NLP

Data-based Strategies to Low-resource MT

Graham Neubig

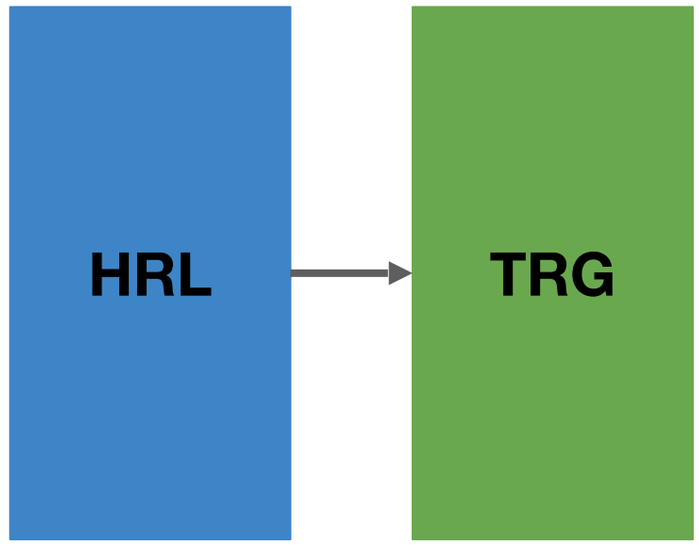
Carnegie Mellon University
Language Technologies Institute

Site
http://phontron.com/class/multiling2022/

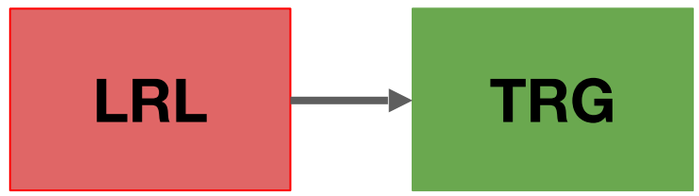
Many slides from:
Xia, Mengzhou, et al. "Generalized data augmentation for low-resource translation." *ACL 2019.*

Data Challenges in Low-resource MT

- MT of high-resource languages (HRLs) with large parallel corpora → relatively 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.

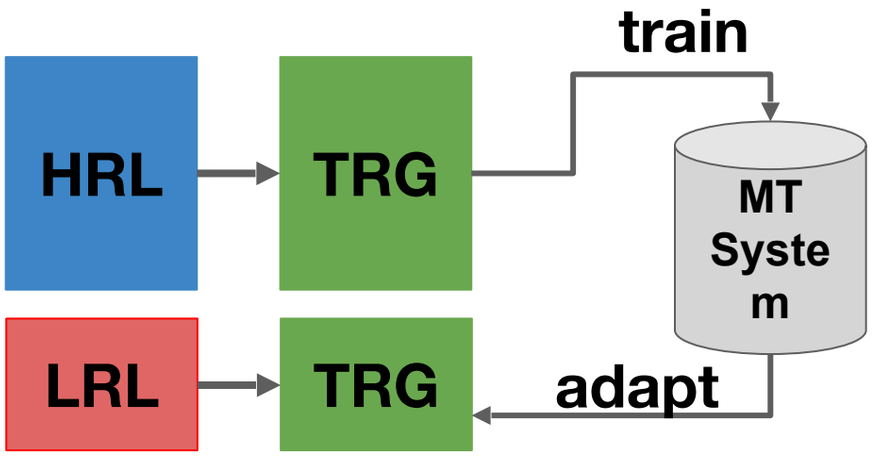
translation - So I’m going to became a lot of people.

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

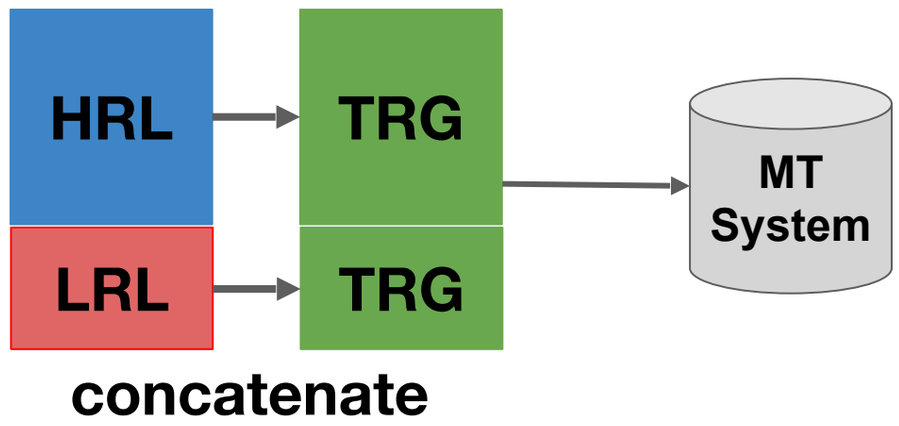
Does not convey the correct meaning at all.

Multilingual Training Approaches

- 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)



- **Problem:** Suboptimal lexical/syntactic sharing.
- **Problem:** Can't leverage monolingual data.

Data Augmentation

Available Resources

TRG-M

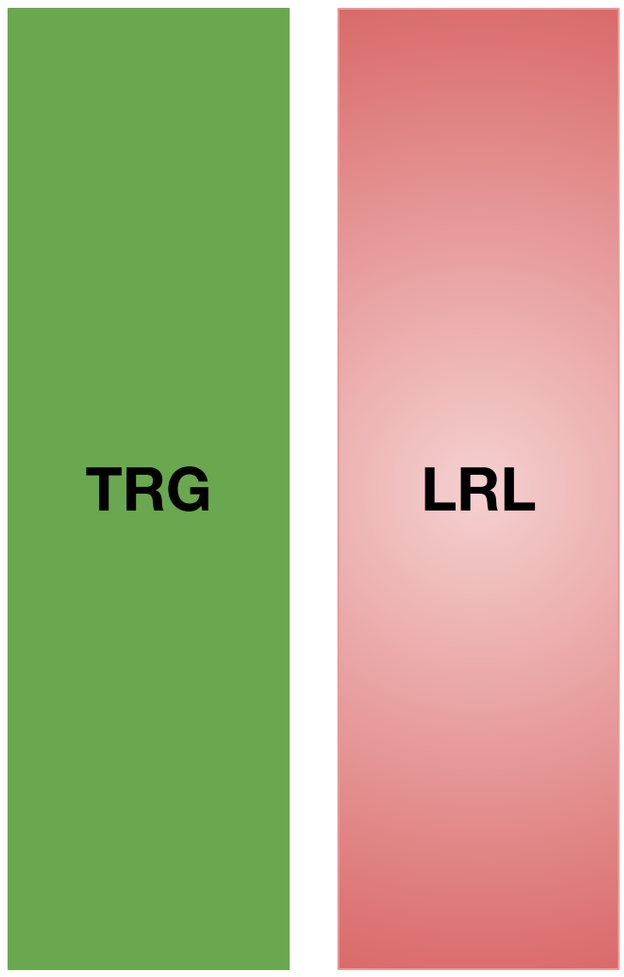
TRG-H

HRL

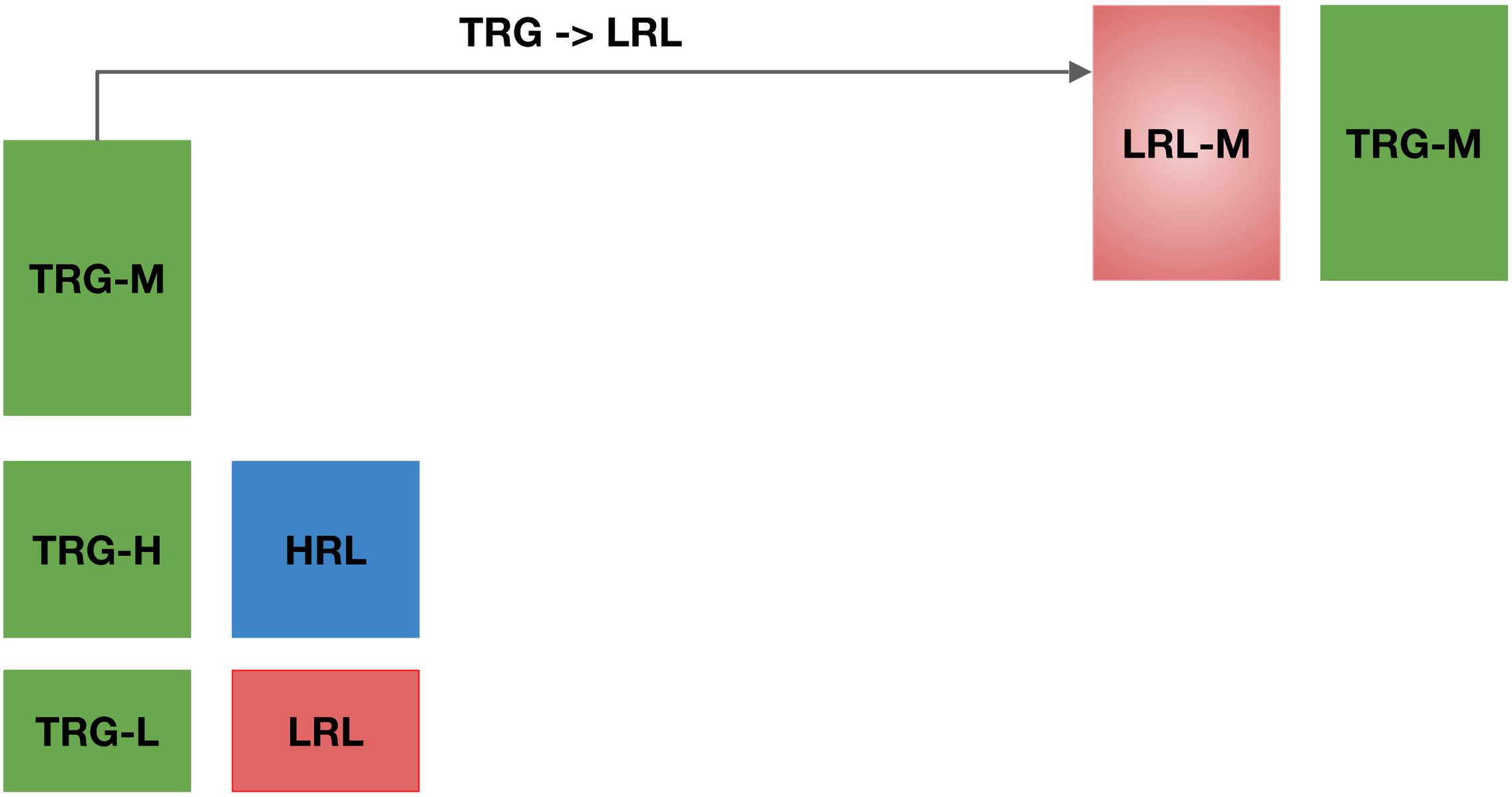
TRG-L

LRL

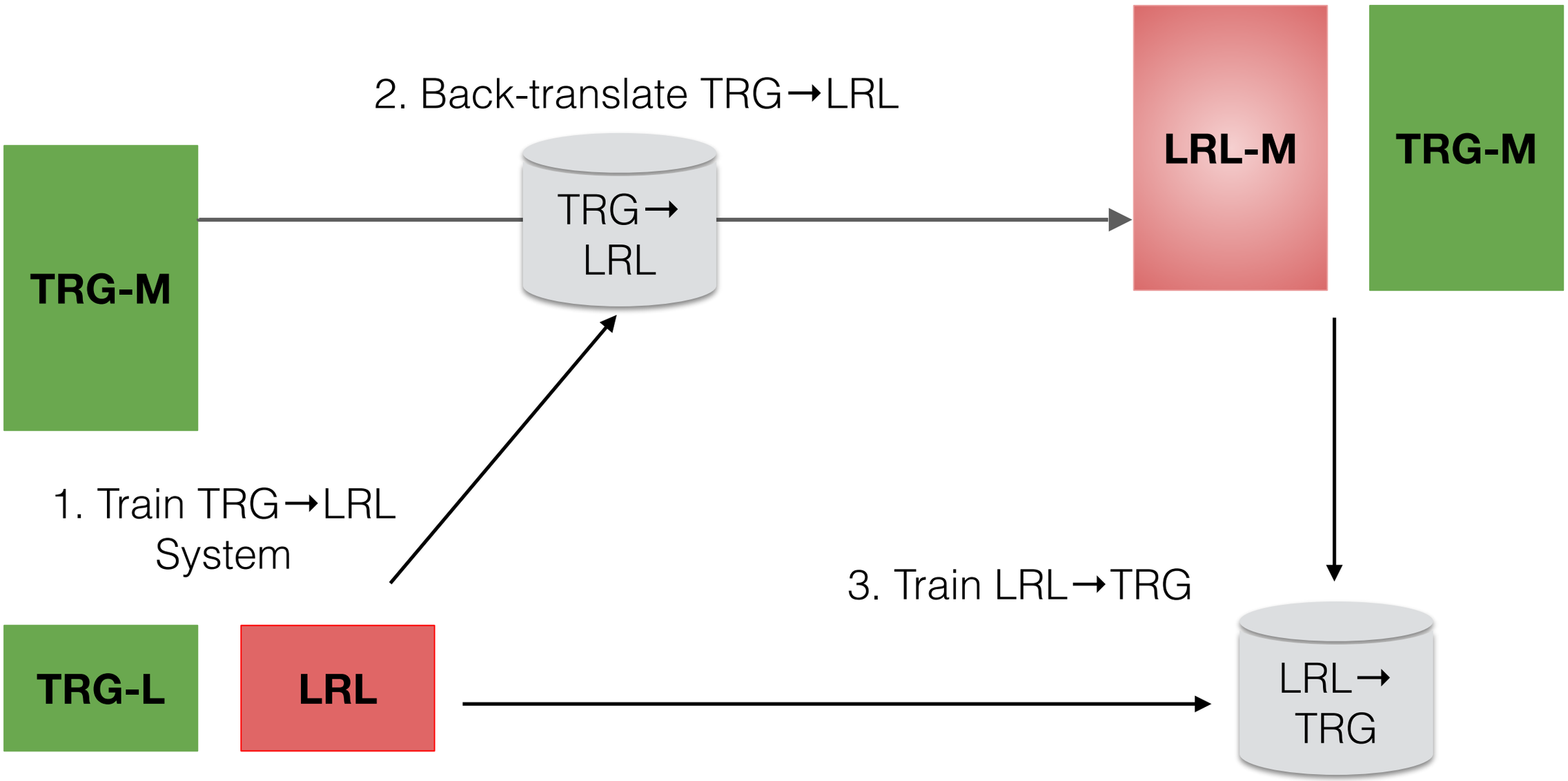
Augmented Data



Data Augmentation 101: Back Translation



Back Translation Idea



- Some degree of error in source data is permissible!

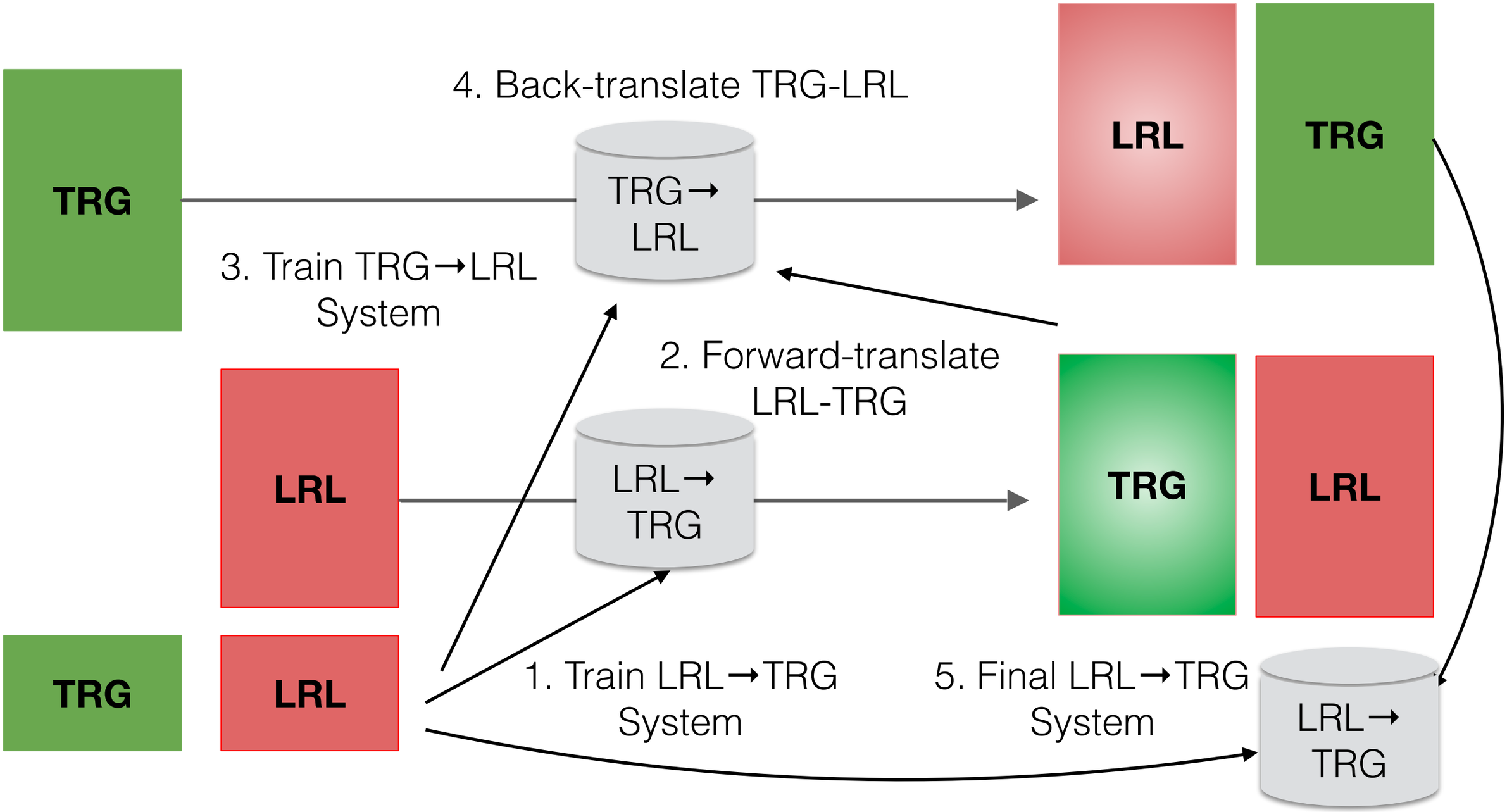
Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." *ACL 2016*.

How to Generate Translations

- How to generate translations?
- **Beam search** (Sennrich et al. 2016)
  - Select the highest scoring output
  - Higher quality, but lower diversity, potential for data bias
- **Sampling** (Edunov et al. 2018)
  - Randomly sample from back-translation model
  - Lower overall quality, but higher diversity
- Sampling has shown to be more effective overall, can be viewed as modeling data distribution (Pham et al. 2021)

Understanding Back-Translation at Scale. Sergey Edunov, Myle Ott, Michael Auli, David Grangier. EMNLP 2018.
Meta Back-translation. Hieu Pham, Xinyi Wang, Yiming Yang, Graham Neubig. ICLR 2021.

Iterative Back-translation

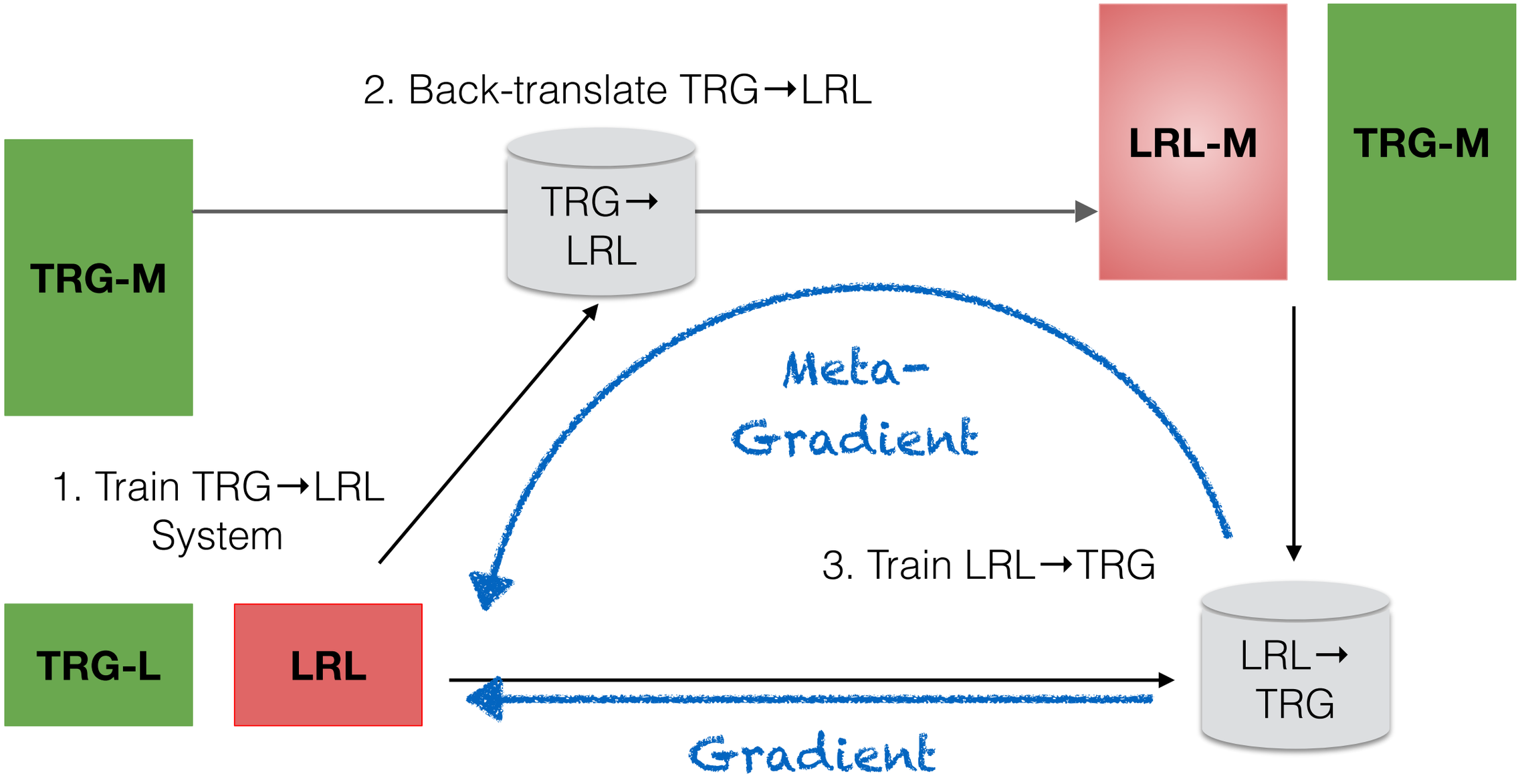


Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn. "Iterative Back-Translation for Neural Machine Translation" WNGT 2018.

Meta Back-translation

(Pham et al. 2021)

- Train back-translation model to improve forward translation model



Meta Back-translation. Hieu Pham, Xinyi Wang, Yiming Yang, Graham Neubig. ICLR 2021.

Back Translation Issues

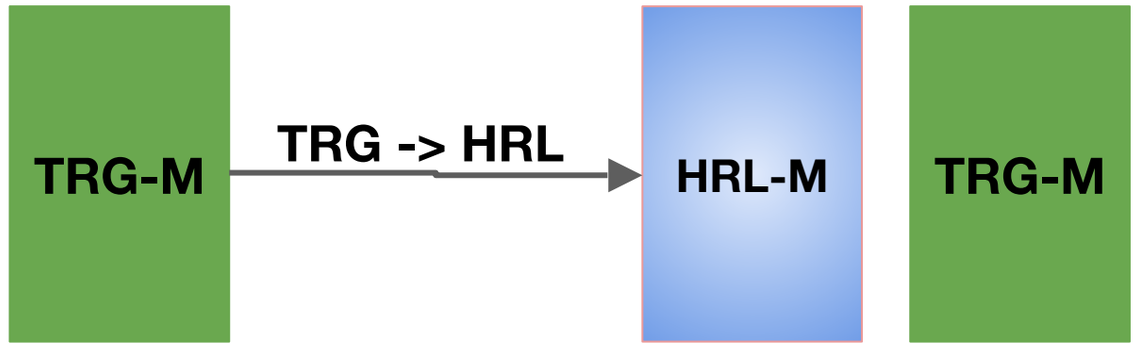
- Back-translation fails in low-resource languages or domains
- Use other high-resource languages
- Combine with monolingual data (maybe with denoising objectives, covered in following class)
- Perform other varieties of rule-based augmentation

Using HRLs in Augmentation

Xia, Mengzhou, et al. "Generalized data augmentation for low-resource translation." *ACL 2019.*

English -> HRL Augmentation

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



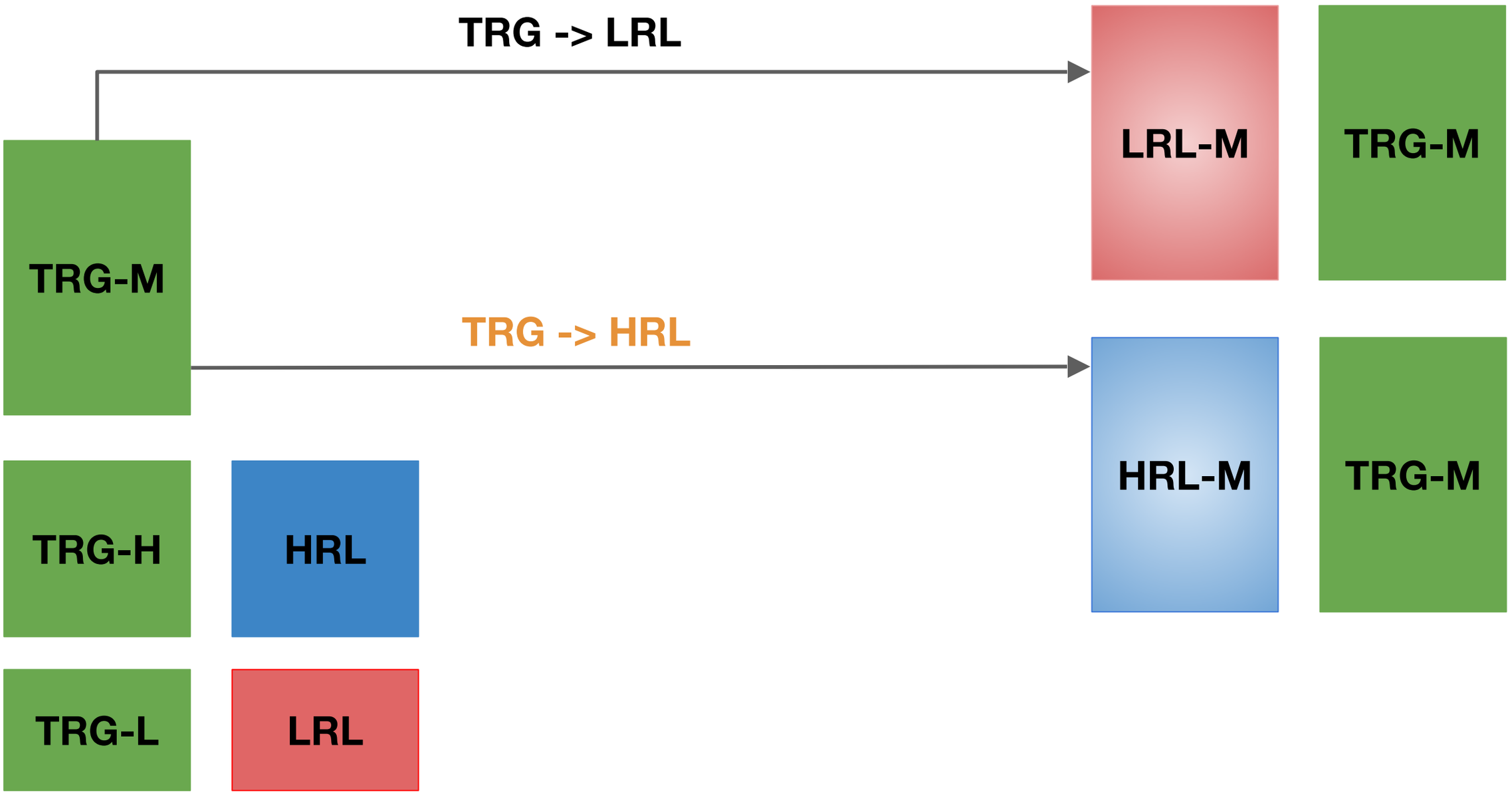
- **Idea: also back-translate into HRL**
  - more sentence pairs
  - vocabulary sharing of source-side
  - syntactic similarity of source-side
  - improves target-side LM

TRG: Thank you very much.

~~**AZE**: Hə Hə Hə.~~

TUR: Çok teşekkür ederim.

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

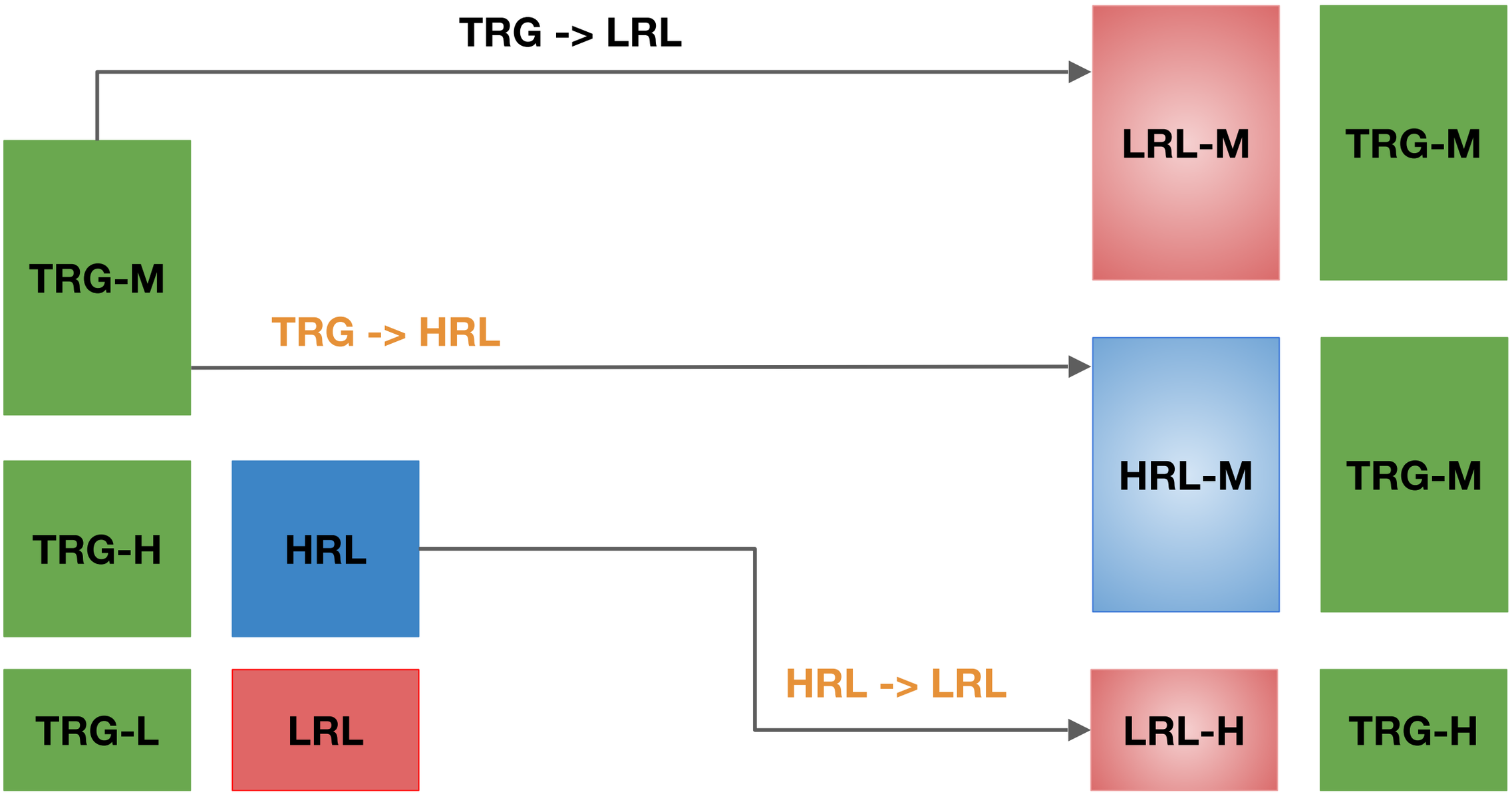


Augmentation via Pivoting

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

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

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



Back-Translation by Pivoting

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

TRG: Thank you so much.

↓

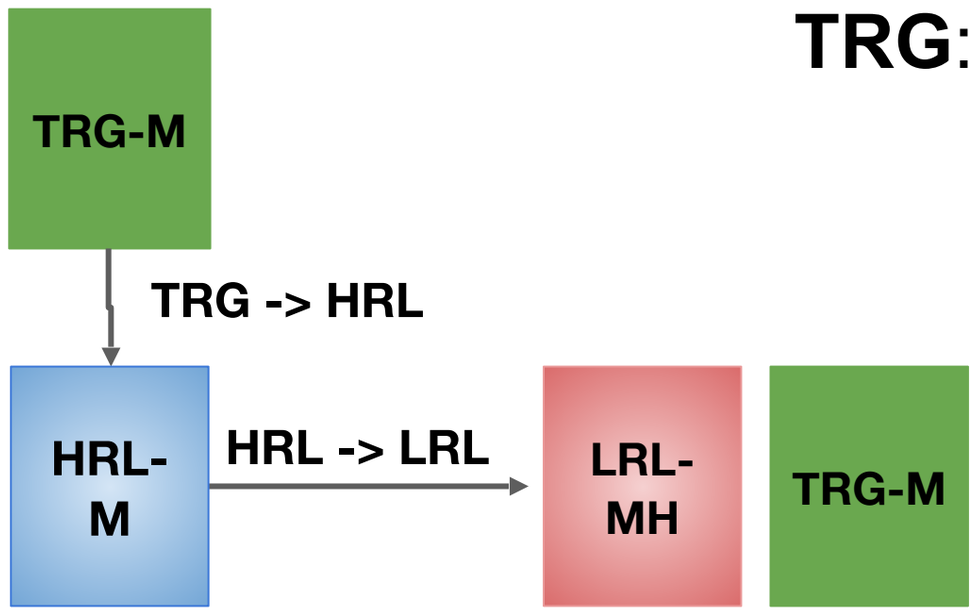
TUR: Çok teşekkür ederim.

↓

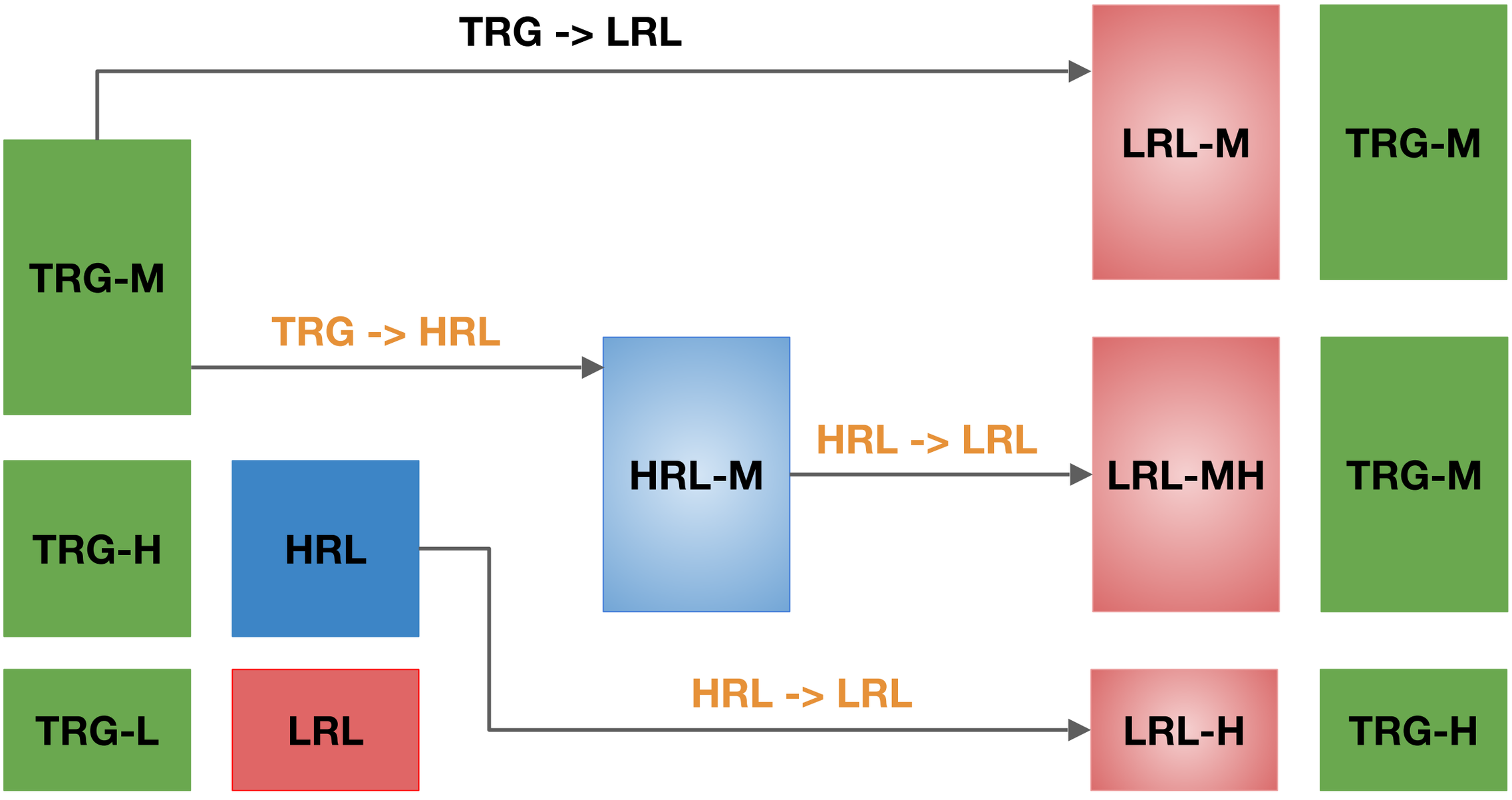
AZE: Çox sağ olun.

TRG: Thank you so much.

- **Idea: TRG-HRL-LRL**
  - Large amount of English monolingual data can be utilized



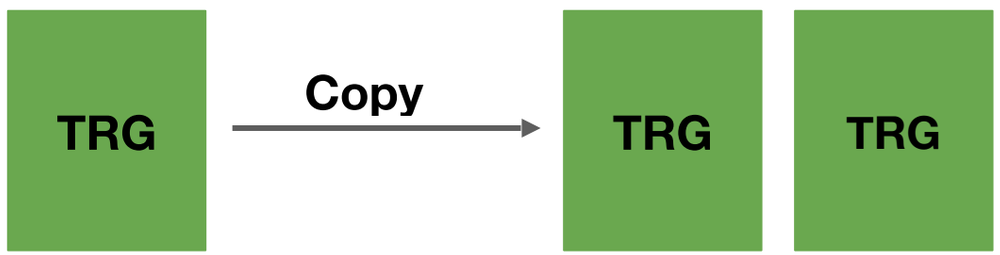
Data w/ Various Types of Pivoting



Monolingual Data Copying

Monolingual Data Copying

- **Problem:** Back-translation may help with structure, but fail at terminology
- **Idea: Use monolingual data *as-is***
  - Helps encourage the model to not drop words
  - Helps translation of terms that are identical across languages



TRG: Thank you so much. → **SRC**: Thank you so much.
TRG: Thank you so much.

Anna Currey, Antonio Valerio Miceli Barone, Kenneth Heafield. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. WMT 2018.

Heuristic Augmentation Strategies

Dictionary-based Augmentation

1. Find rare words in the source sentences
2. Use a language model to predict another word that could appear in that context

Sentence [original / **substituted**]	Plausible
My sister drives a [car / **motorbike**]	yes
My uncle sold his [house / **motorbike**]	yes
Alice waters the [plant / **motorbike**]	no (semantics)
John bought two [shirts / **motorbike**]	no (syntax)

3. Replace, and aligned word with translation from dictionary

Marzieh Fadaee, Arianna Bisazza, Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. ACL 2017.

An Aside: Word Alignment

- Automatically find alignments between source and target words for dictionary learning, analysis, supervised attention etc.

- **Traditional symbolic methods:** word-based translation models trained using EM algorithm
  - GIZA++: https://github.com/moses-smt/giza-pp
  - FastAlign: https://github.com/clab/fast_align

- **Neural methods:** use model like multilingual BERT or translation and find words with similar embeddings
  - Awesome-Align: https://github.com/neulab/awesome-align

Word-by-word Data Augmentation

- Even simpler, translate sentences word-by-word into target sentence using dictionary

J'ai acheté une nouvelle voiture
↓ ↓ ↓ ↓ ↓
I bought a new car

- **Problem:** what about word ordering, syntactic divergence?

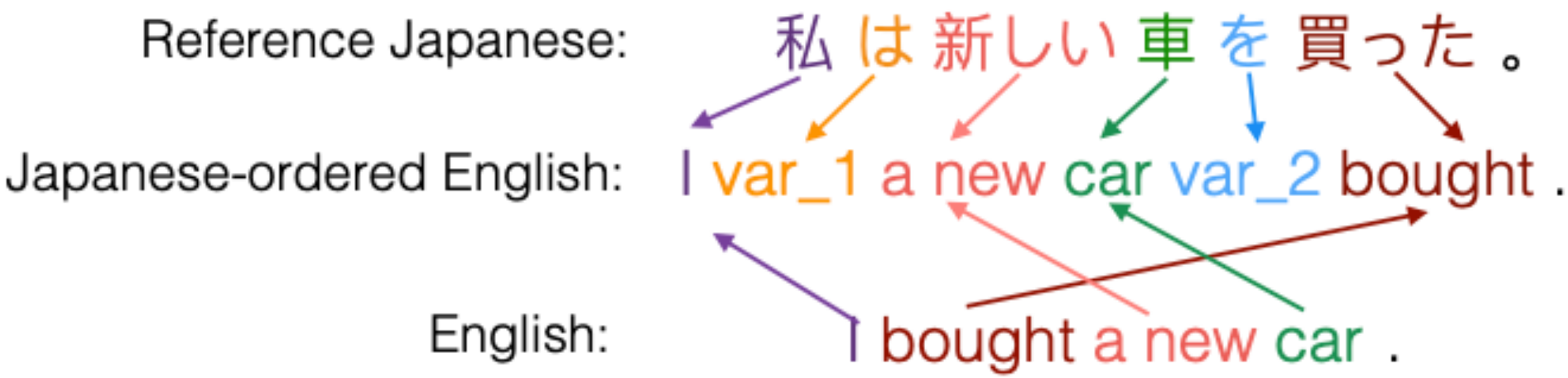
私 は 新しい 車 を 買った
↓ ↓ ↓ ↓ ↓ ↓
I the new car a bought

Lample, Guillaume, et al. "Unsupervised machine translation using monolingual corpora only." *arXiv preprint arXiv:1711.00043* (2017).

Word-by-word Augmentation w/ Reordering

- **Problem:** Source-target word order can differ significantly in methods that use monolingual pre-training

- **Solution:** Do re-ordering according to grammatical rules, followed by word-by-word translation to create pseudo-parallel data



Zhou, Chunting, et al. "Handling Syntactic Divergence in Low-resource Machine Translation." arXiv preprint arXiv:1909.00040 (2019).

In-class Assignment

In-class Assignment

- Read one of the cited papers on heuristic data augmentation

Marzieh Fadaee, Arianna Bisazza, Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. ACL 2017.

Zhou, Chunting, et al. "Handling Syntactic Divergence in Low-resource Machine Translation." EMNLP 2019.

- Try to think of how it would work for one of the languages you're familiar with

- Are there any potential hurdles to applying such a method? Are there any improvements you can think of?