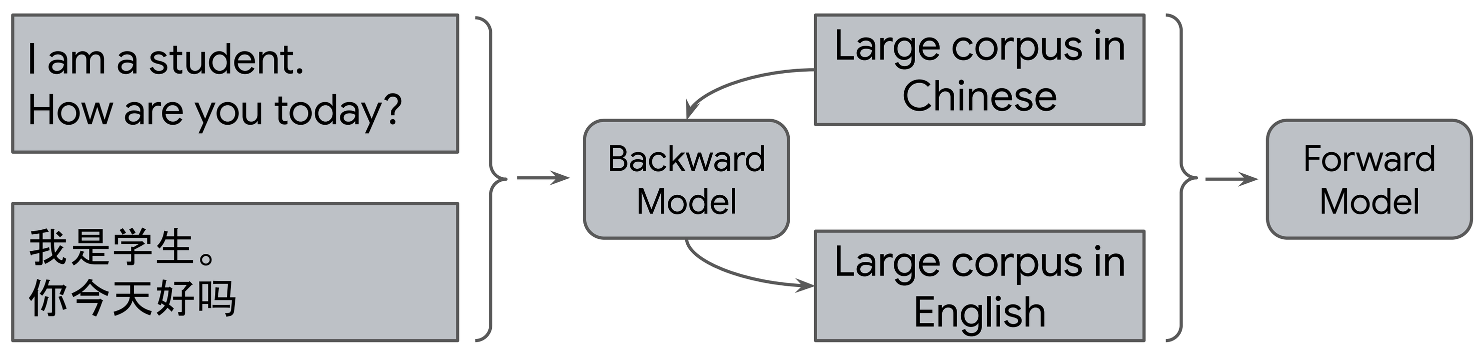
Meta Back-Translation

Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, Quoc V. Le
{hyhieu,zihangd,qizhex,thangluong,qvl}@google.com

Google Research
Carnegie Mellon University
Language Technologies Institute

Introduction

Suppose we want to train a model to translate **English** into **Chinese**



Original parallel data *Pseudo parallel data*

Back-Translation (BT)

- Very effective for using large monolingual corpora, but...

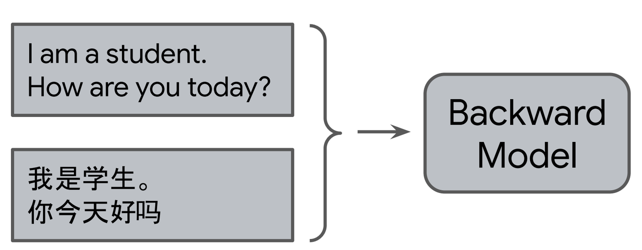
Limitations of Back-Translation

- Backward model's is constrained by the amount of parallel data
- It is unclear how to sample pseudo parallel data to train the best forward model.

Meta Back-Translation (Meta BT) resolves both limitations.

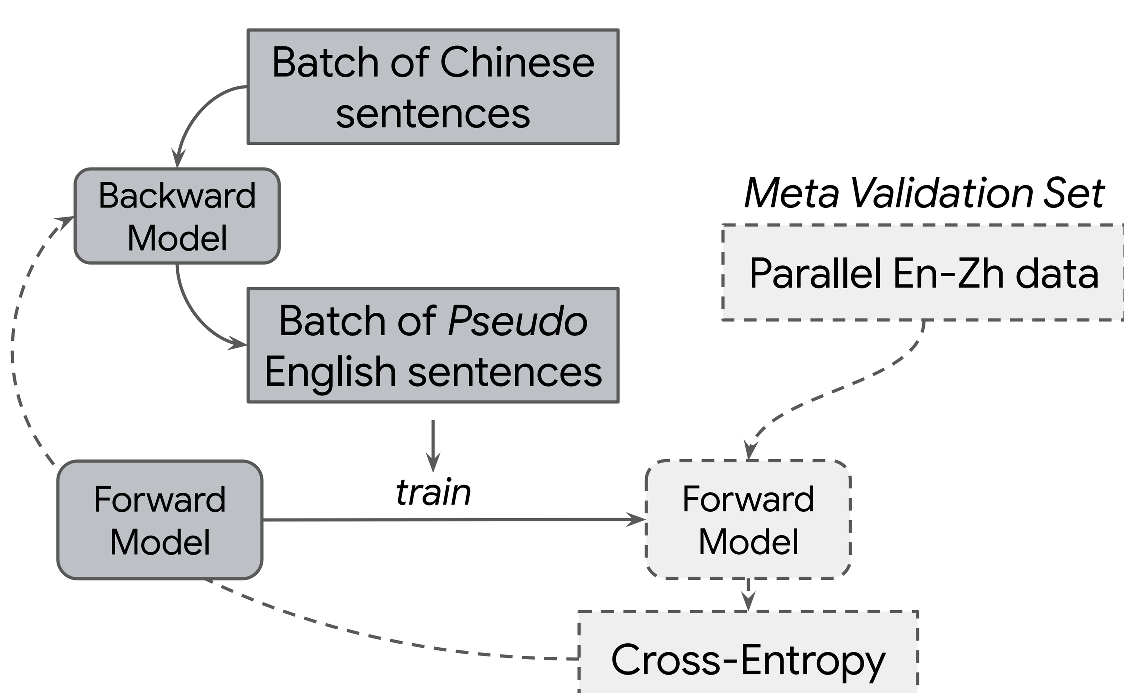
High-level Intuition

Step 1: Train a backward model using all available parallel data.



Step 2: Use the backward model to generate pseudo for a forward model, *but*:

- At each step, measure the forward model's cross-entropy on a *Meta Validation Set* and use it to update the backward model.



Intuition: the backward model should generate pseudo data so that *by learning on such data, the forward model generalizes well*

How to update the backward model?

Notations

ψ_t	backward model's params at step t
θ_t	forward model's params at step t
η_θ, η_ψ	learning rates for θ, ψ
$(\widehat{x}, y)$	pseudo source and target sentences
$J(\theta; \widehat{x}, y)$	cross-entropy for θ on $(\widehat{x}, y)$
$J_{\text{MetaVal}}(\theta)$	cross-entropy for θ *meta validation* data

Update the forward model's params:

$$\theta_t = \theta_{t-1} - \eta_\theta \nabla_\theta J(\theta; \widehat{x}, y)$$

Since $\widehat{x}$ is generated by the backward model, we have the *dependency* $\theta_t = \theta_t(\psi)$
This means that the *meta validation loss* depends on ψ

$$J_{\text{MetaVal}}(\theta_t) = J_{\text{MetaVal}}(\theta_t(\psi))$$

Using second-ordered gradients, we can derive:

$$\nabla_\psi J_{\text{MetaVal}}(\theta_t) \approx \left[\nabla_\theta J_{\text{MetaVal}}(\theta_t)^\top \cdot \nabla_\theta J(\theta_{t-1}; \widehat{x}, y)\right] \cdot \nabla_\psi \log P(\widehat{x}|y; \psi)$$

This is an efficient approximation of ∇_ψ. Using it, we update the backward model with gradient descent.

Multilingual Training

Problem: The backward model's quality depends on the amount of parallel training data.

Solution: Since the backward model will continue training along with the forward model, we can:

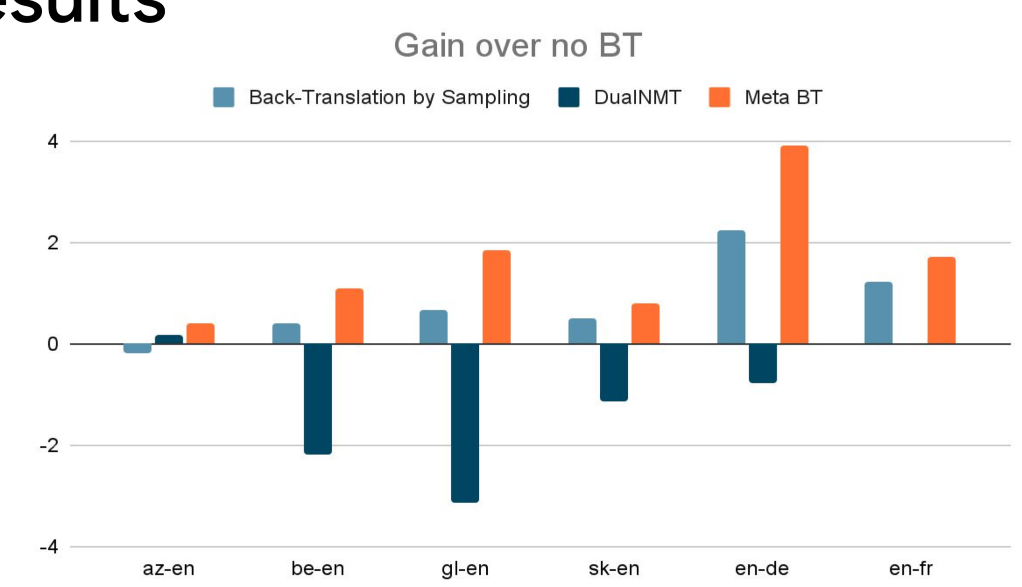
1. Train the backward model on a related language with more data.
   - *Examples (low-high):* az-tr, be-ru, gl-pl, sl-cs
2. Adapt the backward model throughout the forward model's learning.

Experiments

Setup

- Standard BT
  - WMT 2014 en-de en-fr; WMT news monolingual data
- Multilingual BT
  - Low-resource: az, be, gl,sl; paired with high-resource: tr, ru, pl, cs

Results



Analysis

Meta BT avoids overfitting and underfitting

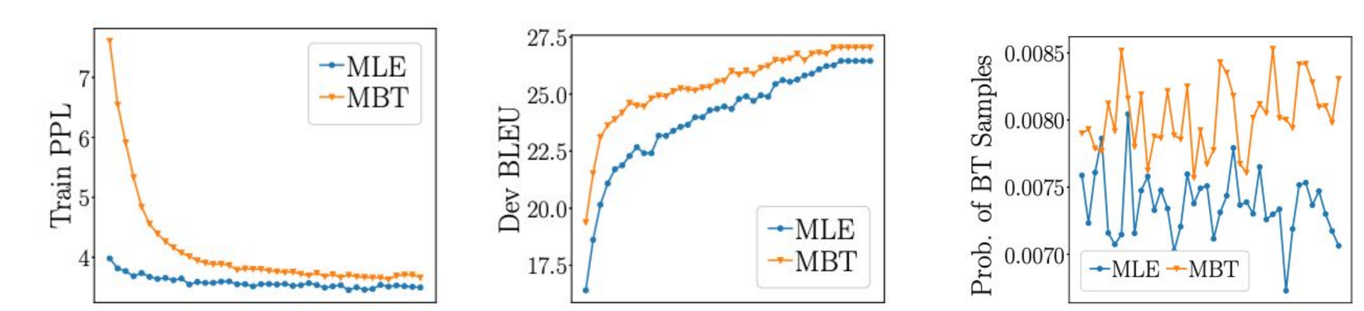


Figure. MBT leads to higher training PPL but better better dev BLEU

Figure. MBT helps underfitting by decreasing the diversity of BT samples

Meta BT samples data closer to meta validation set

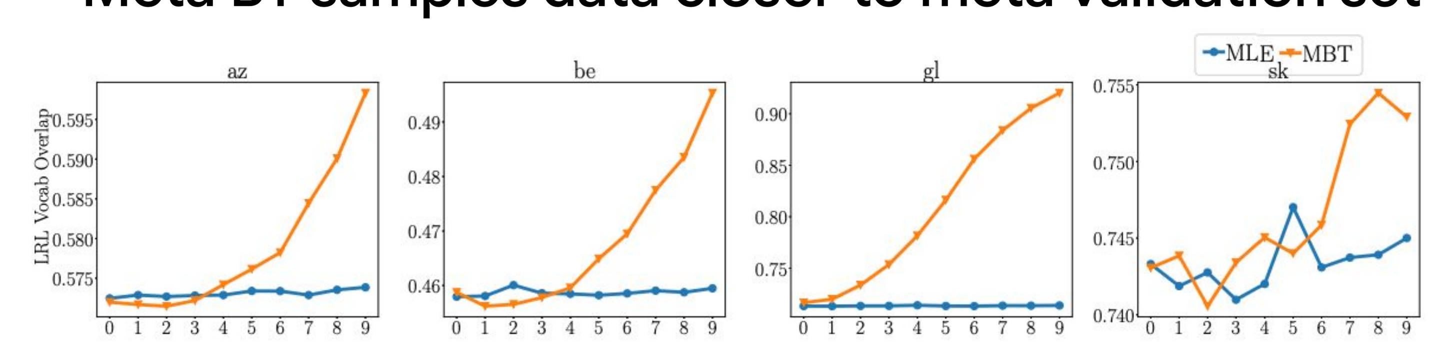


Figure. MBT learns to favor data more similar to the low-resource languages

References

- R. Sennrich, B. Haddow, A. Birch. *Improving neural machine translation models with monolingual data.* ACL 2016
- S. Edunov, M. Ott, M. Auli, David Grangier. *Understanding back-translation at scale.* EMNLP 2018
- Y. Xia, D. He, T. Qin, L. Wang, N. Yu, TY. Liu, WY. Ma. *Dual Learning for Machine Translation.* NIPS 2016