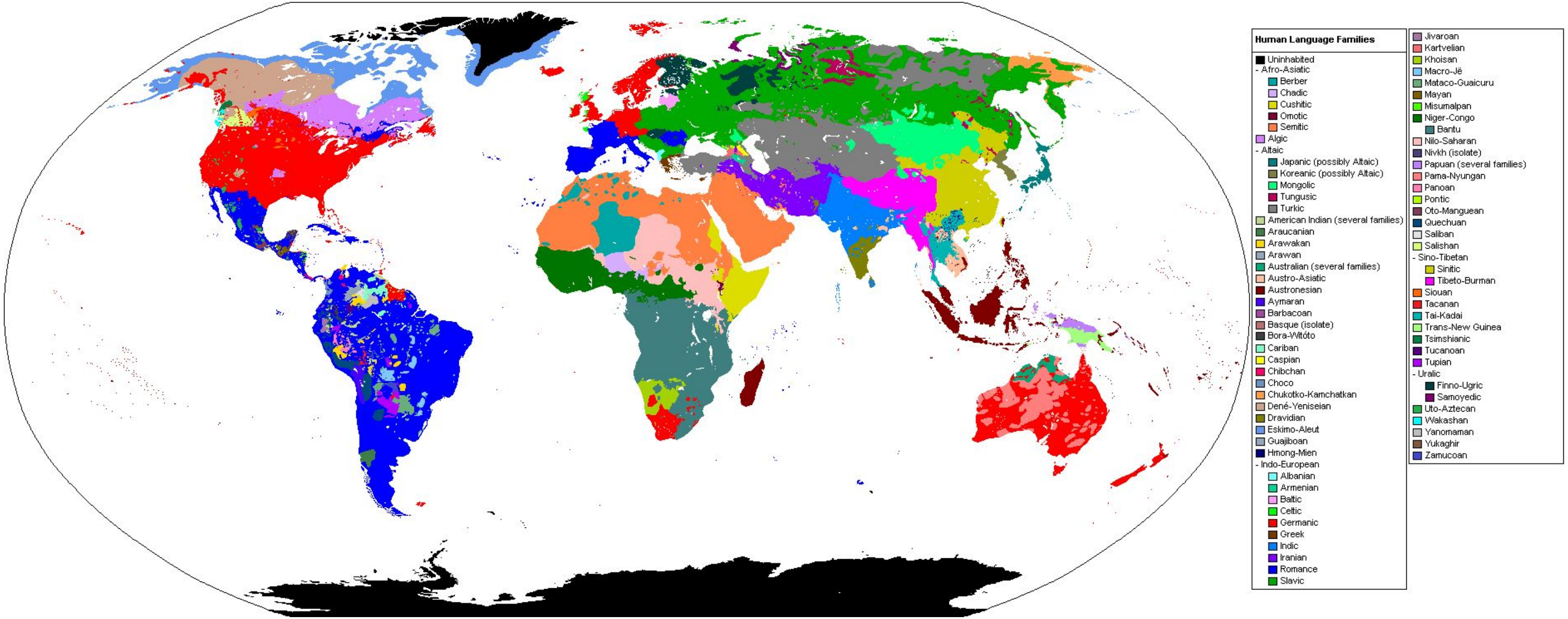
# Multilingual Machine Translation

Surafel M. Lakew

April 26, 2022 | University of Luxembourg (Virtual Presentation)




Ethnologue: > 7000 languages are spoken in the world. Figure source: Wikimedia (CC)

# Talk Outline

- Overview of Multilingual/Neural Machine Translation
- Translation Tasks & Progress in NMT
- NMT Methods, Experiments, Results and Key Takeaways
- Conclusion, Current and Future Trends (Q&A’s)

# Overview

Neural Machine Translation

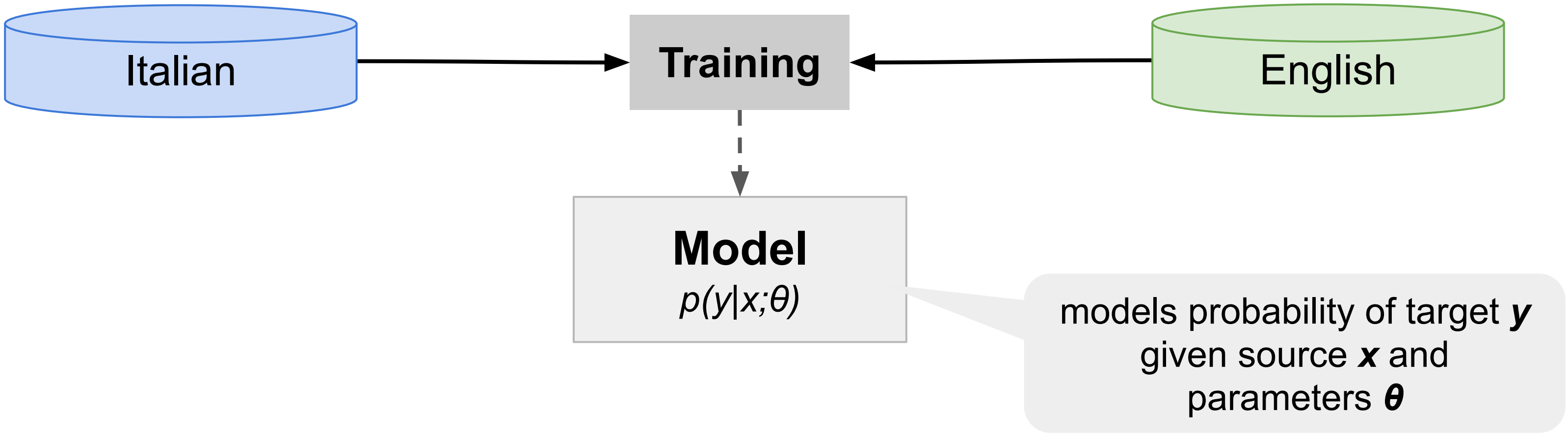
Multilingual
Neural Machine Translation

Task Overview

# Neural Machine Translation

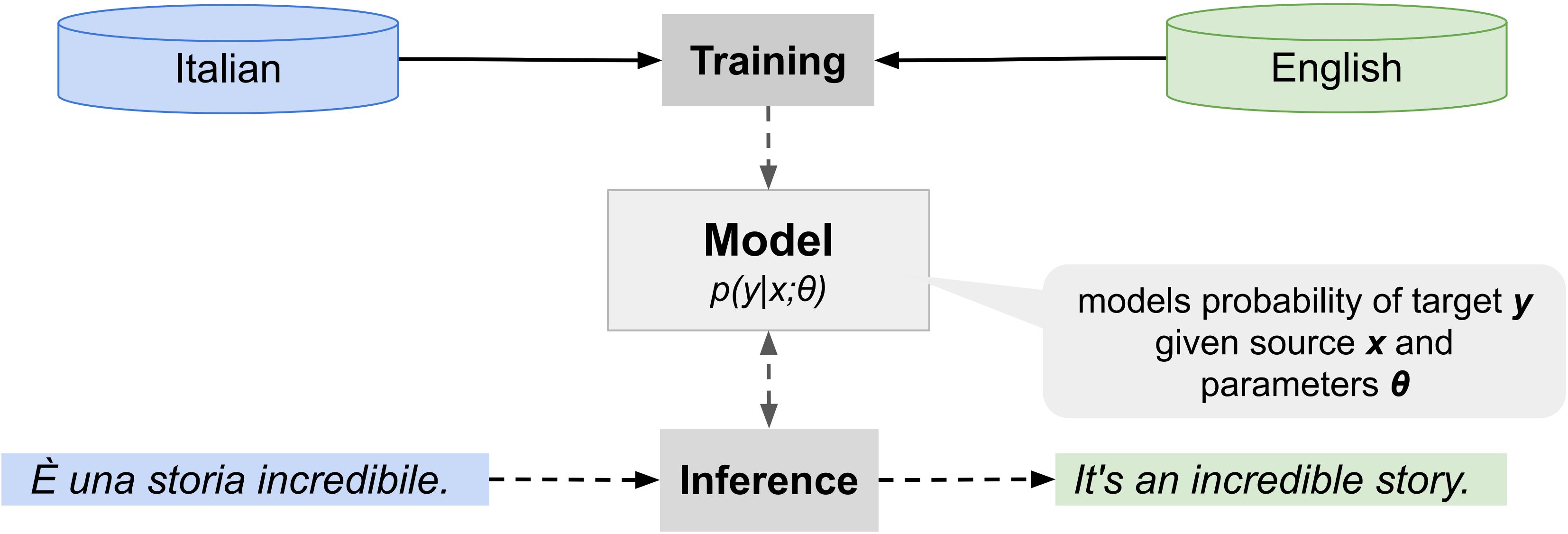

Requires sentence aligned parallel training data between the source (Italian) and target(English) language pairs.

# Neural Machine Translation




The NMT model is parameterized with a **seq2seq (encoder-decoder)** neural network.

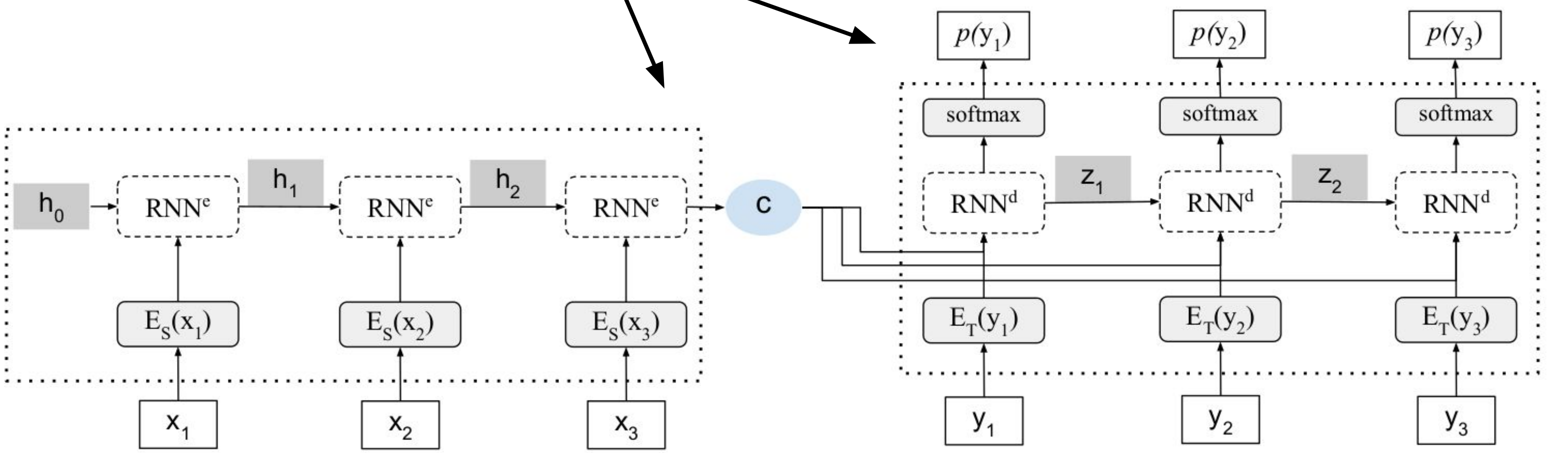
# Neural Machine Translation




The NMT translation performance depends on several factors, ranging from the training data size and domain to the model type and capacity.

# Machine Translation with Recurrent NN

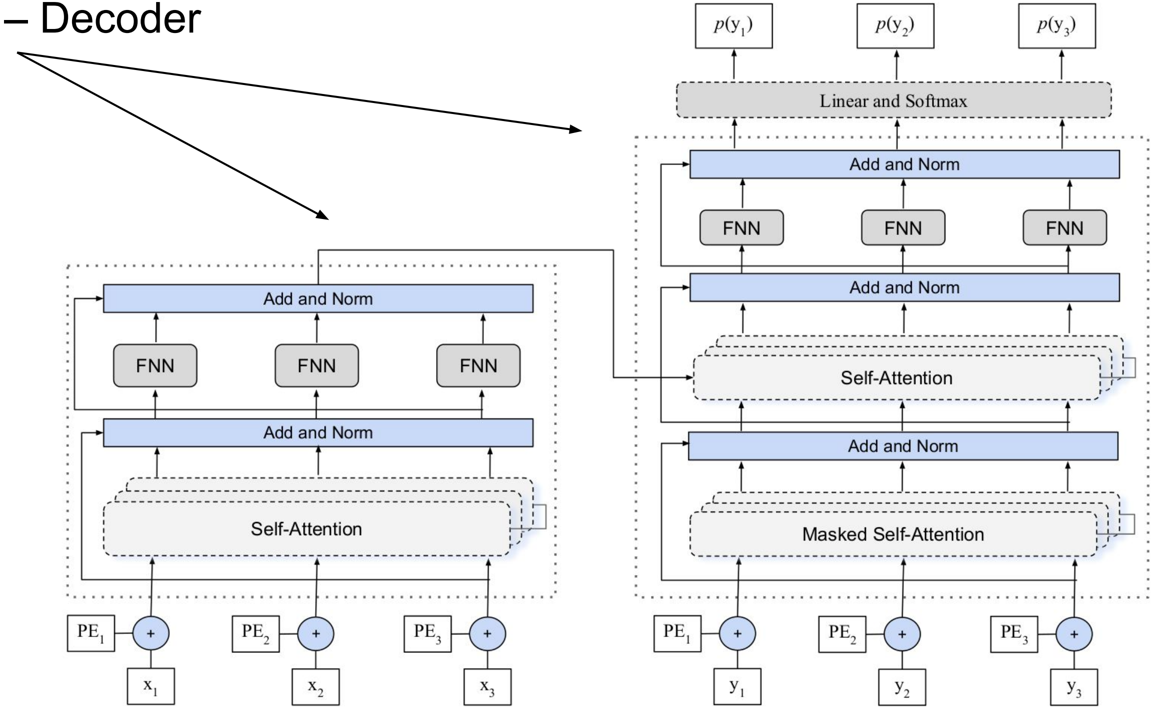
Encoder – Decoder




Sutskever et al. 2014

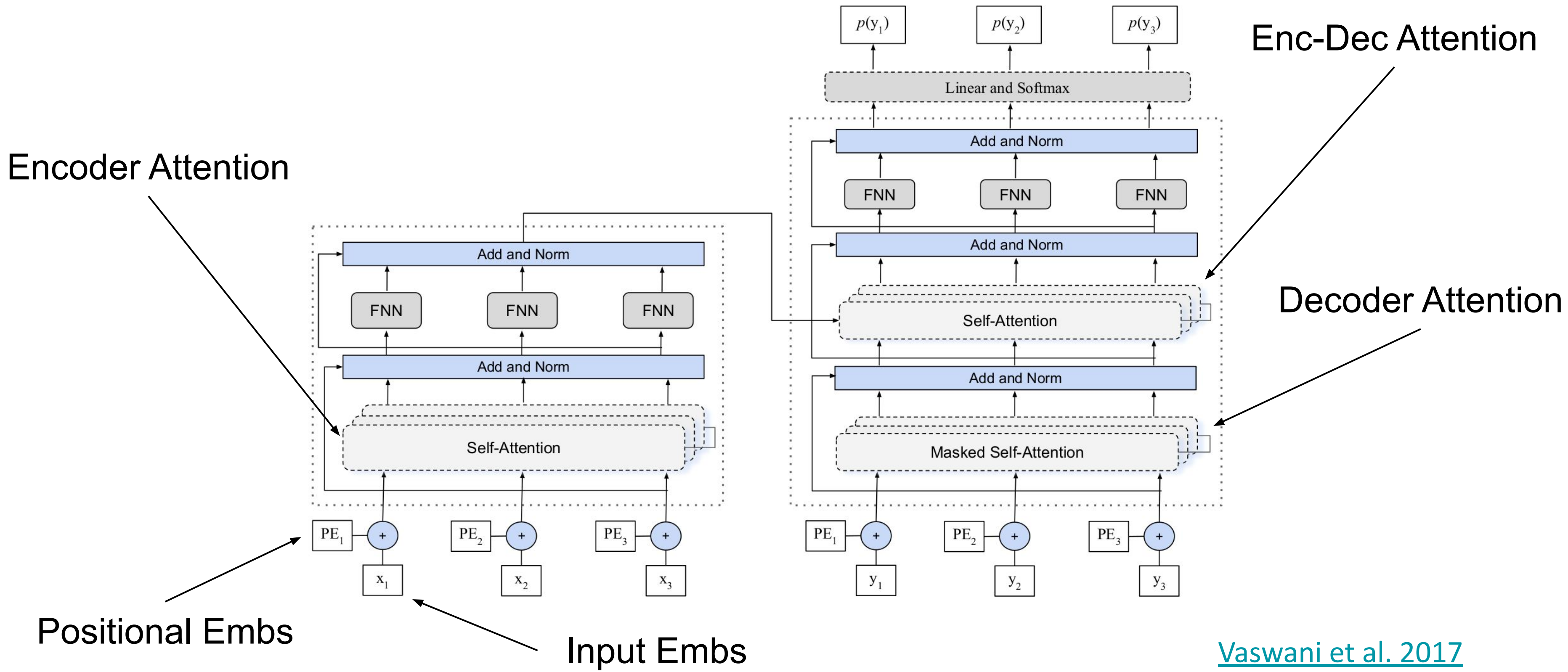
# Machine Translation with Transformer NN

Encoder – Decoder




Vaswani et al. 2017

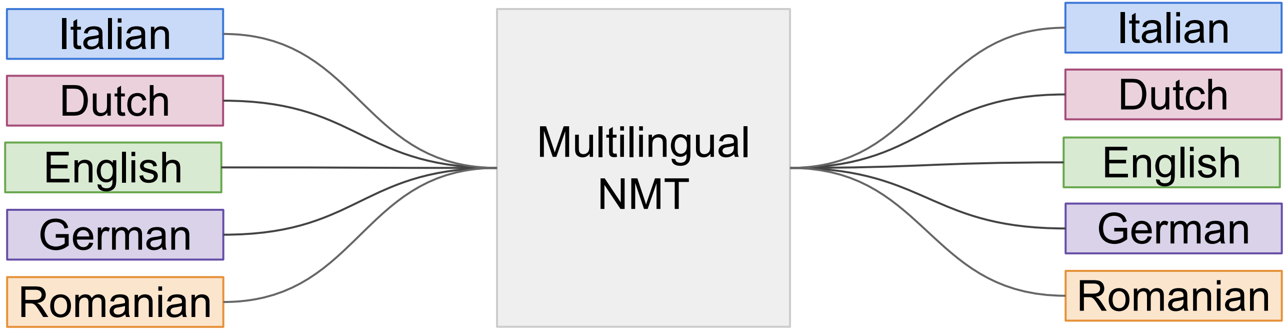
# Machine Translation with Transformer NN




Vaswani et al. 2017

# Overview of Tasks

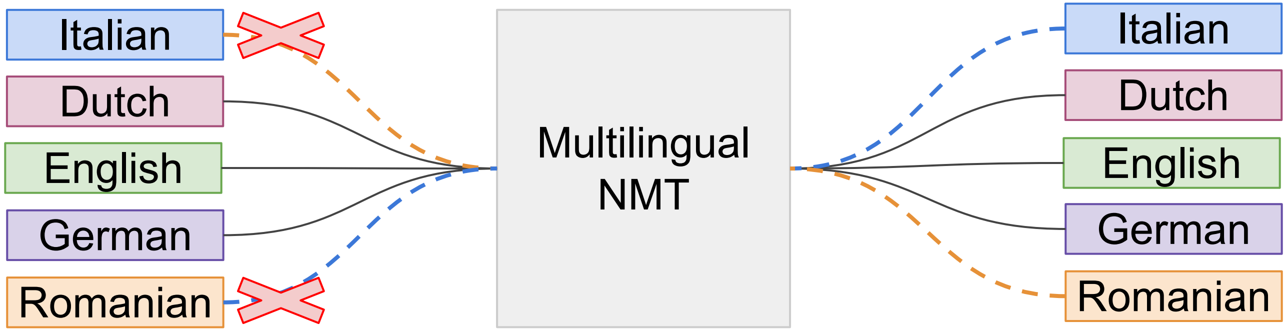
# Multilingual Neural Machine Translation




Modeling a single NMT model to translate between multiple languages

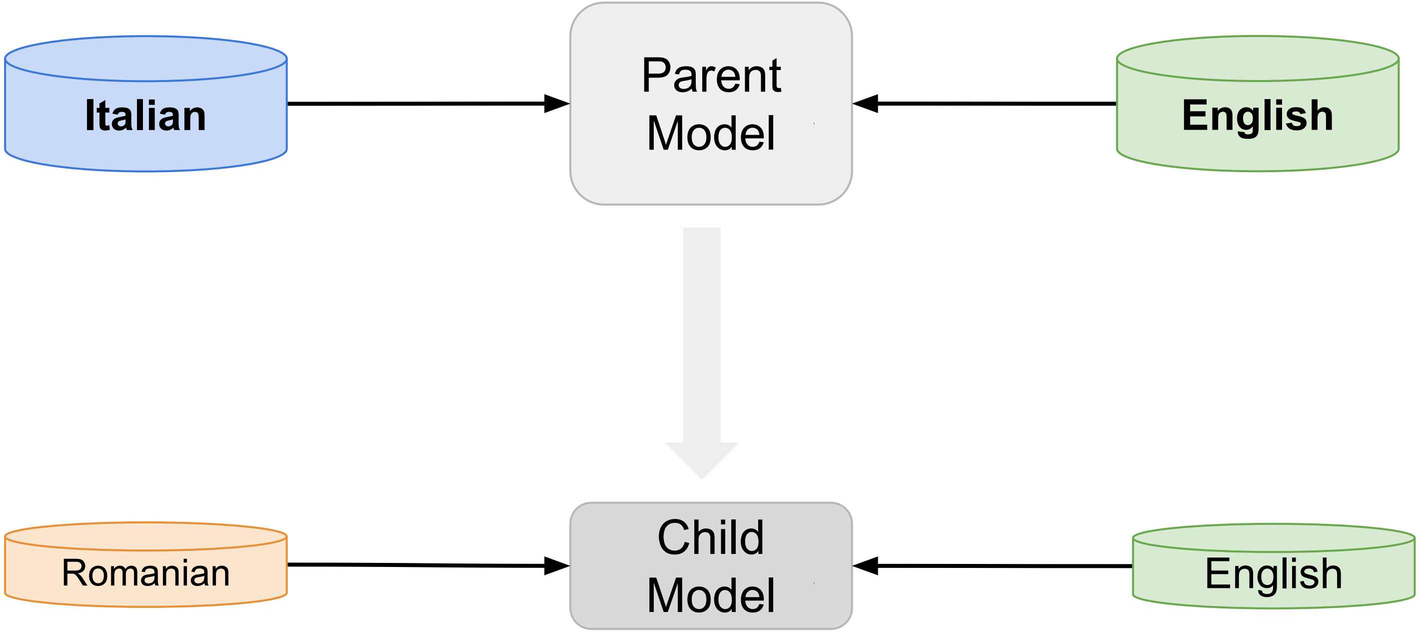
Johnson et al. 2016, Ha et al., 2016

# Low/Zero Resource NMT




In the absence of parallel examples, we use monolingual or/and multilingual data.

# Transfer-Learning in NMT

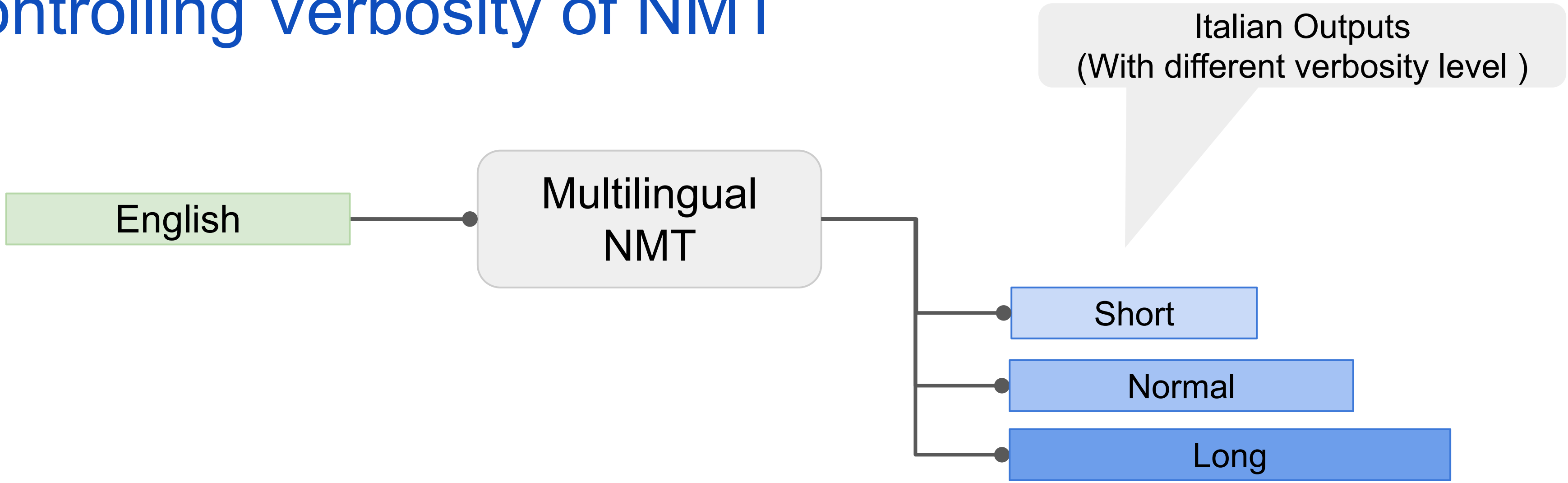



Improving low-resource tasks by leveraging high-resource language pairs.

# NMT into Language Varieties

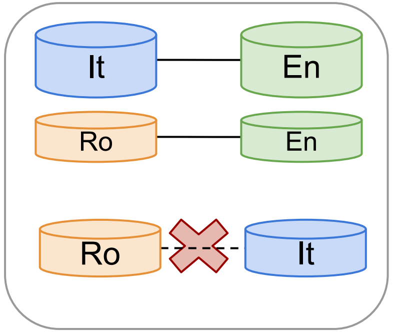

Multilingual NMT repurposed to translate into language varieties.

# Controlling Verbosity of NMT



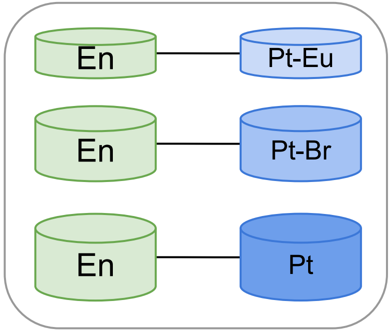

Multilingual NMT repurposed to translate into different verbosity level.

# Overview of Tasks: what makes them similar?

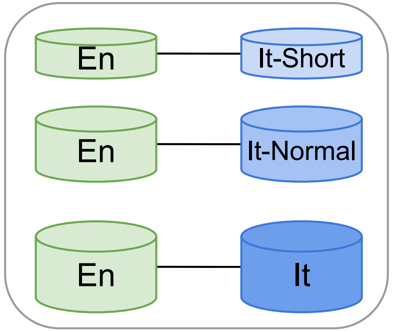



Low-Resource

Zero-Resource



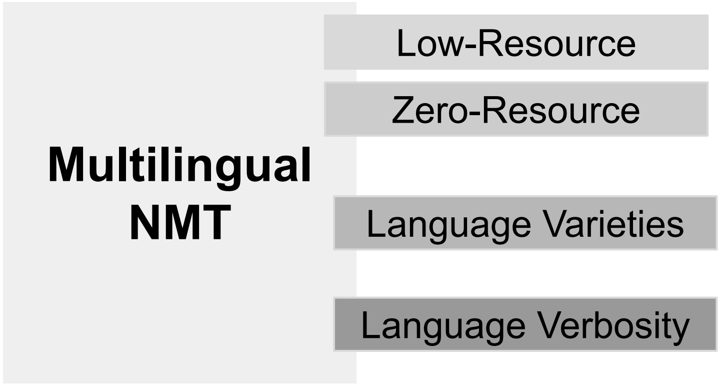

Language Varieties




Language Verbosity

Unbalanced/Unavailable resources across languages, varieties and styles.

# Overview of Tasks: what makes them similar?




Modeling multiple tasks in a single model and enabling positive transfer-learning.

# Progress in NMT

Tasks and Approaches

Low/Zero Resource NMT

Dynamic Transfer Learning for NMT

NMT into Language Varieties

Controlling NMT Verbosity

# Low/Zero-Resource Neural Machine Translation

## Previous Work:

**2015**

### Emergence of Multilingual NMT

Dong et al., an encoder - multi. attn. & decoder.

Luong et al., multiple encoder - decoder.

**2016.v1**

### Towards Multilingual NMT

Firat et al., multi. encoder - attn - multi. decoder
Zoph & Knight., multi-source - attn - decoder
Lee & Cho., many to one, single enc-attn-dec

**2016.v2**

### Single Encoder-Attn-Decoder

Firat et al., zero-resource NMT
Johnson et al., large multilingual & zero-shot
Ha et al., multilingual, pivot and zero-shot

# Low/Zero-Resource Neural Machine Translation

## Previous Work:

**2015**

### Emergence of Multilingual NMT

Dong et al., an encoder - multi. attn. & decoder.

Luong et al., multiple encoder - decoder.

**2016.v1**

### Towards Multilingual NMT

Firat et al., multi. encoder - attn - multi. decoder
Zoph & Knight., multi-source - attn - decoder
Lee & Cho., many to one, single enc-attn-dec

**2016.v2**

### Single Encoder-Attn-Decoder

Firat et al., zero-resource NMT
Johnson et al., large multilingual & zero-shot
Ha et al., multilingual, pivot and zero-shot

*Does multilingual NMT improve in low-resource conditions?*

*Can we further improve Zero-Shot translation of a multilingual NMT?*

*Lakew et al. 2017*

# Transfer Learning for Low-Resource Languages

## Previous Work:

**2016**

### Parent - Child TL

Zoph et al., initialize child model with parent parameters.

**2017**

### TL across related languages

Nguyen et al., transfer across related languages.

**2018**

### TL both for src & tgt languages

Neubig & Hu., pre-train multilingual model.
Kocmi & Bojar., update both src & tgt sides

**2019**

### Data-Augmentation, Multilinguality

Xia et al., combine TL with back-translation.
Aharoni et al., massive multilingual model.

# Transfer Learning for Low-Resource Languages

## Previous Work:

2016

**Parent - Child TL**

**Zoph et al.**, initialize child model with parent parameters.

2017

**TL across related languages**

**Nguyen et al.**, transfer across related languages.

2018

**TL both for src & tgt languages**

**Neubig & Hu.**, pre-train multilingual model.
**Kocmi & Bojar.**, update both src & tgt sides

2019

**Data-Augmentation, Multilinguality**

**Xia et al.**, combine TL with back-translation.
**Aharoni et al.**, massive multilingual model.

*Does dynamic transfer-learning improves over fixed parent model transfer ?*

*Can we do better transfer-learning with relevant data selection ?*

*Lakew et al., 2019*

# NMT into Language Varieties

## Previous Work:

**2017**

### SMT vs. NMT b/n LV

Costa-Jussa., RNN improves Catalan-Spanish MT over PBSMT.

**2018**

### NMT b/n LV

Costa-Jussa et al., NMT b/n EU and BR Portugues.

# NMT into Language Varieties

**Previous Work:**

**2017**

**SMT vs. NMT b/n LV**

Costa-Jussa., RNN improves Catalan-Spanish MT over PBSMT.

**2018**

**NMT b/n LV**

Costa-Jussa et al., NMT b/n EU and BR Portugues.

*Does modeling multiple varieties in a single model is achievable?*

*How to handle unlabeled LV data?*

*Lakew et al., 2018*

# Controlling the Verbosity of NMT

## Previous Work:

**2016**

### NMT, MNMT, Summarization

Sennrich et al.,: politeness in NMT
Kikuchi et al.,: summarization

**2017**

### Summarization

Fan et al.,: conditioning the output on a length token

**2018**

### Customized NMT

Michel & Neubig.: personalized NMT

**2019**

### Summarization, Pos. encoding

Sho & Naoaki.,: adapt positional encoding to length encoding.

# Controlling the Verbosity of NMT

## Previous Work

**2016**

**NMT, MNMT, Summarization**

**Sennrich et al.,:** politeness in NMT
**Kikuchi et al.,:** summarization

**2017**

**Summarization**

**Fan et al.,:** conditioning the output on a length token

**2018**

**Customized NMT**

**Michel & Neubig.:** personalized NMT

**2019**

**Summarization, Pos. encoding**

**Sho & Naoaki.,:** adapt positional encoding to length encoding.

*Can we bias length of an NMT output, while keeping the translation quality ?*

*Can we make it versatile to any pre-trained model?*

**Lakew et al.,** IWSLT, *2019.*

# Approaches & Applications

Methods, Experiments, Results and Findings

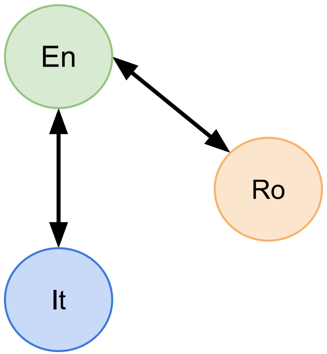
**Zero-Shot NMT Modeling**

Dynamic Transfer Learning

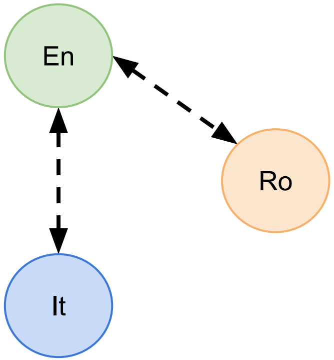
NMT into Language Varieties

Controlling NMT Verbosity

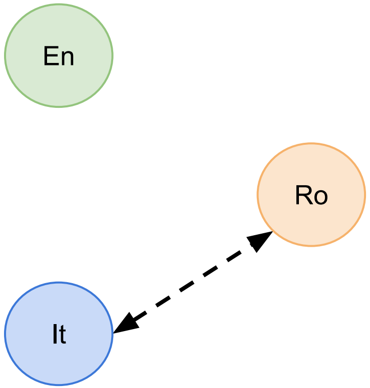
# Zero-Shot Translation




Multilingual Training



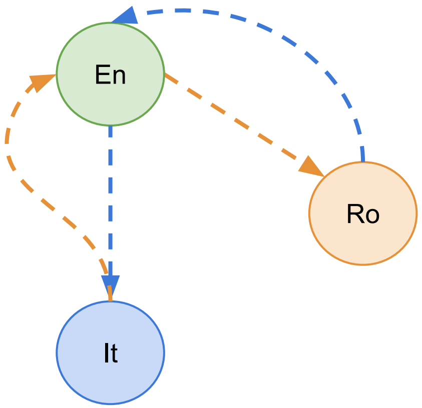

Non Zero-shot Inference




Zero-Shot Inference

Zero-Shot Translation is among the advantages of Multilingual NMT

# Pivoting Translation as Alternative




Pivot (N-step) Inference

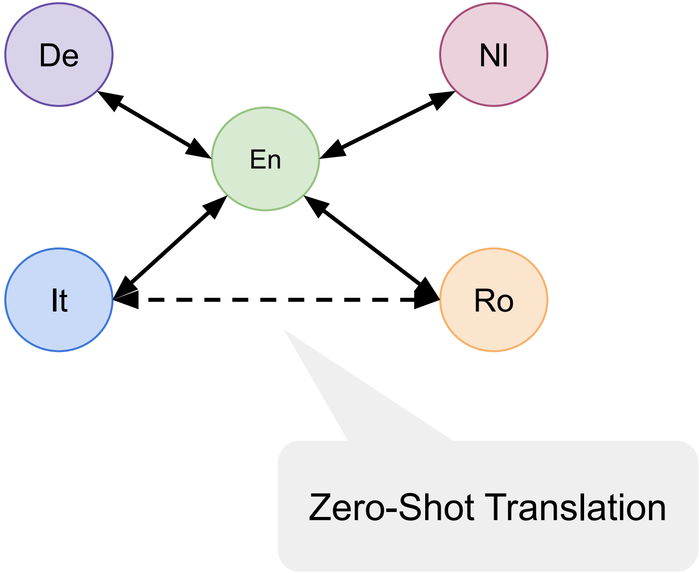
# Research Questions

*Does multilingual NMT improve low/zero-resource translation ?*

*Can we further improve Zero-Shot translation of a multilingual NMT ?*

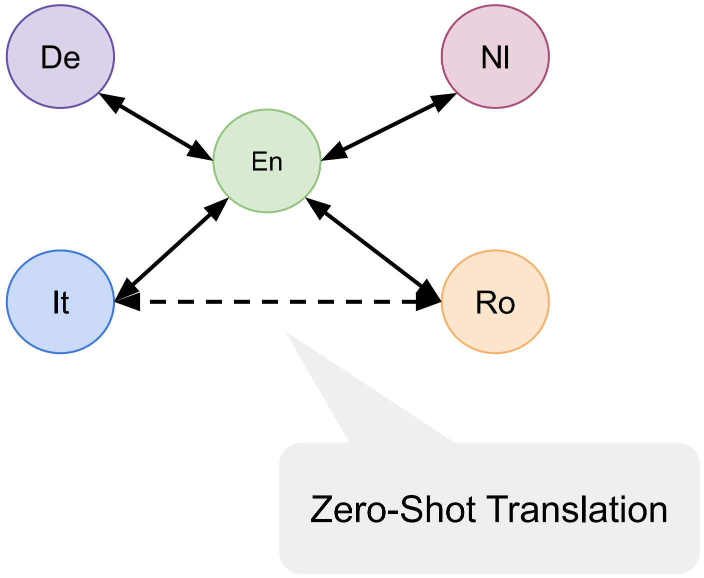
# Zero-Shot NMT Modeling

## Available Resource and ZST Task



# Zero-Shot NMT Modeling

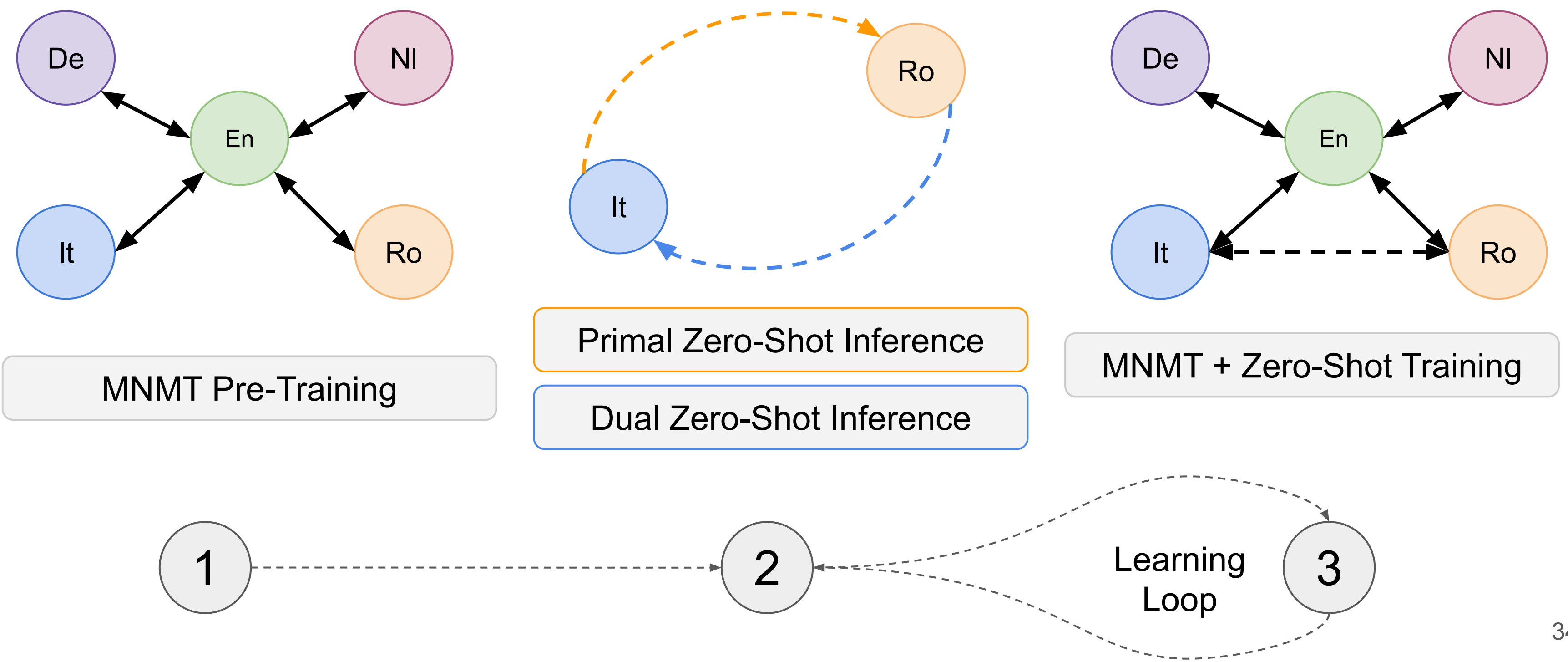
## Available Resource and ZST Task



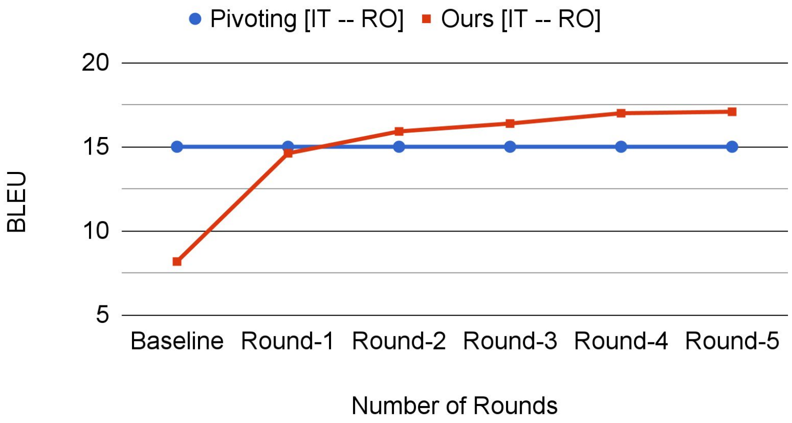

## Zero-Shot NMT Learning Steps

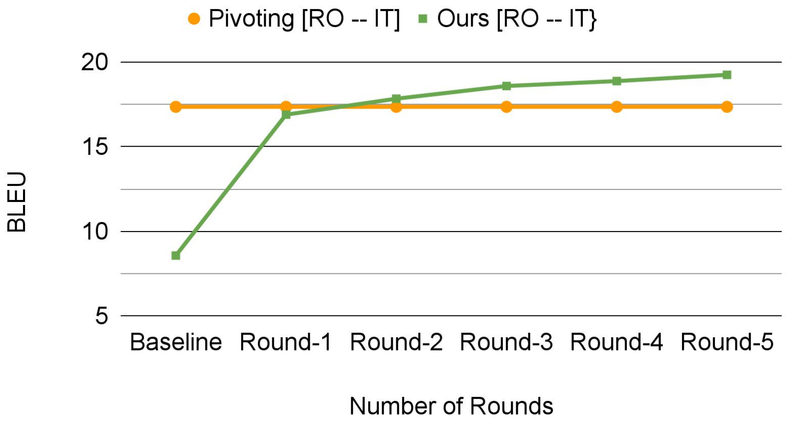
- Leverage monolingual data
- Perform dual back-translation
- Self-Learning using iterative data augmentation & learning with supervised tasks.

# Zero-Shot NMT Modeling: Three Steps



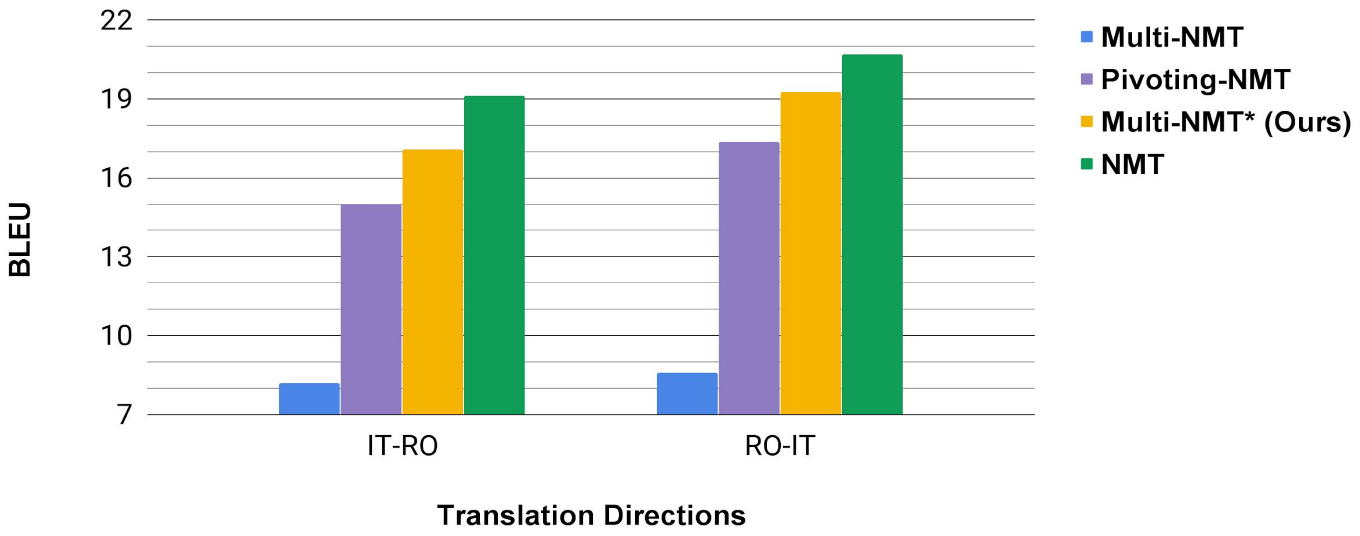
# Results




Results of the Italian <> Romanian zero-shot directions on IWSLT test-2017

# Comparative Results




Zero-Shot NMT modeling outperformed the baseline **Multilingual NMT** and the **Pivoting** mechanism on IWSLT *test-2017.*

# Key Takeaways

- Improves over the initial zero-shot translation only approach
- Learns through the different round of training and inference
- Shows better performance than pivoting
- Signals the universality of multilingual NMT

**Lakew et al.**, *IWSLT, 2017*

# Approaches & Applications

Zero-Shot NMT Modeling

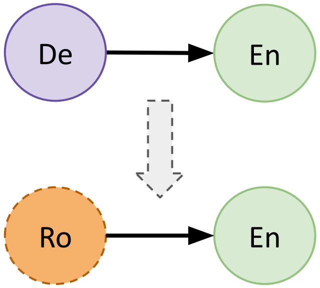
**Dynamic Transfer Learning**

NMT into Language Varieties

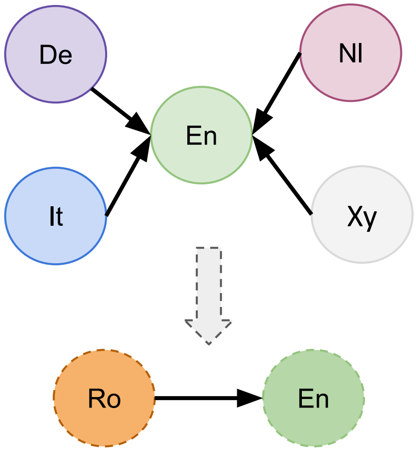
Controlling NMT Verbosity

# Transfer Learning

Parent > Child




Transfer from M-NMT




**Notice:** the parent model parameters are fixed following a one-size-fits-all approach.

# Research Questions

*Does dynamic transfer-learning improves over fixed parent model transfer-learning?*

*Can we expand pre-trained NMT into unseen languages directions?*

*Can we do better transfer-learning with data selection?*

# Dynamic Transfer Learning: Two Approaches

**Progressive Adapt (ProgAdapt)**

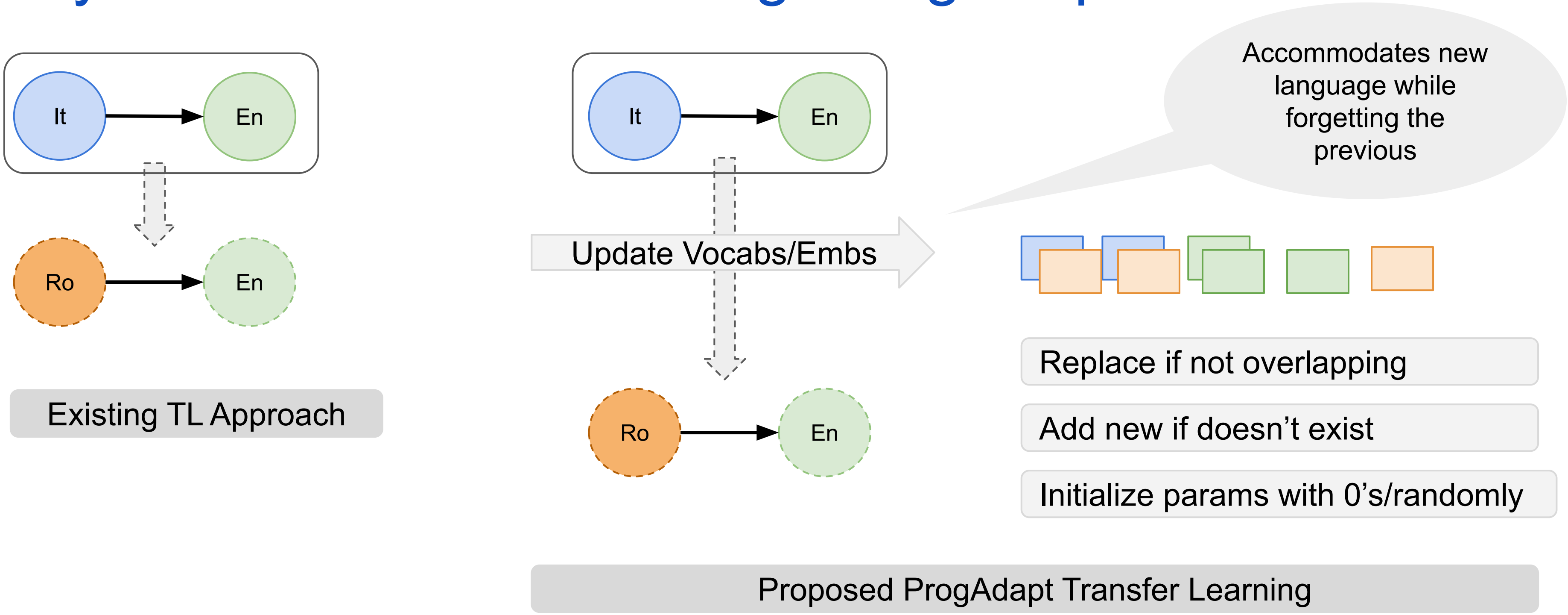
- Transfer parent model parameter to child model with new language pair.

**Progressive Grow (ProgGrow)**

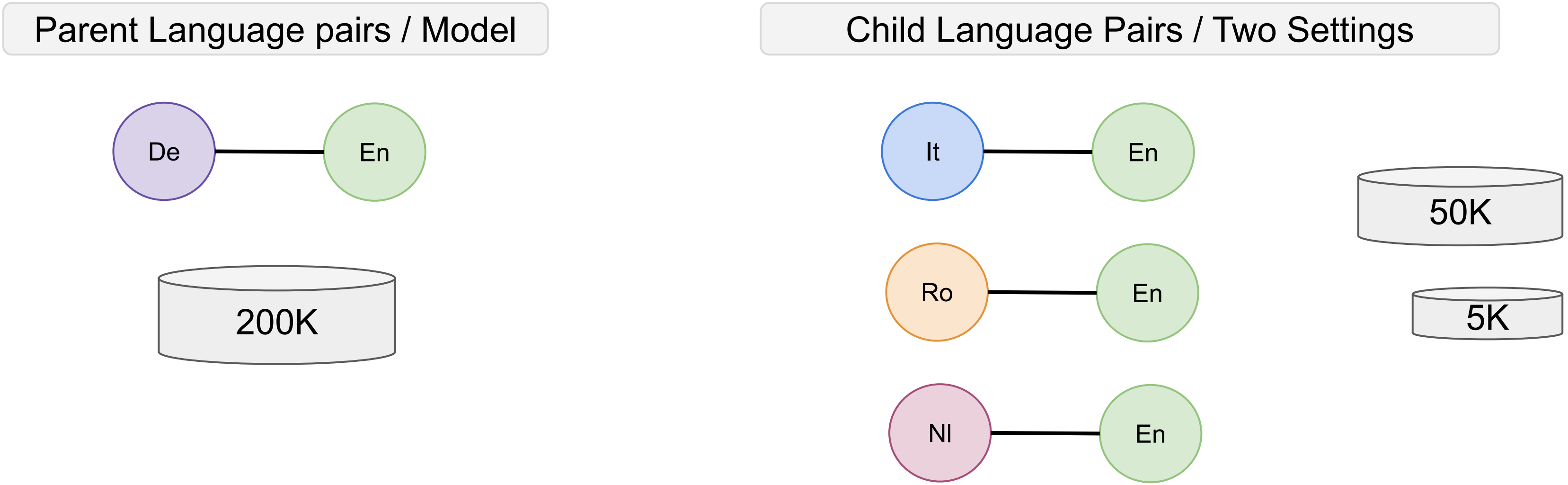
- Accommodate new language pairs when data becomes available.

**Lakew et al.**, *IWSLT, 2018*

# Dynamic Transfer Learning: ProgAdapt

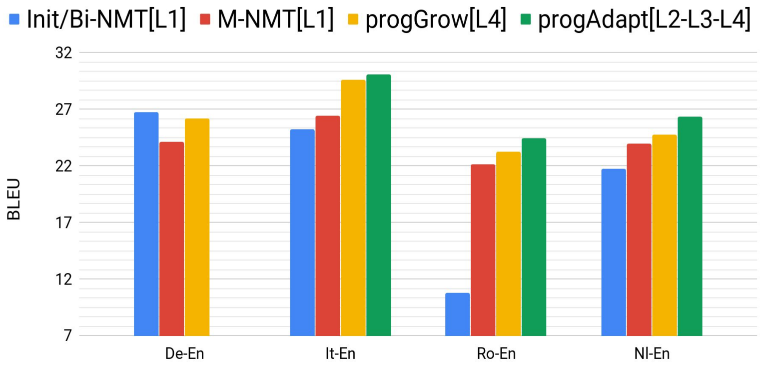


# Experimental Settings



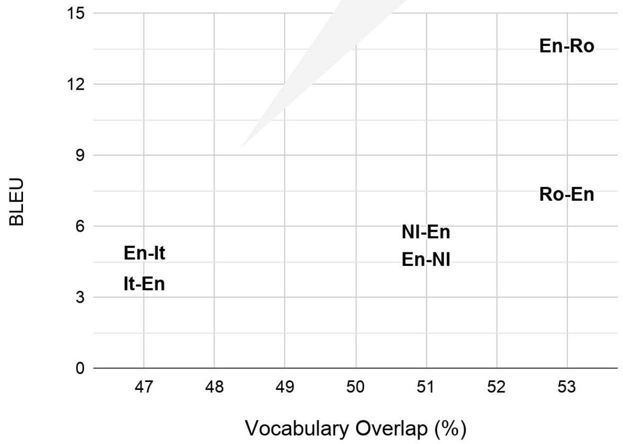
# Results

DTL approach outperform both single-pair NMT and M-NMT approaches




Low-Resource Results

In ProgAdapt parent model vocab size is ~10% larger




**Key Takeaway**

A higher % of shared vocabulary b/n consecutive models in progAdapt shows better gain.

Compared to a model trained from scratch, DTL takes 4% to 20% training steps with significantly higher performance.

# Dynamic Transfer Learning

## Two Additional Proposals

### Multilingual Model

- Train a large scale multilingual parent model to dynamically transfer parameters.

### Data selection for TL

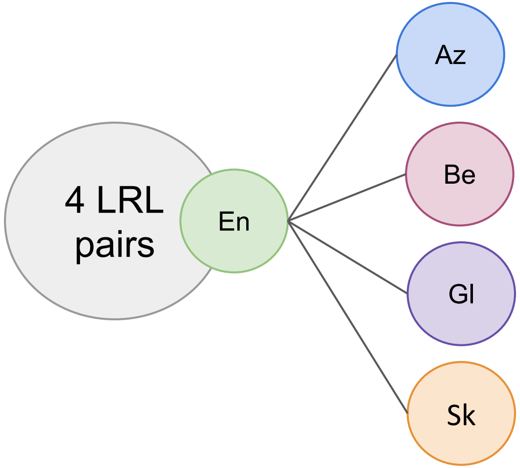
- Train language model on the test language (child) to select relevant data for the TL stage.

Lakew et al., 2019

# Experimental Settings

54 languages paired with English

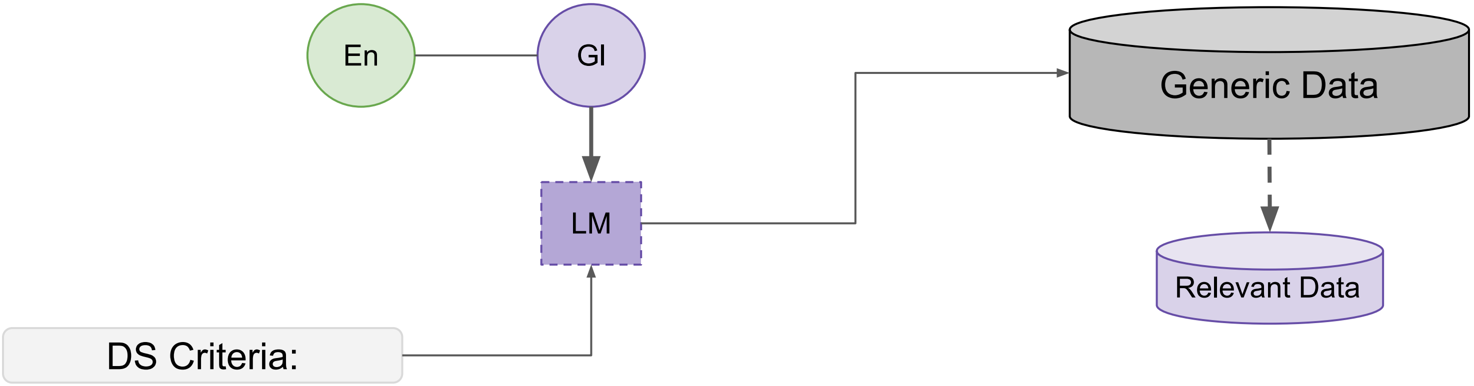
Used to pre-train 108 direction multilingual (parent) model




| | |
|---|---|
| Az | 5.9k |
| Be | 4.5k |
| Gl | 10k |
| Sk | 61.5k |

Low-resource language pairs (child)

# Dynamic Transfer Learning: Data Selection Strategies




**Select-pplx**: select data with lowest pplx from HRLs pool

**Select-one**: select all data from one HRL related to the LRL

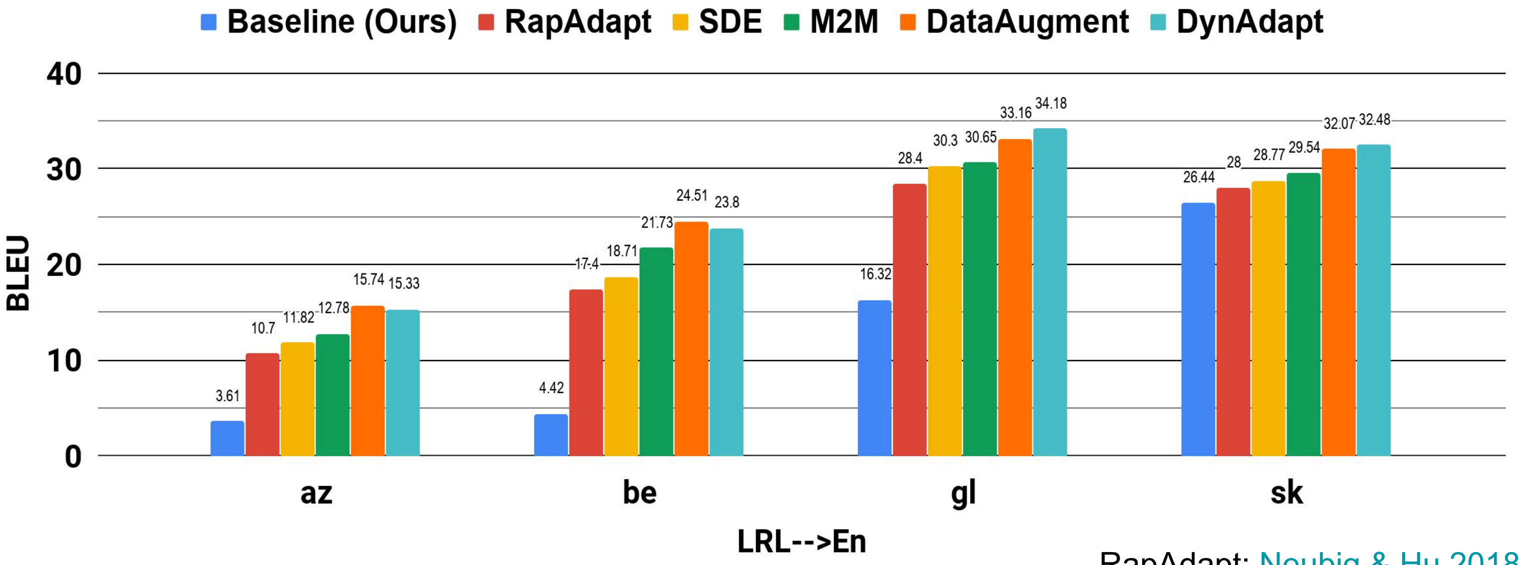
**Select-rand**: randomly sample from HRLs related/family to the LRL

**Select-fam**: select all data from a set of HRLs related/family to LRL

*except for Select-fam, other selection strategies pick equal proportion of data.

# Results

Using Select-One strategy




RapAdapt: Neubig & Hu 2018
SDE: Wang et al., 2018
M2M: Aharoni et al., 2019
DataAugment: Xia et al., 2019
DynAdapt: Lakew et al., 2019

# Key Takeaways

- Utilizing a universal pre-trained multilingual model improves dynamic/TL for LRL’s.
- Relevant data-selection further improves dynamic adaptation & cheaper to acquire.
- DynAdapt + Data selection approaches can provide the best performance over the other data augmentation and transfer-learning approaches.

# Approaches & Applications

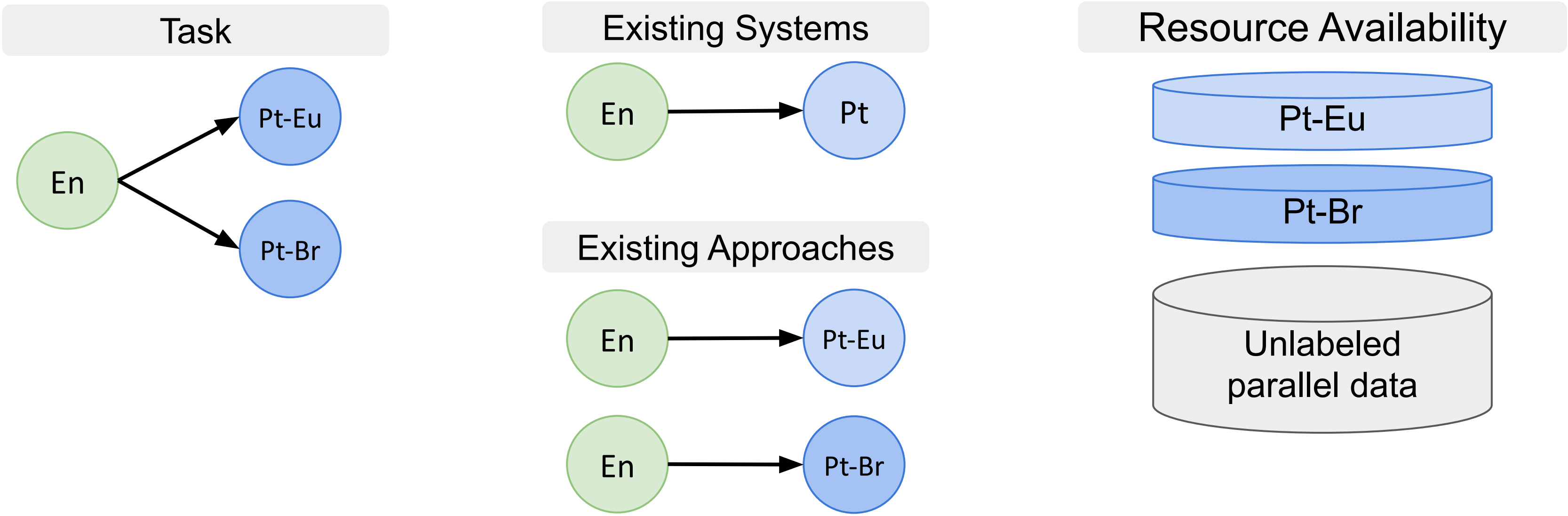
Zero-Shot NMT Modeling

Dynamic Transfer Learning

**NMT into Language Varieties**

Controlling NMT Verbosity

# NMT into Language Varieties: A scenario




A large scale unlabeled data can lead to poor performance when translating to a specific language variety/dialect.

# Research Questions

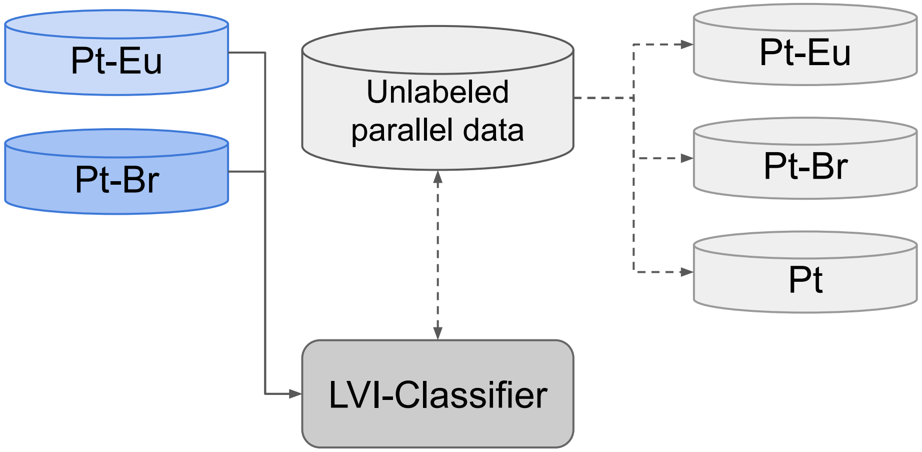
*Can we model translation in to multiple language varieties/dialects using a single model?*

*Can we further improve over the baseline single LV models?*

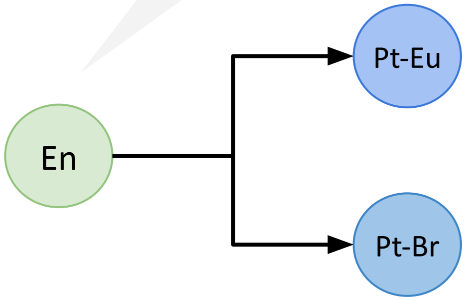
*How to handle majority of LV unlabeled parallel data?*

# Modeling NMT into Language Varieties

We use a similar principles as in multilingual NMT



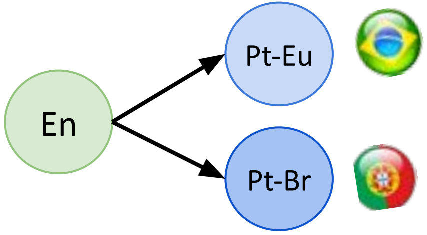

Offline labeling using an LVI classifier



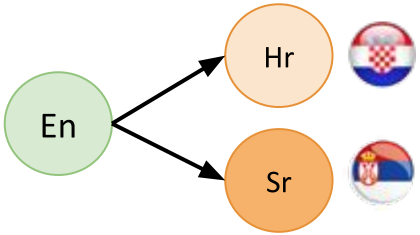

A single LV aware NMT model Training

[Lakew et al., WMT, 2019]

# Experimental Settings: Two Scenarios




Dialects




Closely Related Languages

# Experimental Settings: Data regimes & model types

**Gen:** unsupervised NMT model trained with the union of unlabeled data

**Spec:** supervised models trained with variety specific data

**Mul:** supervised model trained with the union of labeled data

**M-U:** semi-supervised trained with the union of both labeled and unlabeled data

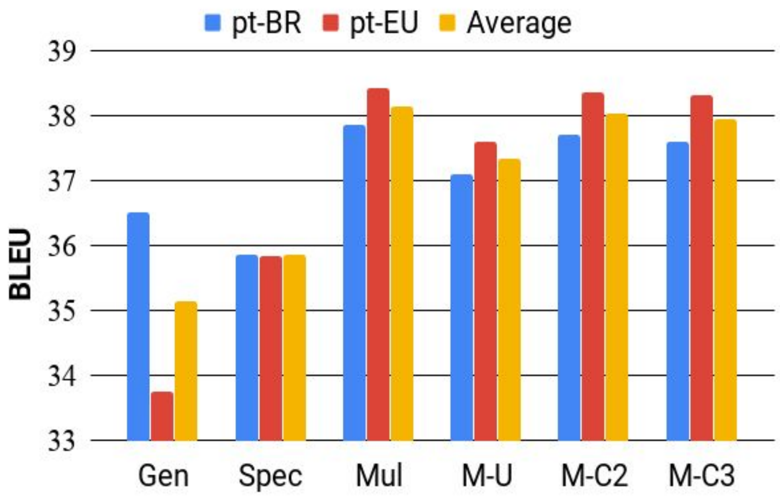
**M-C2:** semi-supervised training using LVI to map the unlabeled segments to variety classes

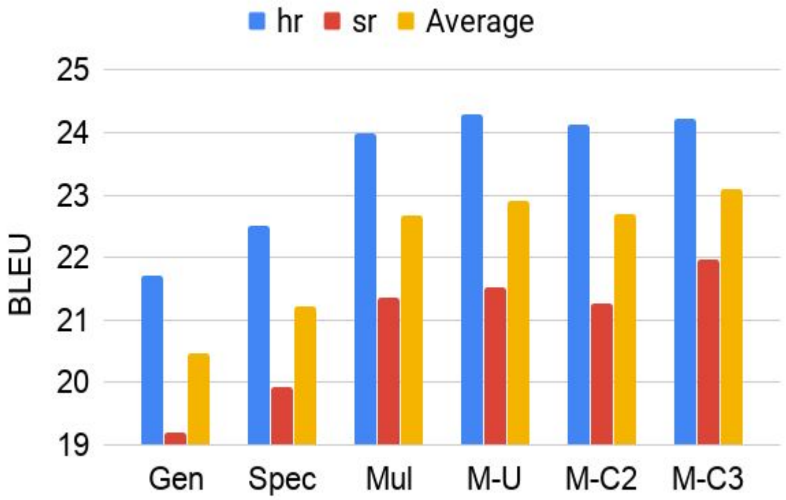
**M-C3:** trained similarly as M-C2, ambiguous sentences with low classifier confidence are not labeled

# Results

Mul (supervised) shows the largest improvement, with comparable performance from semi-supervised (M-C2/3)

Semi-supervised approaches are better than Mul



# Key Takeaways

- Presented NMT from English into dialects & related languages, comparing models that can be trained under unsupervised, supervised, and semi-supervised settings.

- Multilingual model (M-C3) trained using labels from LVI module can perform very close to its supervised (Mul) version.

- The approach keeps resource together maximizing transfer-learning b/n the high and low-resourced variety.

- Delivers simplified modeling, in addition to improved performance & translation quality.

# Approaches & Applications

Zero-Shot NMT Modeling

Dynamic Transfer Learning

NMT into Language Varieties

**Controlling NMT Verbosity**

# Length Control of NMT Outputs: A Scenario

What if translations have to fit a given layout?
E.g. translating subtitles, dubbing script, headlines.

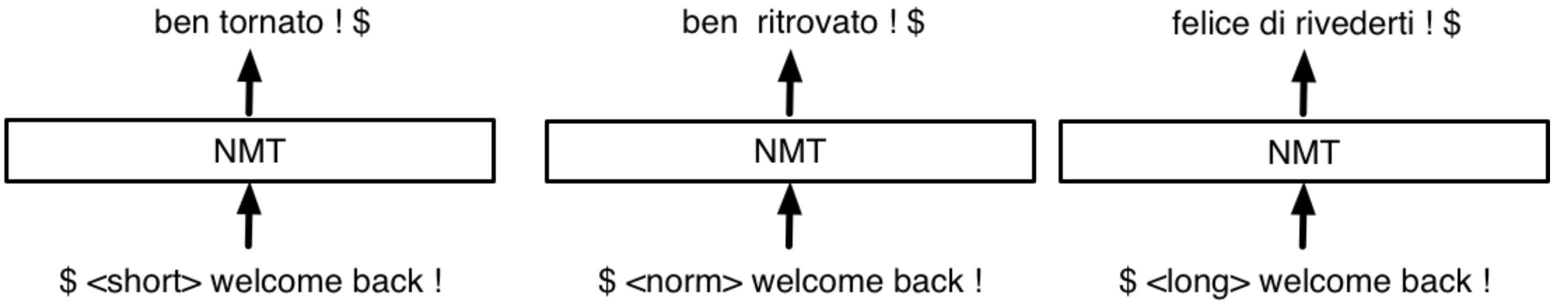
| | |
|---|---|
| SRC | It is actually the true integration of the man and the machine. |
| MT | Es ist tatsächlich die wahre Integration von Mensch und Maschine. |
| MT* | Es ist die wirkliche Integration von Mensch und Maschine.------ |
| SRC | So we thought we would look at this challenge and create an exoskeleton that would help deal with this issue. |
| MT | Quindi abbiamo pensato di guardare a questa sfida e creare un esoscheletro che potesse aiutare ad affrontare questo problema. |
| MT* | Pensavamo di guardare a questa sfida e creare un esoscheletro che potesse aiutare a risolvere il problema.--- |

# Research Questions

*Can we control length of an NMT output, while keeping the translation quality ?*

*Can we make it versatile to any pre-trained model ?*

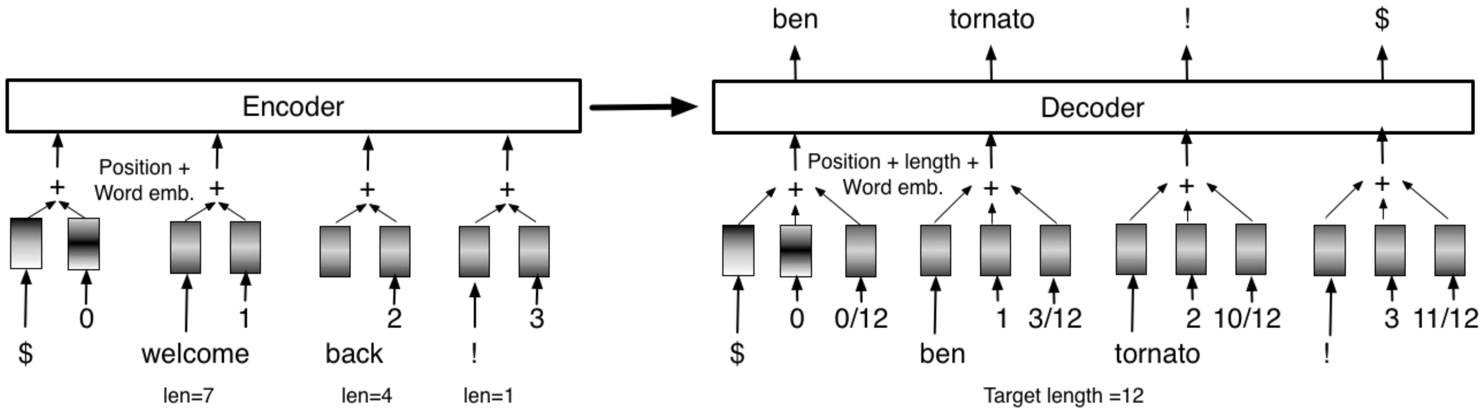
# Controlling Verbosity of NMT: Length-Token




Approach conditions the output of NMT to a given target-source length-ratio class

Lakew et al., 2019

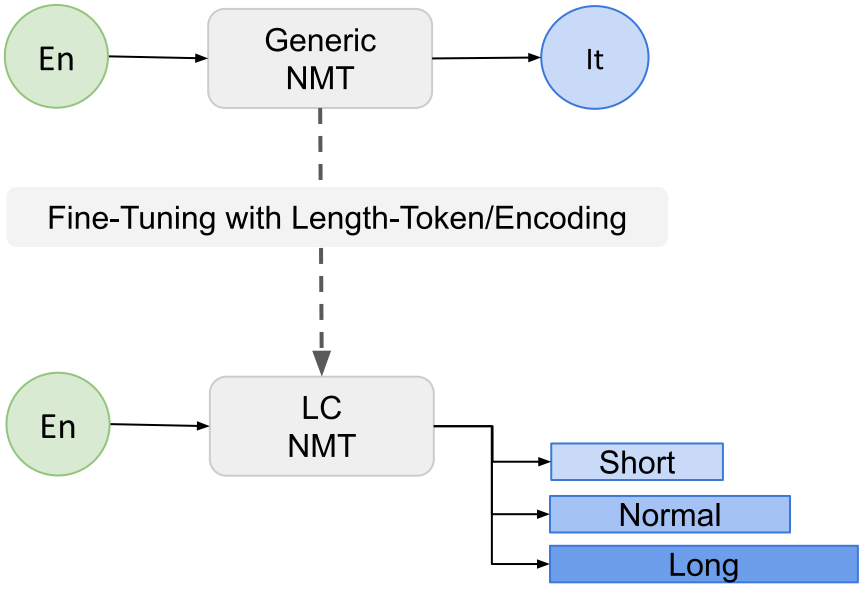
# Controlling Verbosity of NMT: Length-Encoding




Approach enriches the positional embedding of NMT with length information.

Lakew et al., 2019

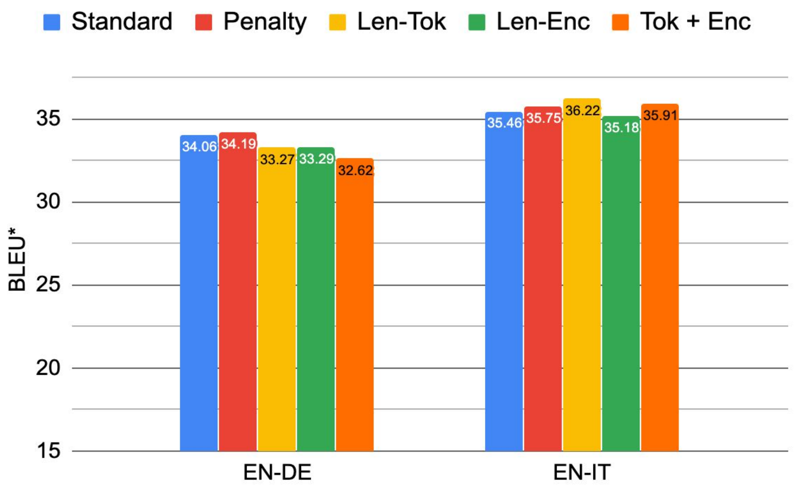
# Controlling Verbosity of NMT: as a Fine-Tuning Task

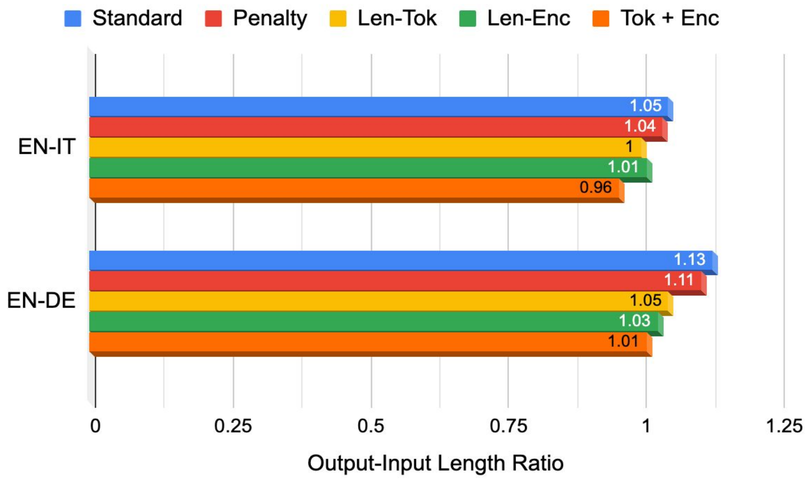



**Advantages:**

- Versatile to any pre-trained model
- Better performance than training from scratch
- Faster training and language independent

# Experimental Results




Models performance (left) with respect to output length (right)

# Examples

Examples of shorter translations obtained by (linguistic variations) paraphrasing, drop of words, and change of verb tense.

| | English > Italian |
|---|---|
| SRC | **And we in the West** couldn't understand |
| NMT | *E noi occidentali* non riuscivamo a capire |
| **LC-NMT** | *In occidente* non riuscivamo a capire ----- |
| SRC | how much **this would restrict** freedom |
| NMT | quanto questo avrebbe limitato la libertà |
| **LC-NMT** | quanto limitasse la libertà ------------------- |
| SRC | **this** is a **really** extraordinary honor for me |
| NMT | questo è  un onore davvero straordinario per me |
| **LC-NMT** | per me è un onore straordinario ---------------------- |

# Key Takeaways

Proposed solutions for controlling the output length of NMT:

**Length-Tok:** coarse-grained control without degradation in quality

**Length-Enc:** fine-grained control with a slight decrease in the translation quality

**Fine-Tuning:** versatile to any pre-trained model

# Conclusions

**Multilingual**
**Neural Machine Translation**

# Conclusions

- Multilingual NMT enables zero-shot translation, and
- Leveraging monolingual data improves zero-shot translation by large margin.
- Dynamic transfer-learning that tailors the parent (multilingual) model to the child model increases translation performance.
- Multilingual model can be repurposed to enable translation into language varieties (dialects), and verbosity of the NMT model outputs.

# Current and Future Trends

Two primary directions

- Specialization
  - Improving a model performance on specific task or  language.
- Generalization
  - Improving a model performance on several tasks or languages.

# Current and Future Trends

Among current trends (reading material):

- Large scale model training/pre-training
    - Kim et al., 2021 Scalable and Efficient MoE Training for Multitask Multilingual Models
- Multimodal/Multi-task training
    - Bapna et al., 2022 mSLAM: Massively multilingual joint pre-training for speech and text
- Self-Learning for multilingual training
    - Siddhant et al., 2022 Towards the Next 1000 Languages in Multilingual Machine Translation

# Thank You! Q&A’s …

Slides: https://github.com/surafelml/talks/machine-translation/

Contact: Surafel M. Lakew  surawinfo@gmail.com