Multilingual Neural Machine Translation for Low-Resource Languages

Surafel Melaku Lakew
University of Trento & Fondazione Bruno Kessler

Advisor: Marcello Federico
Amazon AI & Fondazione Bruno Kessler

Doctoral Thesis Defence | April 20th 2020 | Trento, Italy

Thesis Presentation Outline

- Introduction / Overview of the Thesis /
- Problem Statements / Challenges and Motivations /
- Thesis Contributions / Methods, Experiments, Results and Findings /
- Thesis Conclusion
- Questions and Answers

Introduction

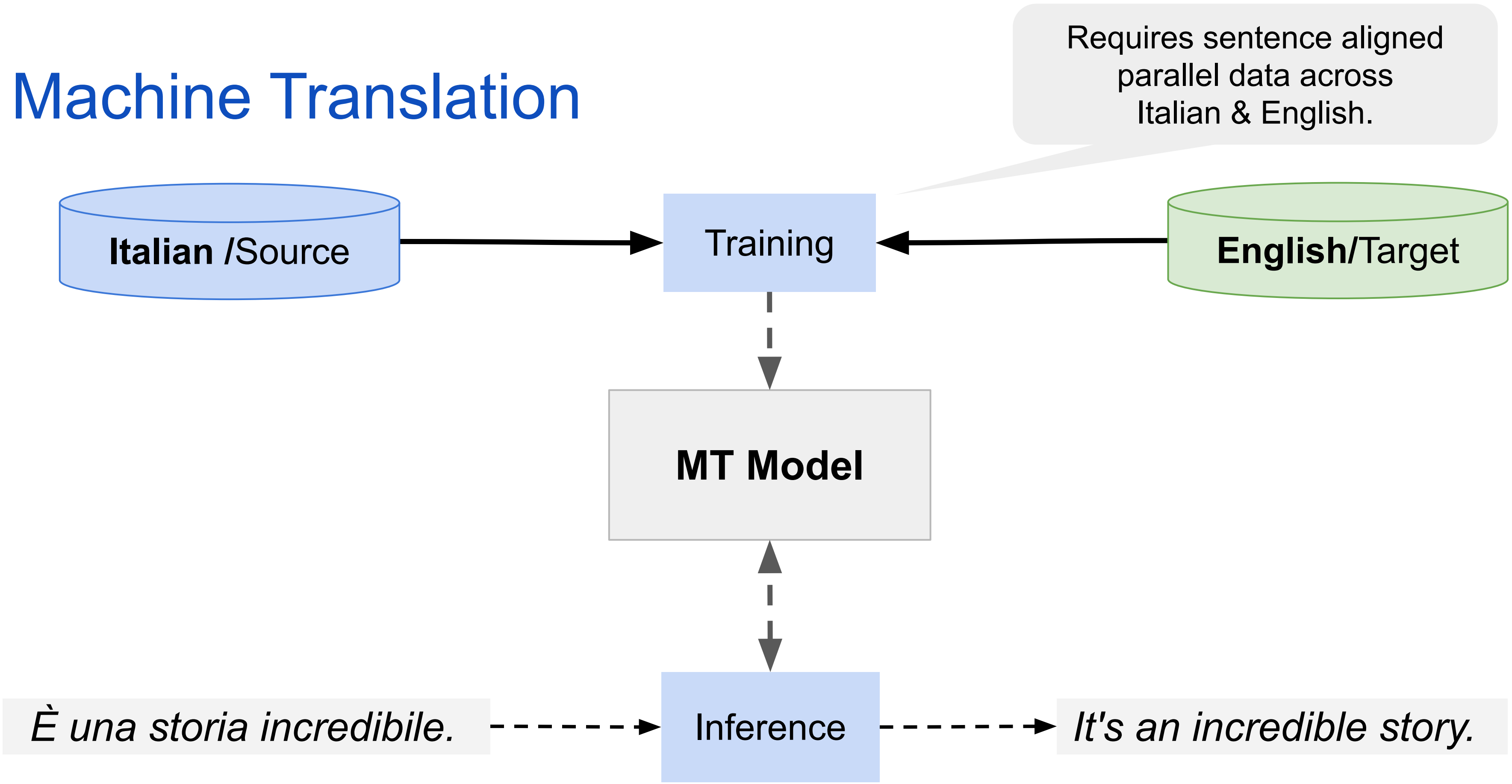
Overview of the Thesis

Neural Machine Translation

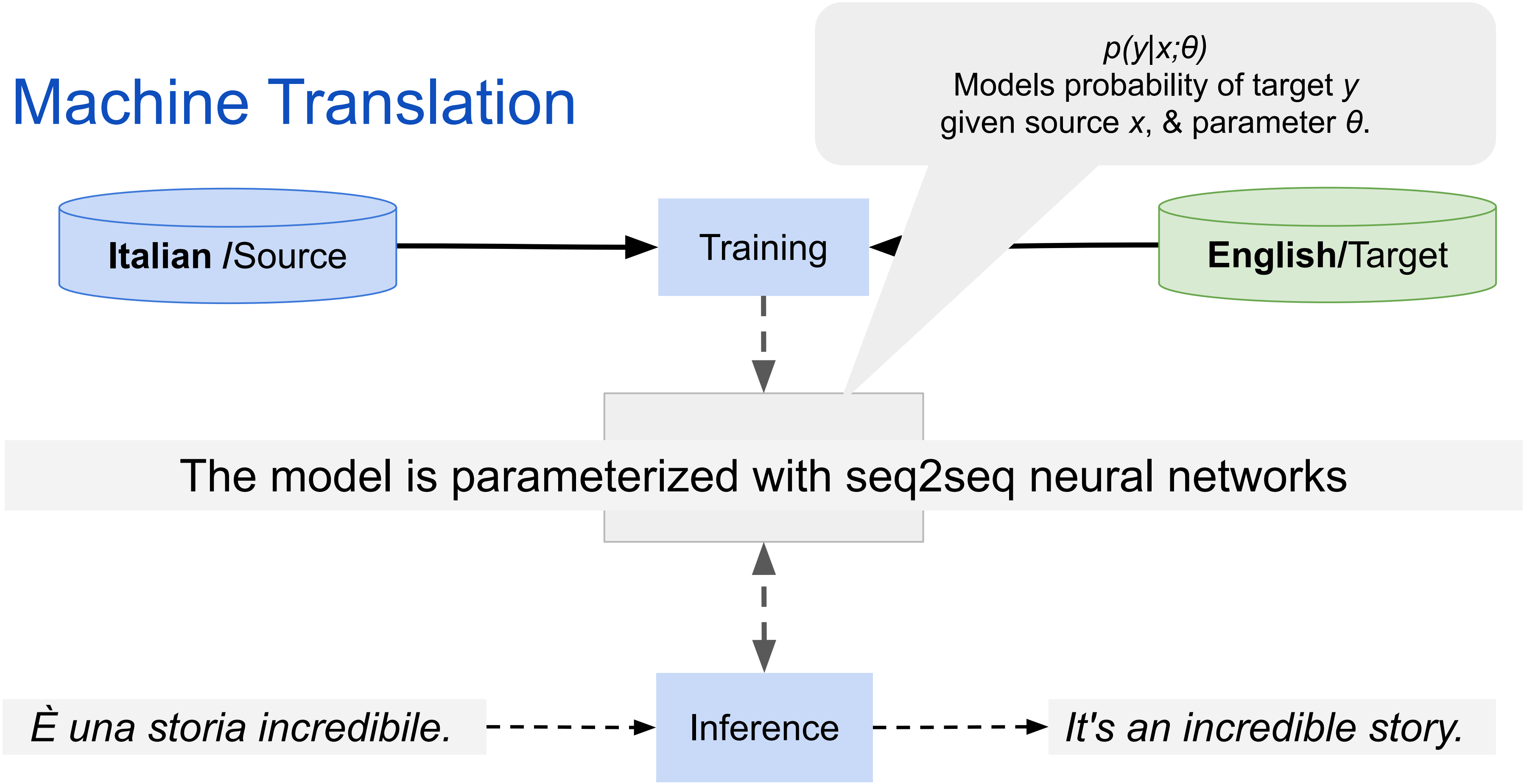
Multilingual
Neural Machine Translation

Task Overview

Machine Translation

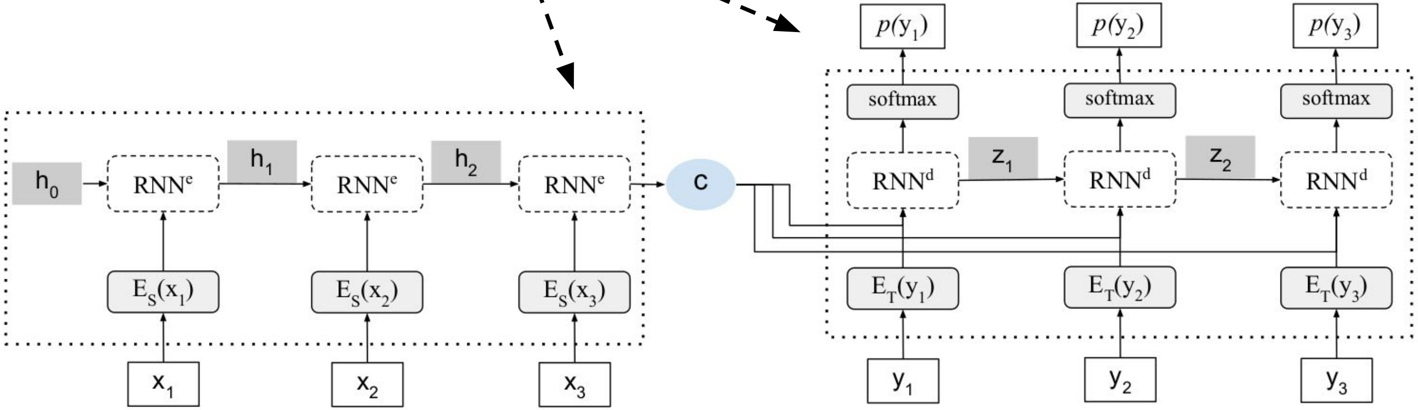


Machine Translation



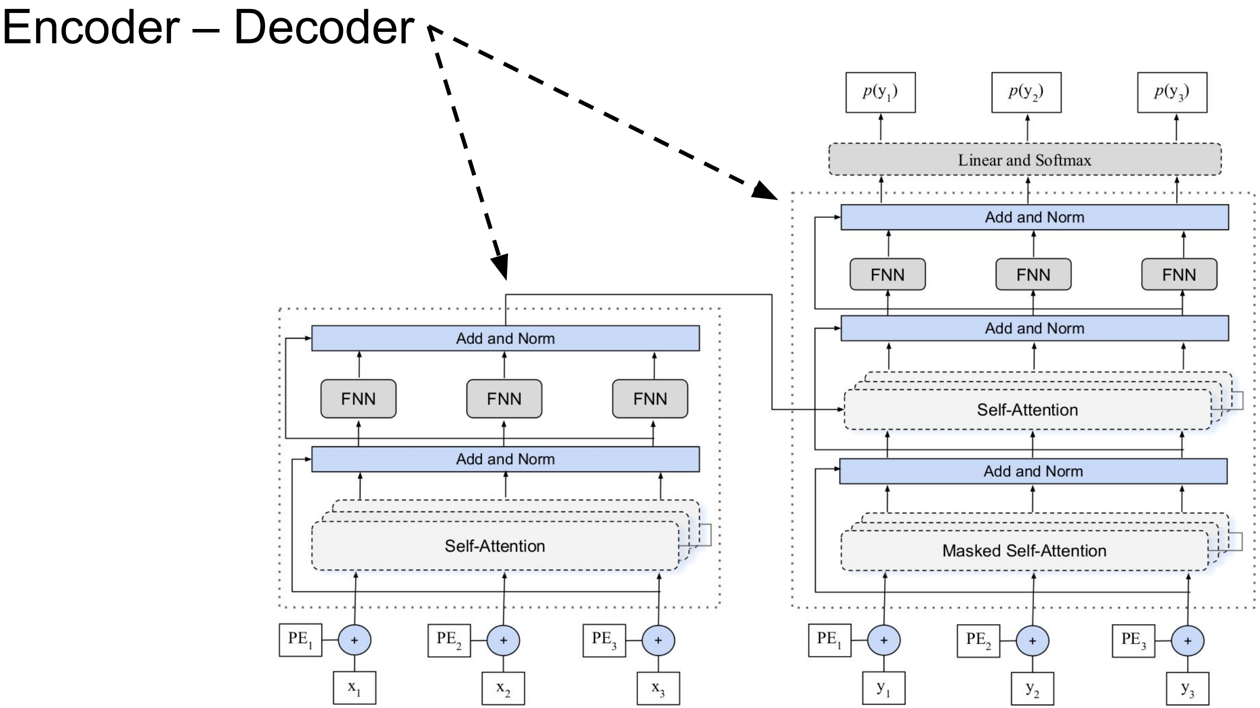
Neural Machine Translation: **Recurrent NN**

Encoder – Decoder



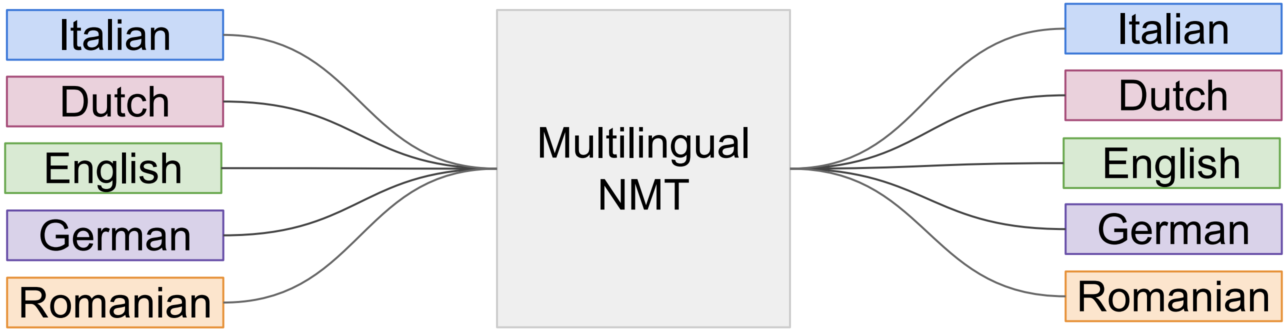
Sutskever et al. 2014

Neural Machine Translation: **Transformer NN**



[Vaswani et al. 2017]

Neural Machine Translation: **Multilingual**

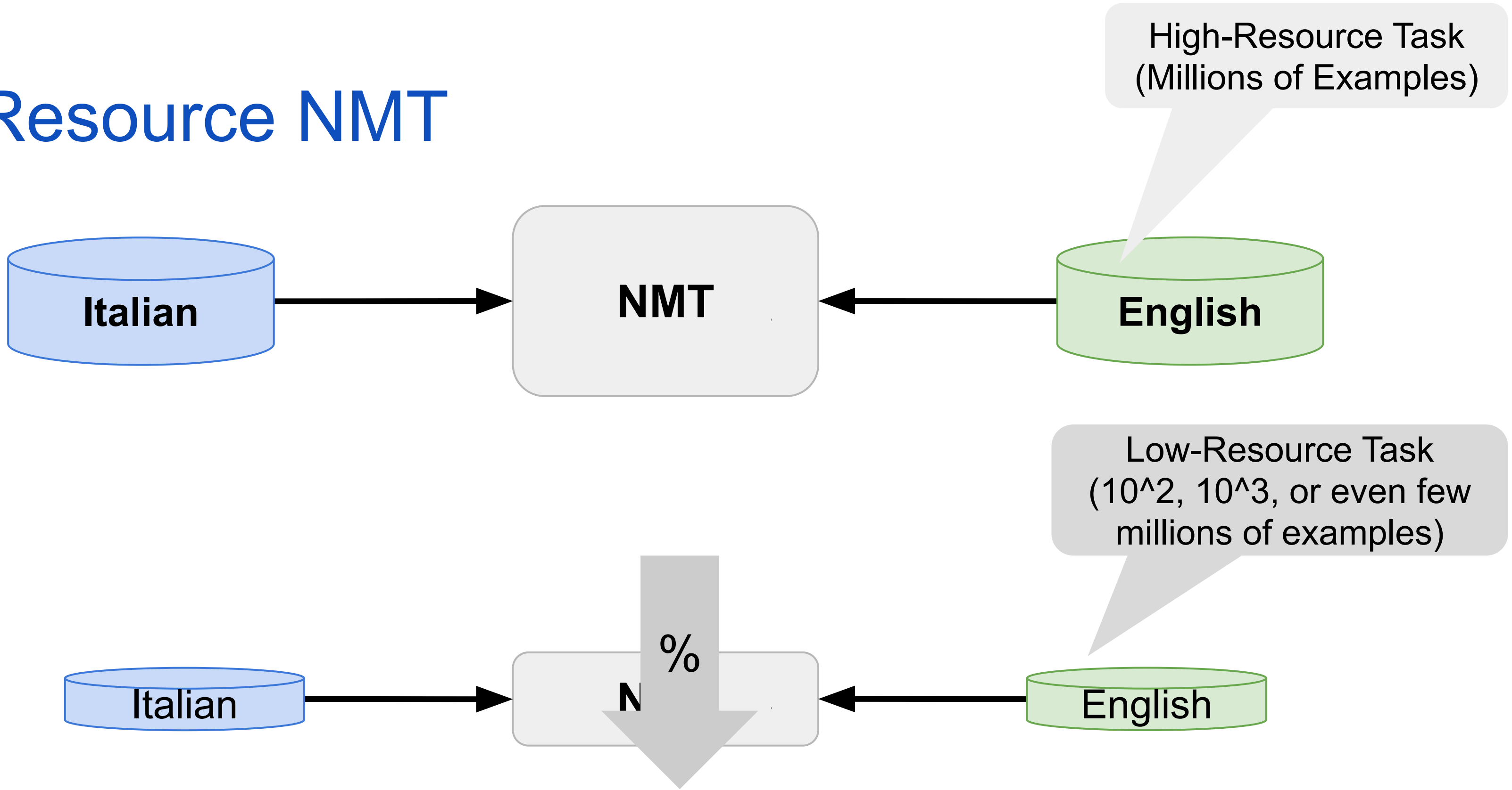


Modeling a single NMT model to translate between multiple languages

Johnson et al. 2016

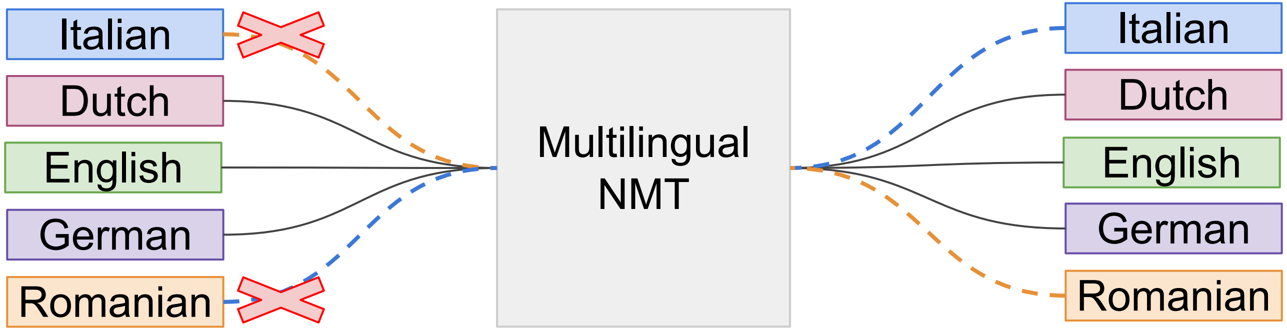
Overview of Tasks

Low-Resource NMT



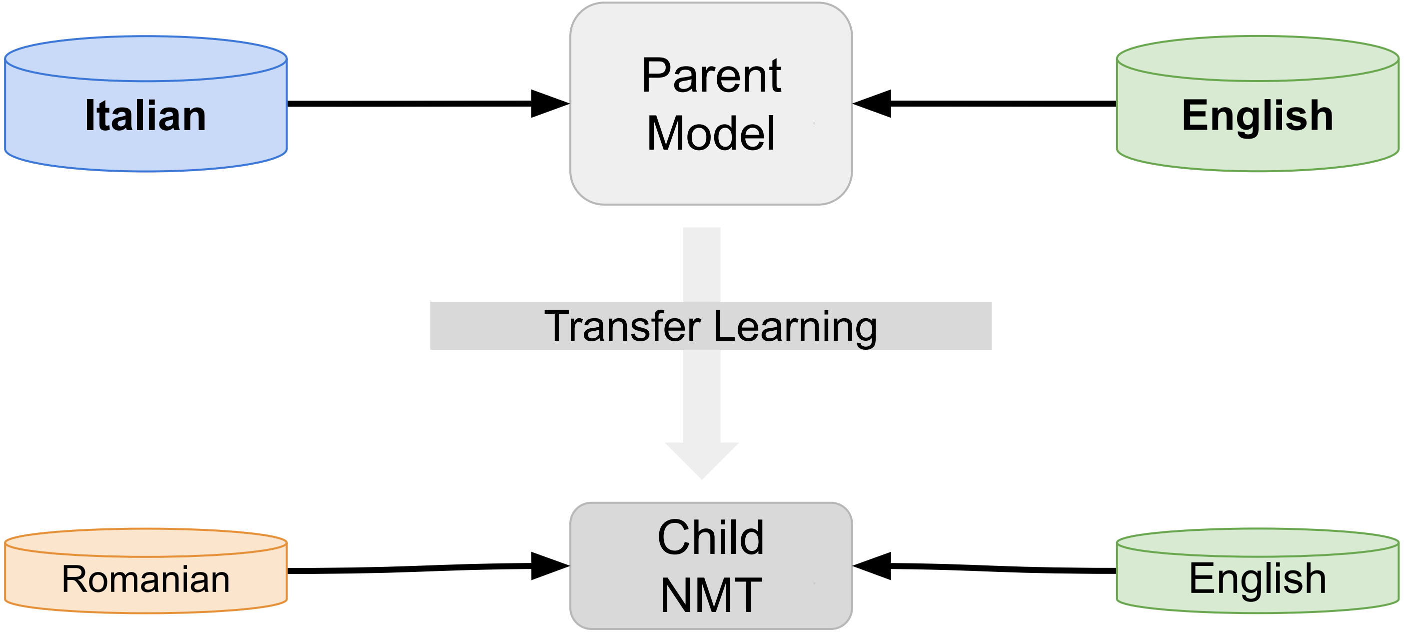
NMT model performance depends on the amount of available examples

Low/Zero-Resource NMT



In the absence of parallel examples, using monolingual & multilingual data

Transfer-Learning in NMT

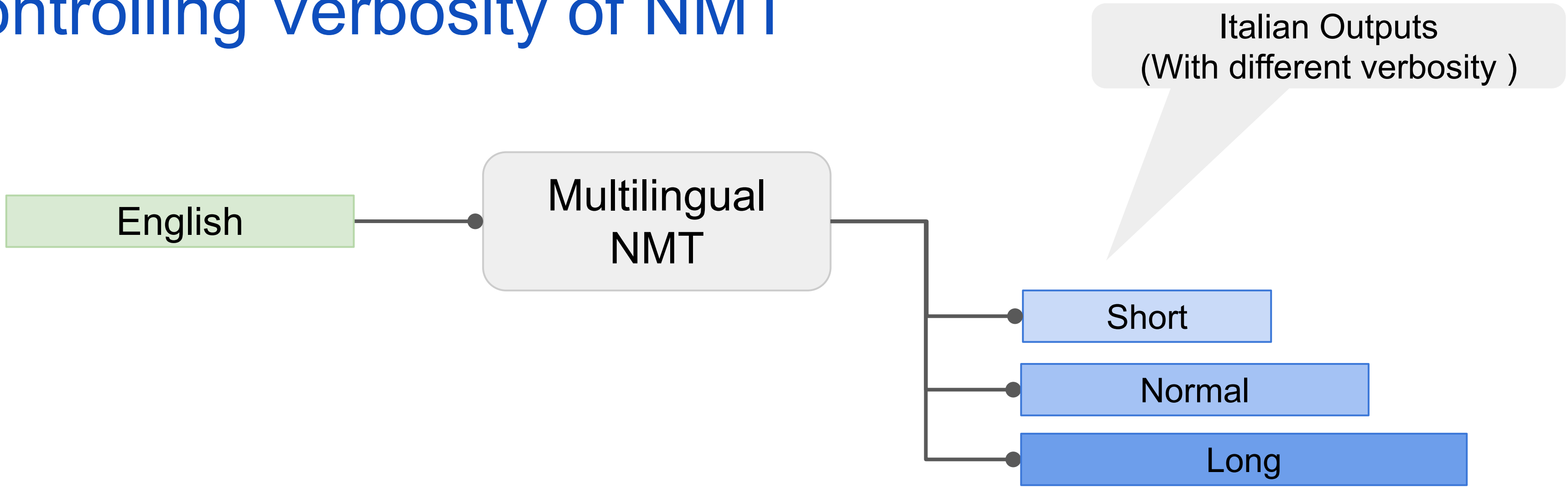


Improving low-resource tasks by leveraging high-resource language pairs

NMT into Language Varieties

NMT can be purposed to translate into several verbosity levels

Controlling Verbosity of NMT

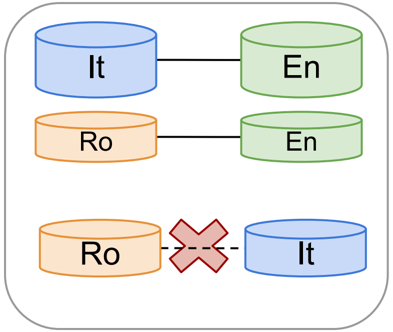


NMT can be purposed to translate into different output length

What connects the tasks together?

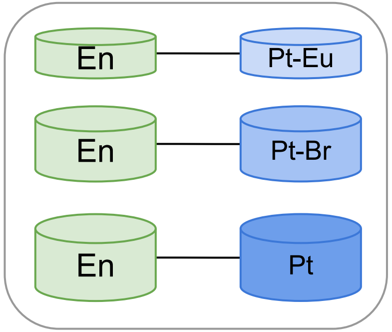
Overview of Tasks: what makes them similar ?

Unbalanced/Unavailable resources across: languages, varieties and styles

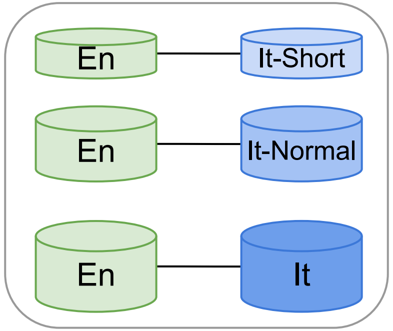


Low-Resource

Zero-Resource



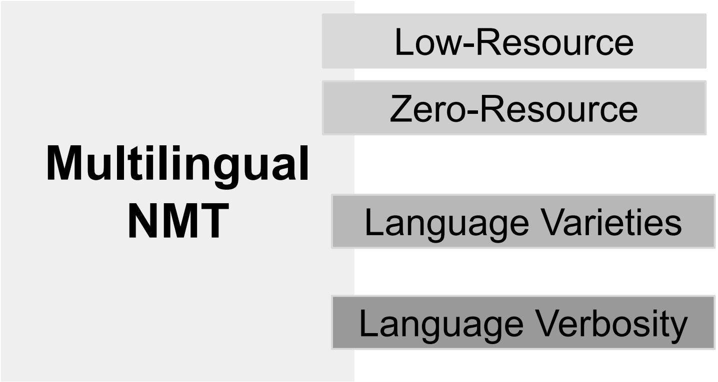
Language Varieties



Language Verbosity

Overview of Tasks: what makes them similar ?

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



Problem Statements

Challenges and Motivations

Low-Resource NMT/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

Mainstream 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

Mainstream 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

We Ask (2016):

Does multilingual NMT improve in low-resource conditions ?

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

Lakew et al., Clic-It, *2017.*

Lakew et al., *IWSLT, 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.

We Ask (2018, 2019):

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

Can we grow NMT into unseen languages directions ?

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

Lakew et al., *IWSLT, 2018.*

Lakew et al., *IWSLT, 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.

We Ask (2018):

Does modeling multiple varieties in a single model is achievable ?

Can we further improve over the baseline single LV models ?

How to handle majority of LV unlabeled parallel data ?

Lakew et al., *EMNLP-WMT, 2018.*

Controlling the Verbosity of NMT

Previous Work:

2016

NMT, MNMT, Summarization

Sennrich et al.,: politeness in NMT
Johnson et al., Ha et al.,: Multilingual NMT
Kikuchi et al.,: summarization

2017

Summarization

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

2018

Customized NMT

Lakew et al.,: NMT into language varieties
Michel & Neubig.: personalized NMT
Niu et al.,: translation styles

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
Johnson et al., Ha et al.,: Multilingual NMT
Kikuchi et al.,: summarization

2017

Summarization

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

2018

Customized NMT

Lakew et al.,: NMT into language varieties
Michel & Neubig.: personalized NMT
Niu et al.,: translation styles

2019

Summarization, Pos. encoding

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

We Ask (2018):

Does modeling multiple length/verbosity level of NMT achievable ?

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.*

Thesis Contributions

Methods, Experiments, Results and Findings

Zero-Shot NMT Modeling

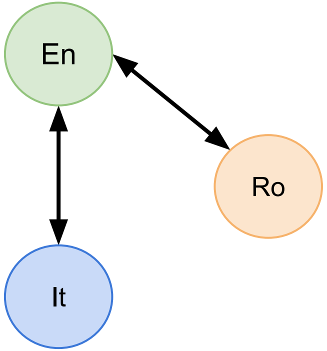
Dynamic Transfer Learning

NMT into Language Varieties

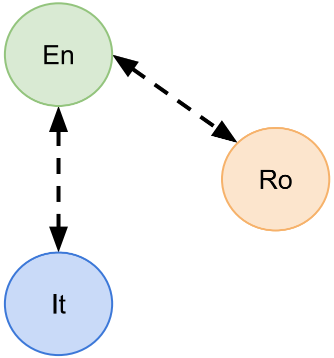
Controlling NMT Verbosity

FONDAZIONE BRUNO KESSLER

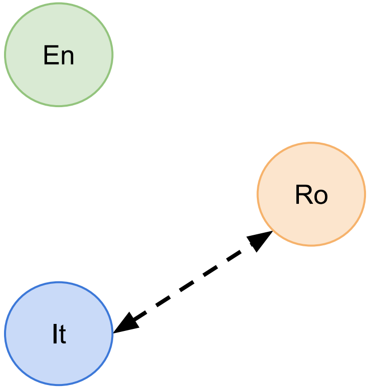
Zero-Shot Translation



Multilingual Training



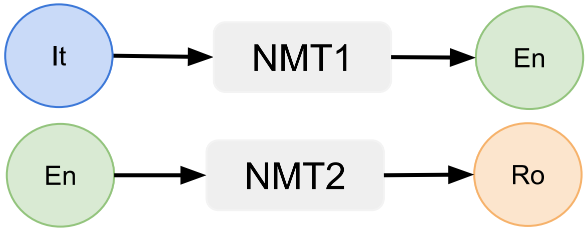
Inference



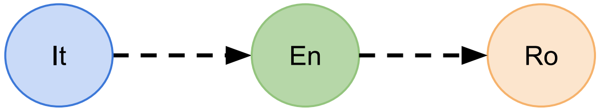
Zero-Shot Inference

Zero-Shot Translation - among the advantages of Multilingual NMT

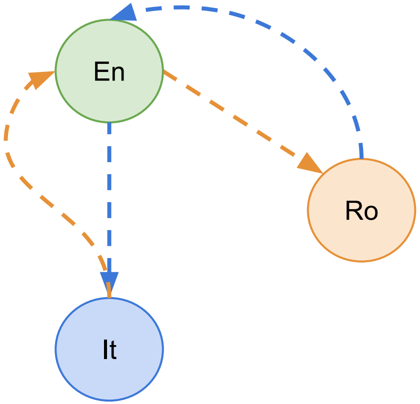
Pivoting Translation as Alternative



NMT Training



Pivot Inference



Pivot (Multilingual) Inference

Our Research Questions

Does multilingual NMT improve low/zero-resource translation ?

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

Multilingual NMT in Low-Resource Condition

Does multilingual modeling improve the pairs with parallel examples ?

Does zero-shot translation work as expected for zero-resource pairs ?

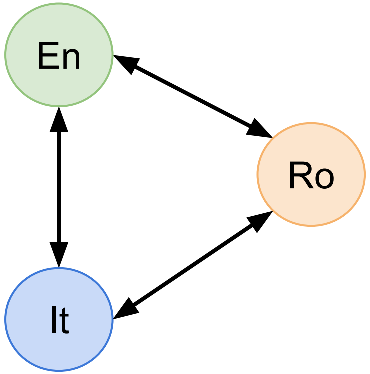
Is pivoting translation with multilingual NMT an effective alternative ?

Multilingual NMT in Low-Resource Condition

Language Direction	Training Data Size
En - De	197,489
En - It	221,688
En - Nl	231,669
En - Ro	211,508
It - Ro	209,668

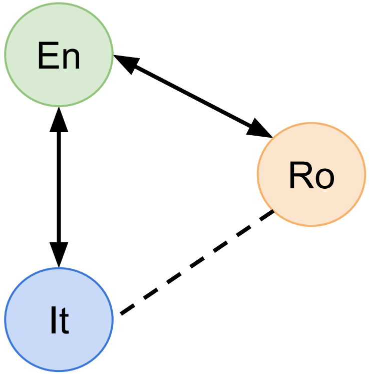
Experimental setting as a low-resource condition (~ 200k examples).

Multilingual NMT: Model Types



Training on 6 pairs

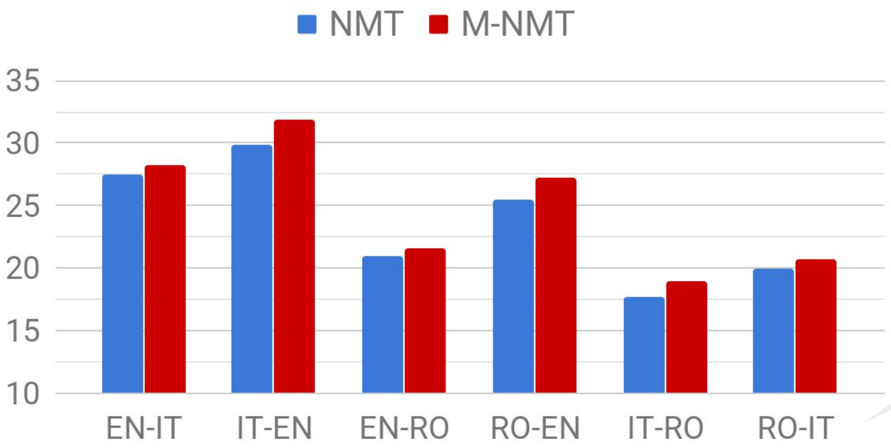
6 standard Inferences



Training on 4 pairs

4 standard & 2 ZT/Pivot Inferences

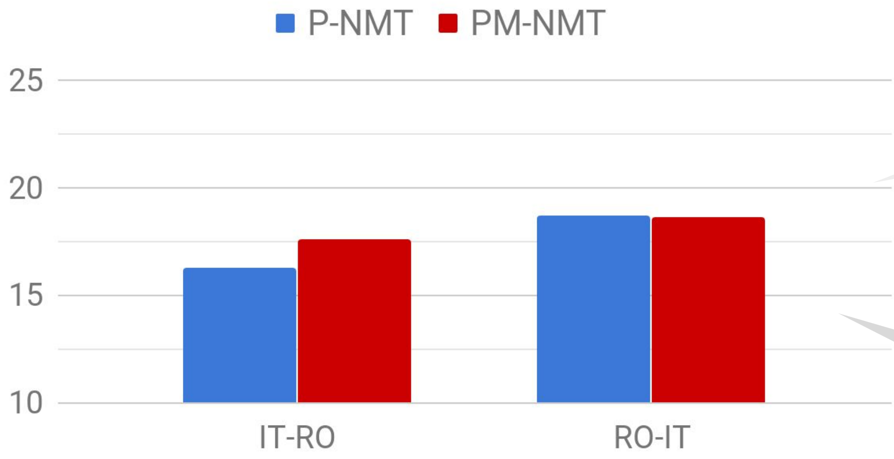
Multilingual NMT: Improves over single pair NMT’s



Results on test-2017

Multilingual NMT
outperforms
six single pair NMT’s

Multilingual NMT: Fails at direct zero-shot translation



Pivoting results on test-2017

Zero-shot translation results in < 1 BLEU in both directions.

Instead, multilingual pivoting becomes a better alternative

Takeaway/Findings

Confirmations:

- Better performance against single pair NMT
- Zero-shot (implicit bridging) is weaker than Pivoting (explicit bridging) (confirming both Johnson et al., and Ha et al.,)

Findings:

- Pivoting is better using a multilingual model than traversing two NMT's
- Zero-shot is way poorer in a low-resource multilingual setting

Lakew et al., Clic-It, *2017.*

Zero-Shot NMT Modeling

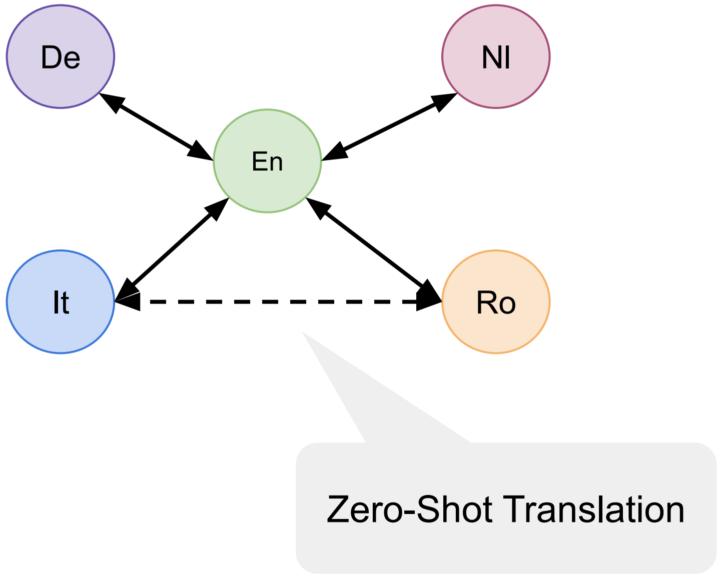
Aims at improving zero-shot translation in a multilingual model.

Why ?

- Zero-shot is (was) just one time inference
- Meaning, translation only, no learning!
- Multilingual model for LRL pairs is still weak
- *Resulting even weaker zero-shot translations*

Zero-Shot NMT Modeling

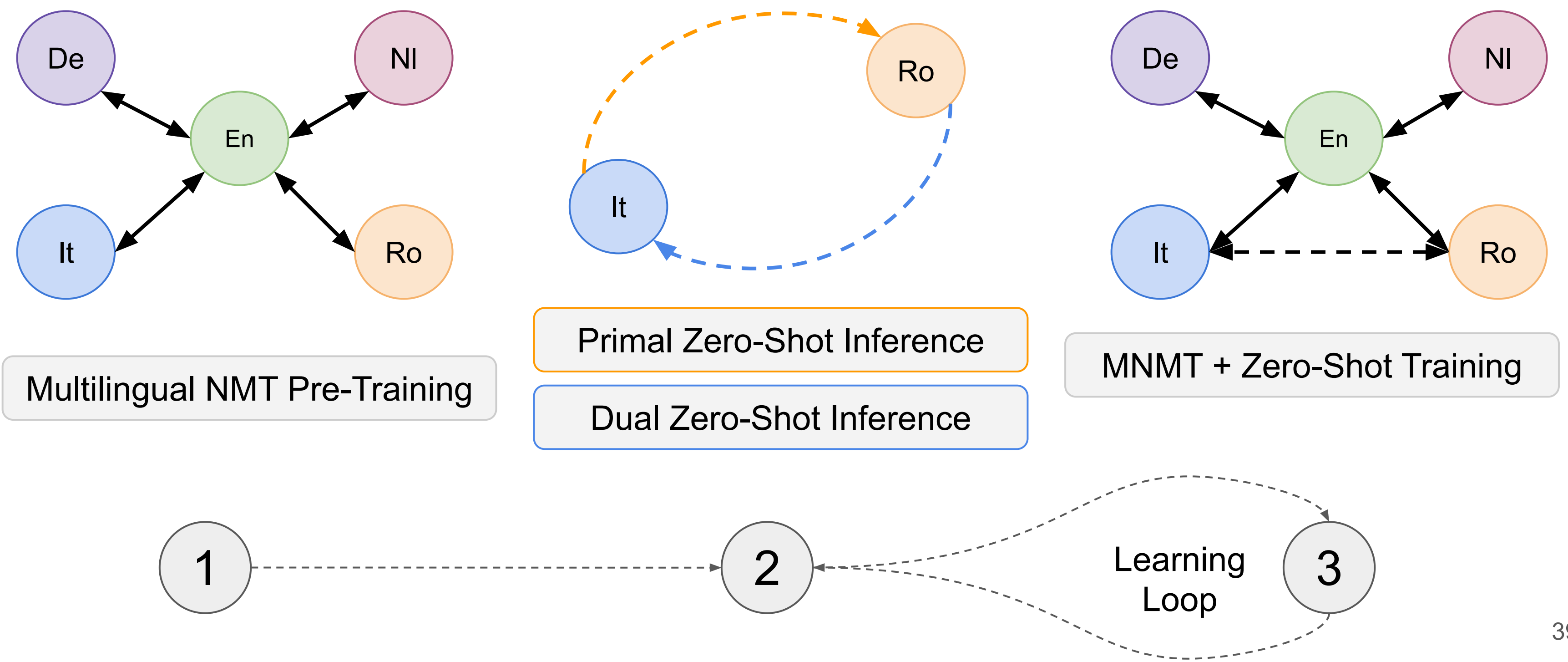
Available Resource and ZST Task



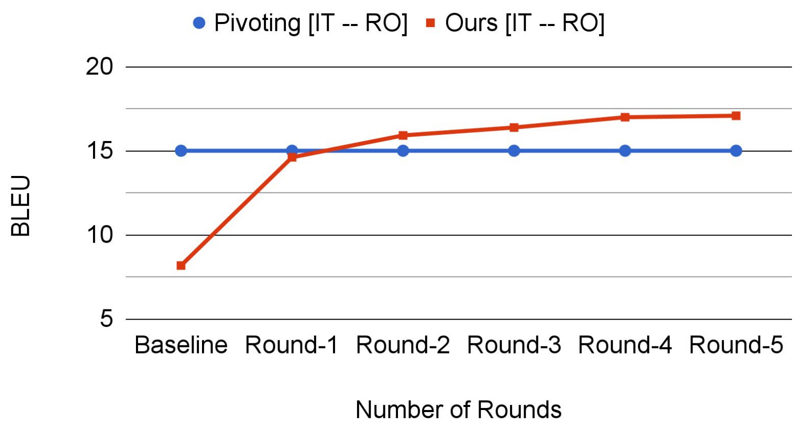
Zero-Shot Learning Principles

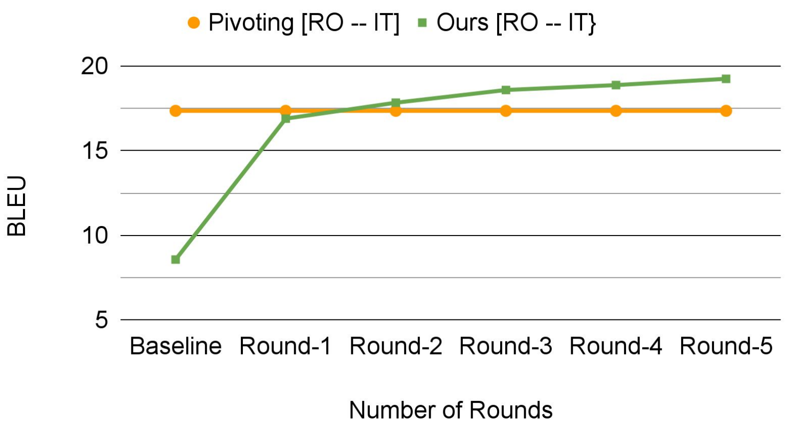
- Leverage monolingual data
- Perform dual back-translation
- Self-Learning using Iterative Data Augmentation & Learning with supervised tasks.

Zero-Shot NMT Modeling: Three Stages



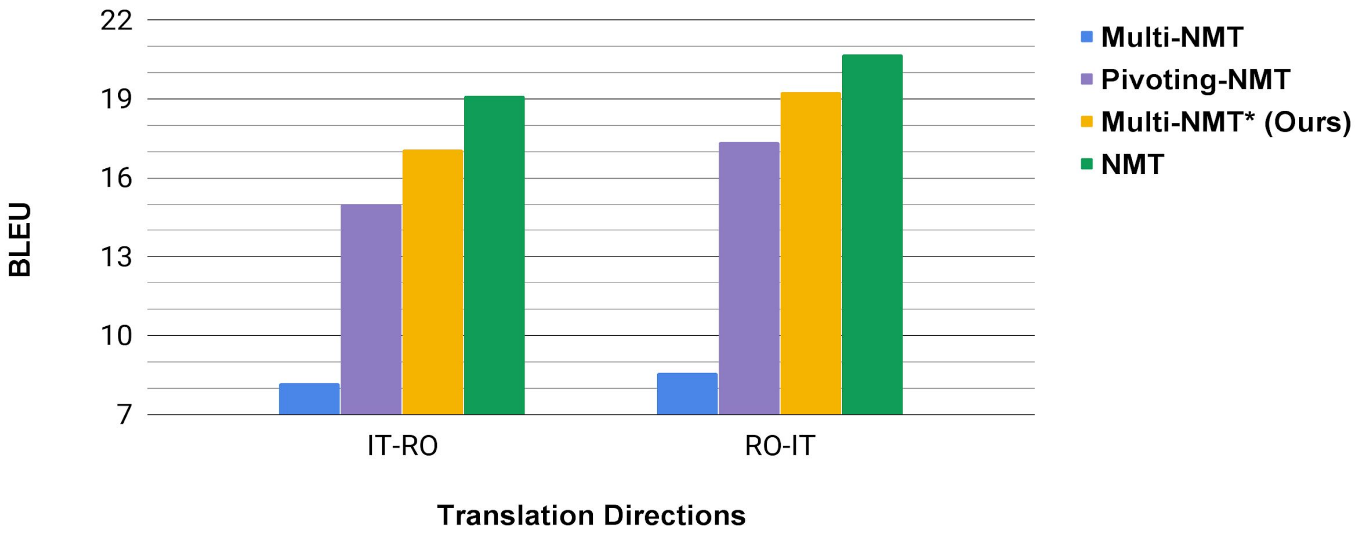
Results





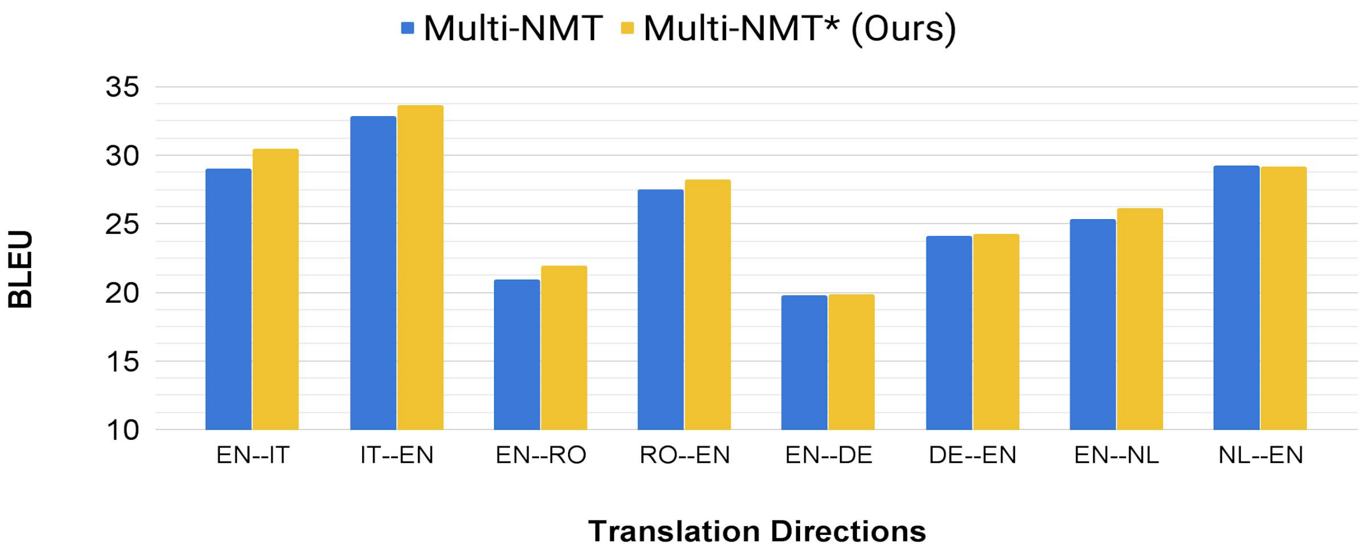
Results of the Italian <> Romanian zero-shot directions on test2017

Results in comparison with Pivoting



Zero-Shot NMT modeling outperformed the baseline **Multi-NMT** and the **Pivoting** mechanism on *test2017*

Results: for the non Zero-shot Directions



Our proposed approach slightly improves the baseline **Multi-NMT** on *test2017*

Examples

	Zero-shot: Italian > Romanian
Source	... che rafforza la corruzione, l'evasione fiscale, la povertà, l'instabilità.
Pivot	… poarta de bază, evazia fiscală, sărăcia, instabilitatea.
Multi-NMT	... restrânge corrupția, fiscale de evasion, poverty, instabilitate.
Multi-NMT*	... care rafinează corupția, evasarea fiscală, sărăcia, instabilitatea.
Reference	... care protejează corupţia, evaziunea fiscală, sărăcia şi instabilitatea.

Takeaway

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

Lakew et al., *IWSLT, 2017.*

Thesis Contributions

Zero-Shot NMT Modeling

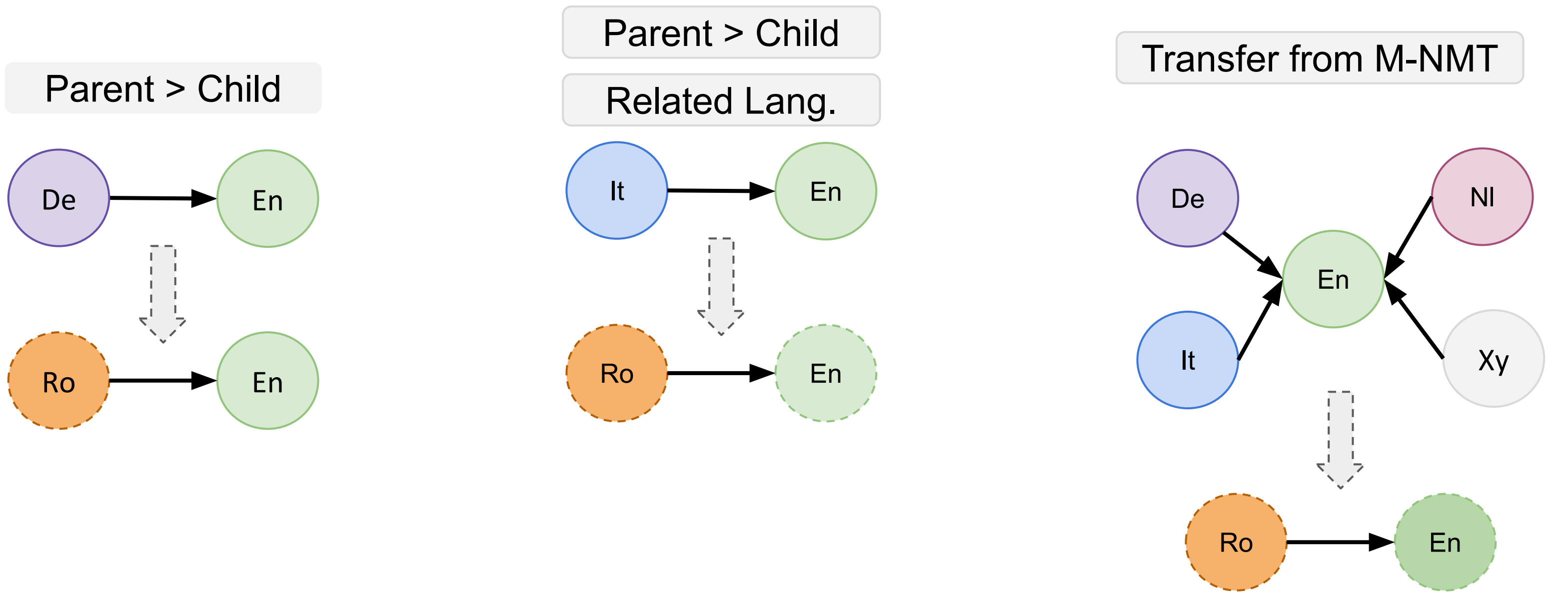
Dynamic Transfer Learning

NMT into Language Varieties

Controlling NMT Verbosity

FONDAZIONE BRUNO KESSLER

Transfer Learning



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 grow NMT into unseen languages directions ?

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

Dynamic Transfer Learning

Aim at maximizing the positive transfer-learning from a Parent to the Child model.

Our Transfer-Learning Principle

- Tailor the parent model parameters (vocabulary, embedding) to the child model languages.

Dynamic Transfer Learning: Two Approaches

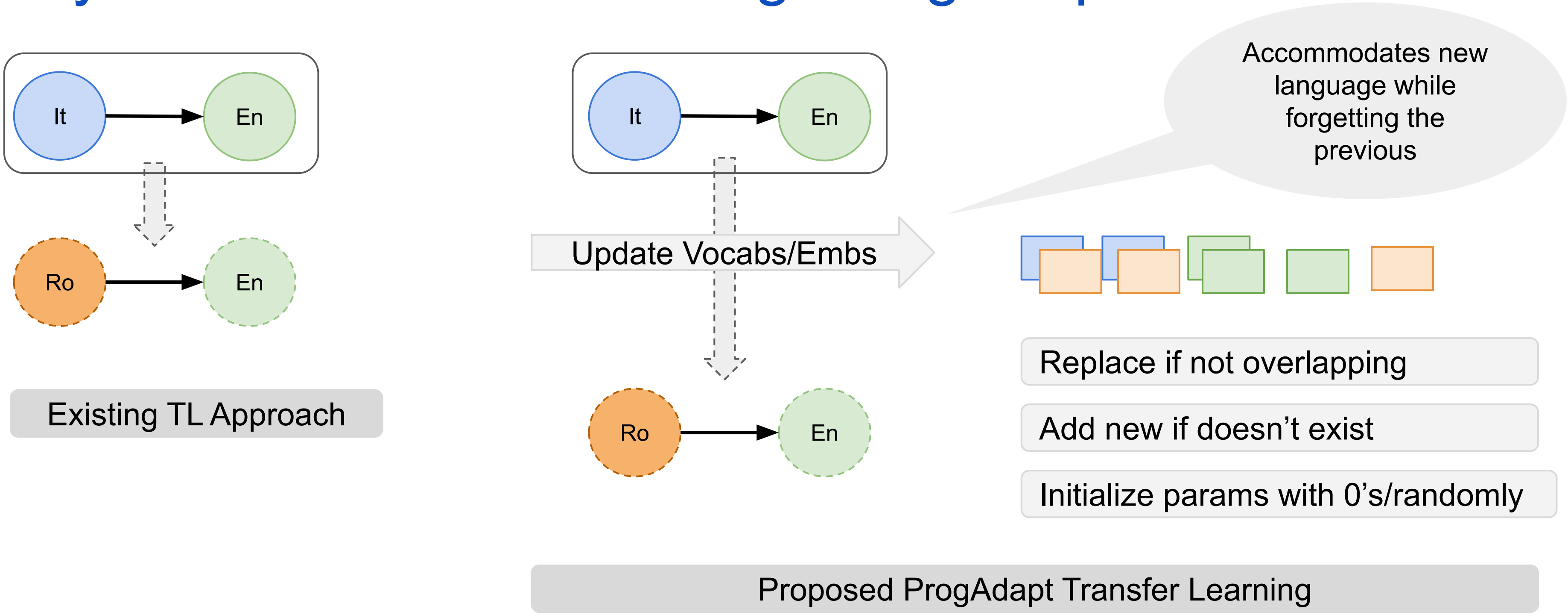
Progressive Adapt (ProgAdapt)

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

Progressive Grow (ProgGrow)

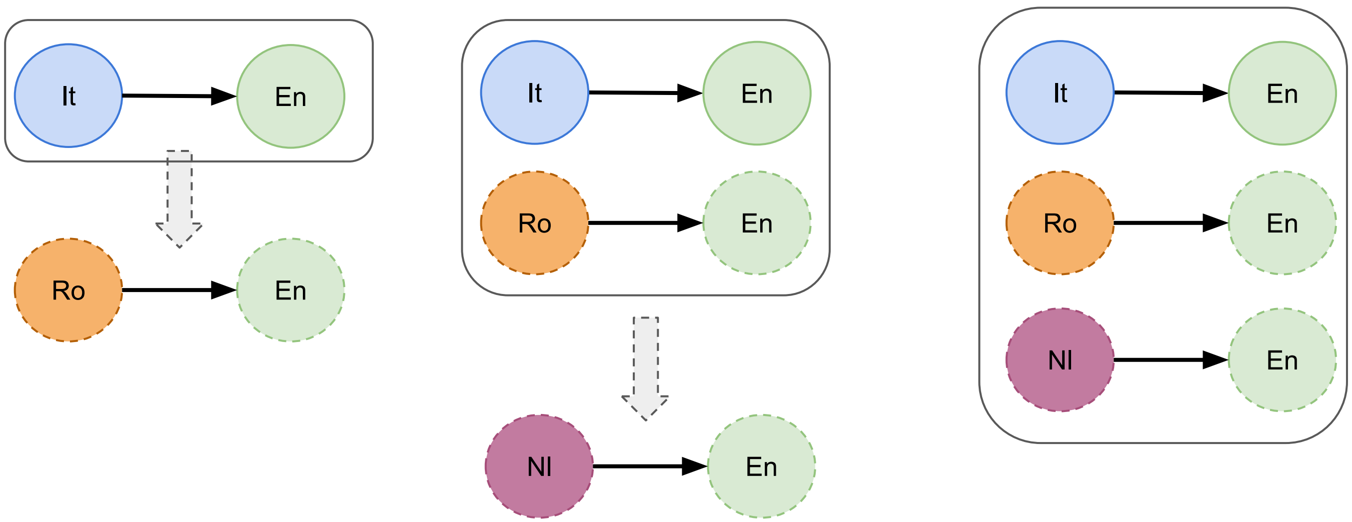
- Accommodate new language pairs when data becomes available

Dynamic Transfer Learning: ProgAdapt



Dynamic Transfer Learning: ProGrow

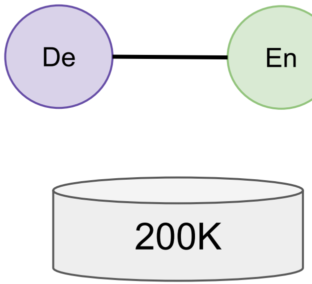
Similar approach as progAdapt, except keeping previous language pairs.



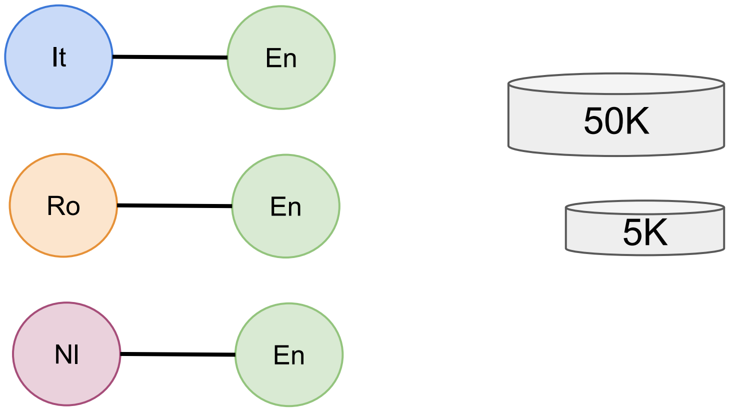
Proposed ProgGrow Transfer Learning in 3 stages.

Experimental Settings

Parent Language pairs / Model



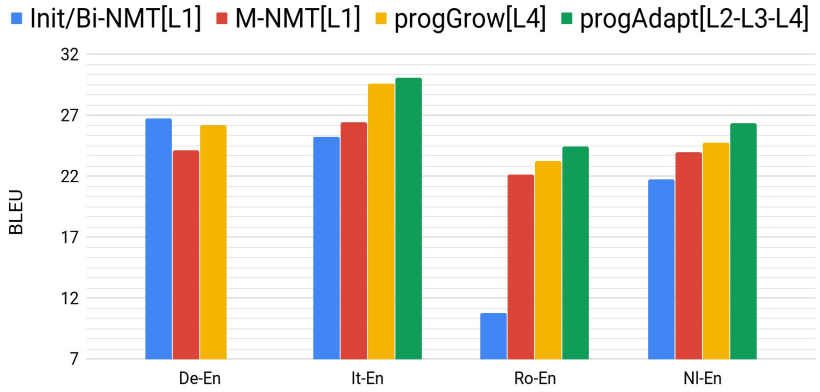
Child Language Pairs / Two Settings



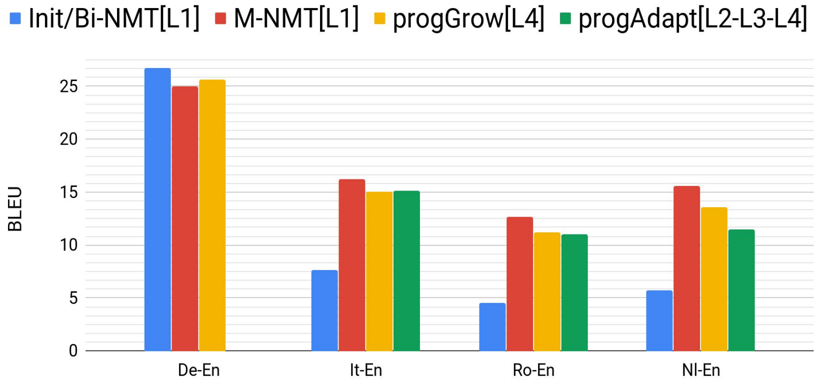
Results

Outperform both single-pair NMT and M-NMT approaches

Outperform single-pair NMT



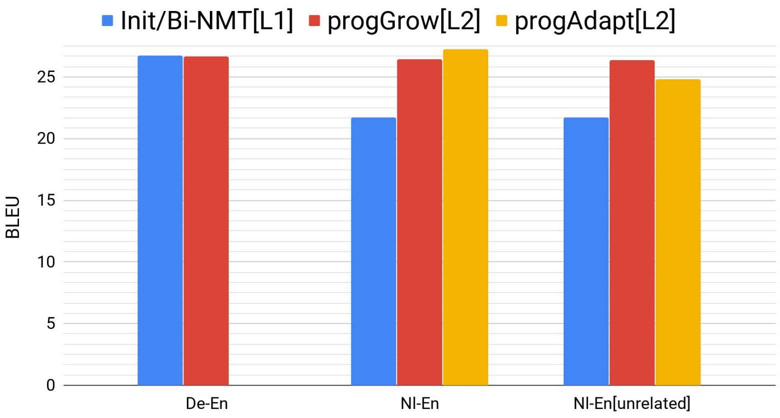
Low-Resource Results



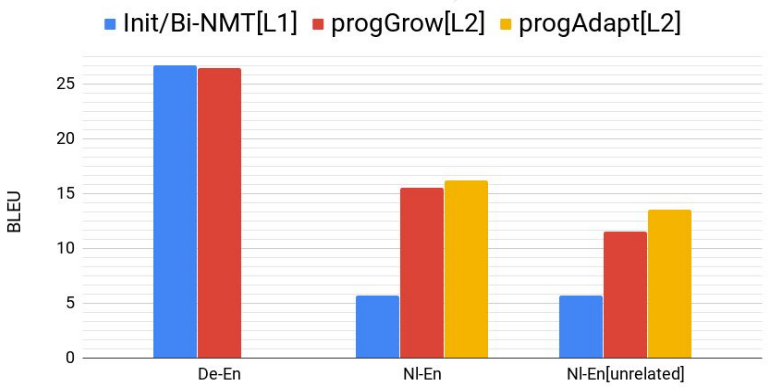
Extremely Low-Resource Results

Results: Role of language relatedness

Large improvement
if Parent model pair
is related to Child

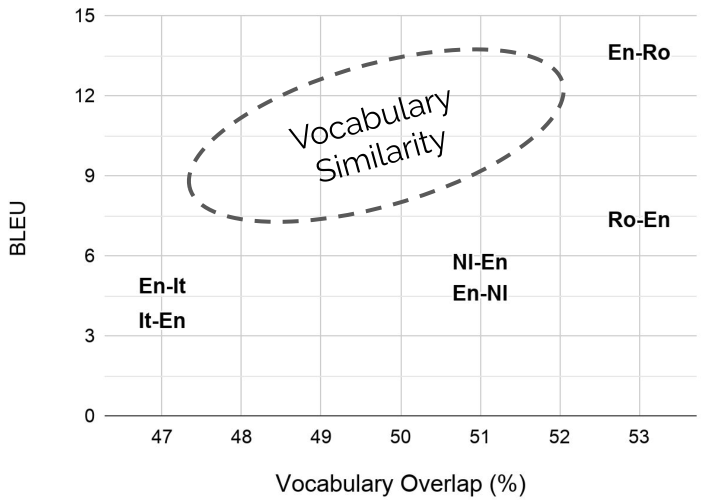


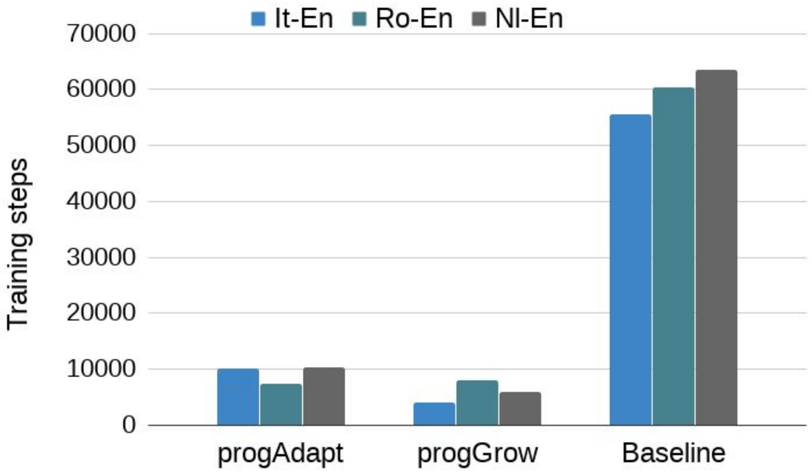
Low-Resource Results



Extremely Low-Resource Results

Results: Vocabulary Overlap & Time for TL





Dynamic Transfer Learning

Two Additional Proposals

Multilingual Model

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

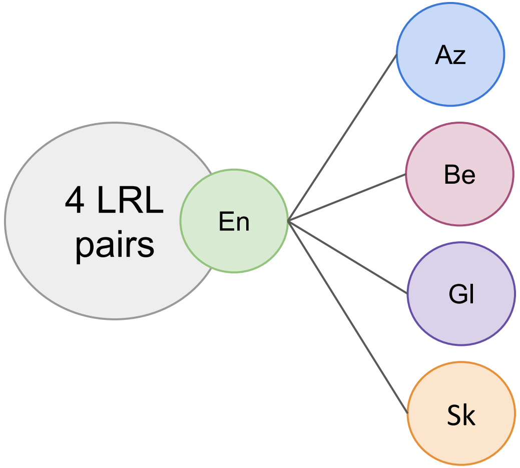
Data selection for TL

- We train a LM on the test language (child) data to select relevant data for the transfer-learning stage.

Experimental Settings

54 Languages paired with English

Used to pre-trained 108 directions multilingual model



5.9k

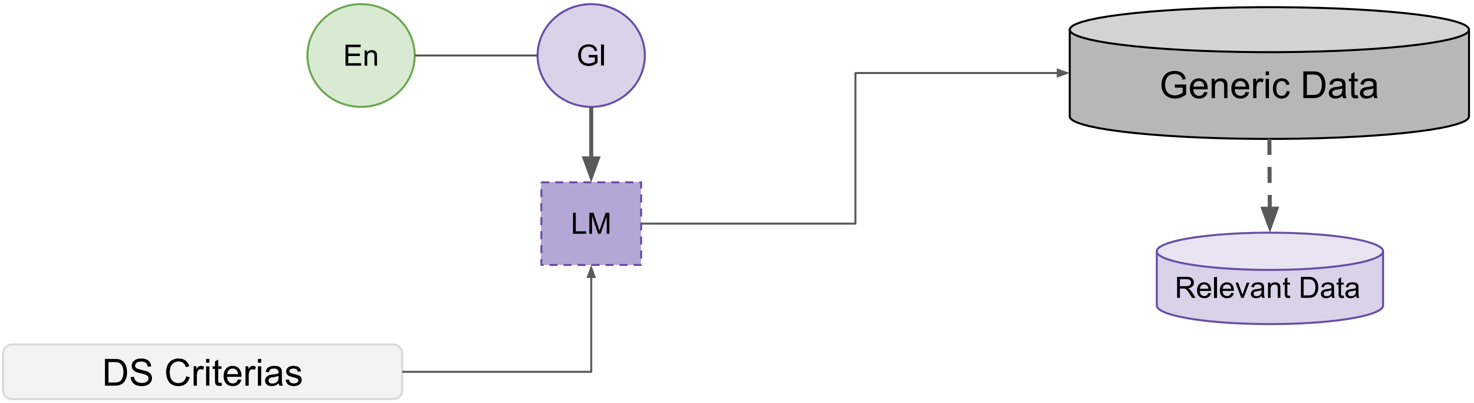
4.5k

10k

61.5k

Low-Resource Language Pairs

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

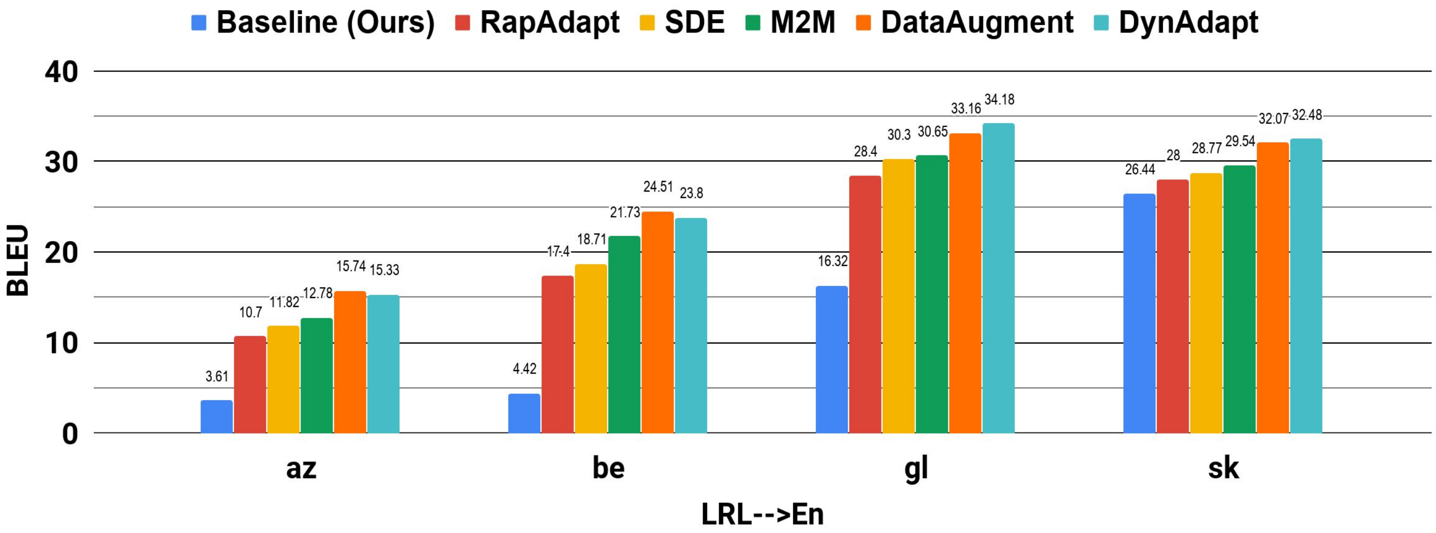
Select-rand: randomly sample from HRLs related/family to the LRL

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

*except for Select-fam, the rest approaches pick an equal proportion of selected data.

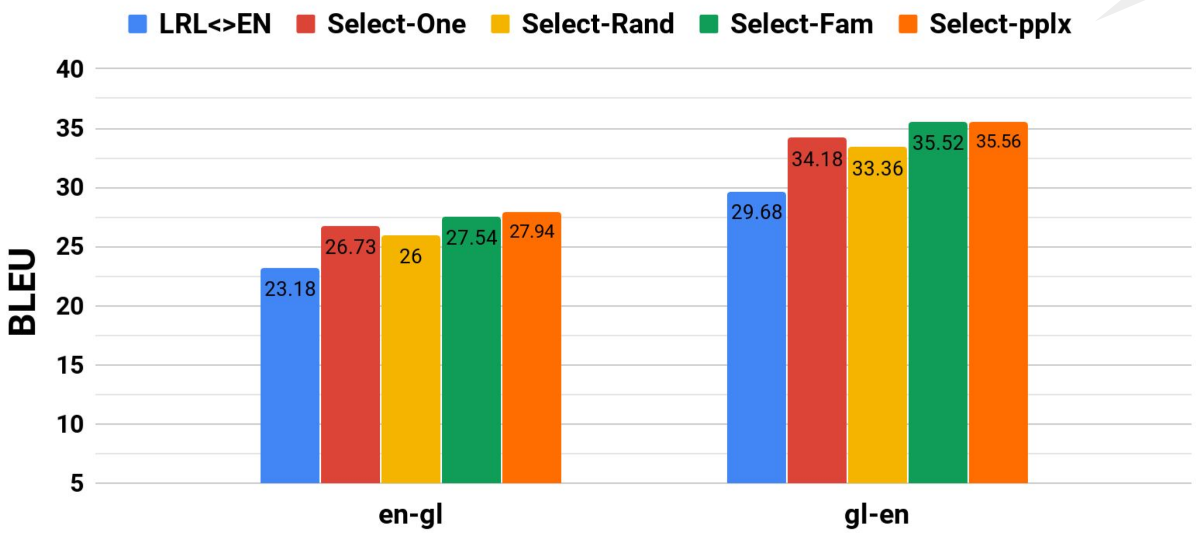
Results

Our DynAdapt
using Select-One
strategy



Results

Comparison of the DS strategies with DynAdapt.
Winner: Select-pplx



Takeaway

- Utilizing a universal pre-trained multilingual model improves TL for LRLs.
- Relevant data-selection further improves dynamic adaptation & cheaper to acquire.
- With up to + 17.0 BLEU improvements over baselines, our approach outperformed related work on the same test sets.
- Our DynAdapt + Data Selection is SOTA on this benchmarks without further data augmentation.

Lakew et al., *IWSLT, 2018.*

Lakew et al., *IWSLT, 2019.*

Thesis Contributions

Zero-Shot NMT Modeling

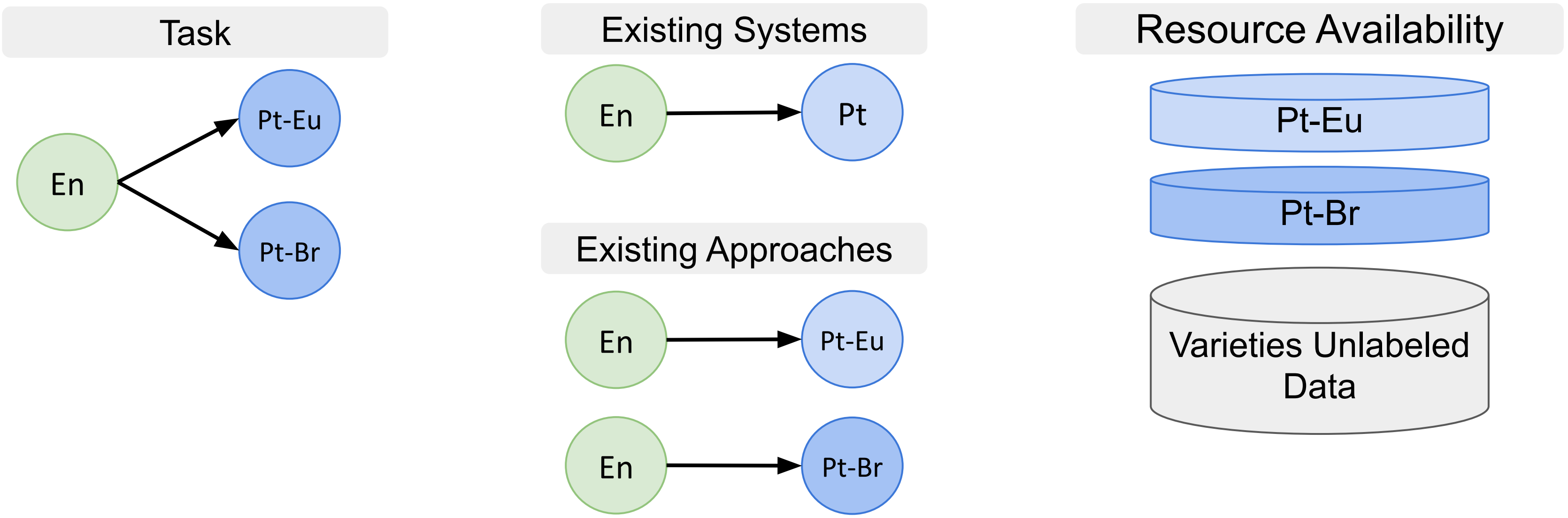
Dynamic Transfer Learning

NMT into Language Varieties

Controlling NMT Verbosity

FONDAZIONE BRUNO KESSLER

NMT into Language Varieties: A scenario



A large scale unlabeled data leads to poor specific language varieties models

Research Questions

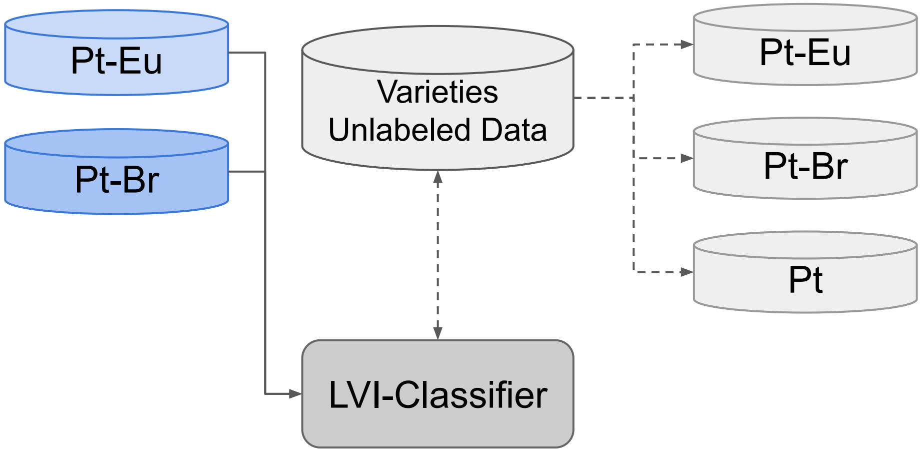
Does modeling multiple varieties in a single model achievable ?

Can we further improve over the baseline single LV models ?

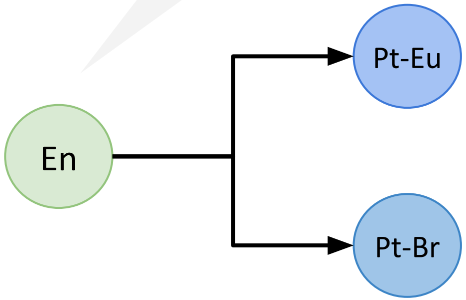
How to handle majority of LV unlabeled parallel data ?

Modeling NMT into Language Varieties

We use a similar principles as in multilingual NMT

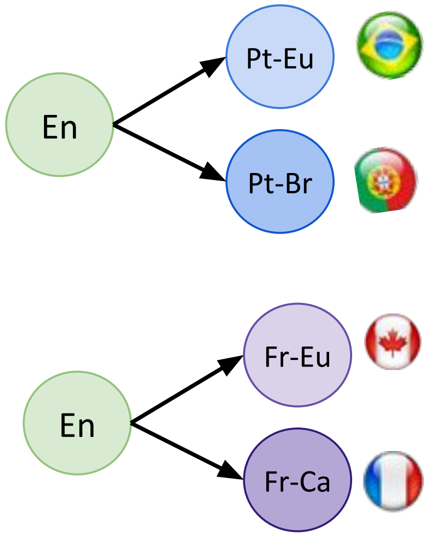


Offline labeling using a LVI

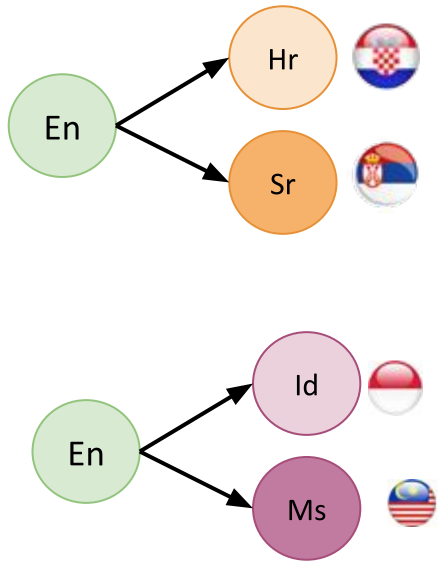


A single LV-NMT model Training

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 varieties labeled data

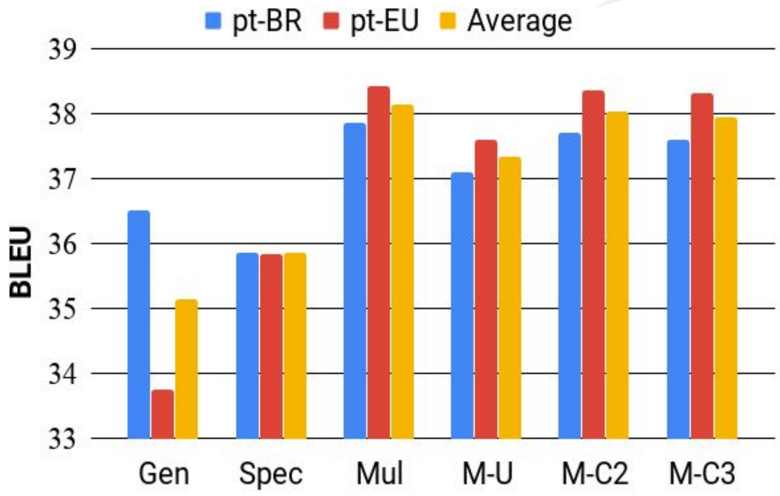
M-U: semi-supervised trained with the union of both labeled and unlabeled data

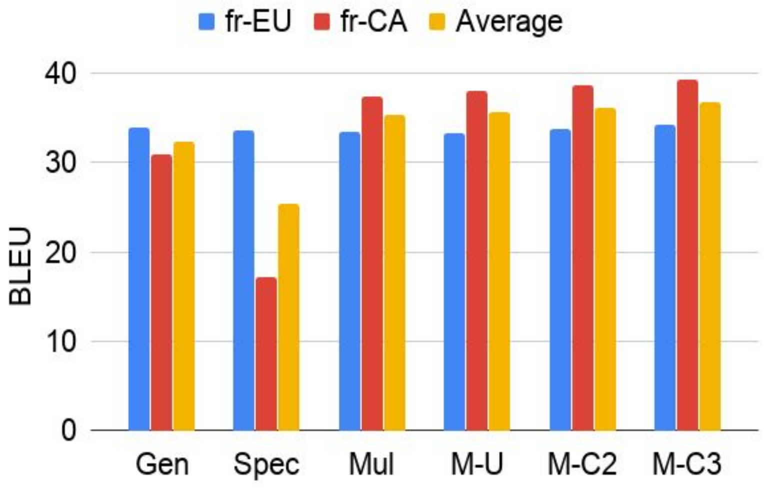
M-C2: semi-supervised training using LVI to map the unlabeled segments to variety classes

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

Results

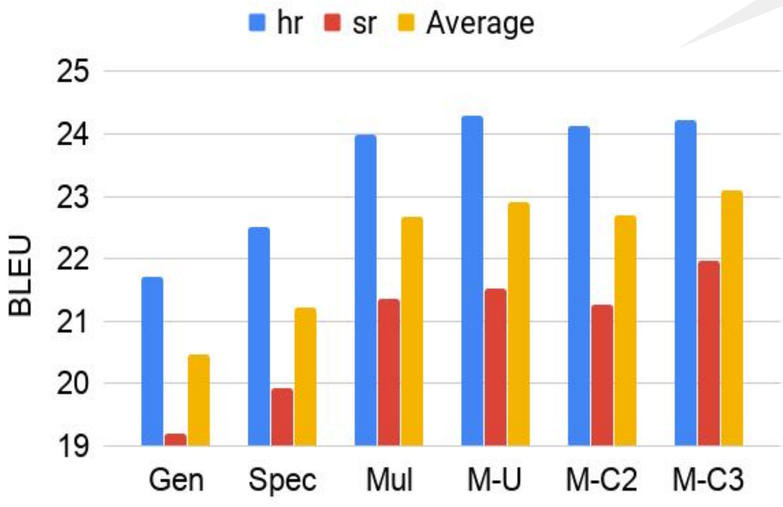
Mul (supervised) shows the largest improvement, with comparable performance from semi-supervised (M-C2/3)

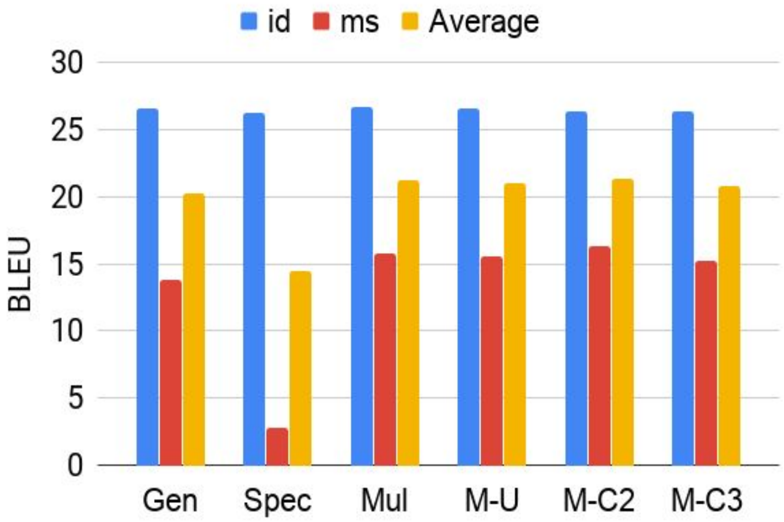




Results

Semi-supervised approaches are better competitive with Mul





Examples

Translation comparison of our large scale LV-NMT model for the Eu/Br portugues against Google Translate

English (source)	I'm going to the gym before breakfast. No, I'm not going to the gym.
pt (Google Translate)	Eu estou indo para a academia antes do café da manhã. Não, eu não vou ao ginásio.
pt-BR (M-C2)	Eu vou á academia antes do café da manhã. Não, eu não vou à academia.
pt-EU (M-C2)	Vou para o ginásio antes do pequeno-almoço. Não, não vou para o ginàsio.
pt-BR (M-C2_L)	Vou à academia antes do café da manhã. Não, não vou à academia.
pt-PT (M-C2_L)	Vou ao ginásio antes do pequeno-almoço. Não, não vou ao ginásio.

Underlined English terms are shown both with ***pt-BR*** & ***pt-EU*** translation variants.

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 similarly to its supervised (Mul) version.

- The approach keeps resource together for a within a single model transfer-learning.

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

Lakew et al., *EMNLP-WMT, 2018.*

Thesis Contributions

Zero-Shot NMT Modeling

Dynamic Transfer Learning

NMT into Language Varieties

Controlling NMT Verbosity

FONDAZIONE BRUNO KESSLER

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.---

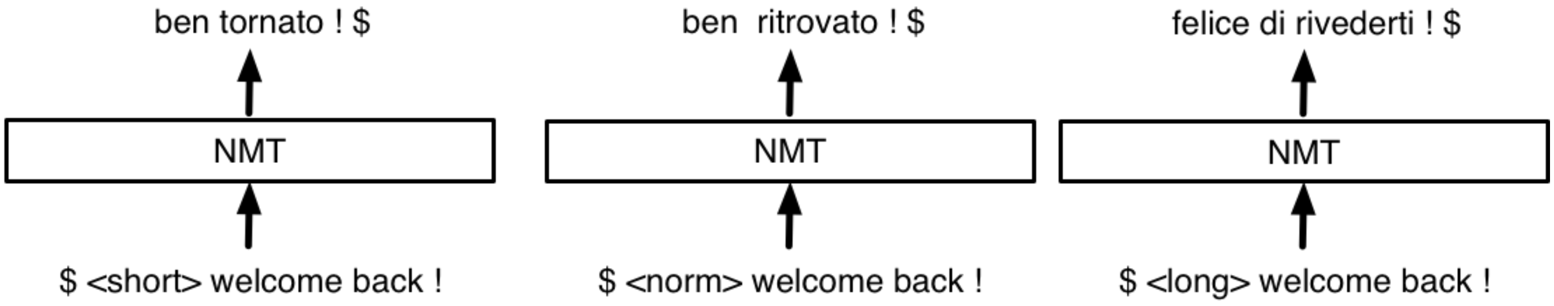
Our Research Questions

Does modeling multiple length/verbosity level of NMT achievable ?

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

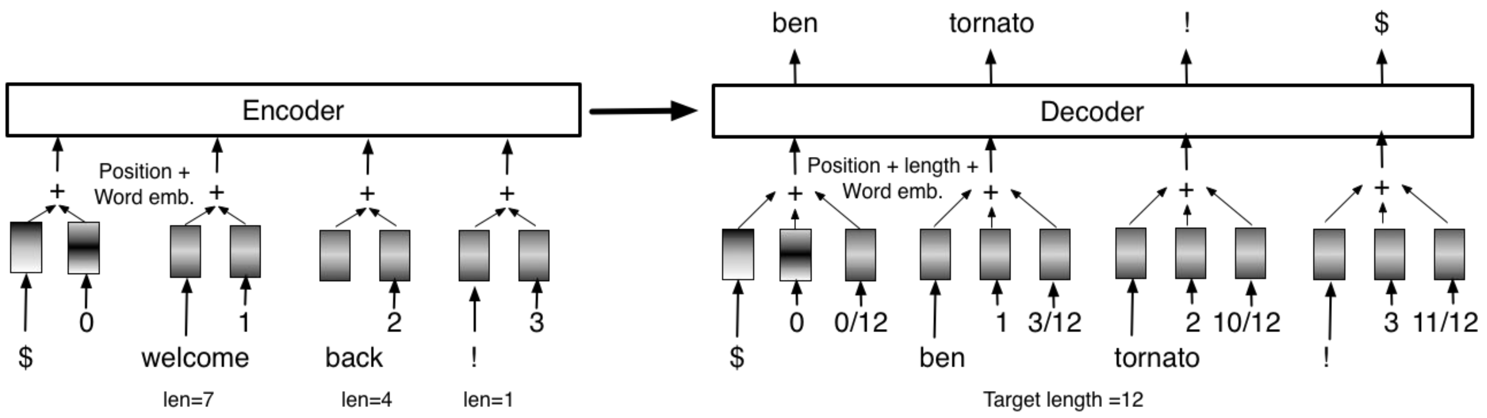
Can we make it versatile to any pre-trained model ?

Controlling Verbosity of NMT: Length-Token



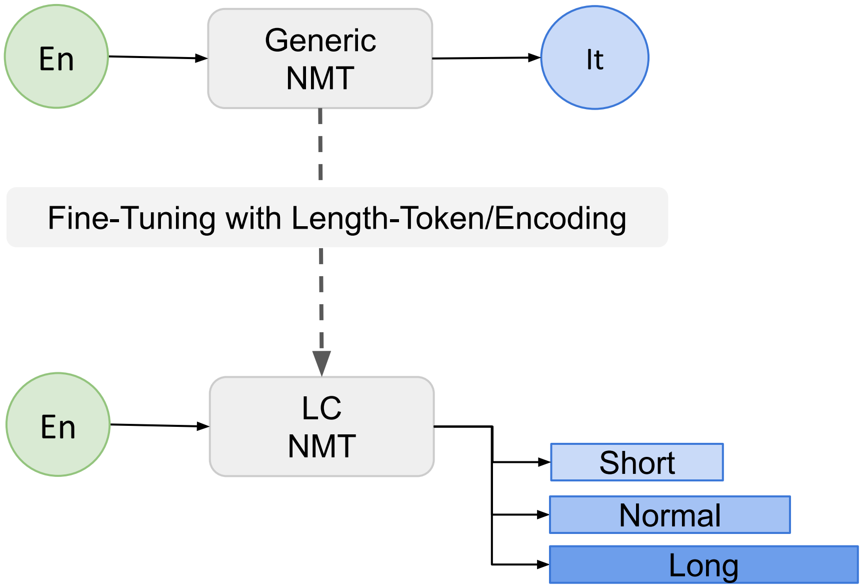
Approach conditions the output of NMT to a given target-source length-ratio class

Controlling Verbosity of NMT: Length-Encoding



Approach enriches the positional embedding of NMT with length information.

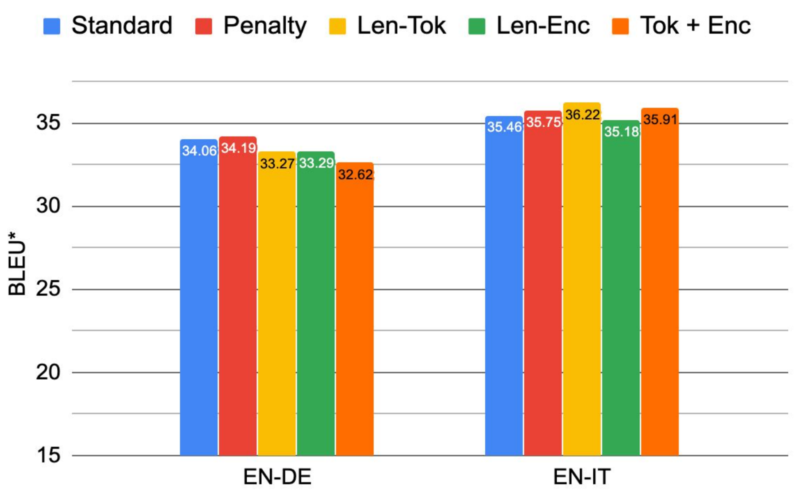
Controlling Verbosity of NMT: as a Fine-Tuning Task

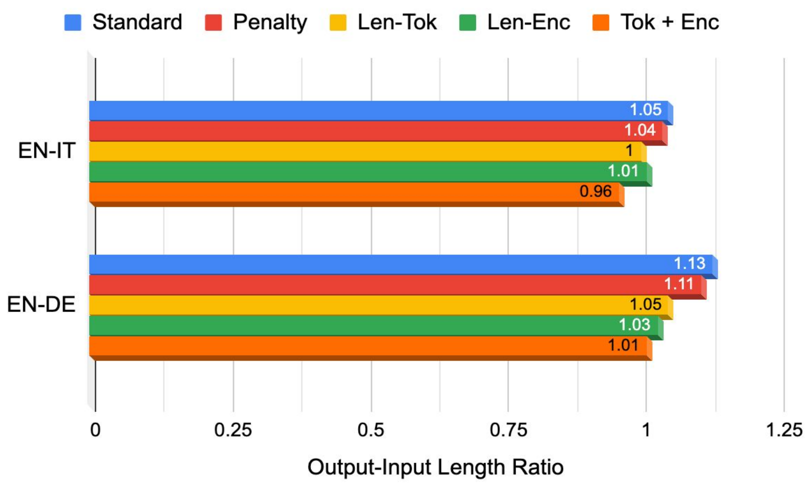


Advantages:

- Versatile to any pre-trained model
- Better performance than training from scratch
- Faster training to converge
- Language independent

Experimental Results





Model performance (left) with respect to output length with *short* condition.

Examples

Examples of shorter translation fragments obtained by paraphrasing, drop of words, and change of verb tense.

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 of speech
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
SRC	And this was done
NMT	E questo è stato fatto in modo che
LC-NMT	E questo fu fatto in modo che

Takeaway

Human evaluation:

- Confirms the translation quality observed with BLEU score

- Linguistics variations of the model to generate short translations, includes:
    - Abbreviations & Paraphrases.
    - Simple verb tenses over compound.
    - Avoiding adjectives, adverbs, pronouns & articles.

Lakew et al., IWSLT, *2019.*

Takeaway

Proposed two solutions for controlling output length of NMT:

Length-Tok: allows a coarse-grained control over the length without degradation in quality.

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

Fine-Tuning: works in a versatile with any pre-trained model.

Lakew et al., IWSLT, *2019.*

Conclusions

Conclusions

Low-Resource Multilingual NMT:

- We confirmed multilingual model improves performance in low-resource settings, and showed how pivoting using multilingual model can be beneficiary, when direct zero-shot fails.

Zero-Resource NMT with Zero-shot NMT Modeling:

- We Proposed a Zero-Shot NMT modeling approach using monolingual data that improves the baseline multilingual zero-shot by a larger margin.

Conclusions

Transfer-Learning:

- We showed a dynamic transfer-learning that trailers the parent model with the child model language characteristics improves the performance by encouraging better positive-transfer and reducing the negative-transfer.

- We showed relevant data selection from other high-resourced languages further improve the transfer-learning from the parent to the child model.

Conclusions

Neural Machine Translation into Language Varieties:

- We showed the possibility of modeling a single model that can generate several varieties translation. We further showed how to incorporate generic data without variety specific label into the training objective.

Controlling the Output of Neural Machine Translation:

- We showed the possibility of modeling a single NMT model that can generate outputs with different level of verbosity, while keeping the performance.

Selected Papers

Lakew, Surafel Melaku, Mattia Antonino Di Gangi, and Marcello Federico. “Multilingual Neural Machine Translation for Low Resource Languages”. In Proceedings of the 4th Italian Conference on Computational Linguistics (CLiCIT), Rome, Italy, 2017.

Lakew, Surafel Melaku, Mauro Cettolo, and Marcello Federico. “A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), New Mexico, USA, 2018.

Lakew, Surafel Melaku, Quintino F Lotito, Negri Matteo, Turchi Marco, and Federico Marcello. “Improving Zero-Shot Translation of Low-Resource Languages”. In 14th International Workshop on Spoken Language Translation (IWSLT), Tokyo, Japan, 2017.

Lakew, Surafel Melaku, Marcello Federico, Matteo Negri, and Marco Turchi. “Multilingual Neural Machine Translation for Low Resource Languages”. In Italian Journal of Computational Linguistics (IJCoL), Rome, Italy, 2018.

Lakew, Surafel Melaku, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. “Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary”. In 15th International Workshop on Spoken Language Translation (IWSLT), Bruges, Belgium, 2018.

Lakew, Surafel Melaku, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. “Adapting Multilingual Neural Machine Translation to Unseen Languages”. In 16th International Workshop on Spoken Language Translation (IWSLT), Hong Kong, 2019

Lakew, Surafel Melaku, Aliia Erofeeva, and Marcello Federico. “Neural Machine Translation into Language Varieties”. In Proceedings of the Third Conference on Machine Translation: Research Papers (WMT), Brussels, Belgium, 2018.

Lakew, Surafel Melaku, Mattia Di Gangi, and Marcello Federico. “Controlling the Output Length of Neural Machine Translation”. In 16th International Workshop on Spoken Language Translation (IWSLT), Hong Kong, 2019.

Thank You!

Questions and Comments are Welcome!

Surafel Melaku Lakew

Advisor: Marcello Federico

Fondazione Bruno Kessler | University of Trento

FONDAZIONE
BRUNO KESSLER

UNIVERSITAS ATHESINA STUDIORUM

End of Presentation