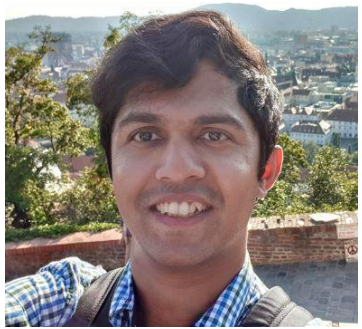
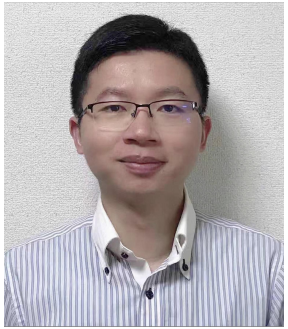
Multilingual Neural Machine Translation



Raj Dabre
NICT
Kyoto, Japan



Chenhui Chu
Kyoto University
Kyoto, Japan

Anoop Kunchukuttan
Microsoft STCI
Hyderabad, India

COLING 2020 Tutorial, December 12th, 2020

Tutorial Material

Tutorial Homepage

https://github.com/anoopkunchukuttan/multinmt_tutorial_coling2020

Survey Paper

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.* 53, 5, Article 99 (September 2020), 38 pages. https://doi.org/10.1145/3406095

Updated Bibliography (the field is moving so fast!)

https://github.com/anoopkunchukuttan/multinmt_tutorial_coling2020/mnmt_bibliography.pdf

Self Introduction (Chenhui Chu)

- Experience
  - 2020-present: program-specific associate professor @ Kyoto University
  - 2017-2020: research assistant professor @ Osaka University
  - 2015-2017: researcher @ Japan Science and Technology Agency
  - 2014-2015: research fellowship for young scientists (DC2) @ JSPS
- Research
  - Machine translation (JSPS DC2, Chinese-Japanese MT practical application project, JSPS research activity start-up)
  - Language and vision understanding (ACT-I, MSRA CORE, JSPS young scientists)

Outline of This Tutorial

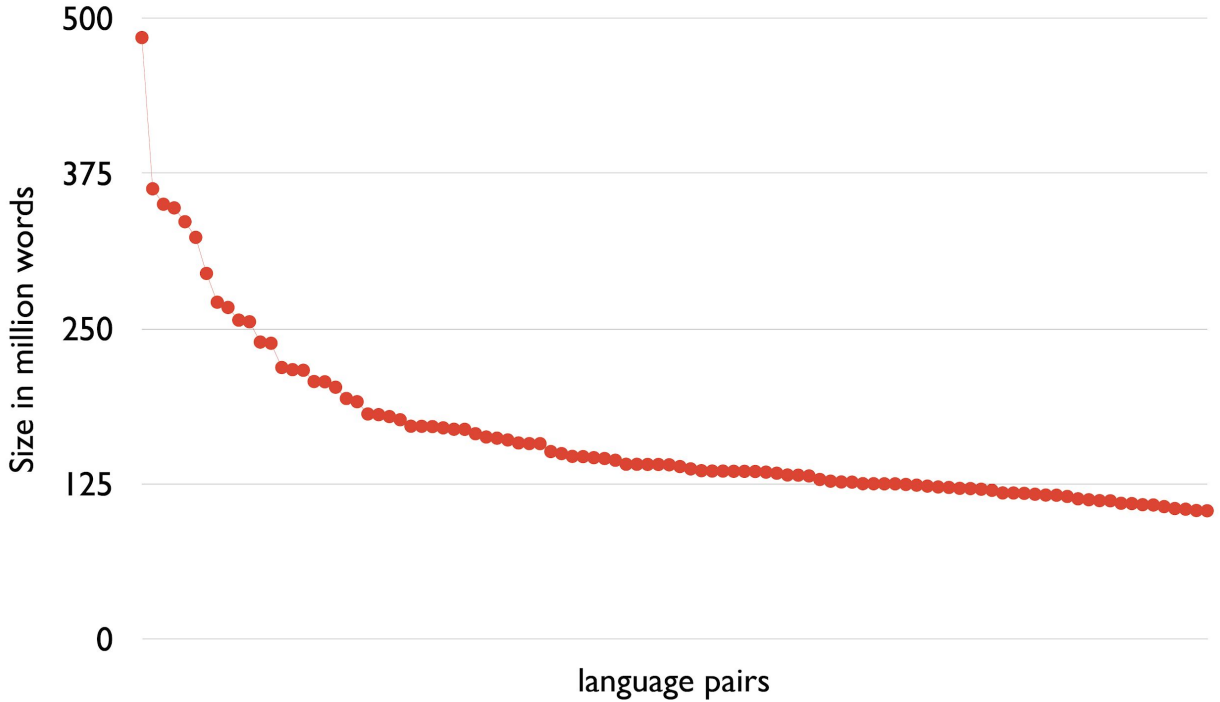
- Overview of Multilingual NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

What is Multilingual NMT (MNMT)?

- Definition
  - Neural machine translation (NMT) systems handling translation between more than one language pair

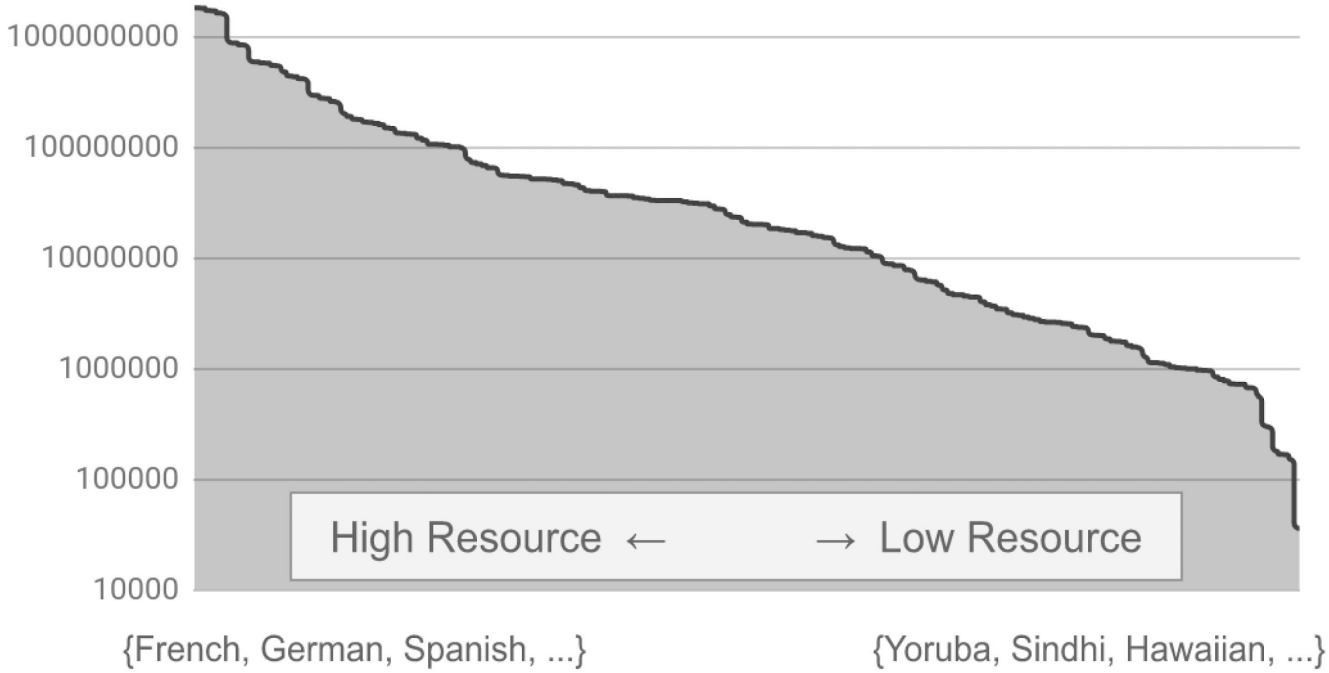
- A research field to study:
  - How can we leverage multilingual data effectively in order to learn distributions across multiple languages so as to improve MT (NLP) performance across all languages?

Why MNMT?



Size of the top-100 language pairs in OPUS (Tiedemann et al., 2012)

Why MNMT?

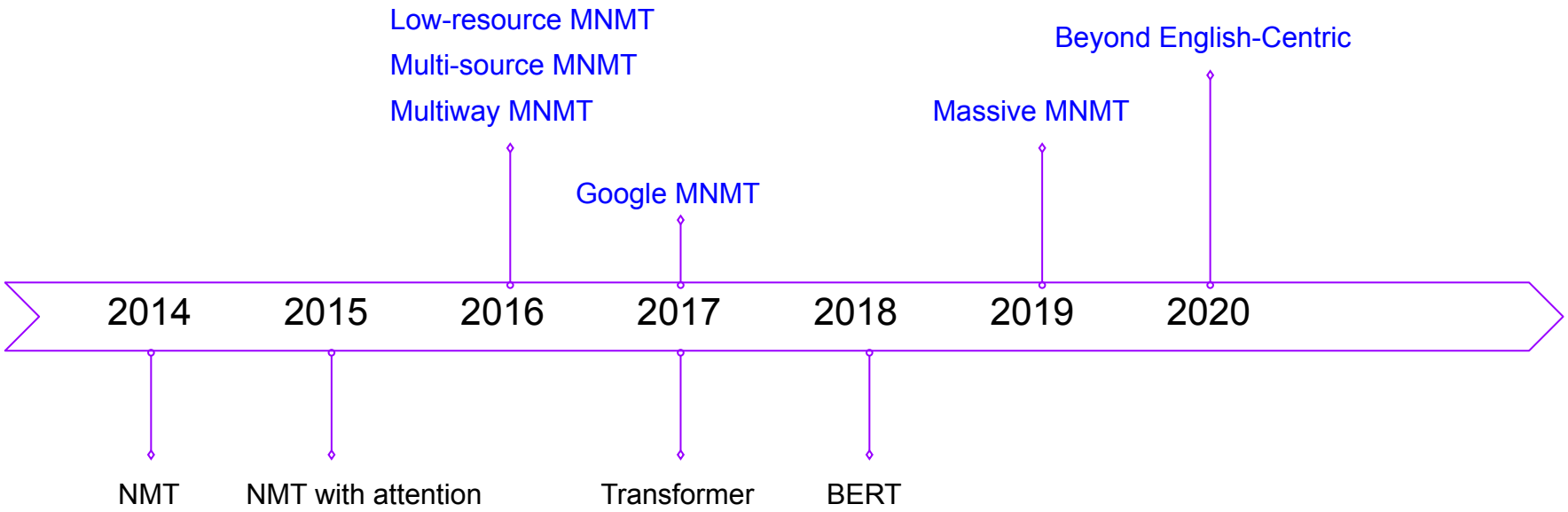


Data Distribution over language pairs (Arivazhagan et al., 2019)

Motivation and Goal of MNMT (Dabre et al., 2020)

- Motivation
  - MNMT systems are desirable because training models with data from many language pairs might help a resource-poor language acquire extra knowledge; from the other languages
  - MNMT systems tend to generalize better due to exposure to diverse languages, leading to improved translation quality compared to bilingual NMT systems
- Goal
  - Have a one-model-for-all-languages solution to MT (NLP) applications
  - Shared multilingual distributed representations help MT (NLP) for low-resource languages

Timeline of MNMT Research



Google's MNMT System (Johnson et al., 2017)

Table 1: Many to One: BLEU scores on for single language pair and multilingual models. ⋆: no oversampling

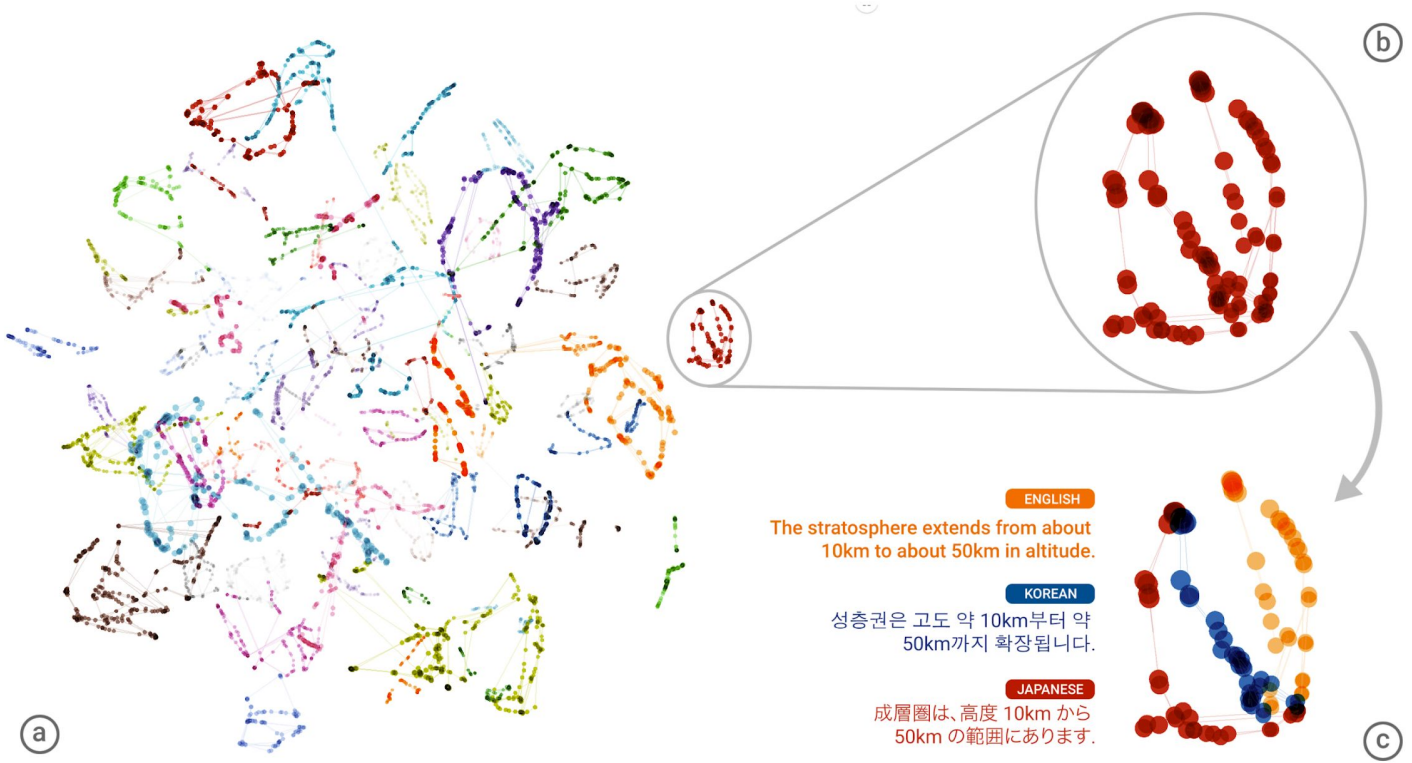
Model	Single	Multi	Diff
WMT De→En	30.43	30.59	+0.16
WMT Fr→En	35.50	35.73	+0.23
WMT De→En⋆	30.43	30.54	+0.11
WMT Fr→En⋆	35.50	36.77	+1.27
Prod Ja→En	23.41	23.87	+0.46
Prod Ko→En	25.42	25.47	+0.05
Prod Es→En	38.00	38.73	+0.73
Prod Pt→En	44.40	45.19	+0.79

Table 2: One to Many: BLEU scores for single language pair and multilingual models. ⋆: no oversampling

Model	Single	Multi	Diff
WMT En→De	24.67	24.97	+0.30
WMT En→Fr	38.95	36.84	-2.11
WMT En→De⋆	24.67	22.61	-2.06
WMT En→Fr⋆	38.95	38.16	-0.79
Prod En→Ja	23.66	23.73	+0.07
Prod En→Ko	19.75	19.58	-0.17
Prod En→Es	34.50	35.40	+0.90
Prod En→Pt	38.40	38.63	+0.23

A single multilingual model achieves comparable performance for English-French and surpasses state-of-the-art results for English-German.

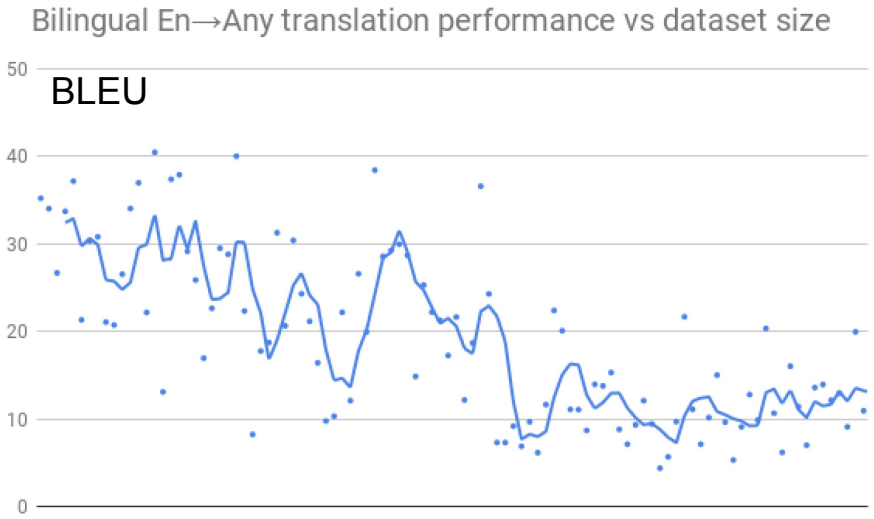
Language Clusters (Johnson et al., 2017)

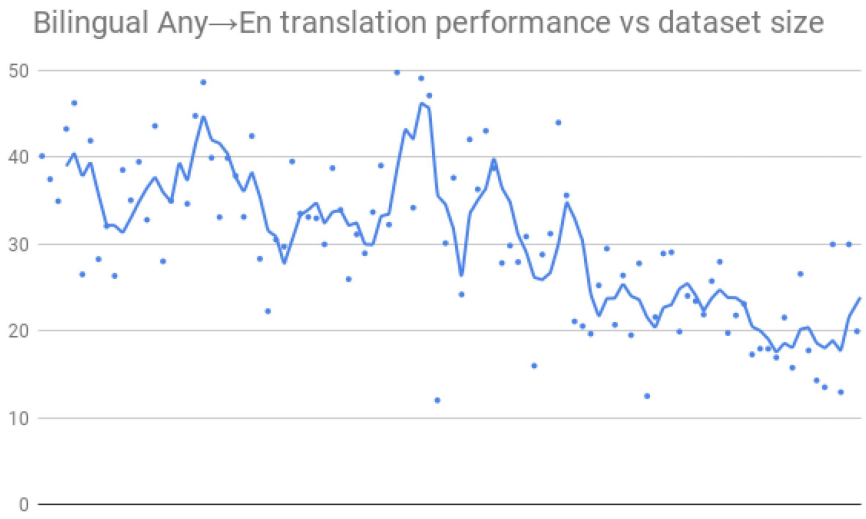


Mixing Target Languages (Johnson et al., 2017)

w_{ko}	I must be getting somewhere near the centre of the earth.
0.00	私は地球の中心の近くにどこかに行っているに違いない。
0.40	私は地球の中心近くのどこかに着いているに違いない。
0.56	私は地球の中心の近くのどこかになっているに違いない。
0.58	私は지구の中心의가까이에어딘가에도착하고있어야한다。
0.60	나는지구의센터의가까이에어딘가에도착하고있어야한다。
0.70	나는지구의중심근처어딘가에도착해야합니다。
0.90	나는어딘가지구의중심근처에도착해야합니다。
1.00	나는어딘가지구의중심근처에도착해야합니다。

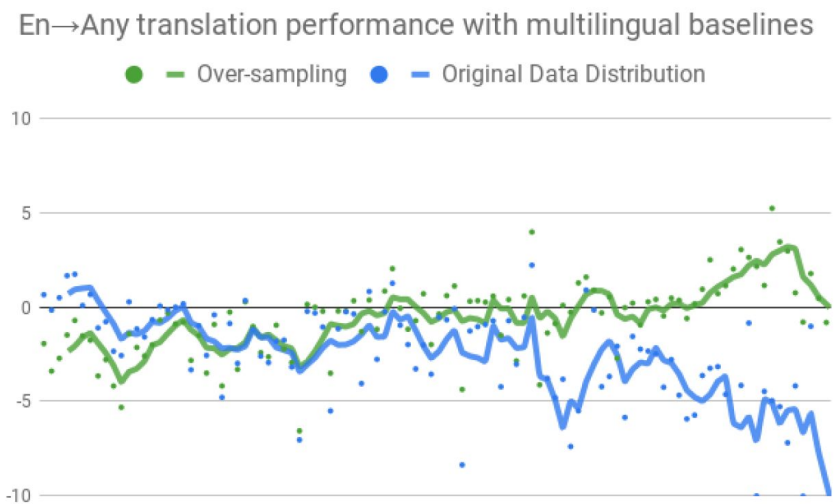
Google's Massive MNMT System (Arivazhagan et al., 2019)

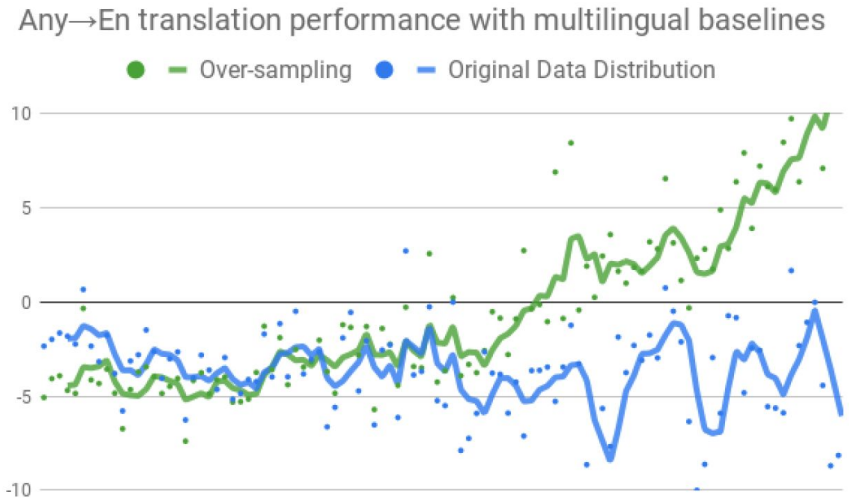




Studied a single massive MNMT model handling
103 languages trained on over 25 billion examples

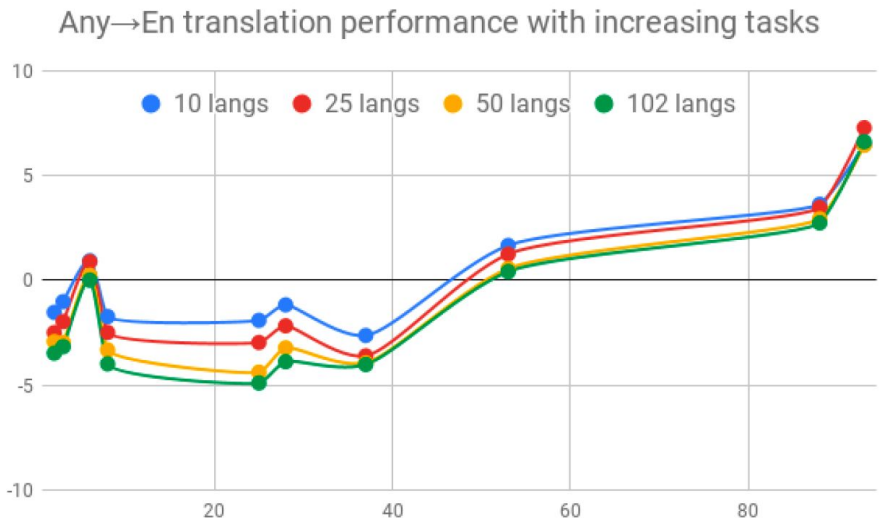
Effect of Massive MNMT (Arivazhagan et al., 2019)

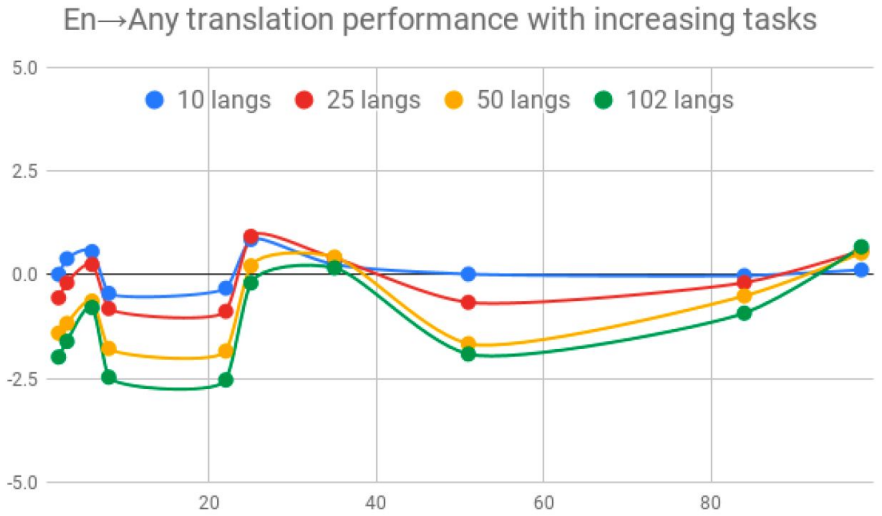




Massive MNMT is helpful for low-resource languages but not high-resource languages

Effect of Task Numbers (Arivazhagan et al., 2019)





Increasing numbers of languages is not always helpful

Facebook's Beyond English-Centric MNMT System (Fan et al., 2020)

- M2M-100: the first MNMT model that can translate between any pair of 100 languages without relying on English data
- It outperforms English-centric systems by 10 points on the widely used BLEU metric for evaluating machine translations
- M2M-100 is trained on a total of 2,200 language directions — or 10x more than previous best, English-centric multilingual models

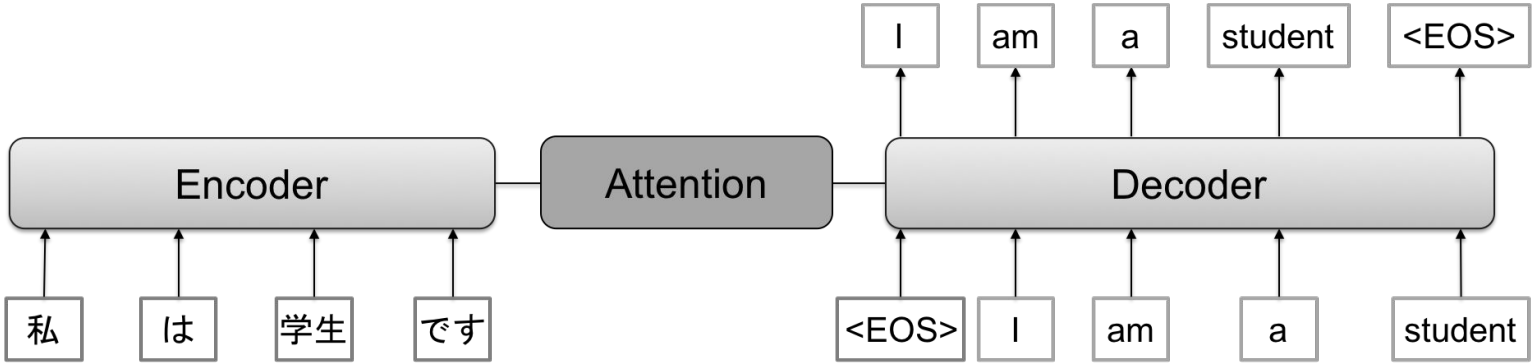
Performance of M2M-100 (Angela et al., 2020)

Model	SUPERVISED Low	Mid	High	Avg	ZERO-SHOT Avg	Avg
Bitext baselines	5.3	20.4	38.2	17.3	—	—
English-centric 1.2B	6.7	12.6	28.5	10.1	5.4	9.1
M2M-100 1.2B	9.9	23.0	37.5	21.1	12.5	17.2
M2M-100	11.1	24.1	39.6	22.3	13.7	18.4

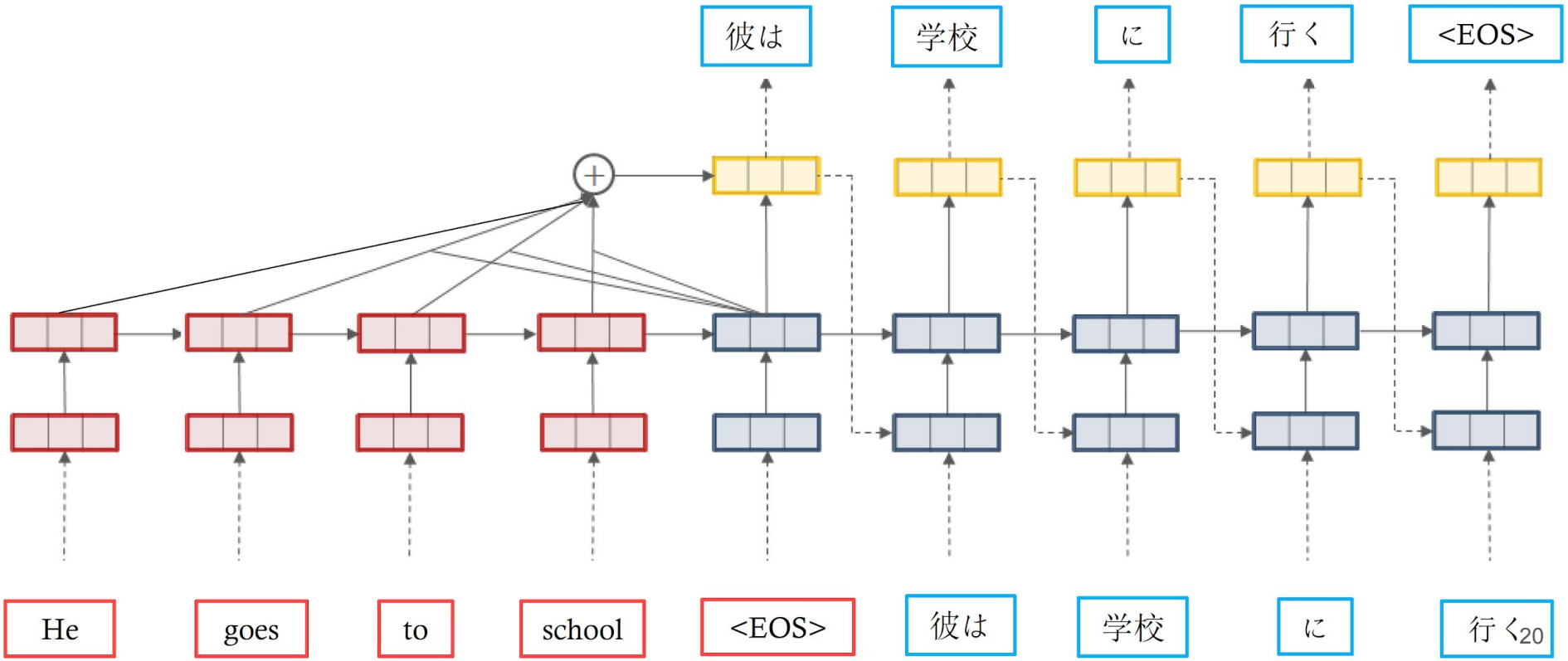
Goal of This Tutorial

- Understand the basics of MNMT
- Cover an in-depth survey of existing literature on MNMT
  - Central use-case, resource scenarios, underlying modeling principles, core-issues and challenges of various approaches
  - Strengths and weaknesses of several techniques by comparing them with each other
- Inspire new ideas for researchers and engineers interested in MNMT

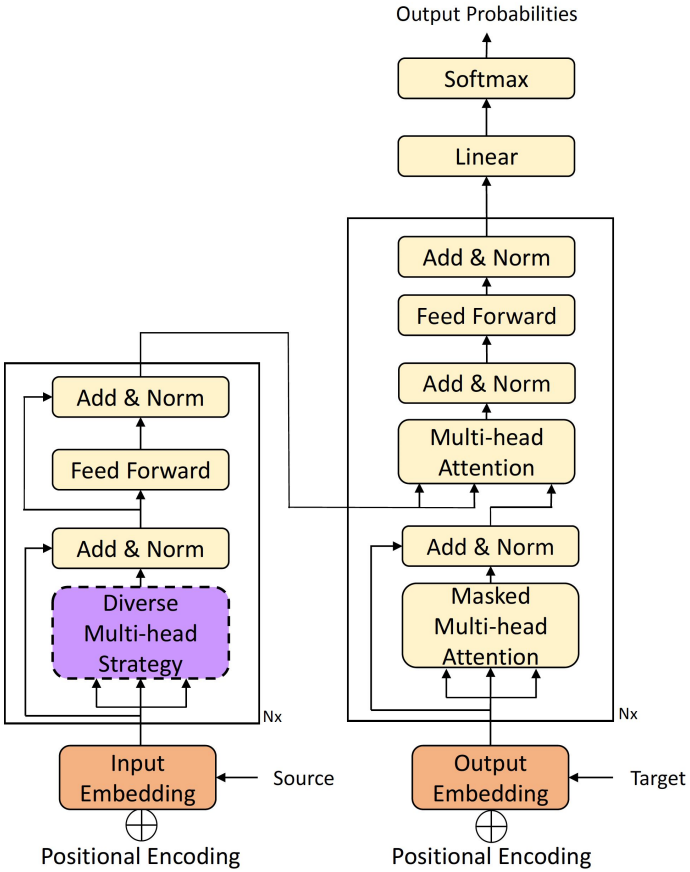
Basic NMT Architecture (Dabre et al., 2020)



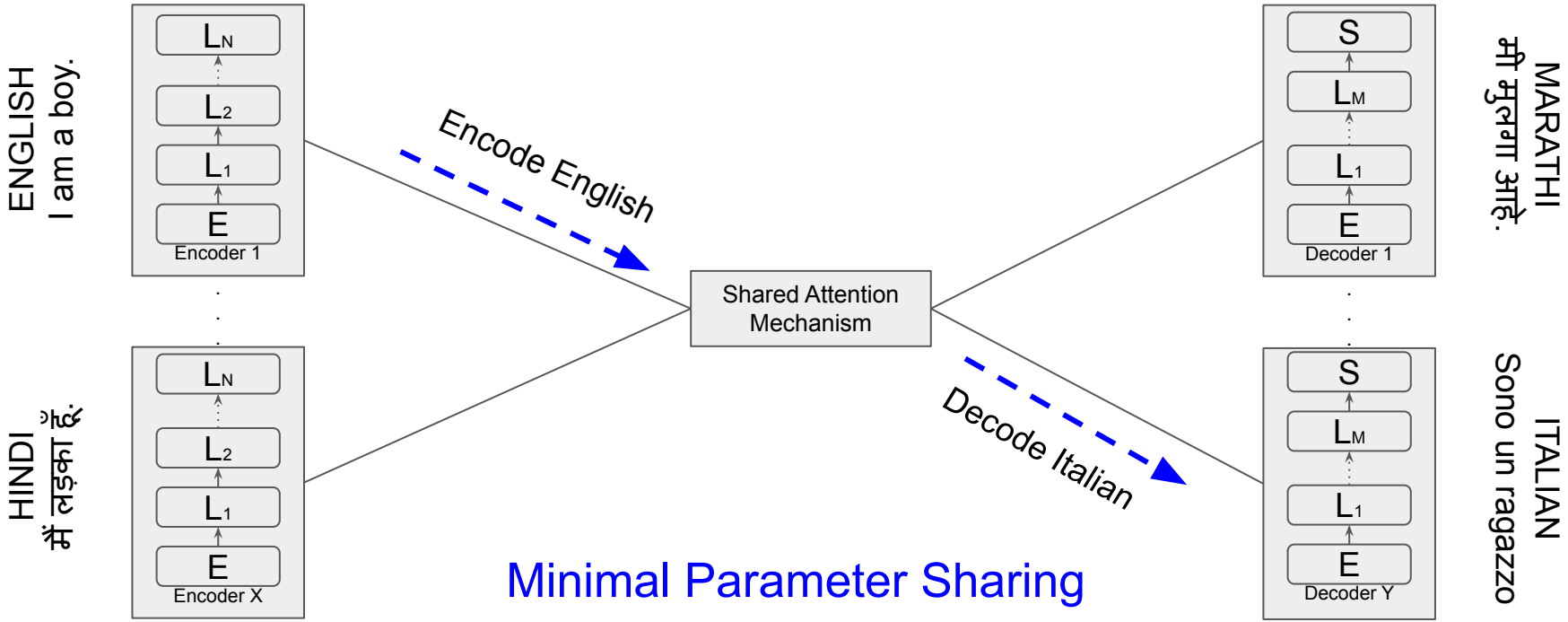
RNN based NMT (Bahdanau et al., 2015)



Self-Attention Based NMT (Vaswani et al., 2017)



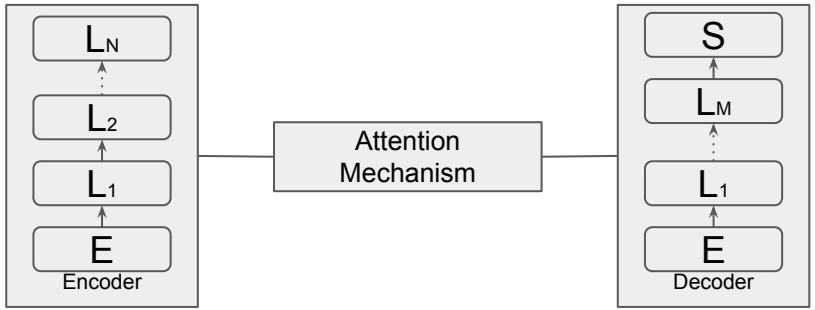
Initial MNMT Models (1/2) (Firat et al., 2016)



Initial MNMT Models (2/2) (Johnson et al., 2017)

1. <2mr> I am a boy.
2. <2mr> मैं लड़का हूँ.
3. <2it> I am a boy.
4. <2it> मैं लड़का हूँ.

.
.

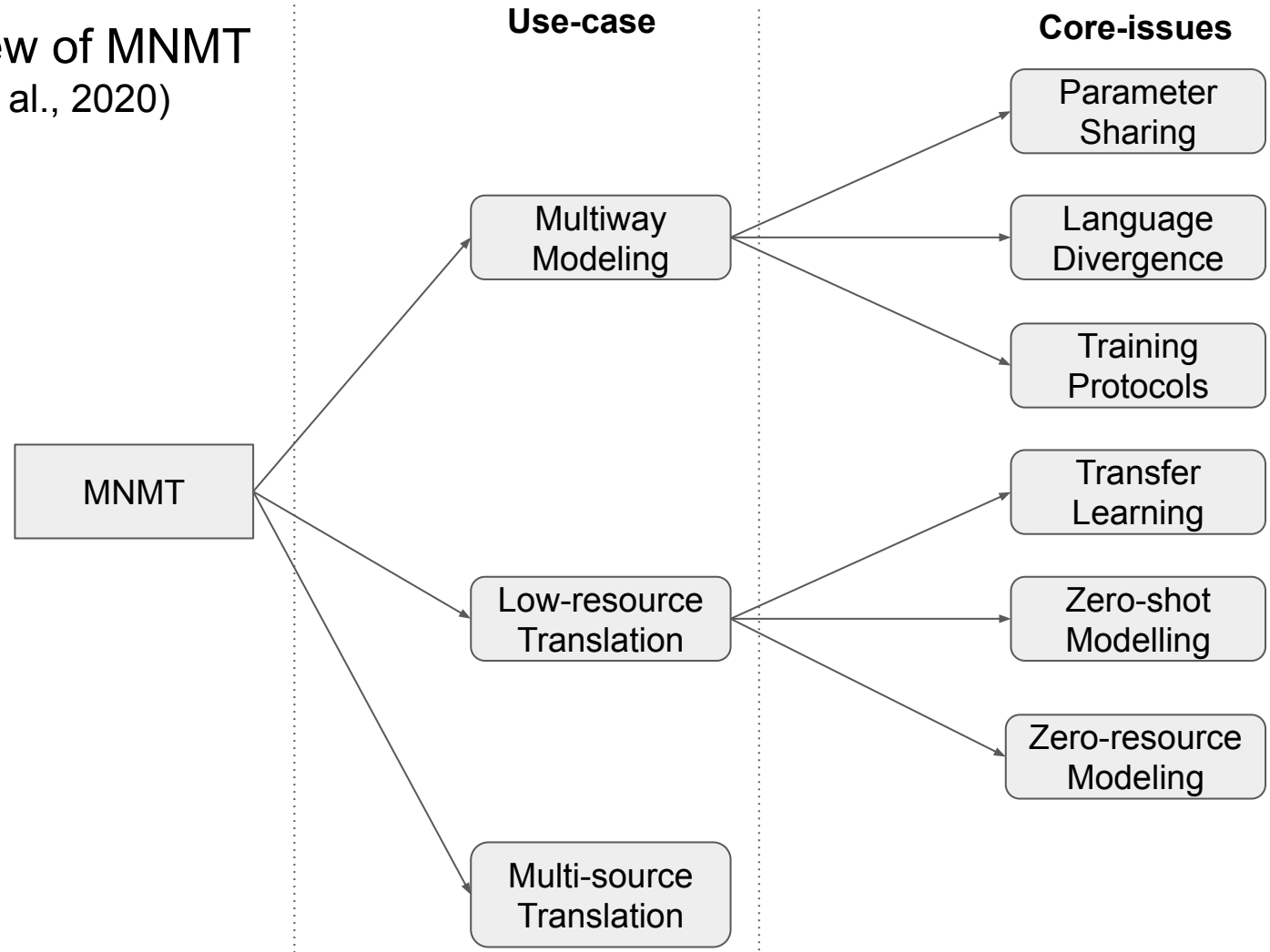


1. मी मुलगा आहे.
2. मी मुलगा आहे.
3. Sono un ragazzo.
4. Sono un ragazzo.

.
.

Complete Parameter Sharing

Overview of MNMT
(Dabre et al., 2020)



Outline of This Tutorial

- Overview of Multilingual NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

Self Introduction (Raj Dabre)

- Experience
  - 2018-present: Researcher at NICT, Japan
  - 2014-2018: MEXT Ph.D. scholar at Kyoto University, Japan
  - 2011-2014: M.Tech. Government RA at IIT Bombay, India
- Research
  - Low-Resource Natural Language Processing
    - **Multilingual Machine Translation: 2012-present**
    - Fundamental Analysis: 2011-2014
  - Efficient Deep Learning:
    - **Compact, flexible and fast models (2018-present)**

Topics to address

- Parameter sharing
- Massively multilingual models
- Language divergence
- Training protocols

Topics to address

- Parameter sharing
- Massively multilingual models
- Language divergence
- Training protocols

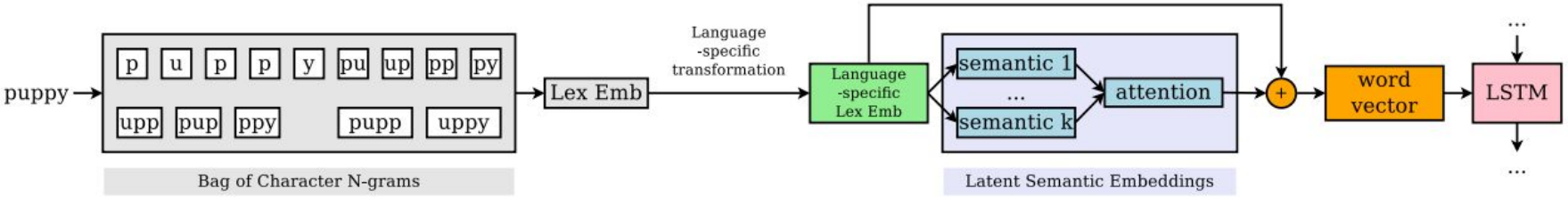
Parameter Sharing: Finding The Right Balance

- Empirically determined component sharing
  - How to maximize impact of shared vocabulary?
  - Universal encoders or decoders?
  - Should attention be universal or language specific?
  - How much sharing is good sharing?

- Parameterized parameters for shared modeling
  - Learn about sharing

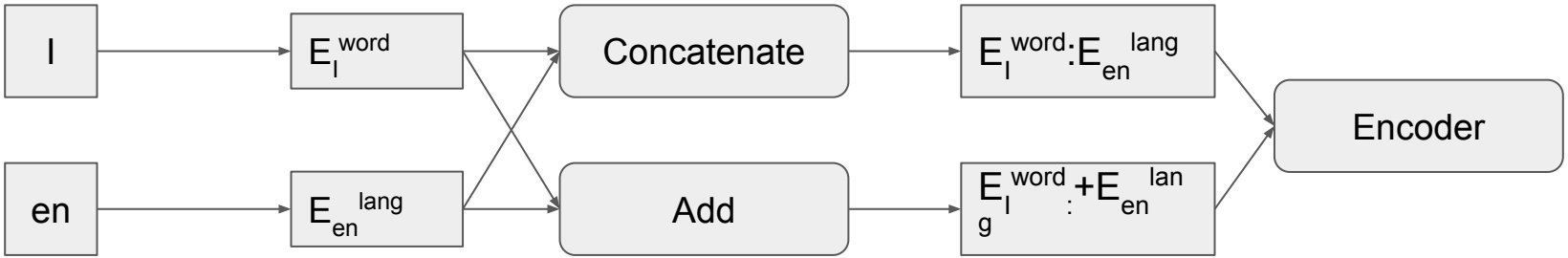
Sharing Vocabularies (1)

- Naive solution: A single vocabulary for all languages (Johnson et al., 2017)
- Adding constraints: Soft-decoupled encoding (Wang et al., 2019)
  - Word or sub-word vocabulary
  - N-gram embeddings, language specific and latent semantic transformation
  - Good for linguistically similar languages in transfer learning setups
- Future investigation
  - Indic languages
  - **Groups of language families (for massively multilingual NMT)**
    - Chung et al. 2020, Lyu et al. 2020



Sharing Vocabularies (2)

- Language sensitive embeddings (Wang et al., 2019)
  - Augment shared embeddings with language specific indicators (factors)
  - One hot, learnable, direction specific
- Language family shared embeddings
  - Separate shared embeddings for different families
  - Avoid accidental n-gram vocabulary overlaps



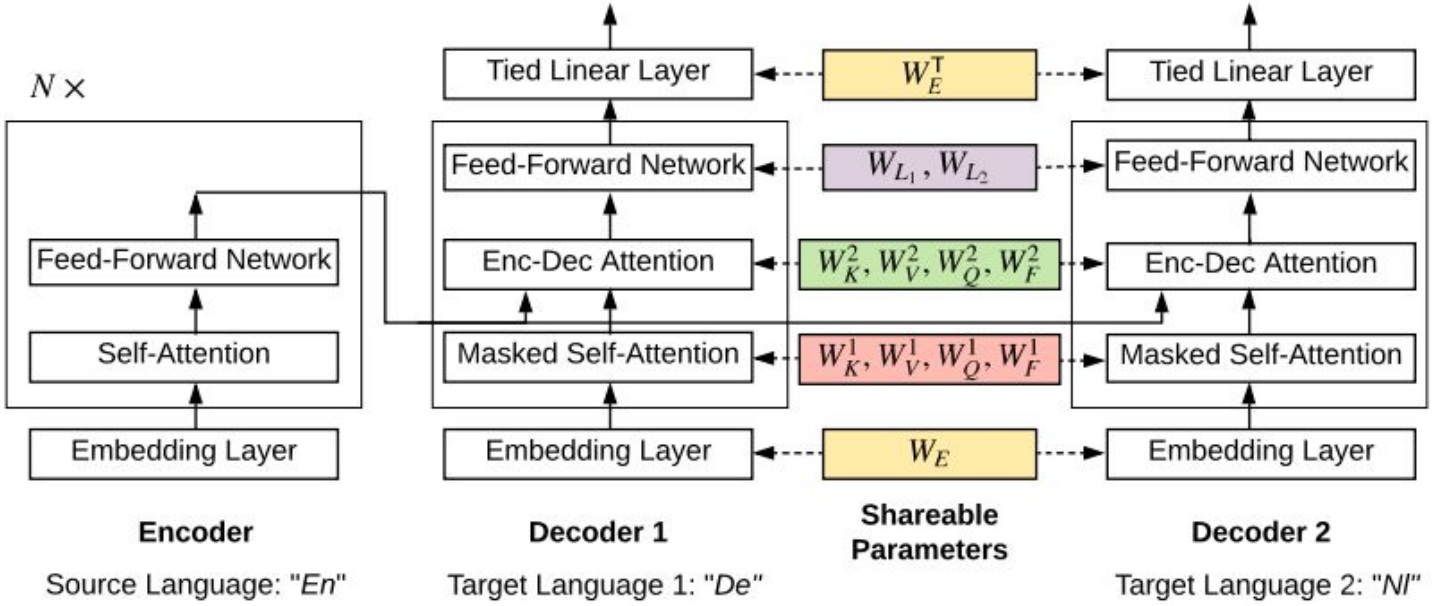
Sharing In Encoders and Decoders (1)

- Most works agree on a universal encoder
  - Dong et al. 2015; Sachan et al. 2018; Wang et al. 2019
  - Lack of cross-lingual components

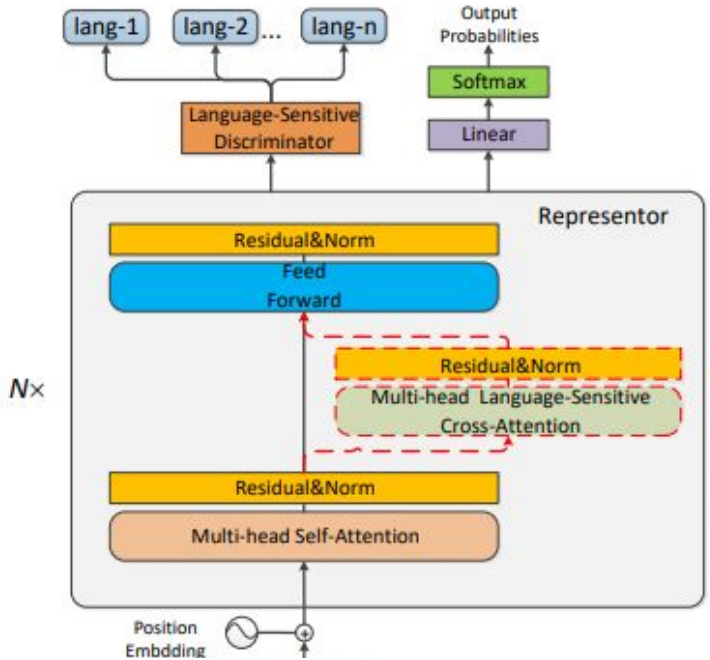
- Component sharing in decoders is tricky
  - Balance between universal and distinct representation (more on this later)
  - Cross-lingual components

Sharing In Encoders and Decoders (2)

- Sachan et al. 2018 get optimal results with *sharing attention and embedding*
- *Distinct FF parameters* for language specific adaptation
- *Reduce MNMT model sizes by 20-30%* but slightly improve quality



Sharing In Encoders and Decoders (3)

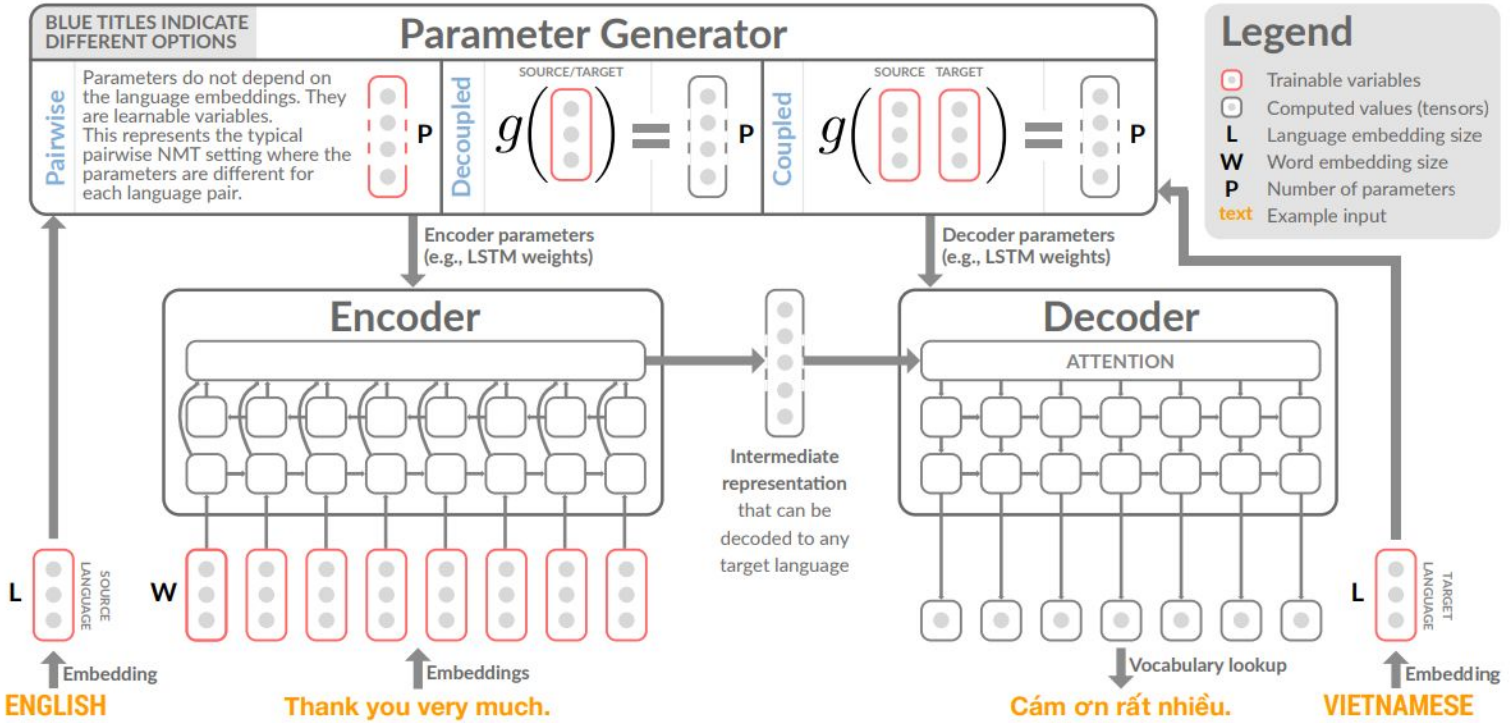


- Wang et al. 2019 focus on language sensitization
  - *Sharing self-attention and FF parameters*
  - Shared encoder/decoder →*representor*
    - Also in Dabre et al. 2019
- Language sensitization strategies:
  - **Language sensitive embeddings**
  - **Language pair specific cross-attention**
  - **Language discrimination**
- General improvement of 1-2 BLEU

A Note On Attention Sharing (Blackwood et al. 2018)

- Universal shared attention
  - Firat et al. 2016
  - *Too much load* on a single set of parameters
- Source specific attention
  - *Reduces load* but not too effective
- Target specific attention
  - *Works best* (shows importance of unshared components in decoders)
  - Orthogonal to Sachan et al. 2018 where shared attention was optimal
- Source-target specific attention
  - *Least load* but **prevents zero-shot**

Contextual Parameters (Platanois et al. 2018)

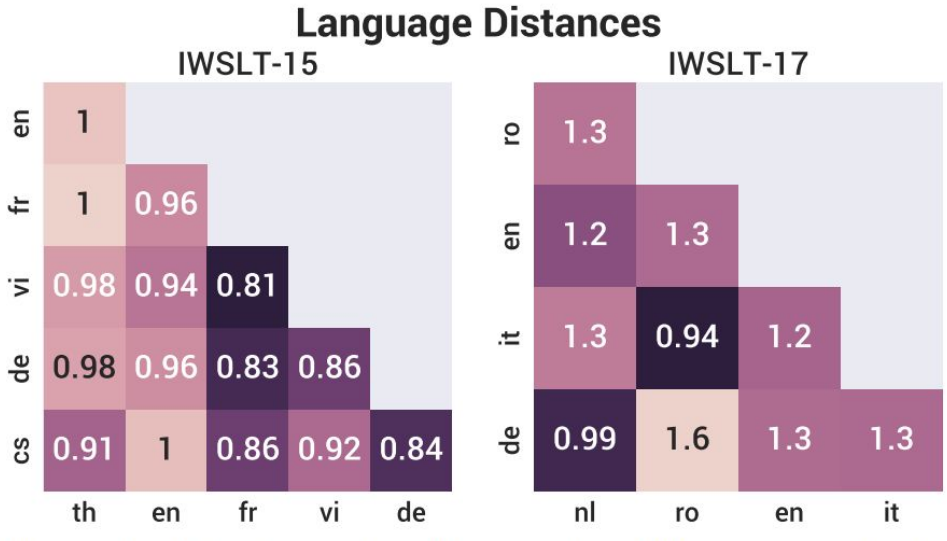


Generate parameters using parameters!

Contextual Parameters (Platanois et al. 2018)

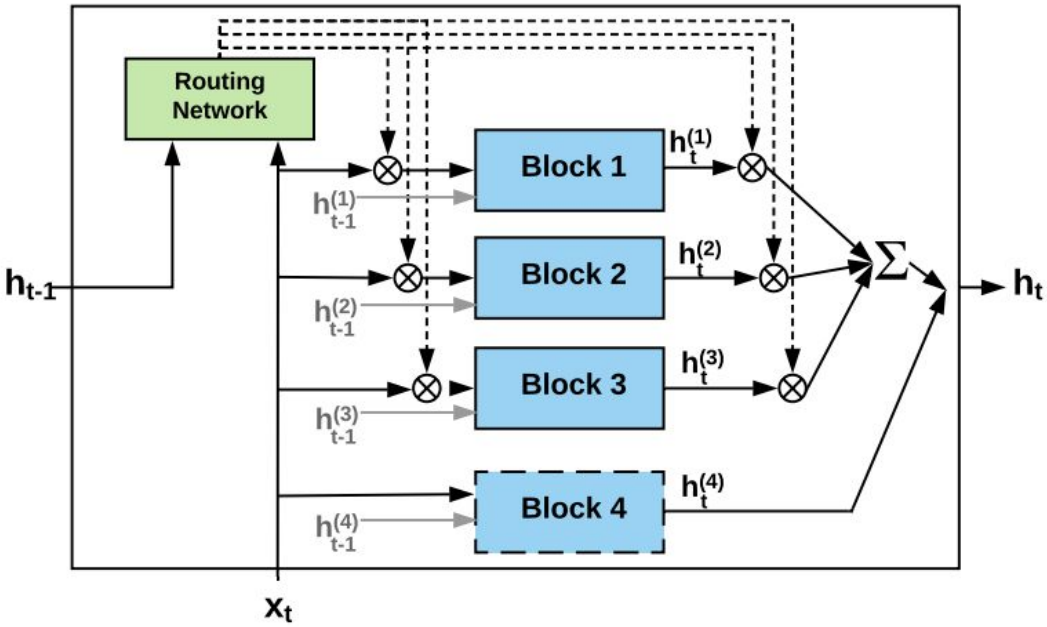
- Meta-parameterization
  - Parameters generate parameters: Θ = F(ƛ)
- Can support parameter groups
  - Potential for *self-discovering optimal language families*
  - Related work on family specific sharing by Lyu et al. 2020
- Significant model compression
  - Separate models: O(L*L*P + 2*L*L*W*V)
  - This model: O(P*M + L*W*V)
- Significant improvement in translation quality

Contextual Parameters (Platanois et al, 2018)



- *Automatically learns language similarity*
  - IT similar to RO
  - VI similar to FR?
- Make groups of contextual parameters
  - Two step meta-learning
  - Step 1: Discover groups
  - Step 2: Optimal sharing
- Next steps:
  - Conditional computation
  - Neural architecture search
  - Tensor routing (Zaremoudi et al. 2018)

Tensor routing (Zaremoudi et al. 2018)



- Augmented encoder block
  - Language specific block (dotted)
  - Shared blocks (solid)
  - Routing controls relevance of blocks
- Another perspective
  - **Each block is an expert**
  - Related to MOE and Gshard works

Topics to address

- Parameter sharing
- Massively multilingual models
- Language divergence
- Training protocols

Massively Multilingual Models

- Google's work
  - MNMT *In the wild*
  - Adaptor Layers
  - GShard
- University of Edinburgh's work
  - Reproducible
- Bottlenecks
    - Hardware/computational/cost
    - Representational

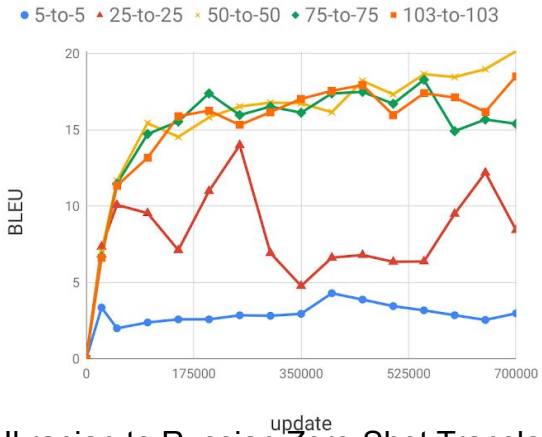
Google's Massively MNMT (Aharoni et al. 2019)

- 59-lingual low-resource models (6-layer transformer base)
- One to many models better than many-to-many for non-English targets
  - 2-3 BLEU improvement
  - *Relative over-representation of English*
- Many to one models worse than many-to-many for English targets
  - 2-3 BLEU drop
  - *Relative under-representation of English*

Google's Massively MNMT (Aharoni et al. 2019)

- 103-lingual model (6-layer variation of transformer-big)
- Many-to-one and one-to-many are both better than many-to-many
  - *N-way corpora is detrimental to many-to-one performance*
- Supervised NMT quality degrades with more languages
- Zero-shot NMT quality increases with more languages

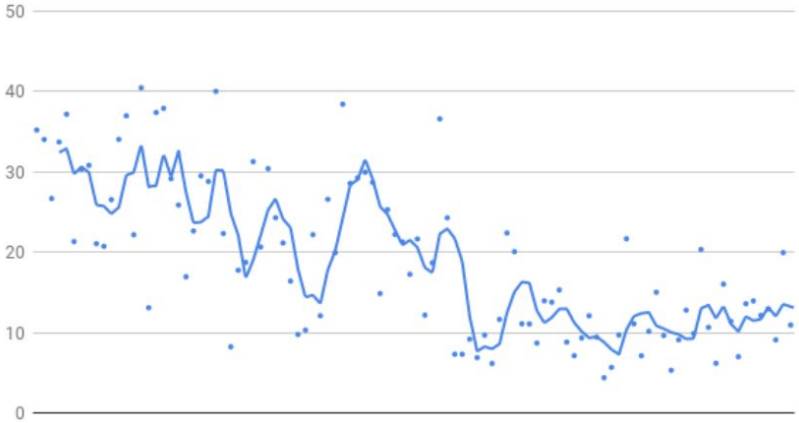
	Ar	Az	Be	De
baselines	23.34	16.3	21.93	30.18
many-to-one	**26.04**	**23.68**	**25.36**	35.05
many-to-many	22.17	21.45	23.03	**37.06**



Ukranian to Russian Zero-Shot Translation

Google's Massively MNMT Model In the Wild

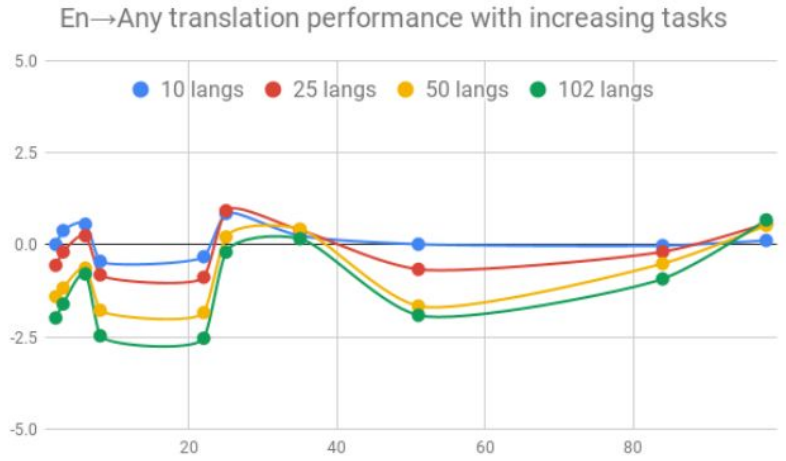
- Extension of Aharoni et al. 2019 by Arivazhagan et al. 2019
- Main focus
  - Temperature based data sampling for transfer-interference balance
  - Pushing number of supported language pairs to limit
  - Pushing MNMT performance with 1+ billion parameters
- Starting point: Quality is directly proportional to data size

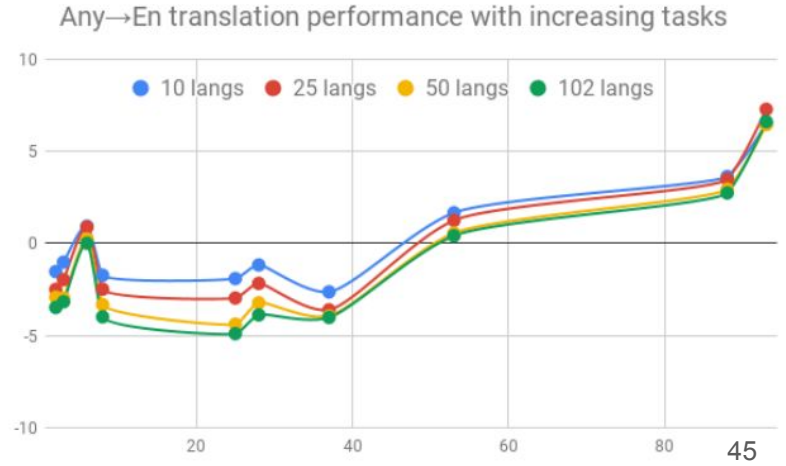


En→Any	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
Any→En	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63

More Languages and Directions

- Generic drop in performance for resource rich pairs
- Performance degrades as number of pairs increase
  - 25 languages seems to be optimal
- Separate one-to-all and all-to one models preferred
  - All to All models degrade translation into English



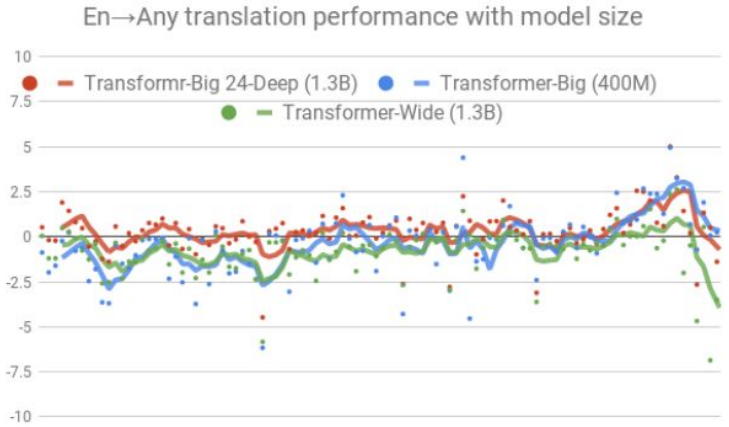


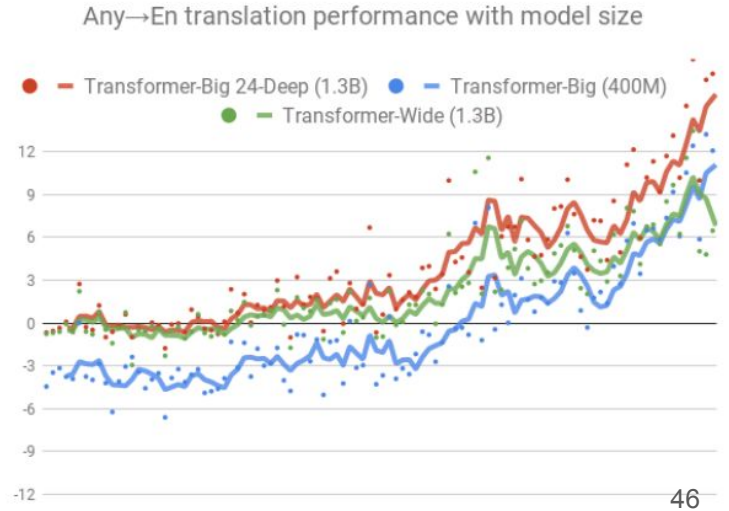
Any→En	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63
All→All	33.85	30.25	26.96
Any→En	36.61	33.66	30.56

En→Any	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
All→All	28.03	16.91	12.75
En→Any	28.75	17.32	12.98

Are Larger Models Helpful?

- 400M parameters
  - Transformer Big:
    - 6-layers, 1024-4096 hidden-filter sizes
    - 16 attention heads
- 1.3B parameters
  - Transformer Deep:
    - 24-layers, 1024-4096 hidden-filter sizes
    - 16 attention heads
    - Low-resource performance
  - Transformer Wide:
    - 12-layers, 2048-16384 hidden-filter sizes
    - 32 attention heads
    - High-resource performance
- 128 TPUs and 4M tokens per batch ($$$)



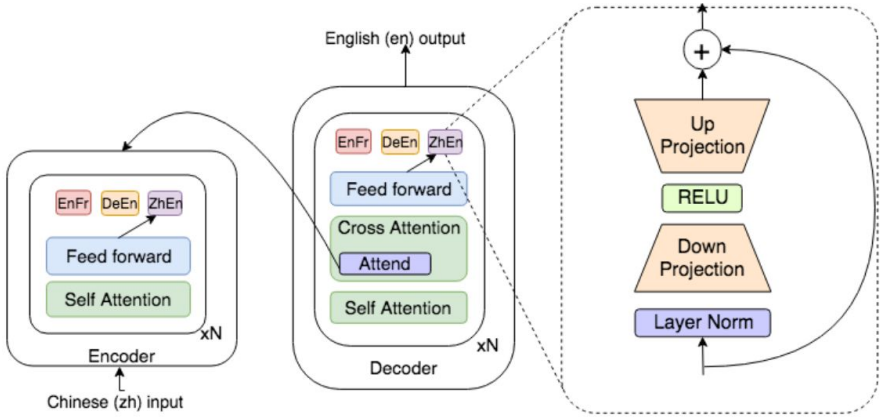


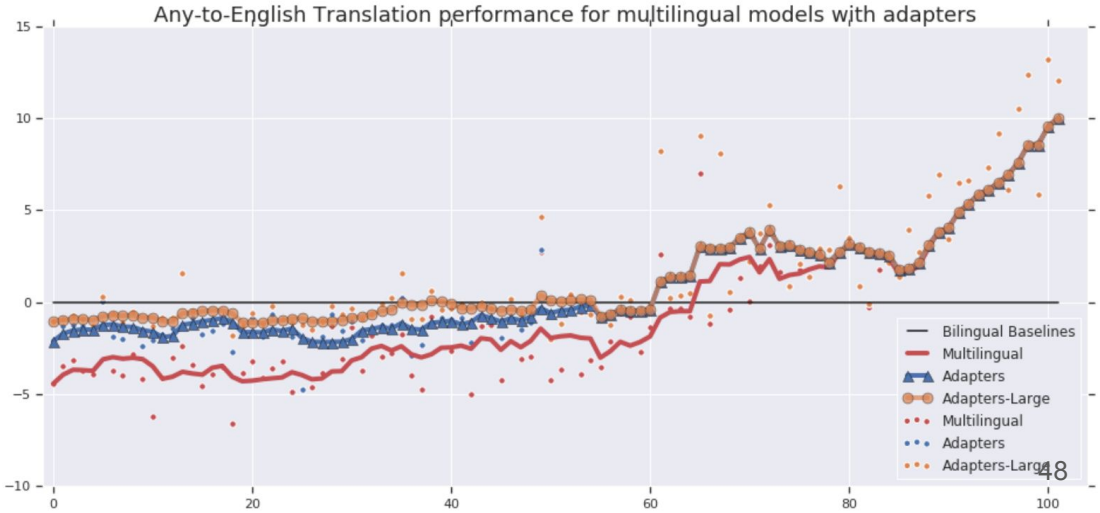
Language Aware Multilingualism

- Zhang et al. 2020
  - **OPUS 100 dataset: 55M sentence pairs**
- Deep transformers: Up to 2 BLEU better using 12 layers instead of 6 layers
  - Depth scaled initialization: Better initialization to handle deep stacking
  - Merged attention: Self and cross attention side by side
- Expressibility-Scalability: Improves by 2-3 BLEU
  - Target language aware layer normalization
  - Language aware linear transformation b/w Encoder and Decoder
- Random online back-translation: Boosts zero-shot quality by up to 10 BLEU
  - Back-translate monolingual corpora and use it to train

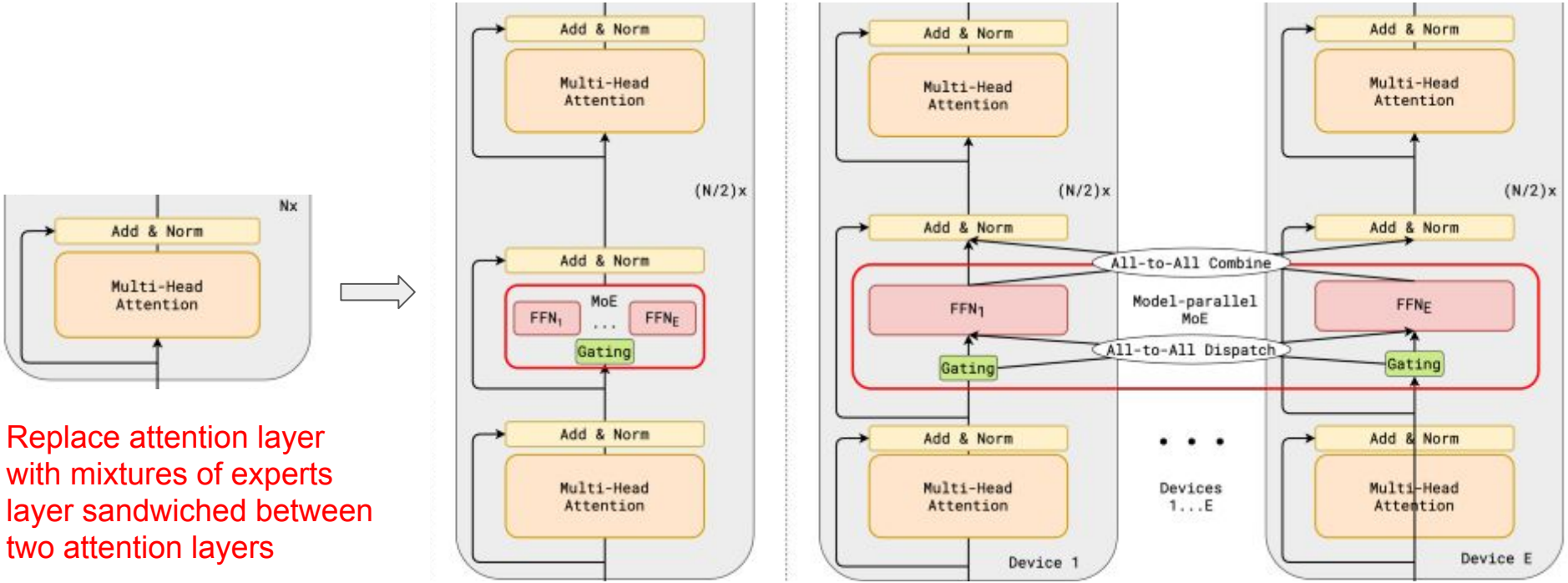
Adapting Previously Trained Models

- Feed forward layers to refine outputs
  - Bapna et al. 2019
  - **Partial solution to bottleneck**
  - Language pair specific
    - Zero-shot performance?
- 13.5% larger models
- Improved high-resource pair performance
  - Low-resource performance kept
- *Questions:*
  - Multiple adapters?
  - Dynamic adapter size?
  - Adapter-Base Model ratio in massively multilingual situation?





Pushing Limits: Mixtures Of Experts (Gshard; Lepikhin et al. 2020)



Replace attention layer with mixtures of experts layer sandwiched between two attention layers

Explosive growth in parameters (600B) as number of experts grow (2048).

Use model sharding and dynamic routing with large number of accelerators (2048 TPUs).

Engineering + Research = Ultimate Solution?

Id	Model	Cores	Steps per sec.	Batch sz. (Tokens)	TPU core years	Training time (days)	BLEU avg.
(1)	MoE(2048E, 36L)	2048	0.72	4M	22.4	**4.0**	**44.3**
(2)	MoE(2048E, 12L)	2048	2.15	4M	7.5	1.4	41.3
(3)	MoE(512E, 36L)	512	1.05	1M	15.5	11.0	43.7
(4)	MoE(512E, 12L)	512	3.28	1M	4.9	3.5	40.0
(5)	MoE(128E, 36L)	128	0.67	1M	6.1	17.3	39.0
(6)	MoE(128E, 12L)	128	2.16	1M	1.9	5.4	36.7
*	T(96L)	2048	-	4M	~235.5	~42	36.9

- Advances in gating, tensor routing and load balancing to increase experts
  - Language agnostic work
- Shallower models with large number of experts are recommended
  - Avoid deep modeling issues partially or completely
  - Save 10 times or more training time compared to inefficient 96L models
  - ~7 BLEU improvement globally over deepest model

Bottlenecks: Hardware, Cost and Energy Efficiency

- **Needs heavy hardware ($$$)**
- Most existing work by Google with TPUs
  - TPUs are faster than GPUs
  - Typical to use up to 100s of TPUs
  - Equivalent GPU setting not yet known
- May not be possible to do this in a **university setting :(((((**

Bottlenecks: Hardware, Cost and Energy Efficiency

- **More devices and data = More training time (effective) = $$$$**
  - Are the BLEU gains worth the mammoth models?
  - Do BLEU gains translate into human evaluation gains?
  - **How do we deploy these models?**
    - **One model on a 2048 TPU cluster?**
  - How about continual learning?
- Future: Bigger models or better language aware neural modeling?
  - Model size (*representational capacity*) is no longer the bottleneck

Topics to address

- Parameter sharing
- Massively multilingual models
- Language divergence
- Training protocols

Language Divergence

- Vocabularies
  - Sufficient and fair representation for languages
- Lessons from visualization and language families
  - Understanding what is learned where
- Token based control
  - Augmenting input with features

On Vocabulary in MNMT

- Core point: Fair vocabulary representation for all languages
- Difficulty: Skew in data and hence word vocabulary
- Key Points:
  - Use large monolingual corpora
  - Oversample smaller vocabularies
  - Temperature sampling
    - $p_{VL}^{(1/i)}$ where **i** is the sampling temperature for vocabulary item **V** for language **L**
  - Adjust vocabulary sizes (32K to 128K sub-word vocabularies work well)
  - Trade-offs
    - *Sequence length, softmax computation time, enough sub-words per language*

Key Observations In Practice

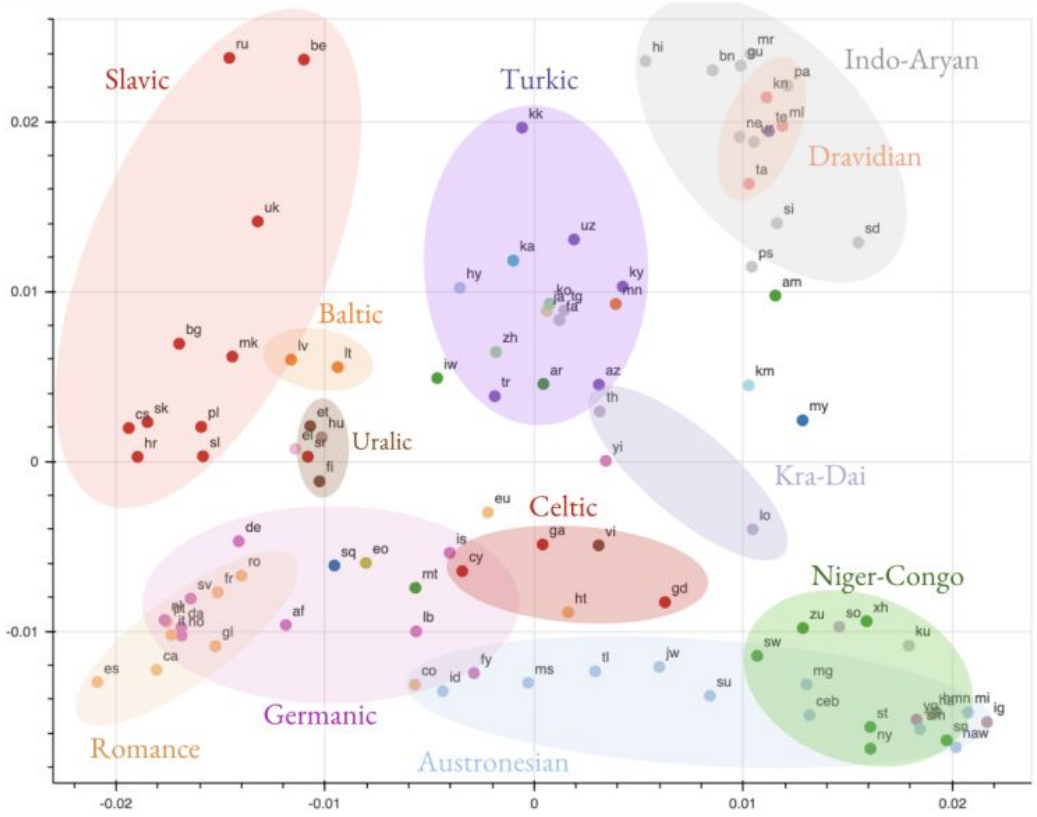
- Larger vocabularies do not always bring proportionate improvements
  - Larger vocabularies = Shorter sequences
  - Larger vocabularies = Slower softmaxes
- Do your cost-benefit analysis!

En→Any	High 25	Med. 52	Low 25
32k Vocab	27.69	16.84	12.90
64k Vocab	28.03	16.91	12.75
Any→En	High 25	Med. 52	Low 25
32k Vocab	33.24	29.40	26.18
64k Vocab	33.85	30.25	26.96

- (Almost) equal representation is important
- Moderate temperature sampling with fixed vocabulary size is crucial
- Is T=5 a golden rule?
  - Arivazhagan et al.; Bapna et al. 2019

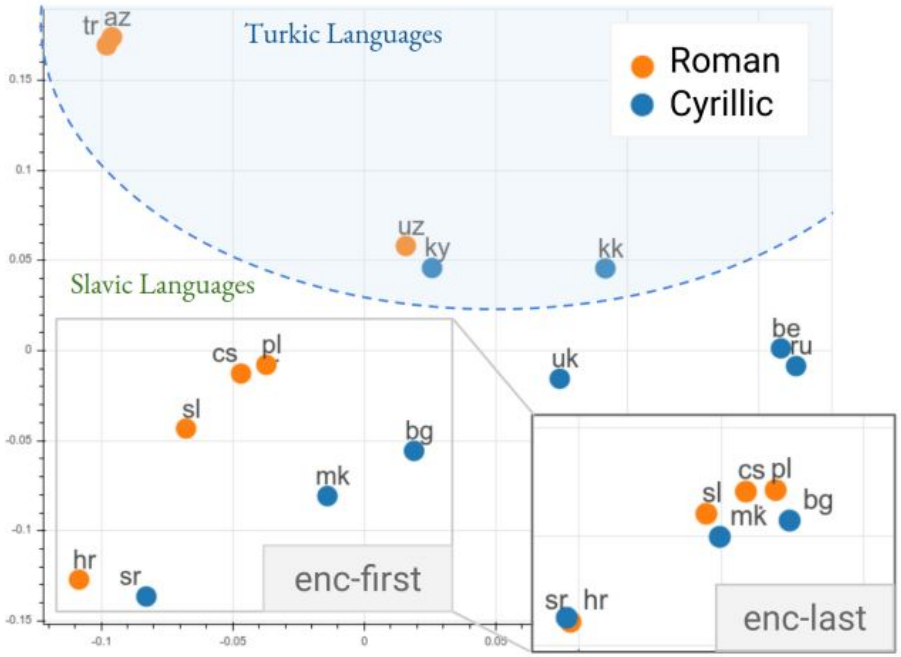
Any→En	High 25	Med. 52	Low 25
$T_V = 1$	33.82	29.78	26.27
$T_V = 100$	33.70	30.15	26.91
$T_V = 5$	33.85	30.25	26.96

Visualization Of MNMT Representations



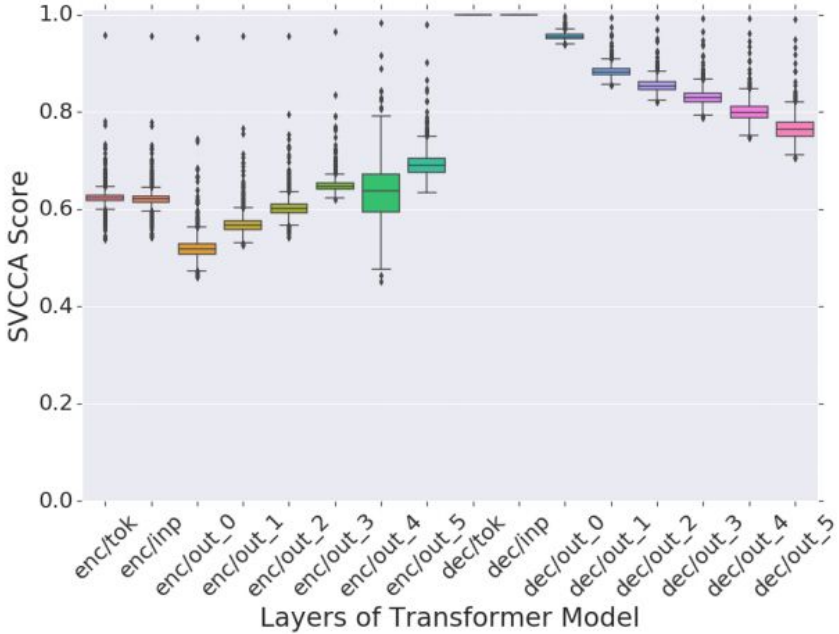
- SVCCA similarity between representations (Kuduganta et al. 2019)
  - Also see Dabre et al. and Johnson et al. 2017
- Encoder representations cluster sentences into language families
  - Regardless of script sharing
  - Script sharing for stronger clustering
- High resource languages cause partition
  - Low-resource languages ride the wave
- Evidence of representation invariance when fine-tuning
  - Explains poor zero shot quality between distant pairs

Representation Similarity Evolution

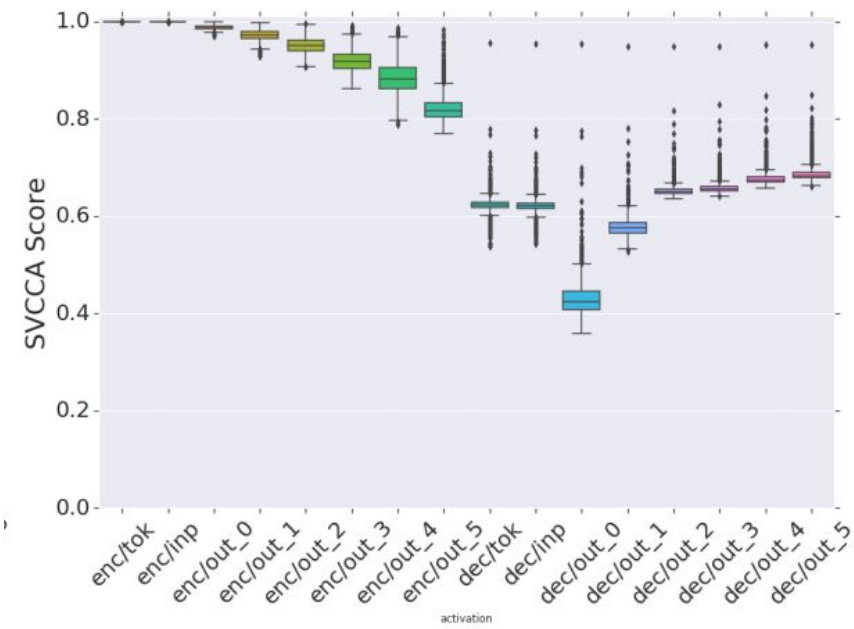


- Representation similarity varies with depth
  - Shallower layers cluster by script
  - Deeper layers cluster by family
- Case study: Turkic and Slavic
  - Some use roman script
    - Turkish, Uzbeki and Azerbaijani
  - Some use cyrillic script
    - Kyrgyz and Kazakh
  - First encoder layer separates by script and last layer closes the gap
    - Distinguishes from slavic languages
    - Such languages use roman or cyrillic

Representation Evolution With Depth



(a) X-En Language Pairs



(b) En-X Language Pairs

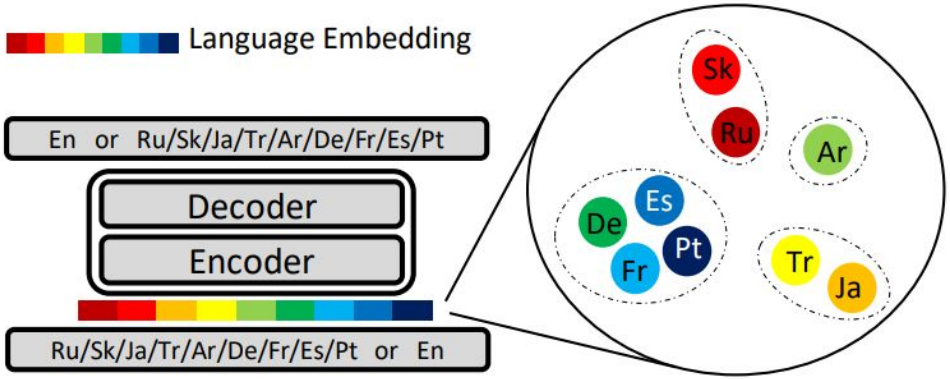
For Many to English: Encoder representations converge and decoder representations diverge

For English to Many: Encoder representations of English diverge based on target language

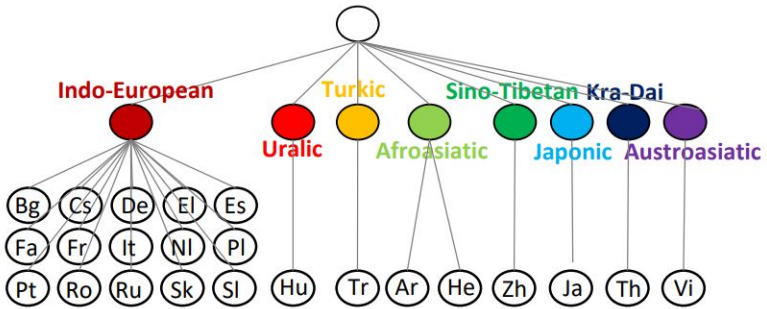
Question: Is divergence a good thing? Does it cause a learning overhead? Should it?

Empirically Determined Language Families

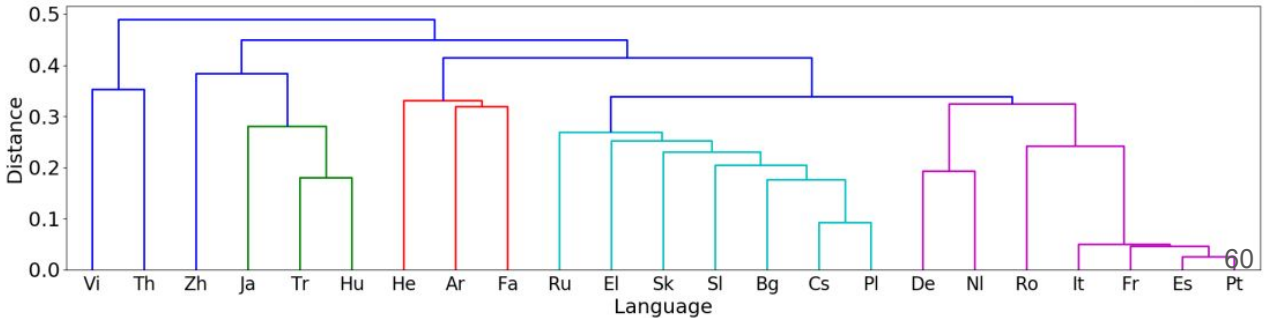
- Train many-to-many model with language tokens
- Hierarchical clustering of tokens
  - Set number of clusters by elbow-sampling
- **Tan et al. 2019**
- **Also see Oncevay et al. 2020**



Predetermined language families



Empirically determined language families via embedding clustering



Is There An Optimal Number of Languages

- Does empirical clustering help? (Upper table)
  - **Mostly yes**
  - Random clustering gives poorer results
  - **Predetermined clustering is equally good**
- Language family specific models (Bottom table)
  - **Universal model < Individual models**
  - **Family specific model > Individual models**
  - Related to observations by Dabre et al. 2017 and 2018
- Next steps
  - Family specific adaptor layers (Bapna et al. 2019)
  - Family specific vocabulary and decoder separation
  - Behavior in extremely low-resource settings (<20k pairs; Dabre et al, 2019)

Language	Ar	Bg	Cs	De
Random	22.90	32.18	28.88	30.67
Family	25.02	**32.75**	30.27	31.09
Embedding	**25.27**	32.52	**30.97**	**31.33**

Language	Ar	Bg	Cs	De
Data size	180K	140K	110K	180K
Individual (23)	**25.43**	**32.87**	29.15	**32.18**
Universal (1)	23.26	32.47	28.86	30.42
Embedding (7)	25.27	32.52	**30.97**	31.33

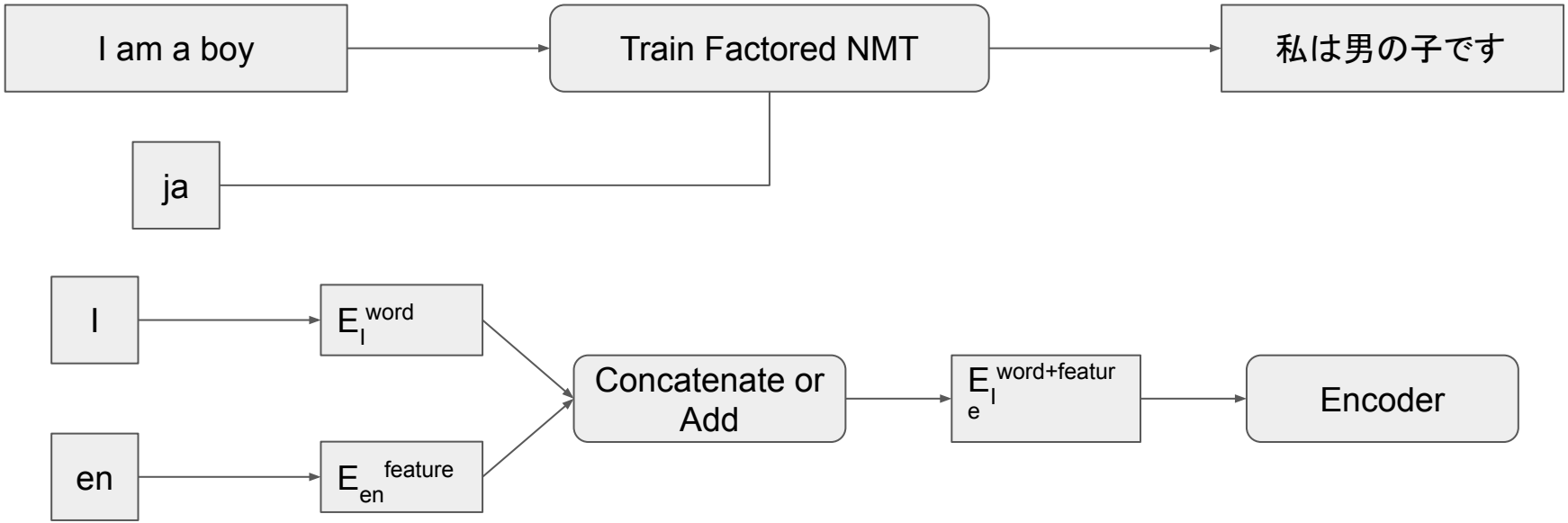
On Language Tags: Embeddings vs Features

Johnson et al. 2017: prime the encoder's output using a <2xx> token.
<2xx> is a single token with its embedding. Black box approach..

Ha et al. 2016: distinguish between shared vocabulary units. Encoder is primed with source as well as target language information.

Blackwood et al. 2019: Use language tokens at beginning and end.

On Language Tags: Embeddings vs Features



Ha et al. 2017 and Hokamp et al. 2019: keep word embeddings independent of task (target language) via features

Topics to address

- Parameter sharing
- Massively multilingual models
- Language divergence
- Training protocols

Training Protocols

- MNMT training fundamentals
- Training schedules
  - Batching strategies
  - Importance of sampling
- Leveraging bilingual models
  - Distillation
- Model expansion and incremental learning

On MNMT Training

- Fundamentally the same as standard NMT: Minimizing negative log likelihood

$$L_\theta = \frac{1}{L}\sum_{l=1}^{N} L^{src_l - tgt_l}(\theta)$$

  Where the individual language pair negative log-likelihood is $L^{src_l - tgt_l}(\theta)$

- Challenges:
  - Good training schedule
  - Language equality

Training Schedule: Joint Training

- One pair at a time (Firat et al. 2016, Dong et al. 2015)
  - Cycle through corpora (L_1-L_2 →L_3-L_4→L_5-L_6→.....→L_1-L_2)
- Useful for models with *separate encoders and/or decoders*
- Potential forgetting of language pair information in a previous batch
  - *Catastrophic forgetting!*

Training Schedule: Joint Training

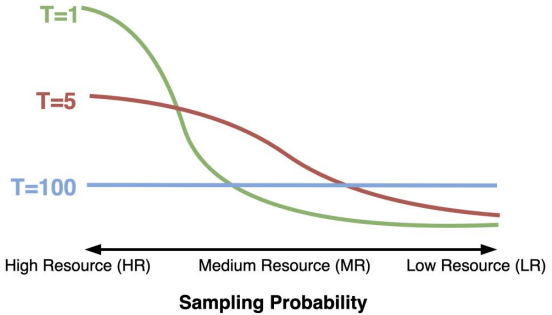
- Mixed language pairs batch (Johnson et al 2017)
  - Mix all corpora, shuffle and then choose batches
- Useful for fully shared models
- For models with separate language encoders/decoders
  - *Shard batch and feed to appropriate components*

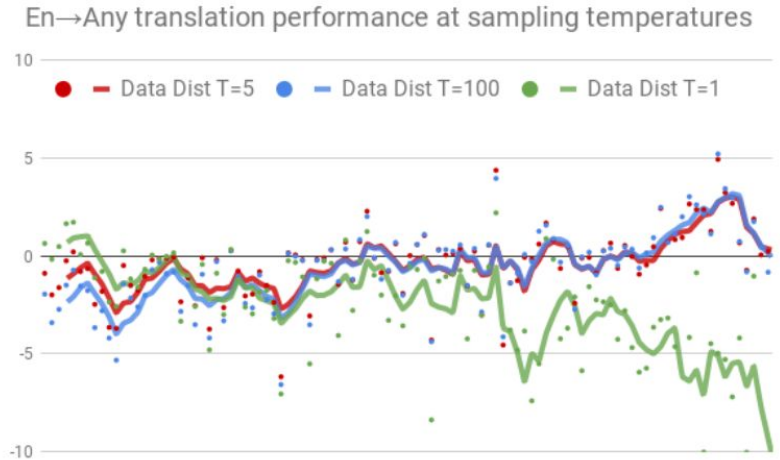
Training Schedule: Addressing Language Equality

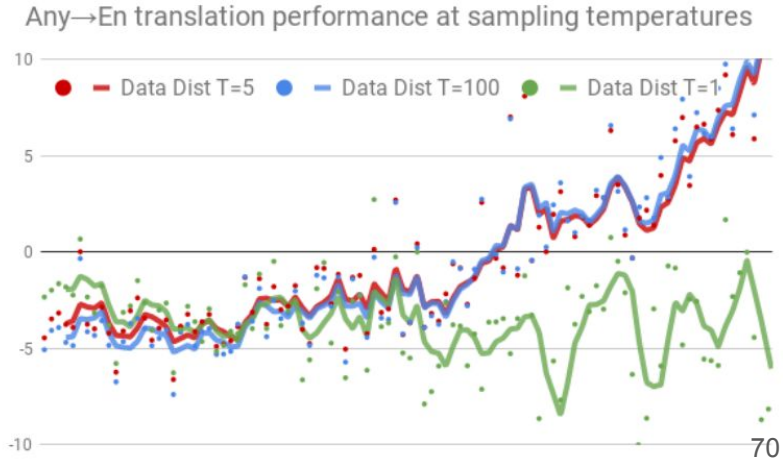
- Source of inequality: Corpora size skew
- Solutions: Oversampling smaller corpora
- Oversampling **before training or during training**?
  - Matter of implementation choice
  - Oversampling prior to training creates large duplicated corpora

Importance Of Temperature Based Sampling (Arivazhagan et al. 2019)

- Naive approaches:
  - Ignore corpora size distributions
  - Sample from all corpora equally
- New approach: Temperature based sampling ($p_L^{(1/i)}$)
  - Where p_L is the probability of sampling a sentence from a corpus
  - i is the sampling temperature
  - Strongly benefits low-resource pairs

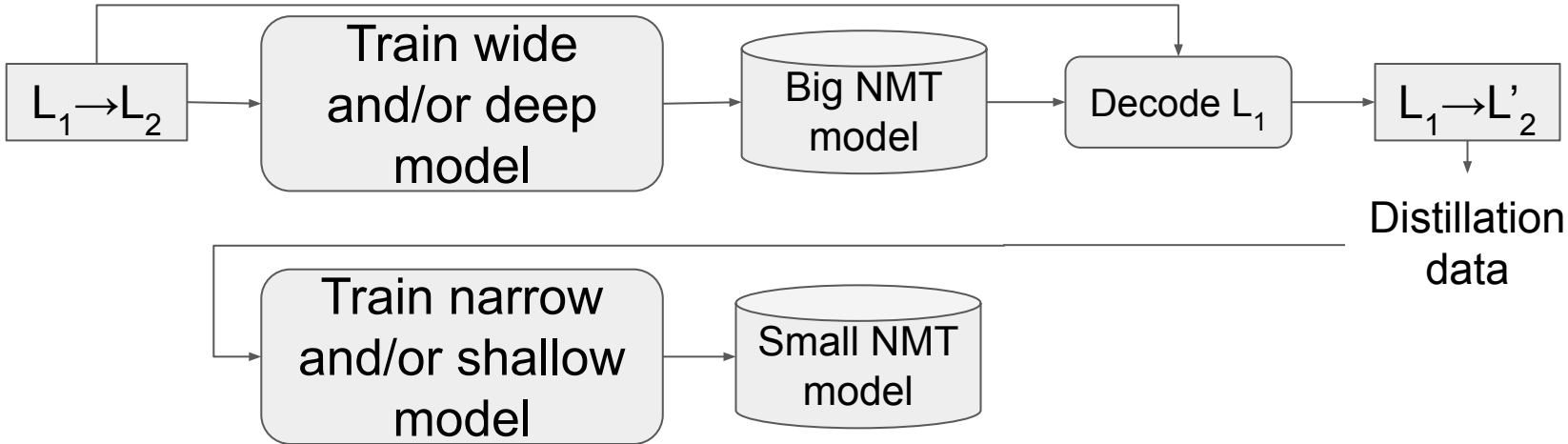




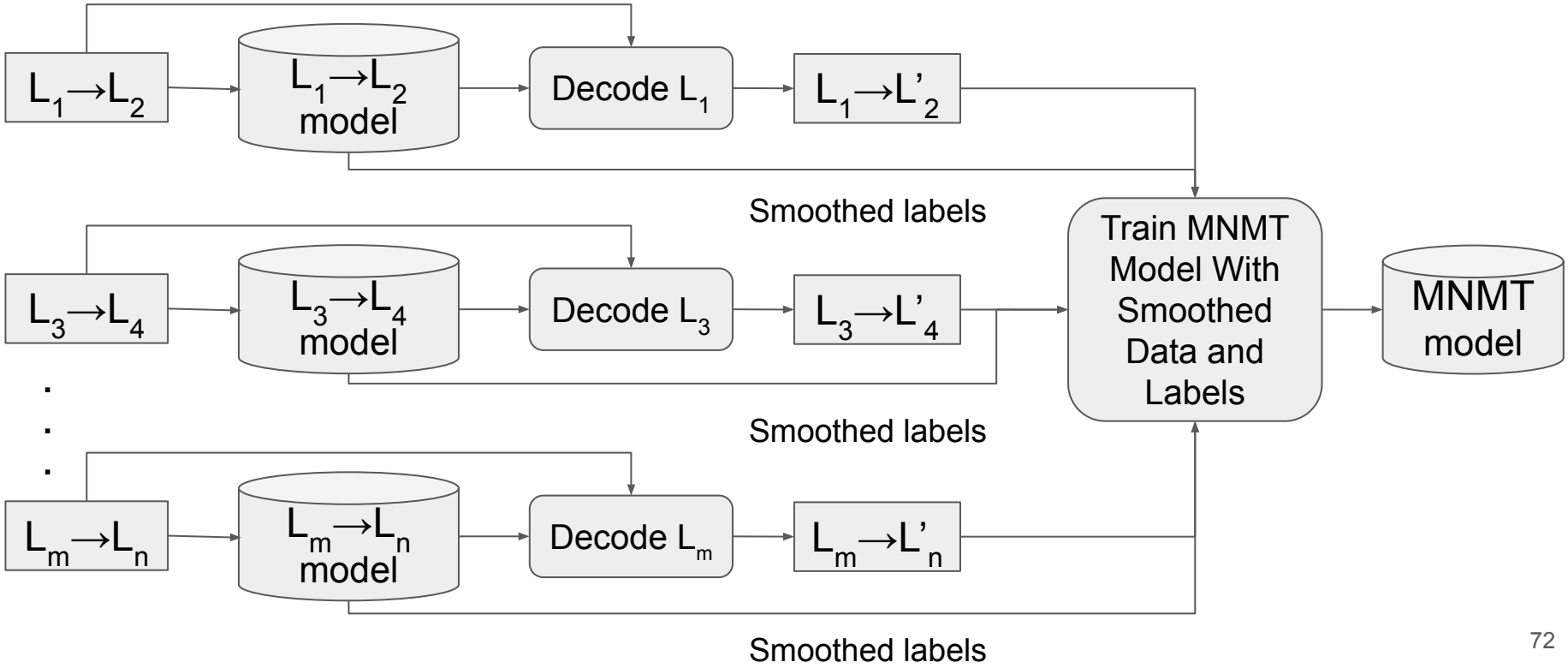


Leveraging Bilingual Models

- Training from scratch is challenging
  - Multitude of pairs
  - Complexity of task
  - Interference of languages
- Solution: Leverage bilingual models or smaller multilingual models
  - Key principle: Transfer learning via sequence knowledge distillation (Kim et al. 2016)



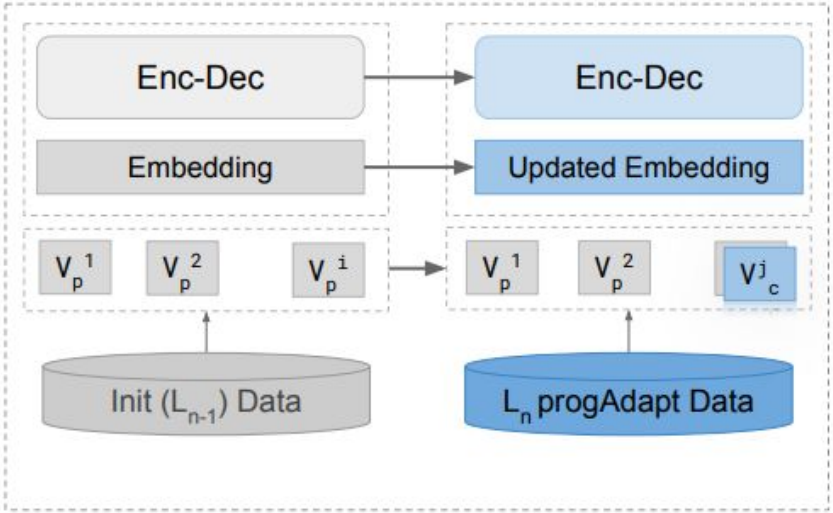
Distillation For MNMT Training (Tan et al. 2019)

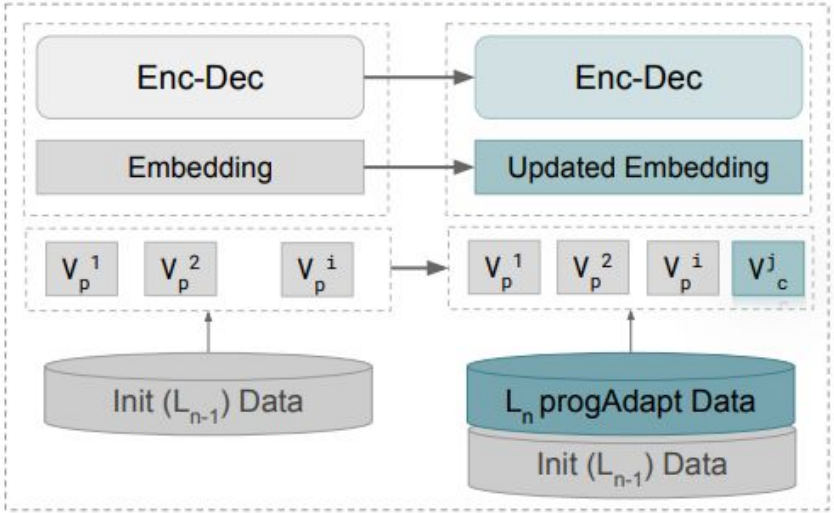


Incremental Training

- Objective: Expand the capacity of existing models with maximum reusability
- Promising approaches
  - **Vocabulary expansion**
    - Surafel et al. 2018
  - **Gradual capacity expansion**
    - Escolano et al. 2019
  - **Adaptor layers (experts) on top of pre-trained models**
    - Bapna et al. 2019 (already discussed)

Incorporating New Languages (Surafel et al. 2018)





- Language specific transfer
  - **Replace vocabulary**
  - Fine-tune on new data
- Similar to *Zoph et al. 2016*

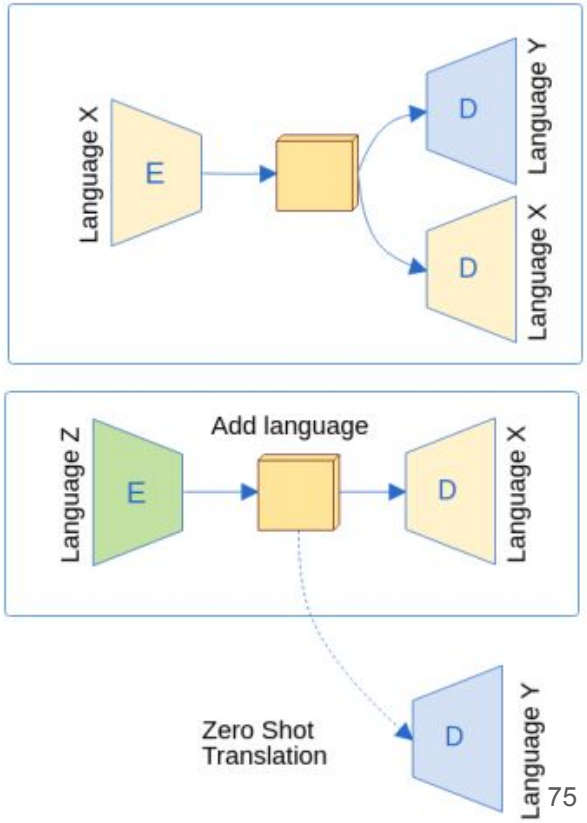
- Expanding to new languages
  - **Expand vocabulary**
  - Fine-tune on old+new data
- *Increase computational capacity?*

Capacity Expansion Of Existing Models (Escolano et al. 2019)

- Add new components while freezing existing components
  - Lightweight training BUT
  - Previous components may not be aware of new languages
    - Poor transfer learning
  - Potential zero-shot learning
    - Will it work for distant languages

System	ES-EN	EN-ES	FR-EN	DE-EN
Baseline	32.60	32.90	31.81	28.96
Joint	29.70	30.74	-	-
Added lang	-	-	30.93	27.63

- Lessons from Sachan et al. 2018; Firat et al. 2016a/b; Bapna et al. 2019
  - Deepen encoders and decoders
  - Only train new components with old and/or new data
    - Vocabulary expansion by Surafel et al. 2018 will help



Correlated Topics Not Well Addressed Yet

- Diverse language learning rates
  - Language pairs are learned at different rates
- Jean et al. 2019 set task weights using model performance
  - Weights decide sampling (task importance)
- Lessons from Wang et al. 2018
  - Adding or removing instances from training set

Correlated Topics Not Well Addressed Yet

- Catastrophic forgetting
  - Thompson et al. 2019 on domain adaptation
  - Addressing forgetting for incremental learning?
- MNMT model convergence
  - Currently report average performance over all pairs
    - I'm looking at you GShard
  - Ignoring individual pair's performance is unwise
  - Weighted performance metric to the rescue?
  - Better evaluation metrics for MNMT settings?

Outline of This Tutorial

- Overview of Multilingual NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

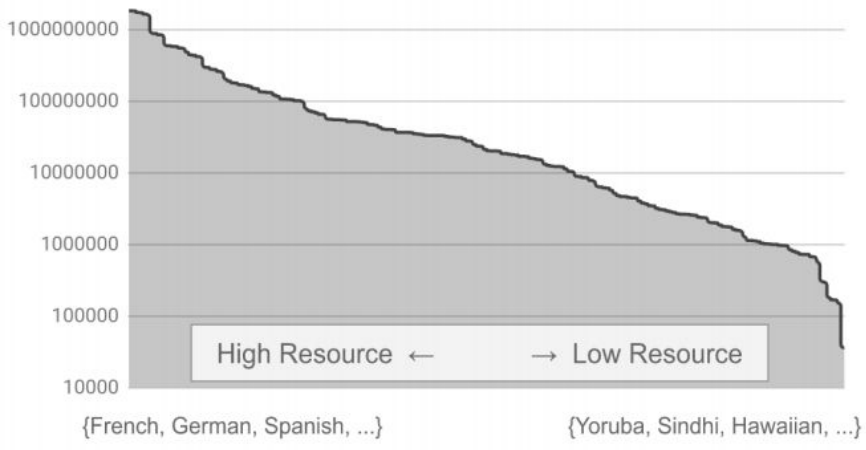
Self Introduction (Anoop Kunchukuttan)

- Experience
  - 2018-present: Senior Applied Researcher, MT Group, Microsoft India
  - 2012-2017: Ph.D. scholar at IIT Bombay, India
  - 2008-2011: ML/NLP Lead, Life Sciences Group, Persistent Systems
  - 2006-2008: M.Tech. IIT Bombay, India
- Research Interests
  - Multilingual NLP
  - Machine Translation, Transliteration
  - Representation Learning for NLP
  - Indian language NLP

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Translation for low-resource languages



(Arivazhagan et al., 2019b)

A large skew in parallel corpus availability

Long tail of low-resource languages

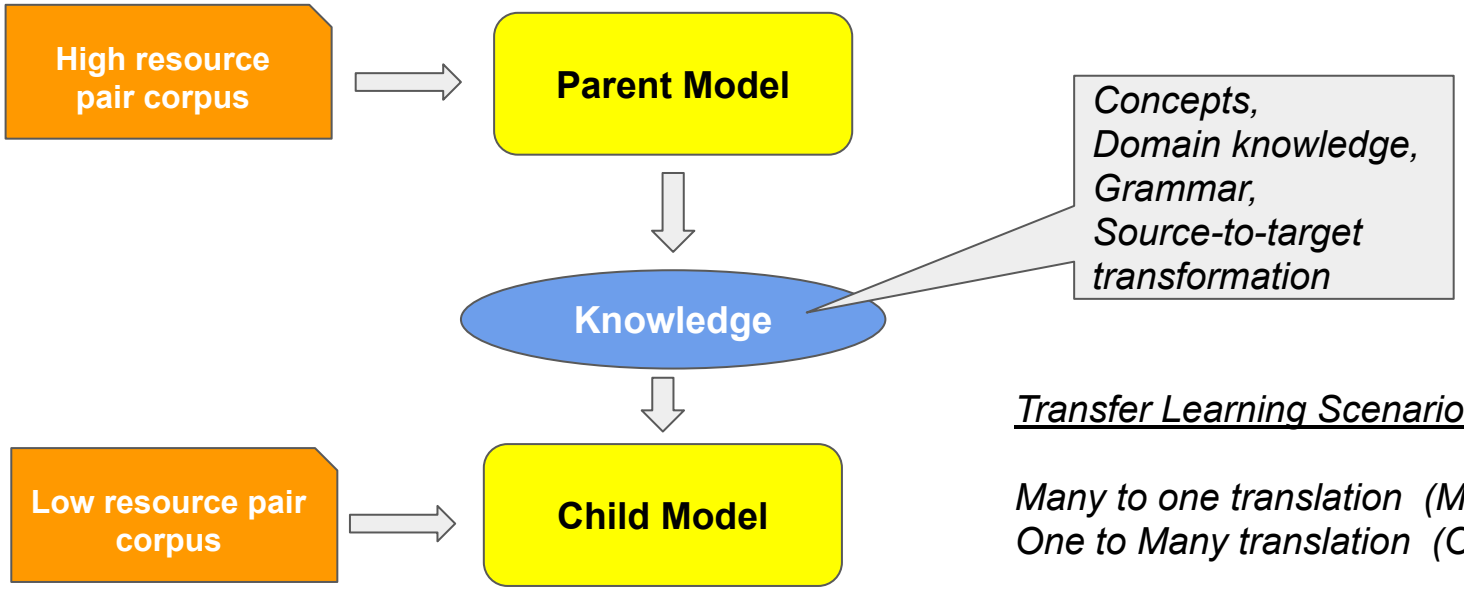
Difficult to obtain corpora for many languages

Can high-resource languages help low-resource languages?

Transfer learning

Storing knowledge gained while solving one problem & applying it to a different but related problem.

(Pan & Yang, 2010)



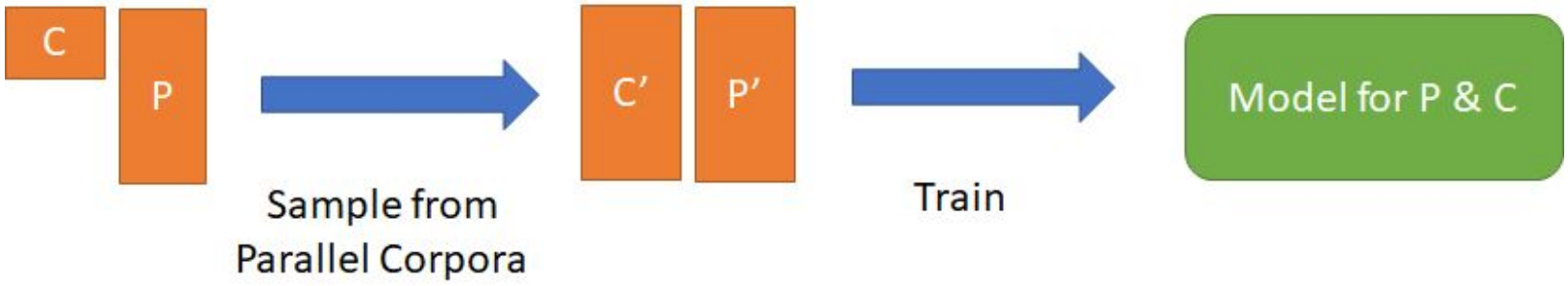
Transfer Learning Scenarios

Many to one translation (M2O)
One to Many translation (O2M)

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Joint Training *(Ha et al., 2016; Johnson et al, 2017)*



Oversample child/undersample parent to balance training

Low-resource languages performance

Many-to-One direction ⇒ major gains

One-to-Many direction ⇒ minor gains

Fine-tuning *(Zoph et al., 2016; Tan et al., 2019b)*

- *Fine-tuning vs. Joint Training*
  - *Fine-tuning better in O2M setting and vice-versa*
- *While fine-tuning, specific set of parameters can be tuned*
  - *Many to one ⇒ Encoder bottom layers*
  - *One to many⇒ Decoder top-layers*

Small low-resource corpus ⇒ *overfitting might occur* ⇒ *Use mixed-finetuning*

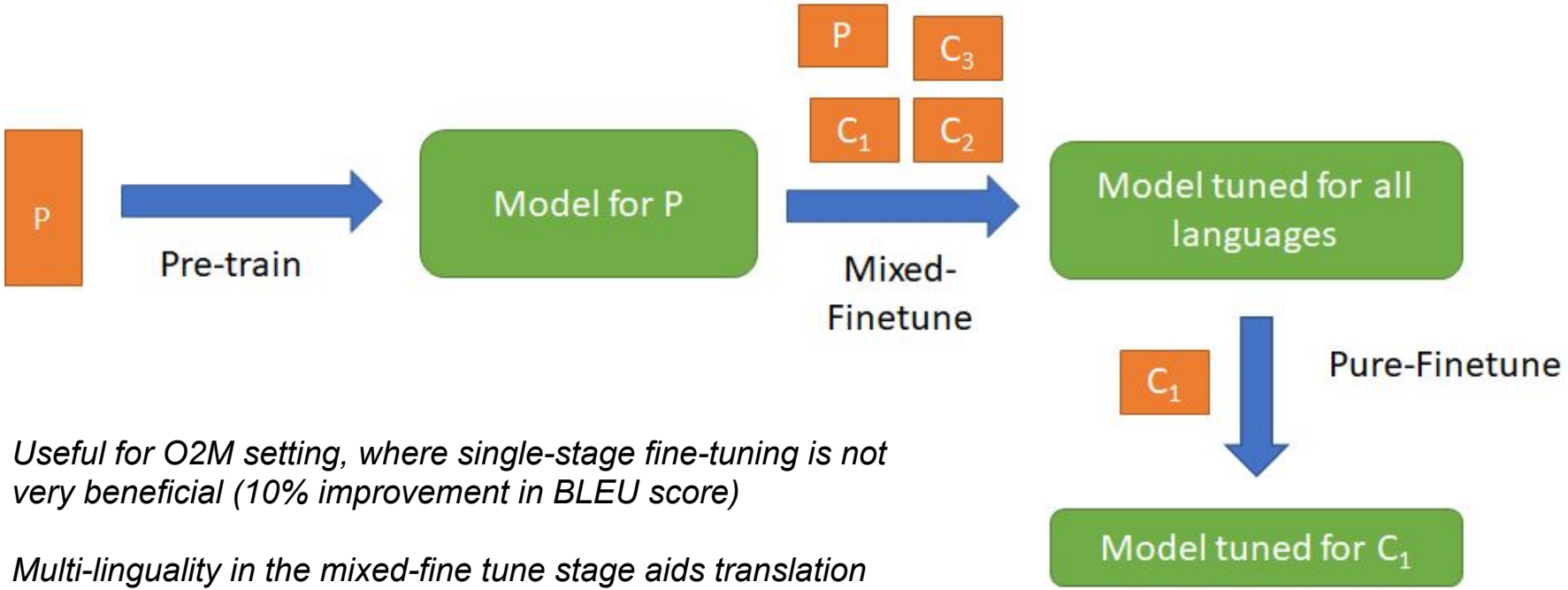
(Dabre et al., 2019)

Transfer from multiple parents *(Neubig and Hu et al., 2018)*

Pre-train a multilingual NMT model on a representative set of high-resource languages

Useful for rapid-adaptation to new languages

Transfer to multiple children *(Dabre et al., 2019)*



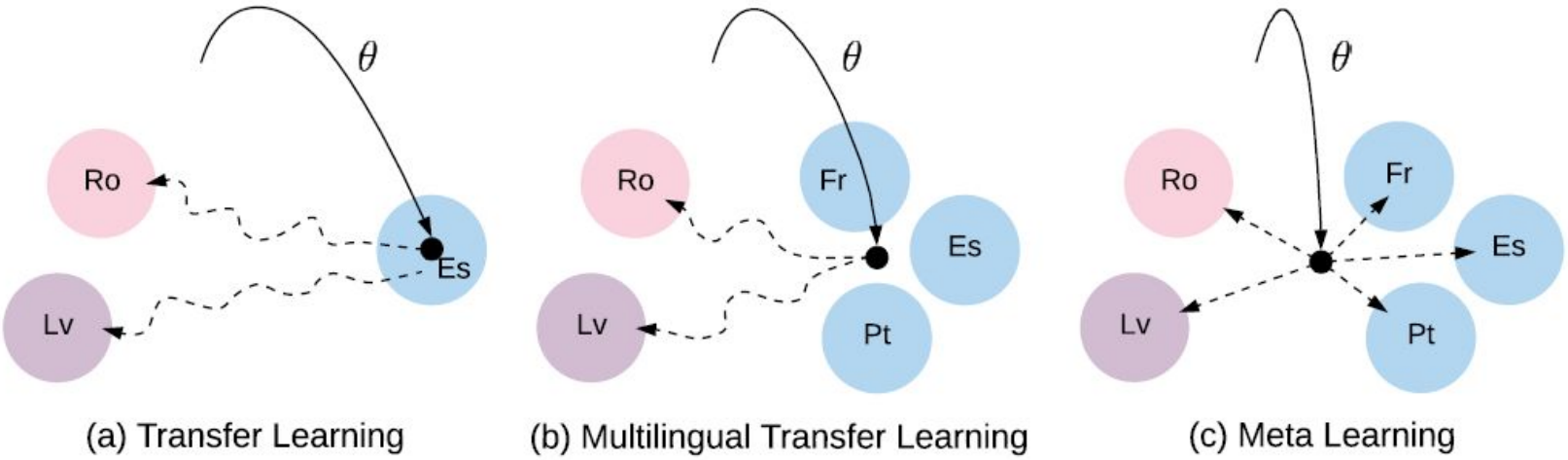
Useful for O2M setting, where single-stage fine-tuning is not very beneficial (10% improvement in BLEU score)

Multi-linguality in the mixed-fine tune stage aids translation

What is the objective of the parent-model training?

- *Optimize performance on parent tasks?*
- *Optimize performance on child tasks?*
- *Enable few-shot learning?*

Meta-learning *(Finn et al., 2017)*

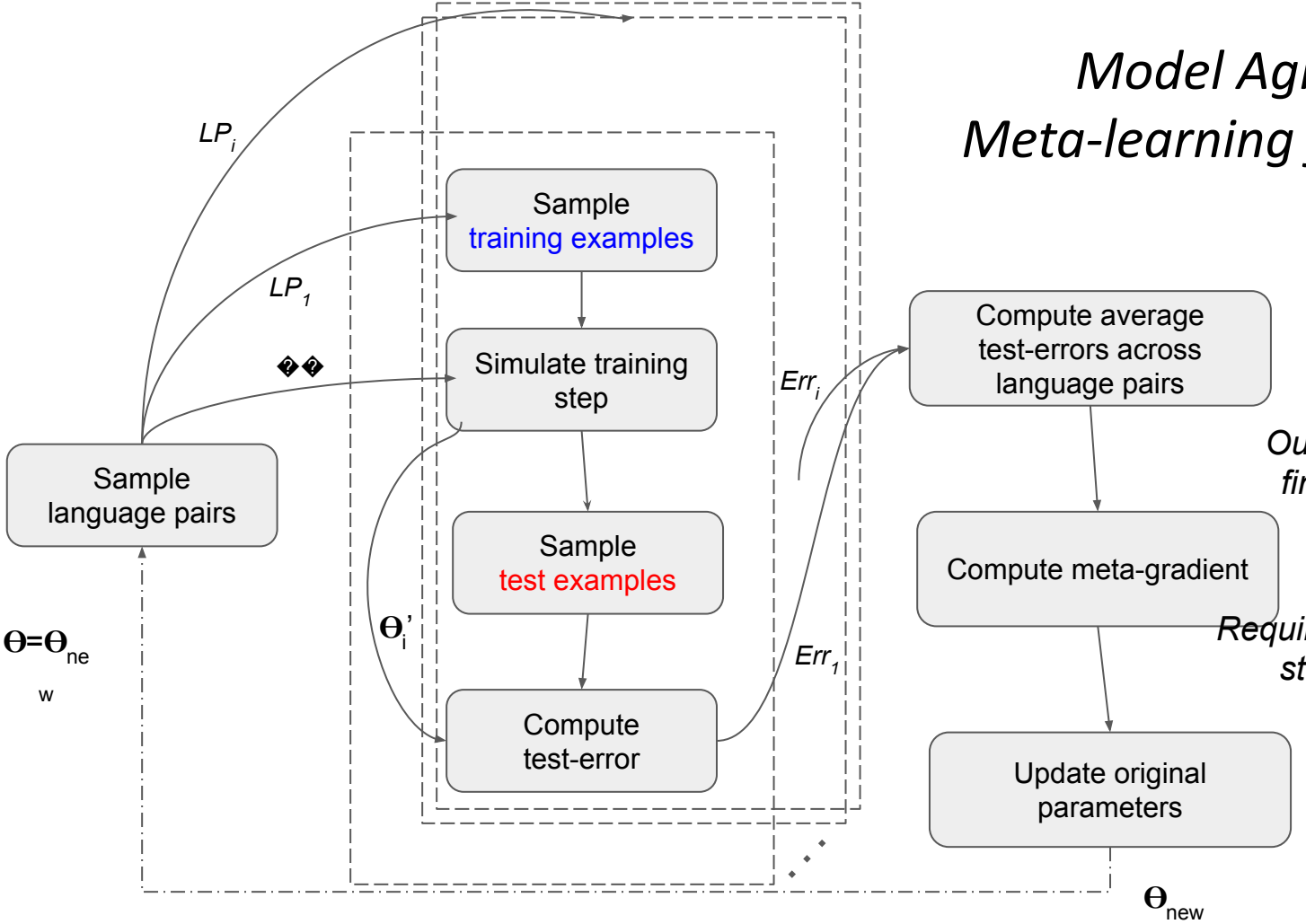


Learning to learn

Learn an initialization from which only a few examples are required to learn the child-task

Model Agnostic Meta-learning for MNMT

(Gu et al., 2018b)



Outperforms multilingual fine-tuning strategy on unseen pairs

Requires far fewer adaptation steps for comparable performance

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

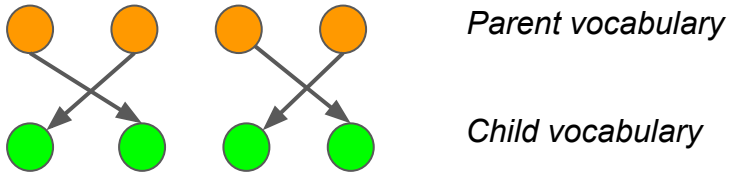
Lexical Transfer

Initialize child token embeddings prior to fine-tuning

How do we initialize child token embeddings?

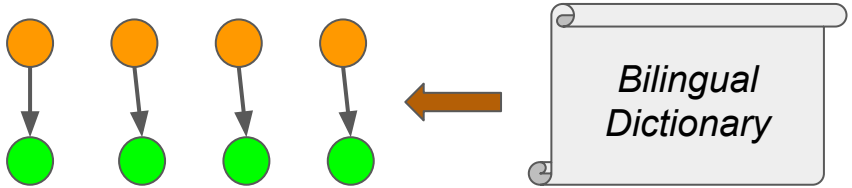
Random Assignment

(Zoph et al., 2016)



Dictionary Initialization

(Zoph et al., 2016)



Dictionary Init faster than random assign, but random init eventually achieves comparable performance

Use bilingual embeddings to map parent and child embeddings to a common space

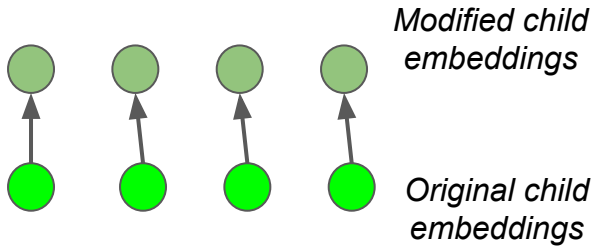
Linear mapping functions can be learnt using small bilingual dictionaries

$$\sum_{(f,f')\in S} \|W E^{\text{mono}}_{\text{child}}(f) - E^{\text{src}}_{\text{parent}}(f')\|_2$$

Map child embeddings to parent embeddings

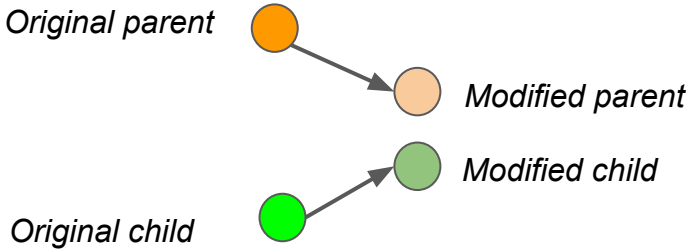
$$E^{\text{src}}_{\text{parent}} \leftarrow W E^{\text{mono}}_{\text{child}}$$

(Kim et al., 2019a)



Map parent and child embeddings to a common space

(Gu et al., 2018a)



Significant improvements over random assignment of embeddings

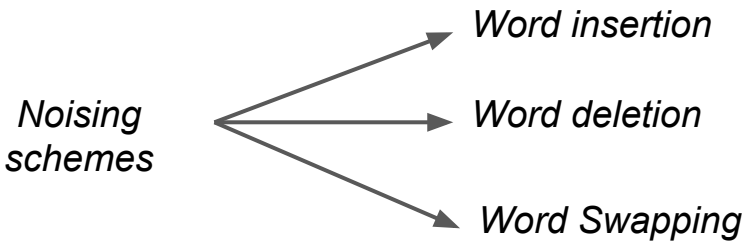
Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

What if parent and child have different word orders?

Introduce noise in parent-source sentences *(Kim et al., 2019a)*

Prevents over-optimization to parent-source language

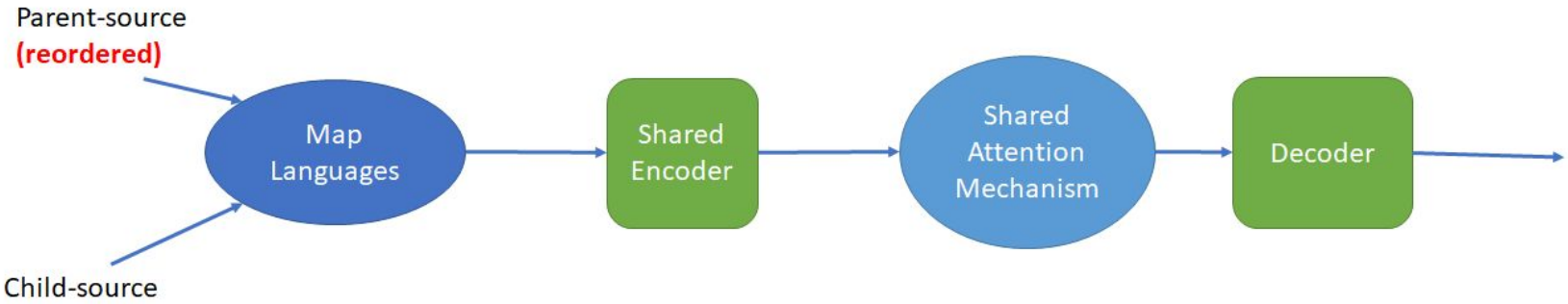


Simple methods, gives modest improvements over baseline finetuning

Reorder parent-source sentence to match child-source sentence

(Murthy et al., 2019)

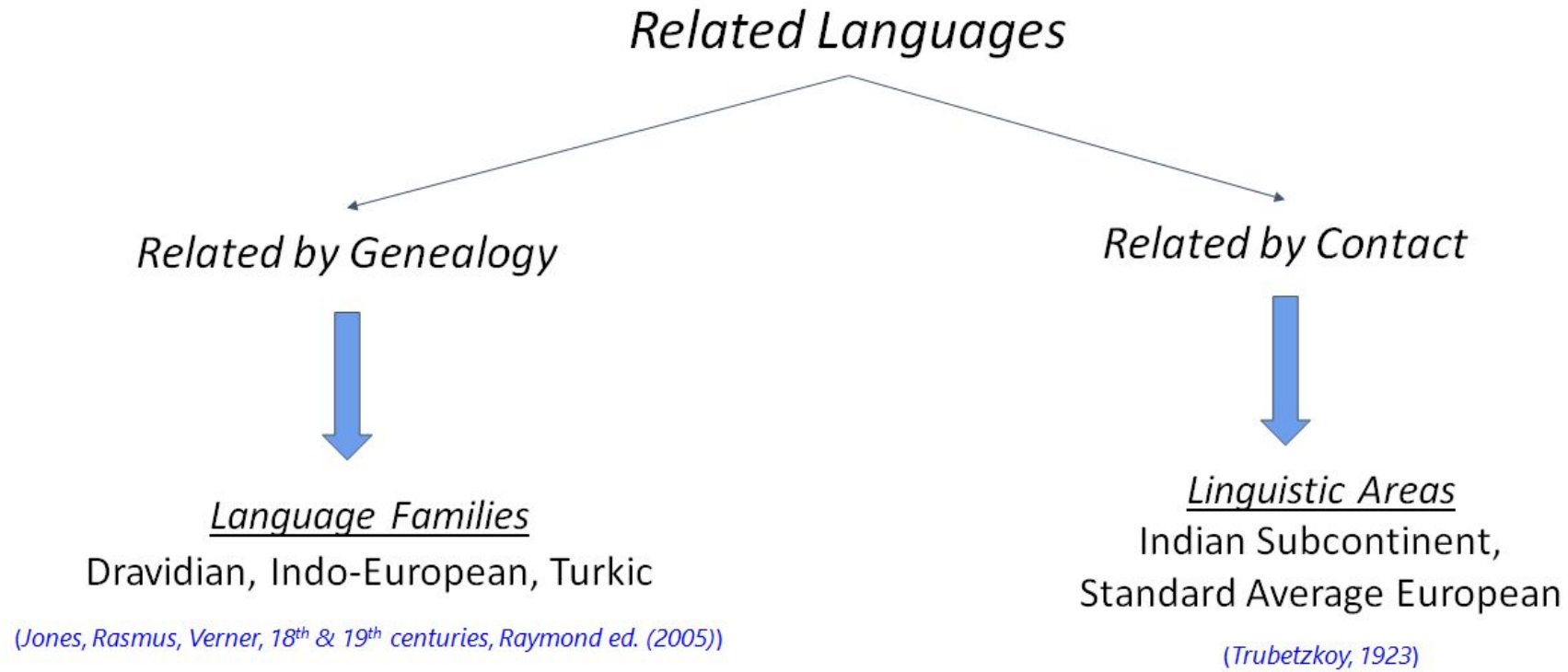
Ensures better alignment of encoder contextual embeddings



Significant improvements over baseline finetuning, but needs a parser and re-ordering system

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT


Related Languages
Related by Genealogy
Related by Contact
Language Families
Dravidian, Indo-European, Turkic
(Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005))
Linguistic Areas
Indian Subcontinent,
Standard Average European
(Trubetzkoy, 1923)

Key Similarities between related languages

(Kunchukuttan & Bhattacharyya., 2020)

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

Marathi

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

Hindi

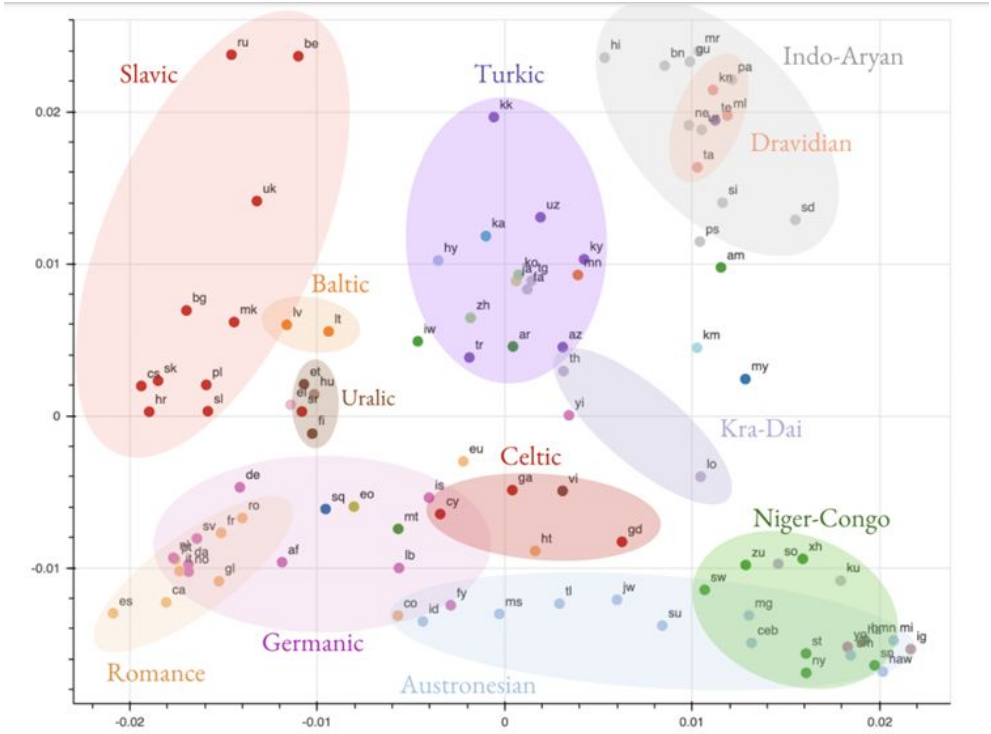
Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Transfer Learning works best for related languages

(Zoph et al., 2016; Dabre et al, 2017b)



(Kudungta et al., 2019)

Encoder Representations cluster by language family

Contact relationship is also captured

Related languages using different scripts also cluster together

Utilizing lexical similarity *(Nguyen and Chiang, 2017)*

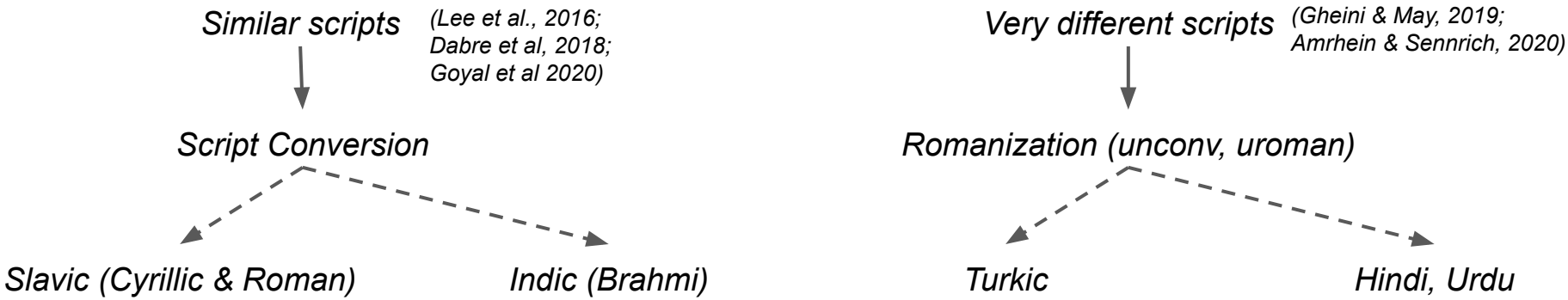
Subword-level vocabulary improves transfer

Improves parent-child vocabulary overlap encouraging better transfer

child	subword	BLEU score		parent-child overlap (%)	
		baseline	transfer	train	dev
tur-eng	word	8.1	8.5	3.9	3.6
	BPE	12.4	**13.2**	58.8	25.0
uyg-eng	word	8.5	10.6	0.5	1.7
	BPE	11.1	**15.4**	57.2	48.5

(uzb-eng is the parent language pair)

Transfer between related languages using different scripts works well



Transfer works without script conversion → but script conversion provides improvements

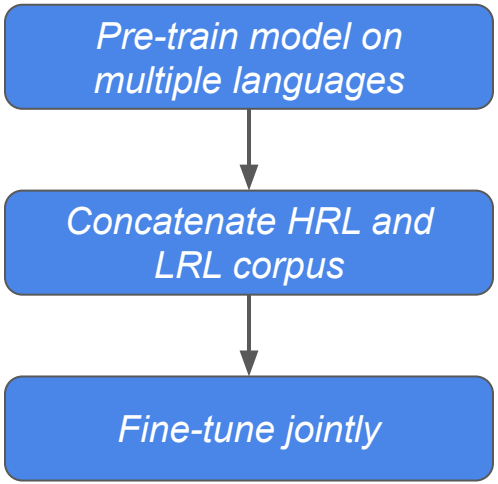
Transfer can also be done between languages related by contact (Goyal et al. 2020)

Dravidian and Indo-Aryan languages form a linguistic area in the Indian subcontinent (Emeneau, 1956)

Similar language regularization

(Neubig & Hu, 2018; Chaudhary et al, 2019)

Very small low resource language ⇒ Overfitting on finetuning



Joint fine-tuning outperforms just target language fine-tuning

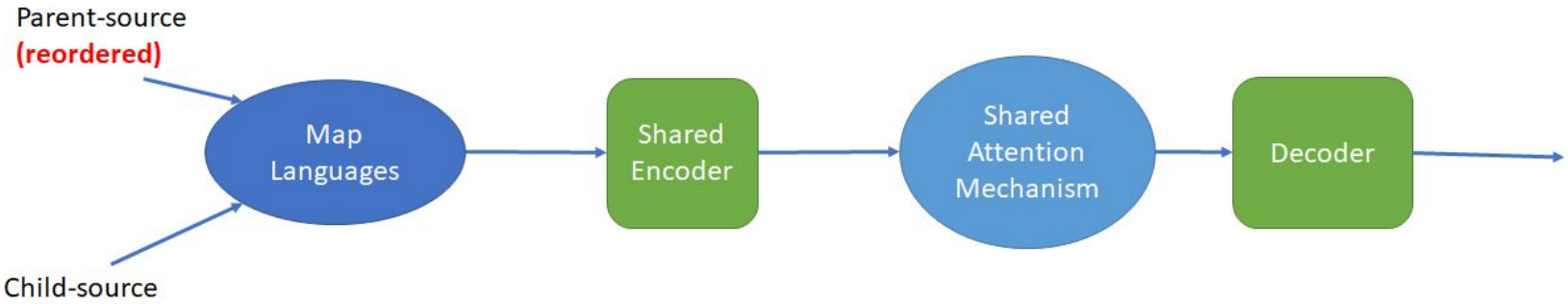
Pre-training with multiple languages is better than single language

Similar idea is also used for knowledge distillation *(Dabre & Fujita et al, 2020)*

Utilizing Syntactic Similarity

Reorder parent-source sentence to match child-source sentence

Significant improvements over baseline finetuning, but needs a parser and re-ordering system



Reordering rules can be reused if the child-source languages have the same word order

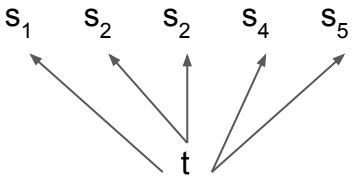
(Murthy et al., 2019)

Parent Data Selection (Wang et al., 2019)

Which examples in the parent language-pair are most helpful for transfer?

Let us look at the case of Many-to-one translation

(s_i,t) is parent sentence pair from (H,E)



Score s_i by the probability that is belongs to low-resource source language L

Scoring Functions

- *score(s_i,L) ∝ vocab-overlap of s_i with L*
- *score(s_i,L) ∝ $P_{LM\text{-}L}(s_i)$*

Sample examples using this score

- *Can be extended to multiple parent languages*
- *Can be extended to language-level similarity score*

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Pivot Translation

aaja bahuta ThaNDa hai

hi-en model

It is very cold today

en-ml model

inn vaLar.e taNuppAN

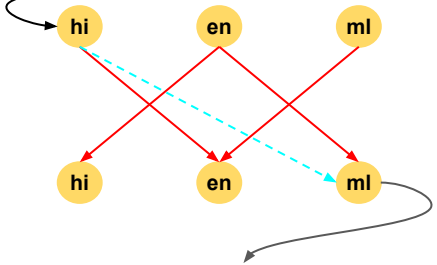
Multiple decoding steps

Multiple translation systems

Zero-shot Translation

(Johnson et al., 2016)

aaja bahuta ThaNDa hai



inn vaLar.e taNuppAN

Zero-resource Translation

(Firat et al., 2016b)

many-2-many model

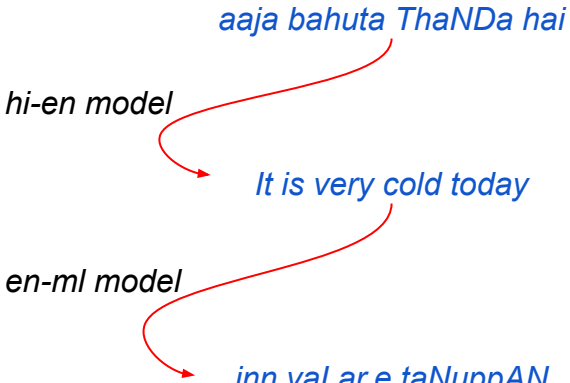
Single decoding step

Single translation system

Section Overview

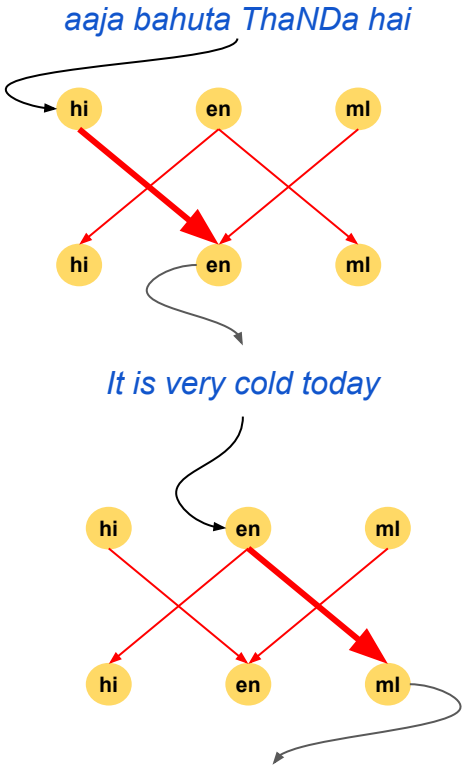
- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Bilingual Pivot



Multilingual Pivot

(Johnson et al., 2016)



Multilingual pivot generally outperforms bilingual pivot *(Firat et al., 2016)*

Pivot translation is a strong baseline

Limitations:

- *Cascading errors*
- *Decode time* *(function of path length)*

Reduce cascading errors using n-best translations

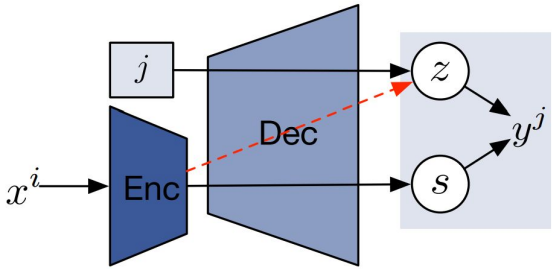
Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Naive Zeroshot-performance significantly lags behind pivot translation

(Arivazhagan et al., 2019a; Gu et al., 2019)

	de→fr	fr→de
Pivot	26.25	20.18
Zeroshot	16.80	12.03



Output is generated in the wrong language

	en	de	fr
$de \rightarrow fr$	14%	25%	60%
$fr \rightarrow de$	12%	54%	34%

(Results from Arivazhagan et al., 2019a)

- Spurious correlation between source input & target language
- Model always associates english output with any input
- Copying behaviour

Performance gap reduces when evaluation restricted to correct language output

	de→fr	fr→de
Pivot	19.71	24.33
Zeroshot	19.22	21.63

Code-mixing is rare

Once a wrong language token is generated, all subsequent tokens are generated in that language

Vocabulary construction and control

Restrict decoder to output only target-language vocabulary items

(Ha et al., 2017)

	German-Dutch	German-Romanian
baseline	14.95	10.83
+vocab filter	**16.02**	**11.00**

Language-specific subword model learning *(Rios et al., 2020)*

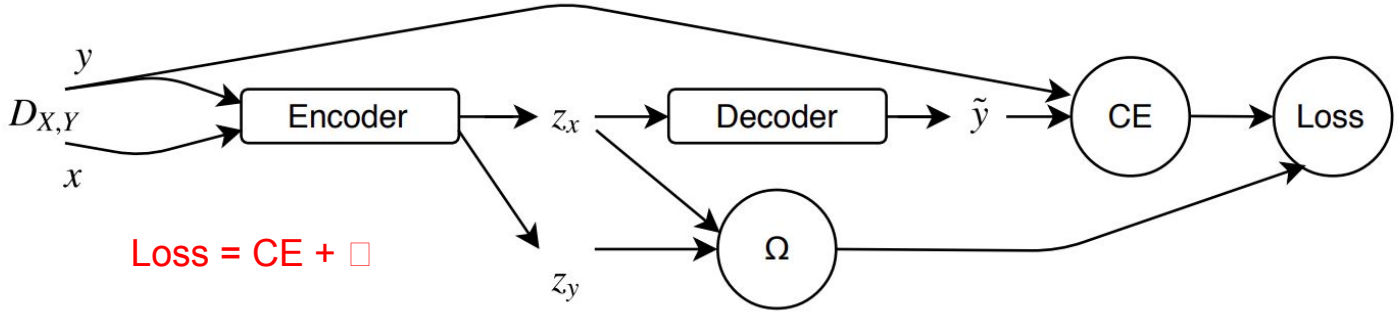
\+ *overlapping model vocab*

Reduces copying behaviour and bias towards English output

Joint vocab	15.4
Language-specific vocab	**20.5**

Prevents generation of wrong language, perfomance still lags pivot baseline

Minimize divergence between encoder representations



□*(z_x,z_y): distance between encoder representations of source (x) and target (y)*

Supervised Objectives

- Cosine distance (Arivazhagan et al., 2019a)
- Euclidean distance (Pham et al., 2019)
- Correlation distance (Saha et al., 2016)

$$\sum_{i=1}^{N} s(h_X(x_i))s(h_Z(z_i))^T$$

Unsupervised Distribution Matching
(Arivazhagan et al., 2019a)

Use a domain-adversarial loss

Competitive with pivot and improves over baseline MNMT

Avoid pooling to generate encoder representations

(Pham et al., 2019)

Encoder output is variable length

Decoder receives fixed input at every timestep

Minimize divergence between auto-encoded and true target at different points in the decoder

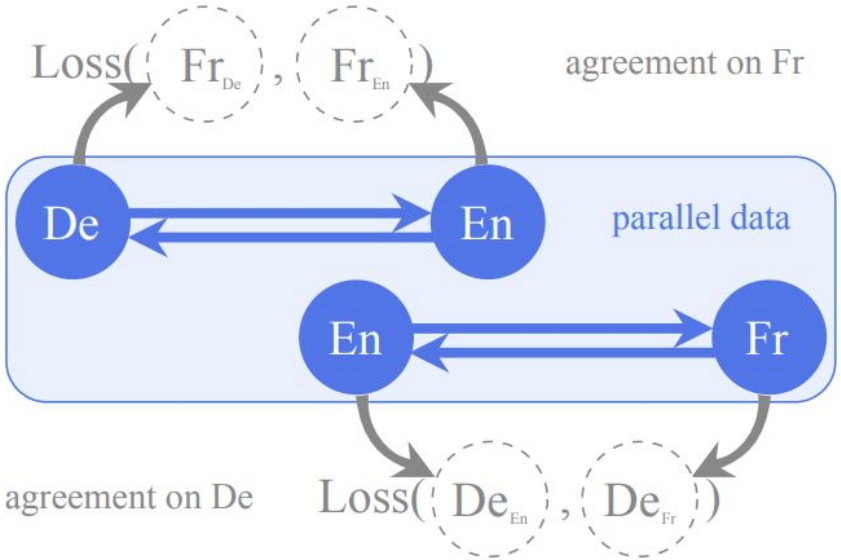
Attention Forcing at bottom decoder layer

$$R(X,Y) = -\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}(Att^n(Y_t,X) - Att^n(Y_t,Y))^2$$

Improvements over directly minimizing encoder divergence

Encourage output agreement

(Al-Shedivat and Parikh, 2019)



Encourage equivalent sentence in two languages to generate similar output in an auxiliary language

Decode to language L using both inputs s and t

$$\mathbf{Z}_{a\leftarrow s} \leftarrow \text{Decode}\left(\mathbf{Z}_a \mid f_\theta^{\text{enc}}(\mathbf{X}_s, L_a)\right)$$
$$\mathbf{Z}_{a\leftarrow t} \leftarrow \text{Decode}\left(\mathbf{Z}_a \mid f_\theta^{\text{enc}}(\mathbf{X}_t, L_a)\right)$$

Score output of one input using other input

$$\ell^t_{a\leftarrow s} \leftarrow \log \mathbb{P}_\theta\left(\mathbf{Z}_{a\leftarrow s} \mid \mathbf{X}_t\right)$$
$$\ell^s_{a\leftarrow t} \leftarrow \log \mathbb{P}_\theta\left(\mathbf{Z}_{a\leftarrow t} \mid \mathbf{X}_s\right)$$

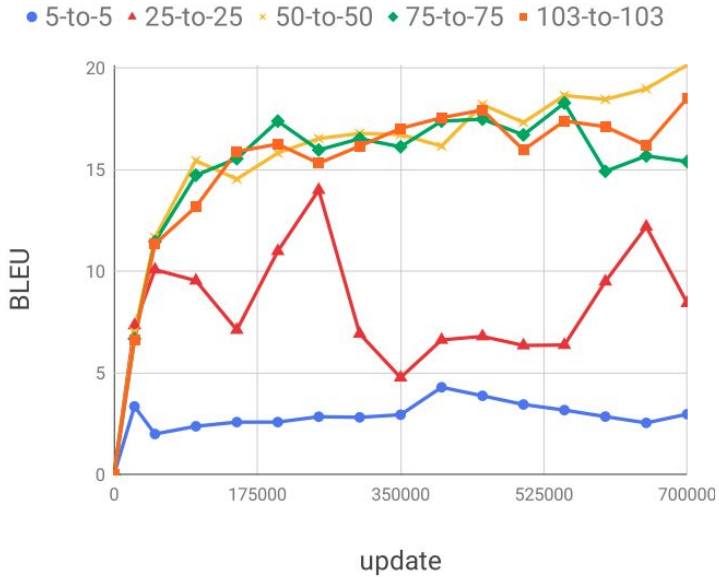
Add loss term force output agreement

$$\mathcal{L}^{\text{total}}(\theta) \leftarrow \mathcal{L}^{\text{sup}}(\theta) + \gamma(\ell^t_{a\leftarrow s} + \ell^s_{a\leftarrow t})$$

Competitive with pivot and no-loss in supervised directions

Effect of number of languages & corpus size

Zeroshot performance improves with the number of languages
(Aharoni et al., 2019; Arivazhagan et al., 2019b)



(Aharoni et al., 2019)

	$De \rightarrow Fr$	$Be \rightarrow Ru$	$Yi \rightarrow De$	$Fr \rightarrow Zh$	$Hi \rightarrow Fi$	$Ru \rightarrow Fi$
10 langs	11.15	36.28	8.97	15.07	2.98	6.02
102 langs	14.24	50.26	20.00	11.83	8.76	9.06

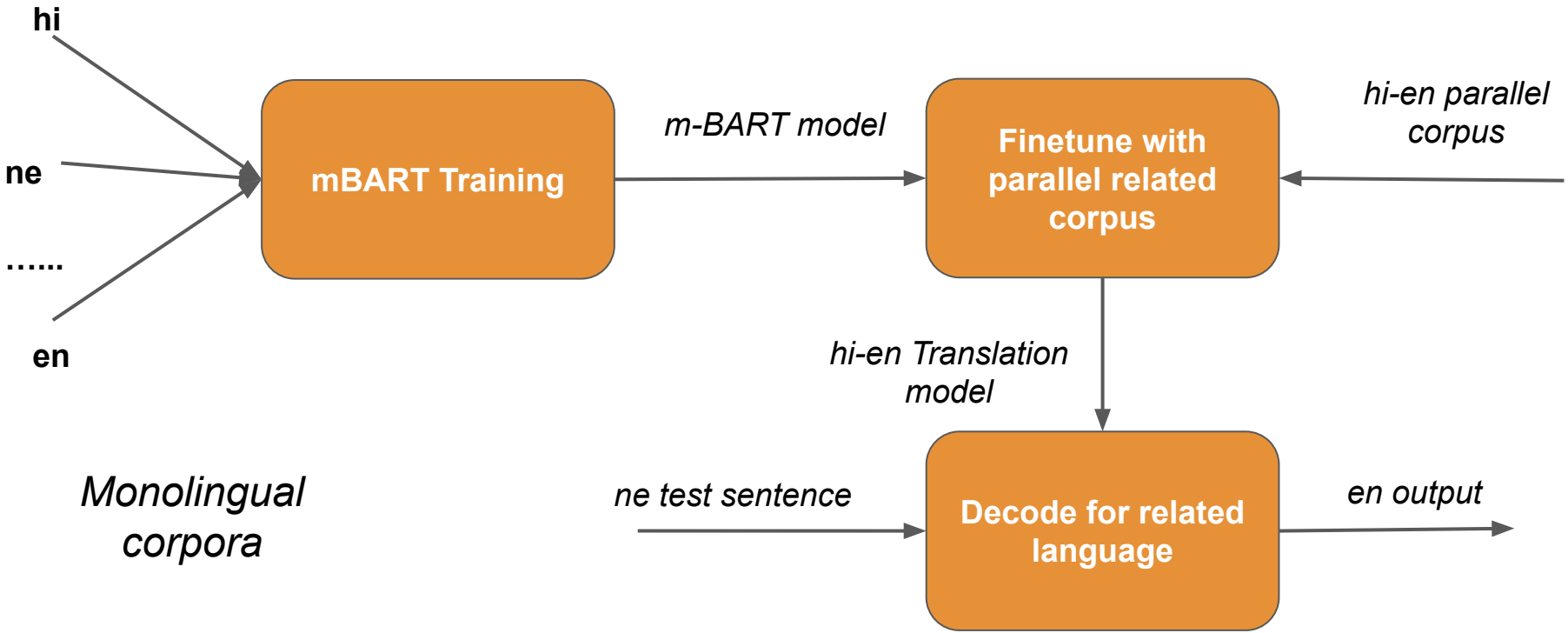
(Arivazhagan et al., 2019b)

Zero-shot translation may work well only when the multilingual parallel corpora is large

(Mattoni et al., 2017; Lakew et al., 2017)

Can monolingual pre-training help zero-shot translation?

(Liu et al., 2020)



Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Zero-resource Translation *(Firat et al., 2016b)*

Zero-shot translation & Zero-resource → No parallel corpus between unseen languages

Zero-shot → no specific training for unseen language pairs

Zero-resource → training takes into account unseen language pairs of interest
e.g use synthetic parallel corpus

Zero-resource NMT can be used to tune the NMT model for some unseen language pairs of interest

Creating Synthetic Parallel Corpus

(Firat et al., 2016b; Lakew et al., 2017; Gu et al., 2019; Currey & Heafield, 2019)

Expose M2M model to zeroshot directions

Create synthetic source to real target parallel

$\{ (X'_1 Y_1), (X'_2 Y_2), (X'_3 Y_3), (X'_4 Y_4) \}$

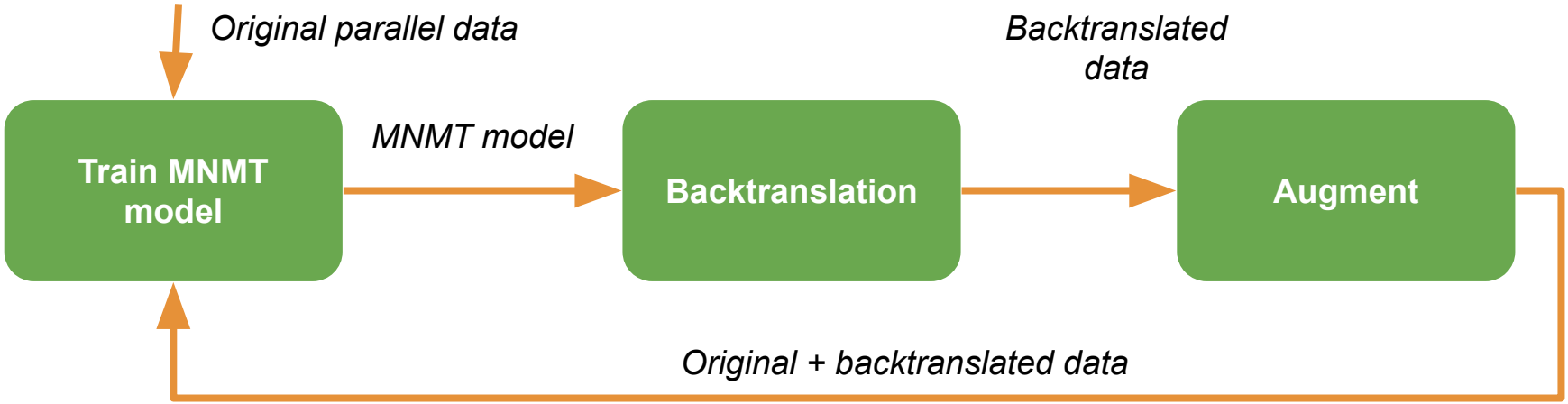
Via back-translation using the M2M model

Via pivot language
$Y \rightarrow E \rightarrow X'$
$Y' \leftarrow E \rightarrow X'$

Iterative Refinement

(Lakew et al., 2017)

Backtranslation quality depends on quality of underlying translation models



Iterative reinforcement learning approaches which reward original translation directions based on language modelling and reconstruction losses in zeroshot directions (Sestorain et al., 2018)

Scaling BT to multiple translation directions

Expensive to generate BT data for $O(n^2)$ language pairs

Random Online backtranslation *(Zhang et al., 2020)*

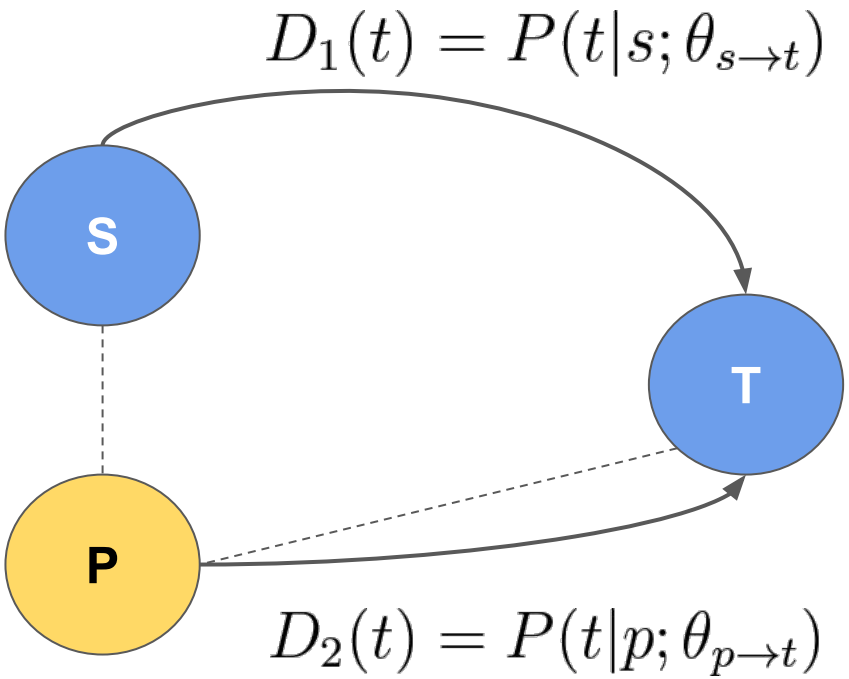
For every real sentence pair (x,y) in lang-pair (s,t)

- *Sample a source language s' → t*
- *Generate backtranslation pair (x',y) in pair (s',t)*
- *add backtranslated pairs to the minibatch*

Only doubles the effective corpus size

Results approach pivot baseline

Teacher-student training *(Chen et al., 2017)*



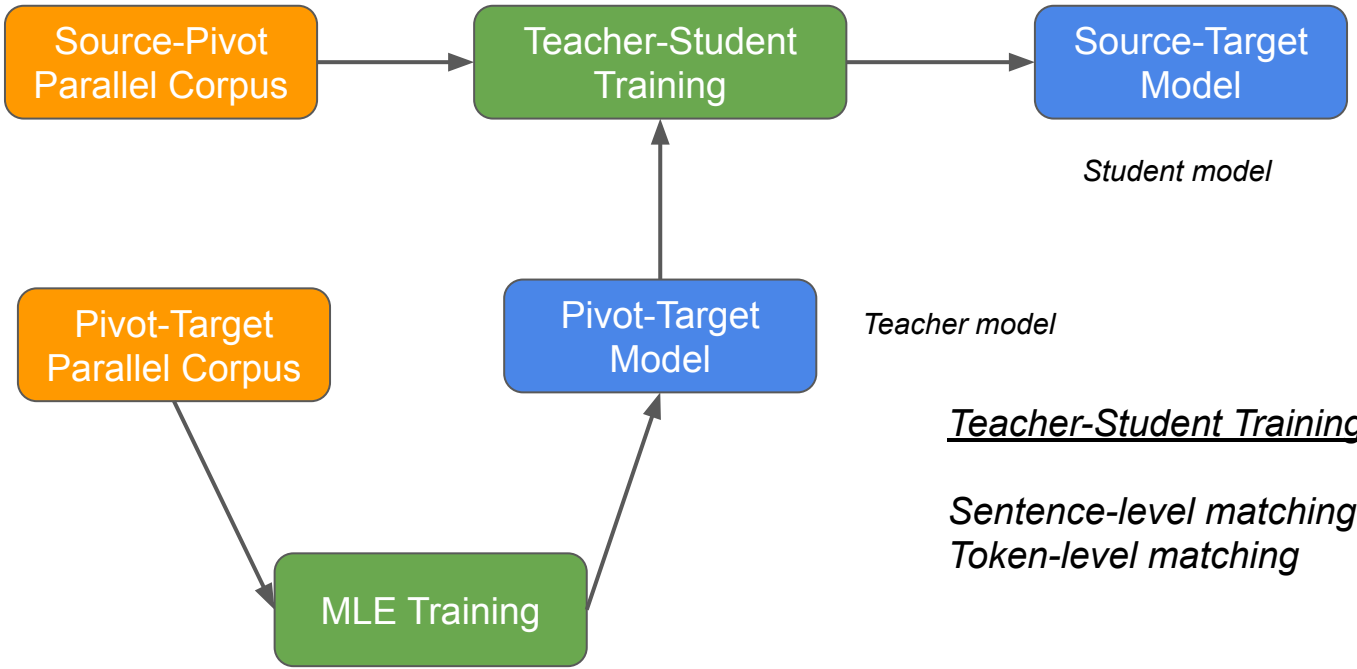
***Assumption**: The following distributions are similar*

$$P(t|s;\theta_{s\to t}) \qquad P(t|p;\theta_{p\to t})$$

Minimize

$$KL(D_2(t), D_1(t))$$

Given: Source-Pivot and Pivot-Target Parallel Corpus



Teacher-Student Training Approaches

Sentence-level matching
Token-level matching

Section Overview

- Transfer Learning for MNMT
  - Training Methods
  - Lexical Transfer
  - Syntactic Transfer
  - Language Relatedness
- Translation between unseen language pairs
  - Pivot Translation
  - Zeroshot Translation
  - Zero-resource Translation
  - Multibridge MNMT

Can't we simply add direct parallel corpora between non-English languages?

How do we acquire such parallel data?

How do we address data imbalance?

Single-bridge vs. Multi-bridge systems

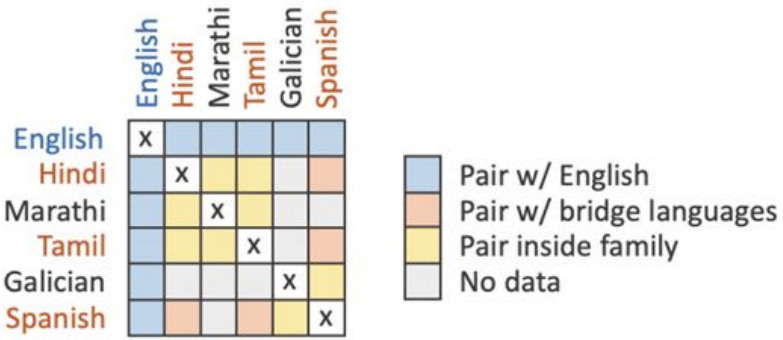
Mine parallel corpus from monolingual corpora

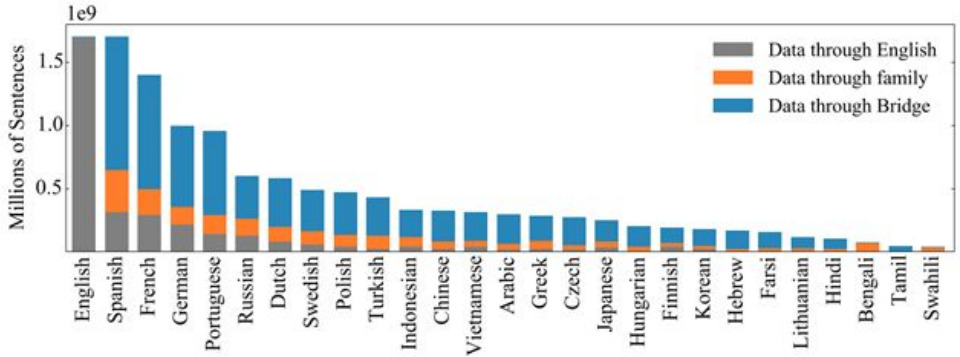
(Fan et al., 2020)

Expensive to mine from all 100 x 99 pairs

Which are the most promising language pairs to mine from?

Cluster languages and select bridge languages



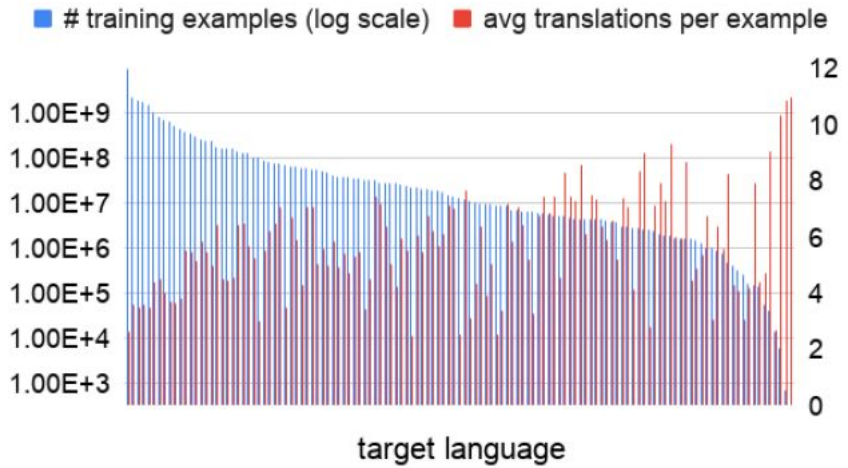


Extraction from English-centric parallel corpora

(Freitag & Firat, 2020; Rios et al., 2020)

X-Y	*Bleib sicher* ↔ Stay safe
Z-Y	Mantente segura ↔ Stay safe
X-Y-Z	*Bleib sicher* ↔ Mantente segura ↔ Stay safe

	cs	de	en	es	fr	ru
cs		0.7	47	0.8	1	0.9
de	0.7		4.5	2.3	2.5	0.3
en	47	4.5		13.1	38.1	33.5
es	0.8	2.3	13.1		10	4.4
fr	1	2.5	38.1	10		4.8
ru	0.9	0.3	33.5	4.4	4.8	



Data Sampling Strategies

Sampling strategies used for English-centric datasets have limitations

- *Large sample space → quadratic number of pairs to sample*
- *Pairwise sampling will be biased in favour of instances containing English*

	cs	de	en	es	fr	ru
cs		0.7	47	0.8	1	0.9
de	0.7		4.5	2.3	2.5	0.3
en	47	4.5		13.1	38.1	33.5
es	0.8	2.3	13.1		10	4.4
fr	1	2.5	38.1	10		4.8
ru	0.9	0.3	33.5	4.4	4.8	

Just ~15% sentence pairs exclusively non-English

Sampling independently on source and target marginal is also biased.

Solution 1: (Freitag & Firat, 2020)

1. *Temperature-based sampling of target language first using marginal distribution of target languages*
2. *Sample source language uniformly*

Solution 2: (Fan et al., 2020)

Sample from probability matrix while ensuring source and target marginals follow a temperature based schedule

These approaches improve over standard temperature-based sampling for non-English directions

Major Results

- Significant improvement in translation quality over pivot baseline for:
  - Newly trained directions
  - Zero-shot directions
- Multi-bridge translation does not adversely impact English-centric directions
- Multi-bridge translation and synthetic data augmentation provide complementary benefits

	Fan et al., 2020		**Rios et al., 2020**	
Language Pairs →	**New Train**	**Unseen**	**New Train**	**Unseen**
Single-bridge	5.4	7.6	20.0	20.9
Single-bridge pivot	9.8	12.4	22.9	23.7
Multi-bridge	**12.3**	**18.5**	**25.1**	**24.0**
+ *synth-data*				**25.2**

zeroshot

Word of caution: Human evaluations gains are not as significant (Fan et al., 2020)

Section Summary

MNMT has helped make significant advances in low-resource MT

- Novel methods for transfer learning and zeroshot translation

Transfer Learning

- Optimize the right objective for improved transfer
- Language-relatedness plays a key role in successful transfer
- Transfer works better in M2O setting than O2M setting
- Lexical transfer is easier to achieve than syntactic transfer

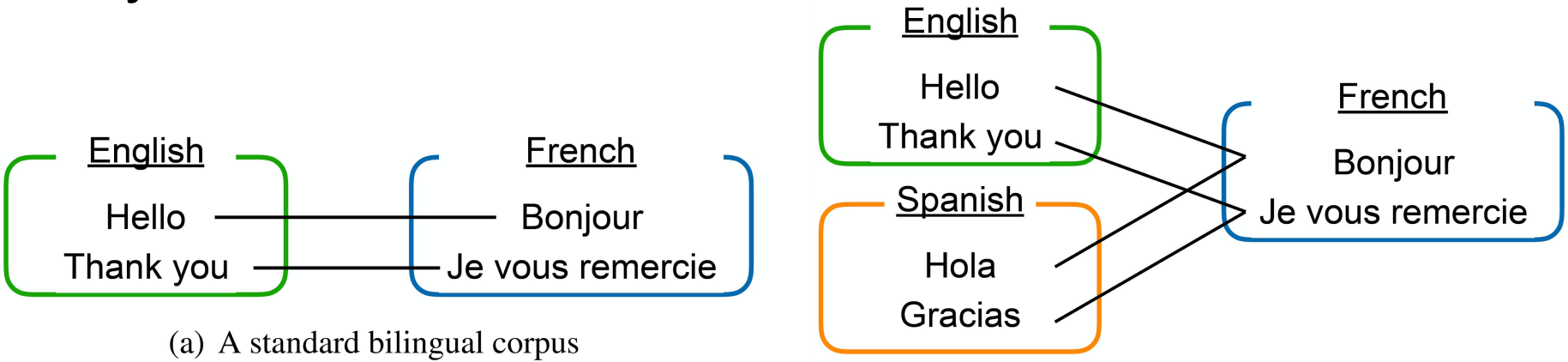
Translation between unseen languages

- Pivot translation is a strong baseline
- Zeroshot ⇒ spurious correlation between input representation & output language
  - Reducing divergence between internal representations of different languages
  - Multi-bridge systems
- Zero-resource translation can use synthetic data to reduce spurious correlations

Outline of This Tutorial

- Overview of Multimodal NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

Why Multi-source MT?



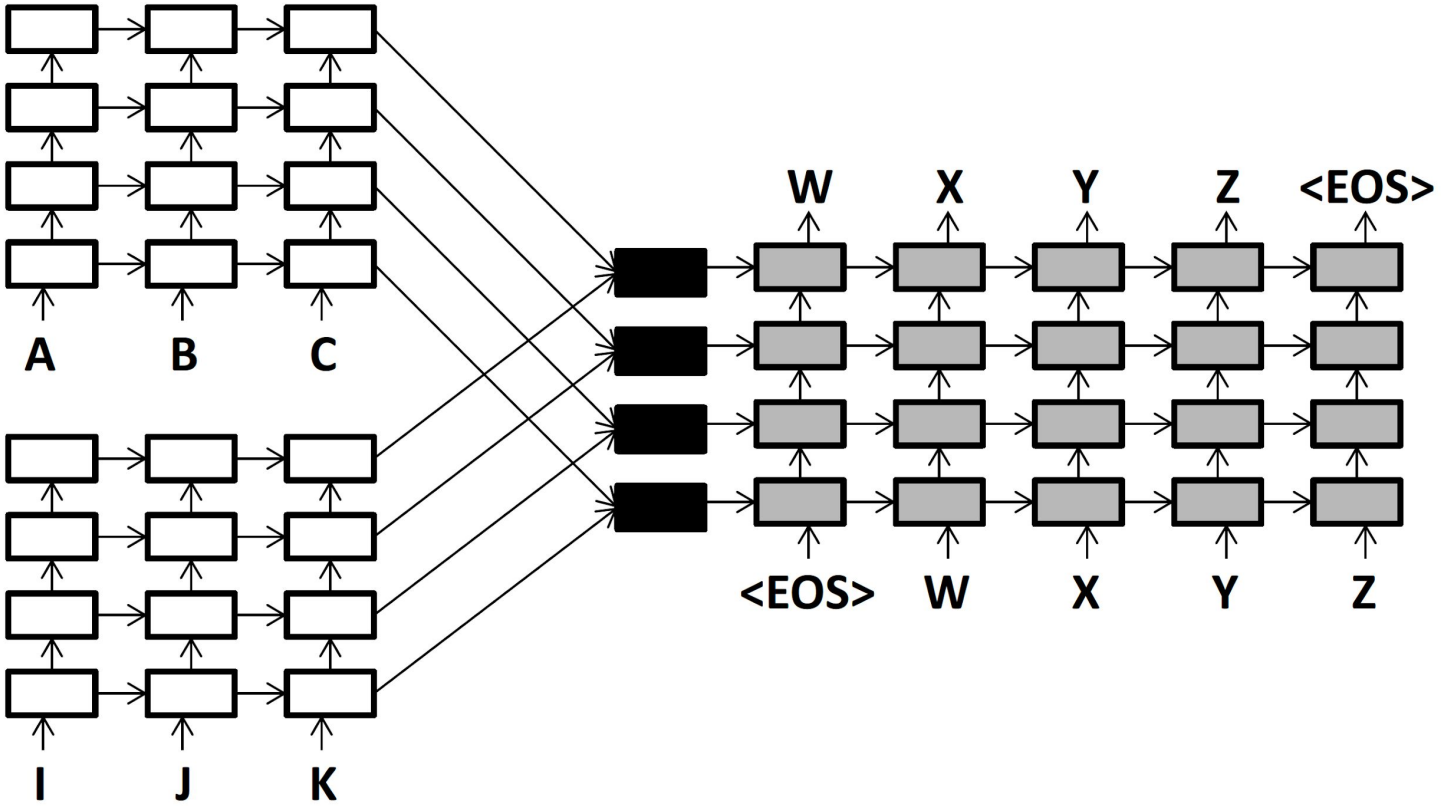
(a) A standard bilingual corpus

(b) A complete multi-source corpus

Figures from (Nishimura et al., 2018)

If a source sentence has already been translated into multiple languages then these sentences can be used together to improve the translation into the target language: such as EU, and UN.

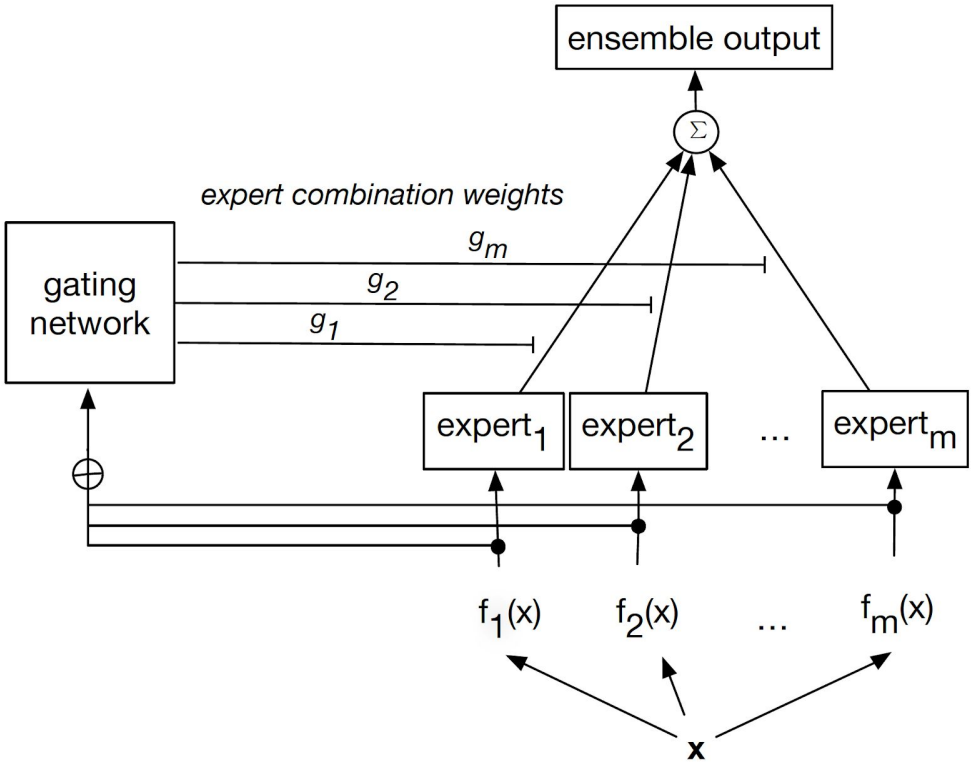
Multi-Source Available: Multi-source Encoder (Zoop et al., 2016)



Results of Multi-source Encoder (Zoph et al., 2016)

Target = English			
Source	**Method**	**Ppl**	**BLEU**
French	—	10.3	21.0
German	—	15.9	17.3
French+German	Basic	8.7	23.2
French+German	Child-Sum	9.0	22.5
French+French	Child-Sum	10.9	20.7
French	Attention	8.1	25.2
French+German	B-Attent.	5.7	30.0
French+German	CS-Attent.	6.0	29.6

Multi-Source Available: Ensembling (Garmash et al., 2016)



outputs of individual NMT systems $(y^1, y^2, \ldots, y^m)$

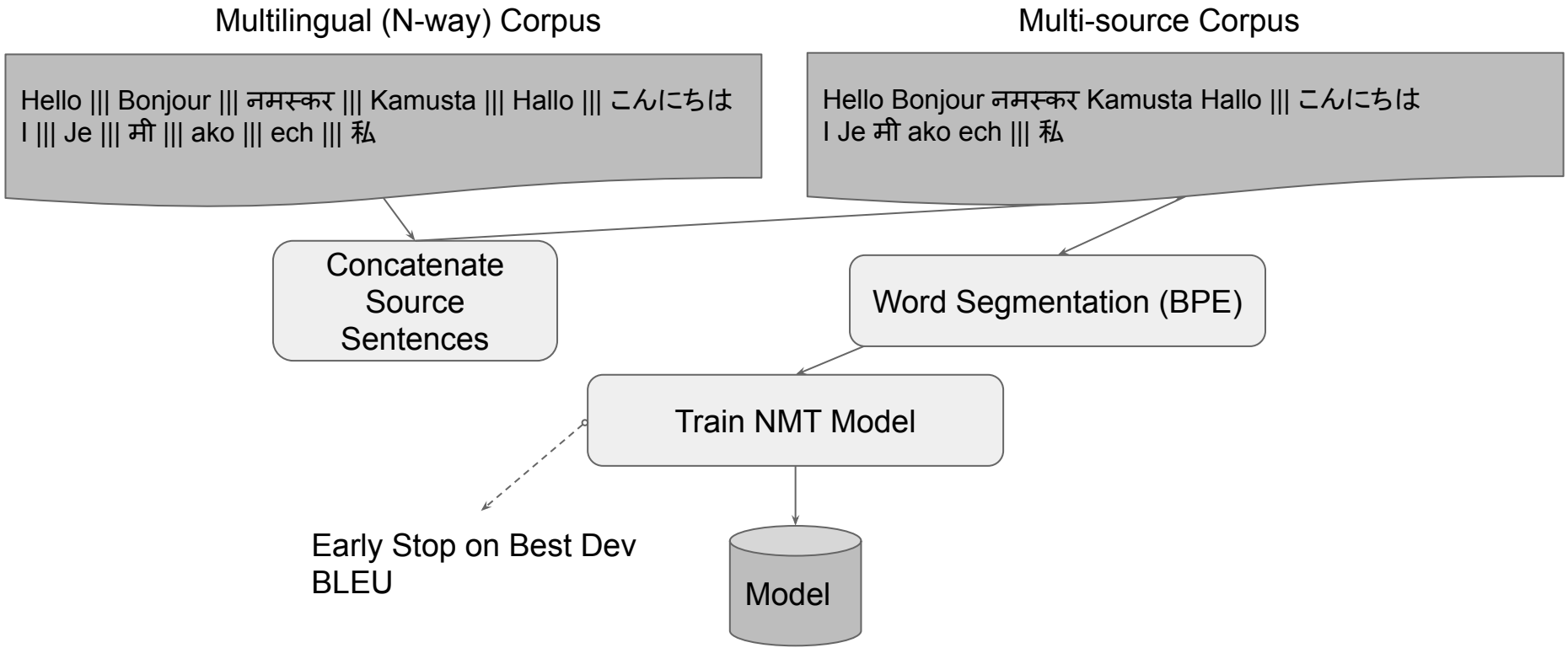
expert-specific state representations (hidden non-recurrent state of the NMT model)

abstract decoding state

Results of Ensembling (Garmash et al., 2016)

Ensemble set	Combination type					
	uniform		global		context-dependent	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
$de_1^2 \rightarrow en$	21.8	50.4	21.8	50.4	21.8	50.3
$de_1^4 \rightarrow en$	21.8	50.4	21.8	50.4	-	-
$\{de_1^2, de_3^4\} \rightarrow en$	-	-	21.8	50.4	**22.8**	**51.0**
$fr_1^2 \rightarrow en$	29.2	58.0	29.2	58.1	29.3	58.1
$fr_1^4 \rightarrow en$	29.2	58.0	29.2	58.1	-	-
$\{fr_1^2, fr_3^4\} \rightarrow en$	-	-	29.2	58.1	**30.2**	**59.0**
de,fr$\rightarrow$ en	**29.5**	**58.3**	**29.9**	**58.7**	30.3	59.2
de_1,de_2,fr_1,$fr_2 \rightarrow en$	29.4	58.3	29.3	58.2	-	-
$\{\{de_1,fr_1\},\{de_2,fr_2\}\} \rightarrow en$	-	-	29.2	57.9	**31.5**	**60.3**

Multi-Source Available: Concatenation (Dabre et al., 2017)



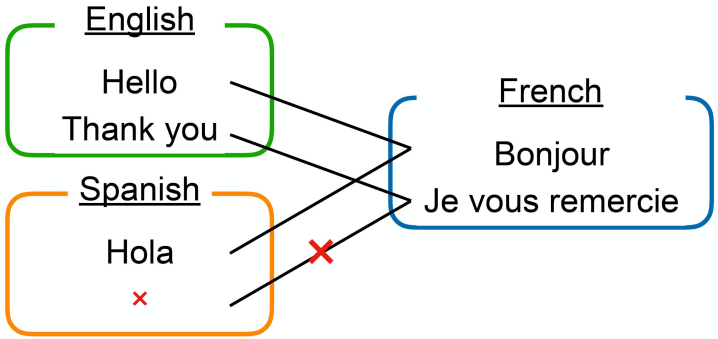
Results of Concatenation (Dabre et al., 2017)

Corpus Type	Language Pair	tst2010	tst2013	Number of sources	tst2010	tst2013
3 lingual 191381 lines	**Fr-En**	19.72	22.05	**2 sources**	**22.56***/18.64/22.03	**24.02***/18.45/23.92
	De-En	16.19	16.13			
4 lingual 84301 lines	**Fr-En**	9.02	7.78	**3 sources**	**11.70**/12.86*/10.30	**9.16**/9.48*/7.30
	De-En	7.58	5.45			
	Ar-En	6.53	5.25			
5 lingual 45684 lines	**Fr-En**	6.69	6.36	**4 sources**	**8.34**/9.23*/7.79	**6.67***/6.49/5.92
	De-En	5.76	3.86			
	Ar-En	4.53	2.92			
	Cs-En	4.56	3.40			

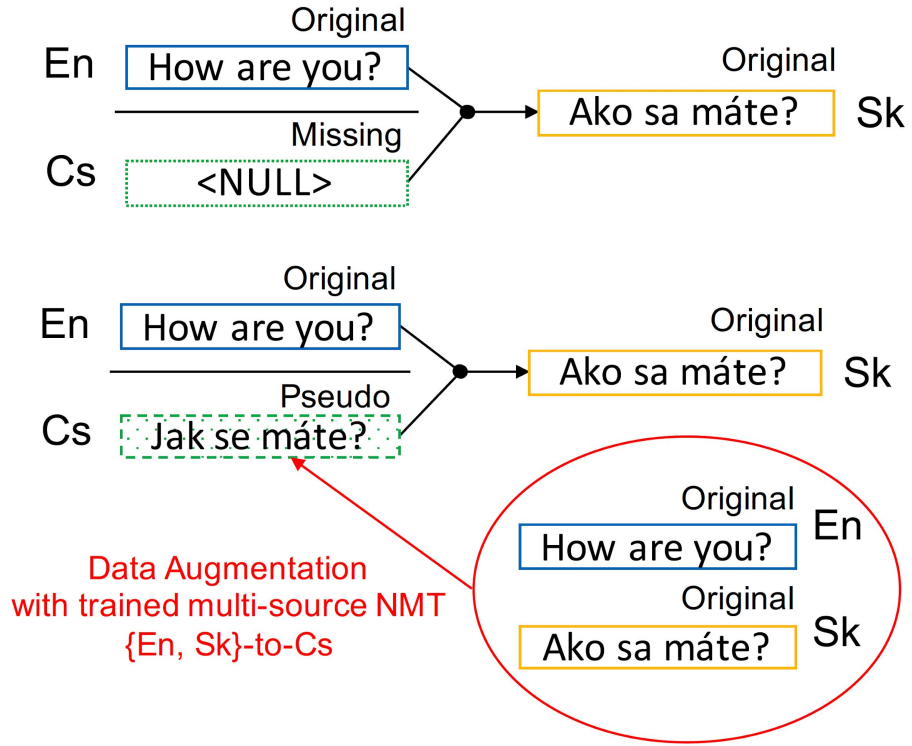
* Concatenation (bold)/Ensembling/Multi-source Encoder

Missing Source Sentences (Nishimura et al., 2018)

Scenario:



Methods:



Results of Missing Source Sentences (Nishimura et al., 2018)

Pair	Trg	baseline method			proposed method		
		one-to-one (En-to-Trg)	multi-encoder NMT (fill up with symbol)	multi-encoder NMT (back translation)	fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	20.21	28.18	27.57	29.17	29.37	**29.40**
	sr	16.42	23.85	22.73	24.41	**24.96**	24.15
en-sk/cs	sk	13.79	20.27	19.83	20.26	20.43	**20.59**
	cs	14.72	19.88	19.54	20.78	**20.90**	20.61
en-vi/id	vi	24.60	25.70	26.66	**26.73**	26.48	26.32
	id	24.89	**26.89**	26.34	26.40	25.73	26.21

Summary of Multi-source Approaches

Multi-Source Approach		N-way data needed	Solutions	Concerns	Benefits
1	**Vanilla**	Yes	Multi or Shared Encoder model	Long training times Bulky Models	Expanding N-way corpora
2	**Ensembling**	No	Ensemble multiple bilingual models	Learning ensemble functions Need small N-way corpora	Reuse existing bilingual models
3	**Synthetic data**	No	Generate missing source sentences	Infeasible for real time translation	Applicable for post-editing

Outline of This Tutorial

- Overview of Multimodal NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

Multiway Datasets

Languages	**Corpora**
European languages	Europarl, JRC-Aquis, DGT-Aquis, DGT-TM, ECDC-TM, EAC-TM etc.
Asian languages	WAT shared tasks etc.
Indic languages	CVIT-PIB/MKB, PMIndia, IndoWordNet etc.
Massive	WikiMatrix, JW300 etc.
Others	UN, TED, Opensubtitles etc.

Refer to catalogs like OPUS and the IndicNLP catalog for comprehensive listings of parallel corpora resources.

Low or Zero-Resource Multiway Datasets

Corpus	Domain	Languages
FLORES	Wikipedia	English, Nepali, Sinhala
XNLI	Caption	15 languages
CVIT-Mann ki Baat/PIB	General	10 Indian Languages
Indic parallel corpus	Wikipedia	6 Indian Languages
WMT shared tasks	Web	German, Upper Sorbian

+All the multiway datasets listed in the previous slide can be used for testing

Multi-source Datasets

Corpus	N-way	Domain	Languages
Europarl	11	Politics	European languages
TED	5	Spoken	French, German, Czech, Arabic and English
UN	6	Politics	Arabic, Chinese, English, French, Russian and Spanish
ILCI	11	Tourism/health	Indian languages + English
ALT	9	News	South-East Asian languages + English, Japanese
Bible	1,000	Religion	Most major languages

Outline of This Tutorial

- Overview of Multimodal NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

Exploring Pre-trained Models

- Pre-training embeddings, encoders and decoders have been shown to be useful for NMT
- But
  - How pre-training can be incorporated into different MNMT architectures?
  - How to address techniques to maximize the impact of transfer in fine-tuning?
- Unsupervised pre-training and unsupervised NMT might be worth investing

Unseen Language Pair Translation

- Previous work on unseen language pair translation has only addressed cases where the pivot language is related to or shares the same script with the source language
- The pivot language (mostly English) is unlikely to be related to the source and target languages and this scenario requires further investigation (especially for zero-shot translation).
- New approaches need to be explored to significantly improve over the simple pivot baseline.

Joint Multilingual and Multi-Domain NMT

- When extending an NMT system to a new language, the parallel corpus in the domain of interest may not be available
- Transfer learning in this case has to span languages and domains
- It might be worthwhile to explore adversarial approaches where domain and language invariant representations can be learned for the best translations

Multilingual Speech-to-Speech NMT

- An interesting research direction would be to explore multilingual speech translation, where the ASR, translation and TTS modules can be multilingual
- Interesting challenges and opportunities may arise in the quest to compose all these multilingual systems in an end-to-end method.
- Multilingual end-to-end speech-to-speech translation would also be a future challenging scenario

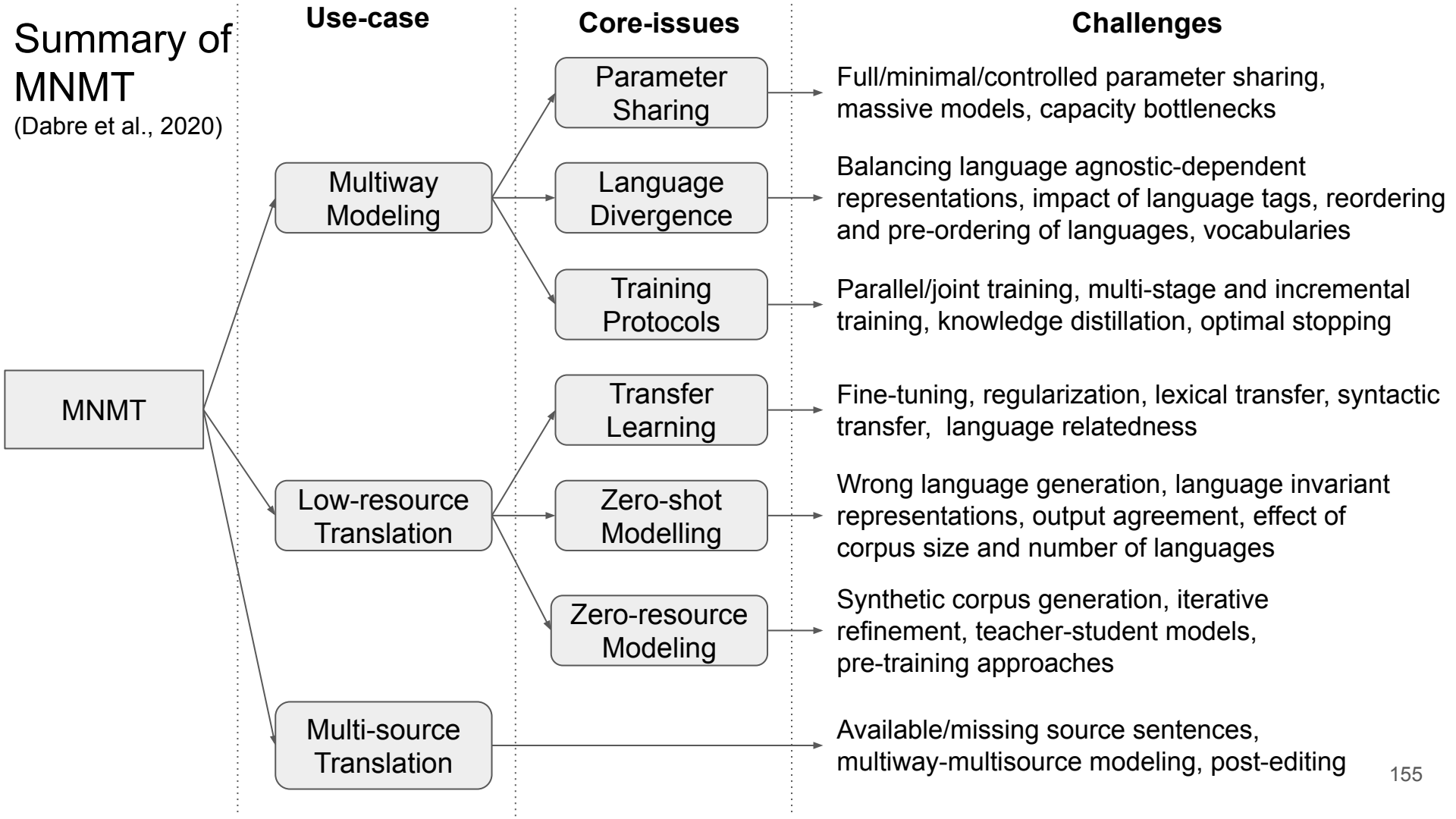
Your ideas?

Outline of This Tutorial

- Overview of Multimodal NMT (30 min by Chenhui Chu)
- Multiway Modeling (1 hour by Raj Dabre)
- Low-resource Translation (1 hour by Anoop Kunchukuttan)
- Multi-source Translation (10 min by Chenhui Chu)
- Datasets, Future Directions, and Summary (20 min by Chenhui Chu)

Summary of MNMT

(Dabre et al., 2020)



Thank You!

Contact:

prajdabre@gmail.com
chu@i.kyoto-u.ac.jp
anoop.kunchukuttan@microsoft.com

Tutorial Material: https://github.com/anoopkunchukuttan/multinmt_tutorial_coling2020

Survey Paper: https://dl.acm.org/doi/10.1145/3406095