A Journey on Multilingual Neural Machine Translation

SHAOMU TAN |  UNIVERSITY OF AMSTERDAM
Language Technology Lab

This Lecture

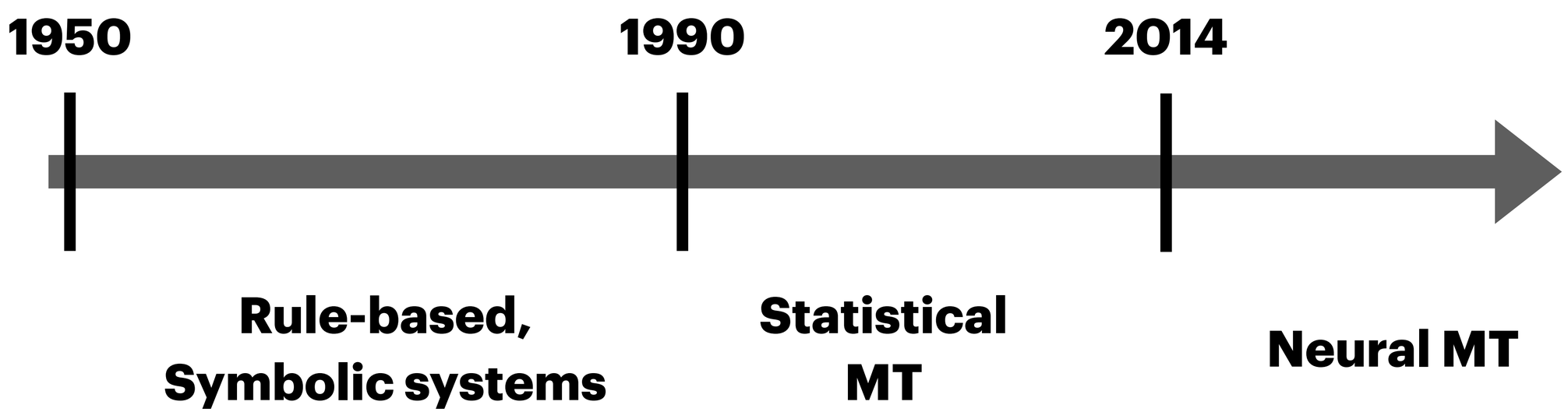
Part I:

- Machine Translation Background (20 mins)
- Multilingual Machine Translation (25 mins)
- Break (15 mins)

Part II:

- Neuron Specialization (25 mins)
- Zero-Shot Machine Translation (20 mins)

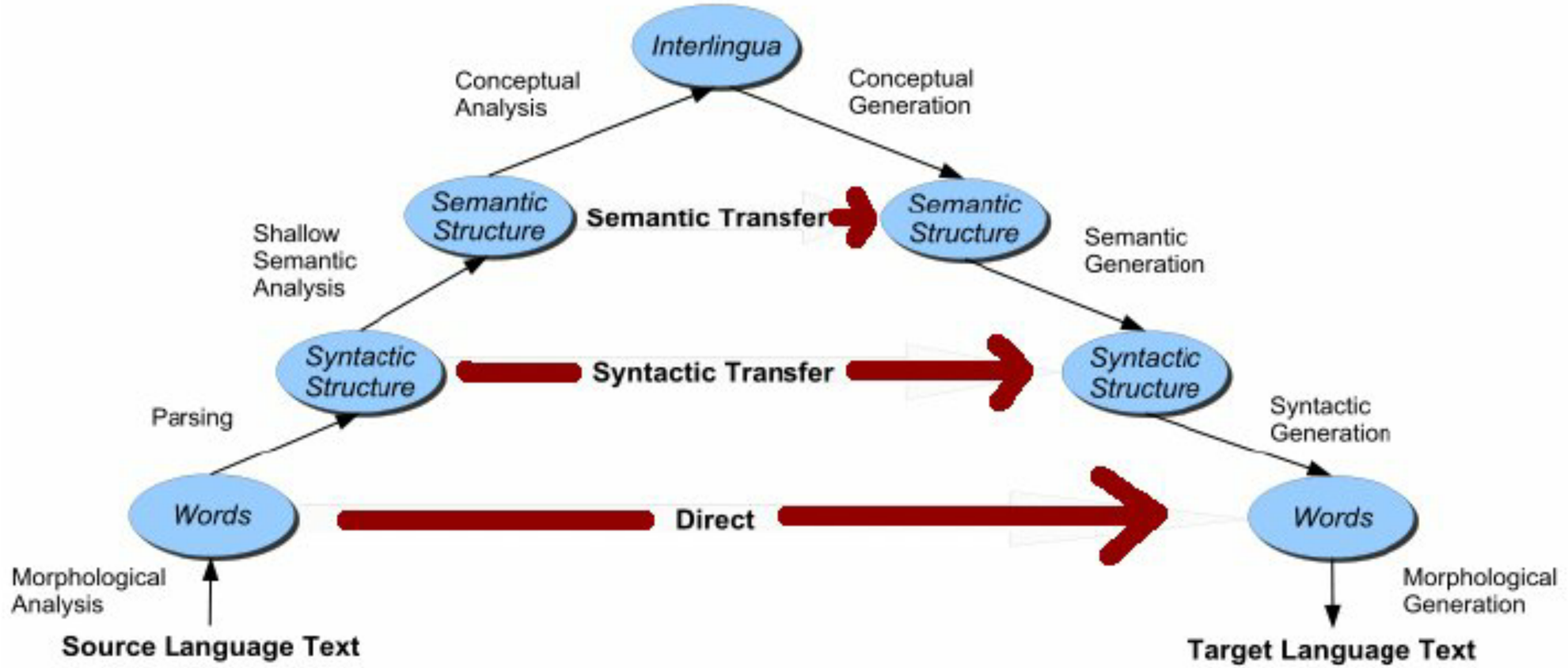
Machine Translation History



Classical Methods for Machine Translation

1950 - 1990

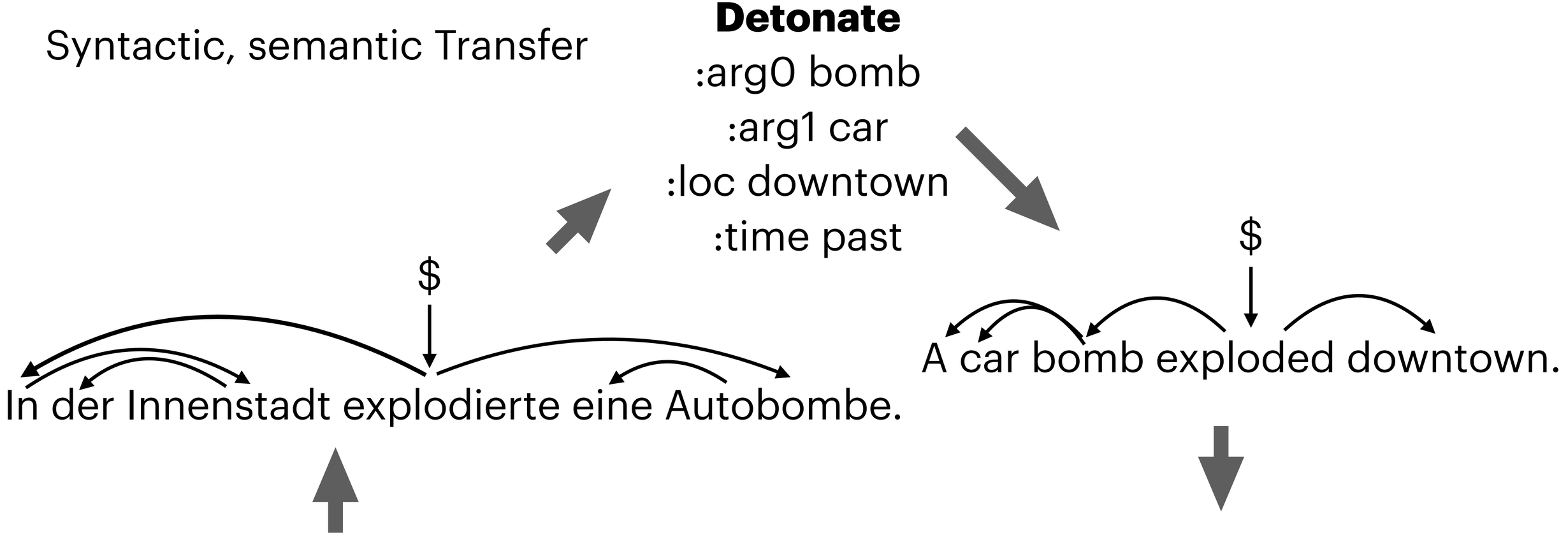
The Vauquois Triangle (1986)



Classical Methods for Machine Translation

1950 - 1990

Syntactic, semantic Transfer



Example from: Graham Neubig, CMU Multilingual NLP 2022

Classical Methods for Machine Translation

1950 - 1990

- **Kummerspeck** (German) — The weight you put on due to being sad or depressed, especially after heartbreak. It literally means “grief bacon”.

Statistical Machine Translation

1990 - 2013

- Powered by Statistical Machine Learning
- Translation is done by two separate components.
- Given a sentence x in one language, we find the English sentence y:

$$y^* = \underset{y}{\operatorname{argmax}}\ p(y \mid x)$$

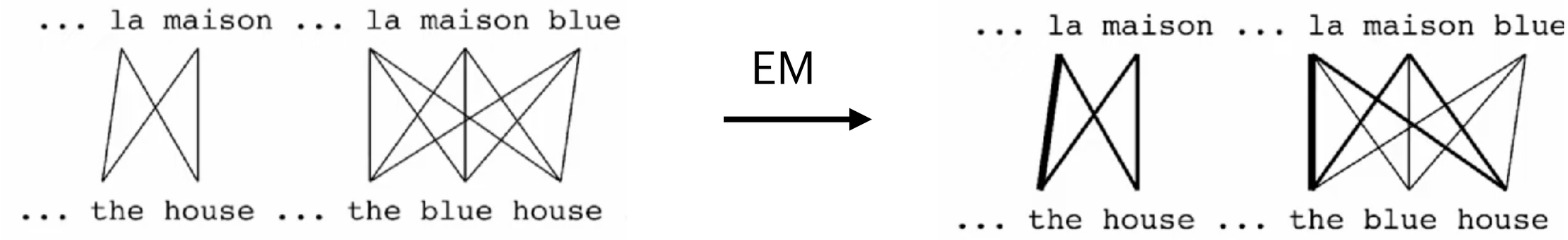
$$= \underset{y}{\operatorname{argmax}}\ \underbrace{p(y)}_{\text{Langauge Model}} \cdot \underbrace{p(x \mid y)}_{\text{“Translation” Model}}$$

word / phrase alignment

Statistical Machine Translation

1990 - 2013

- The “Translation” Model learns the alignment by Expectation Maximization (EM) algorithm.



Example from: Philipp Koehn, Machine Translation 2020

Neural Machine Translation

2014 - now

Language Modeling

$$p(y_1, y_2, \ldots, y_n) = \prod_{t=1}^{n} p(y_t \mid y_{<t})$$

Contexts

Machine Translation
(Conditional LM)

$$p(y_1, y_2, \ldots, y_n \mid x) = \prod_{t=1}^{n} p(y_t \mid y_{<t}, x)$$

Source
sentence

Neural Machine Translation

2014 - now

- Utilizing Encoder-Decoder Architecture based on Neural Networks.

- Encoder: map a sentence into a continuous representation.

- Decoder: generate the target sentence given the representation.

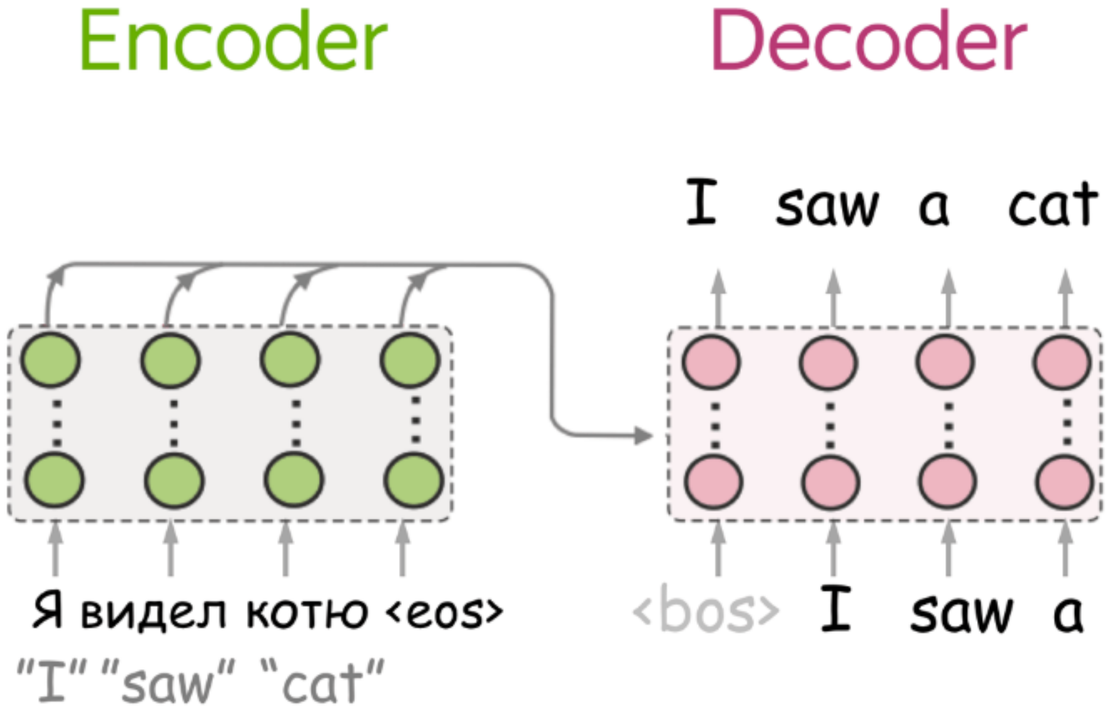


Figure from: Elena Voita, NLP course for you

Machine Translation as Seq-2-Seq Modeling

Sutskever et al. (2014) utilizes LSTM to train MT models:

- Encoder -> LSTM
- Decoder -> LSTM

How are the Encoder and Decoder connected?

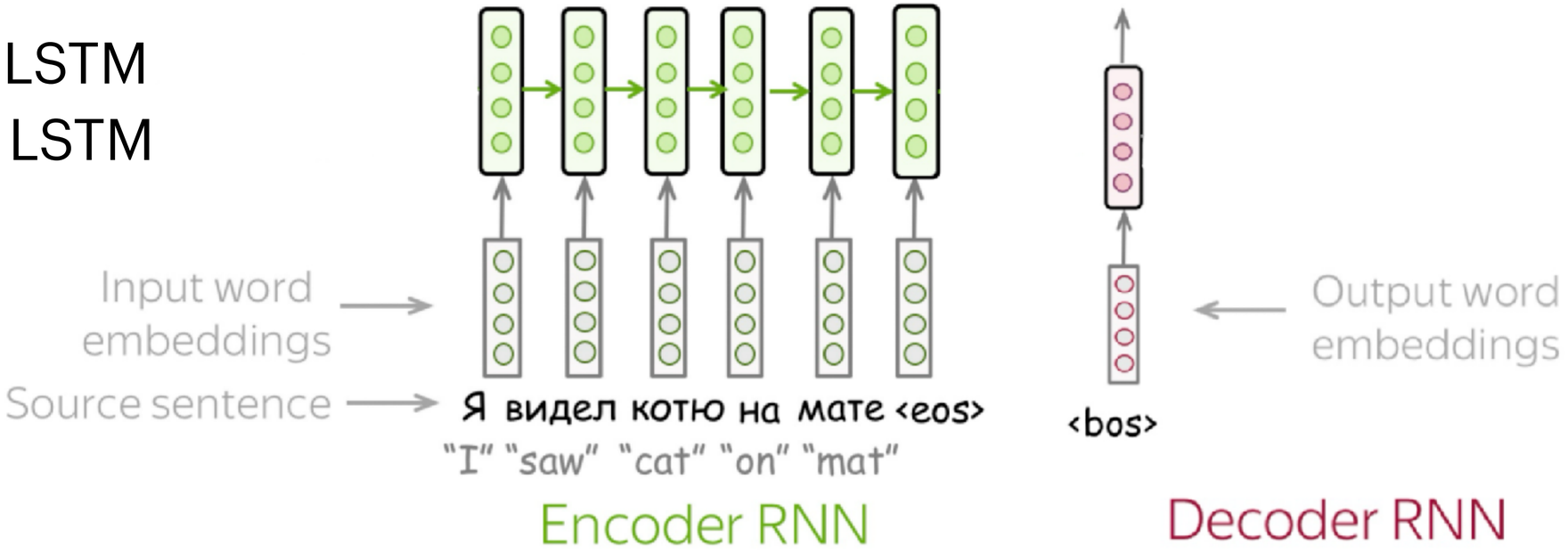
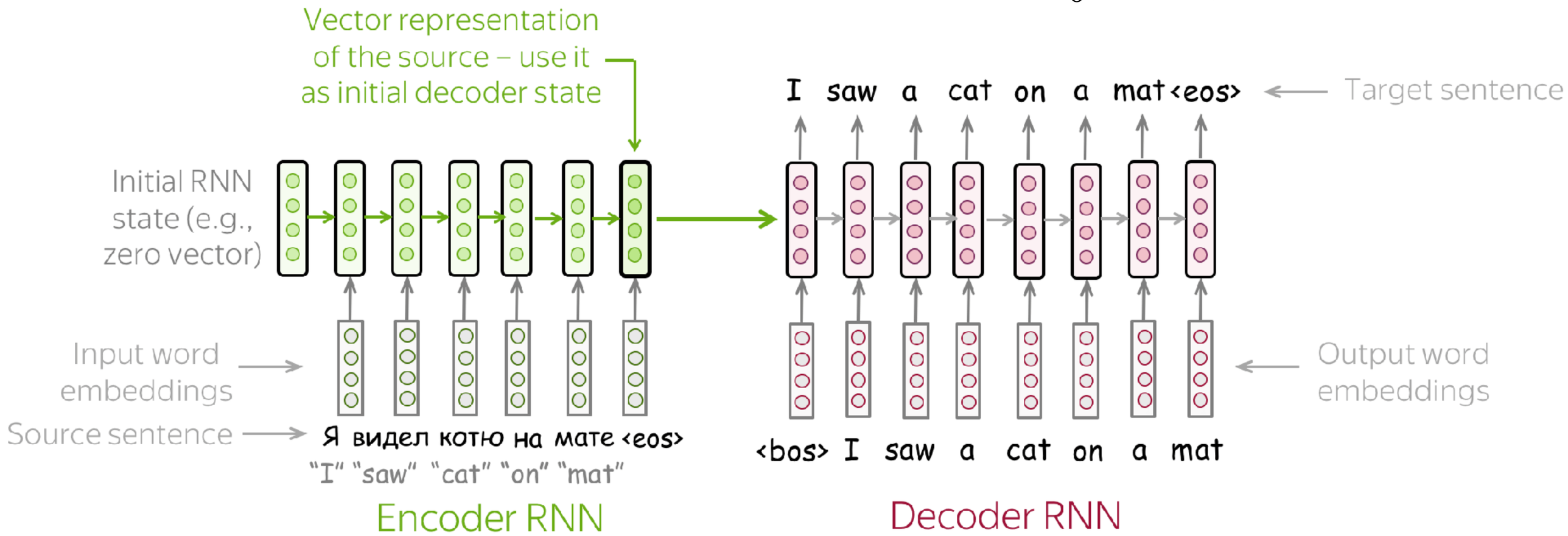


Figure from: Elena Voita, NLP course for you

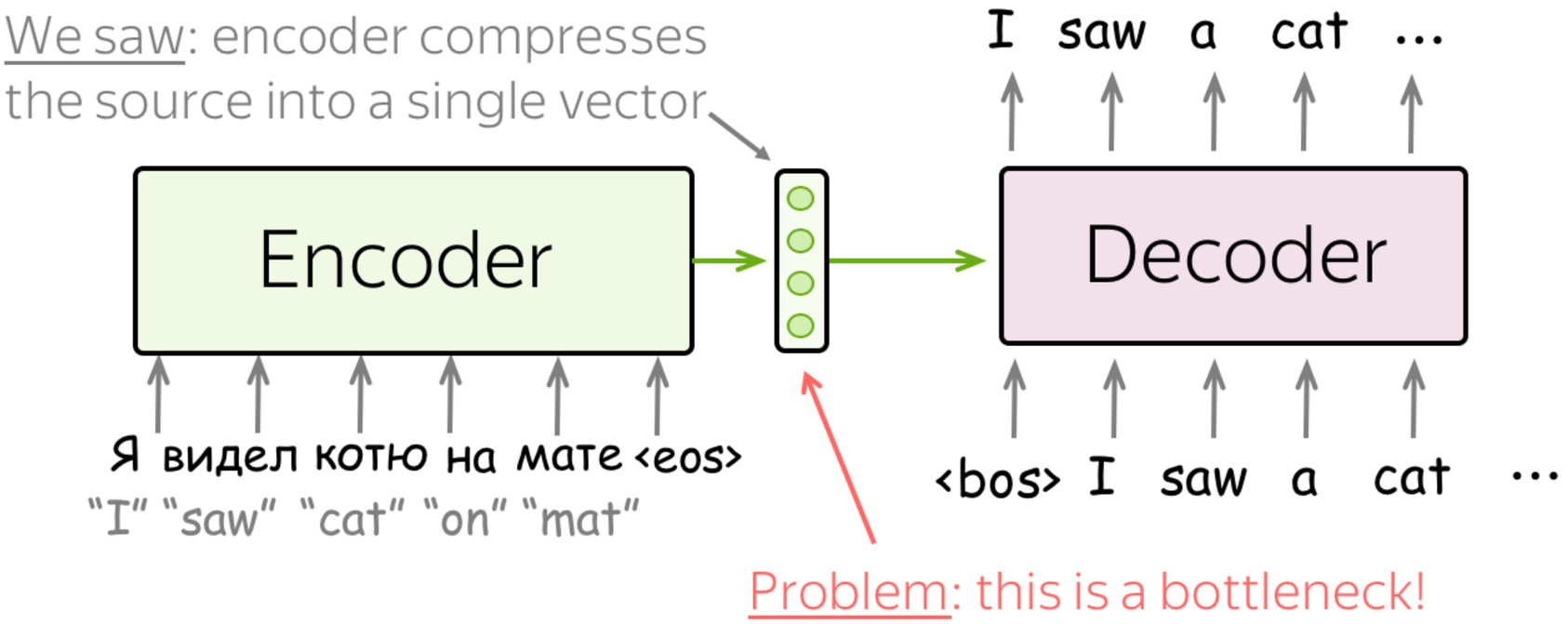
Machine Translation with RNNs

Sutskever et al. (2014) **initialise the first decoder LSTM state with the last state of the encoder LSTM.** $h_0^{dec} = h_n^{enc}$



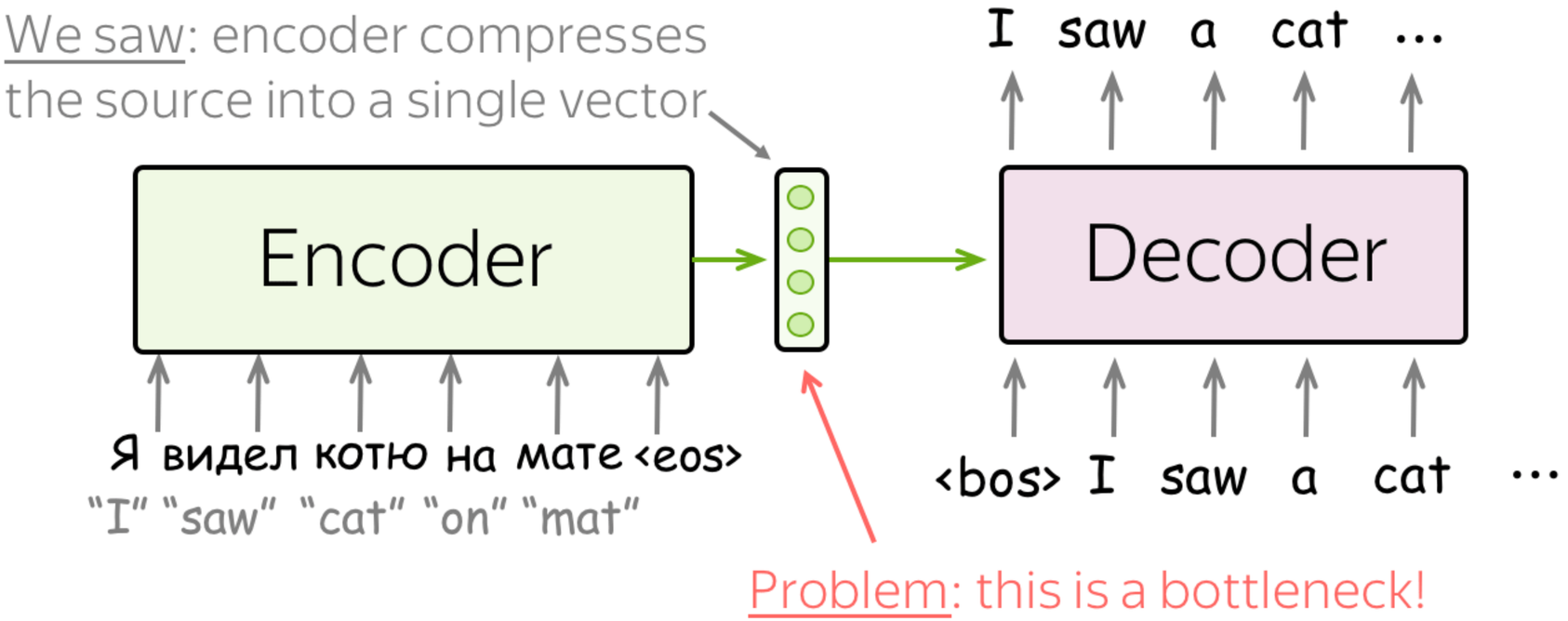
Limitations behind the RNN-MT approaches

RNN approaches store the full information in a **Single vector**.



Limitations behind the RNN-MT approaches

The **source representation is fixed** when generating target tokens.



Attention Mechanism

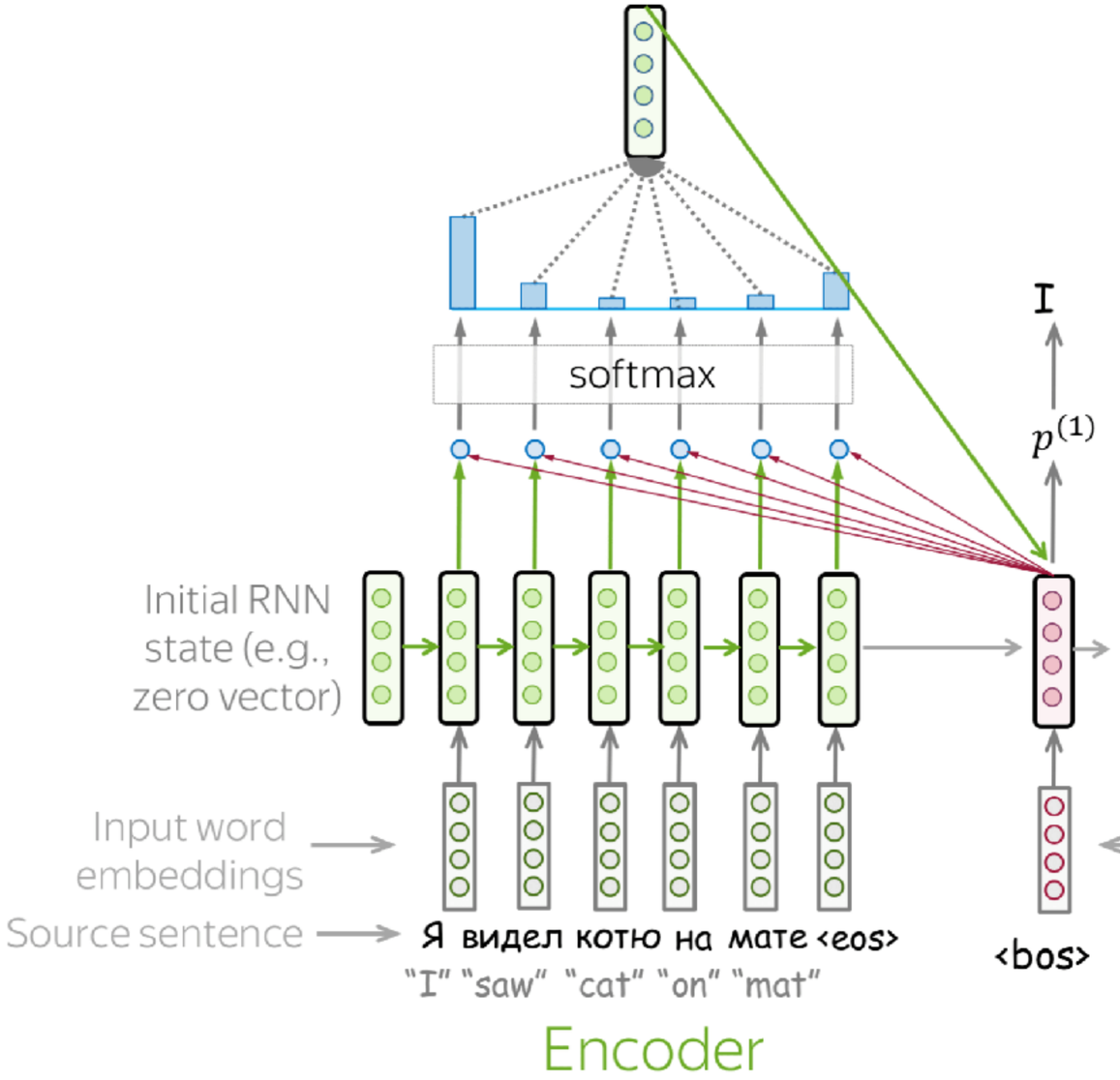
Bahdanau et al. (2014)

The intuition behind Attention:

~~Encode all tokens into in a single vector.~~
Encode each token into a vector, e.g.: all RNN states.

~~Remember the whole source information (fixed source embedding)~~
At each step, model could focus on more important parts, e.g.: word alignments.

Attention Mechanism

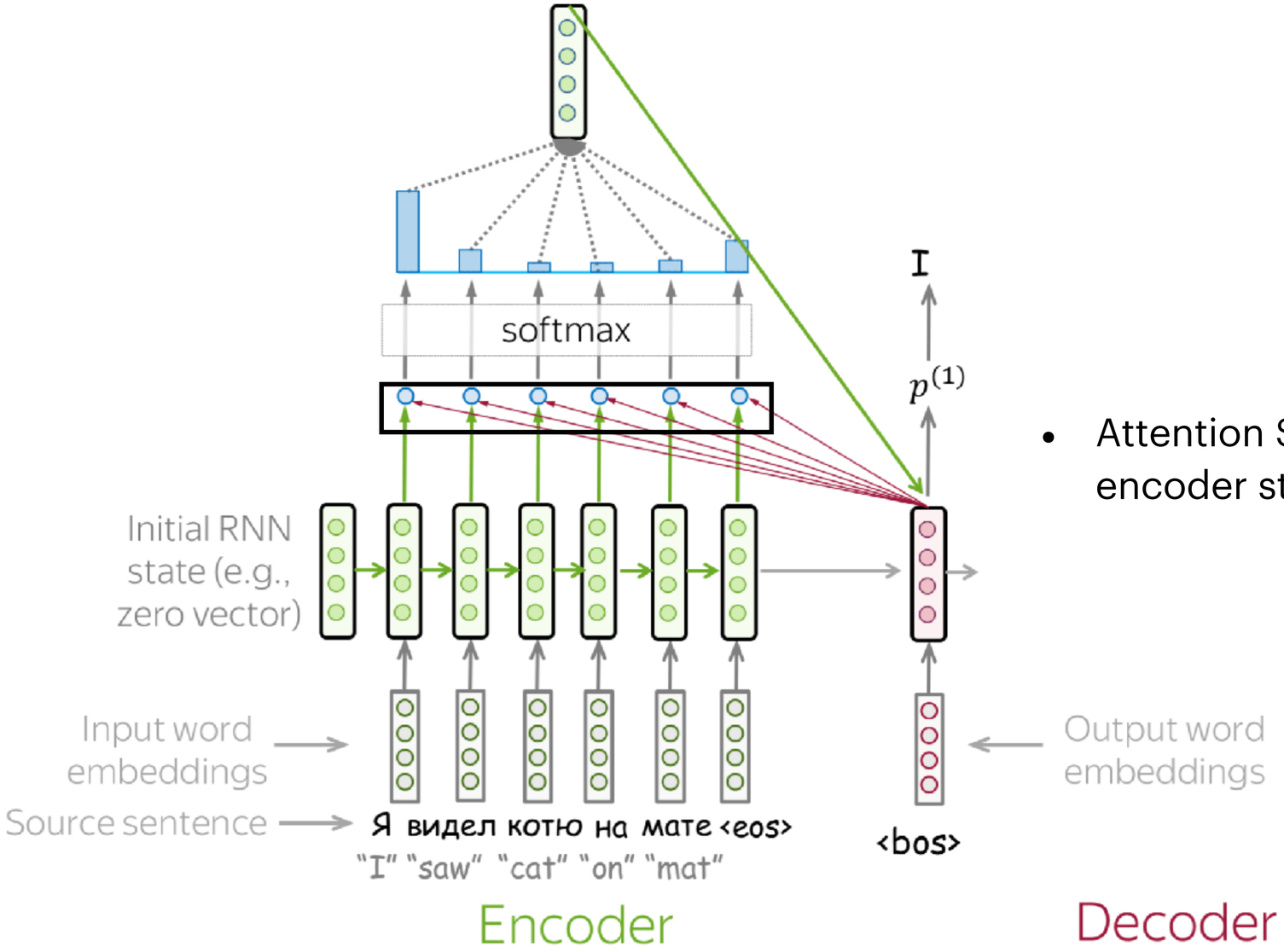


$s_1, s_2, \ldots, s_m$ all encoder states

h_t one decoder state

- Given the decoder state h_t, and encoder states $\{s_1, s_2, \ldots, s_n,\}$.

Attention Mechanism

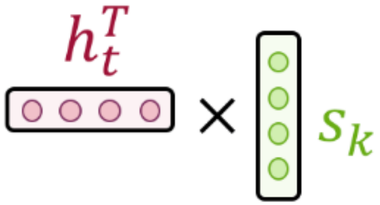


- Attention Scores: measure how "relevant" h_t to an encoder state s_k.

Attention Mechanism

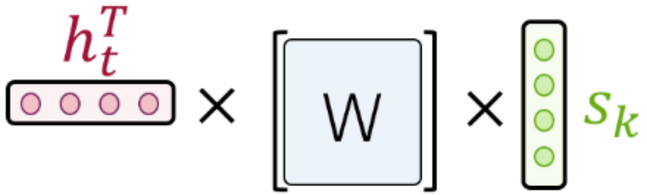
How to compute Attention Score:

Dot-product



$$\text{score}(h_t, s_k) = h_t^T \, s_k$$

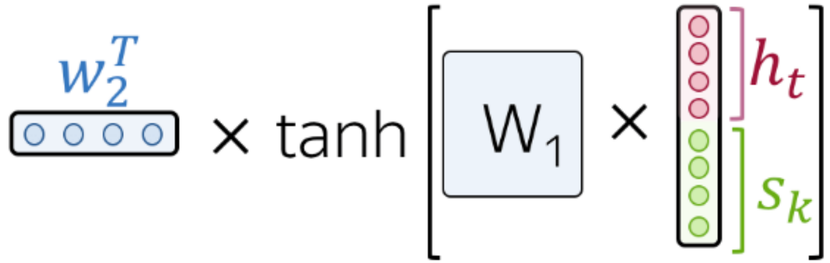
Bilinear



$$\text{score}(h_t, s_k) = h_t^T \, W s_k$$

"Luong Attention"
Luong, et al (2015)

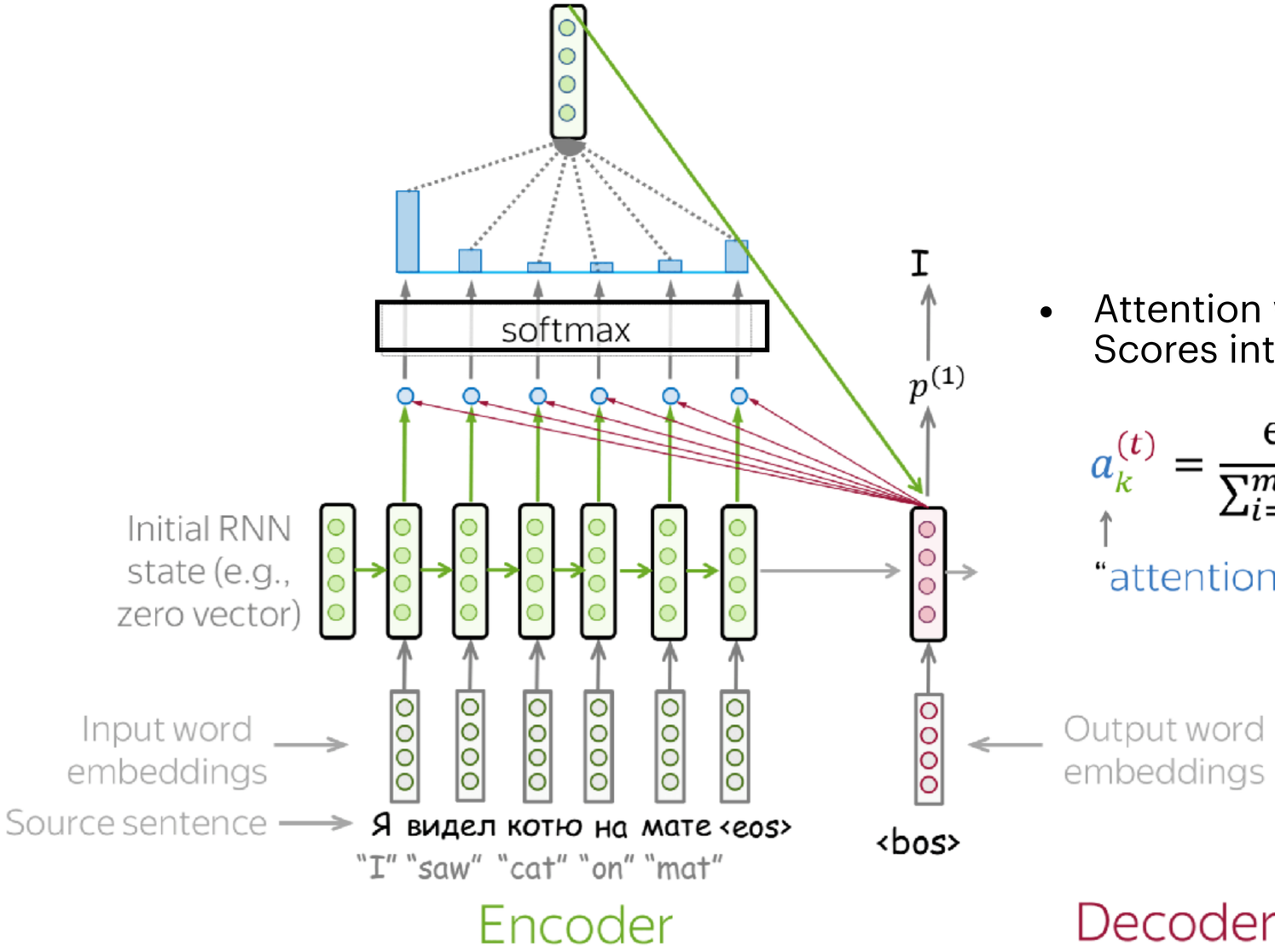
Multi-Layer Perceptron



$$\text{score}(h_t, s_k) = w_2^T \cdot \tanh(W_1[h_t, s_k])$$

"Bahdanau Attention"
Bahdanau, et al (2015)

Attention Mechanism

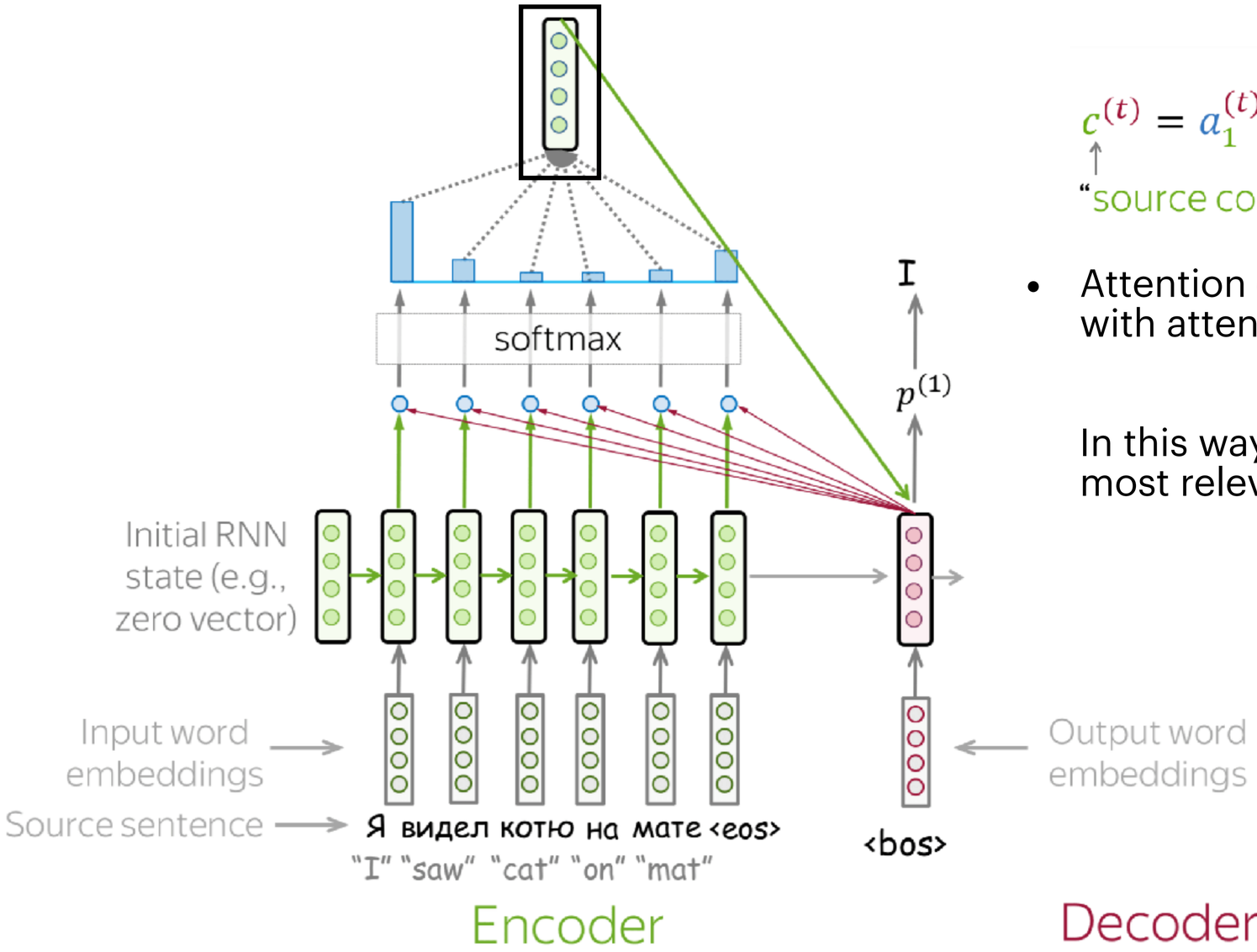


- Attention weights: apply Softmax to convert Attention Scores into probability.

$$a_k^{(t)} = \frac{\exp(\text{score}(h_t, s_k))}{\sum_{i=1}^{m} \exp(\text{score}(h_t, s_i))}, \text{k} = 1..\text{m}$$

"attention weight for source token k at decoder step t"

Attention Mechanism



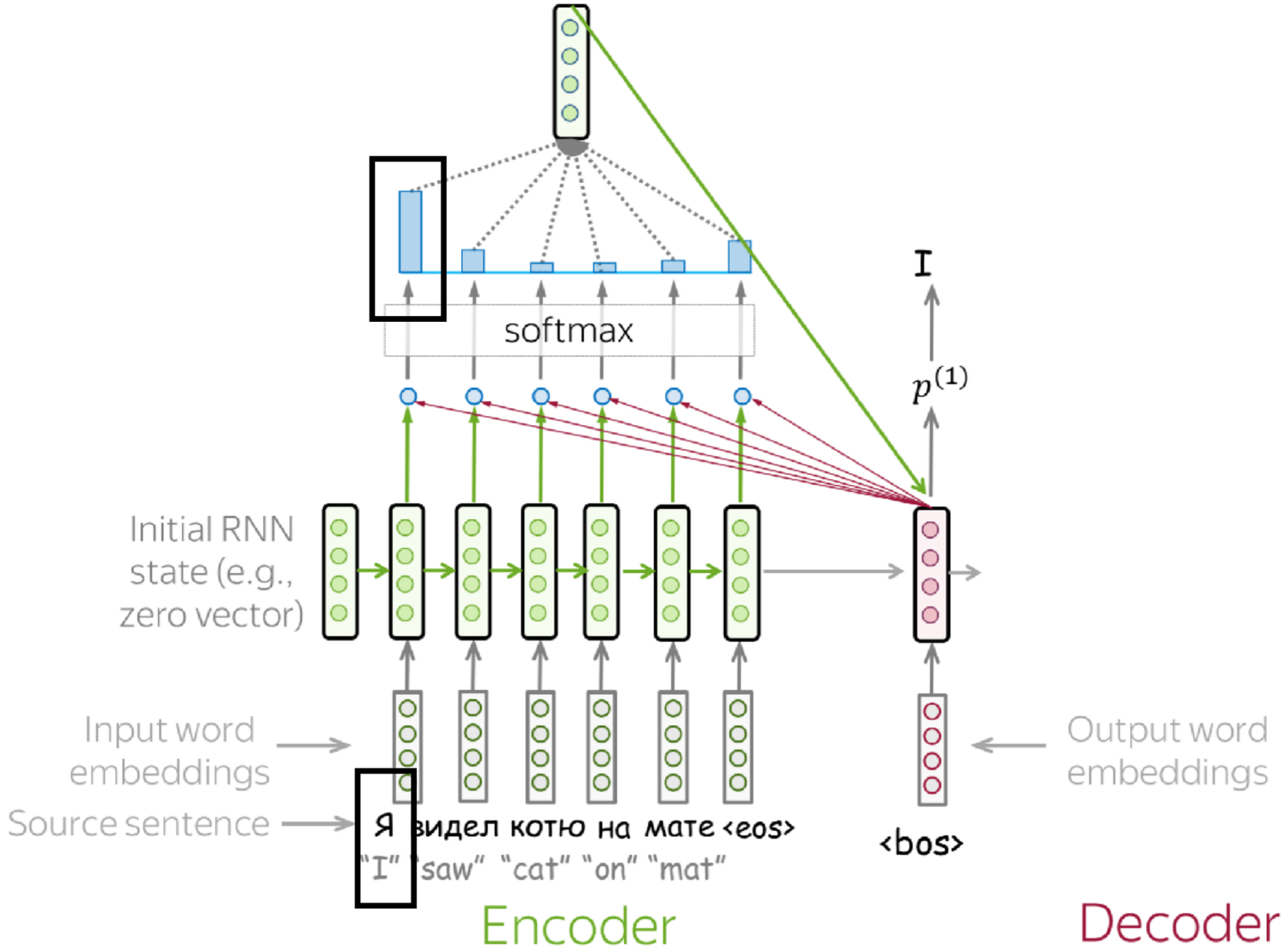
$$c^{(t)} = a_1^{(t)} s_1 + a_2^{(t)} s_2 + \cdots + a_m^{(t)} s_m = \sum_{k=1}^{m} a_k^{(t)} s_k$$

"source context for decoder step t"

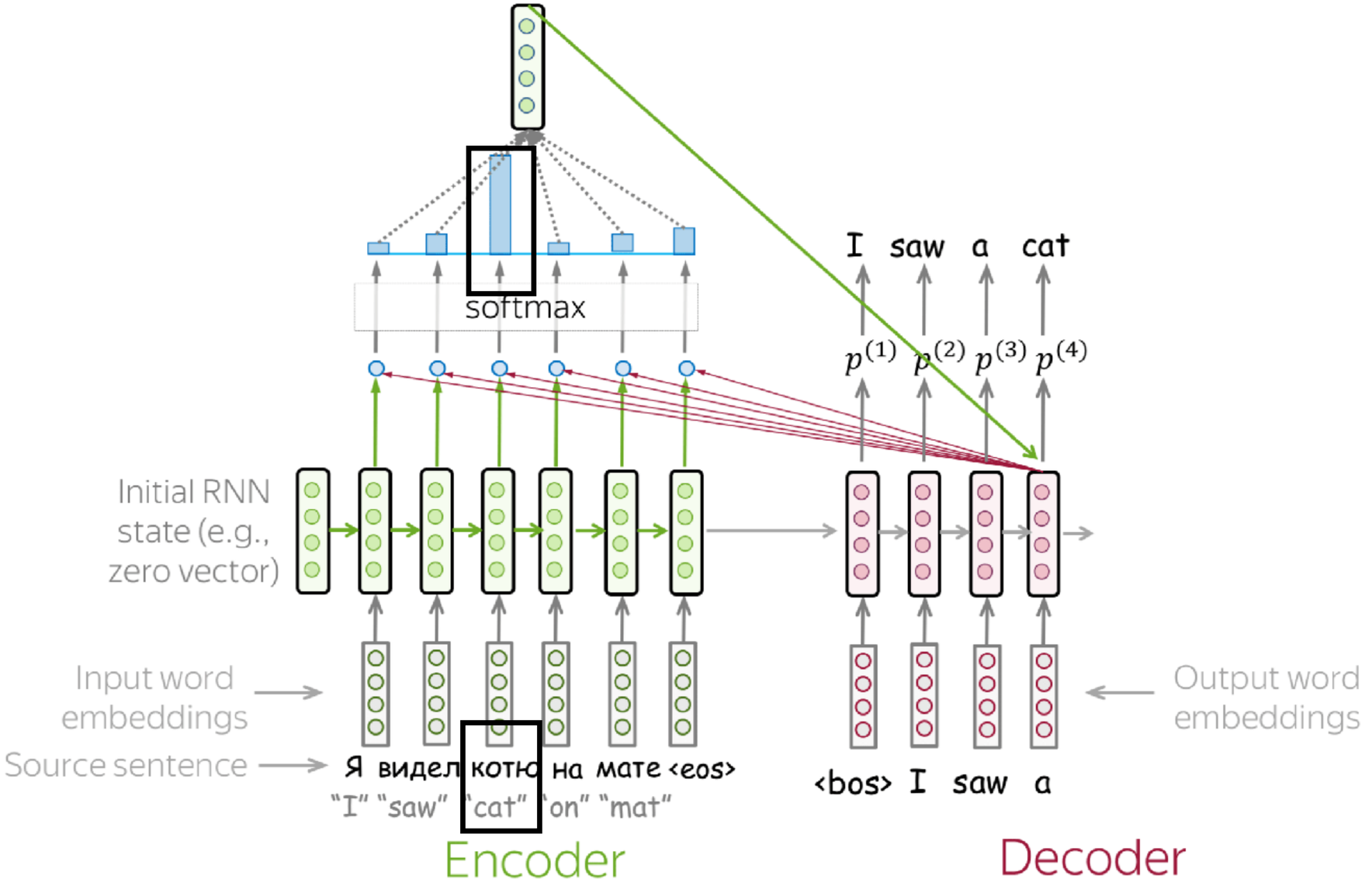
- Attention output: weighted sum of encoder states with attention weights.

In this way, the model learns to pay attention to the most relevant source tokens.

Attention Mechanism



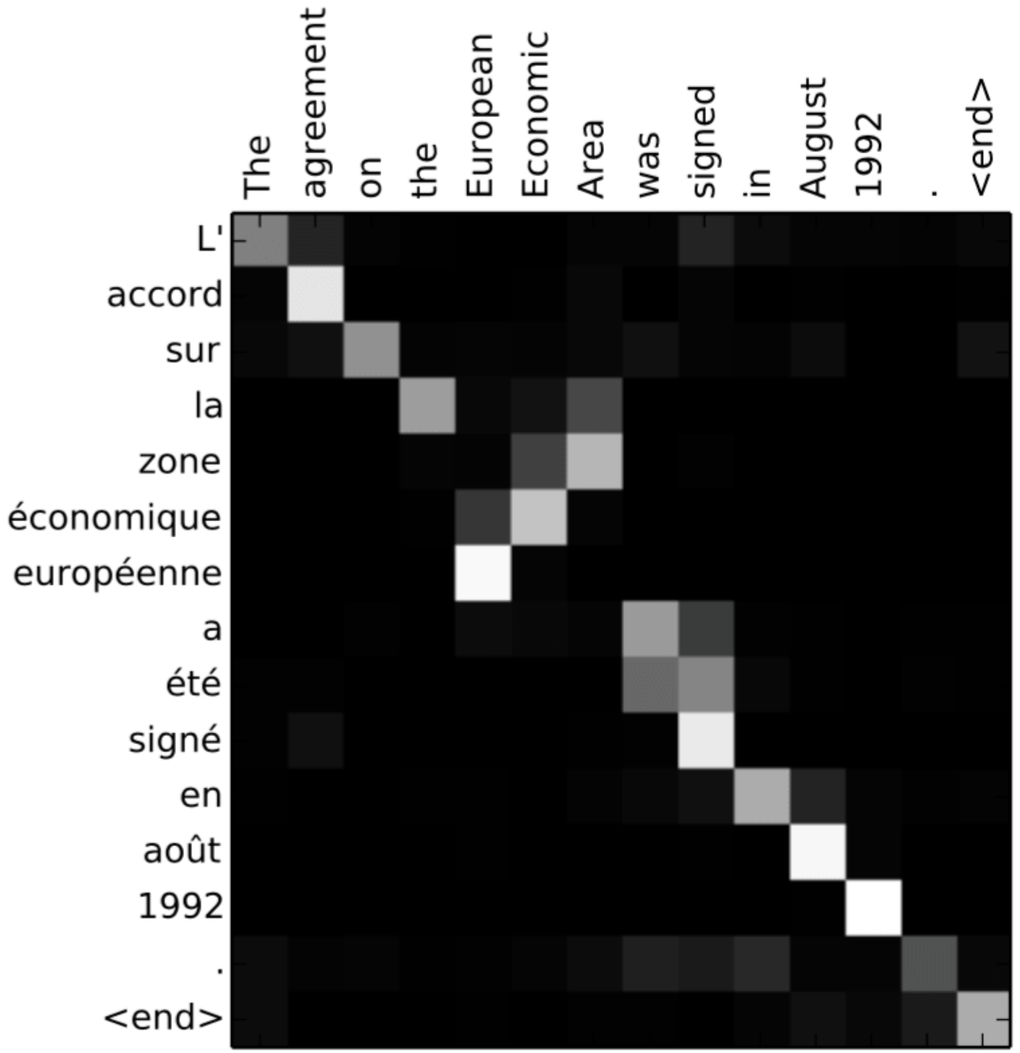
Attention Mechanism



Attention Mechanism

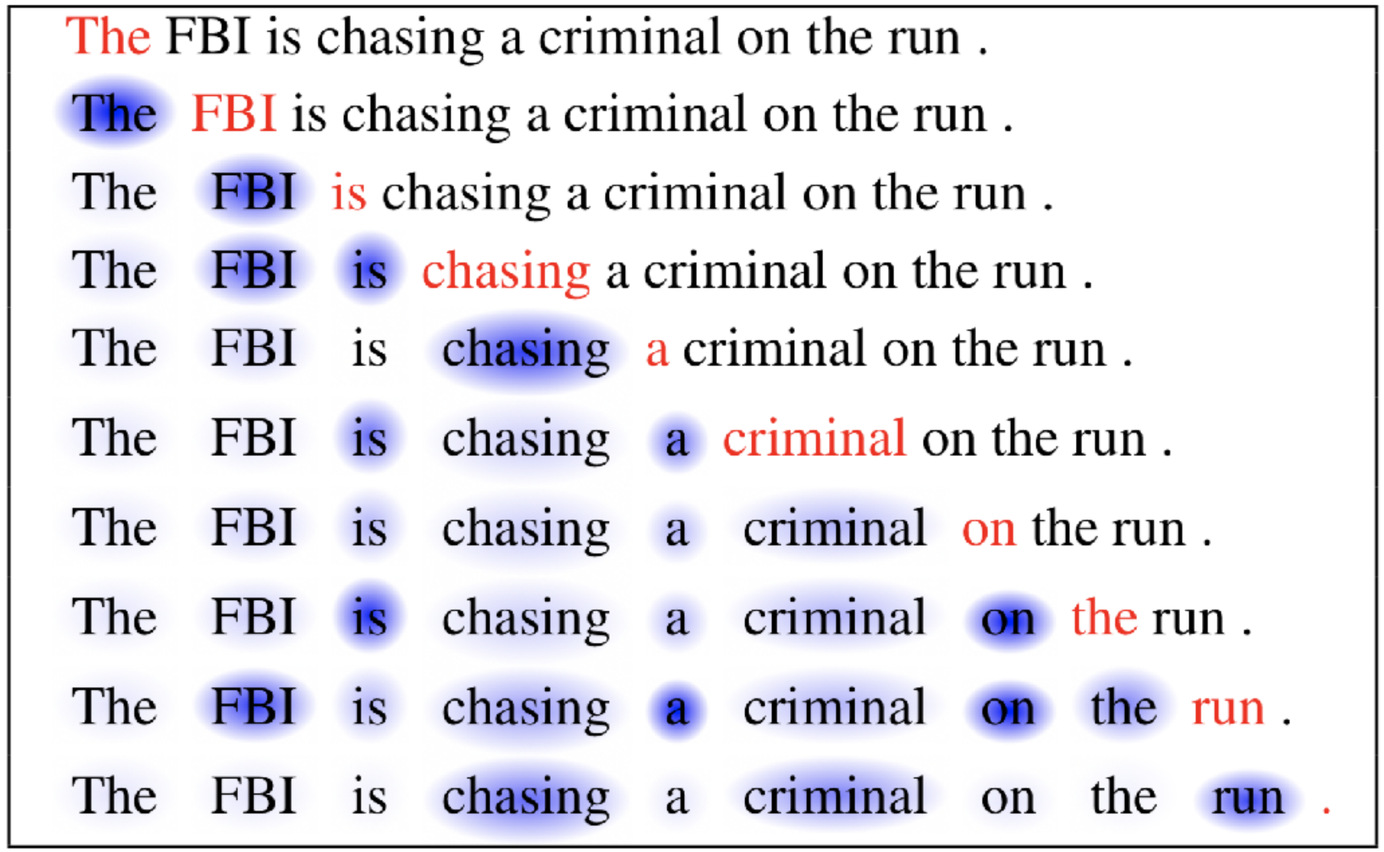
What does attention learn?

Bahdanau, et al (2015):
By visualizing attention weights, we see they learn **word alignments**.



Self-Attention

Self-Attention allows the inputs to interact with each other.

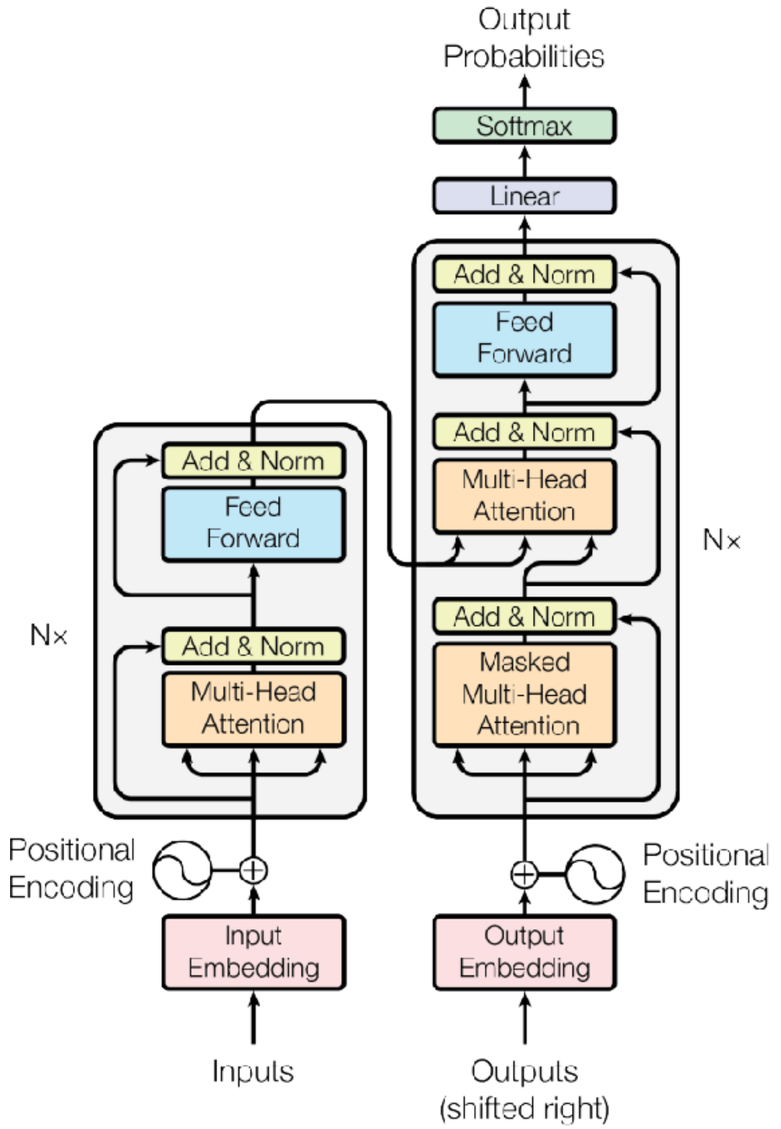


Cheng, Jianpeng, Li Dong, and Mirella Lapata. "Long short-term memory-networks for machine reading."

Transformer Architecture

Vaswani et al. (2017)

- Encoder Self-Attention
- Feed-Forward Networks
- Add & Norm
- Cross Attention / Enc-Dec Attention
- Decoder Masked Attention



Transformer Architecture

Vaswani et al. (2017)

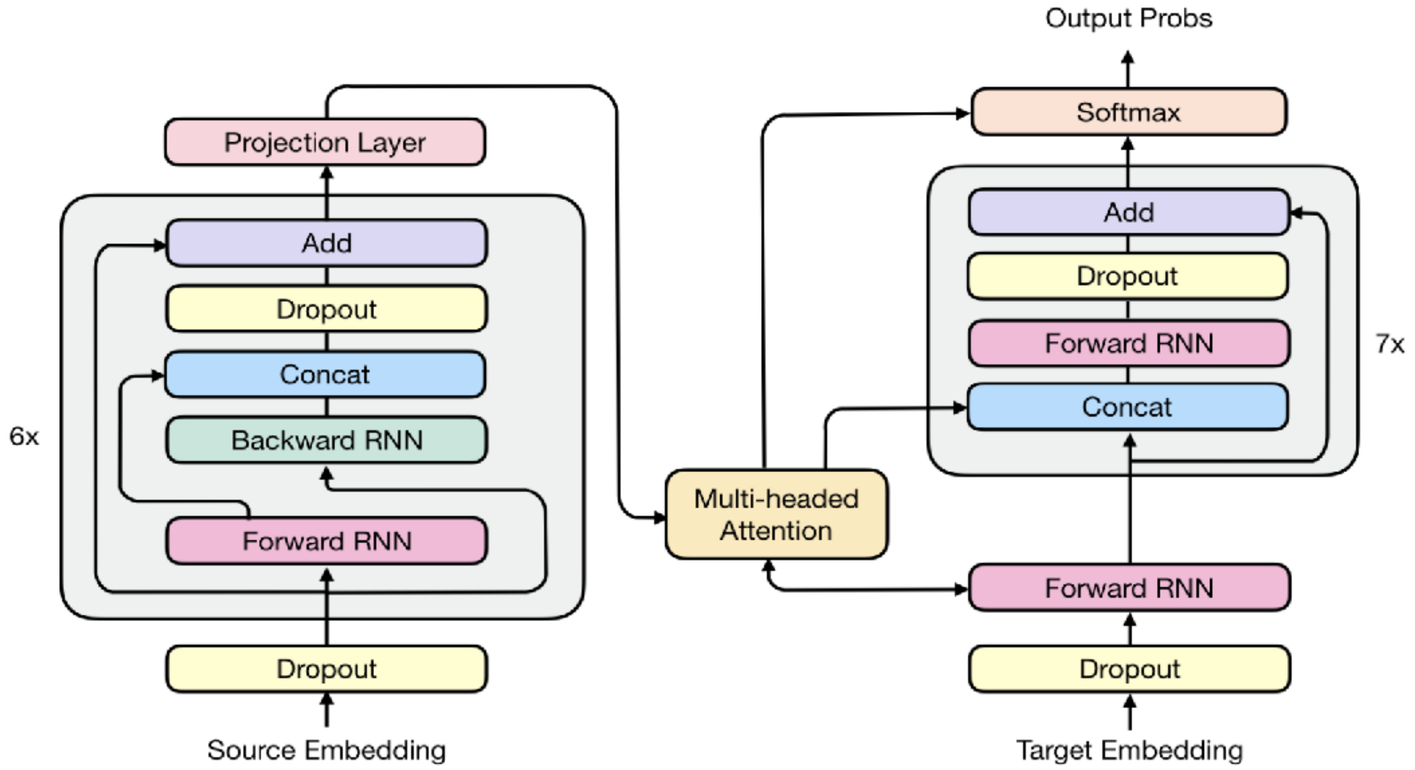
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Transformer vs RNMT+ (RNN instead of Self-Attn)

Chen et al. (2018)

Transformer: Attention is all you need.

But, is Attention the main reason for improved performance?



Replace self-attention by RNNs

Transformer vs RNMT+ (RNN instead of Self-Attn)

Chen et al. (2018)

Translation Quality

Model	Test BLEU	Epochs	Training Time
GNMT	38.95	-	-
ConvS2S [7]	39.49 ± 0.11	62.2	438h
Trans. Base	39.43 ± 0.17	20.7	90h
Trans. Big [8]	40.73 ± 0.19	8.3	120h
RNMT+	41.00 ± 0.05	8.5	120h

Inference Efficiency

Model	Examples/s	FLOPs	Params
ConvS2S	80	15.7B	263.4M
Trans. Base	160	6.2B	93.3M
Trans. Big	50	31.2B	375.4M
RNMT+	30	28.1B	378.9M

Decoding in Machine Translation

Our goal is to find the argmax of the target sentence.

$$y' = \underset{y}{\operatorname{argmax}}\ p(y \mid x) = \underset{y}{\operatorname{argmax}} \prod_{t=1}^{n} P(y_t \mid y_{<t}, x)$$

Greedy Decoding: select the most probable token at each step.

Decoding in Machine Translation

Our goal is to find the argmax of the target sentence.

$$y' = \underset{y}{\text{argmax}}\ p(y \mid x) = \underset{y}{\text{argmax}} \prod_{t=1}^{n} P(y_t \mid y_{<t}, x)$$

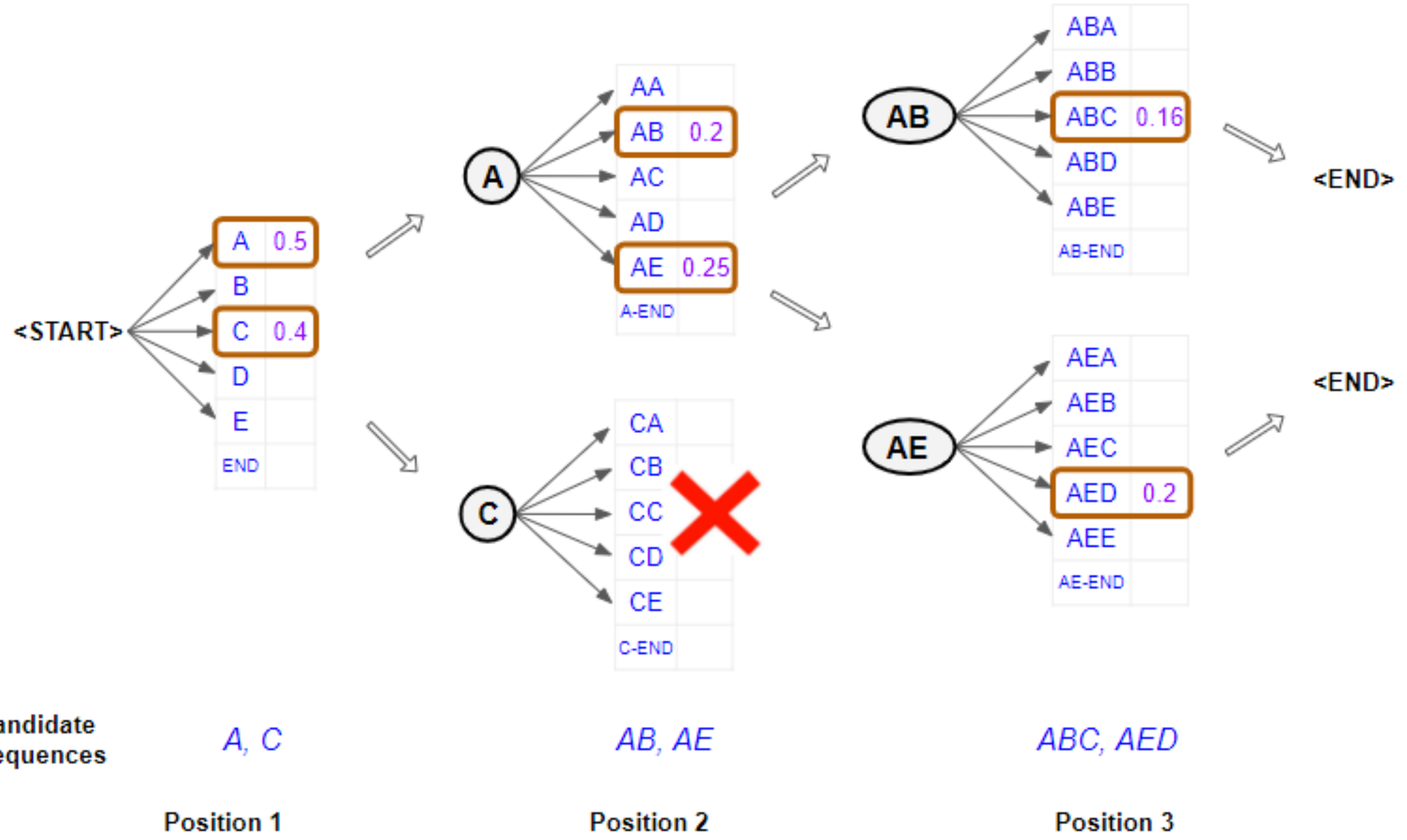
Greedy Decoding: select the most probable token at each step.

$$\underset{y}{\text{argmax}} \prod_{t=1}^{n} P(y_t \mid y_{<t}, x) \neq \prod_{t=1}^{n} \underset{y_t}{\text{argmax}} P(y_t \mid y_{<t}, x)$$

The “best” token at the current step does not necessarily lead to the best sequence.

Decoding in Machine Translation

Beam Decoding / Search: explore multiple alternatives in parallel.



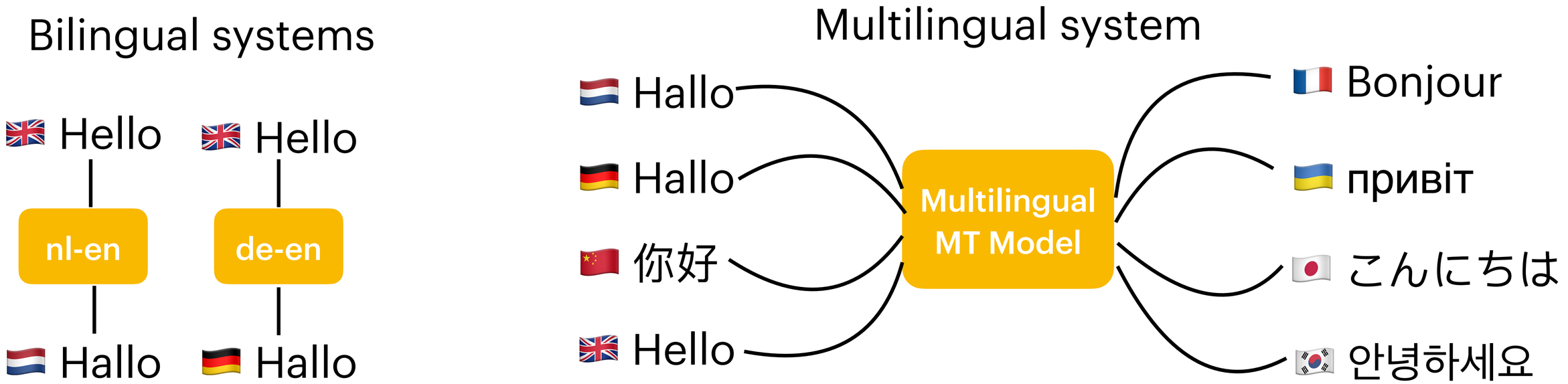
Multilingual Machine Translation

SHAOMU TAN |  UNIVERSITY OF AMSTERDAM
Language Technology Lab

Multilingual Neural Machine Translation (MNMT)

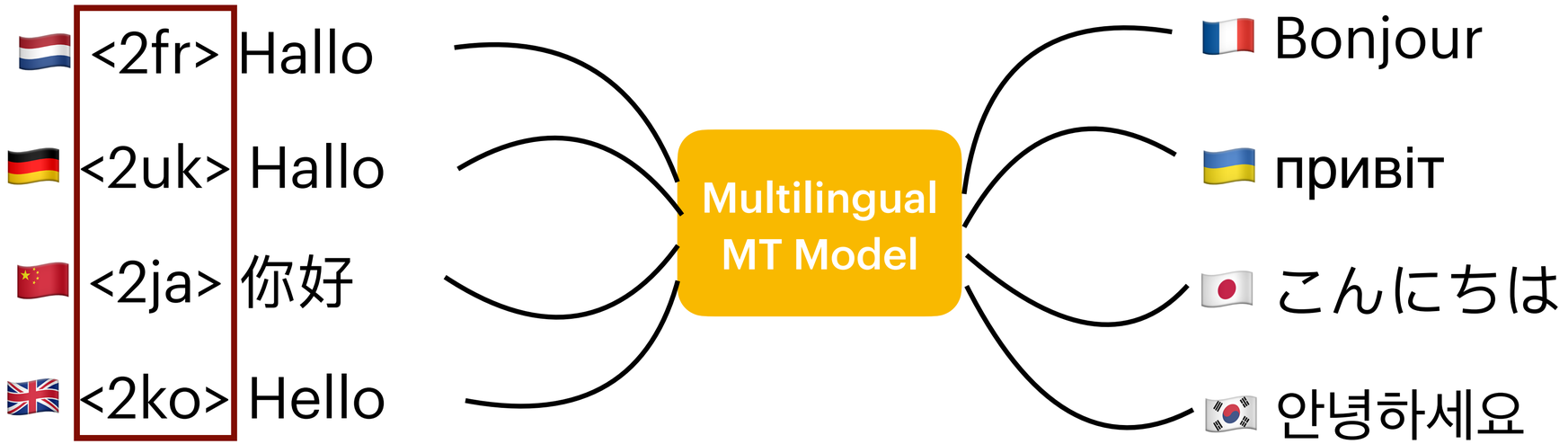
-> Training a unified model on a mixed dataset from multiple languages.

-> Efficient: One model for all languages.



Specifying Languages in MNMT

Johnson et al (2016): Simply adding a language tag to specify target languages.



Multilingual Vocabulary

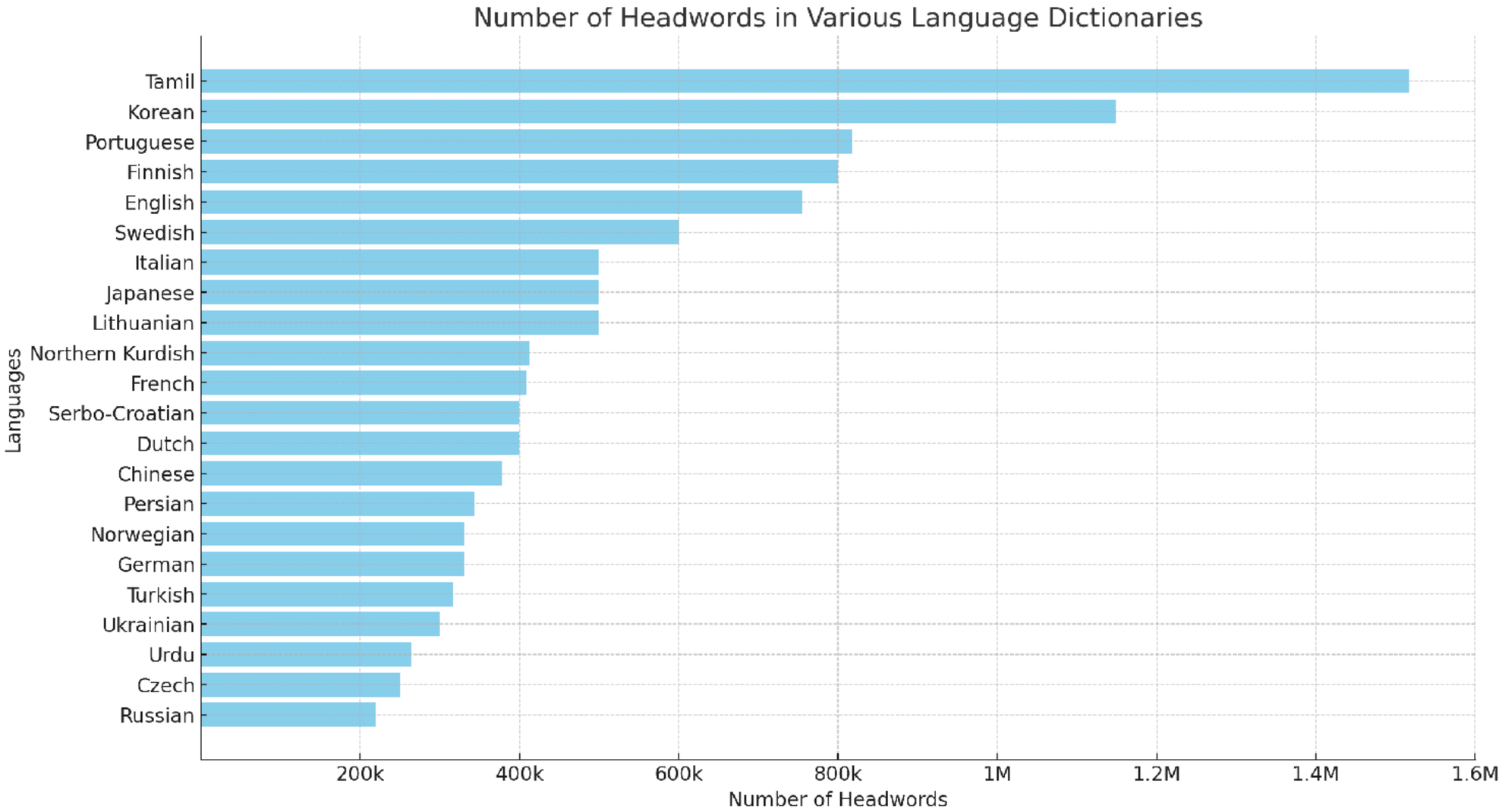
Multilingual Models

Joint Multilingual Vocab

Bilingual Models

En Vocab | De Vocab | Nl Vocab | Zh Vocab | Ja Vocab | Ar Vocab

Vocabulary Problem



Sub-word Segmentation

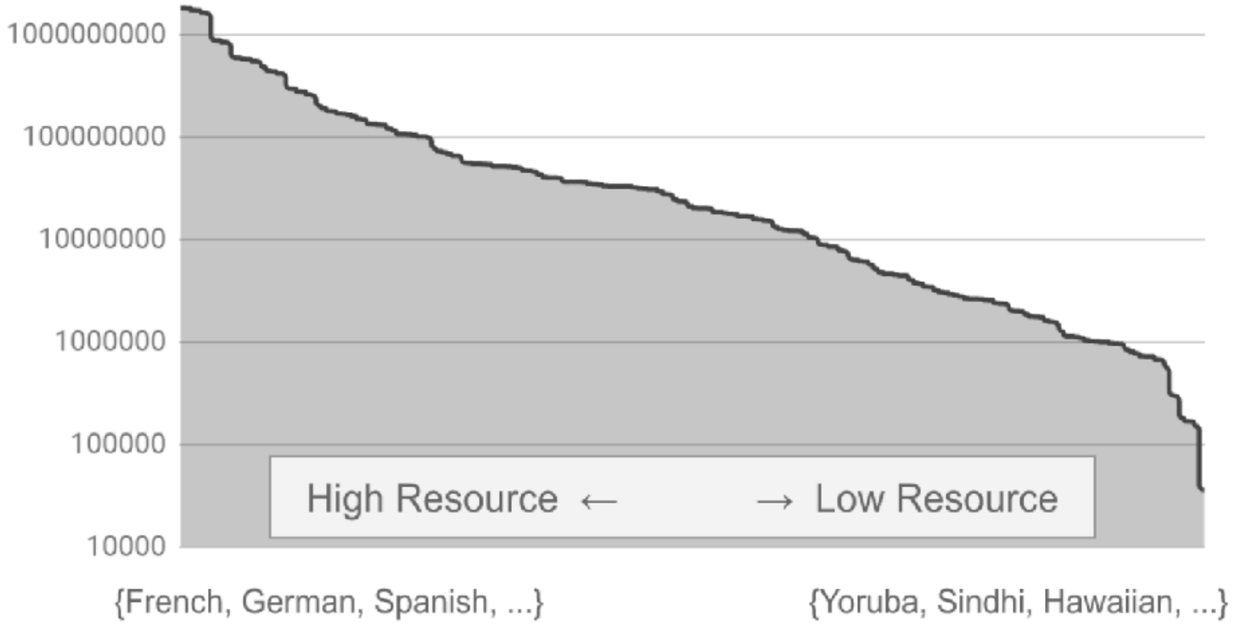
- Spanish:
  - transformación
- Portuguese:
  - transformação
- Italian:
  - trasformazione

- Common subwords:
  - "transform"

Sub-word solves Out-Of-Vocab

- Spanish:
  - _transform ación
- Portuguese:
  - _transform ação
- Italian:
  - _trasform azione

Imbalanced Data in Multilingual Machine Translation



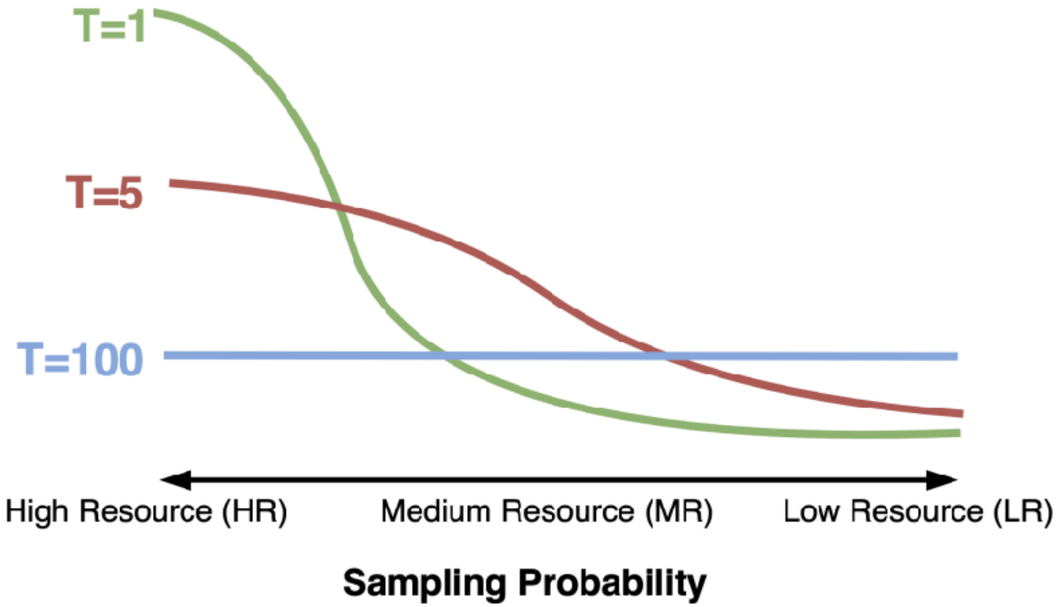
High-resource Languages have much more data than low-resource ones.
Deep learning is still a data-driven approach. -> more data, better performance.

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Imbalanced Data in Multilingual Machine Translation

Temperature Sampling

$$\Pr(v_k) = \frac{e^{l_k/T}}{\sum_i e^{l_i/T}}$$

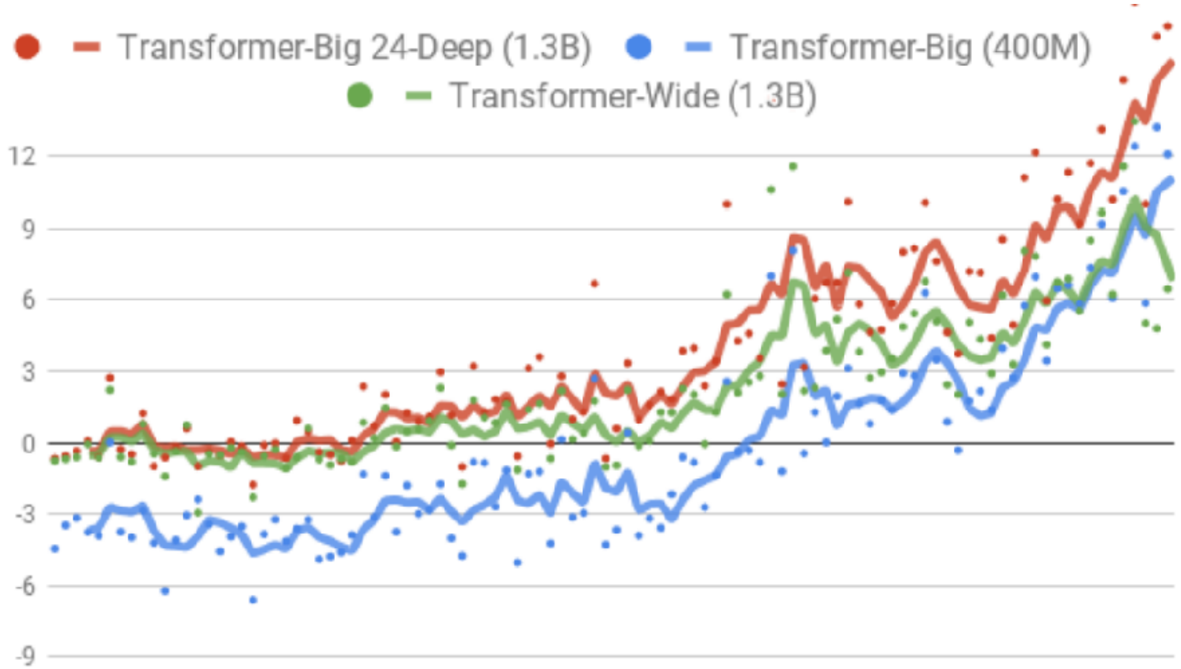


Up-sampling low-resource data to balance the data distribution.

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

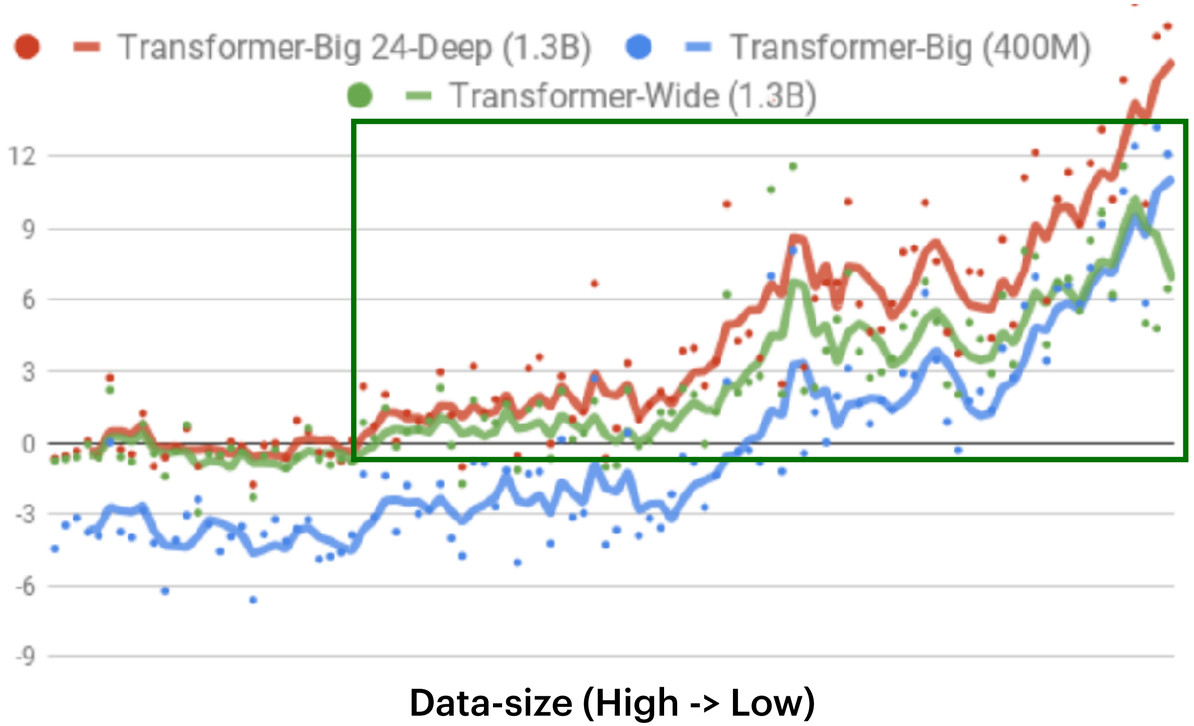
facilitates Knowledge Transfer



Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

facilitates Knowledge Transfer



Knowledge transfer benefits **low-resource** languages

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

is not a free lunch

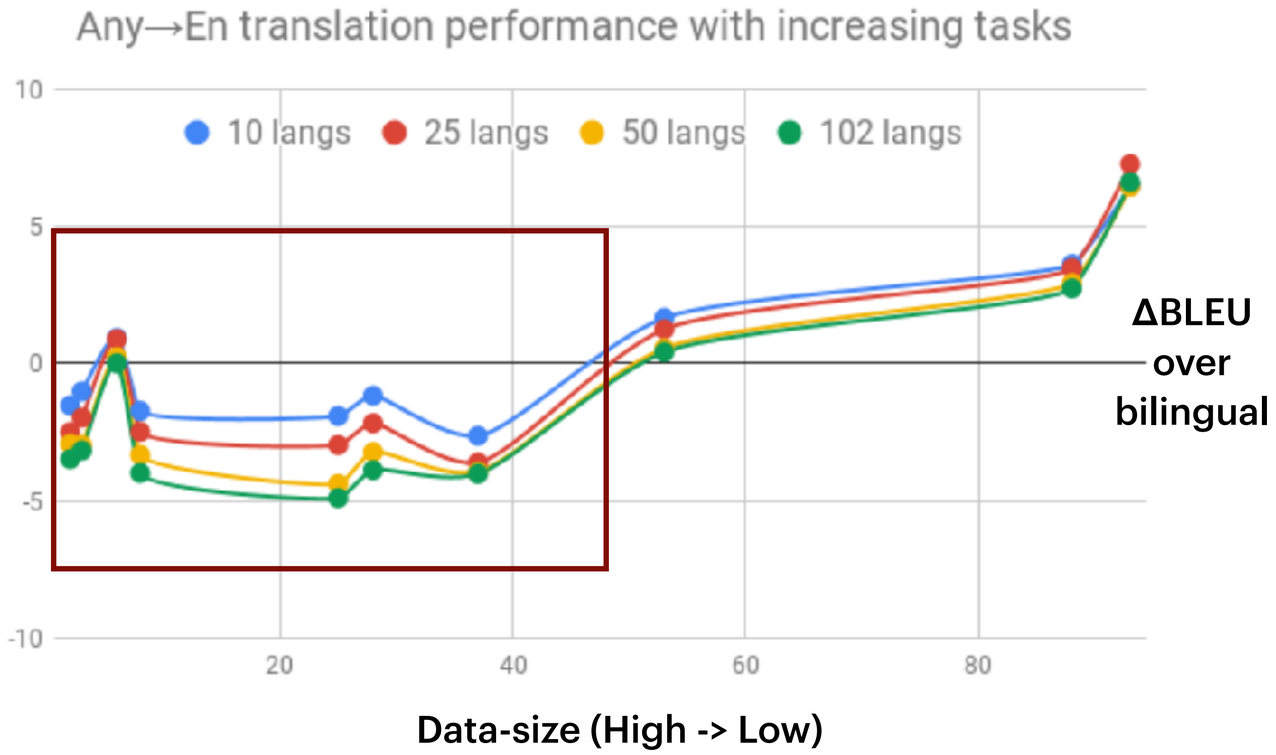
Joint Multilingual Training brings Synergy
but also **Interference** (negative transfer)

Multilingual Machine Translation

is not a free lunch

Interference

compromises performance
(Especially for high-resource languages)



Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

is not a free lunch

Interference

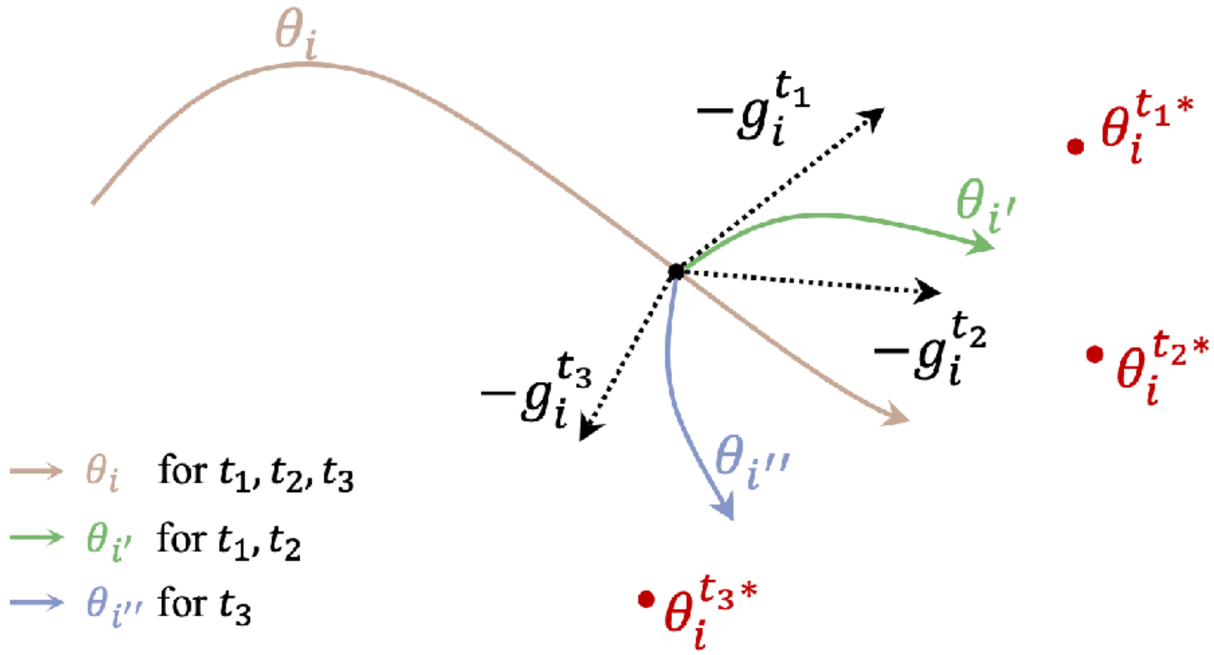
rooted in **Conflicting** optimization demands of various tasks

Multilingual Machine Translation

is not a free lunch

Interference

rooted in **Conflicting** optimization demands of various tasks



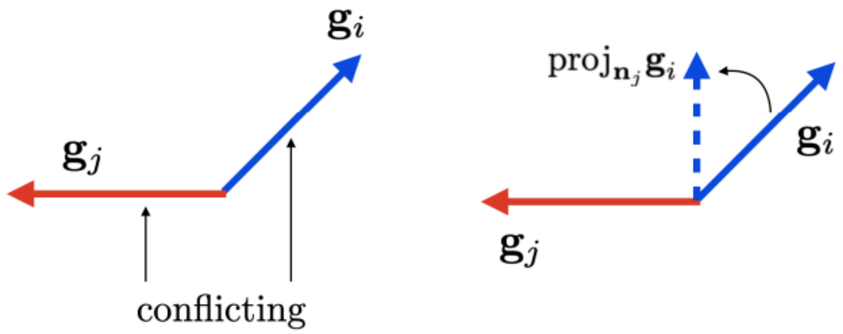
Gradient Conflicts

Wang, Qian, and Jiajun Zhang. "Parameter differentiation based multilingual neural machine translation."

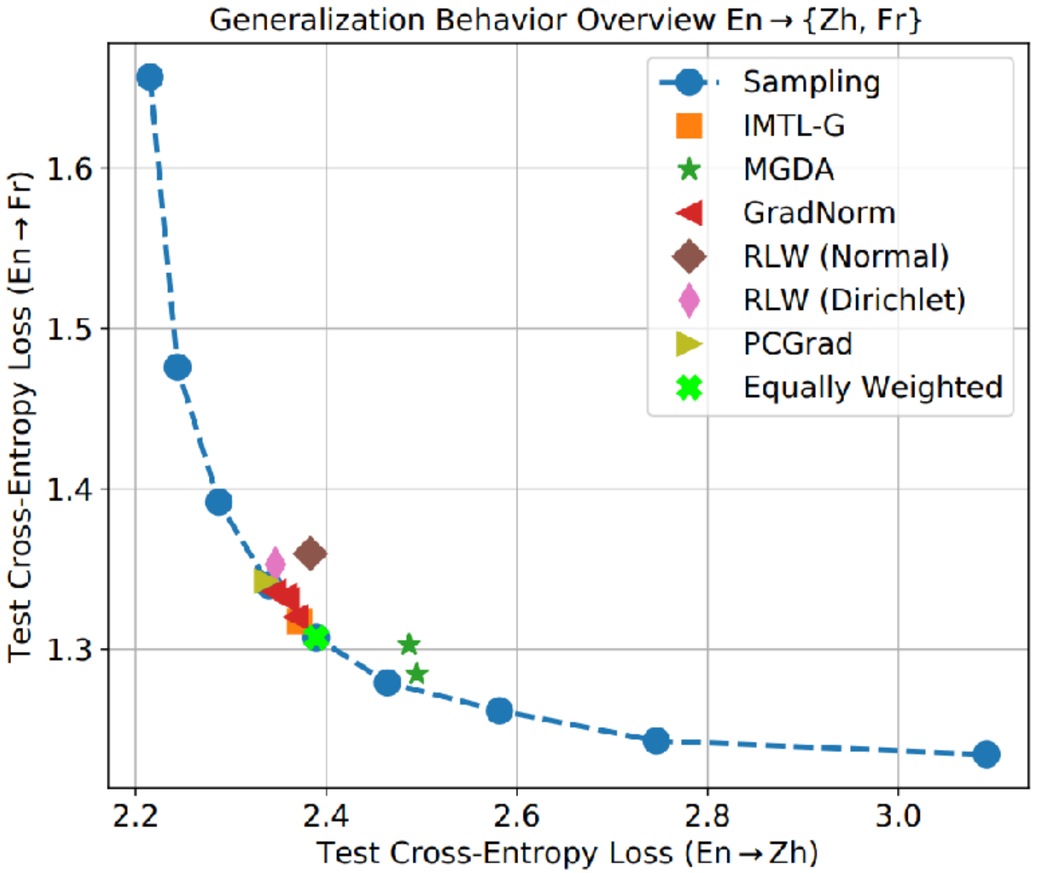
Recent work in Reducing Interference

How to Reduce Interference

Multi-task Optimization



Gradient-Sergery to project conflicted task gradient onto the normal plane

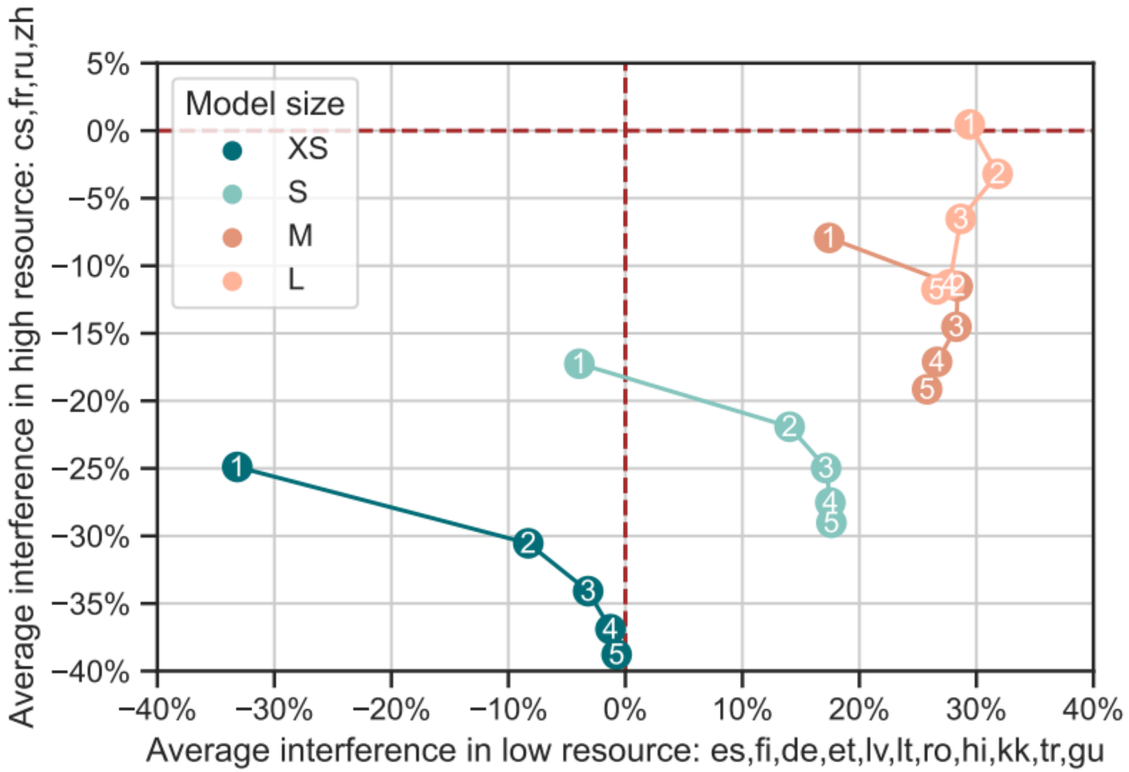


Yu, Tianhe, et al. "Gradient surgery for multi-task learning."

Xin, Derrick, et al. "Do current multi-task optimization methods in deep learning even help?."

How to Reduce Interference

Scaling up?



Scaling up model size alleviate Interference

Shaham, Uri, et al. "Causes and cures for interference in multilingual translation."

How to Reduce Interference

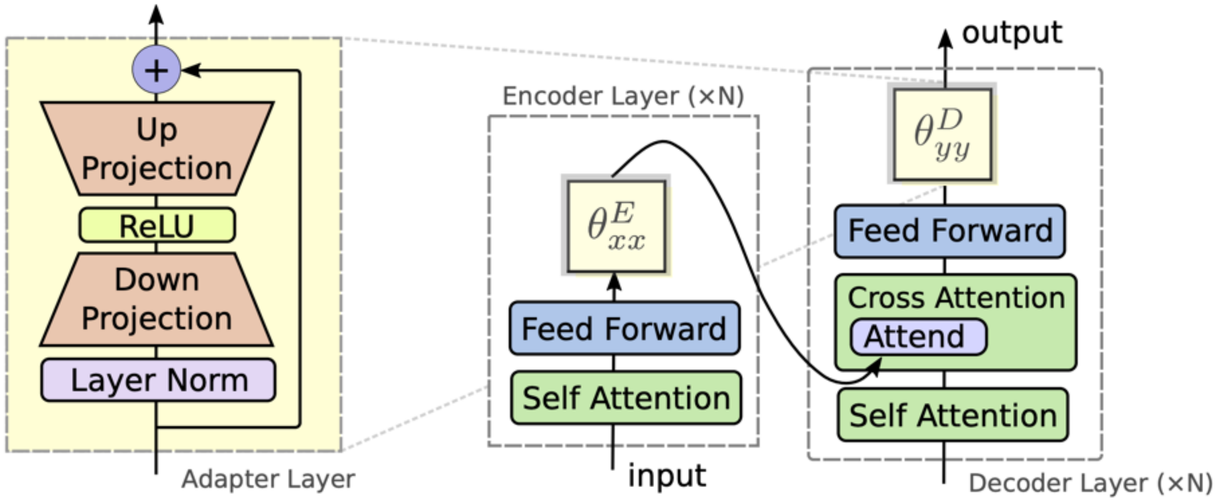
Modular Deep Learning

Multilingual Model

Language-Specific (Ls) Modules

How to Reduce Interference

Modular Deep Learning

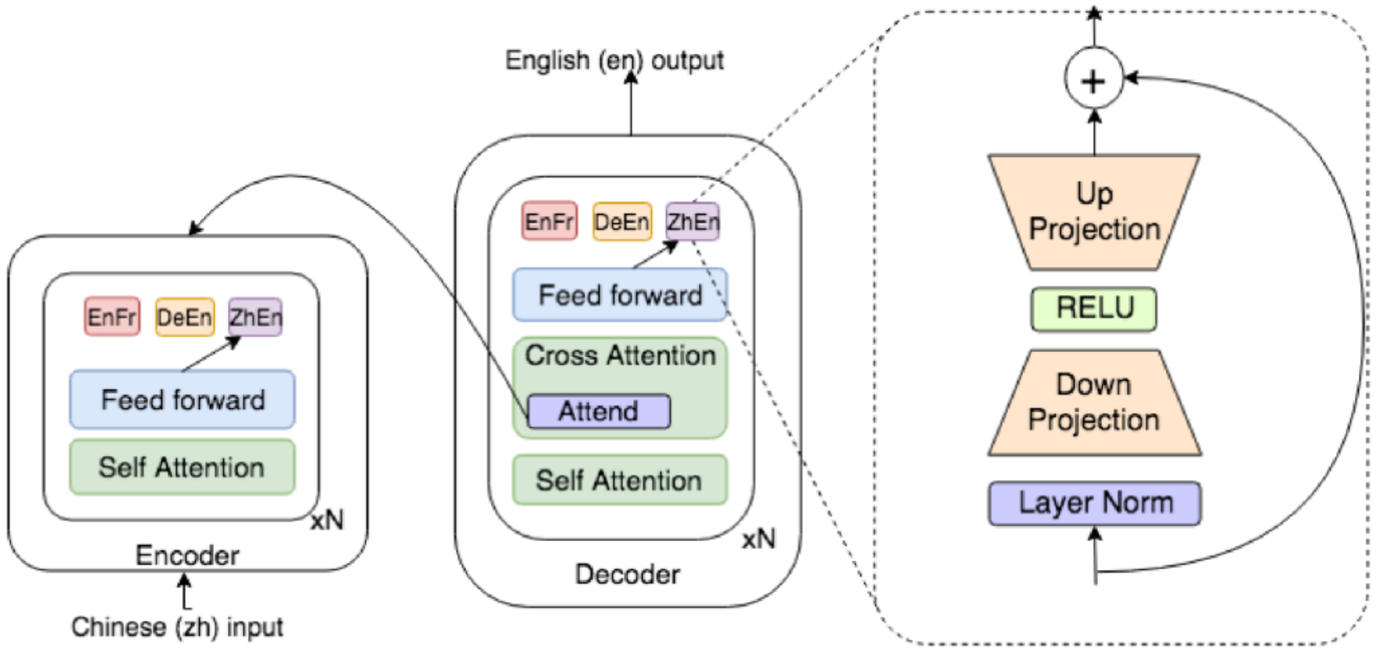


Adapters: light-weight module that can be inserted into pre-trained Transformer models

Philip, Jerin, et al. "Monolingual adapters for zero-shot neural machine translation."

How to Reduce Interference

Modular Deep Learning

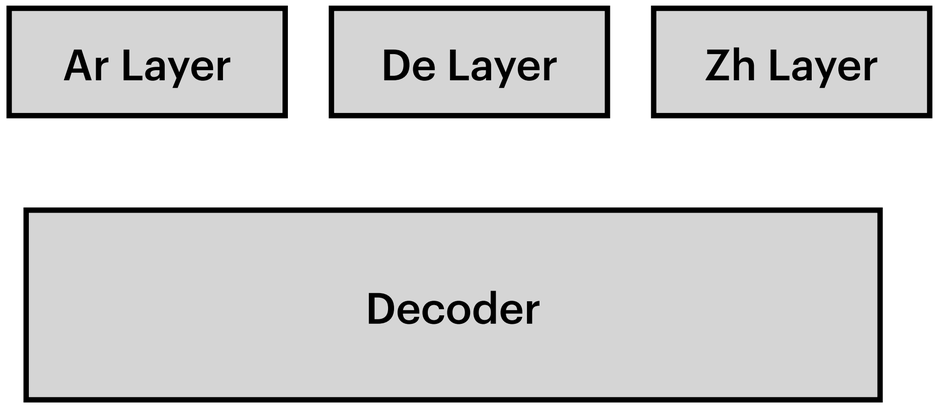


Language Pair Adapters: insert adapters conditioned on language pairs to add language-specific capacities.

Bapna, Ankur, and Orhan Firat. "Simple, Scalable Adaptation for Neural Machine Translation."

How to Reduce Interference

Modular Deep Learning



Language-Dependent FFN[1] or Norm[2] Layer

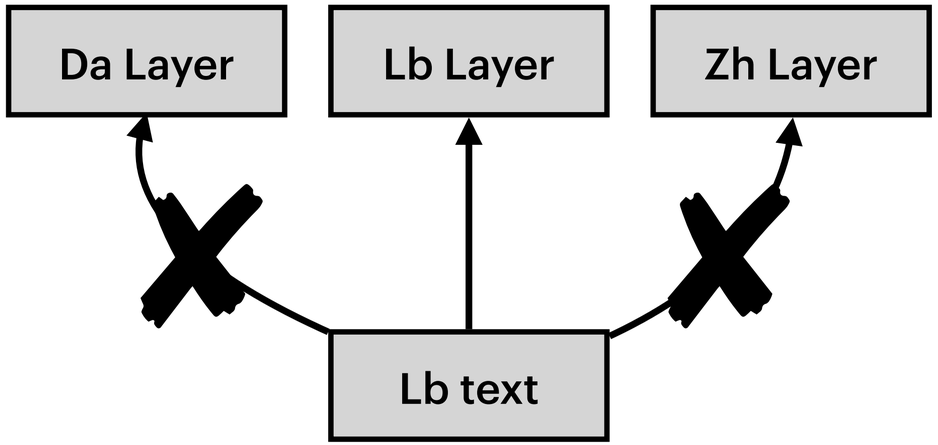
Remain efficacy when scaling up[1]

1) Fan, Angela, et al. "Beyond english-centric multilingual machine translation."
2) Zhang, Biao, et al. "Improving massively multilingual neural machine translation and zero-shot translation."

How to Reduce Interference

Limitations - Modular Deep Learning

Adapters, Language-Specific Modules, are **Language-Dependent** that **operates in isolation**



Such Design fundamentally **dis-encourages cross-lingual Transfer** especially for low-resource languages

How to Reduce Interference

Limitations - Modular Deep Learning

Trade-Off: Efficiency & Performance

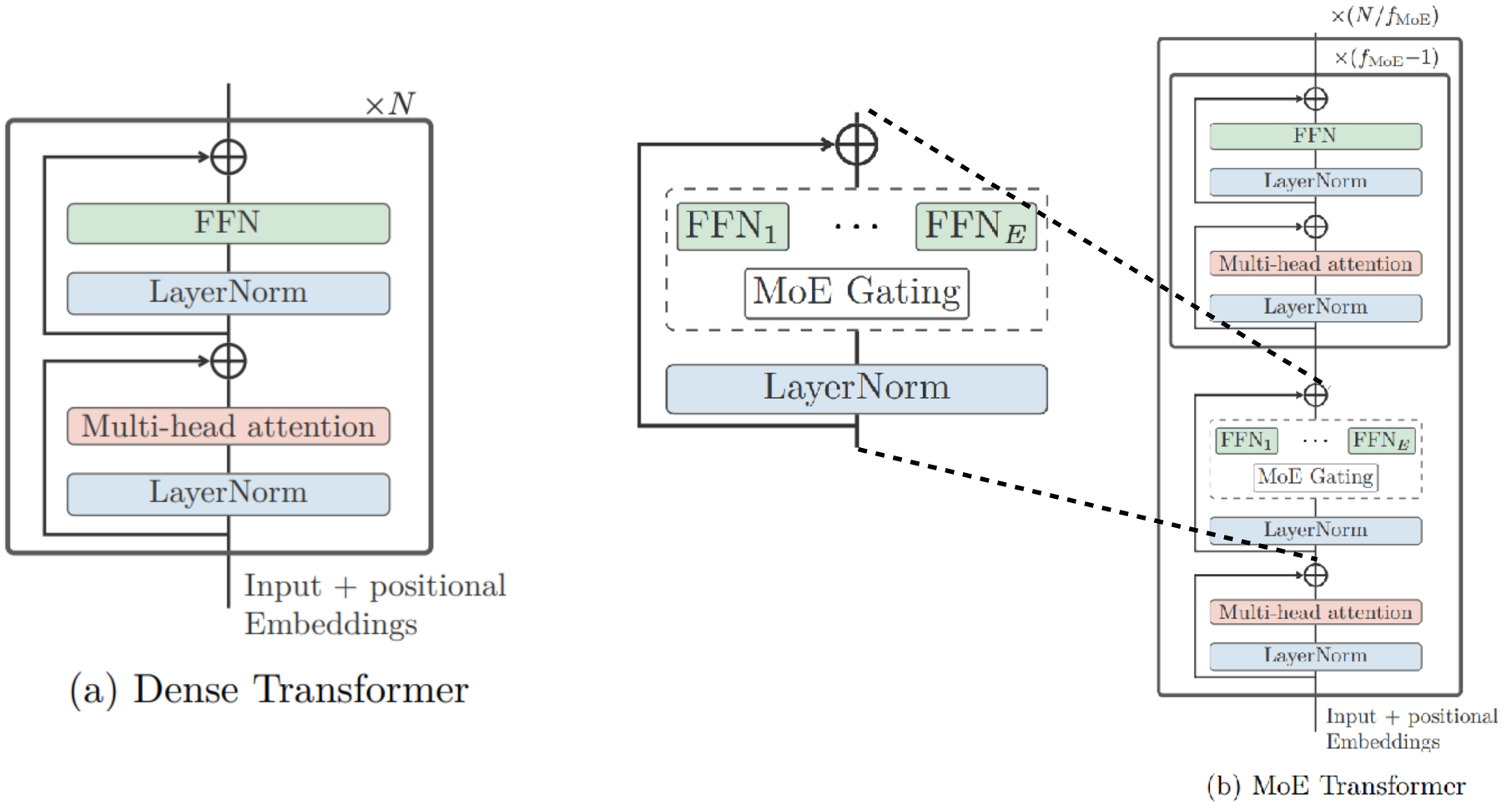
a. increase substantial parameters when many languages are involved

b. memory[1] and latency[2] issue

1) Liao, Baohao, Shaomu Tan, and Christof Monz. "Make your pre-trained model reversible: From parameter to memory efficient fine-tuning."
2) Liao, Baohao, Yan Meng, and Christof Monz. "Parameter-efficient fine-tuning without introducing new latency."

How to Reduce Interference

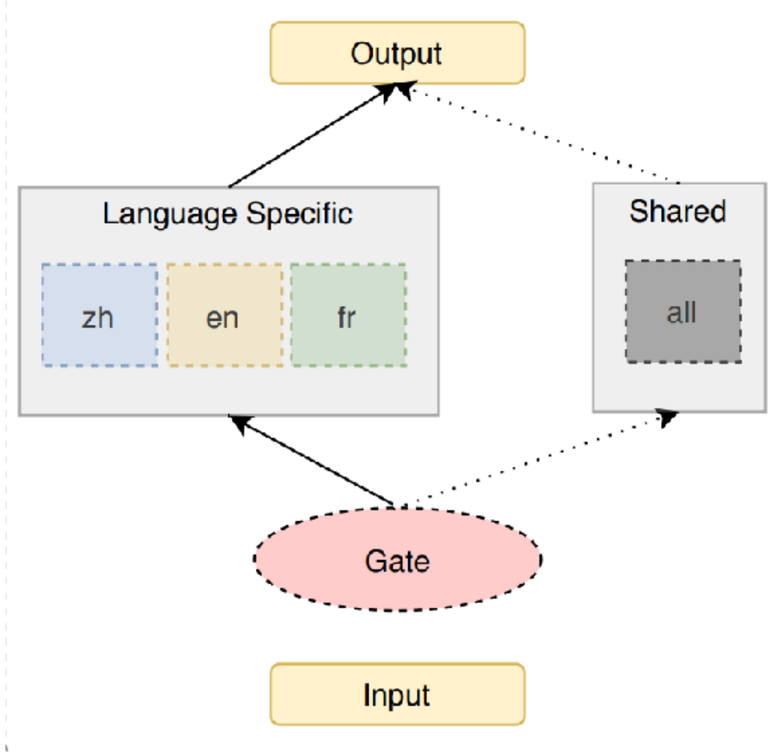
Mixture-of-Experts (MoE)



Costa-jussà, Marta R., et al. "No language left behind: Scaling human-centered machine translation."

How to Reduce Interference

Router + LS Module

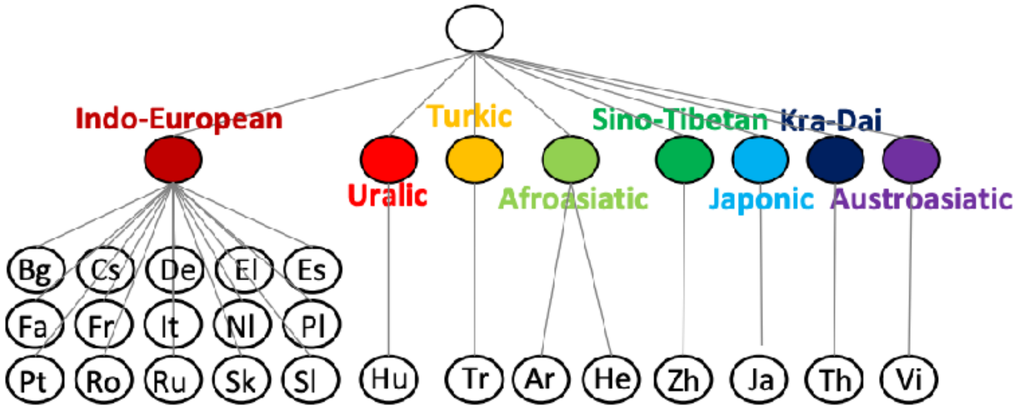


Mix of shared and language-specific Network

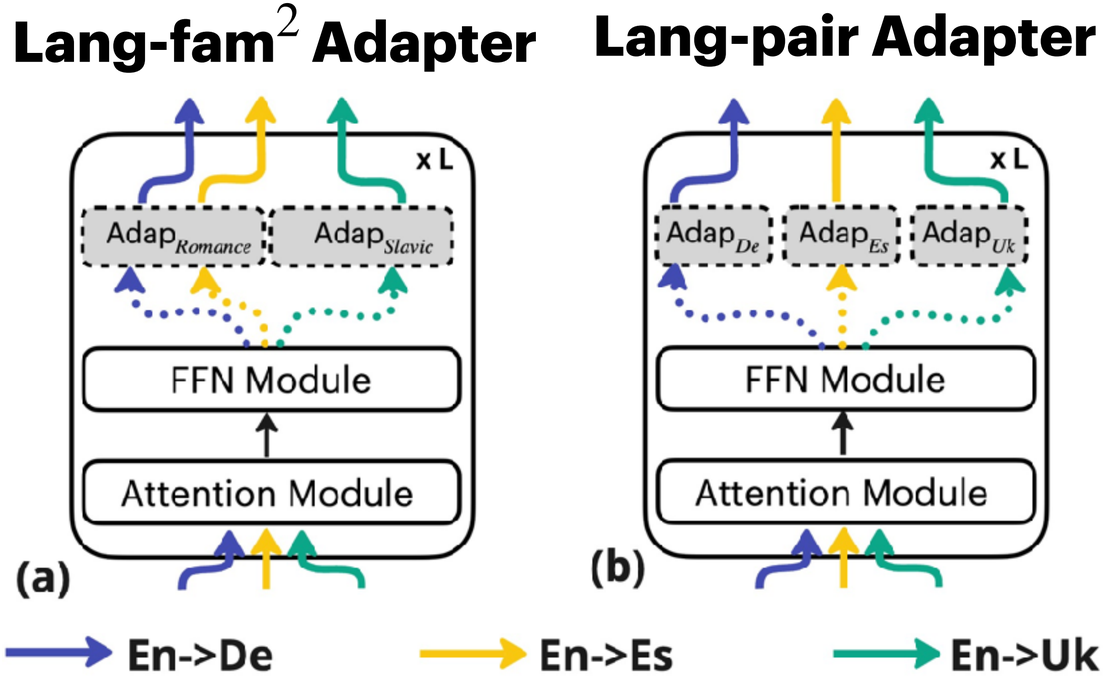
Zhang, Biao, et al. "Share or not? learning to schedule language-specific capacity for multilingual translation."

How to Reduce Interference

Leveraging Priori Linguistic Knowledge



Language cluster Training[1]**:** Train one multilingual model for one language cluster.



(1) Tan, Xu, et al. "Multilingual neural machine translation with language clustering."
(2) Chronopoulou, et al "Language-family adapters for low-resource multilingual neural machine translation."

How to Reduce Interference

Limitations - Leveraging Priori Linguistic Knowledge

a. Heavily rely on priori knowledge, e.g.: linguistic knowledge.

b. lack clear inductive bias, thus heavy reliance on heuristics.

c. show strong effects for low-resource languages, less or no effects on high-resource ones.

Neuron Specialization

Leveraging Intrinsic Task Modularity for Multilingual Machine Translation

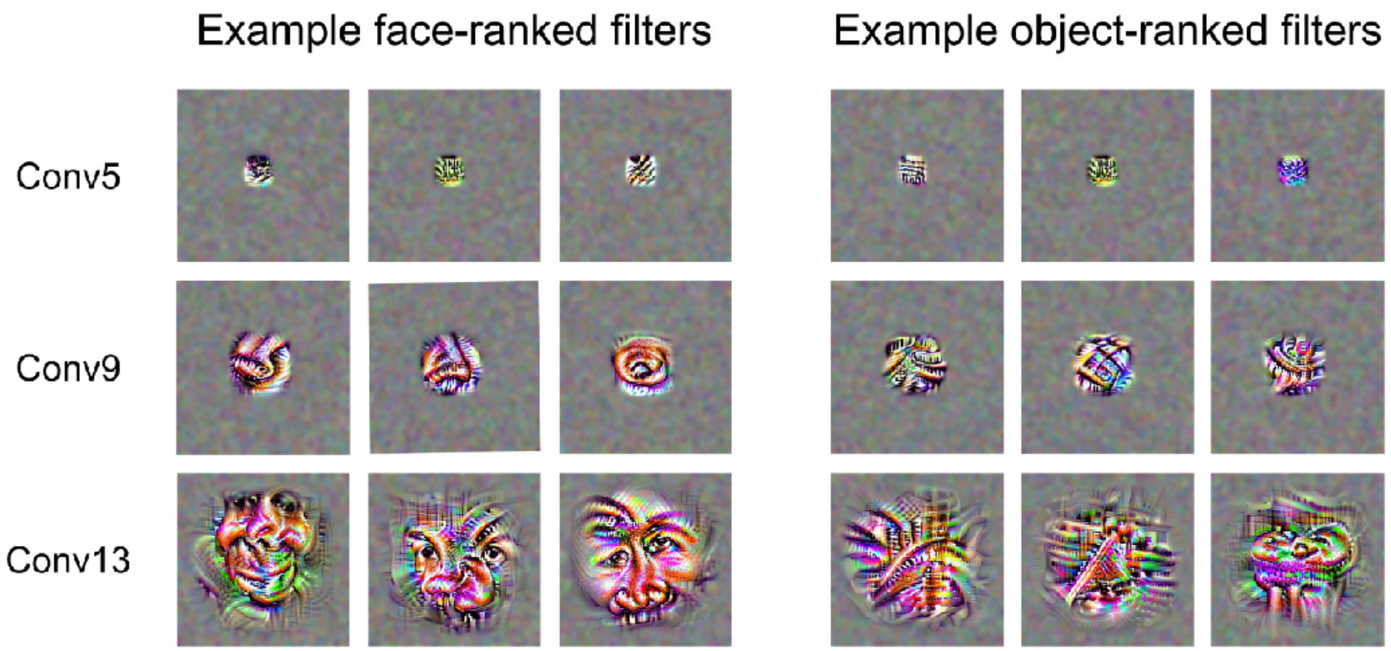
— Shaomu Tan, Di Wu, Christof Monz

SHAOMU TAN |  UNIVERSITY OF AMSTERDAM
Language Technology Lab

Intrinsic Modularity

in Multi-task Networks

Intrinsic Modularity
in Multi-task Vision Networks



Multi-task training develops task-specific **functional specialization**

Dobs, Katharina, et al. "Brain-like functional specialization emerges spontaneously in deep neural networks." *Science advances*

Locating Intrinsic Modularity

in MNMT Models

Prior Studies attempts to identify Task-Specific **Sub-networks** inside trained Multi-task Models

Locating Intrinsic Modularity in MNMT Models

Prior Studies attempts to identify Task-Specific **Sub-networks** inside trained Multi-task Models

Lottery Ticket Training:
Utilizing **Iterative Pruning** to identify important components[1]

1) Foroutan, Negar, et al. "Discovering language-neutral sub-networks in multilingual language models."

Locating Intrinsic Modularity in MNMT Models

Prior Studies attempts to identify Task-Specific **Sub-networks** inside trained Multi-task Models

Lottery Ticket Training:
Utilizing Iterative Pruning to identify important components

LaSS: Fine-tuning tasks to see what parameters changed the most[1,2,3]

1) Lin, Zehui, et al. "Learning language specific sub-network for multilingual machine translation."
2) He, Dan, et al. "Gradient-based Gradual Pruning for Language-Specific Multilingual Neural Machine Translation."
3) Choenni, Rochelle, et al. "Cross-Lingual Transfer with Language-Specific Subnetworks for Low-Resource Dependency Parsing."

Locating Intrinsic Modularity
requires Network Modifications

Fine-tuning approaches (LaSS) raise a question:

whether the modularity is inherent to the original model, or simply an **artifact** introduced by network modifications

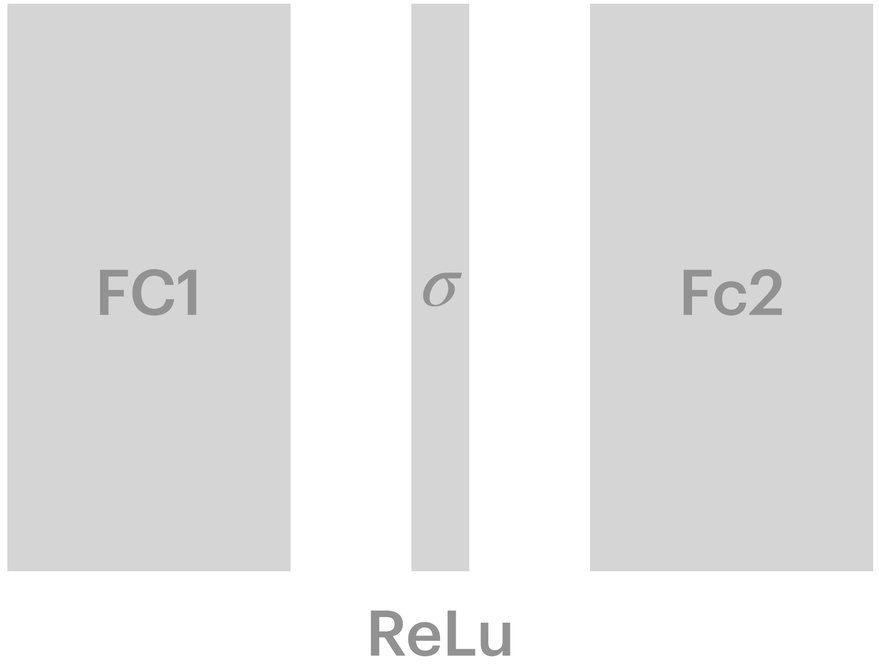
Does Intrinsic Modularity even exist?

Analyzing task Modularity in Multi-task models

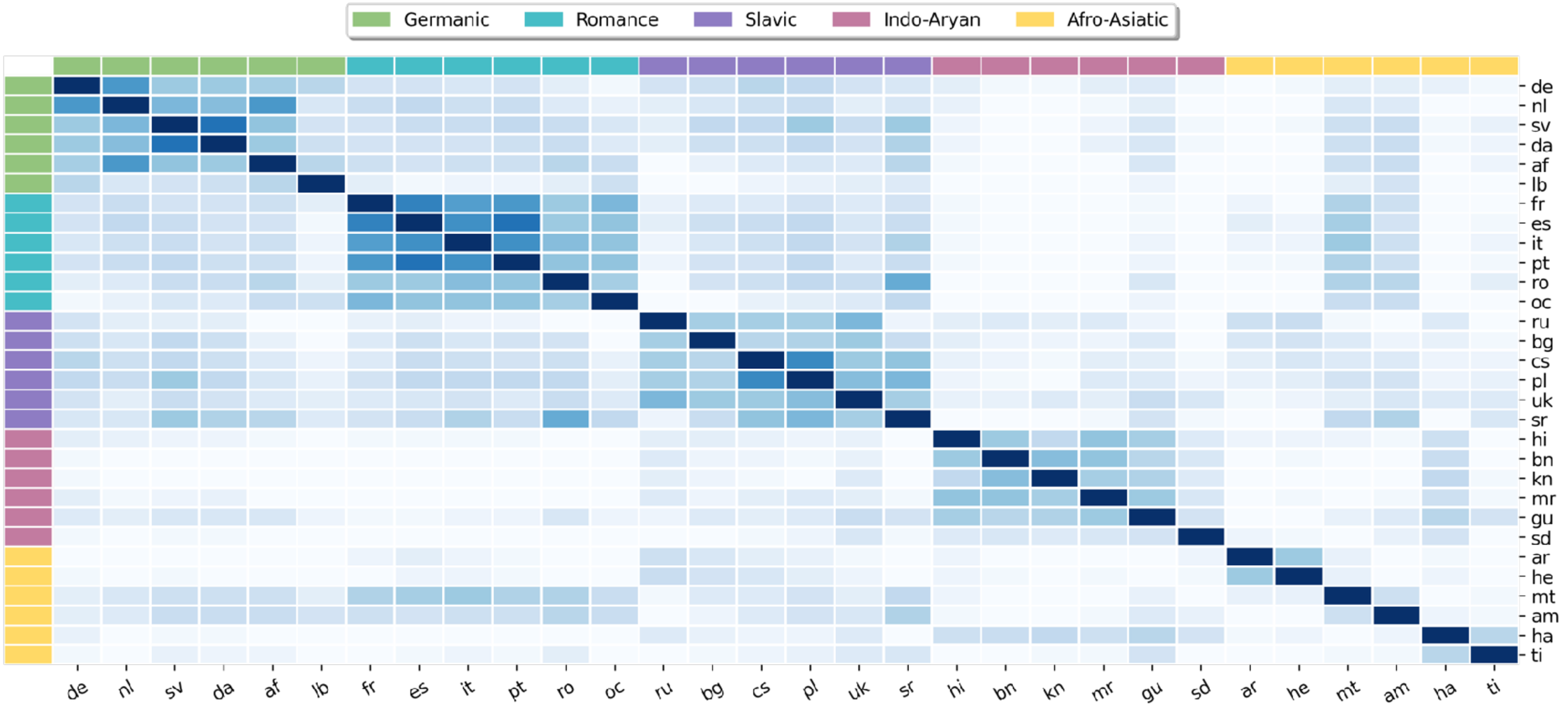
Neuron Specialization

Neuron Structural Analysis - Method

We focus on Neurons:
intermediate activations inside the feed-forward (FFN) blocks

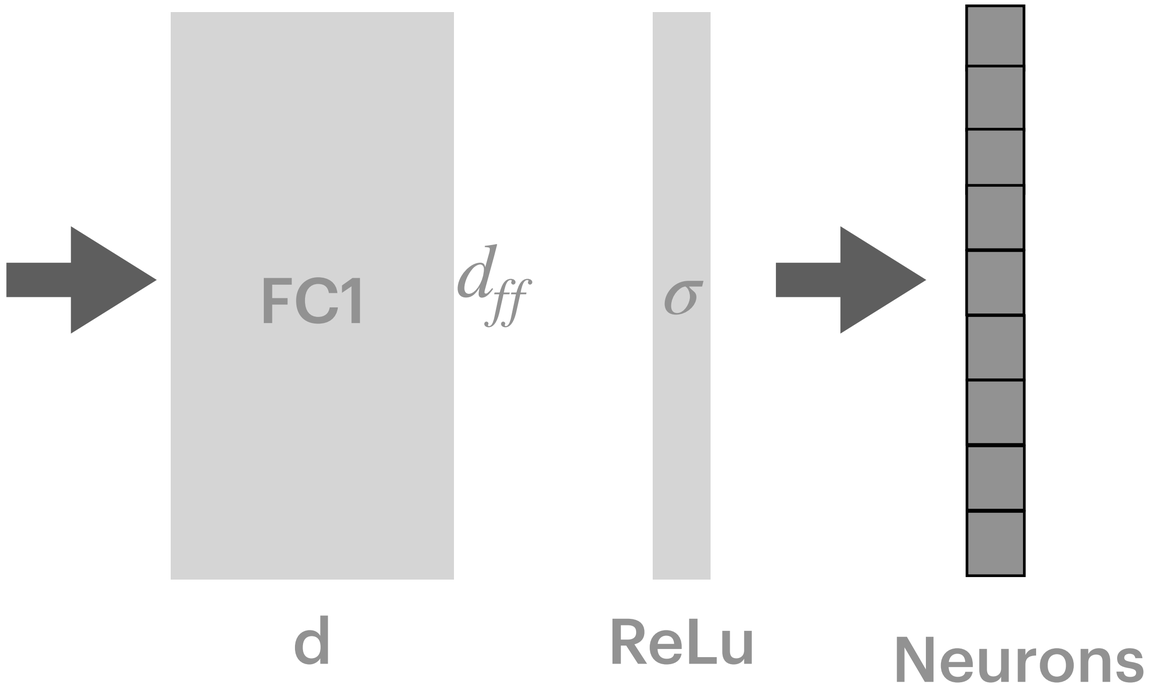


Neuron Specialization



Neuron Specialization

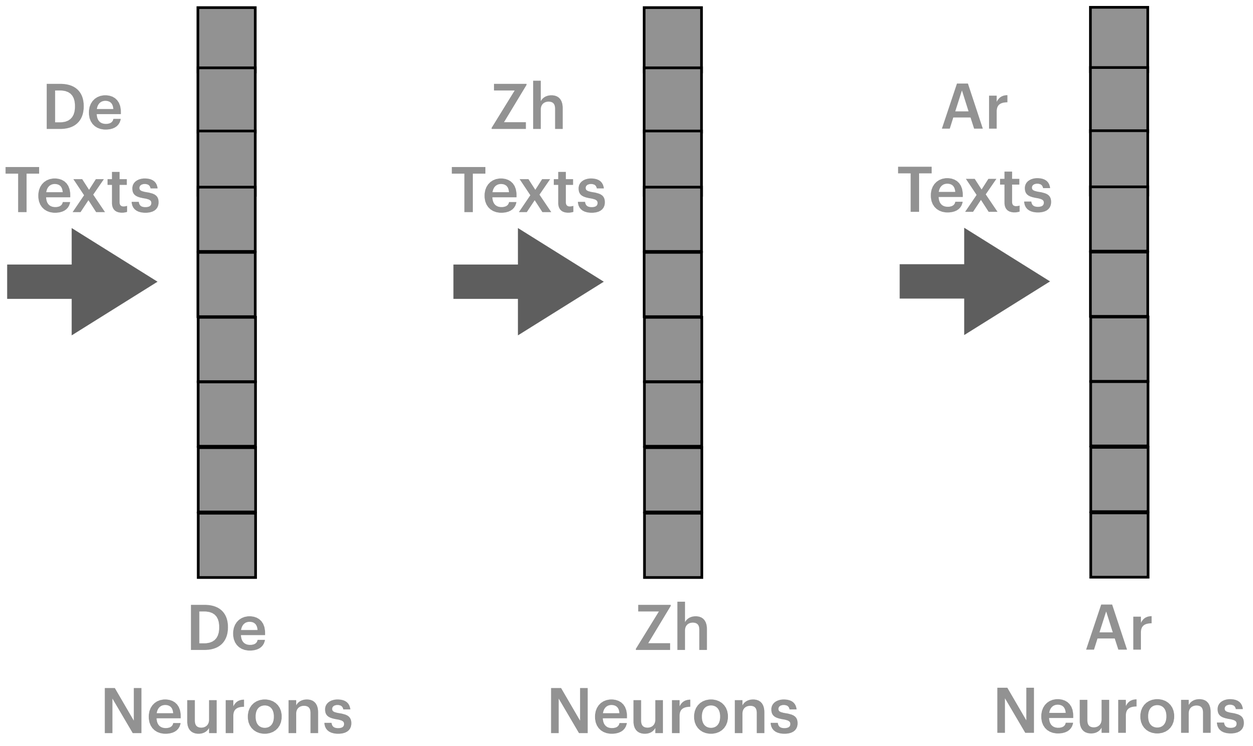
Neuron Structural Analysis - Method



Neurons can only be:
Activated: >0
Non-activated: =0

Neuron Specialization

Neuron Structural Analysis - Method

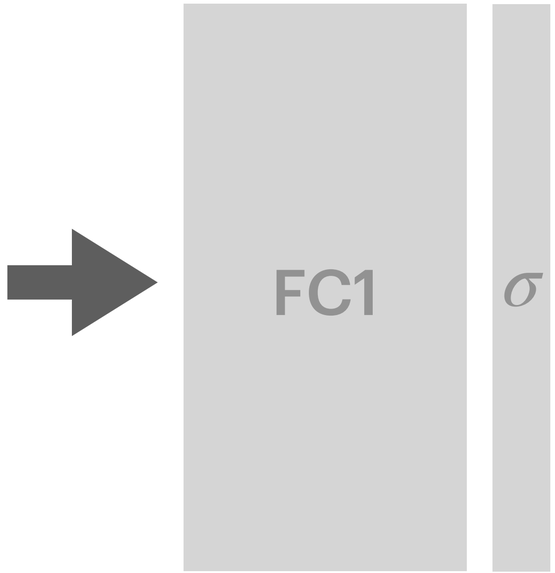


Intuition:
Are neurons task-specific?

Neuron Specialization

Neuron Structural Analysis - Method

Activation Recording



Neuron Specialization

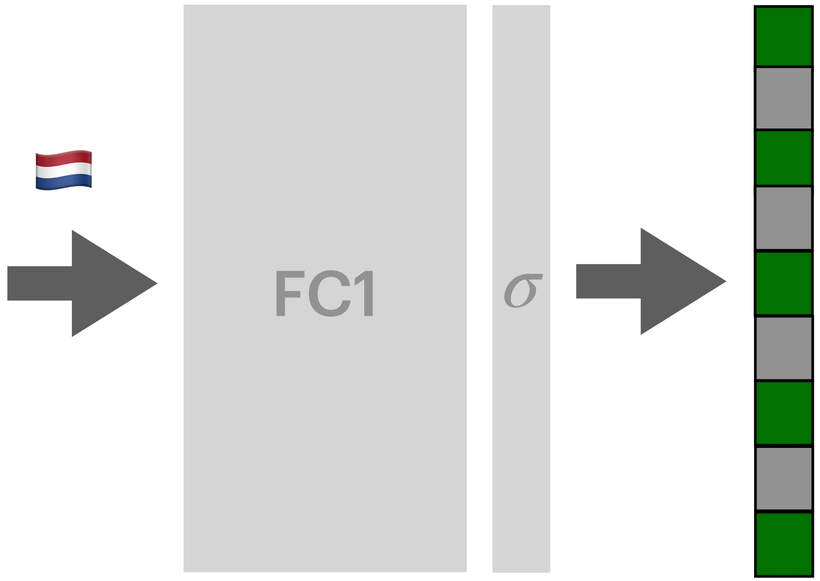
Neuron Structural Analysis - Method

Activation Recording

Neuron Specialization

Neuron Structural Analysis - Method

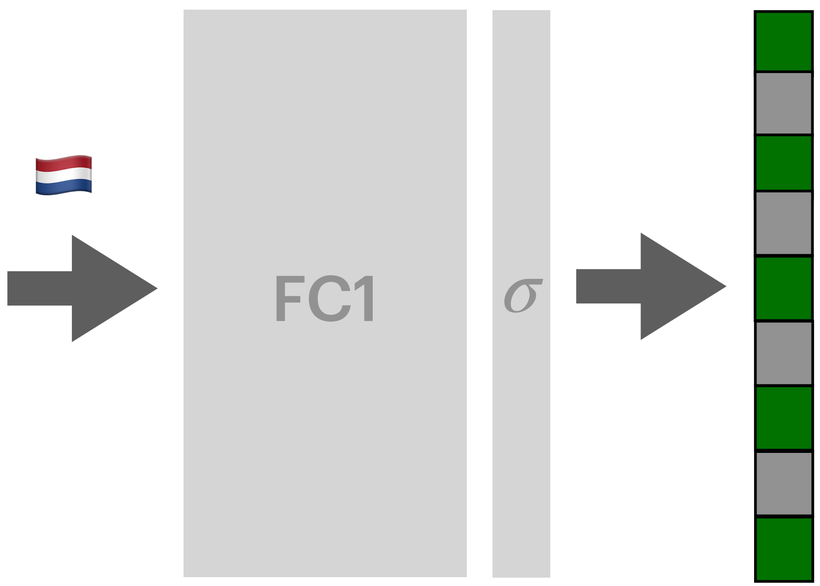
Activation Recording



Neuron Specialization

Neuron Structural Analysis - Method

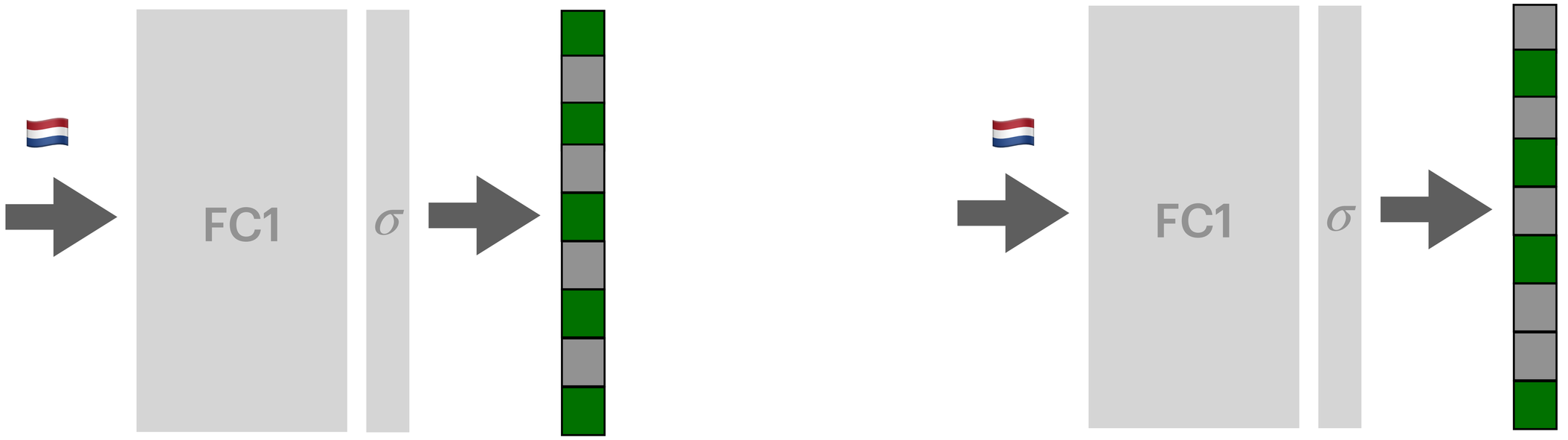
Activation Recording



Neuron Specialization

Neuron Structural Analysis - Method

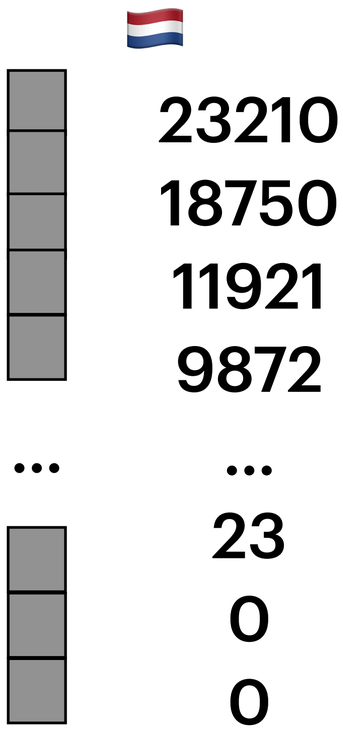
Activation Recording



Neuron Specialization

Neuron Structural Analysis - Method

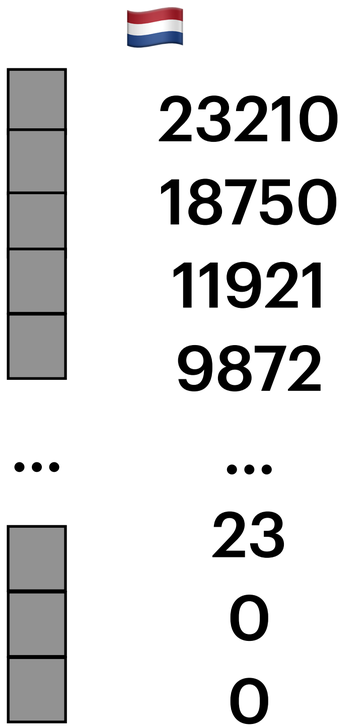
Neuron Activation Frequency



Neuron Specialization

Neuron Structural Analysis - Method

Neuron Activation Frequency



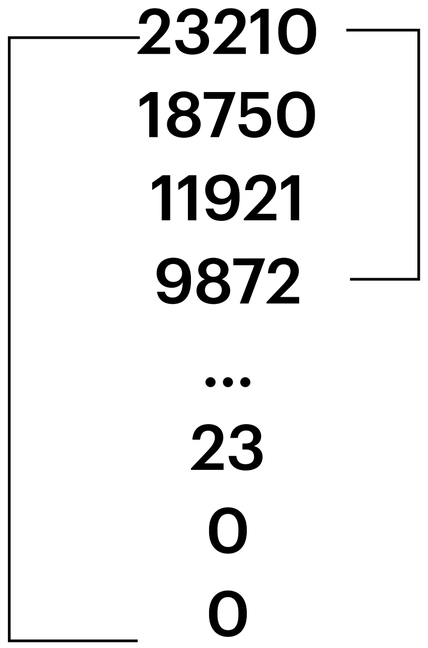
How should we select Specialized Neurons for each language pair?

Neuron Specialization

Neuron Structural Analysis - Method

Specialized Neuron Selection

$$\sum_{n=1}^{dim_{ff}} a_n^{s \to t}$$



$$\sum_{n \in S_k^{s \to t}} a_n^{s \to t} \geq k * \sum_{n=1}^{dim_{ff}} a_n^{s \to t},$$

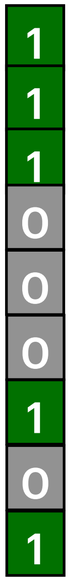
We **dynamically** select neurons based on a cumulative activation threshold $k \in [0,1]$.

Neuron Specialization

Neuron Structural Analysis - Analysis

en->de

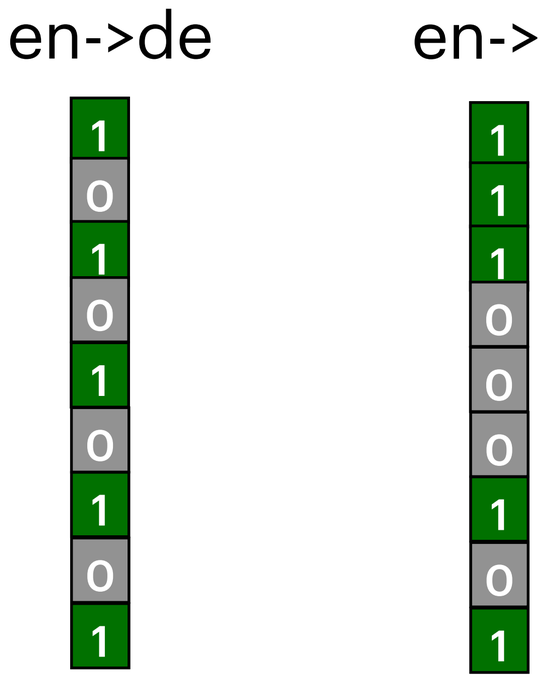
en->nl



$$m_{s \to t}^{l=1} \in \mathbb{R}^{dim_{ff}}$$

Neuron Specialization

Neuron Structural Analysis - Analysis



$$m_{s\to t}^{l=1} \in \mathbb{R}^{dim_{ff}}$$

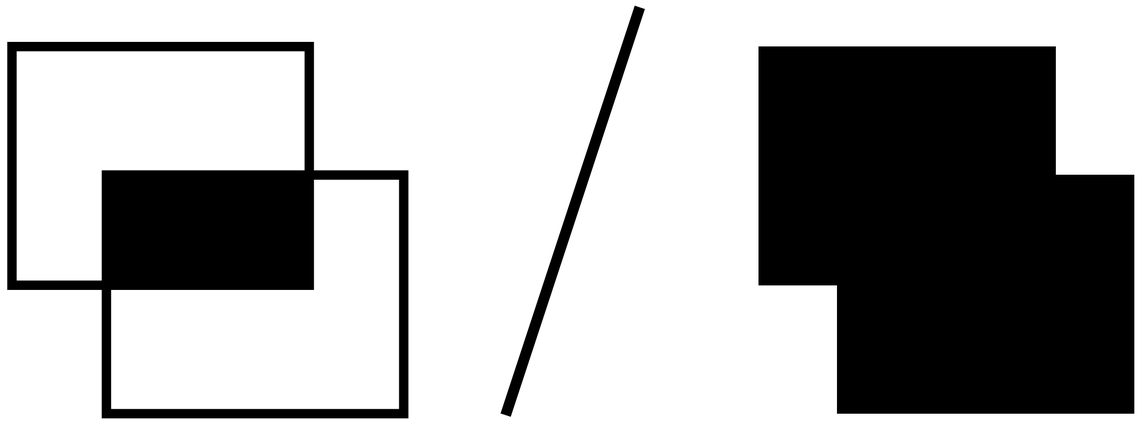
Whether similar languages share Similar Specialized Neurons?

Neuron Specialization

Neuron Structural Analysis - Analysis

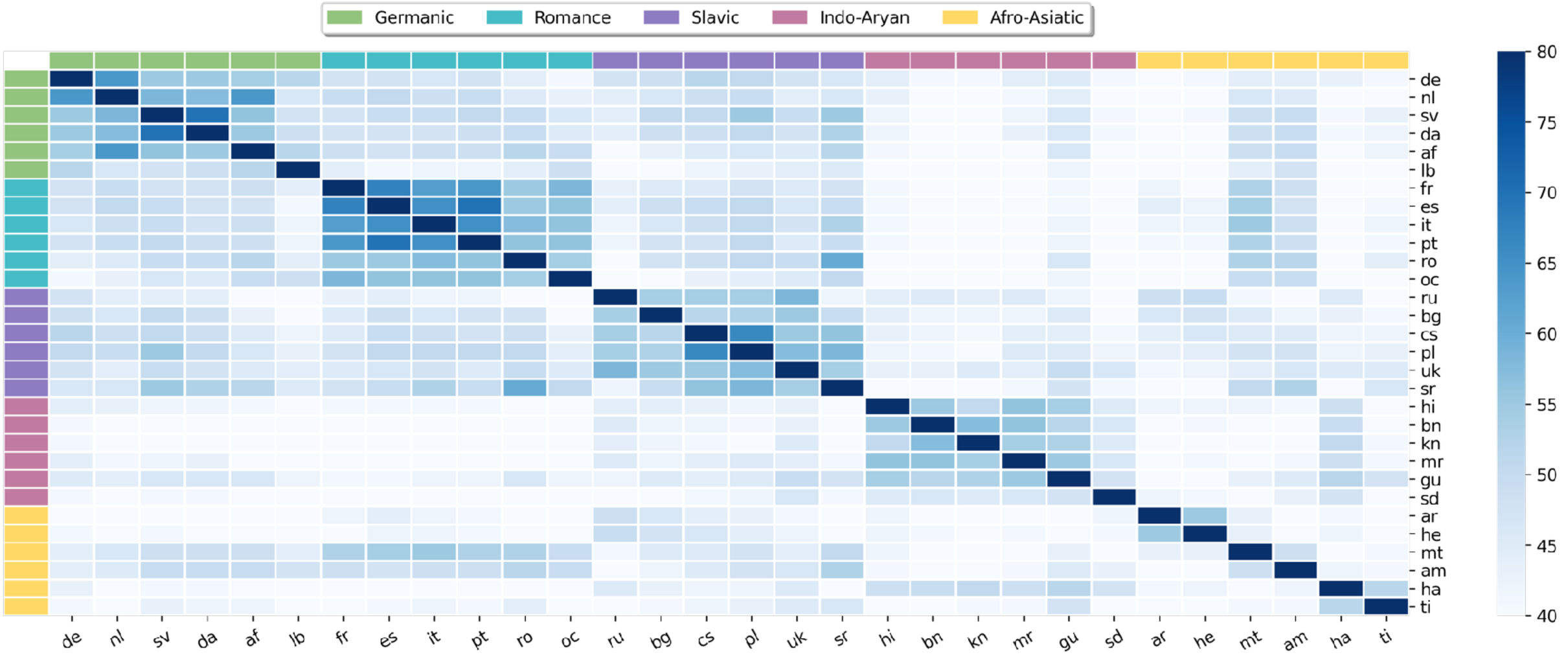
We use Intersection Over Union (IoU) to measure the similarity between two specialized neuron sets.

$$\text{IoU} = \frac{\text{Overlap}}{\text{Union}} =$$



Neuron Specialization

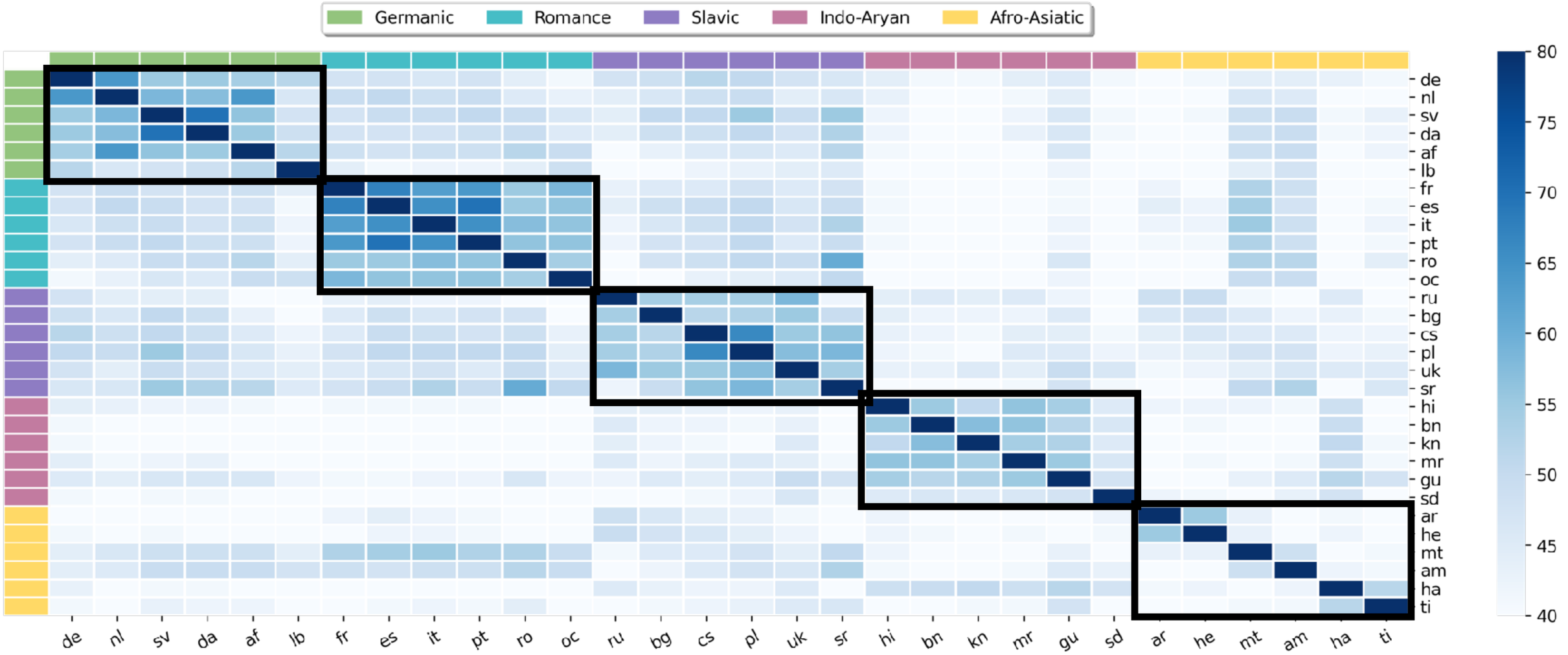
Neuron Structural Analysis - Observations



Specialized Neuron that extracted from En->X in first Decoder layer.

Neuron Specialization

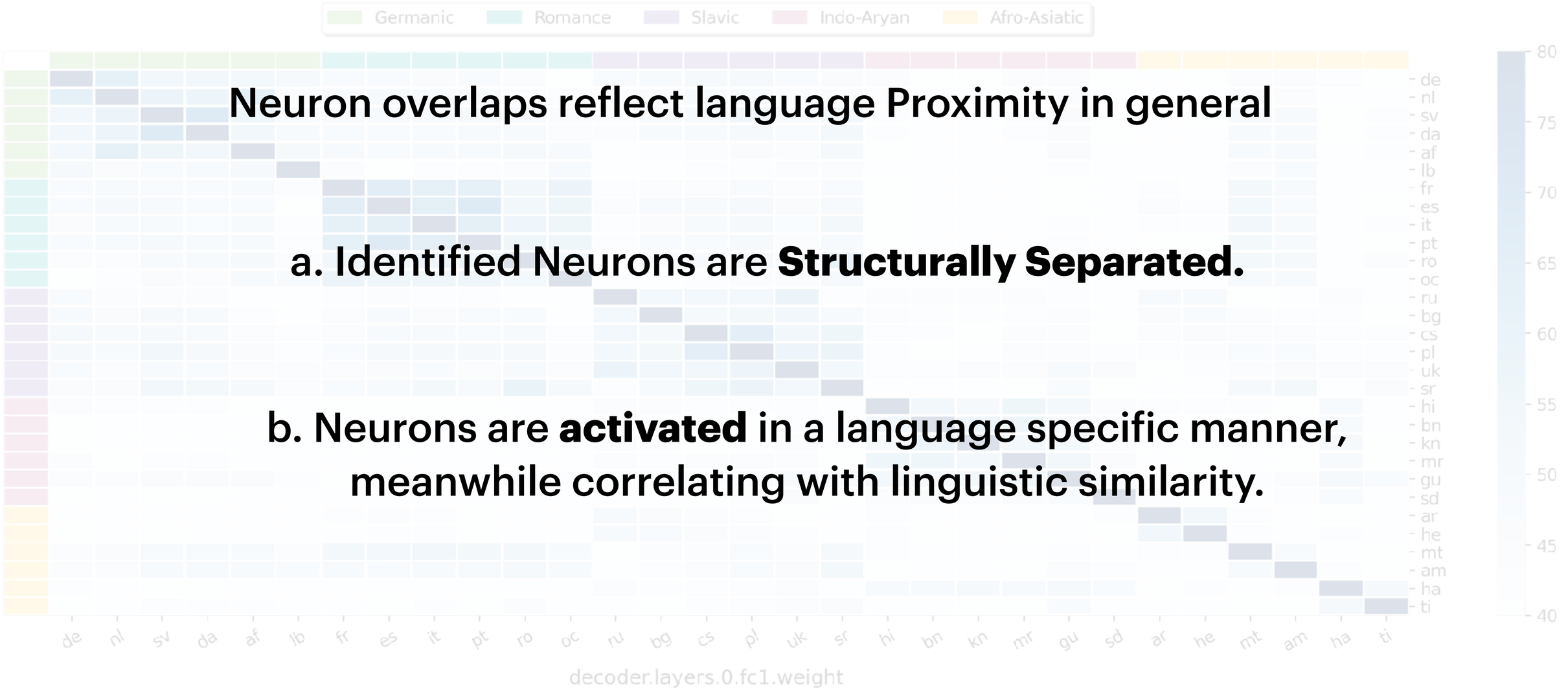
Neuron Structural Analysis - Observations



Specialized Neuron that extracted from En->X in first Decoder layer.

Neuron Specialization

Neuron Structural Analysis - Findings



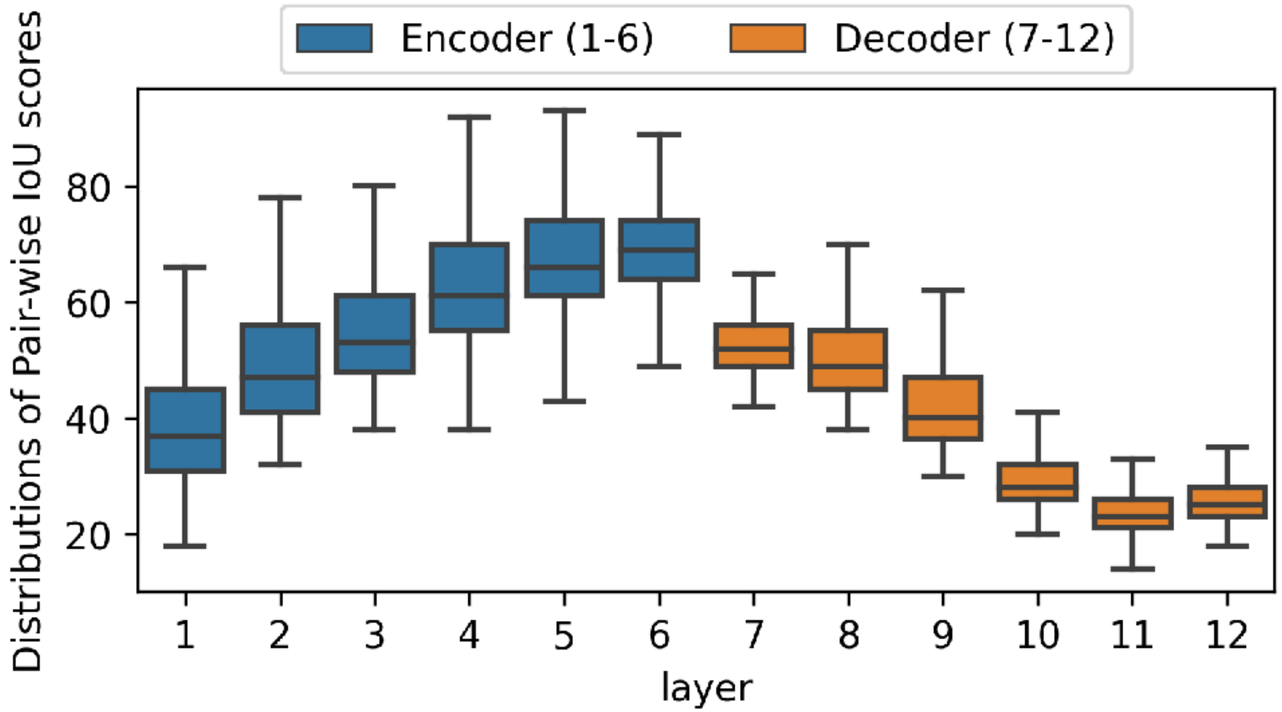
Neuron overlaps reflect language Proximity in general

a. Identified Neurons are **Structurally Separated.**

b. Neurons are **activated** in a language specific manner, meanwhile correlating with linguistic similarity.

Neuron Specialization

Neuron Structural Analysis - Observations



Neuron overlap
progresses across layers

specific -> agnostic in Encoder

agnostic -> specific in Decoder

Similar to prior MNMT
representation study[1]

1) Kudugunta, Sneha Reddy, et al. "Investigating multilingual NMT representations at scale."

We show intrinsic task Modularity exists in Neurons

We show intrinsic task Modularity exists in Neurons

How can we further promote such inherent structural signals?

Neuron Specialization

Method

We further **intensify** such **intrinsic task modularity** with **sparse network**

Neuron Specialization

Method

We further **intensify** such **intrinsic task modularity** with **sparse network**

$$M_{en->de} \in \{0,1\}$$

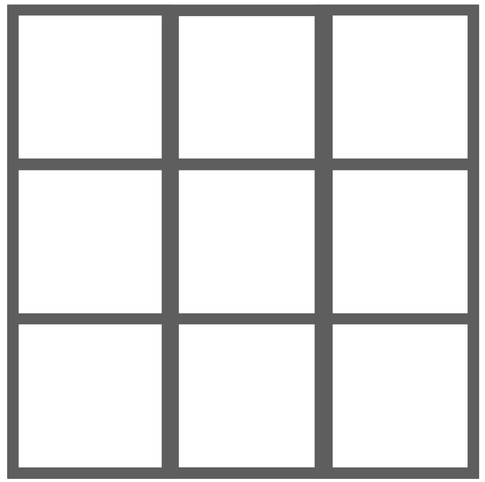
Neuron Specialization

Method

We further **intensify** such **intrinsic task modularity** with **sparse network**

$$M_{en->de} \in \{0,1\}$$

1	0	1

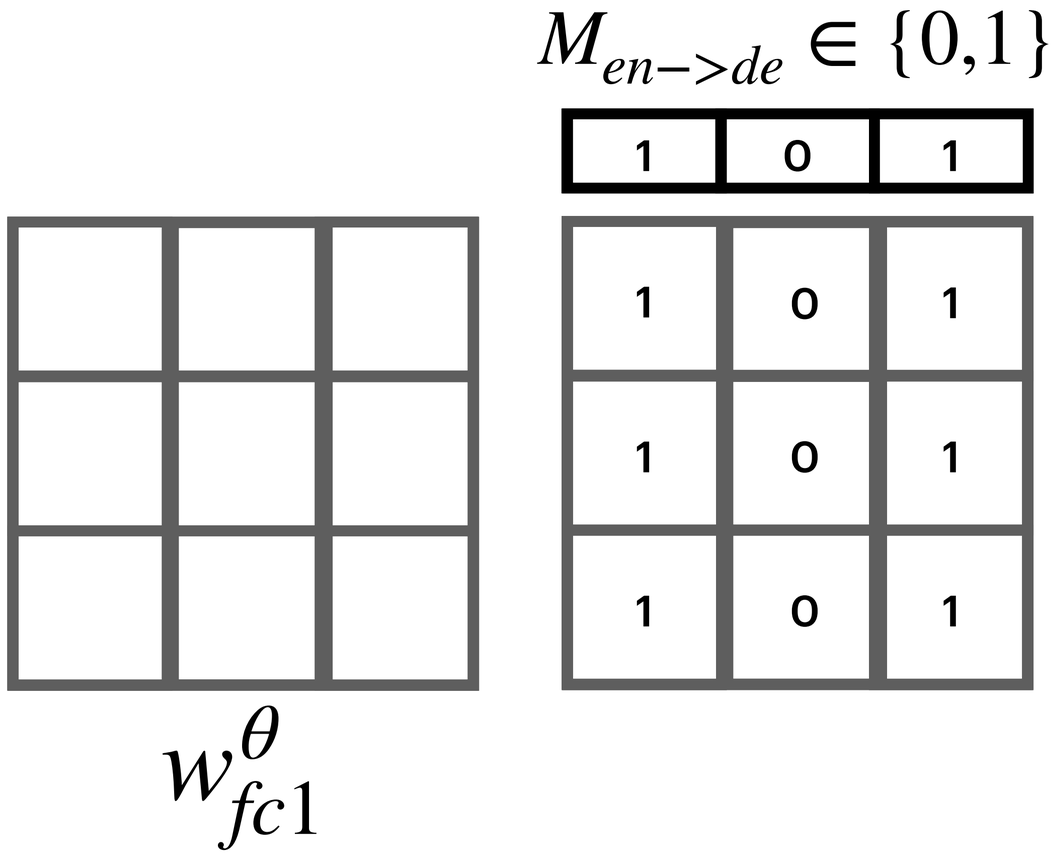


$$w_{fc1}^{\theta}$$

Neuron Specialization

Method

We further **intensify** such **intrinsic task modularity** with **sparse network**



Neuron Specialization

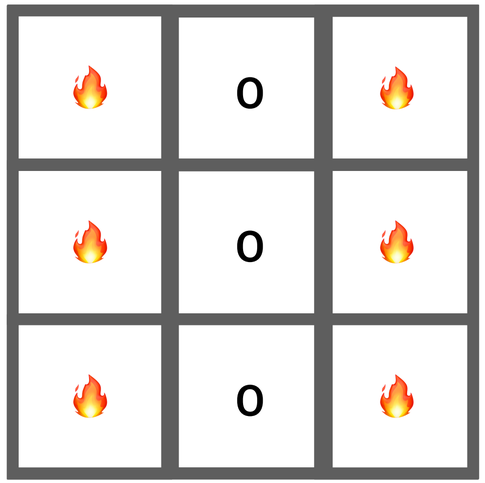
Method

We further **intensify** such **intrinsic task modularity** with **sparse network** via **continue training**

$$\mathrm{FFN}(H) = \mathrm{ReLU}(HW_1)\,W_2.$$

$$FFN(H) = ReLU(H(m_k^t \odot W_1))W_2.$$

$$W_{fc1}^{\theta'}$$



For en->de data

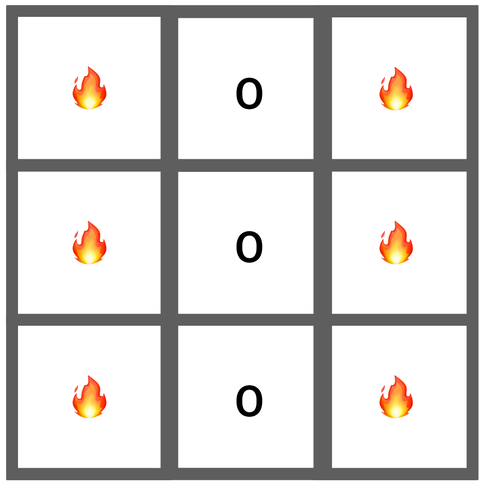
$$M_{en->de} \in \{0,1\}$$

1	0	1

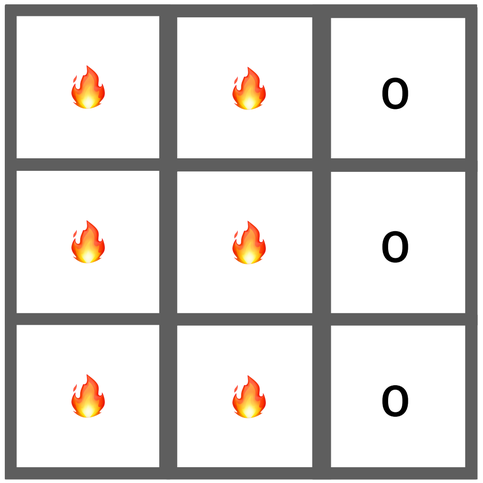
Neuron Specialization

Method

No extra parameters are introduced!



en->de



en->ar

Neuron Specialization

Results - EC30 (large-scale)

Reduce Interference - **Consistent performance gains**

	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Adapter-128	+81%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	**+0.9**	**+1.2**	**+1.5**	**+1.1**	**+1.3**

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Results - EC30 (large-scale)

While maintaining performance on **low-resource** languages

	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Adapter-128	+81%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	**+0.9**	**+1.2**	**+1.5**	**+1.1**	**+1.3**

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Results - EC30 (large-scale)

More Balanced improvement compare to LaSS

	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg	En→X	X→En	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Adapter-128	+81%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	**+0.9**	**+1.2**	**+1.5**	**+1.1**	**+1.3**

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Results - Iwslt14 (small-scale)

Language Size	$\Delta\theta$	Fa 89k	Pl 128k	Ar 139k	He 144k	Nl 153k	De 160k	It 167k	Es 169k	ΔAvg
One-to-Many										
mT-small	-	14.54	9.90	11.99	13.13	17.03	20.6	17.31	18.30	-
Adapter-128	+67%	+0.06	-0.09	+0.38	**+1.39**	**+0.23**	+0.57	+0.08	**+0.37**	**+0.37**
LaSS	0%	-2.64	+0.03	+0.58	+0.66	-0.15	**+0.66**	-0.18	-0.36	-0.18
Ours	0%	**+0.71**	**+0.10**	**+0.94**	+0.64	+0.07	+0.03	**+0.22**	-0.26	+0.31
Many-to-One										
mT-small	-	19.09	19.37	25.65	30.86	30.6	28.07	28.97	33.95	-
Adapter-128	+67%	+0.91	+0.58	+0.85	+1.00	+0.81	+1.03	+0.95	+0.25	+0.80
LaSS	0%	+1.19	+0.61	+0.93	+1.37	+1.06	+1.6	+1.61	+0.83	+1.15
Ours	0%	**+1.59**	**+1.18**	**+1.7**	**+2.02**	**+1.85**	**+2.06**	**+1.8**	**+1.36**	**+1.70**

Bleu Improvements over the baseline system (mT-small)

Neuron Specialization

Results - Iwslt14 (small-scale)

Language Size	$\Delta\theta$	Fa 89k	Pl 128k	Ar 139k	He 144k	Nl 153k	De 160k	It 167k	Es 169k	ΔAvg
One-to-Many										
mT-small	-	14.54	9.90	11.99	13.13	17.03	20.6	17.31	18.30	-
Adapter-128	+67%	+0.06	-0.09	+0.38	**+1.39**	**+0.23**	+0.57	+0.08	**+0.37**	**+0.37**
LaSS	0%	-2.64	+0.03	+0.58	+0.66	-0.15	**+0.66**	-0.18	-0.36	-0.18
Ours	0%	**+0.71**	**+0.10**	**+0.94**	+0.64	+0.07	+0.03	**+0.22**	-0.26	+0.31
Many-to-One										
mT-small	-	19.09	19.37	25.65	30.86	30.6	28.07	28.97	33.95	-
Adapter-128	+67%	+0.91	+0.58	+0.85	+1.00	+0.81	+1.03	+0.95	+0.25	+0.80
LaSS	0%	+1.19	+0.61	+0.93	+1.37	+1.06	+1.6	+1.61	+0.83	+1.15
Ours	0%	**+1.59**	**+1.18**	**+1.7**	**+2.02**	**+1.85**	**+2.06**	**+1.8**	**+1.36**	**+1.70**

Bleu Improvements over the baseline system (mT-small)

Neuron Specialization

Results - Iwslt14 (small-scale)

Language Size	$\Delta\theta$	Fa 89k	Pl 128k	Ar 139k	He 144k	Nl 153k	De 160k	It 167k	Es 169k	ΔAvg
One-to-Many										
mT-small	-	14.54	9.90	11.99	13.13	17.03	20.6	17.31	18.30	-
Adapter-128	+67%	+0.06	-0.09	+0.38	**+1.39**	**+0.23**	+0.57	+0.08	**+0.37**	**+0.37**
LaSS	0%	-2.64	+0.03	+0.58	+0.66	-0.15	**+0.66**	-0.18	-0.36	-0.18
Ours	0%	**+0.71**	**+0.10**	**+0.94**	+0.64	+0.07	+0.03	**+0.22**	-0.26	+0.31
Many-to-One										
mT-small	-	19.09	19.37	25.65	30.86	30.6	28.07	28.97	33.95	-
Adapter-128	+67%	+0.91	+0.58	+0.85	+1.00	+0.81	+1.03	+0.95	+0.25	+0.80
LaSS	0%	+1.19	+0.61	+0.93	+1.37	+1.06	+1.6	+1.61	+0.83	+1.15
Ours	0%	**+1.59**	**+1.18**	**+1.7**	**+2.02**	**+1.85**	**+2.06**	**+1.8**	**+1.36**	**+1.70**

Bleu Improvements over the baseline system (mT-small)

Neuron Specialization

Results - Efficiency Comparison

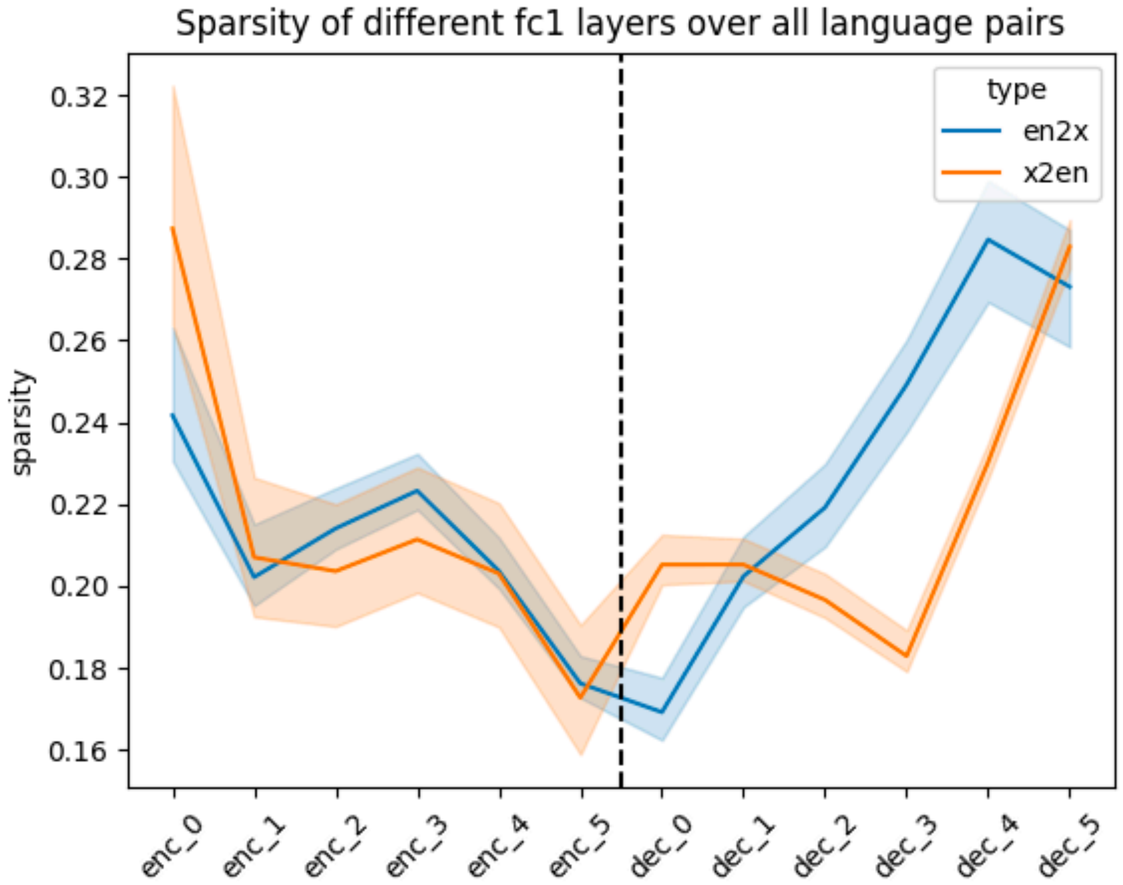
Model	$\triangle\theta$	$\triangle T_{subnet}$	$\triangle$ Memory
Adapter$_{LP}$	+87%	n/a	1.42 GB
LaSS	0%	+33 hours	9.84 GB
Ours	0%	+5 minutes	3e-3 GB

Results reported based on EC30 with 4 A6000 GPUs

Our approach is highly **efficient**, facilitating the **adaptation to massively multilingual models.**

Neuron Specialization

Analysis and Discussion



How does Sparsity changes with our method?

Sparse -> Dense in Encoder

Dense -> Sparse in Decoder

Neuron Specialization

Analysis and Discussion - Neuron specialization on Enc or Dec?

Methods	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Adapter$_{Fam}$	+70%	+0.7	+0.3	+0.5	+0.7	+0.3	+0.5	+1.1	+0.5	+0.8	+0.8	+0.4	+0.6
Adapter$_{LP}$	+87%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Random	0%	+0.9	-0.5	+0.2	+0.5	-0.7	-0.2	-0.3	-1.5	-0.9	+0.5	-0.9	-0.2
Ours-Enc	0%	+1.2	+1.1	+1.1	+1.0	+1.0	+1.0	+0.7	+0.8	+0.8	+1.0	+1.0	+1.0
Ours-Dec	0%	+1.2	+1.1	+1.1	+0.9	+1.1	+1.0	+0.7	+1.1	+0.9	+0.9	+1.1	+1.0
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	**+0.9**	**+1.2**	**+1.5**	**+1.1**	**+1.3**

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Analysis and Discussion — How much we can alleviate interference?

Lang	De	Es	Cs	Hi	Ar	Lb	Ro	Sr	Gu	Am	High	Low
Size	5m	5m	5m	5m	5m	100k	100k	100k	100k	100k	Avg	Avg
One-to-Many												
Bilingual	36.3	24.6	28.7	43.9	23.7	5.5	16.2	17.8	12.8	4.1	31.8	11.3
mT-big	-4.7	-1.5	-3.6	-4.4	-4.7	+9.0	+8.9	+6.2	+13.9	+3.1	-3.7	+8.2
Many-to-One												
Bilingual	39.1	24.5	32.6	35.5	30.8	8.7	19.5	21.3	7.0	8.7	32.7	13.0
mT-big	-1.5	+0.9	+0.2	-1.8	-2.3	+13.7	+11.9	+10.3	+18.2	+12.5	-1.1	+13.3

SacreBleu Improvements over bilingual systems

Evidence of Interference: worse performance on high-resource languages.

Neuron Specialization

Analysis and Discussion — How much we can alleviate interference?

Lang Size	De 5m	Es 5m	Cs 5m	Hi 5m	Ar 5m	Lb 100k	Ro 100k	Sr 100k	Gu 100k	Am 100k	High Avg	Low Avg
One-to-Many												
Bilingual	36.3	24.6	28.7	43.9	23.7	5.5	16.2	17.8	12.8	4.1	31.8	11.3
mT-big	-4.7	-1.5	-3.6	-4.4	-4.7	+9.0	+8.9	+6.2	+13.9	+3.1	-3.7	+8.2
Ours	-2.0	-0.2	-1.7	-2.4	-3.0	+10.8	+10.0	+8.2	+16.4	+3.7	-1.9	+9.8
Many-to-One												
Bilingual	39.1	24.5	32.6	35.5	30.8	8.7	19.5	21.3	7.0	8.7	32.7	13.0
mT-big	-1.5	+0.9	+0.2	-1.8	-2.3	+13.7	+11.9	+10.3	+18.2	+12.5	-1.1	+13.3
Ours	-0.3	+1.7	+1.8	-0.2	-0.3	+15.3	+12.4	+11.3	+19.6	+14.1	+0.3	+14.5

SacreBleu Improvements over bilingual systems

Our method reduces interference while further encouraging knowledge transfer!

Conclusions

Neuron Analysis

Show Intrinsic modularity in multi-task models without modification.

Proposed Method

Presents Consistent Performance

Gains on large-scale experiments.

Neuron Specialization

Efficiency

Introduce 0 extra

Trainable Parameters.

Understanding

fundamental properties

in FFN Modules & Multi-task.

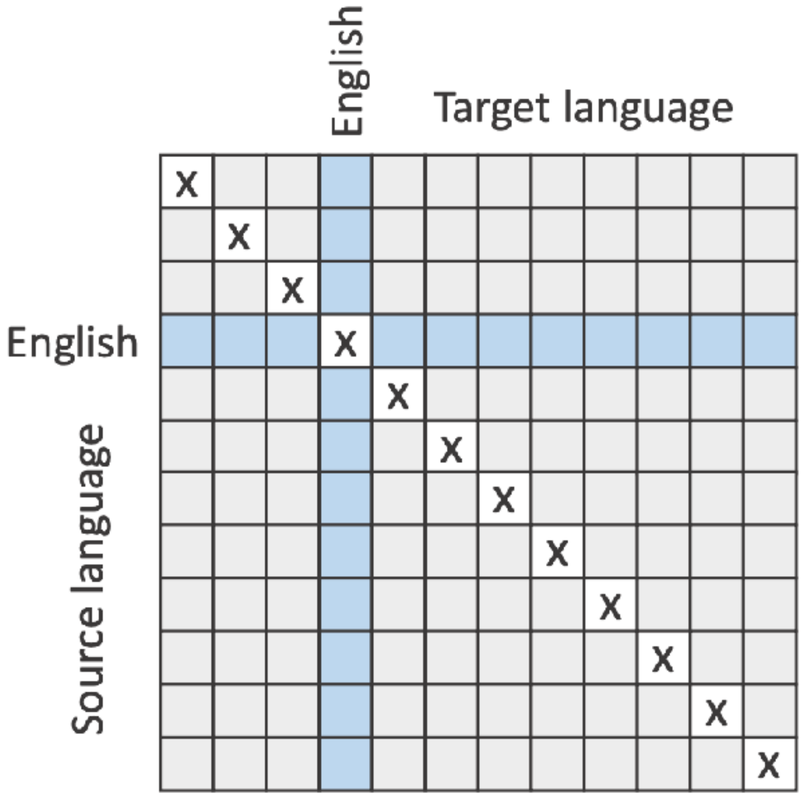
Towards a Better Understanding of Variations in Zero-Shot Neural Machine Translation Performance

— Shaomu Tan, Christof Monz

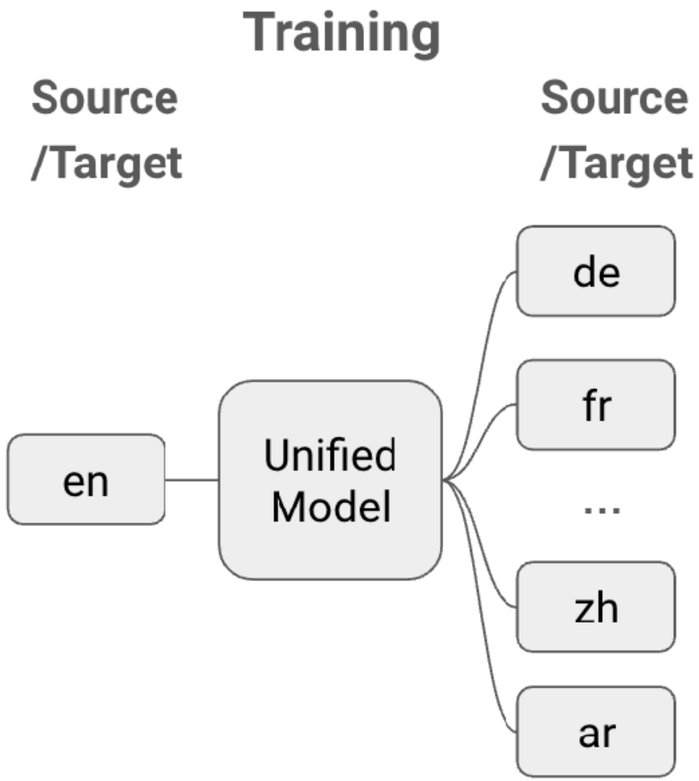
SHAOMU TAN |  UNIVERSITY OF AMSTERDAM
Language Technology Lab

EMNLP 2023

English-Centric Systems



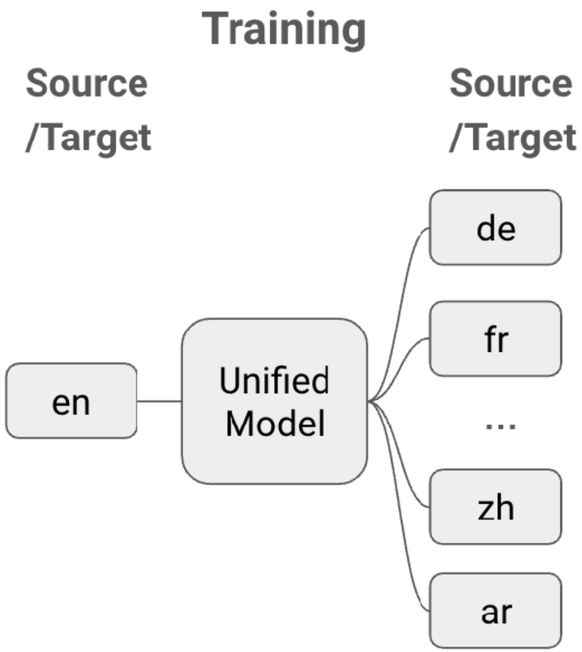
In real-world, most translation data is English-centric

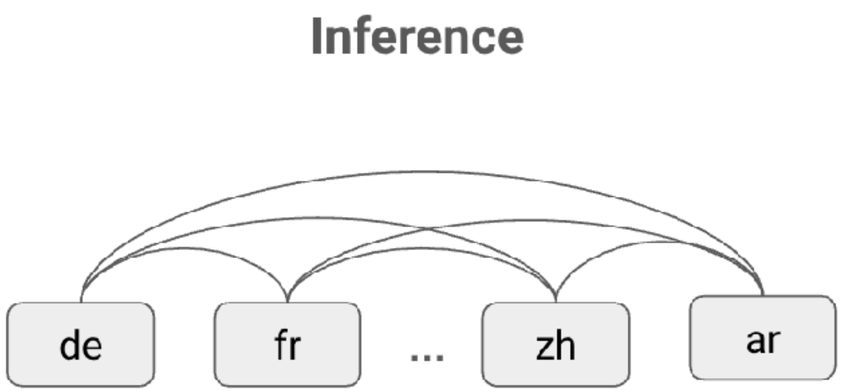


English-centric Multilingual MT approach

Fan, Angela, et al. "Beyond english-centric multilingual machine translation."

Zero-shot Translation





- Efficiency
  - Only trained on N language pairs.
  - Inference on N*N - N directions.

Zero-Shot Translation:
translation between a language pair never seen in training

Zero-shot Translation

	Sacrebleu				Chrf++				SpBleu				Comet			
	En→X	X→En	En↔X	ZS	En→X	X→En	En↔X	ZS	En→X	X→En	En↔X	ZS	En→X	X→En	En↔X	ZS
Averaged Performance																
mT-big	23.1	27.5	25.3	4.9	47.1	52.6	49.9	20.5	29.9	30.6	30.2	7.3	78.4	78.3	78.3	54.7
mBart50	22.7	**29.5**	26.1	6.6	46.8	**53.9**	50.3	23.5	29.6	**32.6**	31.0	9.6	**80.2**	**80.1**	**80.1**	58.8
mT-large	**23.6**	28.7	**26.1**	**7.0**	**47.6**	53.3	**50.4**	**25.3**	**30.5**	31.8	**31.1**	**10.1**	79.2	79.0	79.1	**59.5**

Zero-shot translation qualities are far behind being satisfied when compared to supervised directions (En<->X)

Shaomu Tan, Christof Monz. "Towards a Better Understanding of Variations in Zero-Shot Neural Machine Translation Performance"

Off-target Issues lead poor ZS?

Off-target: Translating into **wrong languages**.
-> Often exists in English-centric systems

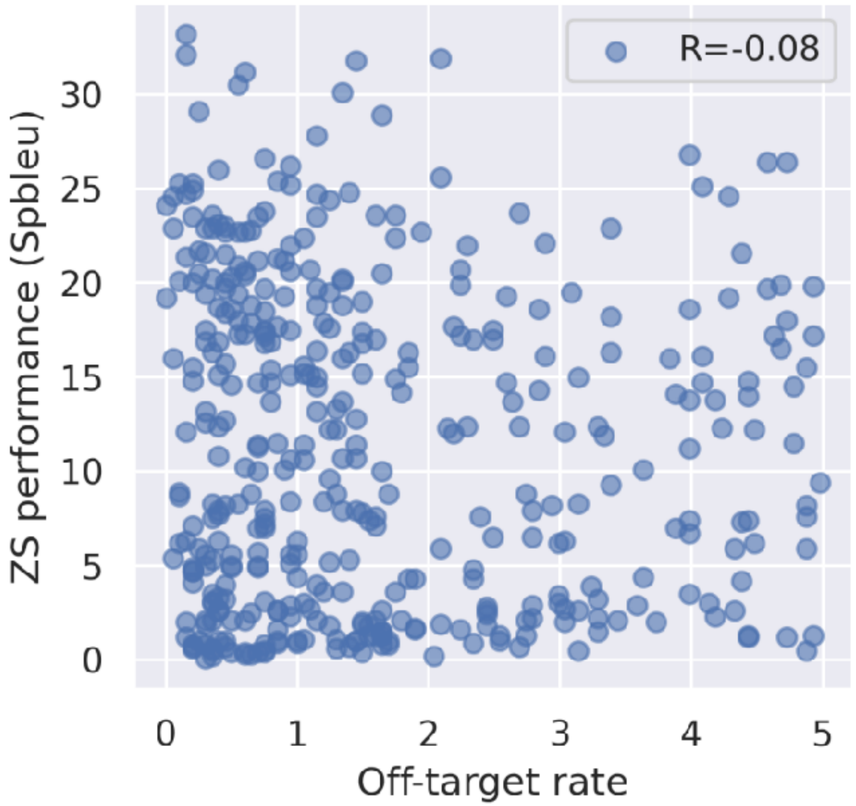
Training: fr->en, en->de;
ZS inference: fr->de

Actual generation: fr->en

Source	Les États membres ont été consultés et ont approuvé cette proposition.
Reference	Die Mitgliedstaaten wurden konsultiert und sprachen sich für diesen Vorschlag aus.
Zero-Shot	Les Member States have been consultedés and have approved this proposal.

Zhang, Biao, et al. "Improving massively multilingual neural machine translation and zero-shot translation."

Off-target Issue is a Symptom



The off-target issue is more likely to be a **symptom** rather than the root cause[1].

Translating into the correct language cannot guarantee decent performance.

1) Shaomu Tan, Christof Monz. "Towards a Better Understanding of Variations in Zero-Shot Neural Machine Translation Performance"

Explaining variations in ZS performance

ID	Features	R-square	MAE	RMSE
	mT-big			
1	En_performance	45.63%	4.03	4.88
2	1 + Vocab-Sim	63.42%	3.70	4.77
3	2 + Linguistic-features	81.17%	3.42	4.37
	mT-large			
4	En_performance	61.34%	4.70	5.59
5	4 + Vocab-Sim	79.75%	3.76	4.84
6	5 + Linguistic-features	81.75%	3.67	4.75

	Tgt resource			
	eLow	Low	Med	High
If Src and Tgt in the same Language Family				
No	2.12	4.82	9.77	9.43
Yes	**3.14***	**7.69***	**13.16***	**12.88***
If Src and Tgt use the same Writing System				
No	1.58	3.97	9.31	8.67
Yes	**3.21***	**8.13***	**11.71***	**12.68***

* represents $p <= 0.05$

What **factors** could lead to **variations** in ZS performance?

- En-centric performance
- Vocabulary Similarity
- Linguistic properties, e.g.: Language families and writing systems

How Far Can 100 Samples Go?

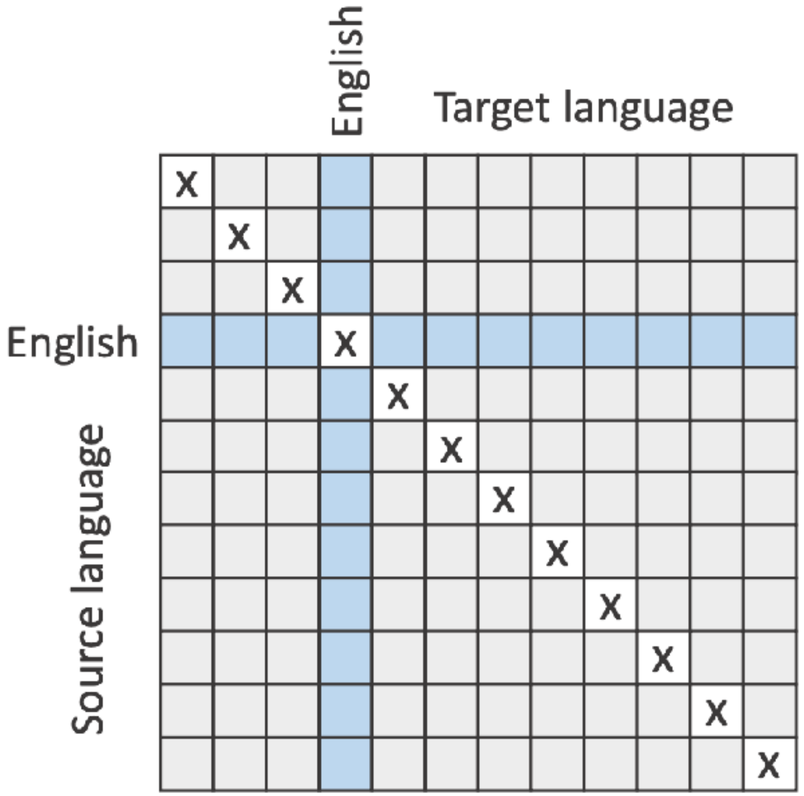
Unlocking Overall Zero-Shot Multilingual Translation via Tiny Multi-Parallel Data

— Di Wu, Shaomu Tan, Yan Meng, David Stap, Christof Monz

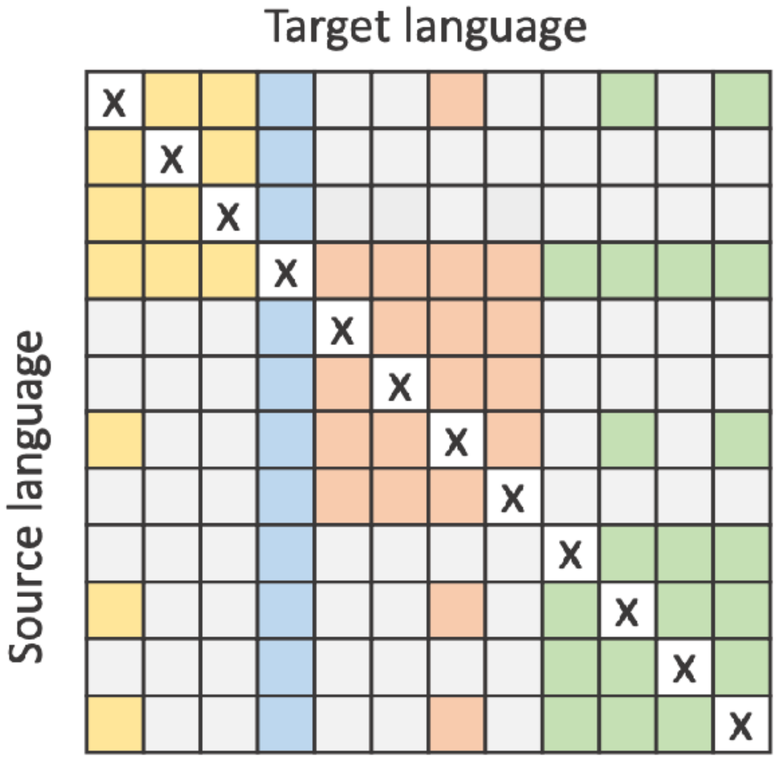
SHAOMU TAN |  UNIVERSITY OF AMSTERDAM
Language Technology Lab

ACL 2024

Bridge Translation



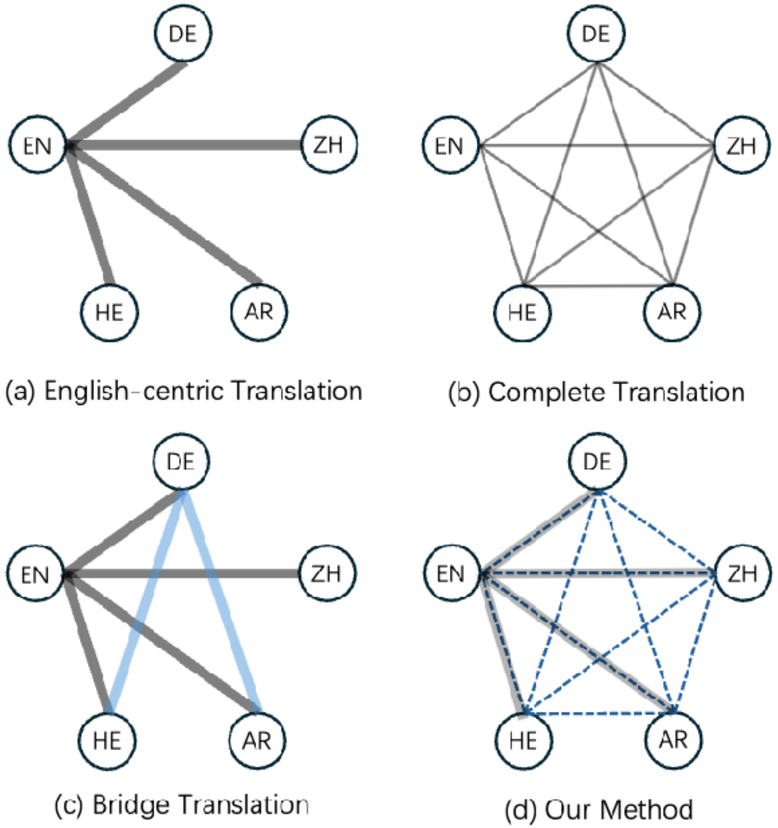
In real-world, most translation data is English-centric



Mining partial non-English data as the bridge language

Fan, Angela, et al. "Beyond english-centric multilingual machine translation."

Beyond English-Centric Multilingual MT



Complete Translation aims to cover all directions but suffers from the small data scale

Bridge Translation: Mining partial non-English data as the bridge language

Our method: fine-tuning an English-centric model with tiny **multi-parallel data**

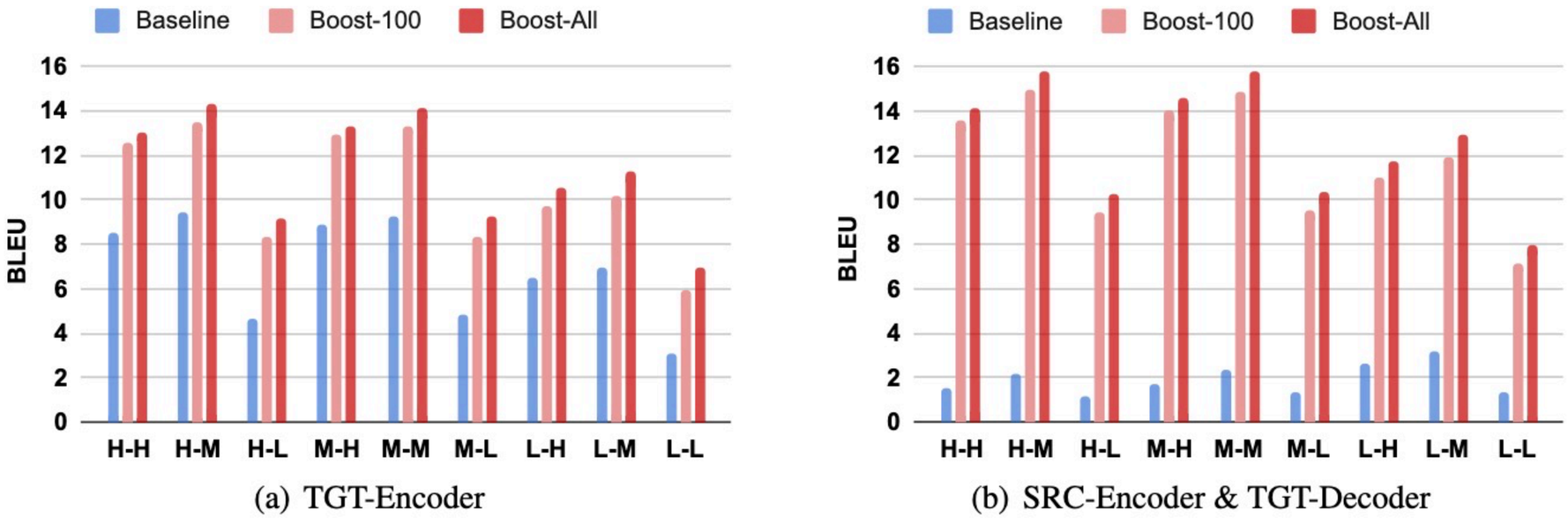
Multi-parallel Translation Data

Ntrex-128: Translating 1997 English sentence into 128 Languages by professional human translators

Semantic Equivalent Dataset

Federmann, Christian, Tom Kocmi, and Ying Xin. "NTREX-128–news test references for MT evaluation of 128 languages."

Boosting Zero-shot Translations



Boosting EC30 Models using Ntrex Multi-parallel data on 870 zero-shot directions.

Catastrophic Forgetting?

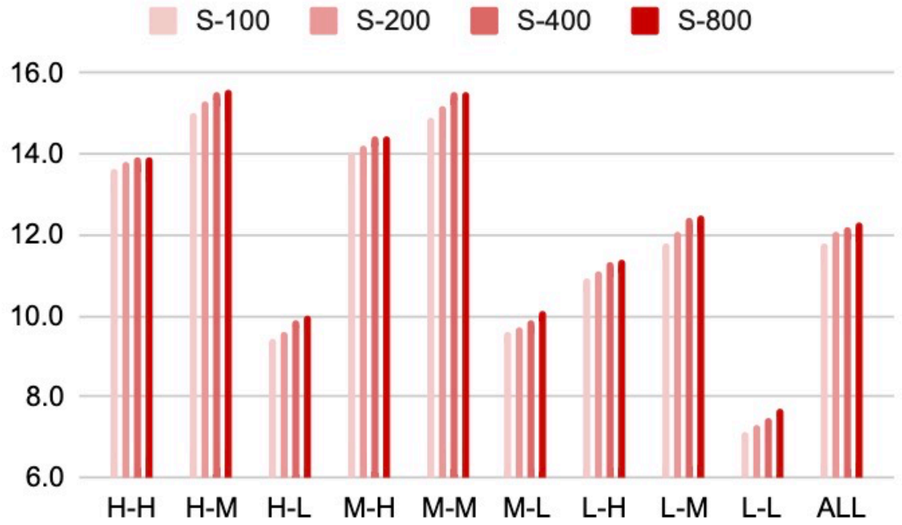
Model	High		Medium		Low		ALL		
	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	AVG
Baseline	28.7	31.1	31.0	31.1	20.1	25.4	26.6	29.2	27.9
Boost-100	27.9	30.6	30.8	31.6	20.0	25.6	26.2	29.3	27.7
Boost-All	27.2	30.1	30.1	31.3	20.1	25.7	25.8	29.0	27.4
Δ-100	-0.8	-0.5	-0.2	+0.5	-0.1	+0.2	-0.4	+0.1	-0.2
Δ-All	-1.5	-1.0	-0.9	+0.2	0.0	+0.3	-0.8	-0.2	-0.5
Baseline	28.3	31.6	30.7	31.5	19.4	25.9	26.1	29.7	27.9
Boost-100	27.6	30.1	30.3	31.3	19.5	25.0	25.8	28.8	27.3
Boost-All	27.4	29.8	30.2	31.3	19.9	25.8	25.8	29.0	27.4
Δ-100	-0.7	-1.5	-0.4	-0.2	+0.1	-0.9	-0.3	-0.9	-0.6
Δ-All	-0.9	-1.8	-0.5	-0.2	+0.5	-0.1	-0.3	-0.7	-0.5

Boosting zero-shot directions almost does not affect supervised (en-centric) directions.

Catastrophic Forgetting does not play an important role in our observations.

More Data or More directions?

(a) Increase the Number of **D**irections

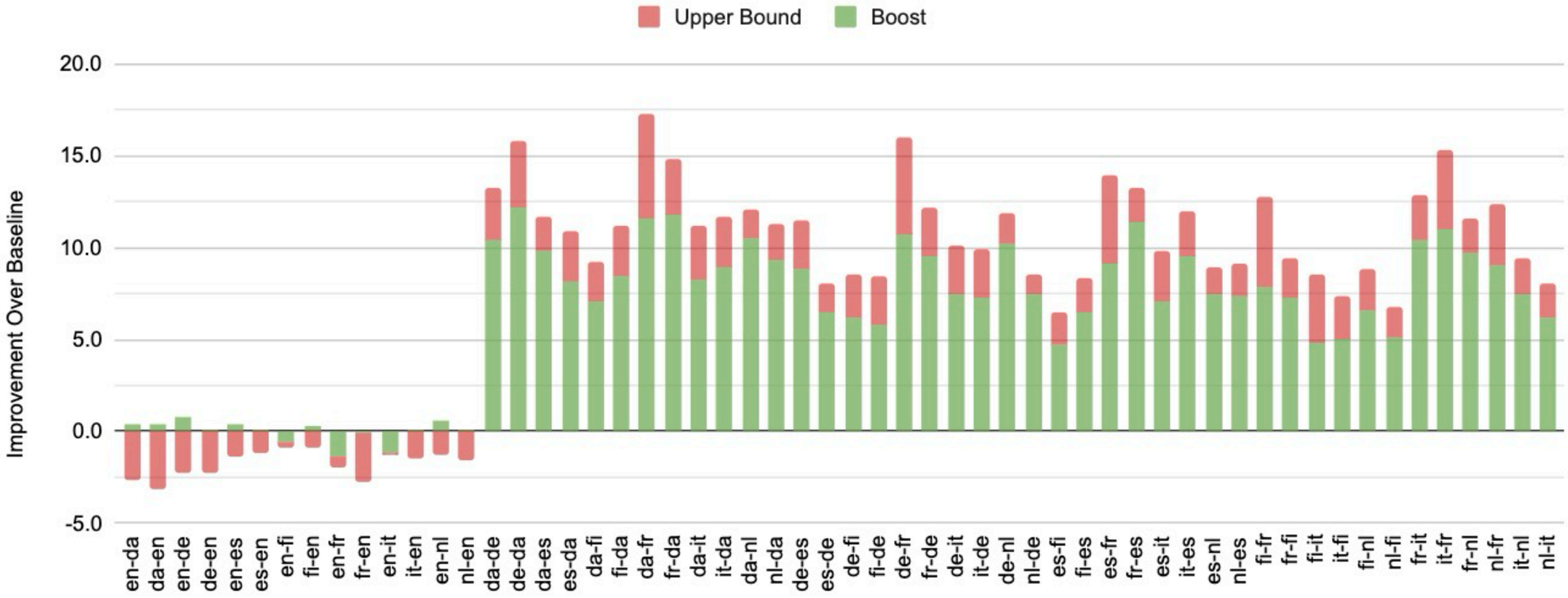


(b) Increase the Number of **S**amples

Incorporating more data leads to better results.

Incorporating 10% directions is almost the same as 100% directions.

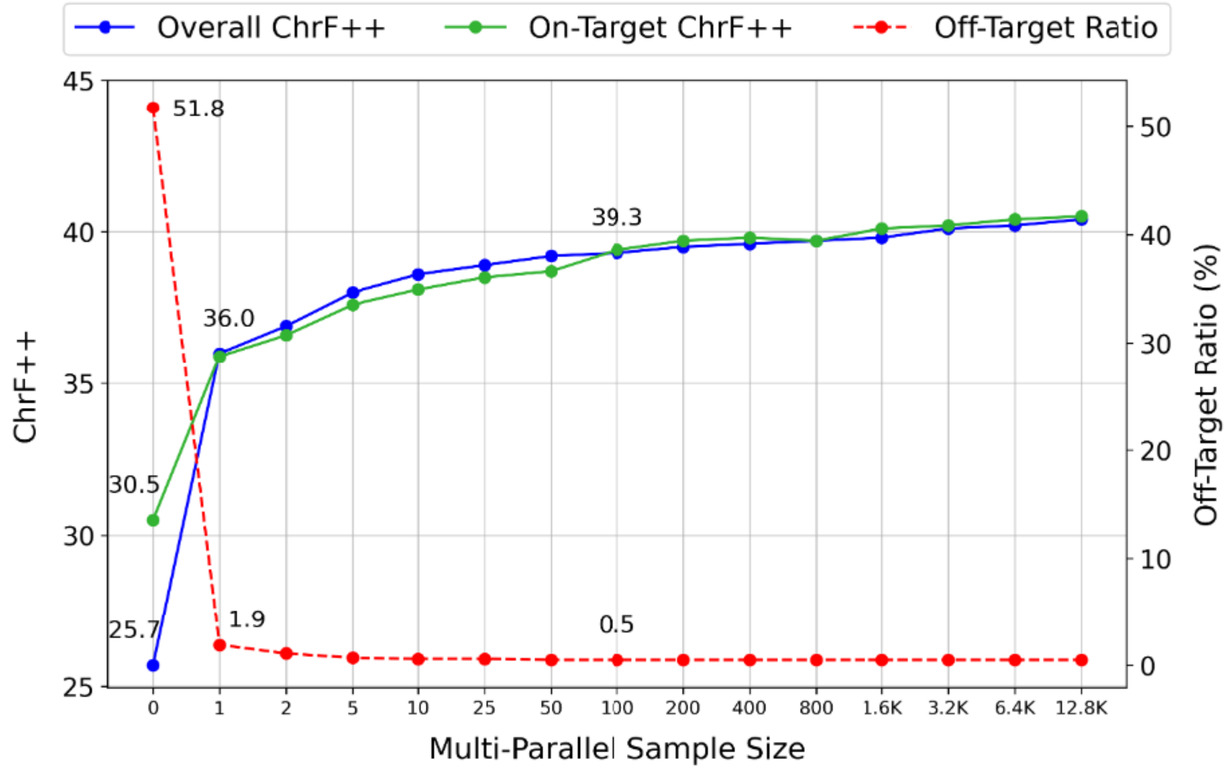
How close to the upper bound (complete Translation)?



Upper-Bound Model: trained with 1.2M multi-parallel / complete training data (Europarl-8).

Boost Model: boost the en-centric model on zero-shot directions.

Effects on off-target issues



Only 1 sample can almost solve the Off-target issue.

Where do the surprising effects come from?

- Does the fine-tuning change tag embeddings a lot? -> Not at all

Where do the surprising effects come from?

- Does the fine-tuning change tag embeddings a lot? -> Not at all
- Does the fine-tuning change word embeddings a lot? -> Not at all

Where do the surprising effects come from?

- Does the fine-tuning change tag embeddings a lot? -> Not at all
- Does the fine-tuning change word embeddings a lot? -> Not at all
- Does Semantic-level information matter? -> Yes!

Boosting with Numbers

[FR] 11, 21, 31
[FR] 12, 22, 32
…
[DE] 11, 21, 31
[DE] 12, 22, 32

[DE] 11, 21, 31
[DE] 12, 22, 32
…
[ES] 11, 21, 31
[ES] 12, 22, 32

Boosting with Word Pairs

[FR] Bonjour
[FR] Monde
…
[DE] Hallo
[DE] Welt

[DE] Hallo
[DE] Welt
…
[ES] Hola
[ES] Mundo

Setting	EN-X	X-EN	Zero-Shot	Off-Target (%)
Baseline	49.8	51.3	25.7	51.8
Numbers	49.9	51.5	26.8	46.1
Words	48.3	50.7	35.8	3.6

Where do the surprising effects come from?

- Does the fine-tuning change tag embeddings a lot? -> Not at all
- Does the fine-tuning change word embeddings a lot? -> Not at all
- Does Semantic-level information matter? -> Yes!
- Does Syntactic-level information matter? -> Yes!

Setting	EN-X	X-EN	Zero-Shot	Off-Target (%)
Baseline	49.8	51.3	25.7	51.8
Numbers	49.9	51.5	26.8	46.1
Words	48.3	50.7	35.8	3.6
NTREX	49.5	50.9	40.0	0.5

Conclusion

- English-centric models can be easily boosted using a tiny amount of non-English data for fine-tuning.
- The boosting effects are quite surprising:
  - High efficiency: 10+BLEU via 100 samples.
  - High effect: Quite close to the upper bound.
  - Off-target: 1 sample can almost solve it.
- We call on the community:
  - To consider the use of fine-tuning as a strong baseline for zero-shot translation.
  - To construct more comprehensive and high-quality multi-parallel data to cover real-world demand.