MACHINE LEARNING DEPARTMENT

Carnegie Mellon University

Contextual Parameter Generation for Universal Neural Machine Translation

Anthony Platanios
e.a.platanios@cs.cmu.edu

Joint work with Mrinmaya Sachan, Graham Neubig, and Tom Mitchell

Machine Translation (MT)

An MT system can translate from one language to another automatically, meaning without requiring human input.

Neural Machine Translation (NMT)

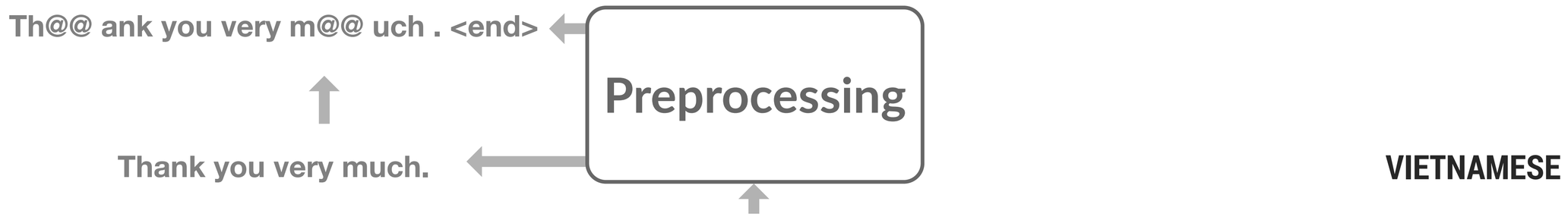
Carnegie Mellon University

Using neural networks to perform machine translation.

Neural Machine Translation (NMT)

Carnegie Mellon University

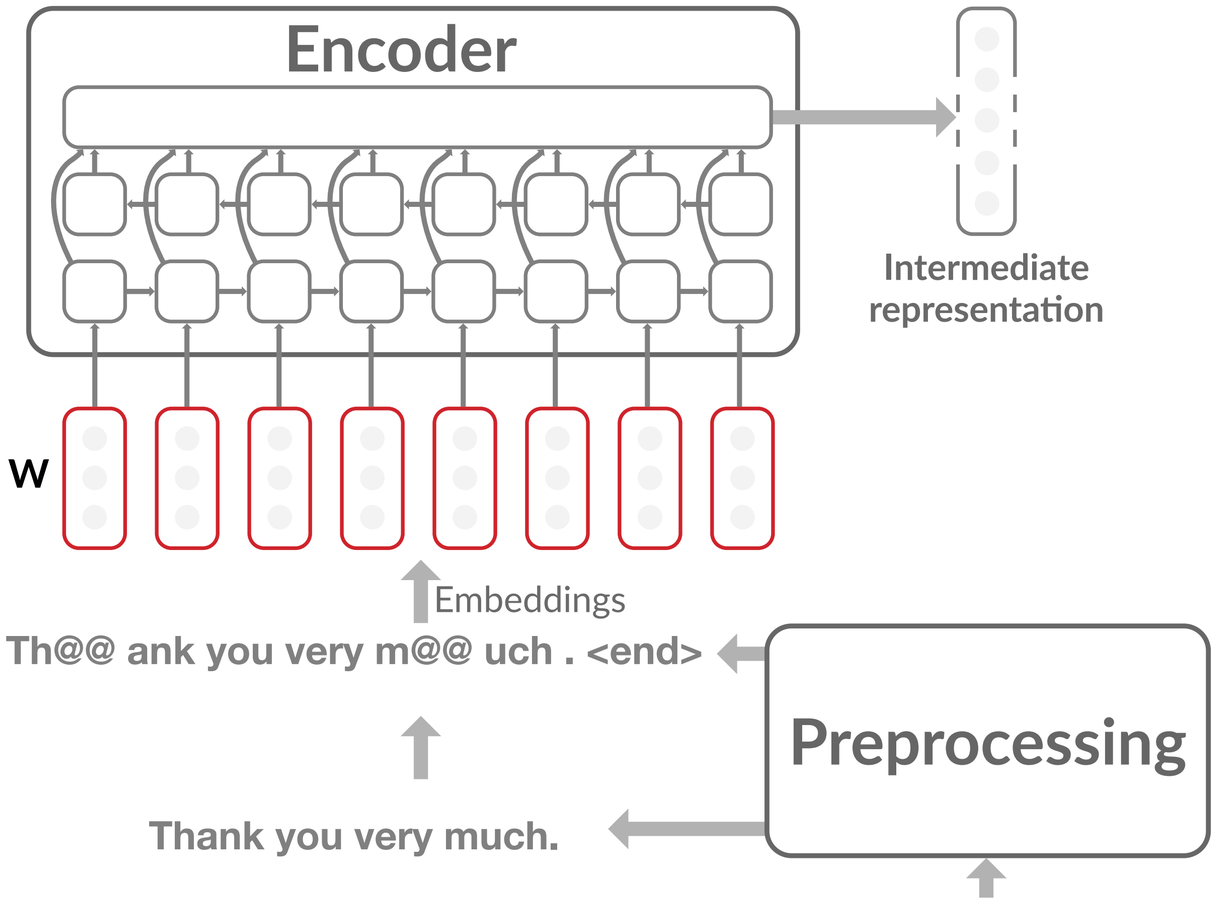
Using neural networks to perform machine translation. For example, recurrent neural networks (RNN) are commonly used:



Neural Machine Translation (NMT)

Carnegie Mellon University

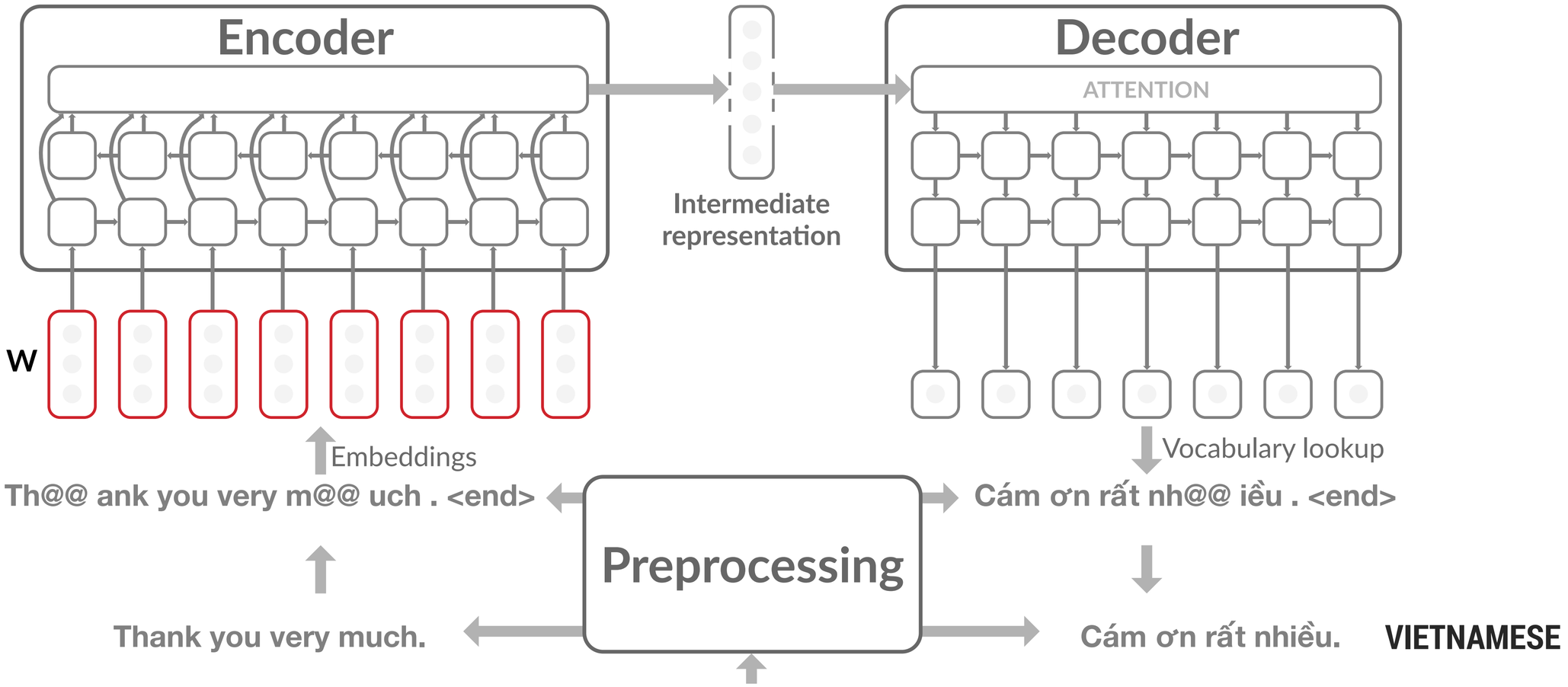
Using neural networks to perform machine translation. For example, recurrent neural networks (RNN) are commonly used:



Neural Machine Translation (NMT)

Carnegie Mellon University

Using neural networks to perform machine translation. For example, recurrent neural networks (RNN) are commonly used:



Neural Machine Translation (NMT)

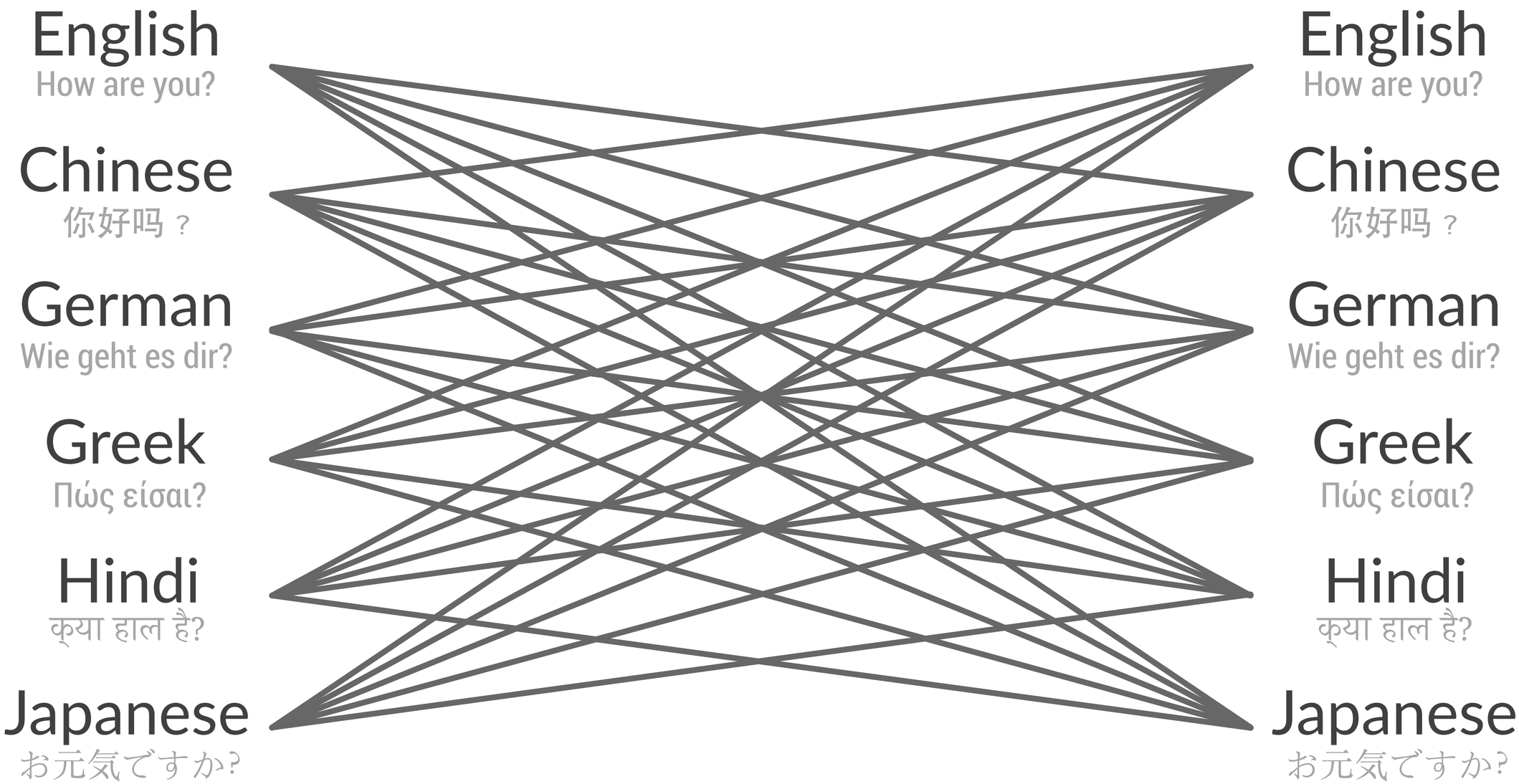
Carnegie Mellon University

More formally, we typically have:

- an **encoder** described by a function $\mathbf{f}^{(enc)}$, parameterized by $\theta^{(enc)} \in \mathbb{R}^{P^{(enc)}}$
- a **decoder** described by a function $\mathbf{f}^{(dec)}$, parameterized by $\theta^{(dec)} \in \mathbb{R}^{P^{(dec)}}$

Translating an input sentence x, consists of simply evaluating $\mathbf{f}^{(dec)}(\mathbf{f}^{(enc)}(x))$.

Multilingual NMT



Multilingual NMT

Carnegie Mellon University

English
How are you?

Chinese
你好吗？

German
Wie geht es dir?

Greek
Πώς είσαι?

Hindi
क्या हाल है?

Japanese
お元気ですか?

We want to be able to
translate between any pair of languages

English
How are you?

Chinese
你好吗？

German
Wie geht es dir?

Greek
Πώς είσαι?

Hindi
क्या हाल है?

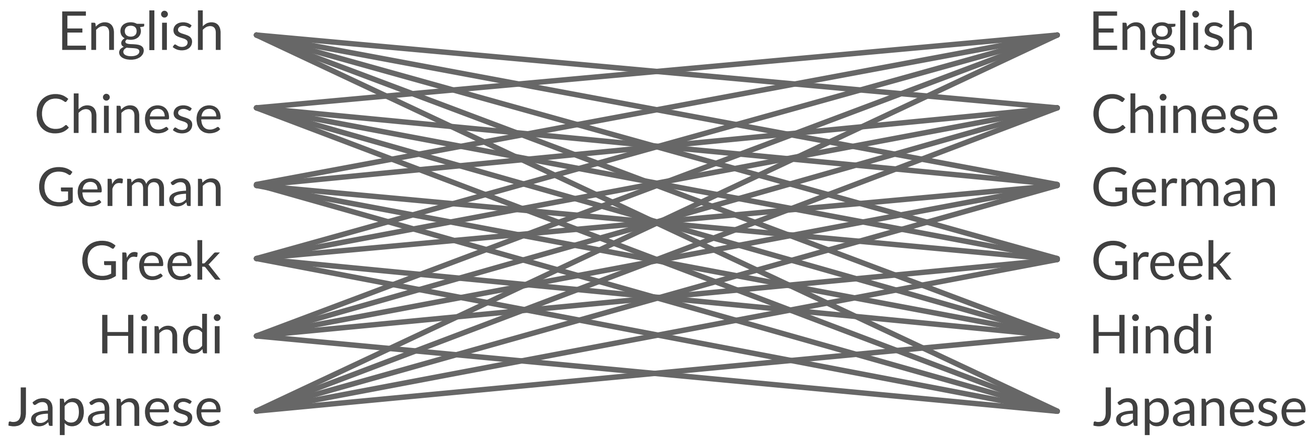
Japanese
お元気ですか?

Multilingual NMT **Approaches**

Carnegie Mellon University

Assuming **L** languages, and **P** parameters in a pairwise model, we can use:

- Pairwise models: Use a separate model pair language pair
  - **O(L²P) parameters** (separate $\theta^{(enc)}$ and $\theta^{(dec)}$ for each language pair)
  - No parameter sharing, over-parameterization, overfitting
  - Bad performance for limited or no training data



Multilingual NMT **Approaches**

Carnegie Mellon University

Assuming **L** languages, and **P** parameters in a pairwise model, we can use:

- Universal model: Uses one shared model
  - **O(P) parameters** (same $\theta^{(enc)}$ and $\theta^{(dec)}$ for all language pairs)
  - Lacks any language-specific parameterization

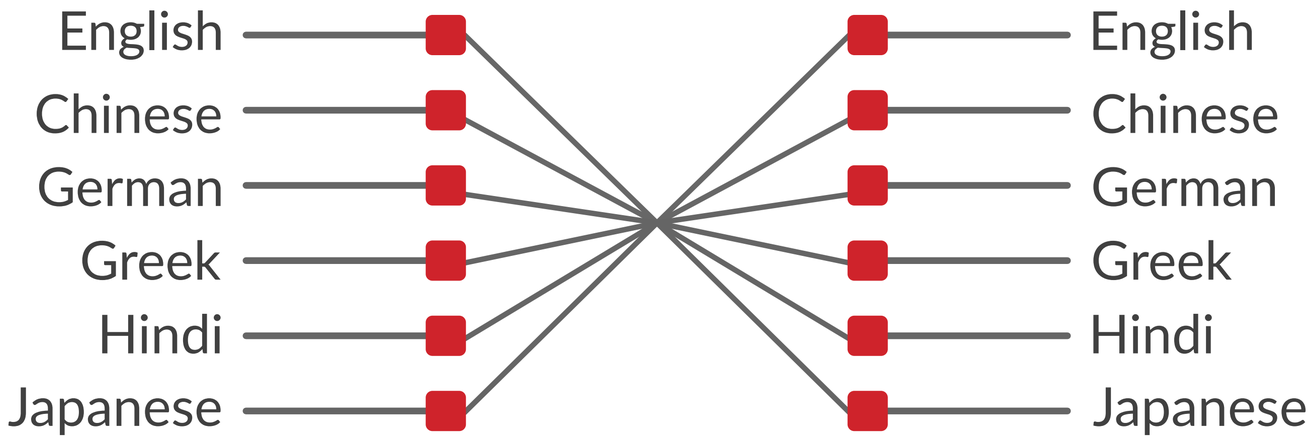
[Ha16, Johnson17]

Multilingual NMT **Approaches**

Carnegie Mellon University

Assuming **L** languages, and **P** parameters in a pairwise model, we can use:

- Per-language encoder/decoder: Uses a different encoder and a different decoder for each language, but the same intermediate representation
  - **O(LP) parameters** (separate $\theta^{(enc)}$ and $\theta^{(dec)}$ for each language)
  - Limited parameter sharing and also makes use of attention difficult



[Luong16, Firat16]

Multilingual NMT **Our Approach**

Carnegie Mellon University

We propose the **contextual parameter generator (CPG)**, which is:

- A generalization of the aforementioned methods
- Simple: It can be applied to most existing NMT systems with minor changes
- Multilingual: Enables multilingual translation using a single model

Multilingual NMT **Our Approach**

Carnegie Mellon University

We propose the **contextual parameter generator (CPG)**, which is:

- A generalization of the aforementioned methods
- Simple: It can be applied to most existing NMT systems with minor changes
- Multilingual: Enables multilingual translation using a single model
- Semi-supervised: Can use monolingual data
- Scalable: Reduces the number of parameters by employing extensive, yet controllable, sharing across languages
- Adaptable: Can adapt to support new languages, without complete retraining

Multilingual NMT **Our Approach**

Carnegie Mellon University

We propose the **contextual parameter generator (CPG)**, which is:

- A generalization of the aforementioned methods
- Simple: It can be applied to most existing NMT systems with minor changes
- Multilingual: Enables multilingual translation using a single model
- Semi-supervised: Can use monolingual data
- Scalable: Reduces the number of parameters by employing extensive, yet controllable, sharing across languages
- Adaptable: Can adapt to support new languages, without complete retraining

We achieve that by **learning embeddings for languages** and using them as **context** for a universal model.

Multilingual NMT **Our Approach**

Carnegie Mellon University

We achieve that by **learning embeddings for languages** and using them as **context** for a universal model.

Let $\mathbf{l_s}$ denote the source language embedding, and $\mathbf{l_t}$ the target one. Then, we can define:

$$\theta^{(enc)} \triangleq g^{(enc)}(\mathbf{l_s})$$
$$\theta^{(dec)} \triangleq g^{(dec)}(\mathbf{l_t})$$

for functions $g^{(enc)}$ and $g^{(dec)}$ that we learn. We call these functions the **parameter generator networks**.

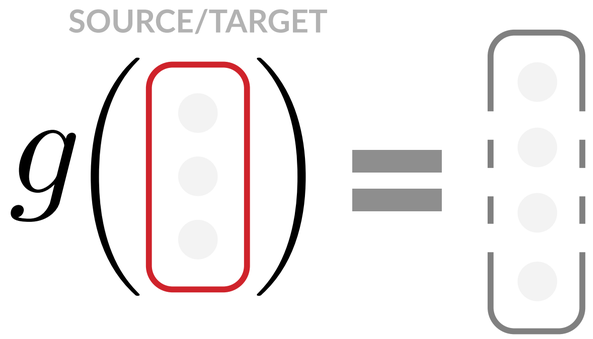
CPG: Parameter Generator Networks

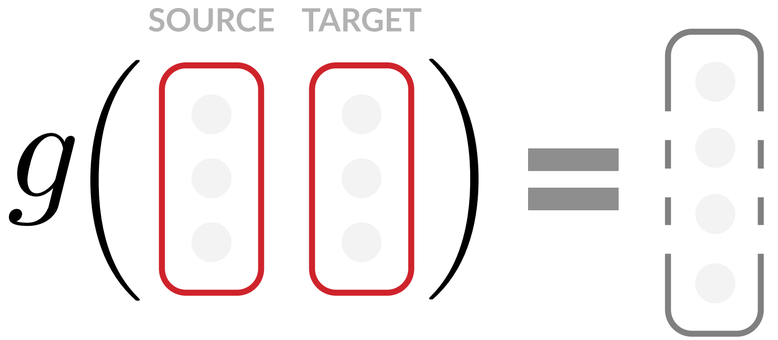
Carnegie Mellon University

Let $\mathbf{l_s}$ denote the source language embedding, and $\mathbf{l_t}$ the target one. Then, we can define:

$$\theta^{(enc)} \triangleq g^{(enc)}(\mathbf{l_s})$$
$$\theta^{(dec)} \triangleq g^{(dec)}(\mathbf{l_t})$$

$$\theta^{(enc)} \triangleq g^{(enc)}(\mathbf{l_s}, \mathbf{l_t})$$
$$\theta^{(dec)} \triangleq g^{(dec)}(\mathbf{l_s}, \mathbf{l_t})$$





CPG: Parameter Generator Networks

Carnegie Mellon University

Let $\mathbf{l_s}$ denote the source language embedding, and $\mathbf{l_t}$ the target one. Then, we can define:

$$\theta^{(enc)} \triangleq g^{(enc)}(\mathbf{l_s})$$
$$\theta^{(dec)} \triangleq g^{(dec)}(\mathbf{l_t})$$

DECOUPLED

$$\theta^{(enc)} \triangleq g^{(enc)}(\mathbf{l_s}, \mathbf{l_t})$$
$$\theta^{(dec)} \triangleq g^{(dec)}(\mathbf{l_s}, \mathbf{l_t})$$

COUPLED

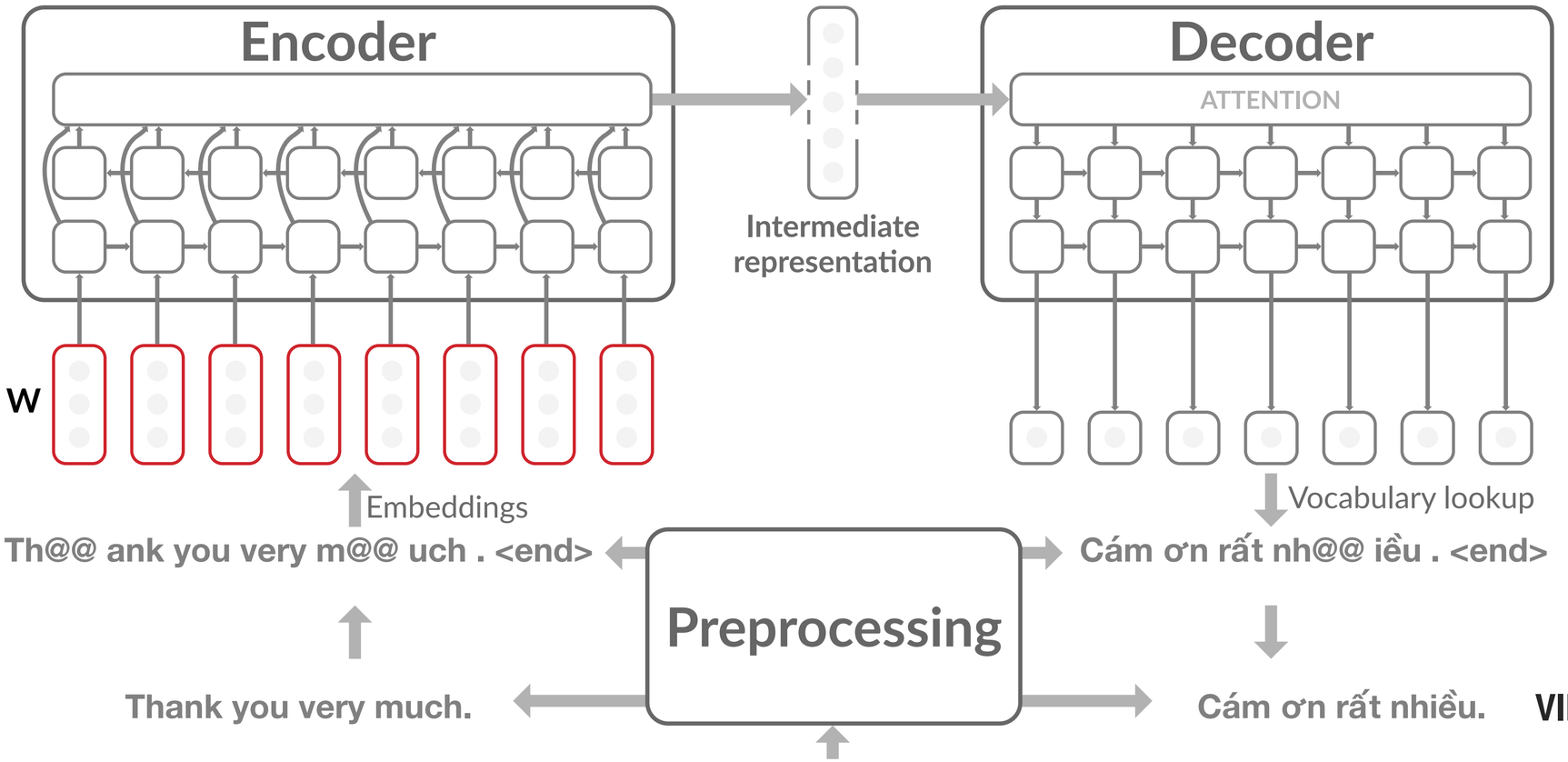
The decoupled form makes for a stronger argument that the representation produced by the encoder could be **approaching a universal interlingua;**

... more so than methods that are aware of the target language when they perform encoding.

CPG: Overview

Carnegie Mellon University

Typical pairwise NMT model:

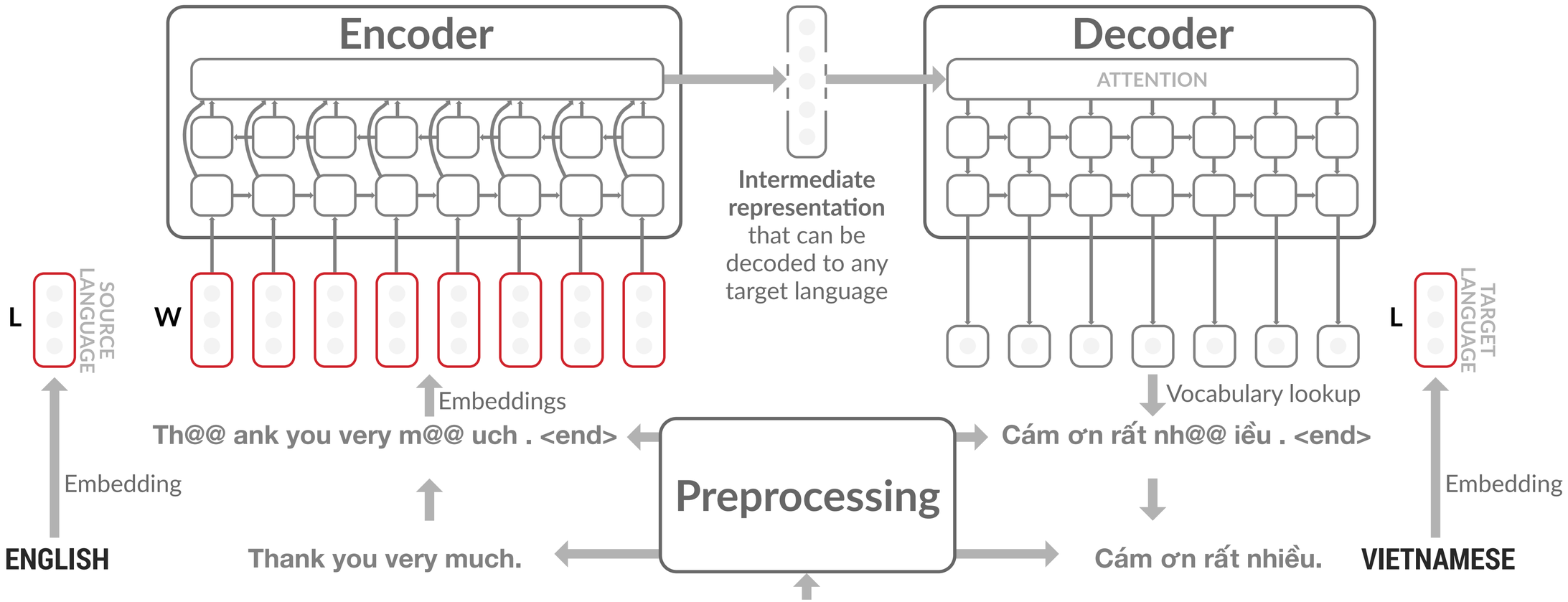


CPG: Overview

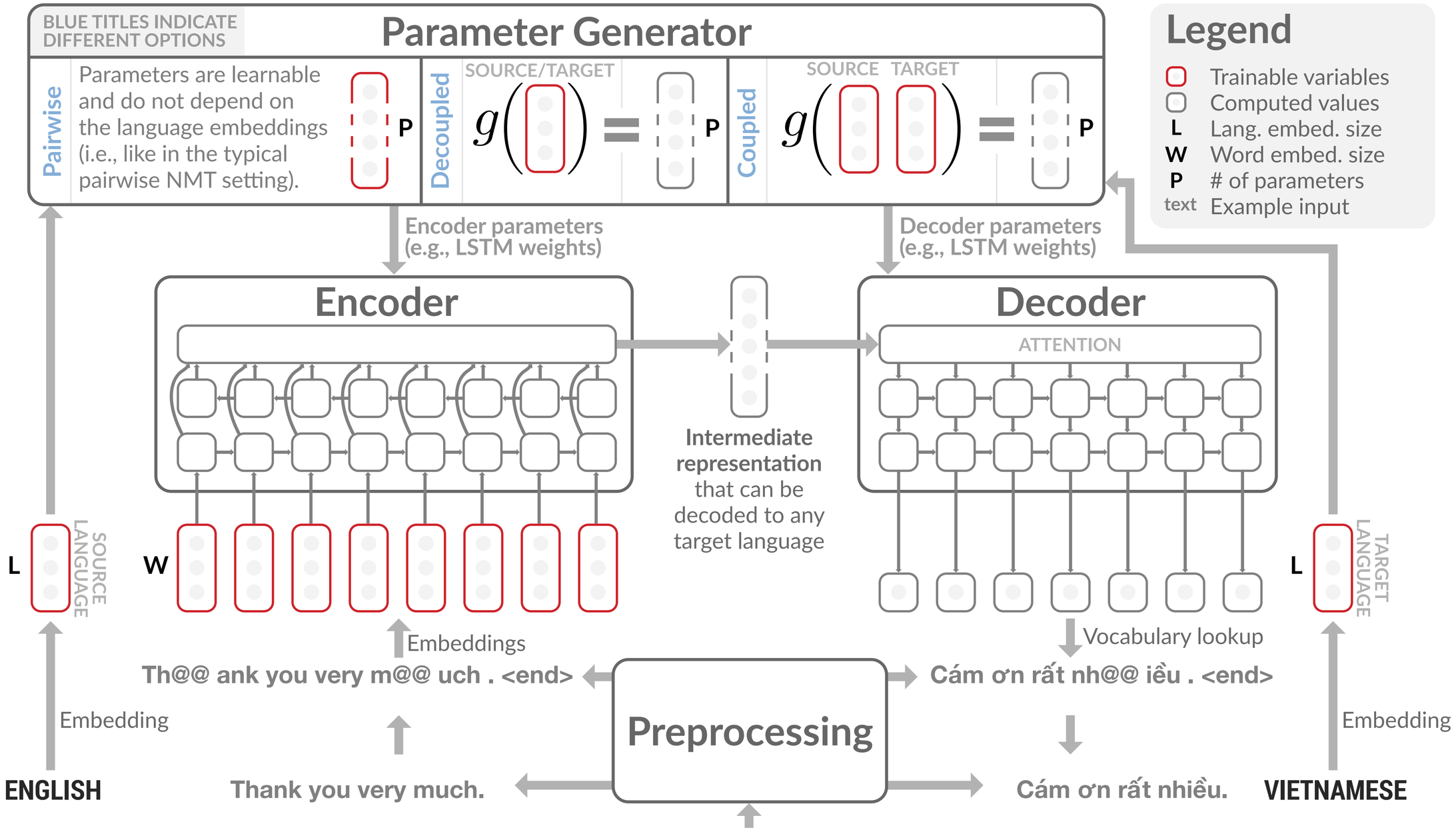
Carnegie Mellon University

Legend

- Trainable variables
- Computed values
- **L** Lang. embed. size
- **W** Word embed. size
- **P** # of parameters
- text Example input



CPG: Overview



CPG: Parameter Generator Networks

Carnegie Mellon University

Our goal is to provide a **simple form** for the parameter generator networks, that works and for which we can reason about. For this reason, we use simple **linear transforms**:

$$g^{(enc)}(\mathbf{l_s}) \triangleq \mathbf{W^{(enc)}}\mathbf{l_s}$$

$$g^{(dec)}(\mathbf{l_t}) \triangleq \mathbf{W^{(dec)}}\mathbf{l_t}$$

CPG: Parameter Generator Networks

Carnegie Mellon University

Our goal is to provide a **simple form** for the parameter generator networks, that works and for which we can reason about. For this reason, we use simple **linear transforms**:

$$g^{(enc)}(\mathbf{l_s}) \triangleq \mathbf{W^{(enc)}}\mathbf{l_s}$$

$$g^{(dec)}(\mathbf{l_t}) \triangleq \mathbf{W^{(dec)}}\mathbf{l_t}$$

For each language, the parameters of the encoder/decoder are defined as a **linear combination of the M columns** of the corresponding weight matrix, where $\mathbf{l_s}, \mathbf{l_t} \in \mathbb{R}^{\mathbf{M}}$.

CPG: Controlled Parameter Sharing

Carnegie Mellon University

- The encoder/decoder parameters often have some ***"natural grouping"*** (e.g., the weight matrix of the first LSTM layer forms a group).
- The language embeddings need to represent all language-specific information and thus may need to be large.
- Only a small part of that information may be relevant for each *"group"*.

CPG: Controlled Parameter Sharing

Carnegie Mellon University

- The encoder/decoder parameters often have some ***"natural grouping"*** (e.g., the weight matrix of the first LSTM layer forms a group).
- The language embeddings need to represent all language-specific information and thus may need to be large.
- Only a small part of that information may be relevant for each *"group"*.

We can use these observations to **control the amount of information sharing across languages.**

CPG: Controlled Parameter Sharing

Carnegie Mellon University

Let $\theta^{(enc)} = \{\theta_j^{(enc)}\}_{j=1}^{G}$ and $\theta_j^{(enc)} \in \mathbb{R}^{P_j^{(enc)}}$, where G is the number of groups.

CPG: Controlled Parameter Sharing

Carnegie Mellon University

Let $\theta^{(enc)} = \{\theta_j^{(enc)}\}_{j=1}^{G}$ and $\theta_j^{(enc)} \in \mathbb{R}^{P_j^{(enc)}}$, where G is the number of groups.

Then, we can define:

$$\theta_j^{(enc)} \triangleq \mathbf{W_j^{(enc)}} \mathbf{P_j^{(enc)}} \mathbf{l_s}$$

where:

$$\mathbf{W_j^{(enc)}} \in \mathbb{R}^{P_j^{(enc)} \times M'}$$

$$\mathbf{P_j^{(enc)}} \in \mathbb{R}^{M' \times M}$$

$$M' < M$$

and similarly for the decoder.

CPG: Controlled Parameter Sharing

Carnegie Mellon University

Let $\theta^{(enc)} = \{\theta_j^{(enc)}\}_{j=1}^{G}$ and $\theta_j^{(enc)} \in \mathbb{R}^{P_j^{(enc)}}$, where G is the number of groups.

Then, we can define:

$$\theta_j^{(enc)} \triangleq \mathbf{W_j^{(enc)}} \mathbf{P_j^{(enc)}} \mathbf{l_s}$$

where:

$$\mathbf{W_j^{(enc)}} \in \mathbb{R}^{P_j^{(enc)} \times M'}$$

$$\mathbf{P_j^{(enc)}} \in \mathbb{R}^{M' \times M}$$

$$M' < M$$

and similarly for the decoder.

If we want to increase the number of per-language parameters, we can increase **M**, while keeping **M'** fixed, and vice-versa.

CPG: Parameter Generator Networks

Carnegie Mellon University

Note that the proposed does not depend on the specific parameter generator network used. It would be interesting to design models that can use side-information about the languages, that may be available.

CPG: Benefits

Carnegie Mellon University

Semi-Supervised

Can use monolingual corpora by learning to translate back-and-forth for each language → Learn language embeddings that encode meaningful priors / language models

CPG: Benefits

Carnegie Mellon University

Semi-Supervised

Can use monolingual corpora by learning to translate back-and-forth for each language → Learn language embeddings that encode meaningful priors / language models

Zero-Shot

Can translate between unsupervised pairs of languages, as long as the languages have been seen in any supervised pairs

CPG: Benefits

Carnegie Mellon University

Semi-Supervised

Can use monolingual corpora by learning to translate back-and-forth for each language → Learn language embeddings that encode meaningful priors / language models

Zero-Shot

Can translate between unsupervised pairs of languages, as long as the languages have been seen in any supervised pairs

Potential for Adaptation

Given a pre-trained model, can adapt to support a new language by just learning the language embedding and fixing the rest of the model

CPG: Number of Parameters

Carnegie Mellon University

Pairwise NMT

$$\mathcal{O}(L^2P + 2L^2WV)$$

CPG NMT

$$\mathcal{O}(PM + LM + LWV)$$

P = Number of parameters of the encoder and the decoder together

W = Word embeddings size

V = Per-language vocabulary size

L = Number of languages

M = Language embeddings size

CPG: Number of Parameters

Carnegie Mellon University

Pairwise NMT	CPG NMT
$\mathcal{O}(L^2P + 2L^2WV)$	$\mathcal{O}(PM + LM + LWV)$
From experiments: ~832M	~124M

P = Number of parameters of the encoder and the decoder together

W = Word embeddings size

V = Per-language vocabulary size

L = Number of languages

M = Language embeddings size

CPG: Number of Parameters

Carnegie Mellon University

Pairwise NMT

$$\mathcal{O}(L^2P + 2L^2WV)$$

CPG NMT

$$\mathcal{O}(PM + LM + LWV)$$

Also depends on the choice of vocabulary (e.g., shared or not).

P = Number of parameters of the encoder and the decoder together

W = Word embeddings size

V = Per-language vocabulary size

L = Number of languages

M = Language embeddings size

Experiments

Carnegie Mellon University

Baseline Model

- 2-layer bidirectional LSTM encoder
- 2-layer LSTM decoder
- 512 hidden units per layer and word embedding size
- AMSGrad optimizer (similar to Adam) with learning rate 0.001
- Label smoothing factor = 0.1
- Batch size = 128
- Beam width = 10 (using GNMT length normalization)

Experiments

Carnegie Mellon University

Baseline Model

- 2-layer bidirectional LSTM encoder
- 2-layer LSTM decoder
- 512 hidden units per layer and word embedding size
- AMSGrad optimizer (similar to Adam) with learning rate 0.001
- Label smoothing factor = 0.1
- Batch size = 128
- Beam width = 10 (using GNMT length normalization)

Vocabulary

- Per-language vocabulary
- 20,000 most frequent words with a cutoff frequency of 5 (i.e., no BPE)

Experiments

Carnegie Mellon University

Baseline Model

- 2-layer bidirectional LSTM encoder
- 2-layer LSTM decoder
- 512 hidden units per layer and word embedding size
- AMSGrad optimizer (similar to Adam) with learning rate 0.001
- Label smoothing factor = 0.1
- Batch size = 128
- Beam width = 10 (using GNMT length normalization)

Vocabulary

- Per-language vocabulary
- 20,000 most frequent words with a cutoff frequency of 5 (i.e., no BPE)

All experiments were run on a machine with a single Nvidia V100 GPU, and 24 GBs of system memory.

The longest experiment required ~10 hours.

Experiments

Carnegie Mellon University

Settings

- Supervised: Training using full parallel corpora
- Low-Resource: Limiting the size of there parallel corpora (still using the rest of the data as monolingual)
- Zero-Shot: No parallel sentences for some language pairs

Experiments

Carnegie Mellon University

Settings

- Supervised: Training using full parallel corpora
- Low-Resource: Limiting the size of there parallel corpora (still using the rest of the data as monolingual)
- Zero-Shot: No parallel sentences for some language pairs

Datasets

- IWSLT-15: Used for supervised and low-resource experiments
  - *Languages:* Czech (`Ch`), English (`En`), French (`Fr`), German (`De`), Thai (`Th`), and Vietnamese (`Vi`)
  - ~90,000-220,000 training sentence pairs, ~500-900 validation, and ~1,000-1,300 test
- IWSLT-17: Used for supervised and zero-shot experiments
  - *Languages:* Dutch (`Nl`), English (`En`), German (`De`), Italian (`It`), and Romanian (`Ro`)
  - ~220,000 training sentence pairs, ~900 validation, and ~1,100 test

Experiments: **IWSLT-15**

Carnegie Mellon University

BLEU Scores

Pairwise NMT baselines

	PNMT	GML	CPG*	CPG
En→Cs	14.89	15.92	16.88	**17.22**
Cs→En	24.43	25.25	26.44	**27.37**
En→De	25.99	15.92	26.41	**26.77**
De→En	30.93	29.60	31.24	**31.77**
En→Fr	38.25	34.40	38.10	**38.32**
Fr→En	37.40	35.14	37.11	**37.89**
En→Th	23.62	22.22	26.03	**26.33**
Th→En	15.54	14.03	16.54	**26.77**
En→Vi	27.47	25.54	28.33	**29.03**
Vi→En	24.03	23.19	25.91	**26.38**
Mean	26.26	24.12	27.30	**27.80**

Experiments: **IWSLT-15**

Carnegie Mellon University

BLEU Scores

Google multilingual model [Johnson17]

	PNMT	GML	CPG*	CPG
En→Cs	14.89	15.92	16.88	**17.22**
Cs→En	24.43	25.25	26.44	**27.37**
En→De	25.99	15.92	26.41	**26.77**
De→En	30.93	29.60	31.24	**31.77**
En→Fr	38.25	34.40	38.10	**38.32**
Fr→En	37.40	35.14	37.11	**37.89**
En→Th	23.62	22.22	26.03	**26.33**
Th→En	15.54	14.03	16.54	**26.77**
En→Vi	27.47	25.54	28.33	**29.03**
Vi→En	24.03	23.19	25.91	**26.38**
Mean	26.26	24.12	27.30	**27.80**

Experiments: **IWSLT-15**

Carnegie Mellon University

BLEU Scores

Trained without auto-encoding (i.e., monolingual data)

	PNMT	GML	CPG*	CPG
En→Cs	14.89	15.92	16.88	**17.22**
Cs→En	24.43	25.25	26.44	**27.37**
En→De	25.99	15.92	26.41	**26.77**
De→En	30.93	29.60	31.24	**31.77**
En→Fr	38.25	34.40	38.10	**38.32**
Fr→En	37.40	35.14	37.11	**37.89**
En→Th	23.62	22.22	26.03	**26.33**
Th→En	15.54	14.03	16.54	**26.77**
En→Vi	27.47	25.54	28.33	**29.03**
Vi→En	24.03	23.19	25.91	**26.38**
Mean	26.26	24.12	27.30	**27.80**

Experiments: **IWSLT-15**

Carnegie Mellon University

BLEU Scores

	PNMT	GML	CPG*	CPG
En→Cs	14.89	15.92	16.88	**17.22**
Cs→En	24.43	25.25	26.44	**27.37**
En→De	25.99	15.92	26.41	**26.77**
De→En	30.93	29.60	31.24	**31.77**
En→Fr	38.25	34.40	38.10	**38.32**
Fr→En	37.40	35.14	37.11	**37.89**
En→Th	23.62	22.22	26.03	**26.33**
Th→En	15.54	14.03	16.54	**26.77**
En→Vi	27.47	25.54	28.33	**29.03**
Vi→En	24.03	23.19	25.91	**26.38**
Mean	26.26	24.12	27.30	**27.80**

> 25.87 [Ha16]

> 28.07 [Huang18]

Experiments: **IWSLT-15**

Carnegie Mellon University

BLEU Scores using only 10% of the parallel corpus

	PNMT	GML	CPG*	CPG
En→Cs	5.71	8.18	8.40	**9.49**
Cs→En	6.64	14.56	14.81	**15.38**
En→De	11.70	14.60	15.09	**16.03**
De→En	18.10	19.02	19.77	**20.25**
En→Fr	24.47	25.15	24.00	**25.79**
Fr→En	23.79	25.02	24.55	**27.12**
En→Th	7.86	15.58	**18.41**	17.65
Th→En	7.13	9.11	**10.19**	10.14
En→Vi	18.01	17.51	**18.92**	18.90
Vi→En	6.69	16.00	16.28	**16.86**
Mean	13.01	16.47	17.04	**17.76**

Experiments: **IWSLT-17**

Carnegie Mellon University

BLEU Scores

M=8

	PNMT	GML	CPG^{8}	CPG$^{8}_{C4}$
De→En	21.78	21.25	**22.56**	20.78
De→It	13.16	13.84	**14.73**	14.34
De→Ro	10.85	11.95	12.24	12.37
En→De	**19.75**	17.06	19.41	19.04
En→It	27.70	25.74	27.57	27.11
En→Nl	24.41	22.46	24.47	**25.15**
En→Ro	19.23	18.60	20.83	**20.96**
It→De	14.39	12.76	14.61	**15.06**
It→En	29.84	27.96	**30.62**	30.10
It→Nl	16.74	16.27	17.99	**18.11**
Nl→En	26.30	24.78	26.31	26.17
Nl→It	16.03	16.10	16.81	**17.50**
Nl→Ro	12.84	12.48	14.01	**14.44**
Ro→De	12.75	12.21	13.58	**13.66**
Ro→En	24.33	22.88	23.83	23.88
Ro→Nl	13.70	14.11	15.34	**15.51**
Mean	18.99	18.15	19.68	**19.75**

Experiments: **IWSLT-17**

Carnegie Mellon University

BLEU Scores

M=8
M'=4

	PNMT	GML	CPG^{8}	CPG$^{8}_{C4}$
De→En	21.78	21.25	**22.56**	20.78
De→It	13.16	13.84	**14.73**	14.34
De→Ro	10.85	11.95	12.24	12.37
En→De	**19.75**	17.06	19.41	19.04
En→It	27.70	25.74	27.57	27.11
En→Nl	24.41	22.46	24.47	**25.15**
En→Ro	19.23	18.60	20.83	**20.96**
It→De	14.39	12.76	14.61	**15.06**
It→En	29.84	27.96	**30.62**	30.10
It→Nl	16.74	16.27	17.99	**18.11**
Nl→En	26.30	24.78	26.31	26.17
Nl→It	16.03	16.10	16.81	**17.50**
Nl→Ro	12.84	12.48	14.01	**14.44**
Ro→De	12.75	12.21	13.58	**13.66**
Ro→En	24.33	22.88	23.83	23.88
Ro→Nl	13.70	14.11	15.34	**15.51**
Mean	18.99	18.15	19.68	**19.75**

Experiments: **IWSLT-17**

Carnegie Mellon University

BLEU Scores

	PNMT	GML	CPG^{8}	CPG^{8}_{C4}	CPG^{8}_{C2}	CPG^{8}_{C1}	CPG^{64}_{C8}	CPG^{512}_{C8}
De→En	21.78	21.25	**22.56**	20.78	22.09	21.23	21.50	22.38
De→It	13.16	13.84	**14.73**	14.34	14.43	13.84	14.34	14.11
De→Ro	10.85	11.95	12.24	12.37	**12.72**	10.37	11.32	11.94
En→De	**19.75**	17.06	19.41	19.04	18.42	17.04	17.46	19.29
En→It	27.70	25.74	27.57	27.11	**28.21**	26.26	27.26	27.48
En→Nl	24.41	22.46	24.47	**25.15**	14.64	23.94	24.48	24.50
En→Ro	19.23	18.60	20.83	**20.96**	18.69	17.23	20.20	20.86
It→De	14.39	12.76	14.61	**15.06**	14.15	13.12	14.18	14.69
It→En	29.84	27.96	**30.62**	30.10	29.44	29.22	29.56	30.18
It→Nl	16.74	16.27	17.99	**18.11**	18.05	17.13	17.71	17.99
Nl→En	26.30	24.78	26.31	26.17	25.74	26.15	**26.33**	26.20
Nl→It	16.03	16.10	16.81	**17.50**	17.03	16.81	16.89	17.09
Nl→Ro	12.84	12.48	14.01	**14.44**	12.56	11.79	12.38	13.66
Ro→De	12.75	12.21	13.58	**13.66**	13.02	12.62	12.96	13.63
Ro→En	24.33	22.88	23.83	23.88	24.20	23.58	**24.65**	23.57
Ro→Nl	13.70	14.11	15.34	**15.51**	15.11	14.65	15.29	15.19
Mean	18.99	18.15	19.68	**19.75**	19.28	18.44	19.16	19.74

Experiments: **IWSLT-17**

Carnegie Mellon University

BLEU Scores for zero-shot translation

	PNMT	GML	CPG^{8}	CPG^{8}_{C4}	CPG^{8}_{C2}	CPG^{8}_{C1}	CPG^{64}_{C8}	CPG^{512}_{C8}
De→Nl	12.75	12.50	12.74	**12.80**	11.65	12.41	12.67	12.75
It→Ro	9.97	9.57	10.57	10.17	10.42	9.65	**10.69**	10.32
Nl→De	11.32	10.47	11.52	11.20	11.28	10.89	**11.63**	11.45
Ro→It	11.69	10.82	11.51	11.40	11.66	11.42	**11.78**	11.27
Mean	11.43	10.84	11.59	11.39	11.25	11.09	**11.69**	11.44

Experiments: **IWSLT-17**

Carnegie Mellon University

BLEU Scores for zero-shot translation

	PNMT	GML	CPG^{8}	CPG^{8}_{C4}	CPG^{8}_{C2}	CPG^{8}_{C1}	CPG^{64}_{C8}	CPG^{512}_{C8}
De→Nl	12.75	12.50	12.74	**12.80**	11.65	12.41	12.67	12.75
It→Ro	9.97	9.57	10.57	10.17	10.42	9.65	**10.69**	10.32
Nl→De	11.32	10.47	11.52	11.20	11.28	10.89	**11.63**	11.45
Ro→It	11.69	10.82	11.51	11.40	11.66	11.42	**11.78**	11.27
Mean	11.43	10.84	11.59	11.39	11.25	11.09	**11.69**	11.44

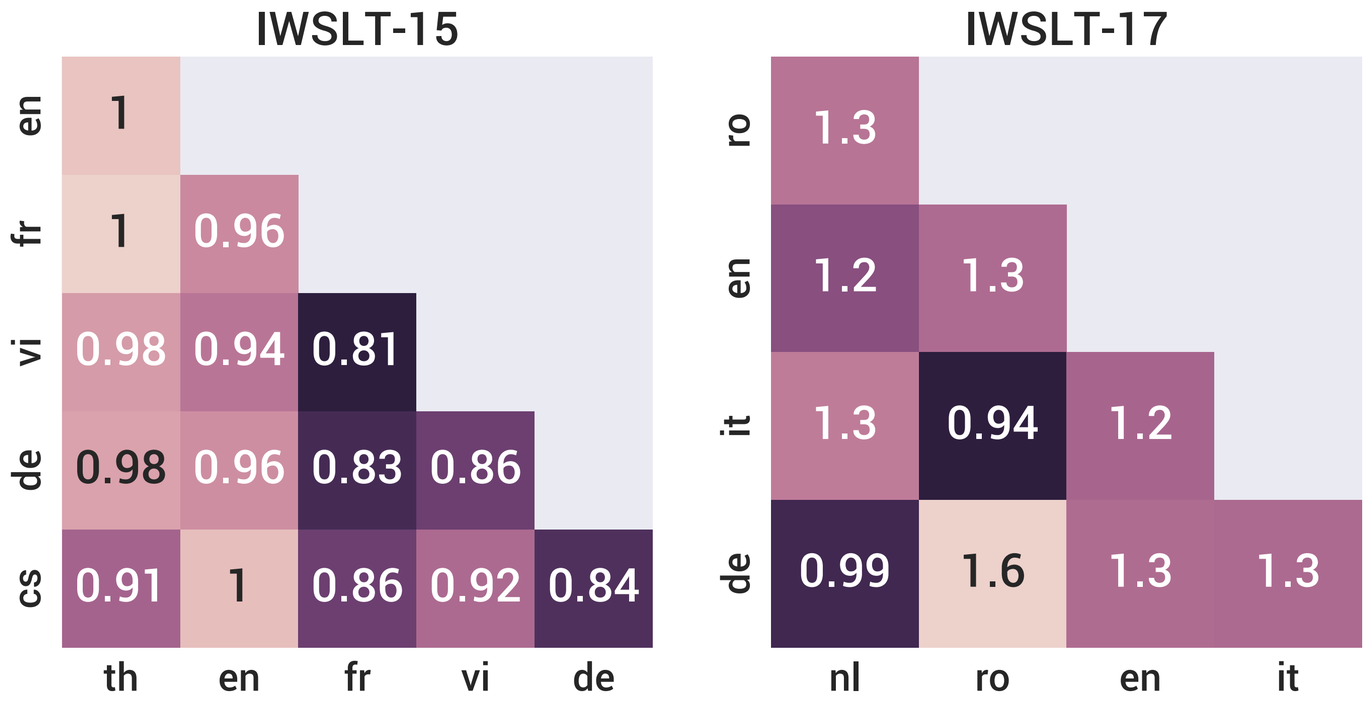
Our results are not comparable to those of the official ISWLT-17 report because we use much **smaller models** and potentially **less training data**.

Note that letting the models train for longer (i.e., 7 days instead of half a day) can result in significant improvements over the reported BLEU scores, in some cases.

Experiments: **Language Distances**

Carnegie Mellon University

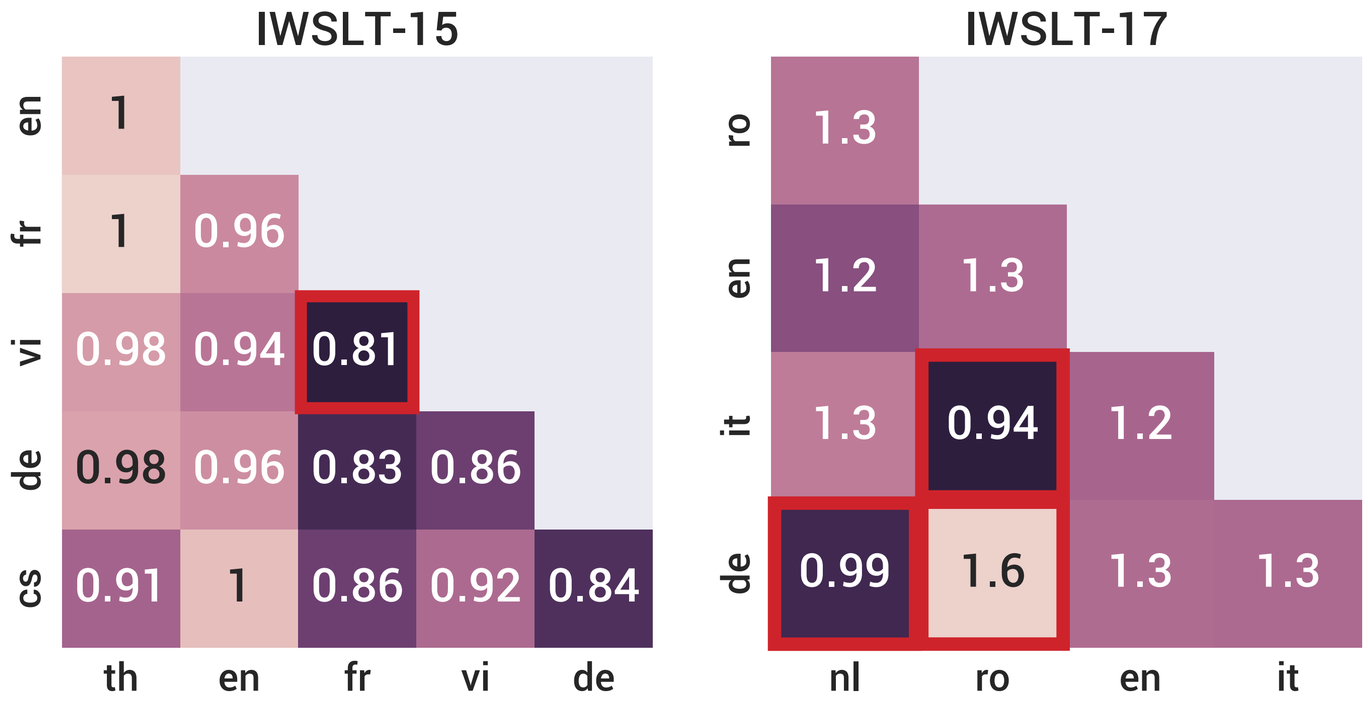
Cosine Distances



Experiments: Language Distances

Carnegie Mellon University

Cosine Distances



Conclusion

Carnegie Mellon University

We have proposed the **contextual parameter generator (CPG)**, which is:

- A generalization of the aforementioned methods
- Simple: It can be applied to most existing NMT systems with minor changes
- Multilingual: Enables multilingual translation using a single model
- Semi-supervised: Can use monolingual data
- Scalable: Reduces the number of parameters by employing extensive, yet controllable, sharing across languages
  - Has a number of parameters that is independent of the number of languages
- Adaptable: Can adapt to support new languages, without complete retraining

We achieve that by **learning embeddings for languages** and using them as **context** for a universal model.