Contextual Parameter Generation
for Universal Neural Machine Translation

Emmanouil Antonios Platanios e.a.platanios@cs.cmu.edu
Mrinmaya Sachan mrinmays@cs.cmu.edu
Graham Neubig gneubig@cs.cmu.edu
Tom M. Mitchell tom.mitchell@cs.cmu.edu

Problem

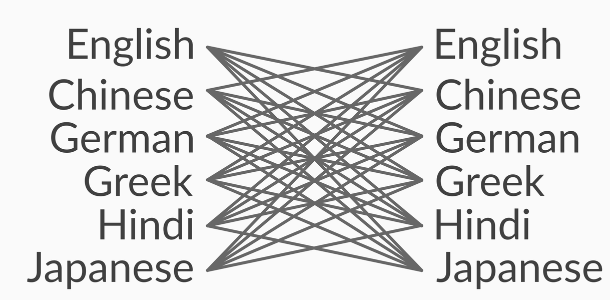
Translate from one language to another.

A multilingual MT system can translate between any pair of languages.

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

PAIRWISE

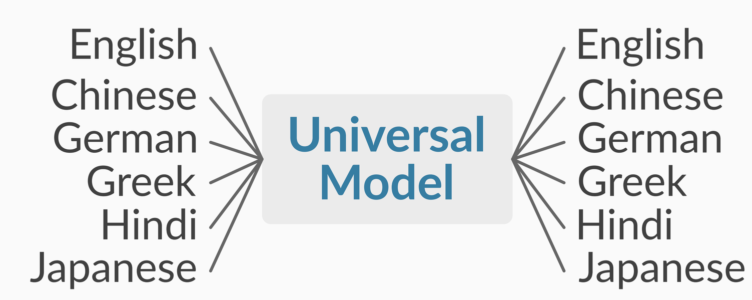
Separate model per language pair:



- **O(L²P) parameters**
- No parameter sharing
- Bad for limited/no training data

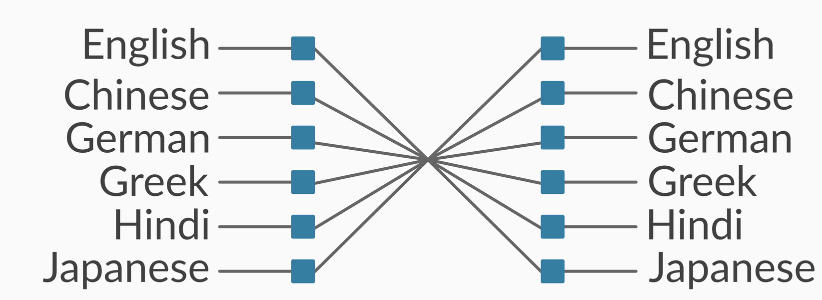
UNIVERSAL [Ha16, Johnson17]

One shared model:



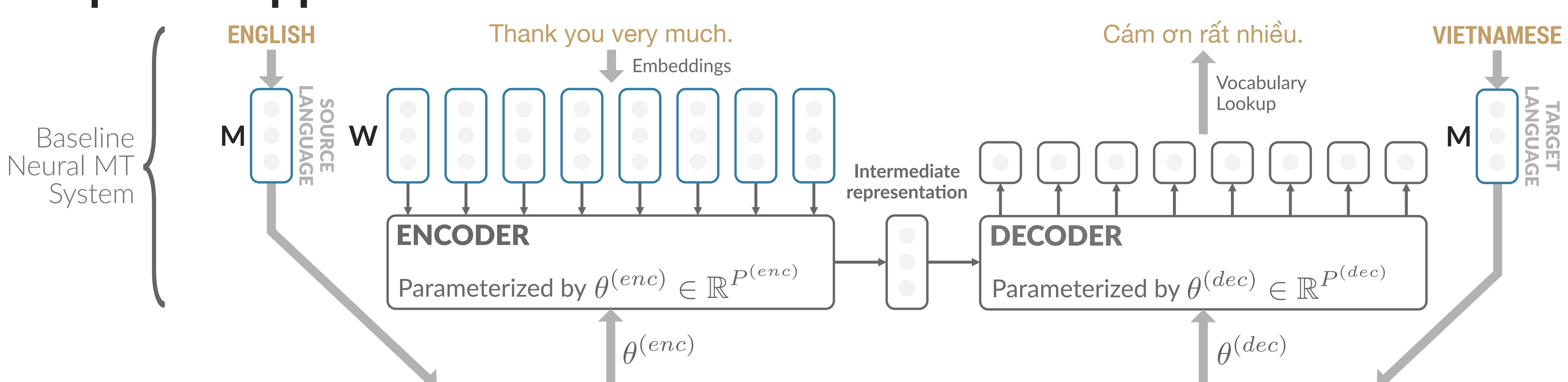
- **O(P) parameters**
- Lacks language-specific parameterization

PER-LANGUAGE ENCODER/DECODER [Luong16, Firat16]



- **O(LP) parameters**
- Limited parameter sharing and use of attention difficult

Proposed Approach



LEGEND

- Trainable variables
- Computed values
- **M** Language embeddings size
- **W** Word embeddings size
- **P** Number of parameters
- text Example input

FEATURES

Scalable
Constant number of parameters - **O(MP)**

Simple & Multilingual
Can be applied to most existing NMT systems with minor changes.

Semi-Supervised
Can use monolingual data by learning to translate back-and-forth → 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.

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

PARAMETER GENERATOR

Generates model parameters at inference time, given some context.

The source and target language represent the context in which translation happens:

$$\theta^{(enc)}, \theta^{(dec)} = g(\text{SOURCE}, \text{TARGET}) = \text{P}$$

We learn **language embeddings**

$$\mathbf{l_s}, \mathbf{l_t} \in \mathbb{R}^{\mathbf{M}}$$

We also *decouple* the encoder and the decoder, thus getting closer to a potential *intelingua*:

$$\theta^{(enc)} = g^{(enc)}(\text{SOURCE})$$

$$\theta^{(dec)} = g^{(dec)}(\text{TARGET})$$

We choose to make g linear for simplicity and interpretability

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

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

For each language, the parameters are defined as a linear combination of the M columns of a weight matrix $\mathbf{W}$*, which makes for better interpretability.*

OBSERVATIONS

- The parameters often have some **"natural grouping"** (e.g., first layer weights).
- Language embeddings represent all language-specific information and may need to be large.
- Only a small part of this information is relevant for each "group".

CONTROLLED SHARING

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

$$\theta_j^{(enc)} \triangleq \mathbf{W}_j^{(enc)}\mathbf{P}_j^{(enc)}\mathbf{l_s}$$

where:

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

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

and M' < M, and similarly for the decoder.

↑ M ⇔ ↑ Per-Language Information
↑ M' ⇔ ↑ Shared Information

Our contribution does not depend on the choice of g. It would be interesting to design models that can use side-information about the languages, that may be available.

PAIRWISE: g picks a different parameter set based on the language pair
UNIVERSAL: g picks the same parameters for all languages
PER-LANGUAGE: g picks different enc/dec parameters based on the languages

The proposed abstraction is a generalization over previous methods

Experiments

Baseline Model

- 2-layer bidirectional LSTM encoder
- 2-layer LSTM decoder
- 512 units per layer / word embedding size
- Per-language vocabulary
- 20,000 most frequent words — no BPE

Settings

- *Supervised:* Train using full parallel data
- *Low-Resource:* Limit the size of the parallel data
- *Zero-Shot:* No parallel data for some language pairs

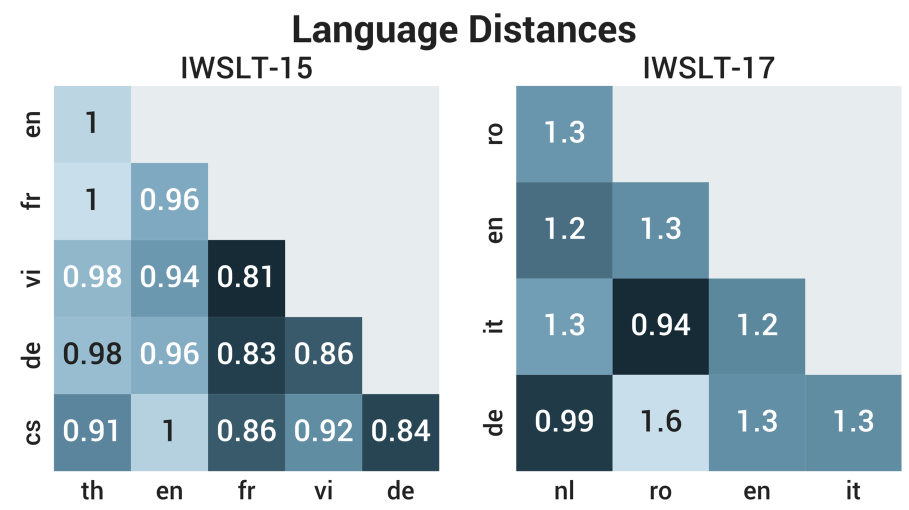
All experiments were run on a machine with a single Nvidia V100 GPU, and 24 GBs of system memory.

The longest experiment required ~10 hours.

Scala MT library and code to reproduce experiments

TensorFlow Scala used for our experiments

Language Distances



IWSLT-15

Pairwise models (PNMT); Google Multilingual (GML); Trained without auto-encoding (CPG*⁸); M=8 (CPG⁸)

		BLEU			
		PNMT	GML	CPG*⁸	CPG⁸
100% Parallel Data	En→Cs	14.89	15.92	16.88	**17.22**
	Cs→En	24.43	25.25	26.44	**27.37**
	En→De	25.99	25.92	26.41	**26.77** (> 25.87 [Ha16])
	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** (> 28.07 [Huang18])
	Vi→En	24.03	23.19	25.91	**26.38**
	Mean	26.26	24.12	27.30	**27.80**
10% Parallel Data	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**

~90,000-220,000 train / ~500-900 val / ~1,000 test

IWSLT-17

CPG⁸: M=8; CPG⁸C4: M=8, M'=4; CPG⁶⁴C8: M=64, M'=8

		BLEU				
		PNMT	GML	CPG⁸	CPG⁸C4	CPG⁶⁴C8
Supervised	De→En	21.78	21.25	**22.56**	20.78	21.50
	De→It	13.16	13.84	**14.73**	14.34	14.34
	De→Ro	10.85	11.95	12.24	12.37	11.32
	En→De	**19.75**	17.06	19.41	19.04	17.46
	En→It	27.70	25.74	27.57	27.11	27.26
	En→Nl	24.41	22.46	24.47	**25.15**	24.48
	En→Ro	19.23	18.60	20.83	**20.96**	20.20
	It→De	14.39	12.76	14.61	**15.06**	14.18
	It→En	29.84	27.96	**30.62**	30.10	29.56
	It→Nl	16.74	16.27	17.99	**18.11**	17.71
	Nl→En	26.30	24.78	26.31	26.17	**26.33**
	Nl→It	16.03	16.10	16.81	**17.50**	16.89
	Nl→Ro	12.84	12.48	14.01	**14.44**	12.38
	Ro→De	12.75	12.21	13.58	**13.66**	12.96
	Ro→En	24.33	22.88	23.83	23.88	**24.65**
	Ro→Nl	13.70	14.11	15.34	**15.51**	15.29
	Mean	18.99	18.15	19.68	**19.75**	19.16
Zero-Shot	De→Nl	12.75	12.50	12.74	**12.80**	12.67
	It→Ro	9.97	9.57	10.57	10.17	**10.69**
	Nl→De	11.32	10.47	11.52	11.20	**11.63**
	Ro→It	11.69	10.82	11.51	11.40	**11.78**
	Mean	11.43	10.84	11.51	11.39	**11.69**

~220,000 train / ~900 val / ~1,100 test