Neural Machine Translation Internals-Part2

By Mohit Kumar

Advance NMT: How to get better performance?

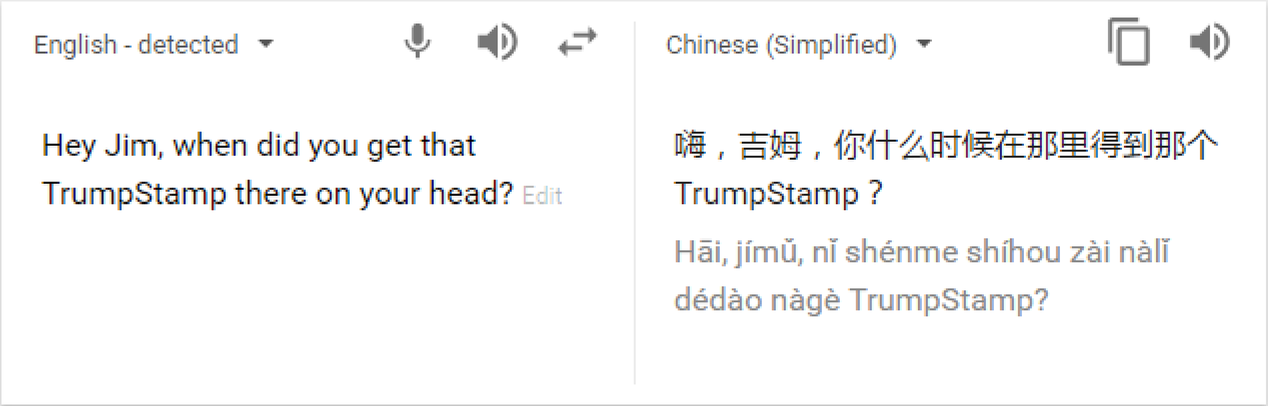
- The **attention** aspect
  - Global and local information

- The **vocabulary** aspect
  - Ensemble decoding
  - **Rare word translations**

- The **data** aspect
  - Monolingual Data
  - Multilingual Data

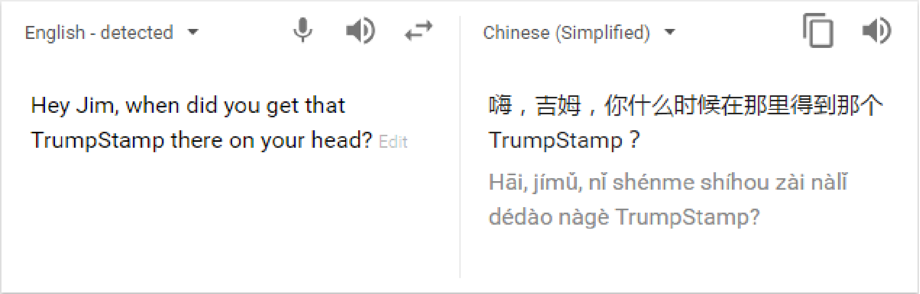
Rare Word Translations

- Inability to translate Out-Of-Vocabulary(OOV) words corr



Rare Word Translations

- Inability to translate Out-Of-Vocabulary(OOV) words correctly
- Types of OOVs
  1. Unseen words in training co
  2. New words



- Solutions
  1. Training on a large vocabulary
  2. Learn on some linguistic features

Solution #1: Large Vocabulary

- Training on a corpus with a very large vocabulary
  - Decrease the number of unseen words
  - Potentially able to solve misspelling
- Computationally expensive
  - Larger corpus -> More training time
  - Larger vocabulary -> More candidates for decoding

Fast large vocabulary NMT system

- **Training**
  - Segment training data in subsets
  - Each subset has exactly n distinct wc
  - Train on one subset at a time
  - Could be GPU parallelized

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

|V| = 5

Jean, Sébastien et al. "[On Using Very Large Target Vocabulary for Neural Machine Translation](). " *ACL* (2015).

Fast large vocabulary NMT system

- **Training**
  - Segment training data in subsets
  - Each subset has exactly n distinct wo
  - Train on one subset at a time
  - Could be GPU parallelized

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

|V| = 5

Fast large vocabulary NMT system

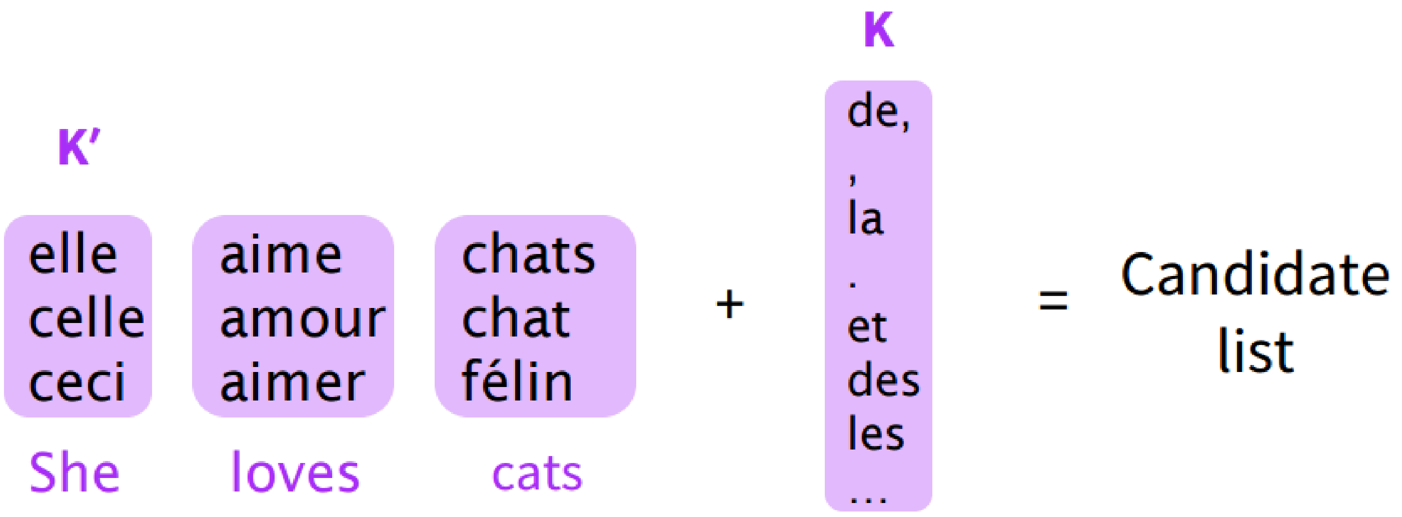
- **Training**
  - Segment training data in subsets
  - Each subset has exactly n distinct wc
  - Train on one subset at a time
  - Could be GPU parallelized

she loves cats
he likes dogs
cats have tails
dogs have tails
dogs chase cats
she loves dogs
cats hate dogs

|V| = 5

Fast large vocabulary NMT system

- **Testing**
  - Choose K most frequent words
  - Choose K' frequent candidates per word from all subsets
  - Use the candidate list to generate translations



Performances

- **Data coverage**

Vocabulary coverage is high

with this approximating algorithm

Size of most frequent words

K	English-French		English-German	
	Train	Test	Train	Test
15k	93.5	90.8	88.5	83.8
30k	96.0	94.6	91.8	87.9
50k	97.3	96.3	93.7	90.4
500k	99.5	99.3	98.4	96.1
All	100.0	99.6	100.0	97.3

Table 1: Data coverage (in %) on target-side corpora for different vocabulary sizes. "All" refers to all the tokens in the training set.

Jean, Sébastien et al. "On Using Very Large Target Vocabulary for Neural Machine Translation." *ACL* (2015).

Performances

- **Decoding speed boosting**

The speed of NMT trained on vocabulary with candidate list **comparable** to the baseline

	CPU*	GPU°
RNNsearch	0.09 s	0.02 s
RNNsearch-LV	0.80 s	0.25 s
RNNsearch-LV +Candidate list	0.12 s	0.05 s

Trained on 30K

Trained on full vocab

Trained on full vocab with candidate list with K = 30K

Table 3: The average per-word decoding time. Decoding here does not include parameter loading and unknown word replacement. The baseline uses 30k words. The candidate list is built with $K = 30k$ and $K' = 10$. ($\star$) i7-4820K (single thread), ($\circ$) GTX TITAN Black

Performances

- **BLEU score boosting**

NMT with candidate list has highe
BLEU score than baseline system
while the speed is much faster

	RNNsearch	RNNsearch-LV	Google	Phrase-based SMT	
Basic NMT	29.97 (26.58)	32.68 (28.76)	30.6*		
+Candidate List	–	33.36 (29.32)	–		
+UNK Replace	33.08 (29.08)	34.11 (29.98)	33.1°	33.3*	37.03•
+Reshuffle (τ=50k)	–	34.60 (30.53)	–		
+Ensemble	–	37.19 (31.98)	37.5°		

(a) English→French

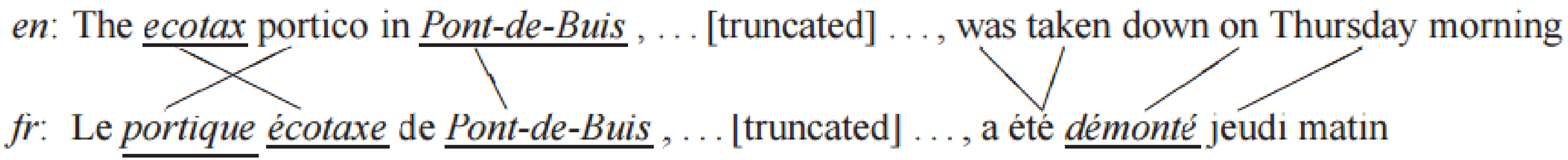
	RNNsearch	RNNsearch-LV	Phrase-based SMT
Basic NMT	16.46 (17.13)	16.95 (17.85)	
+Candidate List	–	17.46 (18.00)	
+UNK Replace	18.97 (19.16)	18.89 (19.03)	20.67⋄
+Reshuffle	–	19.40 (19.37)	
+Ensemble	–	21.59 (21.06)	

(b) English→German

BLEU score obtained on different models

Solution #2: Linguistic Features

- Detect and solve unseen words problem based on some linguistic features, e.g. morphemes
- **Copy Mechanism:**
  - Directly copy the unseen source word into target sentence



nn: Le <unk> <unk> de <unk> a <unk>, …[truncated]…, a ete pris le jeudi matin

nn(with copy): Le <u>ecotax portico</u> in <u>Pont-de-Buid</u>, …, a ete pris le jeudi matin

uong, Minh-Thang, et al. "**<u>Addressing the rare word problem in neural machine translation</u>**." *arXiv preprint arXiv:1410.8206* (201

Solution #2: Linguistic Features

- Detect and solve unseen words problem based on some linguistic features, e.g. morphemes
- **Sub-word unit translation**:
  - Segment each word into the smallest part, translate and re-compound them

Loanwords(differ in alphabets)

English: Claustrophobia
German: Klaustrophobie
Russian: Клаустрофобия

Morphologically complex words(compounds)

English: Solar system(solar+system)
German: Sonnensystem(sonnen+system)
Hungarian: Naprendszer(nep+rendszer)

Performance

- Translation example with segmentation

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
WDict	Forschungsinstitute
C2-50k	Fo|rs|ch|un|gs|in|st|it|ut|io|ne|n
BPE-60k	Gesundheits|forsch|ungsinstitu|ten
BPE-J90k	Gesundheits|forsch|ungsin|stitute
source	asinine situation
reference	dumme Situation
WDict	asinine situation → UNK → asinine
C2-50k	as|in|in|e situation → As|in|en|si|tu|at|io|n
BPE-60k	as|in|ine situation → A|in|line-|Situation
BPE-J90K	as|in|ine situation → As|in|in-|Situation

Word level model with a back-off dictionary(baseline)

Character bigram model

BPE(Byte Pair Encoding) model

Table 4: English→German translation example. "|" marks subword boundaries.

Performance

- **BLEU Score Boosting**

name	segmentation	shortlist	vocabulary source	vocabulary target	BLEU single	BLEU ens-8	CHRF3 single	CHRF3 ens-8	unigram F_1 (%) all	unigram F_1 (%) rare	unigram F_1 (%) OOV
syntax-based (Sennrich and Haddow, 2015)					24.4	-	55.3	-	59.1	46.0	37.7
WUnk	-	-	300 000	500 000	20.6	22.8	47.2	48.9	56.7	20.4	0.0
WDict	-	-	300 000	500 000	22.0	24.2	50.5	52.4	58.1	36.8	**36.8**
C2-50k	char-bigram	50 000	60 000	60 000	**22.8**	**25.3**	51.9	53.5	58.4	40.5	30.9
BPE-60k	BPE	-	60 000	60 000	21.5	24.5	**52.0**	53.9	58.4	40.9	29.3
BPE-J90k	BPE (joint)	-	90 000	90 000	**22.8**	24.7	51.7	**54.1**	**58.5**	**41.8**	33.6

Table 2: English→German translation performance (BLEU, CHRF3 and unigram F_1) on newstest2015.

Systems with sub-word unit translation perform slightly better than baseline system

- Portion of words with explicit sub-word units is only 3-4%
- BPE algorithm may not be ideal

Advance NMT: How to get better performance?

- The **attention** aspect
  - Global and local information

- The **vocabulary** aspect
  - Ensemble decoding
  - Rare word translation

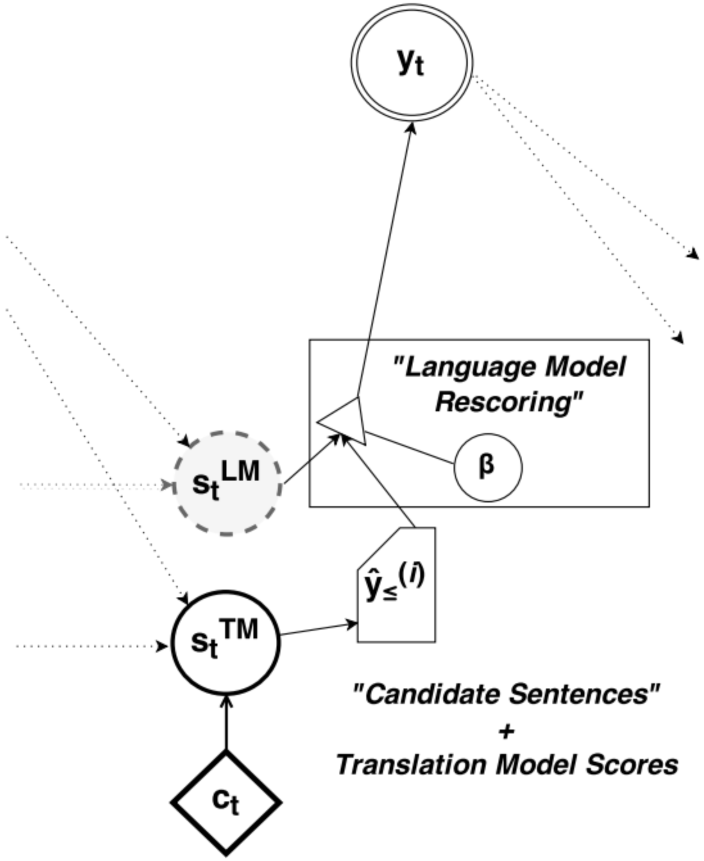
- The **data** aspect
  - **Monolingual data**
  - Multilingual data

Utilize Monolingual Data

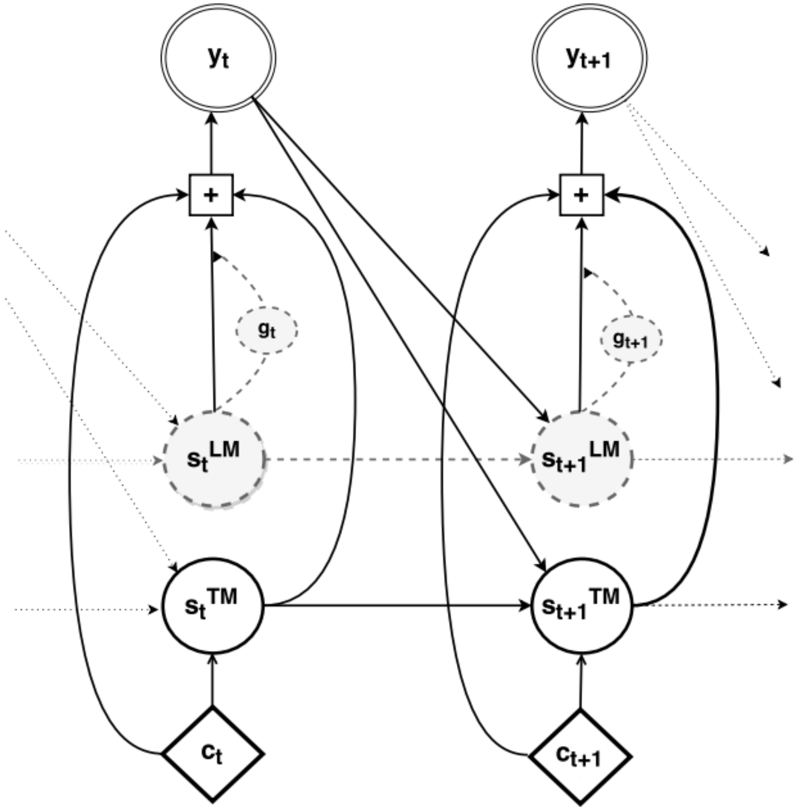
- Higher prior probability (fluency) of target language
- Language model needs to be integrated
  - Shallow Fusion
  - Deep Fusion

Gulcehre, Caglar, et al. "On using monolingual corpora in neural machine translation." *arXiv preprint arXiv:1503.03535* (2015).

Training On Monolingual Target Language



(a) Shallow Fusion (Sec. 4.1)



(b) Deep Fusion (Sec. 4.2)

Training On Monolingual Target Language

- Could make up a parallel corpus
- No need to change the NN architectures
- **Pair** target language with source language

, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." *arXiv preprint arXiv:1511.06709* (2015

Pairing strategies

- **Dummy Source Sentences**

Pair monolingual sentences with a single-word dummy source side <null>.

Source sentence: <null>

Target sentence(in monolingual corpora):Des Teufels liebstes Möbelstück ist die lange Bank.

Pairing strategies

- **Synthetic Source Sentences**

Pair monolingual training instances with a synthetic source sentence (back-translation)

Source sentence(by Google translate): The devil's favorite piece of furniture is the long bench.

Target sentence(in monolingual corpora): Des Teufels liebstes Möbelstück ist die lange Bank.

Performances

- **BLEU Score Boosting**

name	training instances	BLEU newstest2014 single	BLEU newstest2014 ens-4	BLEU newstest2015 single	BLEU newstest2015 ens-4
syntax-based (Sennrich and Haddow, 2015)		22.6	-	24.4	-
Neural MT (Jean et al., 2015b)		-	-	22.4	-
parallel	37m (parallel)	19.9	20.4	22.8	23.6
+monolingual	49m (parallel) / 49m (monolingual)	20.4	21.4	23.2	24.6
+synthetic	44m (parallel) / 36m (synthetic)	**22.7**	**23.8**	**25.7**	**26.5**

Table 3: English→German translation performance (BLEU) on WMT training/test sets. Ens-4: ensemble of 4 models. Number of training instances varies due to differences in training time and speed.

Synthetic model outperforms the other two

- More fluency in target language
- More “golden” parallel sentences

Advance NMT: How to get better performance?

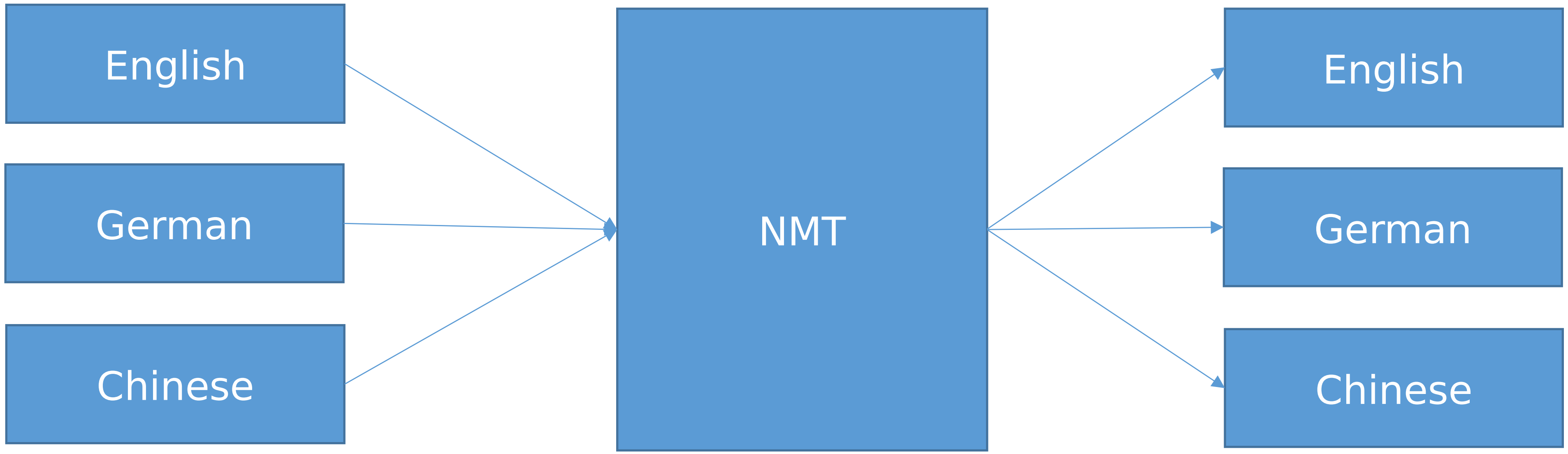
- The **attention** aspect
  - Global and local information

- The **vocabulary** aspect
  - Ensemble decoding
  - Rare word translation

- The **data** aspect
  - Monolingual data
  - **Multilingual data**

Utilize Multilingual Data

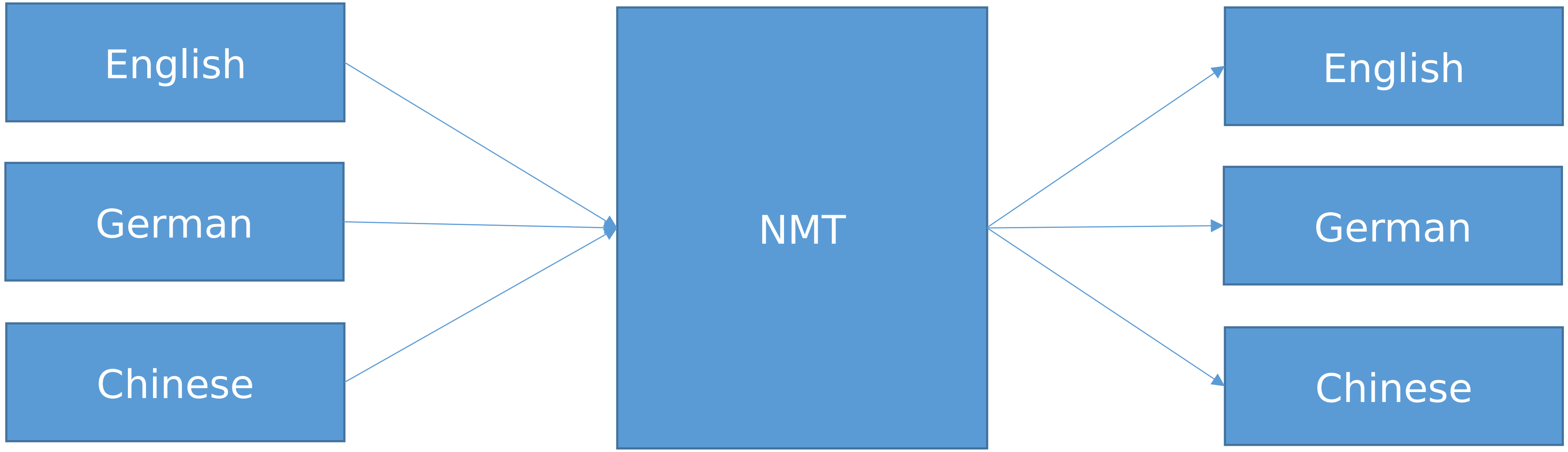
- Using corpus with multiple languages
- Training on each languages pair in the corpus



unghyun Cho, and Yoshua Bengio. "Multi-way, multilingual neural machine translation with a shared attention mechanism." *arXiv preprint arXiv:1601.01073* (20

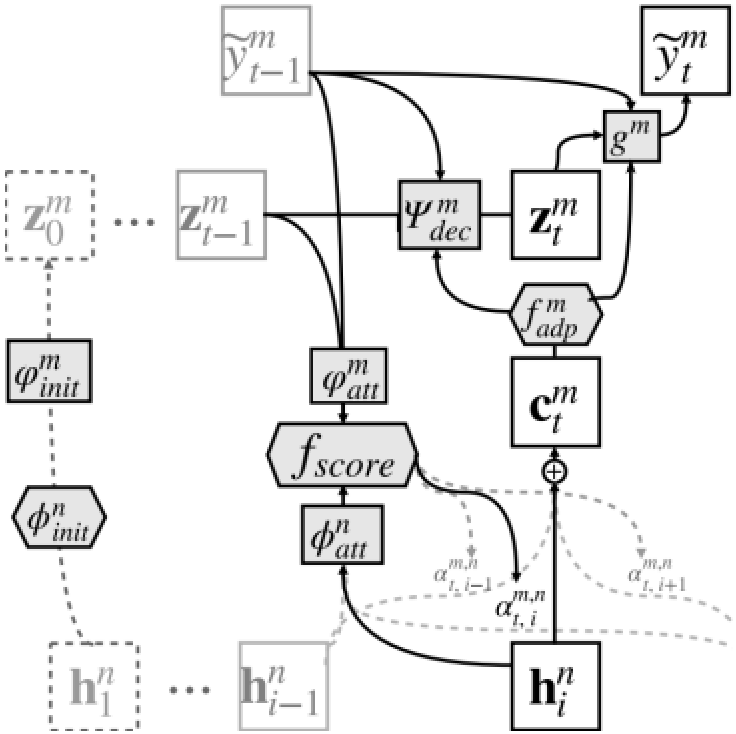
Utilize Multilingual Data

- Encode the source language into a continuous representation
- Decode from the representation into any language



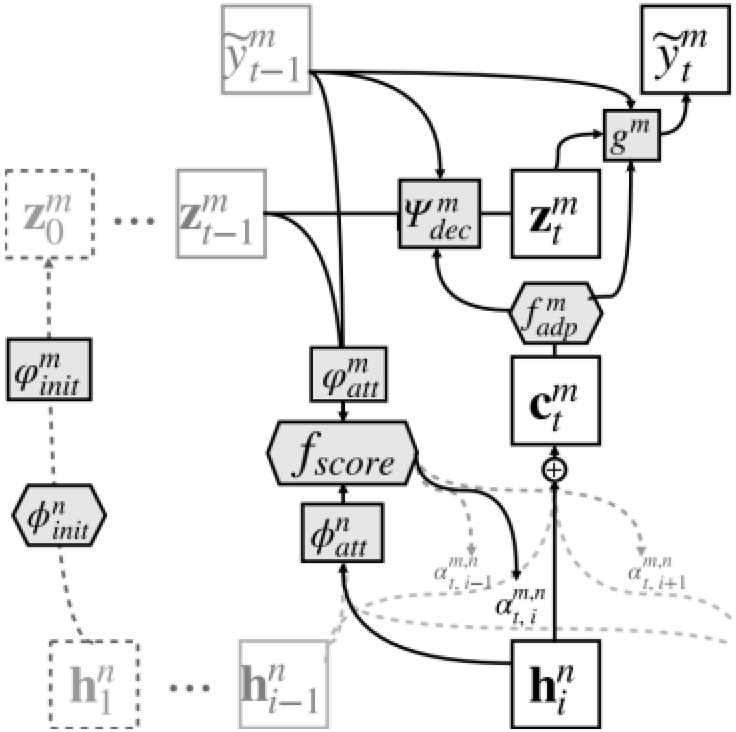
Multi-Way, Multilingual Model

- One encoder and one decoder per source language
- Shared single attention mechanism, with attention s



The nth encoder, mth decoder at time step t in the Multi-way multilingual NMT model

Multi-Way, Multilingual Model



At time t, ith word. For encoder n and decoder m:

Decoder hidden state vector z_{t-1}^{m}

Previously decoded symbol y_{t-1}^{m}

Time-dependent context vector c_t^{m}

Embedding matrix E_y

Recurrent activation function ψ_{dec}

Decoder hidden state $z_t = \psi_{dec}(z_{t-1}^{m},\ E_y^{m}[y_{t-1}^{m}], f_{adp}^{m}(c_t^{m}))$

Performances

- Low-Resource Translation

Multi model outperforms all
the single-pair NMT system

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

Table 2: BLEU scores where the target pair's parallel corpus is constrained to be 5%, 10%, 20% and 40% of the original size.

Performance

- Large-Scale Translation

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98

Multi model either outperforms or is comparable the single pair models.

In translating into English task, it always performs better.

Outline

- **Background**
  - Machine Translation: task overview
- **Basic NMT**
  - An encoder-decoder architecture
- **Advanced NMT**
  - The attention aspect
  - The vocabulary aspect
  - The data aspect
- **State-of-the-art NMT System**
  - **GNMT (Google NMT System)**

State-of-the-art NMT: GNMT

GNMT: Google's NMT System

- **Architecture**
  - Paralleled deep encoder & decoder
- **Mechanisms & Techniques**
  - Residual Connections
  - Model Parallelism
  - Beam Search
  - Segmentation
- **Multi-lingual**

Wu, Yonghui, et al. "
Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Google's Neural Machine Translation System

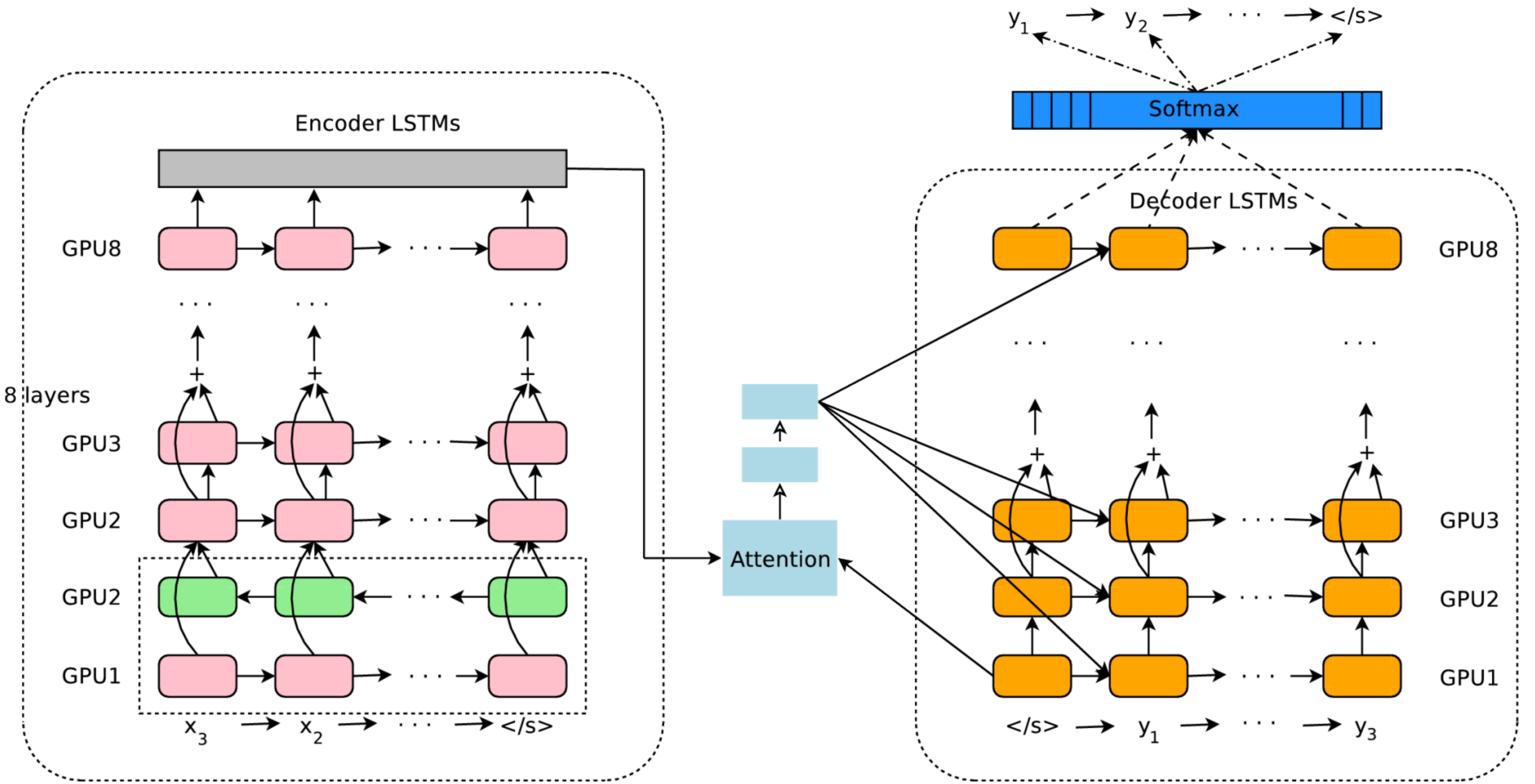
- LSTM network with 8 encoder and 8 decoder
- Residual connections and attention connections
- Low-precision arithmetic during inference
- Sub-word units
- Beam search: length-normalization and coverage penalty

State-of-the-art NMT: GNMT

GNMT: Google's NMT System

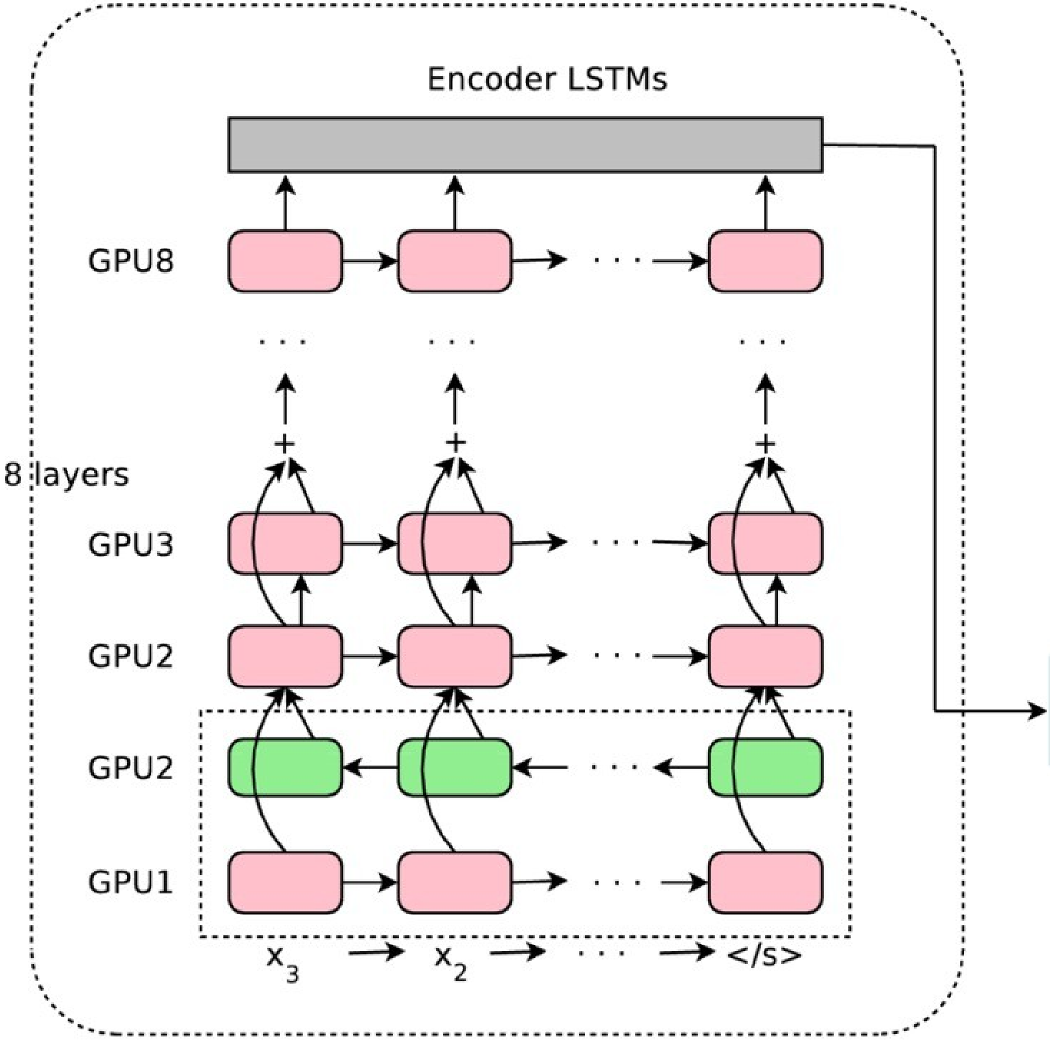
- **Architecture**
  - Paralleled deep encoder & decoder
- **Mechanisms & Techniques**
  - Residual Connections
  - Model Parallelism
  - Beam Search
  - Segmentation
- **Multi-lingual**

Model Architecture



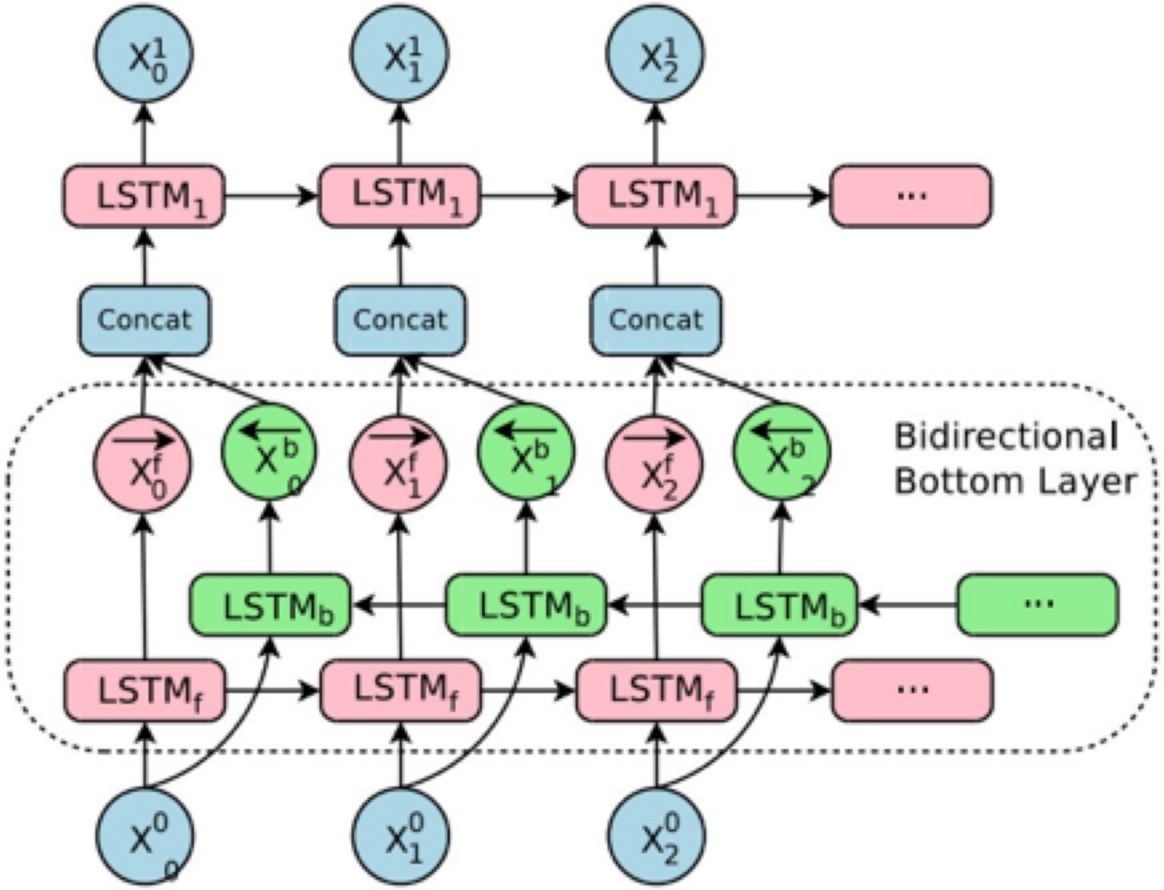
GNMT Encoder

- Model is partitioned 8-ways and is placed on 8 different GPUs
- The bottom bi-directional encoder layers compute in parallel first
- Once both finish, the uni-directional encoder layers can start computing, each on a separate GPU.



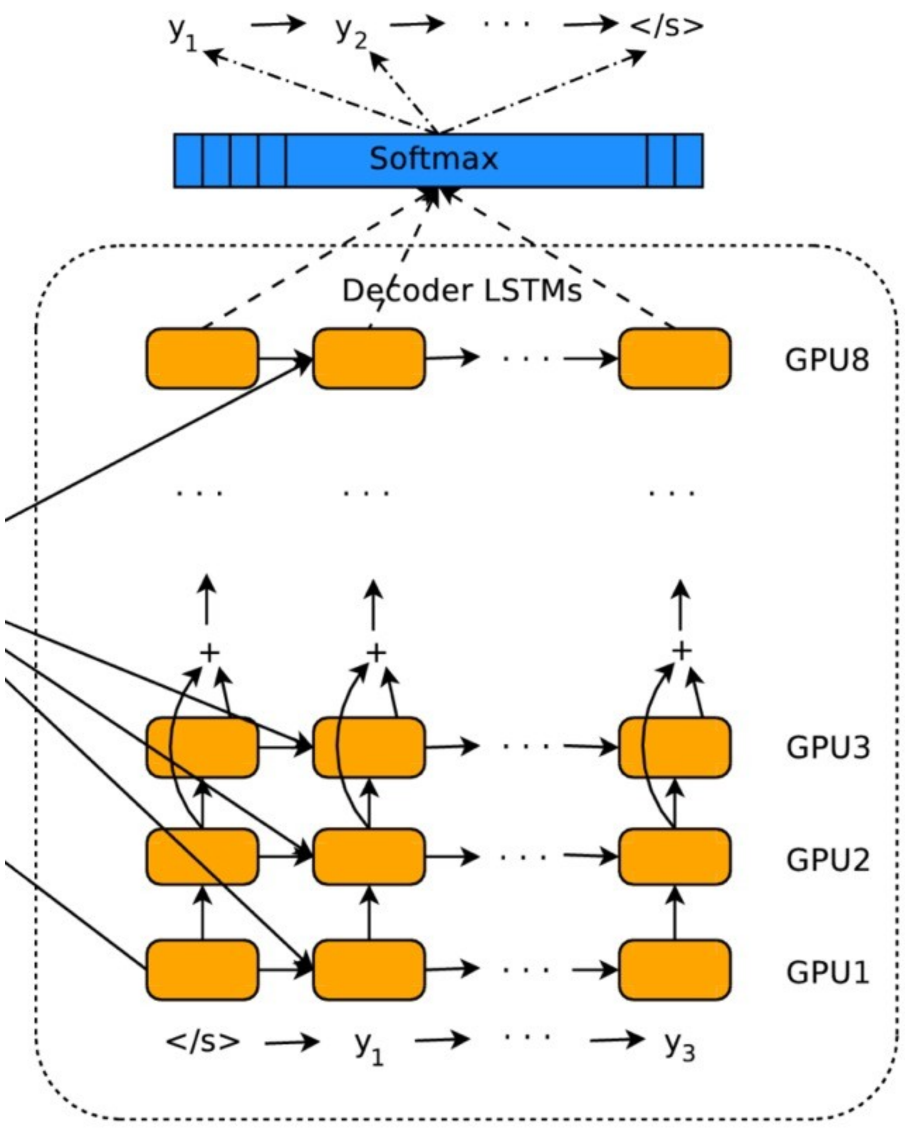
GNMT Bi-directional Encoder: First Layer

- **Bi-directional encoder**
  - The bottom encoder layer is bi-directional
  - Two-side context information
- **Efficiency**:
  - Bi-directional are only used for the bottom encoder layer
  - Allow for maximum possible parallelization during computation



GNMT Decoder

- The bottom decoder layer output only for obtaining recurrent attention context, which is sent directly to all the remaining decoder layers.
- The softmax layer is also partitioned and placed on multiple GPUs.



Training

- **Recall standard training:** maximum-likelihood

$$L(\theta) = \frac{1}{N}\sum_{n} \log p_{\theta}(\vec{\boldsymbol{y}}^{(n)} | \vec{\boldsymbol{x}}^{(n)})$$

- Problems:
  - NOT reflect the task reward function by the BLEU score in translation
  - NOT encourage a ranking among incorrect output sequences

- **Recall standard training:** maximum-likelihood
- Problems:
  - NOT reflect the task reward function by the BLEU score in translation
  - NOT encourage a ranking among incorrect output sequences

Training

- **GMNT objective**: expected rewards

$$\mathcal{O}_{\mathrm{RL}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{Y \in \mathcal{Y}} P_\theta(Y \mid X^{(i)})\, r(Y, Y^{*(i)}).$$

- Reward r(Y, Y ∗(i))
  - per-sentence score
  - compute an expectation over all of the output sentences Y
- Glue Score:
  - record all sub-sequences of 1, 2, 3 or 4 tokens in output and target sequence (n-grams) and then compute the recall and precision

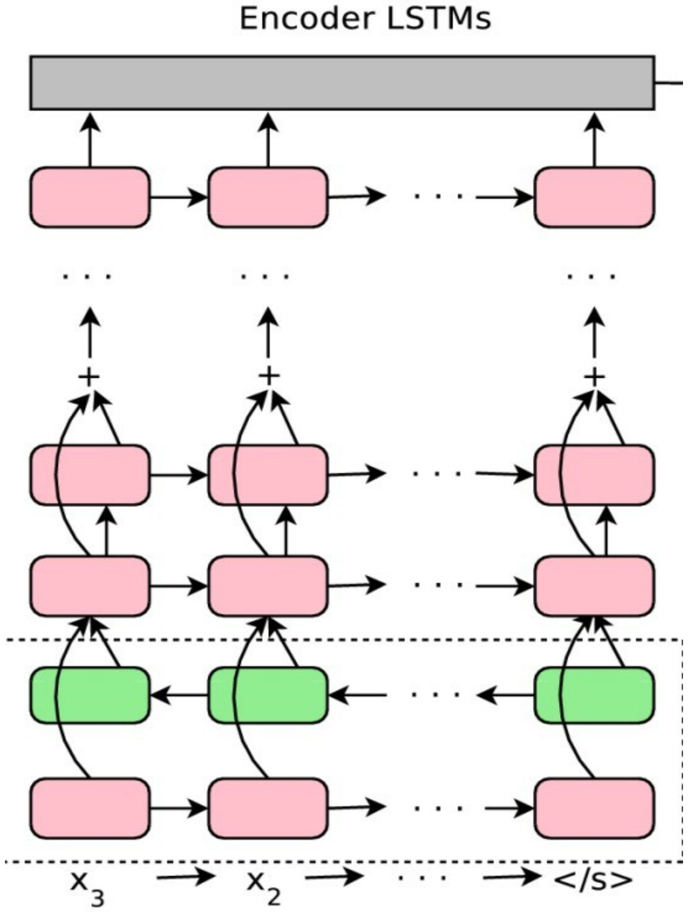
State-of-the-art NMT: GNMT

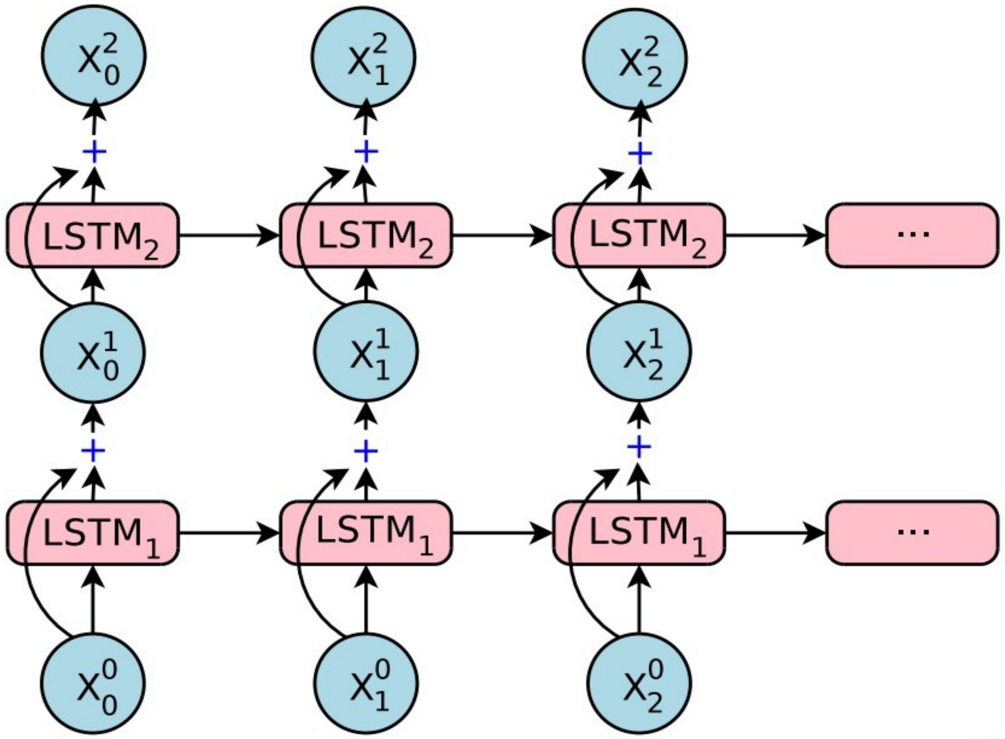
GNMT: Google’s NMT System

- **Architecture**
  - Paralleled deep encoder & decoder
- **Mechanisms & Techniques**
  - Residual Connections
  - Model Parallelism
  - Beam Search
  - Segmentation
- **Multi-lingual**

Residual Connections

- Residual connections start from the layer third
- Residual connections greatly improve the gradient flow

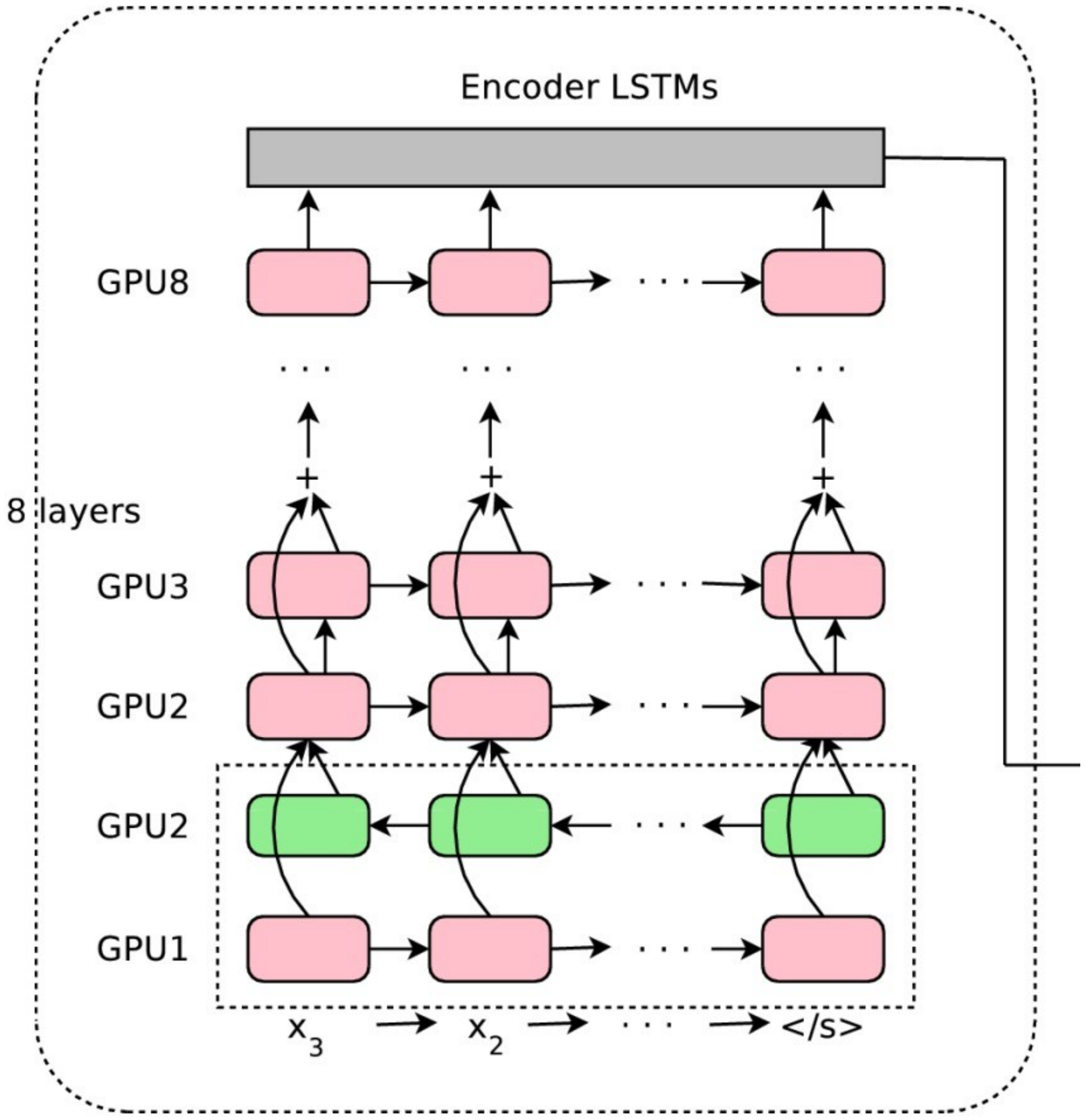




Model Parallelism

- Both model parallelism and data parallelism to speed up training

- Data parallelism: Train n model replicas concurrently using a Downpour SGD algorithm
- Model parallelism: Partitioned along the depth dimension and are placed on multiple GPUs



Decode with Beam Search

- **Coverage penalty**
  - Empirically-better Score function to rank candidate

$$s(Y, X) = \log(P(Y|X))/lp(Y) + cp(X;Y)$$

$$lp(Y) = \frac{(5+|Y|)^\alpha}{(5+1)^\alpha}$$

$$cp(X;Y) = \beta * \sum_{i=1}^{|X|} \log(\min(\sum_{j=1}^{|Y|} p_{i,j}, 1.0)),$$

- **Length normalization**
  - Aim to account for the fact that we have to compare hypotheses of different length.

where p i,j is the attention probability of the j-th target word yj on the i-th source $\sum_{i=0}^{|X|} p_{i,j}$ By construction is equal to 1.

Segmentation Approaches

- Wordpiece Model

Word: Jet makers feud over seat width with big orders at stake

wordpieces: _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

- Mixed Word/Character Model

    convert OOV words into the sequence of its constituent characters

Quantized Inference

- Using 8-bit or 16 bit integer representation to do reduced precision arithmetic
- When it is decoded on TPU, certain operations, such as embedding lookup and attention module, remain on the CPU, and all other quantized operations are off-loaded to the TPU.

	BLEU	Log Perplexity	Decoding time (s)
CPU	31.20	1.4553	1322
GPU	31.20	1.4553	3028
TPU	31.21	1.4626	384

Performances

Table 4: Single model results on WMT En→Fr (newstest2014)

Model	BLEU	Decoding time per sentence (s)
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8K	38.27	0.1919
WPM-16K	37.60	0.1874
WPM-32K	38.95	0.1146
Mixed Word/Character	38.39	0.2774
PBMT [15]	37.0	
LSTM (6 layers) [30]	31.5	
LSTM (6 layers + PosUnk) [30]	33.1	
Deep-Att [43]	37.7	
Deep-Att + PosUnk [43]	39.2	

"WPM-32K", a wordpiece model with a shared source and target vocabulary of 32K wordpieces, performs well on this dataset and achieves the best quality as well as the fastest inference speed.

Performances: Production Data

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.550	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

PBMT: Translation by phrase-based statistical translation system used by Google,
GNMT: Translation by our GNMT system
Human: Translation by humans fluent in both languages.

State-of-the-art NMT: GNMT

GNMT: Google’s NMT System

- **Architecture**
  - Paralleled deep encoder & decoder
- **Mechanisms & Techniques**
  - Residual Connections
  - Model Parallelism
  - Beam Search
  - Segmentation
- **Multi-lingual**

Multiple Language

Multiple source languages and multiple target languages within a single model

- Simplicity
- Low-resource language improvements
- Zero-shot translation

BLEU scores on various data sets for single language pair and multilingual models

Model	Single	Multi	Diff
Prod English→Japanese	23.66	21.10	-2.56
Prod English→Korean	19.75	18.41	-1.34
Prod Japanese→English	23.41	21.62	-1.79
Prod Korean→English	25.42	22.87	-2.55
Prod English→Spanish	34.50	34.25	-0.25
Prod English→Portuguese	38.40	37.35	-1.05
Prod Portuguese→English	44.40	42.53	-1.87
Prod Spanish→English	38.00	36.04	-1.96
Prod English→German	26.43	23.15	-3.28
Prod English→French	35.37	34.00	-1.37
Prod German→English	31.77	31.17	-0.60
Prod French→English	36.47	34.40	-2.07

Johnson, Melvin, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." arXiv preprint arXiv:1611.04558 (2016).

Mixing Language

- What happens when languages are mixed on the source or target side?

- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

Case Study

The Google Translate mobile and web apps are now using GNMT for 100% of machine translations from Chinese to English

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Summary: What have we learned about NMT?

- **Machine Translation :** What is the task and how to evaluate?
- **Basic NMT:** An encoder-decoder architecture
- **Advanced NMT**
  - ***Attention***: attention mechanism for NMT
  - ***Vocab***: Rare word translation
  - ***Data***: Utilize monolingual/multilingual data
- **State-of-art NMT System**
  - GNMT (Google NMT System)

Reading List

Essential:

- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
- Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
- Wiseman, Sam, and Alexander M. Rush. "Sequence-to-sequence learning as beam-search optimization." arXiv preprint arXiv:1606.02960 (2016).
- Luong, Minh-Thang, et al. "Addressing the rare word problem in neural machine translation." arXiv preprint arXiv:1410.8206 (2014).
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).
- Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv preprint arXiv:1609.08144 (2016).

Reading List

Optional:

- Ling, Wang, et al. "Character-based neural machine translation." arXiv preprint arXiv:1511.04586 (2015).
- Johnson, Melvin, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." arXiv preprint arXiv:1611.04558 (2016).
- Gu,Ã, Jiatao, et al. Learning to translate in real-time with neural machine translation, arXiv preprint arXiv:1610.00388(2016).

Other References

- Neural Machine Translation - Tutorial ACL 2016
- Cho, Kyunghyun, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation *arXiv preprint arXiv:1406.1078* (2014).
- Cho, Kyunghyun, et al. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259* (2014).
- Schuster, Mike and Paliwal, Kuldip K. Bidirectional recurrent neural networks. IEEE Transactionson Signal Processing, 45(11):2673–2681, 1997.
- Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- Kalchbrenner, Nal, and Phil Blunsom. "Recurrent convolutional neural networks for discourse compositionality." *arXiv preprint arXiv:1306.3584* (2013).

Thanks !