Recent Advances and the Future of Neural Machine Translation

Orhan Firat[1] Kyunghyun Cho[2]

[1]Middle East Technical University
[2]New York University

Machine Translation Marathon 2016

Before we start!

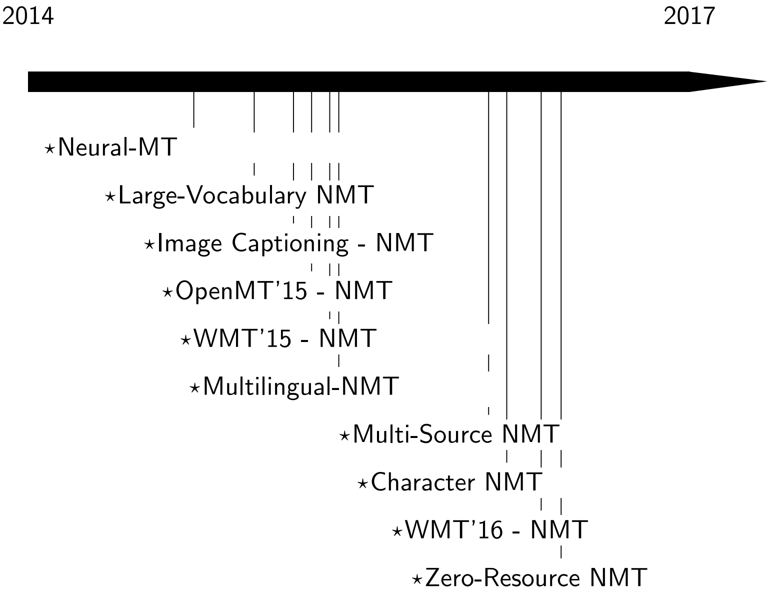
The Fog of Progress[1]

and

Artificial *General* Intelligence

[1] Hinton video-lectures,https://www.youtube.com/watch?v=ZuvRXGX8cY8

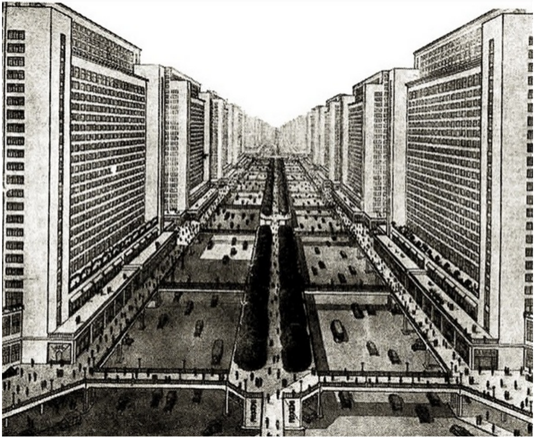
What is going on?



Warren Weaver- "Translation", 1949

Tall towers analogy:

- Do not shout from tower to tower,
- Go down to the common basement of all towers: *interlingua*



Neural Machine Translation - Encoder Decoder

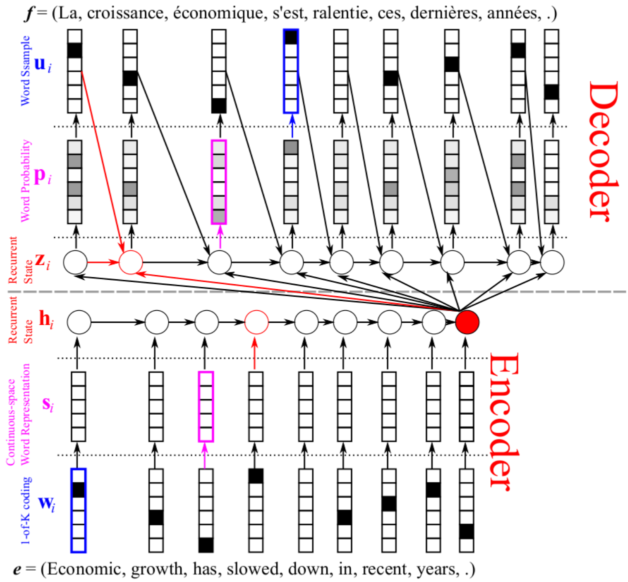
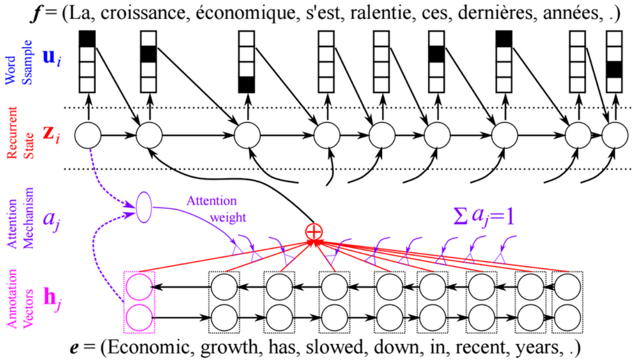


figure credit, Kyunghyun Cho

Encoder-Decoder Architecture with Attention

Bahdanau et al.2015



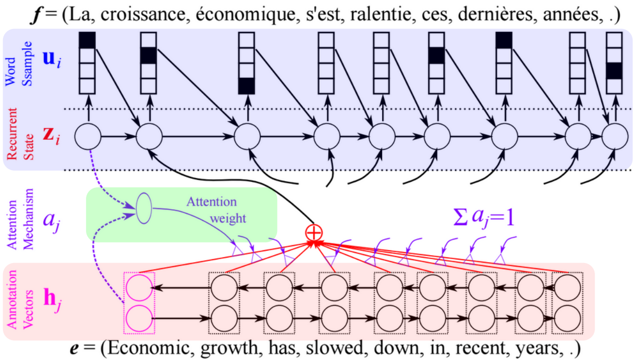
At each timestep in the decoder:

1. Computes a *relevance* score of each annotation
2. Use the weighted sum of the annotations as a *context*

figure credit, Kyunghyun Cho

Encoder-Decoder Architecture with Attention

Bahdanau et al.'15



At each timestep in the decoder:

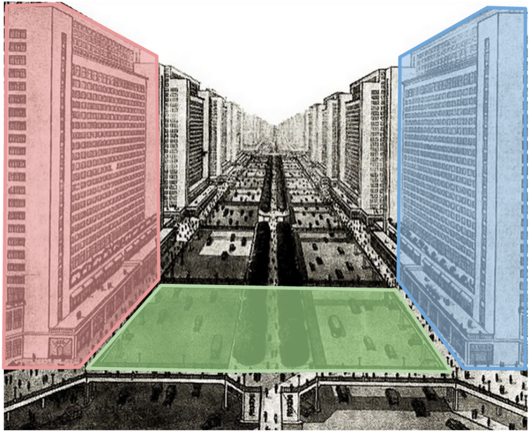
1. Computes a *relevance* score of each annotation
2. Use the weighted sum of the annotations as a *context*

figure credit, Kyunghyun Cho

Warren Weaver- "Translation", 1949

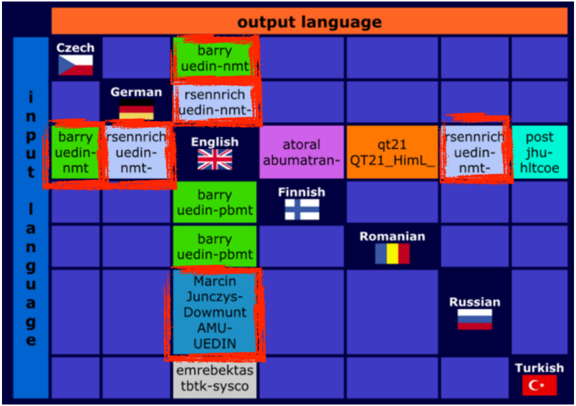
Tall towers analogy:

- Red Tower : source language
- Blue Tower: target language
- Green Car : alignment function



Attention-based NMT at work - WMT'16

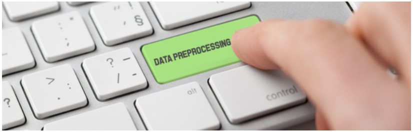
Seems to be working!



Most of the top-rankers used NMT

Neural Machine Translation with Finer Tokens

Let's make a poll on pre-processing!



- Why do we use word-level modelling?
  - Words are basic unit of meaning?!
  - Inherent fear of sparsity!
  - Finer granularities → longer sequences

Why can't an NMT system directly learn from the characters?

[1] www.dataentryoutsourced.com

Neural Machine Translation with Finer Tokens

Issues with tokenization and segmentation

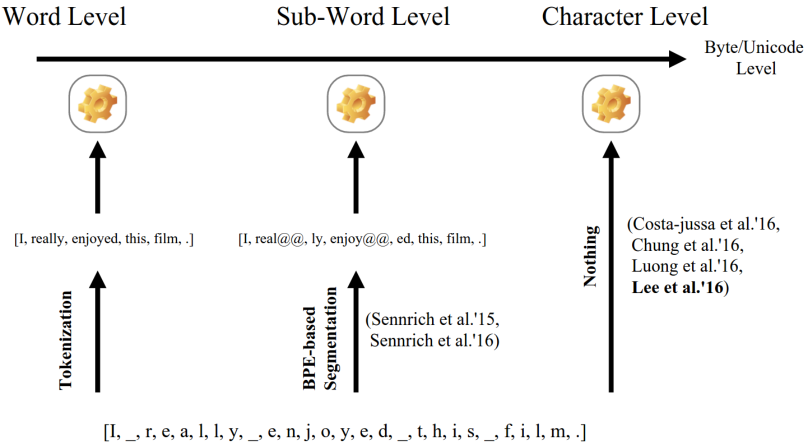
- Ineffective way of handling morphological variants: 'run', 'runs', 'running' and 'runner'
- How are we doing with compound words?

Issues with treating each and every token separately

- Fill the vocabulary with similar words
- Vocabulary size grows linearly w.r.t. the corpus size
- Rare words, numbers and misspelled words: 9/11 is a huge contextual information
- Lose the learning signal of words marked as <UNK>

slide credit, Junyoung Chung

Granularity in Input and Output Spaces (finer tokens)



Character-Level Decoder without Explicit Segmentation

Chung, Cho and Bengio, ACL'16

Model details,

- RNNSearch Model
- Source Side : sub-words (byte pair encoding, BPE)
- Target Side : either sub-words or characters
- Three types of decoders:
  1. Sub-word level *base* decoder
  2. Character level *base* decoder
  3. Character level *bi-scale* decoder

Bi-scale decoder:

- Faster/slower layers for modelling different levels of tokens
- Use soft gating units for differentiability

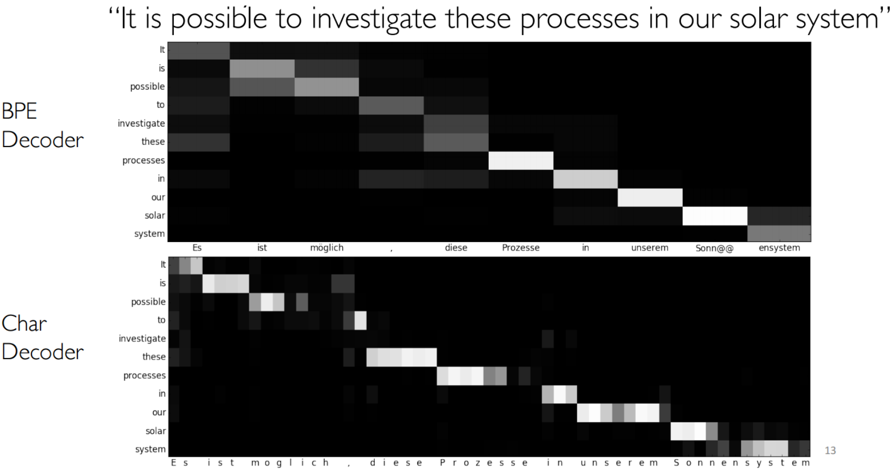
Character-Level Decoder without Explicit Segmentation

Chung, Cho and Bengio, ACL'16

		Src	Trgt	Depth	Attention h^1	Attention h^2	Model	Development Single	Development Ens	Test_1 Single	Test_1 Ens	Test_2 Single	Test_2 Ens
En-De	(a)	BPE	BPE	1	✓		Base	20.78	–	19.98	–	21.72	–
	(b)			2	✓	✓		$21.26^{21.45}_{20.62}$	23.49	$20.47^{20.88}_{19.30}$	23.10	$22.02^{22.21}_{21.35}$	24.83
	(c)		Char	2		✓	Base	$21.57^{21.88}_{20.88}$	23.14	$\mathbf{21.33}^{21.56}_{19.82}$	23.11	$\mathbf{23.45}^{23.91}_{21.72}$	25.24
	(d)			2	✓	✓		20.31	–	19.70	–	21.30	–
	(e)			2		✓	Bi-S	$21.29^{21.43}_{21.13}$	23.05	$21.25^{21.47}_{20.62}$	23.04	$23.06^{23.47}_{22.85}$	25.44
	(f)			2	✓	✓		20.78	–	20.19	–	22.26	–
	(g)			2	✓			20.08	–	19.39	–	20.94	–
	State-of-the-art Non-Neural Approach*							–		20.60(1)		24.00(2)	
En-Cs	(h)	BPE	BPE	2	✓	✓	Base	$16.12^{16.96}_{15.96}$	19.21	$17.16^{17.68}_{16.38}$	20.79	$14.63^{15.09}_{14.26}$	17.61
	(i)		Char	2		✓	Base	$17.68^{17.78}_{17.39}$	19.52	$19.25^{19.55}_{18.89}$	21.95	$\mathbf{16.98}^{17.17}_{16.81}$	18.92
	(j)			2		✓	Bi-S	$17.62^{17.93}_{17.43}$	19.83	$\mathbf{19.27}^{19.53}_{19.15}$	22.15	$16.86^{17.10}_{16.68}$	18.93
	State-of-the-art Non-Neural Approach*							–		21.00(3)		18.20(4)	
En-Ru	(k)	BPE	BPE	2	✓	✓	Base	$18.56^{18.70}_{18.26}$	21.17	$25.30^{25.40}_{24.95}$	29.26	$19.72^{20.29}_{19.02}$	22.96
	(l)		Char	2		✓	Base	$18.56^{18.87}_{18.39}$	20.53	$\mathbf{26.00}^{26.07}_{25.04}$	29.37	$\mathbf{21.10}^{21.24}_{20.14}$	23.51
	(m)			2		✓	Bi-S	$18.30^{18.54}_{17.88}$	20.53	$25.59^{25.76}_{24.57}$	29.26	$20.73^{21.02}_{19.97}$	23.75
	State-of-the-art Non-Neural Approach*							–		28.70(5)		24.30(6)	
En-Fi	(n)	BPE	BPE	2	✓	✓	Base	$9.61^{10.02}_{9.24}$	11.92	–	–	$8.97^{9.17}_{8.88}$	11.73
	(o)		Char	2		✓	Base	$11.19^{11.55}_{11.09}$	13.72	–	–	$\mathbf{10.93}^{11.56}_{10.11}$	13.48
	(p)			2		✓	Bi-S	$10.73^{11.04}_{10.40}$	13.39	–	–	$10.24^{10.63}_{9.71}$	13.32
	State-of-the-art Non-Neural Approach*							–		–		12.70(7)	

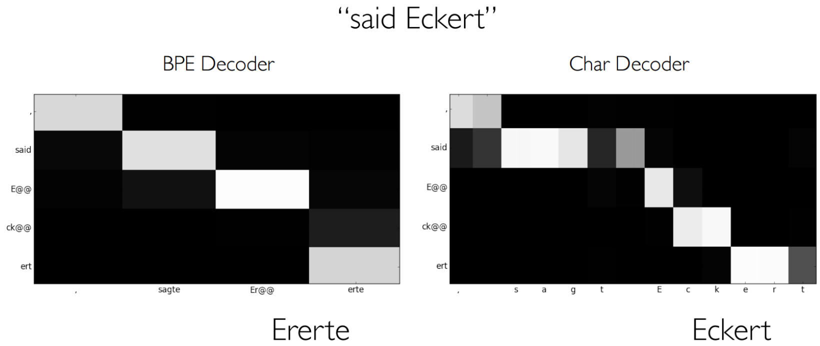
Character-Level Decoder without Explicit Segmentation

Chung, Cho and Bengio, ACL'16



Character-Level Decoder without Explicit Segmentation

Chung, Cho and Bengio, ACL'16



Neural Machine Translation with Finer Tokens

We are still concerned,

- Data sparsity problem will last!
  - but neural nets will less suffer from this issue (Bengio et al.,2003)

- Consequences of increased sequence length!
  - Capturing long-term dependencies
  - Will be harder to train (but wait we have GRU, LSTM and Attention)
  - Speed loss, 2-3 times slower

but ...

- No need to worry about segmentation,
- Open vocabularies, saves us giant matrices or tricks
- Naturally embeds multiple languages (**Lee et al.'16**)

slide credit, Junyoung Chung

Fully Character-Level Multilingual NMT

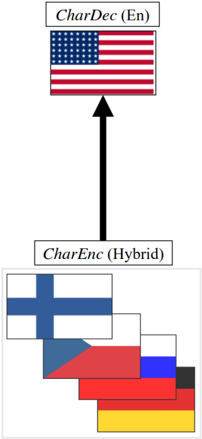
Jason Lee and Kyunghyun Cho, 2016 (in preparation)

Model details,

- RNNSearch model
- Source-Target character level
- CNN+RNN encoder
- *Bi-scale* decoder
- $\{Fi, De, Cs, Ru\} \rightarrow En$

Training,

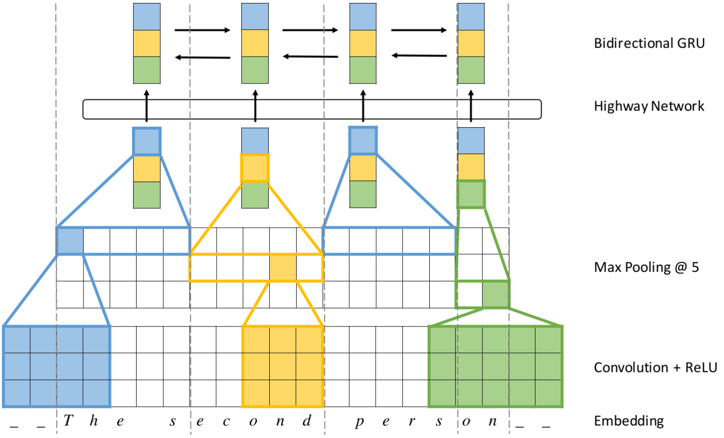
- Mix mini-batches
- Use bi-text only



Fully Character-Level Multilingual NMT

Jason Lee and Kyunghyun Cho, 2016 (in preparation)

Hybrid Character Encoder,



Fully Character-Level Multilingual NMT

Jason Lee and Kyunghyun Cho, 2016 (in preparation)

Preliminary Results, comparison with *BPE* → *Char*

	Model				Valid	Test-1	Test-2
	bpe2char		char2char				
	single	multi	single	multi			
De-En	✓				25.64	24.59	25.27
			✓		**26.03**	**25.80**	**25.77**
		✓			24.28	23.43	24.11
				✓	25.45	24.27	25.06
Cs-En	✓				22.83	23.51	22.46
			✓		22.85	23.38	22.03
		✓			22.76	23.46	21.86
				✓	**24.16**	**24.77**	**22.72**
Fi-En	✓				14.54	13.98	-
			✓		14.18	13.10	-
		✓			14.37	13.71	-
				✓	**15.85**	**15.80**	-
Ru-En	✓				21.68	26.21	**22.83**
			✓		21.75	**26.80**	22.73
		✓			20.91	24.59	21.93
				✓	**22.04**	25.64	22.68

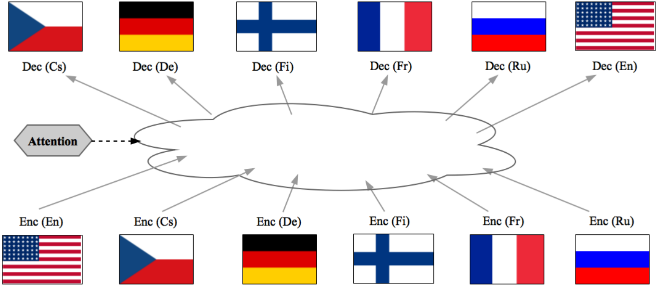
Spoiler - Big Time!

Spoiler - Big Time!

... a better single-pair translation system has never been **the** *goal of neural MT ...*

Kyunghyun Cho

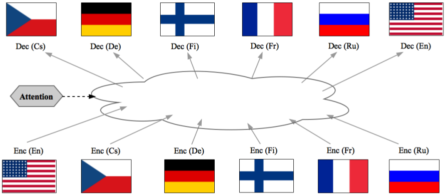
Multi-way, Multilingual Seq2Seq with Attention



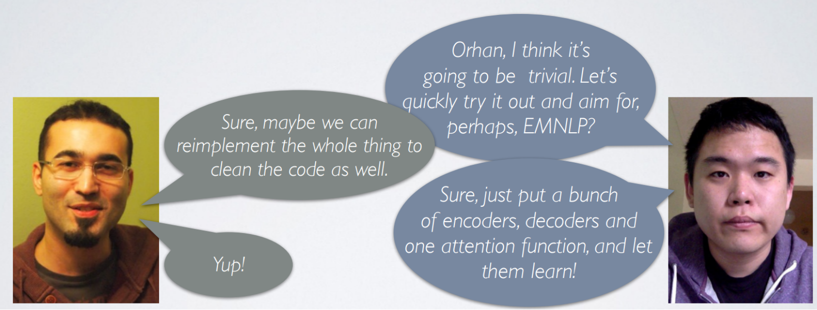
Potential Benefits

1. Positive language transfer across many language pairs/directions
   - Solution to low/zero-resource machine translation
2. # of parameters grows linearly w.r.t. the # of languages
   - as opposed to the quadratic explosion when training many single-pair models.
3. Multi-source translation without requiring any multi-way parallel text
   - inspired by but contrary to Zoph & Knight (2016)

- Super fun to work on!



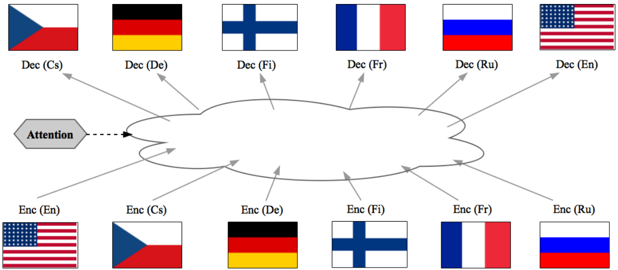
Young and Naive .. March, 2015 (right after WMT'15)



Young, Naive and Ambitious .. March, 2015

Multi-way, Multilingual Neural MT

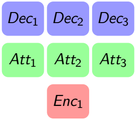
1. Many-to-many translation
2. One shared attention mechanism
3. No need for multi-way parallel text
4. Scalable in terms of # languages, # sentences
5. Extendible to multiple modalities



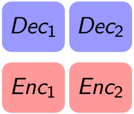
Multilingual (Multi-task) Neural Machine Translation

Recent work (chronologically),

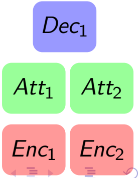
- One-to-many (Dong et al., 2015)
  - Each decoder has it's own attention mechanism.
  - Small scale experiments (Europarl).
  - No support for multiple source sentences.



- Without attention (Luong et al., 2015)
  - Focus on multi-task learning.
  - No attention, single vector space is shared.
  - Multilinguality is not considered in depth (En↔De).



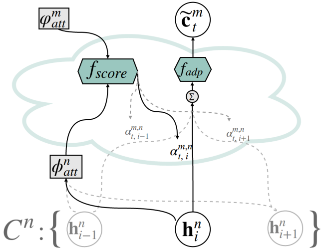
- Many-to-one (Zoph and Knight, 2016)
  - Separate attention for each encoder.
  - Necessitates multi-text.
  - Small scale experiments (WMT'15 subset).

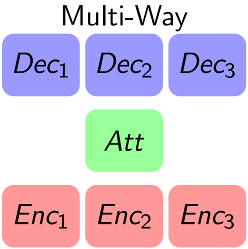


Multi-way NMT - Overview

Firat, Cho and Bengio, 2016

		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
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

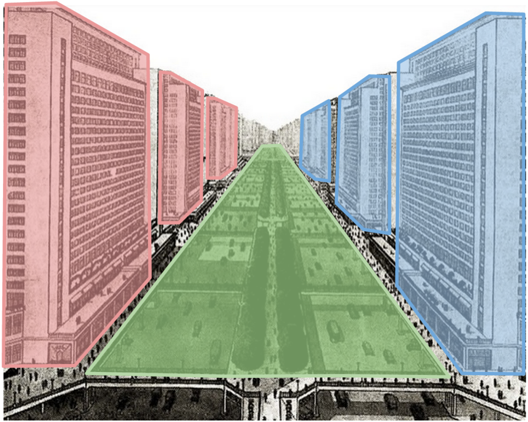




Warren Weaver- "Translation", 1949

Tall towers analogy:

- Do **NOT** model the individual behaviour of a car,
- **Model how the highway works!**



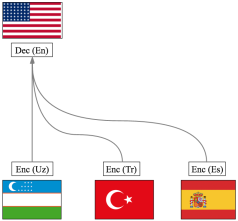
A lot of (if not all) interesting, related questions remain ...

1. What is it good for, other than parameter saving?
2. What if a source sentence is given in multiple languages?
3. What happens with language pairs that are not included during training?
4. How are we going to introduce additional modalities?

ML-NMT is good for Low-Resource (Firat et al., 2016b)

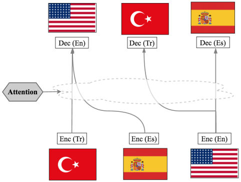
$[Uz \rightarrow En]$

Added Pairs	Test Single	Test Ensemble
Uz→En	42.56 (6.45)	38.56 (8.81)*
+ Tr→En	36.79 (9.34)	34.49 (11.69)*
+ Tr→En + Es→En	**35.39 (10.34)**	33.20 (12.33)*
+ Tr→En + En→Uz + En→Tr	36.28 (9.41)	33.65 (11.30)*
MLNMT Ensemble	**31.77 (12.99)**†	
Conventional SMT	32.38 (9.37)	



$[Tr \rightarrow En]$

Added Pairs	Test Single	Test Ensemble
Tr→En	28.58 (17.28)	24.27 (20.83)*
+ Es→En	27.49 (17.75)	23.94 (20.89)*
+ Es→En + Fr→En	26.77 (18.13)	24.00 (20.90)*
+ Es→En + En→Tr + En→Es	**26.30 (18.66)**	24.28 (20.23)*
MLNMT Ensemble	**21.78 (22.56)**†	
Conventional SMT	23.42 (18.00)	



Where does the improvement coming from?

1. Encoder is shared across *one-to-many* pairs
2. Decoder is shared across *many-to-one* pairs
3. The soft-alignment mechanism is shared across all pairs

What can we do more?

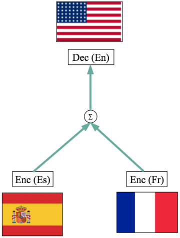
1. Share an encoder for multiple, similar source languages
2. Share a decoder for multiple, similar target languages
3. Perhaps, one recurrent net to rule both source and target languages..? (**Lee et al.'16**)

What if a source sentence is given in multiple languages?

Multi-source neural machine translation

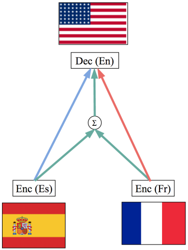
Multi-text given during **training**

1. Train the model for $p(y|x_1, x_2)$ (Zoph and Knight, 2016)
2. May need to device a merger operation



Multi-text given during **test**

1. Two translation strategies (Firat et al., 2016c)
   1.1 Late averaging $p(y|x_1) + p(y|x_2)$
   1.2 Early averaging $p(y|x_1, x_2)$
2. Use existing shared attention for merger operation
   - Simply take the *mean* of representations



What if a source sentence is given in multiple languages?

Multi-source neural machine translation

Multi-text given during **test**

1. Two translation strategies (Firat et al., 2016c)
   1.1 Late averaging $p(y|z_1) + p(y|x_2)$
   1.2 Early averaging $p(y|z_1, z_2)$
2. Use existing shared attention for merger operation
   - Simply take the *mean* of representations

Single-source translation:

	Src	Trgt	Multi Dev	Multi Test	Single Dev	Single Test
(a)	Es	En	30.73	28.32	29.74	27.48
(b)	Fr	En	26.93	27.93	26.00	27.21
(c)	En	Es	30.63	28.41	31.31	28.90
(d)	En	Fr	22.68	23.41	22.80	24.05

Multi-source translation:

		Multi Dev	Multi Test	Single Dev	Single Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

What have we learned from this?

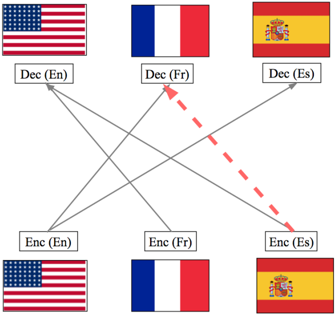
1. No need for multi-way parallel corpora!
2. Because, training with multiple language pairs has encouraged the model to find a common context vector space.
3. Allows us to use simple arithmetic operations in a hopefully flattened manifold.

What more should we do?

1. Finetuning with multi-way parallel corpus helps, but how far can we go?
2. Larger-scale experiments with more source language pairs.

Towards Zero-Resource Language Translation

Can we translate between a pair without any direct resource?

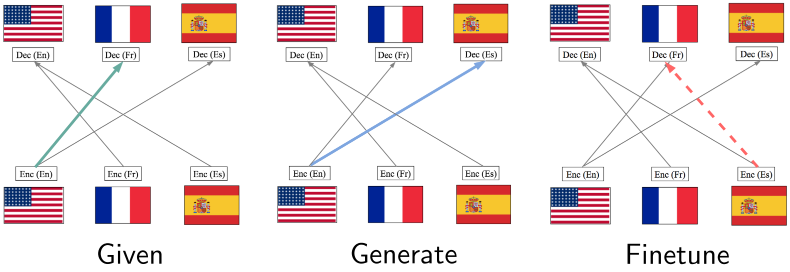


	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	√		20.64	20.4

Unfortunately no! Instead *Es* → *En* → *Fr* is promising

Towards Zero-Resource Language Translation

- *Es* → *Fr*: perhaps we can *generate* a pseudo-parallel corpus (Sennrich et al.,2016)
- Still *zero* direct resource



Towards Zero-Resource Language Translation

Firat et al., 2016c

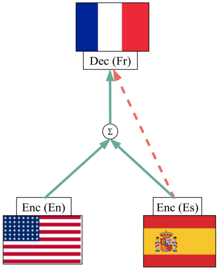
- Generate a pseudo-source (Es) given $En - Fr$ parallel corpus
- Finetune $Es - Fr$ using pseudo-parallel corpus

	Pseudo Parallel Corpus				True Parallel Corpus			
	1k	10k	100k	1m	1k	10k	100k	1m
Dev	–	–	–	–	–	–	11.25	21.32
Test	–	–	–	–	–	–	10.43	20.35
	Dev: 20.64, Test: 20.4				–			
Dev	0.28	10.16	15.61	17.59	0.1	8.45	16.2	20.59
Test	0.47	10.14	15.41	17.61	0.12	8.18	15.8	19.97

Towards Zero-Resource Language Translation

Firat et al., 2016c

Pivot	Many-to-1		Pseudo Parallel Corpus 1k	10k	100k	1m	True Parallel Corpus 1k	10k	100k	1m
√	No Finetuning		Dev: 20.64, Test: 20.4				–			
		Dev	0.28	10.16	15.61	17.59	0.1	8.45	16.2	20.59
		Test	0.47	10.14	15.41	17.61	0.12	8.18	15.8	19.97
√	Early	Dev	19.42	21.08	21.7	21.81	8.89	16.89	20.77	22.08
		Test	19.43	20.72	21.23	21.46	9.77	16.61	20.40	21.7
√	Early+ Late	Dev	20.89	20.93	21.35	21.33	14.86	18.28	20.31	21.33
		Test	20.5	20.71	21.06	21.19	15.42	17.95	20.16	20.9



- **Multi-source comes to aid!**
  - Teacher (multi-source)
  - Student (zero-resource)

What have we learned from this?

1. Don't necessarily need a direct resource.
2. Multilingual NMT naturally embeds multiple translation strategies.

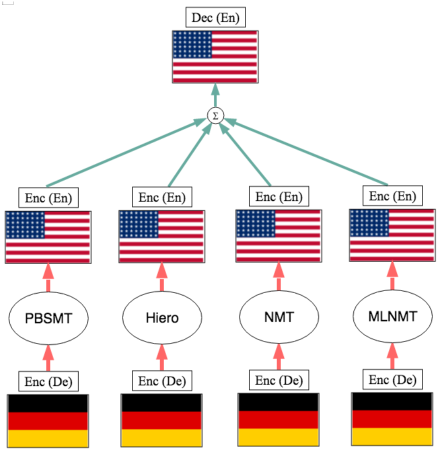
What more should we do?

1. Active learning can bring additional gains.
2. Perhaps, simply more data will constrain the attention mechanism to work with zero-resource pairs automatically.

Trainable Neural System Combination

Firat, Freitag and Cho - ongoing

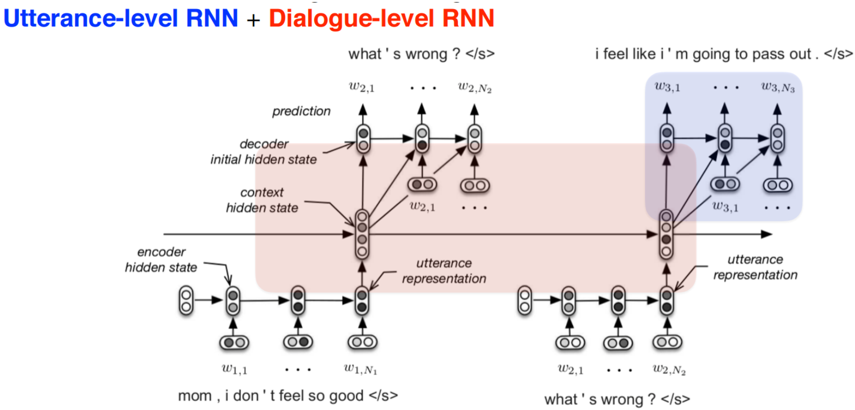
How to combine traditional SMT models with NMT models?



Larger Context NMT - Going beyond sentences!

Hierarchical Recurrent Encoder-Decoder (Serban et al., 2015, Sordoni et al., 2015)

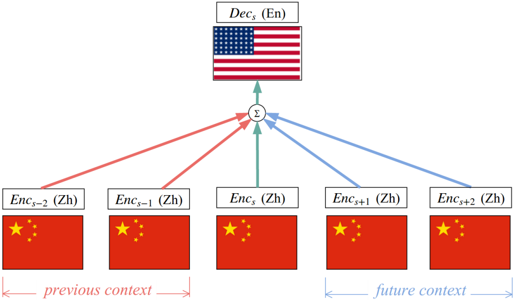
For dialogue modelling, capture previous context.



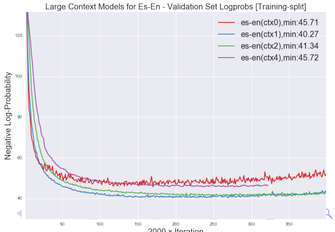
Larger Context NMT - Going beyond sentences!

Firat, Lauly and Cho - ongoing

Extend context to multiple past/future sentences in the document.



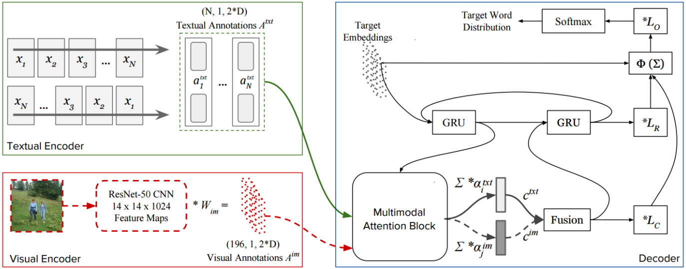
- New released UN Corpus (Es→En and Zh→En)
- **Mean** as the merger op
- Consider context window of size $0, 1, 2$ and 4



What about multi-modal NMT?

“Multi-modal Attention for Neural Machine Translation”

Caglayan, Barrault and Bougares, 2016 (actually, just yesterday)



Model	Attention Type: Fusion	Modality	Decoder	Validation Scores: METEOR	BLEU	CIDEr-D
NMT	-	-	-	34.24 (35.59)	18.64 (21.62)	58.57 (67.93)
IMGTXT	-	-	-	26.80	11.16	31.28
MNMT1	SUM	IND	IND	33.23 (35.42)	18.30 (21.24)	55.45 (65.03)
MNMT2	SUM	IND	DEP	34.17 (35.48)	17.70 (20.70)	53.78 (61.76)
MNMT3	SUM	DEP	IND	34.38 (35.55)	18.42 (20.94)	55.81 (63.37)
MNMT4	SUM	DEP	DEP	33.67 (34.57)	17.83 (20.30)	52.68 (59.63)
MNMT5	CONCAT	IND	IND	33.31 (34.98)	17.50 (20.60)	53.57 (61.46)
MNMT6	CONCAT	IND	DEP	**35.23** (36.79)	19.30 (22.45)	60.62 (69.96)
MNMT7	CONCAT	DEP	IND	**35.11 (37.13)**	**19.72 (23.24)**	**61.04 (72.16)**
MNMT8	CONCAT	DEP	DEP	34.80 (**36.98**)	19.55 (22.78)	60.20 (70.20)

How far we can extend the existing approaches?

Bigger models, complicated architectures!

RNNs can express/approximate a set of Turing machines,

BUT* ...

expressivity $\neq$ learnability

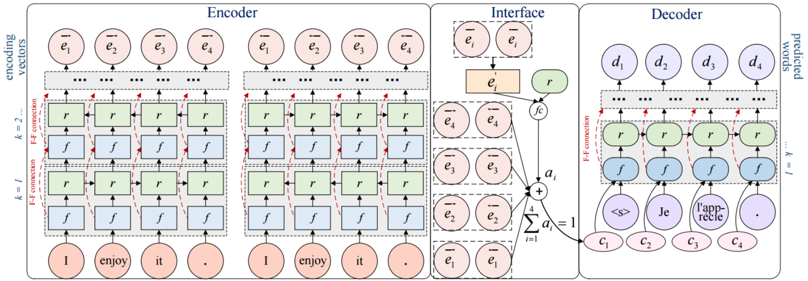
*Edward Grefenstette: Deep Learning Summer School 2016

How far we can extend the existing approaches?

Fast-Forward Connections for NMT, (Zhou et al., 2016)

Bigger models are harder to train!

- Deep topology for recurrent networks (16 layers)
- Performance boost (+6.2 BLEU points)
- Fast-forward connections for gradient flow

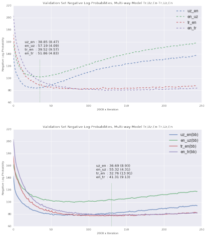


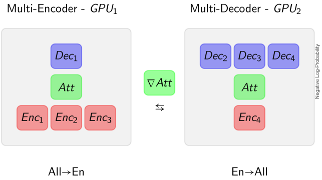
How far we can extend the existing approaches?

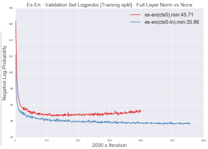
Multi-way Multilingual NMT

Bigger models are harder to train and behave differently!

- Scheduling the learning process
- Preventing the unlearning (catastrophic forgetting)
- Layer Normalization (Kiros et al.,2016)







What are we optimizing?

Explorations on the right objective to be optimized NLL:

- MERT (Och,2003), MRT for NMT (Shen et al.,2016)
- Scheduled Sampling (Bengio et al.,2015), Sequence Level Training (Ranzato et al.,2015), Task Loss Estimation (Bahdanau et al.,2015)
- Actor-Critic (Bahdanau et al.,2016), Reward Augmented ML (Norouzi et al.,2016)
- Seq2Seq as Beam-Search optimization (Wiseman and Rush, 2016)

New territory seems to be using new error signals!

What Lies Ahead?

Perhaps, we've only scratched the surface!

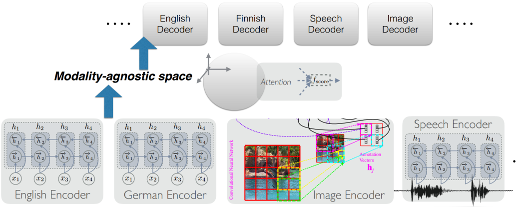
- Language barrier, surpassing human level quality.

Revisiting the new territory:

Character-level Larger-Context Multilingual
Neural Machine Translation

using,

- Multiple modalities
- Better error signals
- and better GPUs 😎



One last thing!

Let's remember the game we were playing once more,

Thank you!

Thanks to: TUBITAK, NSERC, Samsung, IBM, Calcul Quebec, Compute Canada, The Canada Research Chair, CIFAR, NVIDIA, Facebook and Google