Breakthroughs in Neural Machine Translation

Olof Mogren

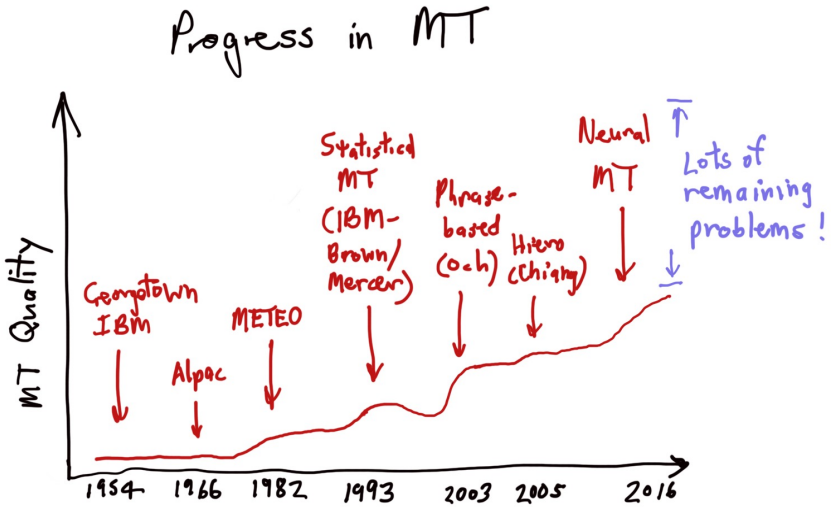
Chalmers University of Technology

2016-09-29

Coming seminars

- Today: Olof Mogren
  Neural Machine Translation
- October 6: John Wiedenhoeft
  Fast Bayesian inference in Hidden Markov Models using Dynamic Wavelet Compression
- October 10: Haris Charalambos Themistocleous
  Linguistic, signal processing, and machine learning approaches in eliciting information form speech

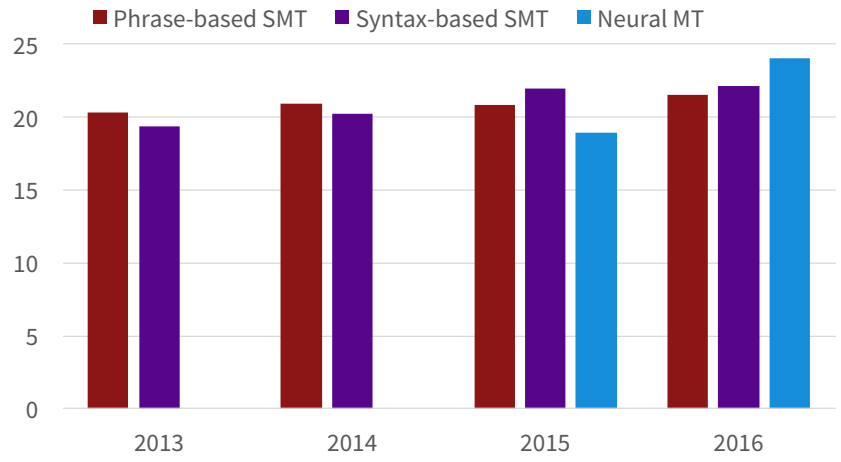
http://www.cse.chalmers.se/research/lab/seminars/


Progress in MT
MT Quality
Georgetown IBM
Alpac
METEO
Statistical MT (IBM-Brown/Mercer)
Phrase-based (o.c.h)
Hiero (Chiang)
Neural MT
Lots of remaining problems!
1954
1966
1982
1993
2003
2005
2016

n.one/

Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



From [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

Phrase-based Statistical Machine Translation

A **marvelous** use of **big data** but … it's mined out?!?

1519年600名西班牙人在墨西哥登陆，去征服几百万人口的阿兹特克帝国，初次交锋他们损兵三分之二。

In 1519, six hundred Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

translate.google.com (2009): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.

translate.google.com (2013): 1519 600 Spaniards landed in Mexico to conquer the Aztec empire, hundreds of millions of people, the initial confrontation loss of soldiers two-thirds.

translate.google.com (2014): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

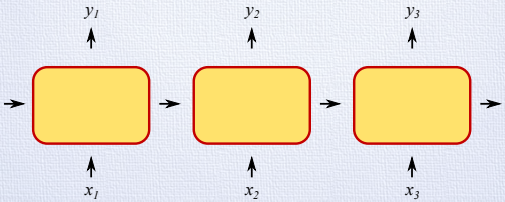
translate.google.com (2015): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

translate.google.com (2016): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

What is Neural MT (NMT)?

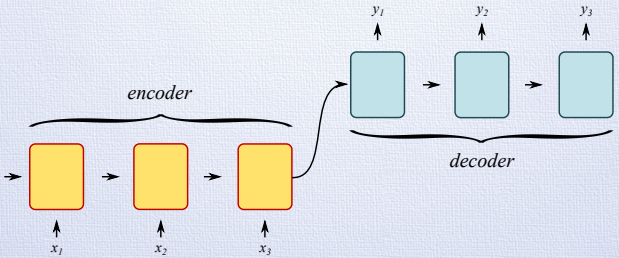
The approach of modelling the entire MT process
via one big artificial neural network.

Modelling language using RNNs



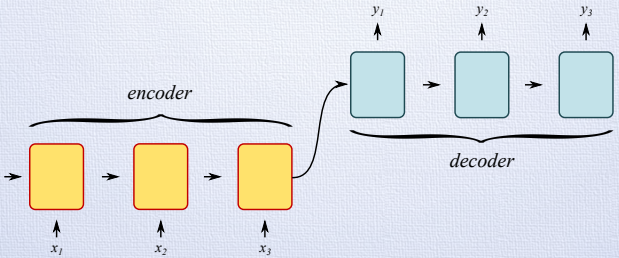
- Language models: $P(word_i|word_1, ..., word_{i-1})$
- Recurrent Neural Networks
- Gated additive sequence modelling:
  LSTM (and variants) details
- Fixed vector representation for sequences
- Use with beam-search for language generation

Encoder-decoder framework



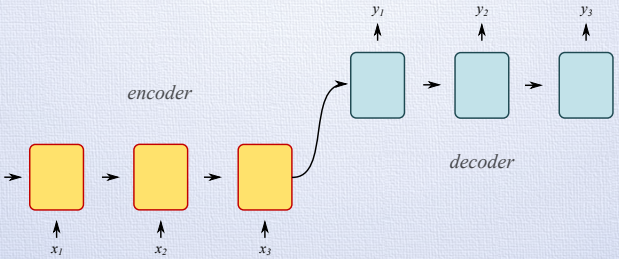
- Sequence to Sequence Learning with Neural Networks
  Ilya Sutskever, Oriol Vinyals, Quoc V. Le, NIPS 2014

Encoder-decoder framework



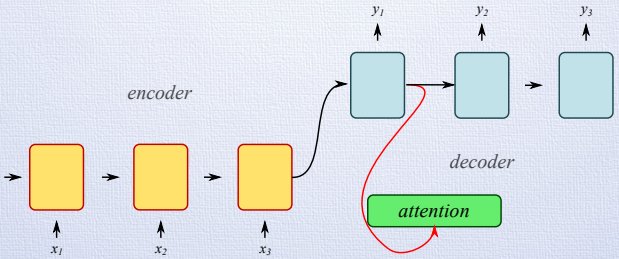
- Sequence to Sequence Learning with Neural Networks
  Ilya Sutskever, Oriol Vinyals, Quoc V. Le, NIPS 2014
- Reversed input sentence!

Encoder-decoder with attention



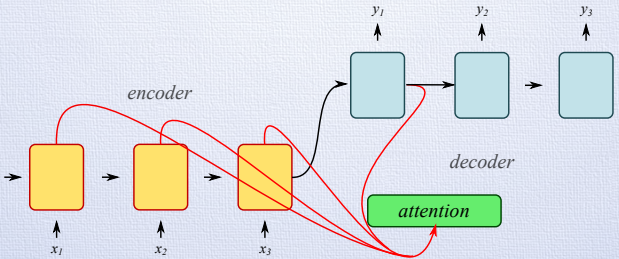
- Neural Machine Translation by Jointly Learning to Align and Translate
  Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio - ICLR 2015

Encoder-decoder with attention



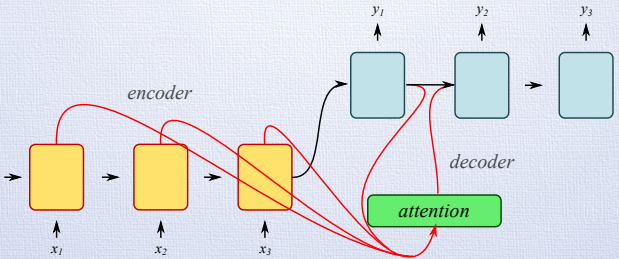
- Neural Machine Translation by Jointly Learning to Align and Translate
  Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio - ICLR 2015

Encoder-decoder with attention



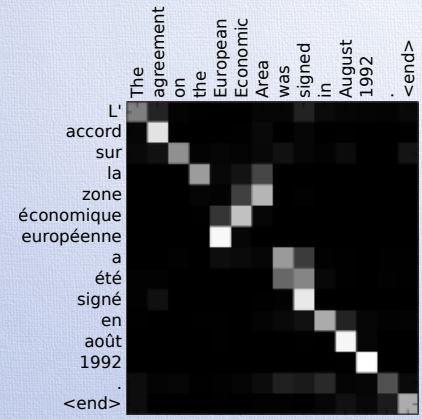
- Neural Machine Translation by Jointly Learning to Align and Translate
  Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio - ICLR 2015

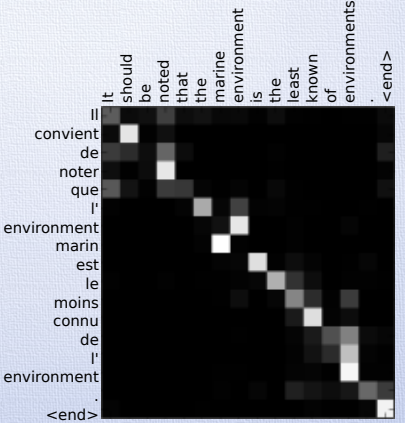
Encoder-decoder with attention



- Neural Machine Translation by Jointly Learning to Align and Translate
  Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio - ICLR 2015

Alignment - (more)





Neural Machine Translation, NMT

- End-to-end training
- Distributed representations
- Better exploitation of context

What's not on that list?

What's been holding NMT back?

- Limited vocabulary
  - Copying
  - Dictionary lookup
- Data requirements
- Computation
  - Training time
  - Inference time
  - Memory usage

Rare words 1: Subword units

- Neural machine translation of rare words with subword units
  Rico Sennrich and Barry Haddow and Alexandra Birch
- A character-level decoder without explicit segmentation for neural machine translation
  Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio, ACL 2016

Byte-pair encoding (BPE):

```
aaabdaaabac
```

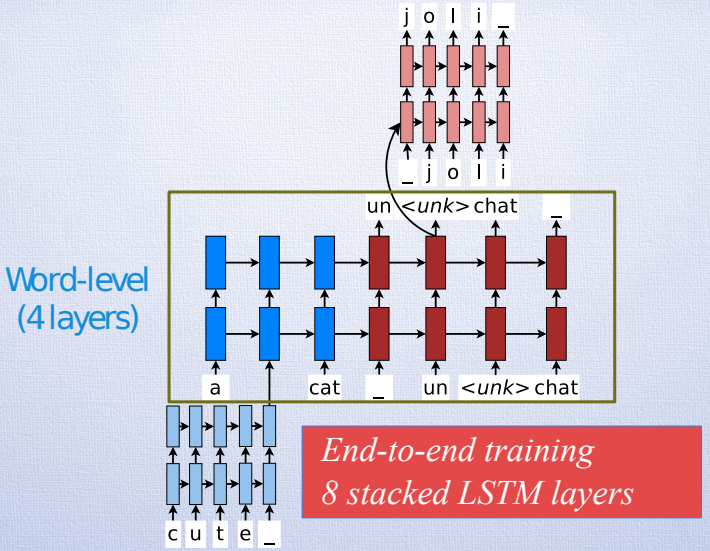
```
ZabdZabac
Z=aa
```

```
ZYdZYac
Y=ab
Z=aa
```

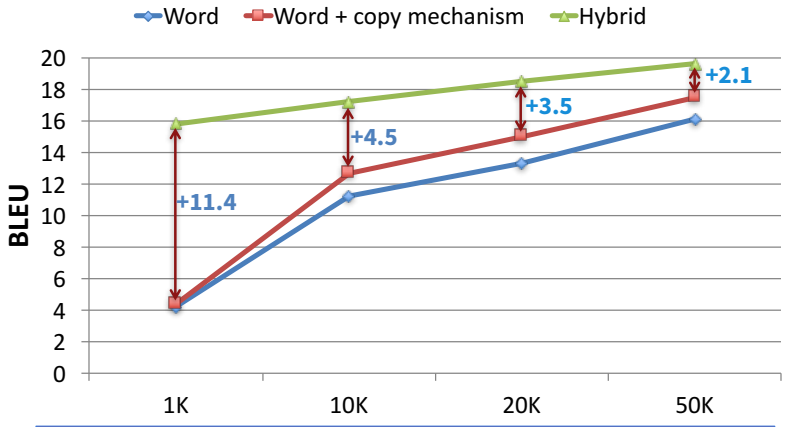
```
XdXac
X=ZY
Y=ab
Z=aa
```

Rare words 2: Hybrid char/word NMT

- Achieving open vocabulary neural machine translation with hybrid word-character models
  Thang Luong and Chris Manning, ACL 2016.
- Hybrid architechture:
  - Word-based for most words
  - Character-based for rare words
  - 2 BLEU points improvement over copy mechanism


Word-level
(4 layers)
j o l i _
_ j o l i
un <unk> chat _
a cat _ un <unk> chat
c u t e _
End-to-end training
8 stacked LSTM layers

Effects of Vocabulary Sizes



More than +2.0 BLEU over copy mechanism!

Rare Word Embeddings

- Word & character-based embeddings.

Training with monolingual data

- Improving neural machine translation models with monolingual data *Rico Sennrich, Barry Haddow, Alexandra Birch, ACL 2016.*
- Backtranslate monolingual data (with NMT model)
- Use backtranslated data as parallell training data

Enriching parallel data

- *Dummy* source sentences

She loves cute cats	Elle aime les chats mignons	(parallel)
<null>	Elle aime les chiens mignons	(mono)

Small gain +0.4-1.0 BLEU.
Difficult to add more mono data.

Rico Sennrich, Barry Haddow, and Alexandra Birch. ***Improving Neural Machine Translation Models with Monolingual Data****. ACL 2016.*

Enriching parallel data

- *Synthetic* source sentences

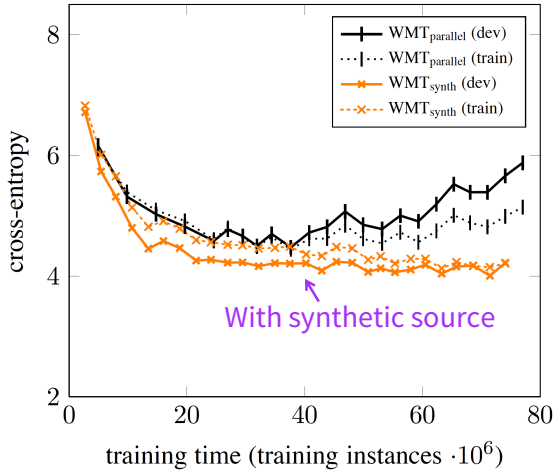
She loves cute cats	Elle aime les chats mignons	(parallel)
She likes cute cats	Elle aime les chiens mignons	(mono)

Back translated

Large gain +2.1-3.4 BLEU.

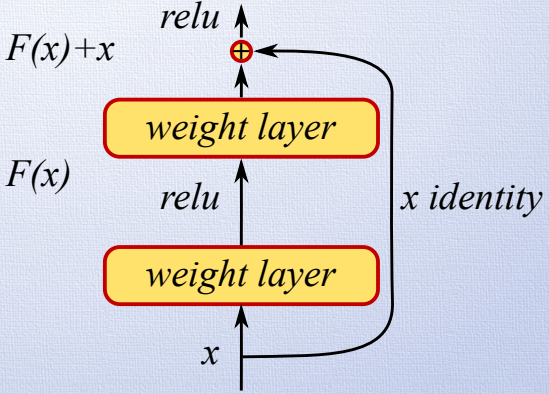
Rico Sennrich, Barry Haddow, and Alexandra Birch. ***Improving Neural Machine Translation Models with Monolingual Data****. ACL 2016.*

Prevent Over-fitting



Residual deep LSTMs

- Deep recurrent models with fast-forward connections for neural machine translation: *Jie Zhou et.al., Baidu research, arXiv preprint, 1606.04199*
- Residual (skip) connections in depth
- 16 layers deep LSTM model

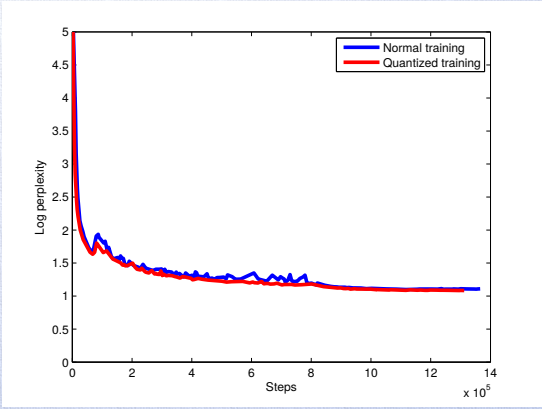


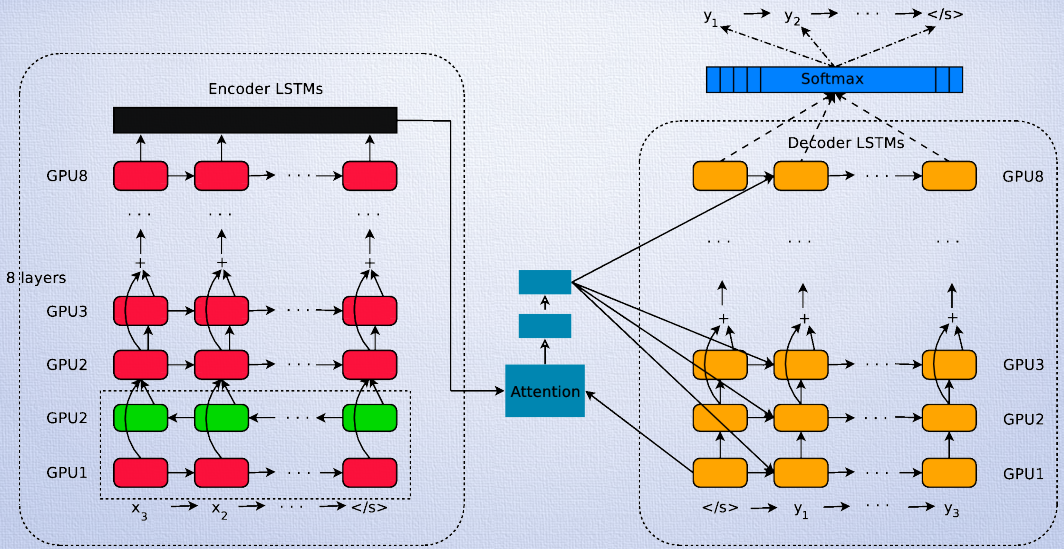
Putting it all together

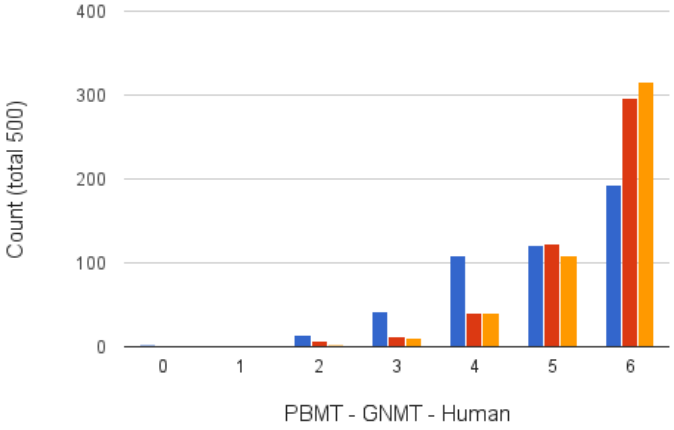
- Google's neural machine translation system:
  Bridging the gap between human and machine translation
  Yonghui Wu, et.al., Google, arXiv preprint, 1609.08144
- Subwords like Sennrich et.al. (BPE)
- 8 layers deep LSTM model.
- Quantizised weights (see next slide)
- Downpour SGD: parallell training
- 8GPUs, one host.

Quantized inference

- Training: real-valued weights
- Limit precision:
  improved inference speed
- Weights $\in -1, 0, 1$
- Extra constraints on training:
  $x_t^i, c_t^i \in [-\delta, \delta]$
- Similar constraints on
  softmax layer.




Encoder LSTMs
GPU8
8 layers
GPU3
GPU2
GPU2
GPU1
x3
x2
</s>
Attention
y1
y2
</s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s>
y1
y3


400
300
200
100
0
Count (total 500)
0
1
2
3
4
5
6
PBMT - GNMT - Human

Single model BLEU scores

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	

Ensemble model BLEU scores

Model	BLEU
WPM-32K (8 models)	40.35
RL-refined WPM-32K (8 models)	41.16
LSTM (6 layers) [30]	35.6
LSTM (6 layers + PosUnk) [30]	37.5
Deep-Att + PosUnk (8 models) [43]	40.4

Single model BLEU scores

Model	BLEU	Side-by-side averaged score
PBMT [15]	37.0	3.87
NMT before RL	40.35	4.46
NMT after RL	41.16	4.44
Human		4.82

Future of (N)MT 1

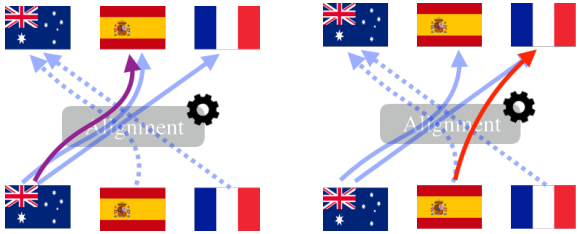
- Larger context (not only one sentence at a time)
  - Attention for long sequences in **speech**:
    Chan, Jaity, Le, Vinyals, ICASSP 2015
  - Tracking states over many sentences in **dialogue systems**:
    Serban, Sordoni, Bengio, Courville, Pineau , AAAI 2015

Future of (N)MT 2

- Multi-language translation models
  - Multi-Task Learning for Multiple Language Translation
    Dong, Wu, He, Yu, Wang, ACL 2015
  - Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism
    Firat, Cho, Bengio, NAACL 2016
  - Improvement for low-resoursce languages
  - Not yet as good for high-resource languages
  - Zero-resource translation (some initial results)

Multilingual Translation: Looking Ahead

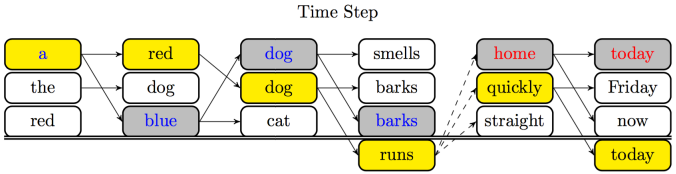
- Zero-resource translation
  - Finetuning with *pseudo*-parallel corpus [Sennrich et al., ACL2016]
  - Closely related to unsupervised learning



Pseudo-corpus Generation **Finetuning** [Firat et al., EMNLP2016]

Beyond Maximum Likelihood

- Maximize the sequence-wise global loss
- Incorporate inference into training
  - Stochastic inference
    - Policy gradient [Ranzato et al., ICLR2016; Bahdanau et al., arXiv2016]
    - Minimum risk training [Shen et al., ACL2016]
  - Deterministic inference
    - Learning to search [Wiseman & Rush, arXiv2016]



mogren@chalmers.se

http://mogren.one/

Appendix

by *ent362* , *ent300* updated 6:06 pm et , thu march 26 , 2015 (*ent300*) the `` *ent321* '' series will have to handcuff a new director . *ent201* , who directed `` *ent71* , '' told *ent286* that she wo n't be back for the sequel , `` *ent100* . '' `` directing ' *ent135* ' has been an intense and incredible journey for which i am hugely grateful , '' she said in a statement to the site . `` while i will not be returning to direct the sequels , i wish nothing but success to whosoever takes on the exciting challenges of films two and three . '' ' *ent71* ' : what fans hoped for ? the first film in the best - selling book series has been hugely successful , pulling in more than $ 550 million worldwide since it premiered in mid-february , but there have been rumbles that creative clashes were in the offing for the sequel . author *ent341* has a great deal of control in how her books are presented on screen , and she made it clear that she wanted to write the screenplay for the second film , *ent184* reported last month . *ent28* wrote the screenplay for `` *ent71* . '' the story behind mr. *ent289* 's suits the film stars *ent344* as billionaire *ent275* -- a man of certain sexual proclivities -- and *ent407* as his romantic partner , *ent389* .

x bows out of the `` *ent321* '' sequel

by *ent339* , *ent42* updated 2:59 pm et , thu march 26 , 2015 (*ent42*) call it `` *ent351* . '' a *ent396* state trooper caught a driver using a cardboard cutout of *ent421* , the *ent364* beer pitchman known as `` *ent397* . '' the driver , who was by himself , was attempting to use the *ent214* . `` the trooper immediately recognized it was a prop and not a passenger , '' trooper *ent367* told the *ent375* . `` as the trooper approached , the driver was actually laughing . '' *ent143* sent out a tweet with a photo of the cutout -- who was clad in what looked like a knit shirt , a far cry from his usual attire -- and the unnamed laughing driver : `` i do n't always violate the *ent303* lane law ... but when i do , i get a $ 124 ticket ! we 'll give him an a for creativity ! '' the driver was caught on *ent300* near *ent327* , *ent396* , just outside *ent53* . `` he could have picked a less recognizable face to put on his prop , '' *ent143* told the *ent375* . `` we see that a lot . usually it 's a sleeping bag . this was very creative . ''

a driver was caught in the **x** with a cutout of `` *ent7* ''

Teaching Machines to Read and Comprehend, Dec 2015

Hermann, Kocisky, Greffenstette,

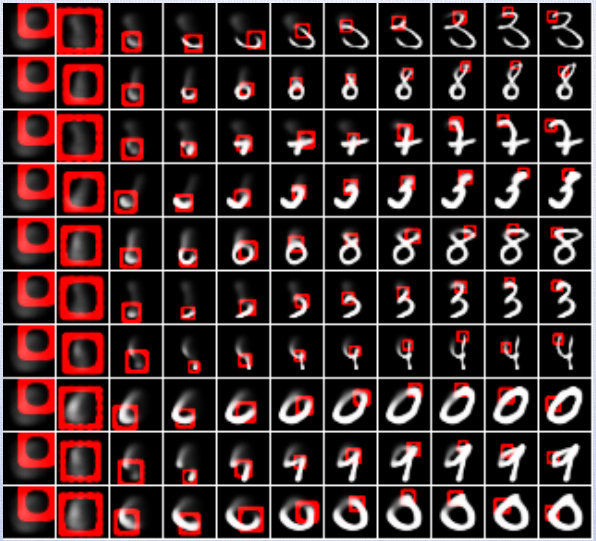
Espeholt, Kay, Suleyman, Blunsom

http://mogren.one/

What wasn't on that list?

- Explicit use of syntactic or semantic structures
- Explicit use of discourse structure, anaphora, etc.
- Black box component models for reordering, transliteration, etc

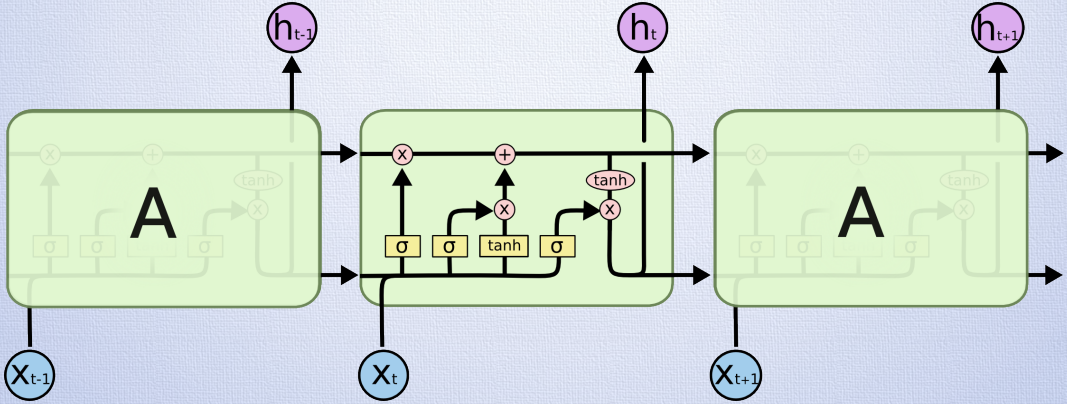
back



DRAW, A Recurrent Neural Network For Image Generation - 2015

Gregor, Danihelka, Graves, Rezende, Wierstra

LSTM



Christopher Olah

back