Neural Machine Translation

Kyunghyun Cho *kyunghyun.cho@umontreal.ca*
Bart van Merrienboer *bart.vanmerrienboer@umontreal.ca*
Caglar Gulcehre *caglar.gulcehre@umontreal.ca*
Dzmitry Bahdanau *d.bahdanau@jacobs-university.de*
Jean Pouget-Abadie *jean.pougetabadie@gmail.com*
Fethi Bougares *fethi.bougares@lium.univ-lemans.fr*
Holger Schwenk *holger.schwenk@lium.univ-lemans.fr*
Yoshua Bengio *yoshua.bengio@umontreal.ca*
and others..

Maybe..

"Syntax isn't a thing."

- Felix Hill, University of Cambridge, UK

Statistical Machine Translation – (1)

Statistical machine translation $\iff$ Maximizing $\log p(f \mid e)$

$$\log p(f \mid e) = \log p(e \mid f) + \log p(f) + C$$

1. $p(f \mid e)$: probability of a target sentence f given a source sentence e
2. $p(e \mid f)$: *so-called* translation model
3. $p(f)$: (target) language model
4. C: constant

Statistical Machine Translation – (2)

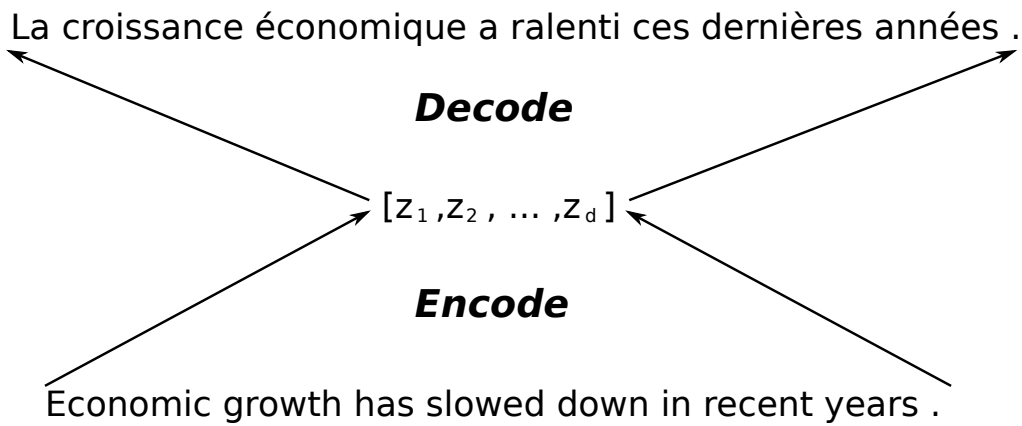
SMT in Reality

$$\arg\max_{f} \sum_{i} \omega_i \log f_i(e, f)$$

1. Build a weak model (log-linear model) for $p(f \mid e)$
2. Feed in lots of features larged based on researchers' intuition and experience
3. Maximize another metric, e.g., BLEU
4. (Use a super-strong language model)

Machine translation from machine learning perspective

Machine Translation
Nothing, but *one Supervised Learning Task*



Properties

- *Variable-length* input/output
- *Many-to-many* mapping

Implications

- *Order-sensitive*
- *Probabilistic*

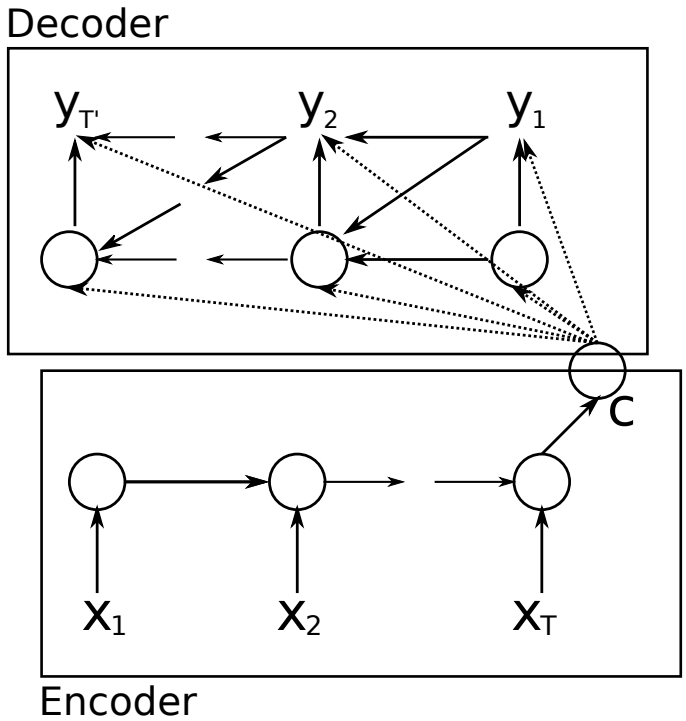
Kalchbrenner&Blunsom (2013), Kalchbrenner et al. (2014), Sutskever et al. (2014), **Cho et al. (2014)**, Grefenstette et al. (2014), **Bahdanau et al. (2014)**

Two obvious questions from the perspective of machine learning research:

1. *Why don't we simply maximize* $\log p(f \mid e)$*?*
2. *Why do we use a log-linear model for* $\log p(f \mid e)$*?*

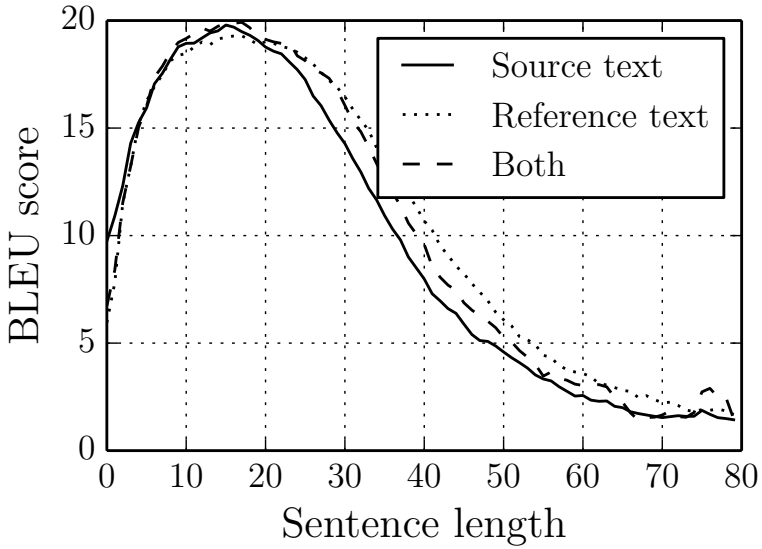
Let's maximize $p(f \mid e)$! – RNN Encoder–Decoder

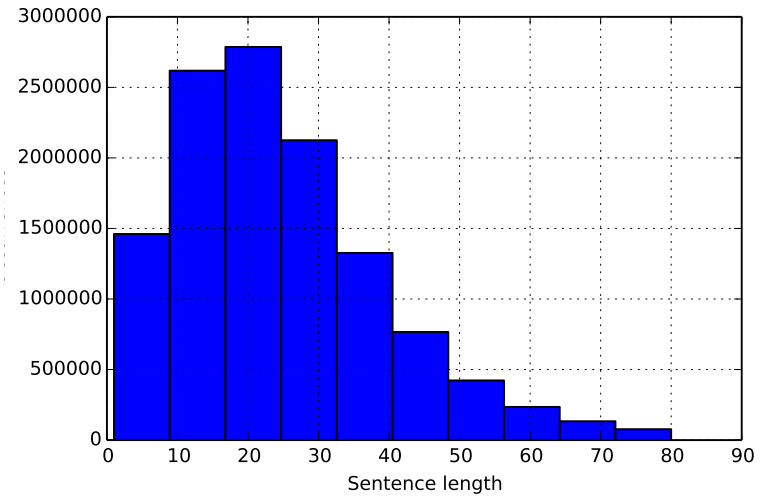
Model $\log p((y_1, \cdots, y_{T'}) \mid (x_1, \cdots, x_T))$ with two RNN's



And, it worked reasonably $\begin{cases} \text{good} & \text{for Google Brain} \\ \text{bad} & \text{for UdeM} \end{cases}$

How did it not work well? – Curse of Sentence Length





Potential Reasons:

1. **Difficult to encode a long sentence into a fixed-length vector**
2. Only few long sentences in the training corpus
3. Difficult to generate a coherent long sentence

Cho et al. (2014a,b), Pouget-Abadie et al. (2014)

Only if we can **align** *a target word with source words...*

Learning to Align and Translate Jointly – (1)

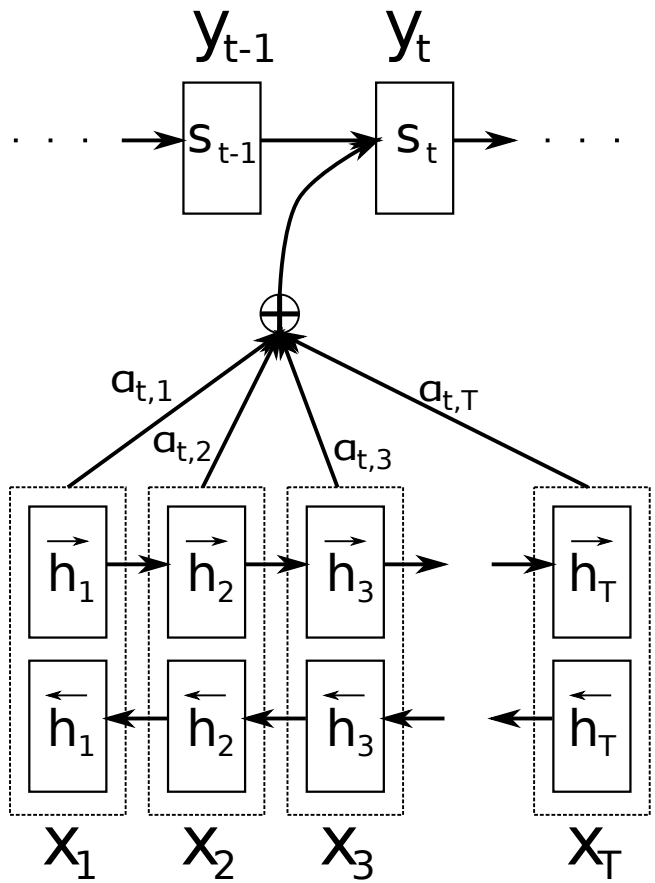
Annotation vector $h_j = \left[\overrightarrow{h}_j; \overleftarrow{h}_j\right]$

- $\overrightarrow{h}_j$: Forward RNN
- $\overleftarrow{h}_j$: Backward RNN

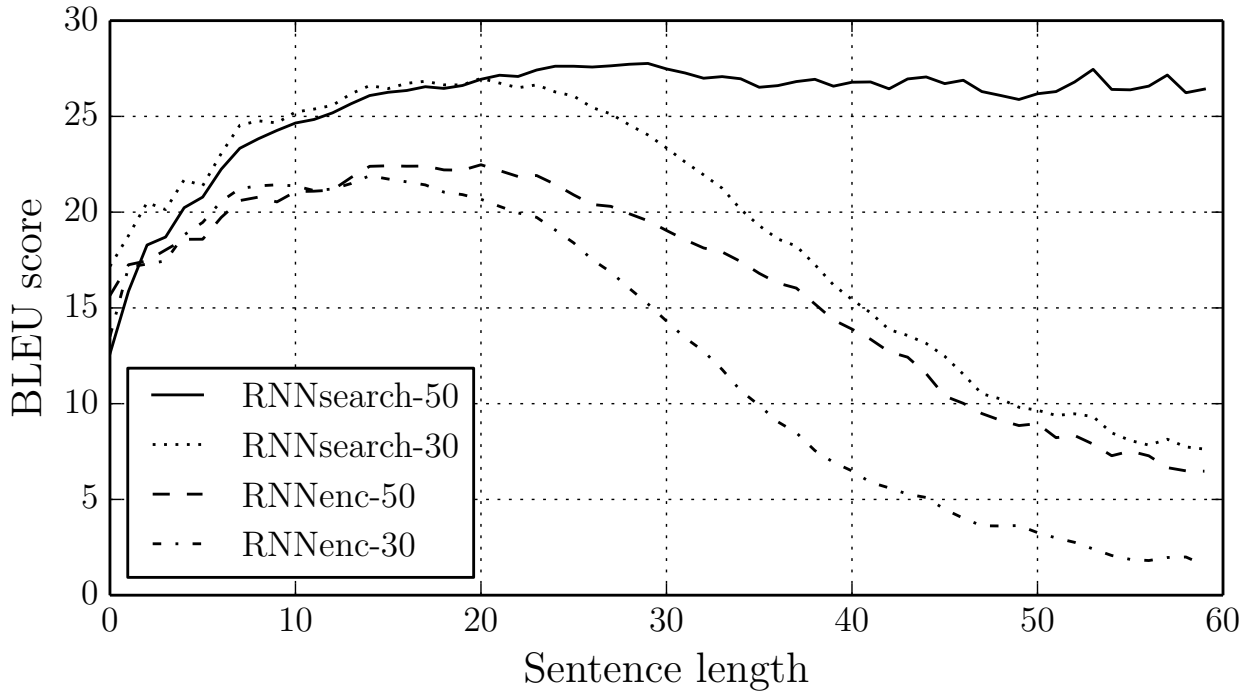
For each target word y_t,

1. Compute $\alpha_{t,j} = f(h_j, s_{t-1})$
   - $\sum_j \alpha_{t,j} = 1$
   - f is a feedforward neural network
2. Get a context vector $c_t = \sum_j \alpha_{t,j} h_j$

And, train the whole model with SGD and backpropagation!



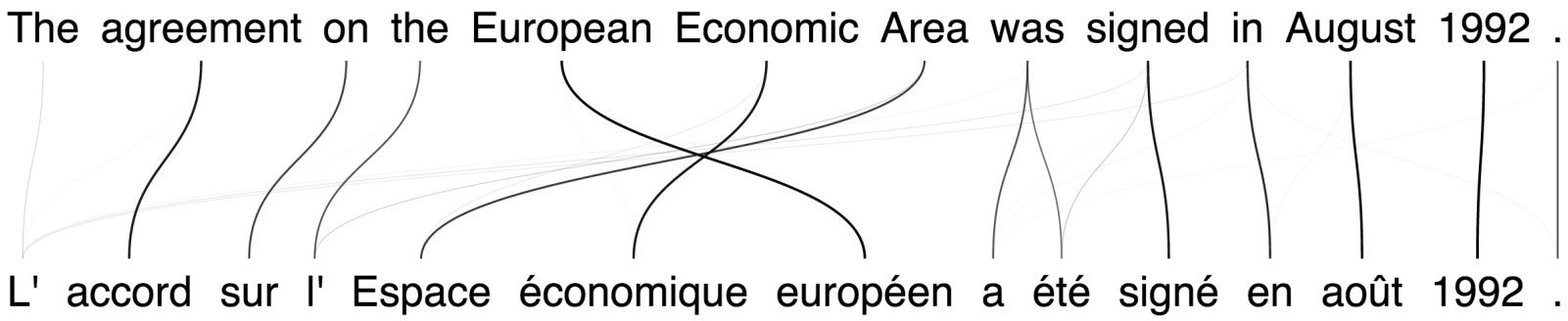
Learning to Align and Translate Jointly – (2)

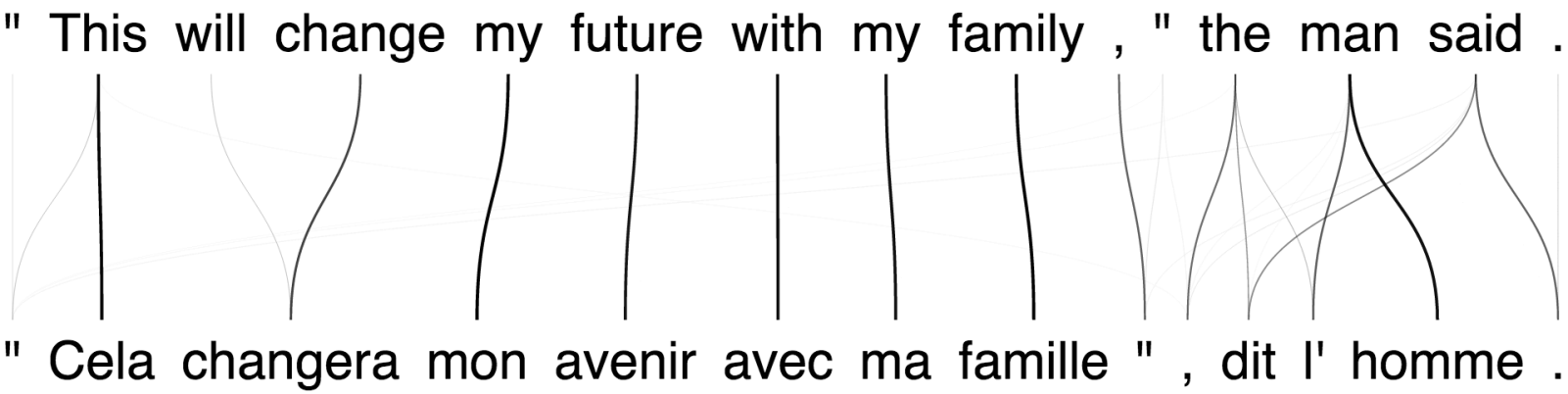


Mission accomplished!

It's more robust to the length of source sentences

Learning to Align and Translate Jointly – (3)





Where are we with Neural MT? – (1)

- Two groups (independently)
  Google Brain, **University of Montreal**
- $\approx$ ~~8 months~~ 9 months
- Encoder–Decoder architecture based on RNN
- English-to-French translation, *mainly*

Where are we with Neural MT? – (2)

Model	All	No UNK
U. Montreal	28.45°	**36.15**
Google Brain (forward*)	26.17	N/A
Google Brain (backward*)	30.59	N/A
Google Brain (backward*, 5 models)	**34.81**	N/A
Moses (no NN, standard setting)	33.30	35.63

Vocabulary sizes: 30k-30k (U. Montreal), 180k-90k (Google Brain)

$\star$: denotes the order in which the encoder reads the source sentence.

$\circ$: *Good news! We get better numbers with effectively infinite vocabularies!*

Where are we with Neural MT? – (3)

On English-to-French Translation

- **Neural MT works as well as phrase-based SMT**
- Much smaller memory footprint: *800MB and that's it!*

Check it out at `http://lisa.iro.umontreal.ca/mt-demo` now!

⋆ `https://github.com/lisa-groundhog/GroundHog`

Where are we with Neural MT? – (4)

English-to-Chinese *without* any segmentation on Chinese:

- Where are we ?
  – 我们在哪儿? (OK!)
- This is another translation.
  – 这又是另一个翻译。(OK!)

On IWSLT 2013 MT Track (TED Subtitle Translation)

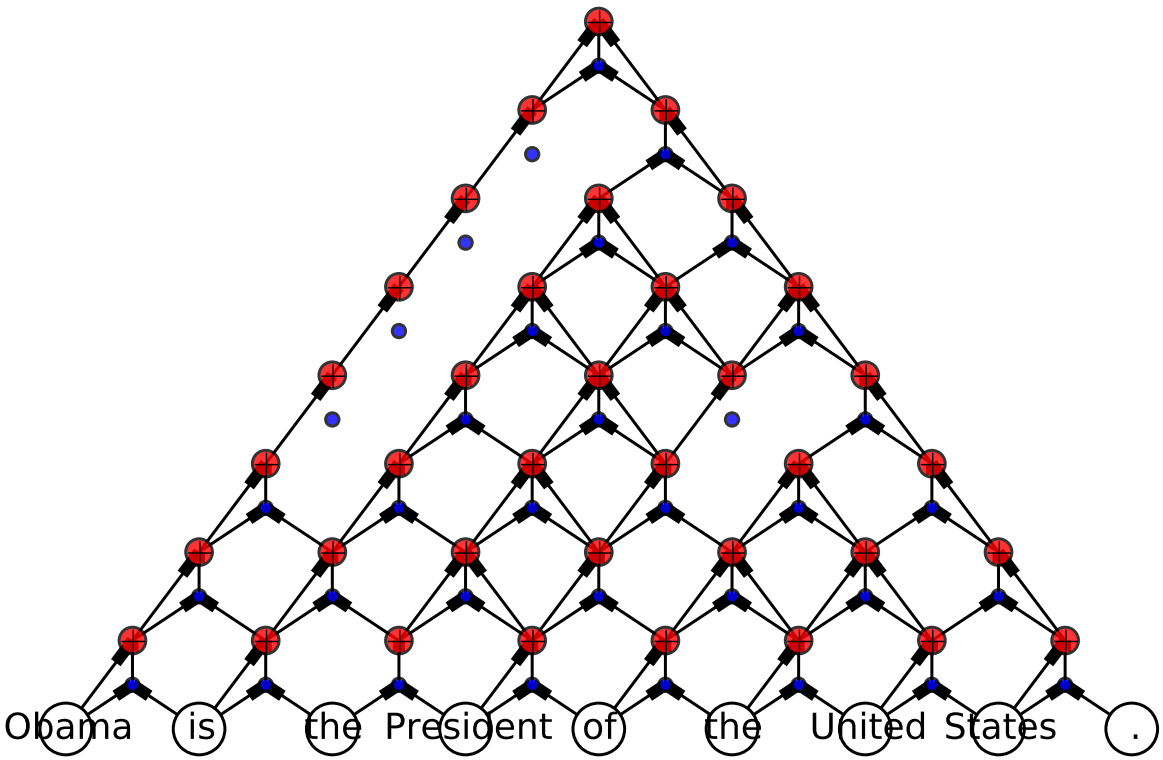
	NMT$^\circ$	Hier. Phrase-base SMT*
BLEU (char)	**23.84**	≈ 20

$\circ$ by Kelvin Xu and Saizheng Zhang
$\star$ Super complex model!

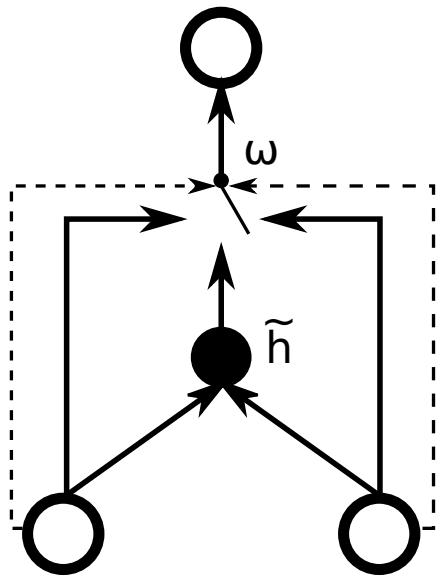
Where and when do syntax and structure come in?

A more interesting, but less powerful encoder: Gated Recursive Convolutional Network

- Encoder: *Gated* Recursive Convolutional Network
- Decoder: Recurrent Neural Network



(Soft) Gating



Cho et al. (2014a)

What does Neural MT say about Syntax and Structure?

Syntax and structure are *consequences*, not *necessary building blocks*.

$\Longleftrightarrow$

Syntax and structure follow naturally by solving actual problems.

Thanks for listening!

Any questions?

1. Any doubt about using super-complicated neural networks
2. Any doubt about divorcing away from traditional NLP
3. Challenges in Neural MT and ML for NLP
4. Challenges in RNN and Deep Learning
5. What's beyond Neural MT?