Statistical Machine Translation LING-462/COSC-482 Week 10: Refinements and Alternative Architectures

Achim Ruopp
achim.ruopp@Georgetown.edu

Agenda

- Language in ten minutes: Logan Peng
- NMT using a sequence-to-sequence RNNs with attention model
  - Refinements/Practical Considerations
  - Adding linguistic information
  - Multiple language pairs/Zero-shot translation
  - Ensemble Decoding

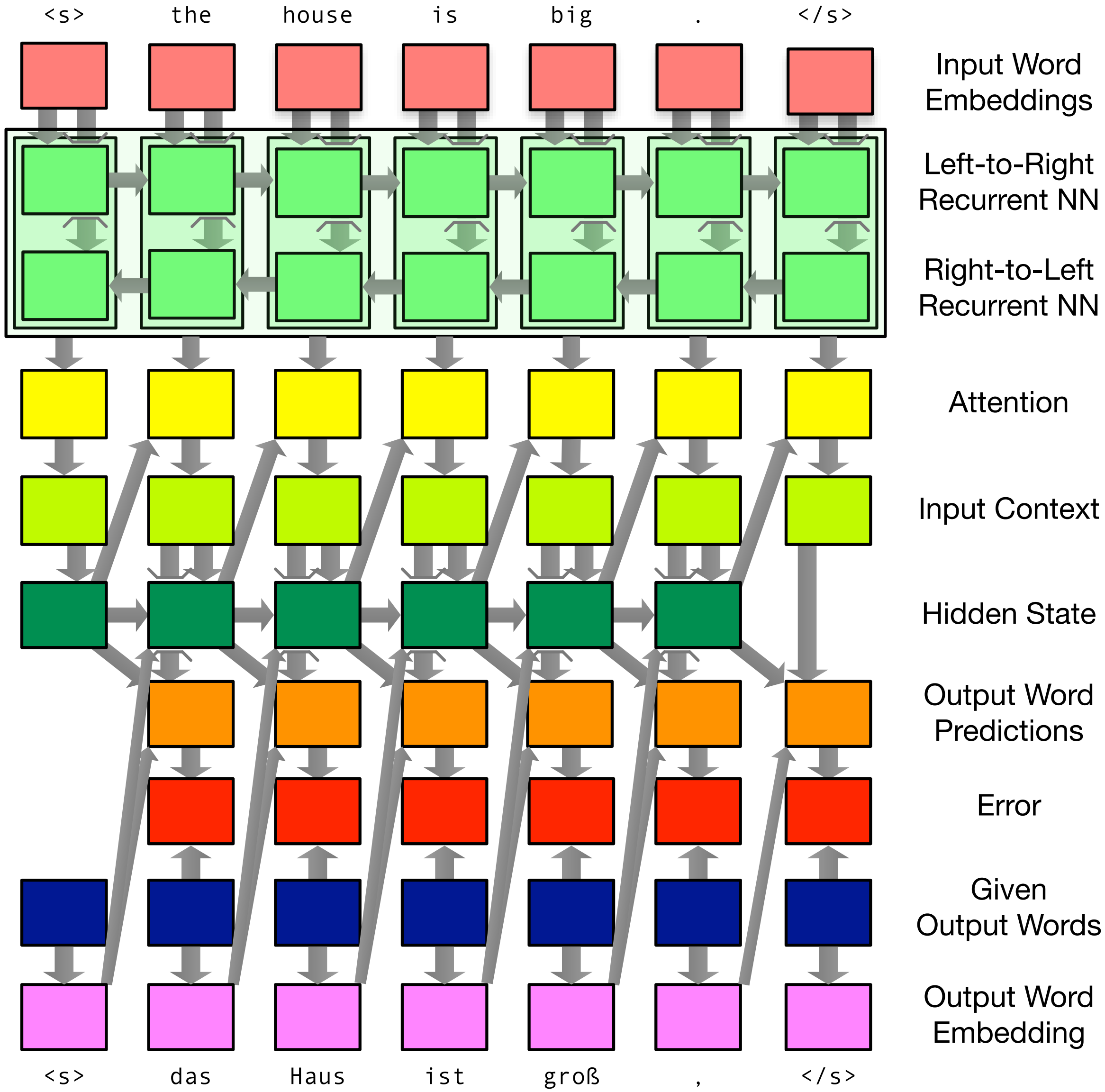
\- Break -

- NMT using convolutional networks with attention
- Transformer NMT model/Attention is all you need
- HW4 questions

NMT Toolkits

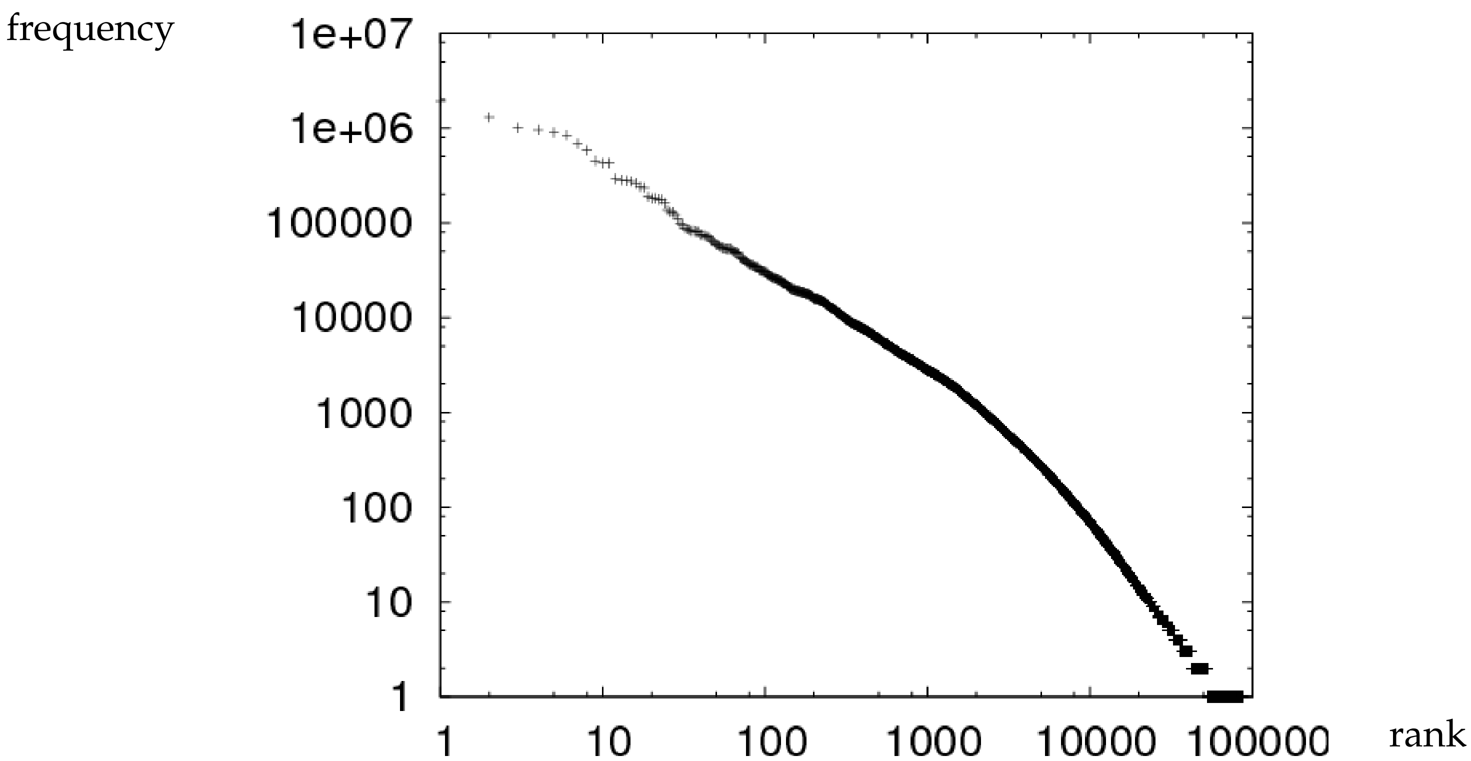
- Number of toolkits still increasing
- List maintained by Jon Dehdari
  - https://github.com/jonsafari/nmt-list
- Latest updates presented this Monday at the AMTA 2018 conference
  - http://www.conference.amtaweb.org/#program

Neural Machine Translation



large vocabularies

Zipf's Law: Many Rare Words



$$\text{frequency} \times \text{rank} = \text{constant}$$

Many Problems

- Sparse data
  - words that occur once or twice have unreliable statistics

- Computation cost
  - input word embedding matrix: $|V| \times 1000$
  - outout word prediction matrix: $1000 \times |V|$

Some Causes for Large Vocabularies

- Morphology

  tweet, tweets, tweeted, tweeting, retweet, ...

  $\rightarrow$ morphological analysis?

- Compounding

  homework, website, ...

  $\rightarrow$ compound splitting?

- Names

  Netanyahu, Jones, Macron, Hoboken, ...

  $\rightarrow$ transliteration?

$\Rightarrow$ Breaking up words into **subwords** may be a good idea

Byte Pair Encoding

- Start by breaking up words into characters

  `t h e ␣ f a t ␣ c a t ␣ i s ␣ i n ␣ t h e ␣ t h i n ␣ b a g`

- Merge frequent pairs

```
t h→th    th e ␣ f a t ␣ c a t ␣ i s ␣ i n ␣ th e ␣ th i n ␣ b a g
a t→at    th e ␣ f at ␣ c at ␣ i s ␣ i n ␣ th e ␣ th i n ␣ b a g
i n→in    th e ␣ f at ␣ c at ␣ i s ␣ in ␣ th e ␣ th in ␣ b a g
th e→the  the ␣ f at ␣ c at ␣ i s ␣ in ␣ the ␣ th in ␣ b a g
```

- Each merge operation increases the vocabulary size
  - starting with the size of the character set (maybe 100 for Latin script)
  - stopping at, say, 50,000

Example: 49,500 BPE Operations

```
Obama receives Net@@ any@@ ahu

the relationship between Obama and Net@@ any@@ ahu is not exactly
friendly .  the two wanted to talk about the implementation of the
international agreement and about Teheran 's destabil@@ ising activities
in the Middle East .  the meeting was also planned to cover the conflict
with the Palestinians and the disputed two state solution .  relations
between Obama and Net@@ any@@ ahu have been stra@@ ined for years .
Washington critic@@ ises the continuous building of settlements in
Israel and acc@@ uses Net@@ any@@ ahu of a lack of initiative in the
peace process .  the relationship between the two has further
deteriorated because of the deal that Obama negotiated on Iran 's
atomic programme .  in March , at the invitation of the Republic@@ ans
, Net@@ any@@ ahu made a controversial speech to the US Congress , which
was partly seen as an aff@@ ront to Obama .  the speech had not been
agreed with Obama , who had rejected a meeting with reference to the
election that was at that time im@@ pending in Israel .
```

using monolingual data

Traditional View

- Two core objectives for translation

Adequacy	**Fluency**
meaning of source and target match	target is well-formed
translation model	language model
parallel data	monolingual data

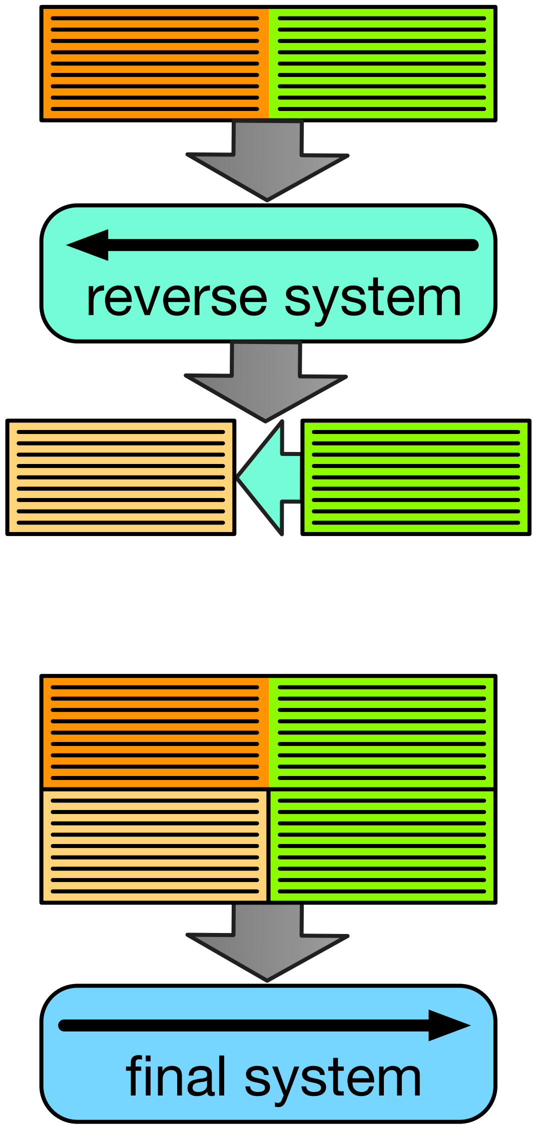
- Language model is key to good performance in statistical models
- But: current neural translation models only trained on parallel data

Integrating a Language Model

- Integrating a language model into neural architecture
  - word prediction informed by translation model and language model
  - gated unit that decides balance
- Use of language model in decoding
  - train language model in isolation
  - add language model score during inference (similar to ensembling)
- Proper balance between models (amount of training data, weights) unclear

Backtranslation

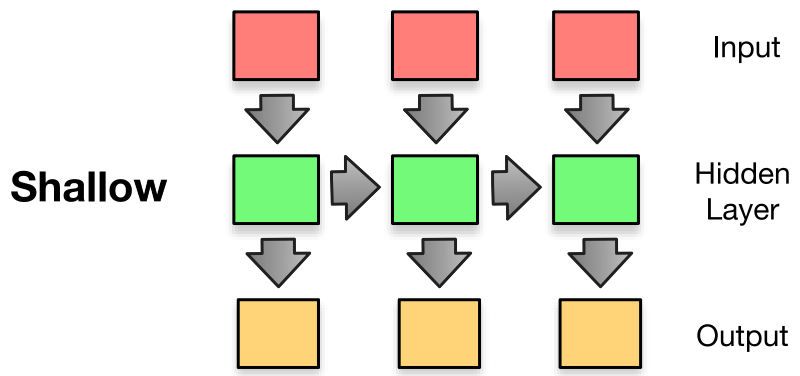
- No changes to model architecture
- Create synthetic parallel data
  - train a system in reverse direction
  - translate target-side monolingual data into source language
  - add as additional parallel data
- Simple, yet effective

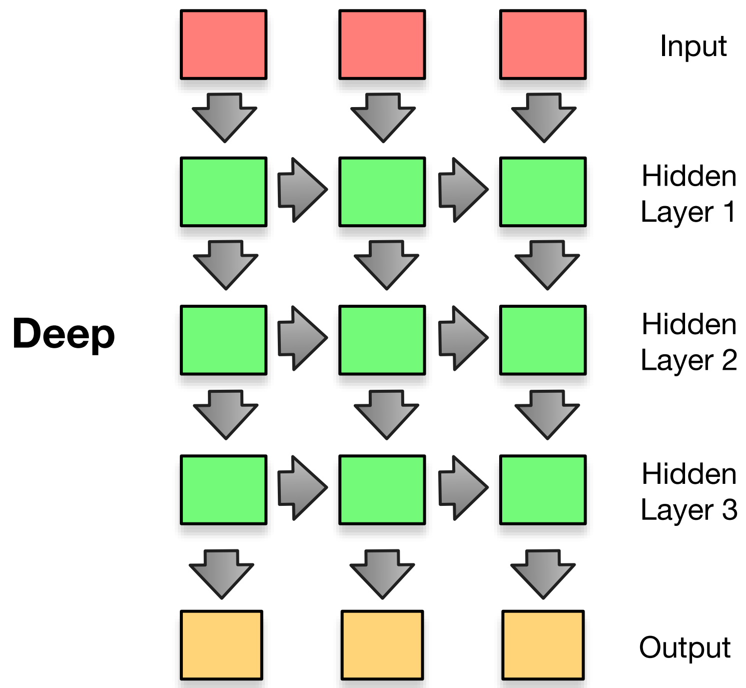


deeper models

Deeper Models

- Encoder and decoder are recurrent neural networks
- We can add additional layers for each step
- Recall shallow and deep language models



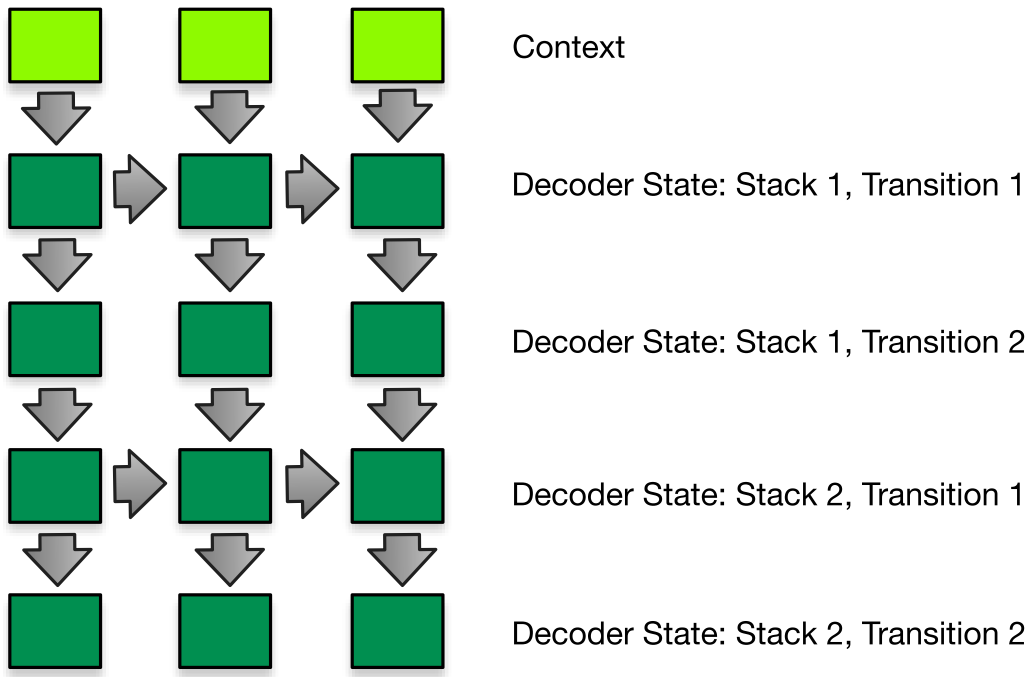


- Adding residual connections (short-cuts through deep layers) help

Deep Decoder

- Two ways of adding layers
  - deep transitions: several layers on path to output
  - deeply stacking recurrent neural networks

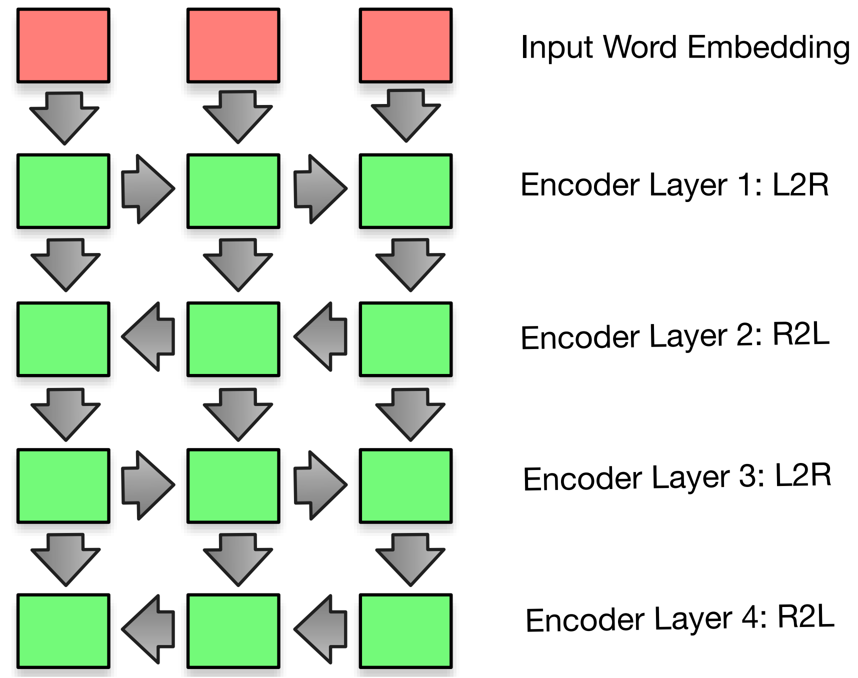
- Why not both?



Deep Encoder

- Previously proposed encoder already has 2 layers
  - left-to-right recurrent network, to encode left context
  - right-to-left recurrent network, to encode right context

⇒ Third way of adding layers



Reality Check: Edinburgh WMT 2017

Table 2: BLEU scores for translating news *into* English (WMT 2016 and 2017 test sets – WMT 2017 dev set is used where there was no 2016 test)

	CS→EN		DE→EN		LV→EN		RU→EN		TR→EN		ZH→EN	
system	**2016**	**2017**	**2016**	**2017**	**2017d**	**2017**	**2016**	**2017**	**2016**	**2017**	**2017d**	**2017**
WMT-16 single system	30.1	25.9	36.2	31.1	—	—	26.9	29.6	—	—	—	—
baseline	31.7	27.5	38.0	32.0	23.5	16.4	27.8	31.3	20.2	19.7	19.9	21.7
+layer normalization	32.6	28.2	38.6	32.1	24.4	17.0	28.8	32.3	19.5	18.8	20.8	22.5
+deep model	33.2	28.9	39.6	33.5	24.4	16.6	29.0	32.7	20.6	20.6	22.1	22.9
+checkpoint ensemble	33.8	29.4	39.7	33.8	25.7	17.7	29.5	33.3	20.6	21.0	22.5	23.6
+independent ensemble	34.6	30.3	40.7	34.4	27.5	18.5	29.8	33.6	22.1	21.6	23.4	25.1
+right-to-left reranking	35.6	31.1	41.0	35.1	28.0	19.0	30.5	34.6	22.9	22.3	24.0	25.7
WMT-17 submission[a]	—	**30.9**	—	**35.1**	—	**19.0**	—	**30.8**	—	**20.1**	—	**25.7**

[a] In some cases training did not converge until after the submission deadline. The contrastive/ablative results shown were obtained with the converged systems; this line reports the BLEU score for the system output submitted by the submission deadline.

Table 3: BLEU scores for translating news *out of* English (WMT 2016 and 2017 test sets – WMT 2017 dev set is used where there was no 2016 test)

	EN→CS		EN→DE		EN→LV		EN→RU		EN→TR		EN→ZH	
system	**2016**	**2017**	**2016**	**2017**	**2017d**	**2017**	**2016**	**2017**	**2016**	**2017**	**2017d**	**2017**
WMT16 single system	23.7	19.7	31.6	24.9	—	—	24.3	26.7	—	—	—	—
baseline	23.5	20.5	32.2	26.1	20.8	14.6	25.2	28.0	13.8	15.6	30.5	31.3
+layer normalization	23.3	20.5	32.5	26.1	21.6	14.9	25.8	28.7	14.0	15.7	31.6	32.3
+deep model	24.1	21.1	33.9	26.6	22.3	15.1	26.5	29.9	14.4	16.2	32.6	33.4
+checkpoint ensemble	24.7	22.0	33.9	27.5	23.4	16.1	27.3	31.0	15.0	16.7	32.8	33.5
+independent ensemble	26.4	22.8	35.1	28.3	24.7	16.7	28.2	31.6	15.5	17.6	35.4	35.8
+right-to-left reranking	26.7	22.8	36.2	28.3	25.0	16.9	–	–	16.1	18.1	35.7	36.3
WMT-17 submission[a]	–	**22.8**	–	**28.3**	–	**16.9**	–	**29.8**	–	**16.5**	–	**36.3**

[a] In some cases training did not converge until after the submission deadline. The contrastive/ablative results shown were obtained with the converged systems; this line reports the BLEU score for the system output submitted by the submission deadline.

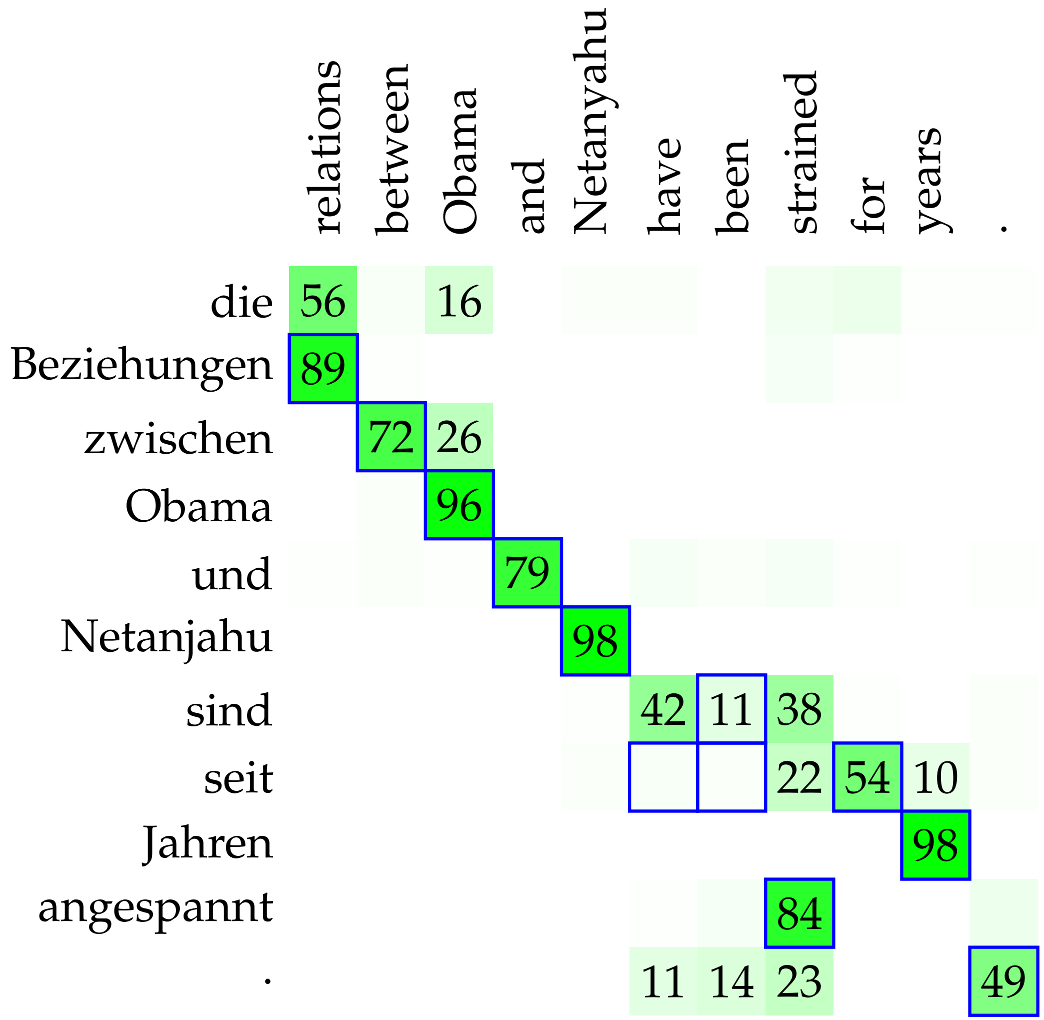
alignment and coverage

Alignment

- Attention model fulfills role of alignment

- Traditional methods for word alignment
  - based on co-occurence, word position, etc.
  - expectation maximization (EM) algorithm
  - popular: IBM models, fast-align

Attention vs. Alignment



Guided Alignment

- Guided alignment training for neural networks
  - traditional objective function: match output words
  - now: also match given word alignments
- Add as cost to objective function
  - given alignment matrix A, with $\sum_j A_{ij} = 1$ (from IBM Models)
  - computed attention α_{ij} (also $\sum_j \alpha_{ij} = 1$ due to softmax)
  - added training objective (cross-entropy)

$$\text{cost}_{\text{CE}} = -\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij} \log \alpha_{ij}$$

Coverage

	in	order	to	solve	the	problem	,	the	"	Social	Housing	"	alliance	suggests	a	fresh	start	.
um	37		33					13										
das					84													
Problem					10	80												
zu				63			12											
lösen				81														
,							40							30				
schlägt														80				
das								71	18									
Unternehmen									86			10						
der									84									
Gesellschaft									80									
für									45	41								
soziale									40	44	10							
Bildung											89							
vor									12			40	10					
.									10			37			11	13		
	43	7	46	161	108	89	62	112	392	121	110	130	26	132	22	19	6	6

Tracking Coverage

- Neural machine translation may drop or duplicate content
- Track coverage during decoding

$$\text{coverage}(j) = \sum_i \alpha_{i,j}$$

$$\text{over-generation} = \max\Big(0, \sum_j \text{coverage}(j) - 1\Big)$$

$$\text{under-generation} = \min\Big(1, \sum_j \text{coverage}(j)\Big)$$

- Add as cost to hypotheses

Coverage Models

- Use as information for state progression

$$a(s_{i-1}, h_j) = W^a s_{i-1} + U^a h_j + V^a \text{coverage}(j) + b^a$$

- Add to objective function

$$\log \sum_i P(y_i|x) + \lambda \sum_j (1 - \text{coverage}(j))^2$$

- May also model fertility
  - some words are typically dropped
  - some words produce multiple output words

linguistic annotation

Example

Words	*the*	*girl*	*watched*	*attentively*	*the*	*beautiful*	*fireflies*
Part of speech	DET	NN	VFIN	ADV	DET	JJ	NNS
Lemma	*the*	*girl*	*watch*	*attentive*	*the*	*beautiful*	*firefly*
Morphology	-	SING.	PAST	-	-	-	PLURAL
Noun phrase	BEGIN	CONT	OTHER	OTHER	BEGIN	CONT	CONT
Verb phrase	OTHER	OTHER	BEGIN	CONT	CONT	CONT	CONT
Synt. dependency	*girl*	*watched*	-	*watched*	*fireflies*	*fireflies*	*watched*
Depend. relation	DET	SUBJ	-	ADV	DET	ADJ	OBJ
Semantic role	-	ACTOR	-	MANNER	-	MOD	PATIENT
Semantic type	-	HUMAN	VIEW	-	-	-	ANIMATE

Input Annotation

- Input words are encoded in one-hot vectors

- Additional linguistic annotation
  - part-of-speech tag
  - morphological features
  - etc.

- Encode each annotation in its own one-hot vector space

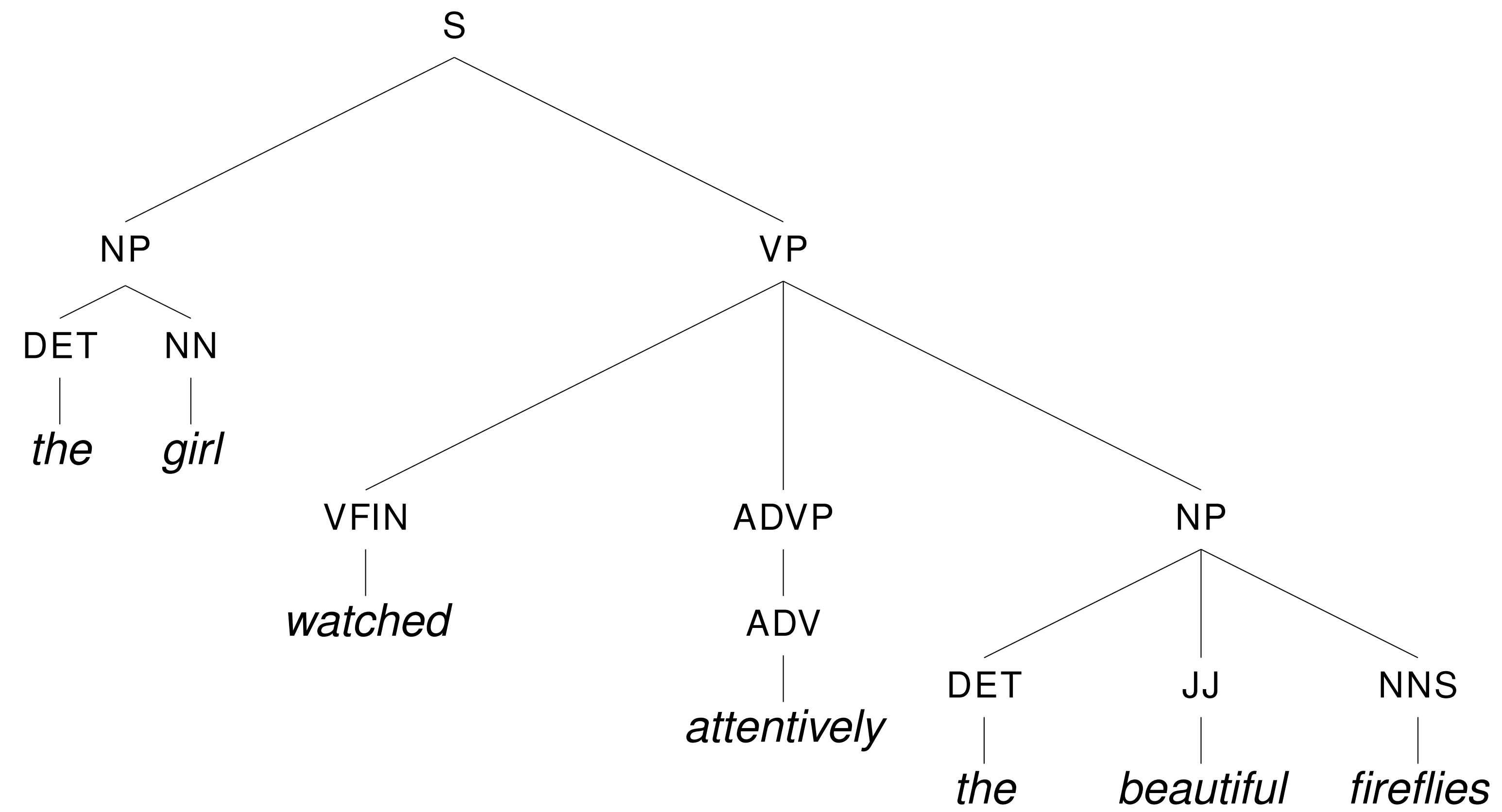
- Concatenate one-hot vecors

- Essentially:
  - each annotation maps to embedding
  - embeddings are added

Output Annotation

- Same can be done for output
- Additional output annotation is latent feature
  - ultimately, we do not care if right part-of-speech tag is predicted
  - only right output words matter
- Optimizing for correct output annotation $\rightarrow$ better prediction of output words

Linearized Output Syntax

Sentence	*the girl watched attentively the beautiful fireflies*
Syntax tree	
Linearized	(S (NP (DET *the*) (NN *girl*)) (VP (VFIN *watched*) (ADVP (ADV *attentively*)) (NP (DET *the*) (JJ *beautiful*) (NNS *fireflies*))))

multiple language pairs

One Model, Multiple Language Pairs

- One language pair $\rightarrow$ train one model
- Multiple language pairs $\rightarrow$ train one model for each
- Multiple language pair $\rightarrow$ train one model for all

Multiple Input Languages

- Given
  - French–English corpus
  - German–English corpus
- Train one model on concatenated corpora
- Benefit: sharing monolingual target language data

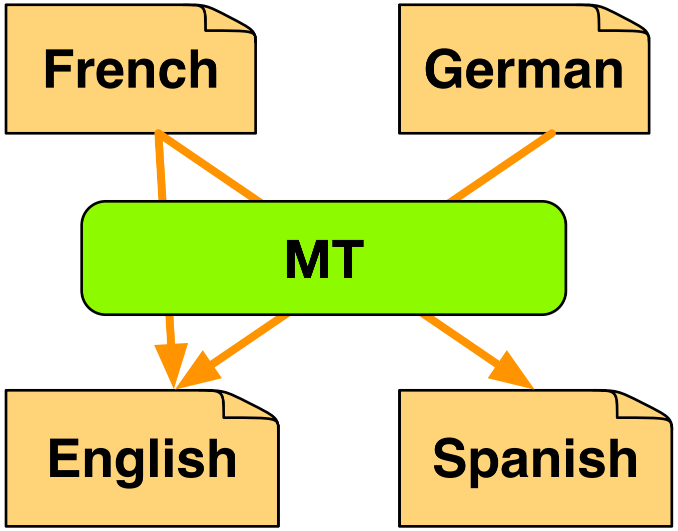
Multiple Output Languages

- Multiple output languages
  - French–English corpus
  - French–Spanish corpus
- Need to mark desired output language with special token

[ENGLISH] *N'y a-t-il pas ici deux poids, deux mesures?*
$\Rightarrow$ *Is this not a case of double standards?*

[SPANISH] *N'y a-t-il pas ici deux poids, deux mesures?*
$\Rightarrow$ *No puede verse con toda claridad que estamos utilizando un doble rasero?*

Zero Shot



- Can the model translate German to Spanish?

[SPANISH] *Messen wir hier nicht mit zweierlei Maß?*
$\Rightarrow$ *No puede verse con toda claridad que estamos utilizando un doble rasero?*

Zero Shot: Vision

- Direct translation only requires bilingual mapping
- Zero shot requires interlingual representation

Algorithms

Google's AI just created its own universal 'language'

The technology used in Google Translate can identify hidden material between languages to create what's known as interlingua

By **MATT BURGESS**
23 Nov 2016

Zero Shot: Reality

Table 5: Portuguese→Spanish BLEU scores using various models.

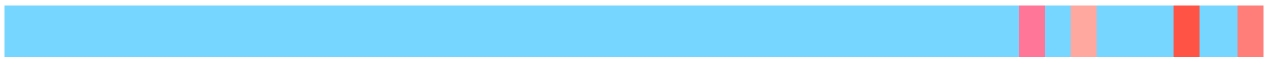
	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

ensembling

Ensembling

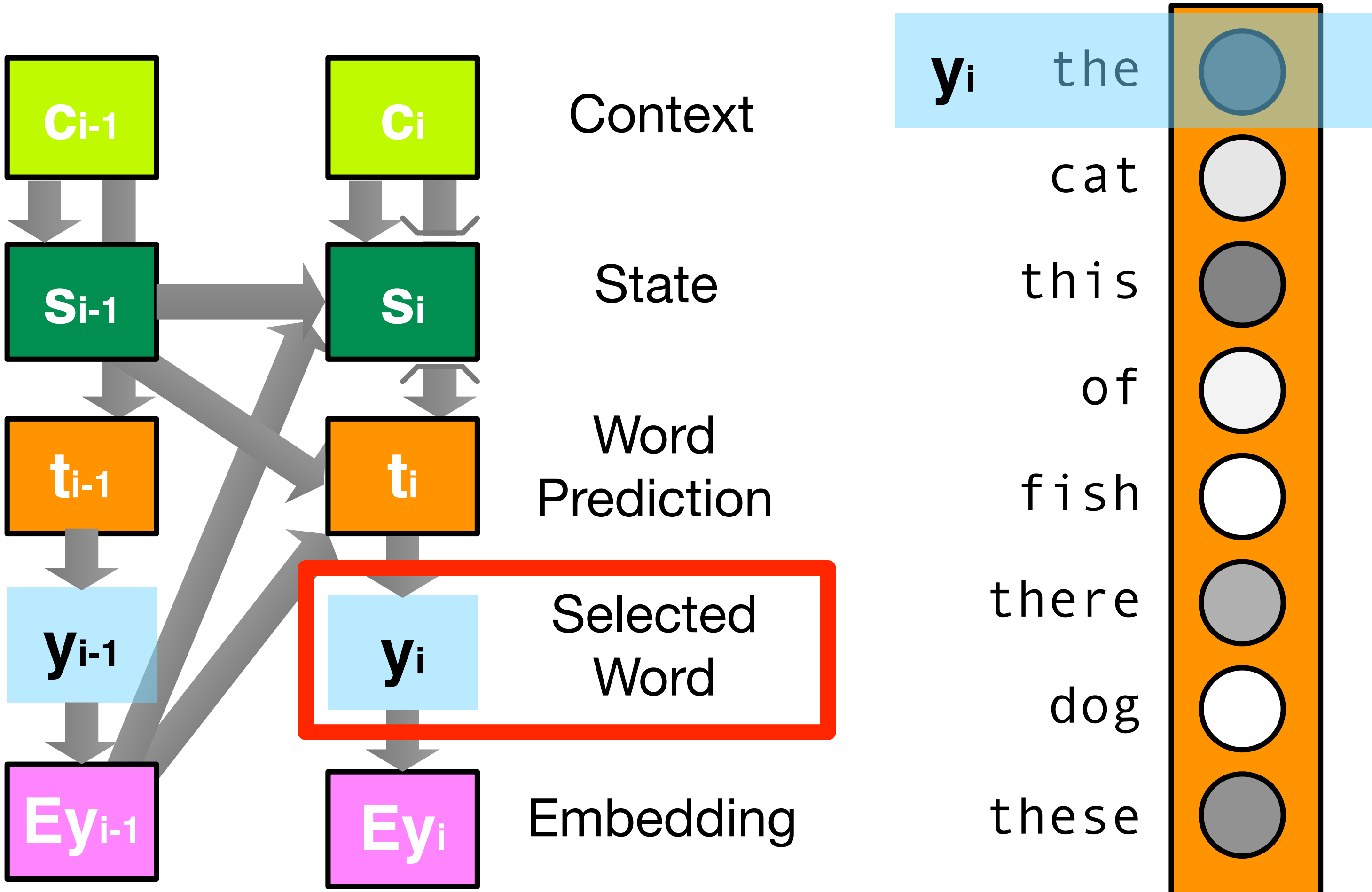
- Train multiple models
- Say, by different random initializations

- Or, by using model dumps from earlier iterations

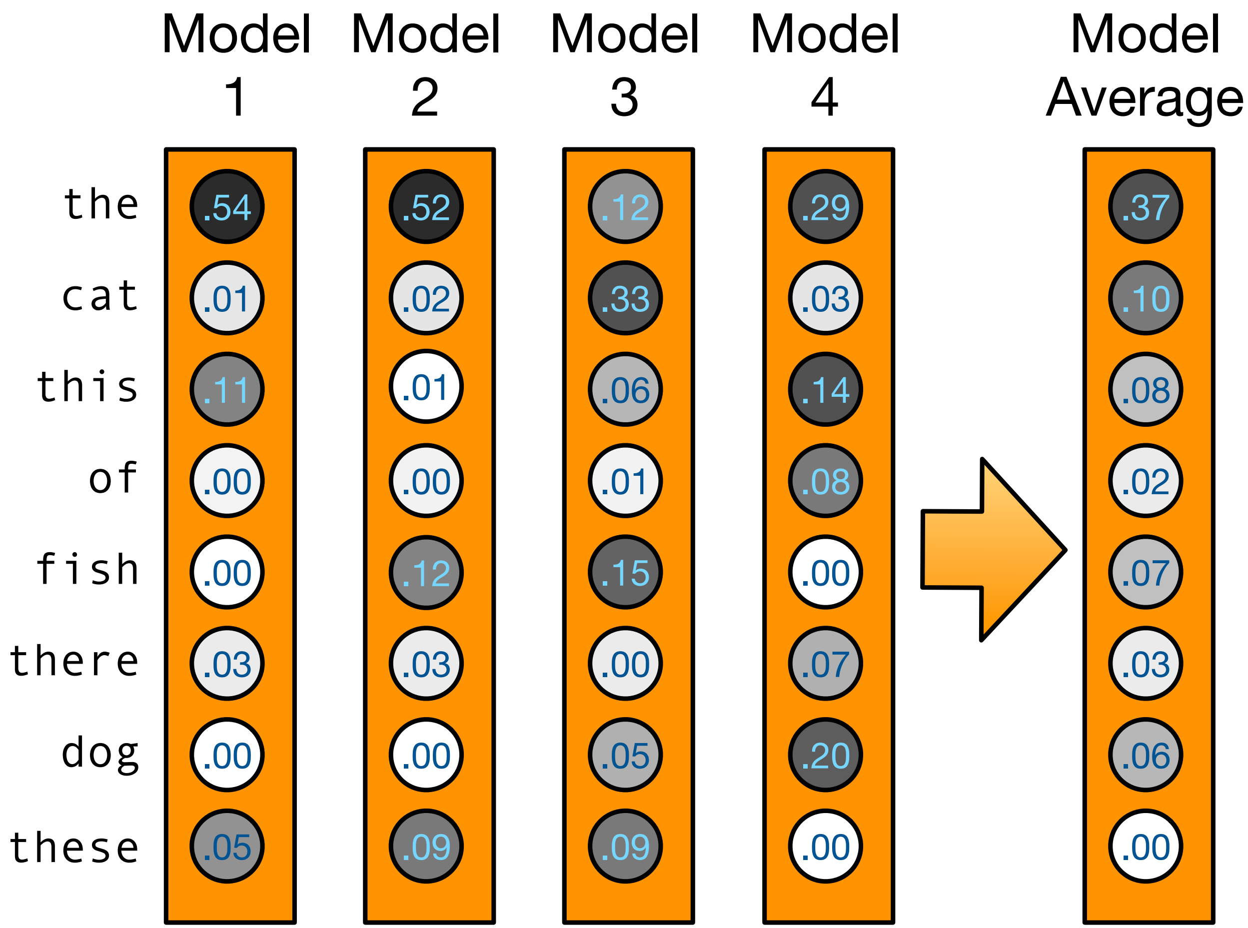


(most recent, or interim models with highest validation score)

Decoding with Single Model



Combine Predictions



Ensembling

- Surprisingly reliable method in machine learning
- Long history, many variants:
  bagging, ensemble, model averaging, system combination, ...
- Works because errors are random, but correct decisions unique

Right-to-Left Inference

- Neural machine translation generates words right to left (L2R)

$$\text{the} \rightarrow \text{cat} \rightarrow \text{is} \rightarrow \text{in} \rightarrow \text{the} \rightarrow \text{bag} \rightarrow .$$

- But it could also generate them right to left (R2L)

$$\text{the} \leftarrow \text{cat} \leftarrow \text{is} \leftarrow \text{in} \leftarrow \text{the} \leftarrow \text{bag} \leftarrow .$$

Obligatory notice: Some languages (Arabic, Hebrew, ...) have writing systems that are right-to-left, so the use of ”right-to-left” is not precise here.

Right-to-Left Reranking

- Train both L2R and R2L model
- Score sentences with both
  $\Rightarrow$ use both left and right context during translation
- Only possible once full sentence produced $\rightarrow$ re-ranking
  1. generate n-best list with L2R model
  2. score candidates in n-best list with R2L model
  3. chose translation with best average score

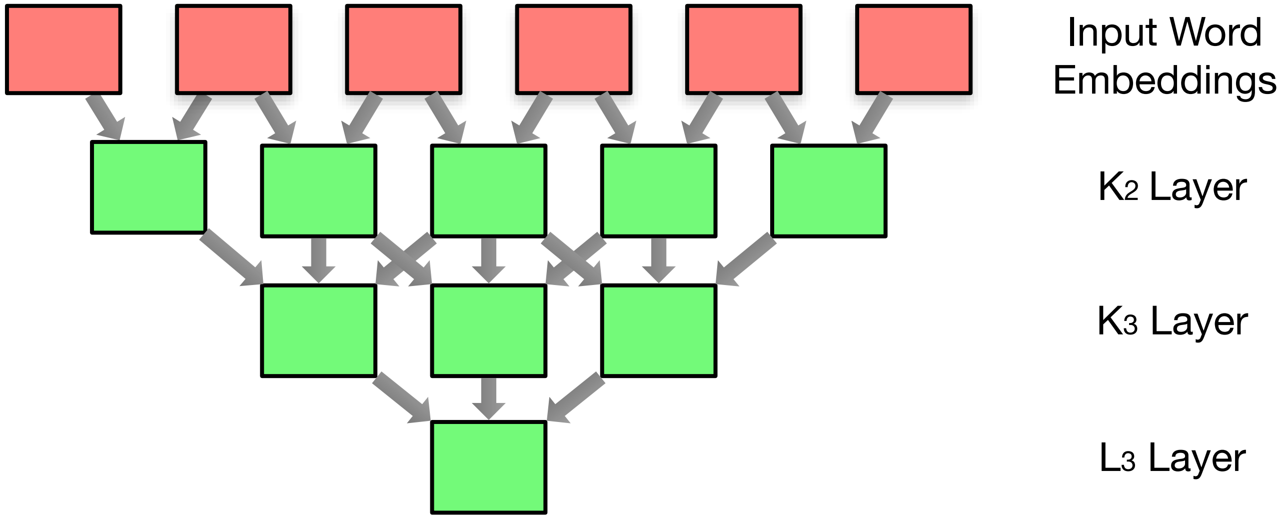
ALTERNATIVE ARCHITECTURES FOR NEURAL MACHINE TRANSLATION

Beyond Recurrent Neural Networks

- We presented the currently dominant model
  - recurrent neural networks for encoder and decoder
  - attention
- Convolutional neural networks
- Self attention

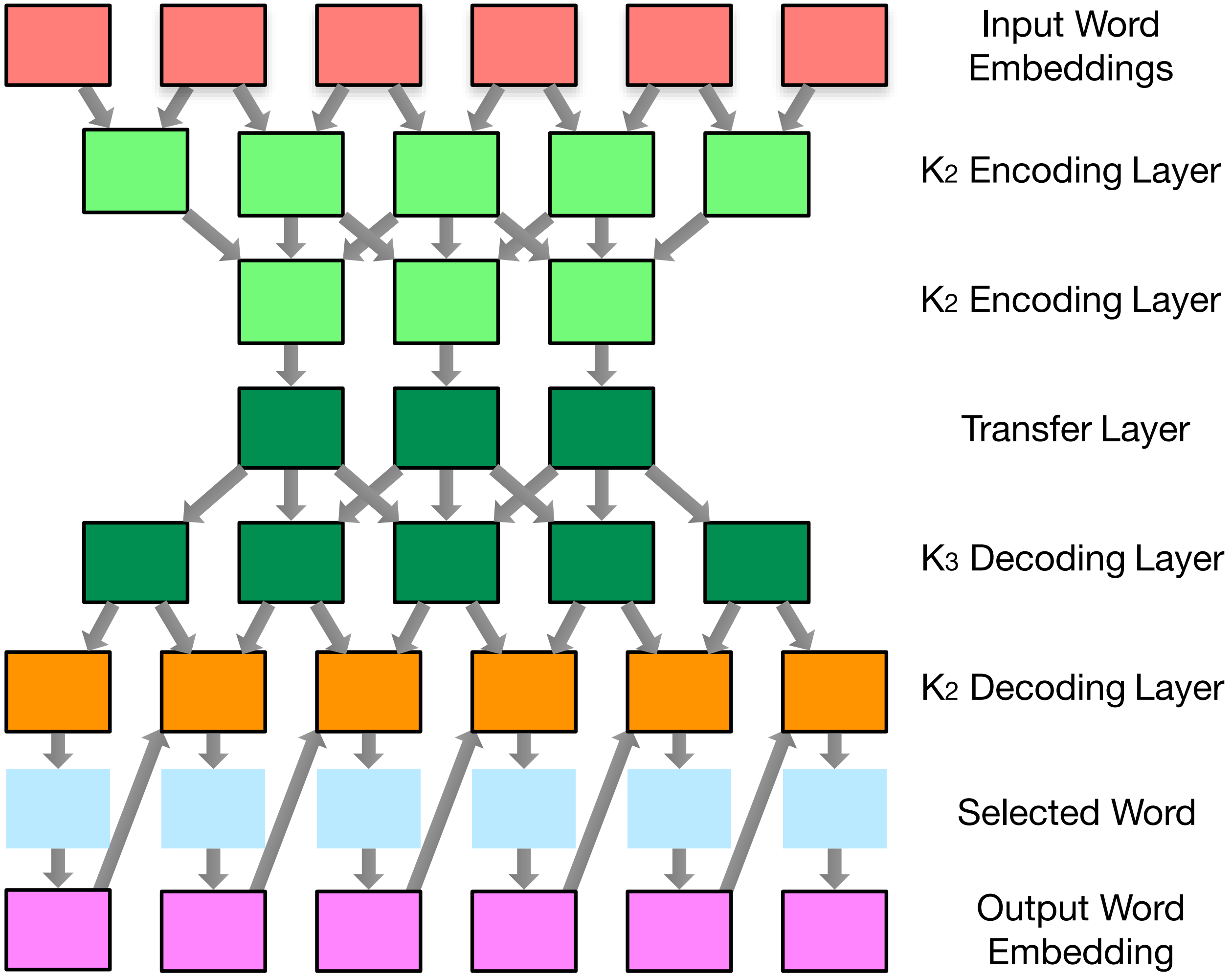
convolutional neural networks

Convolutional Neural Networks



- Build sentence representation bottom-up
  - merge any n neighboring nodes
  - n may be 2, 3, ...

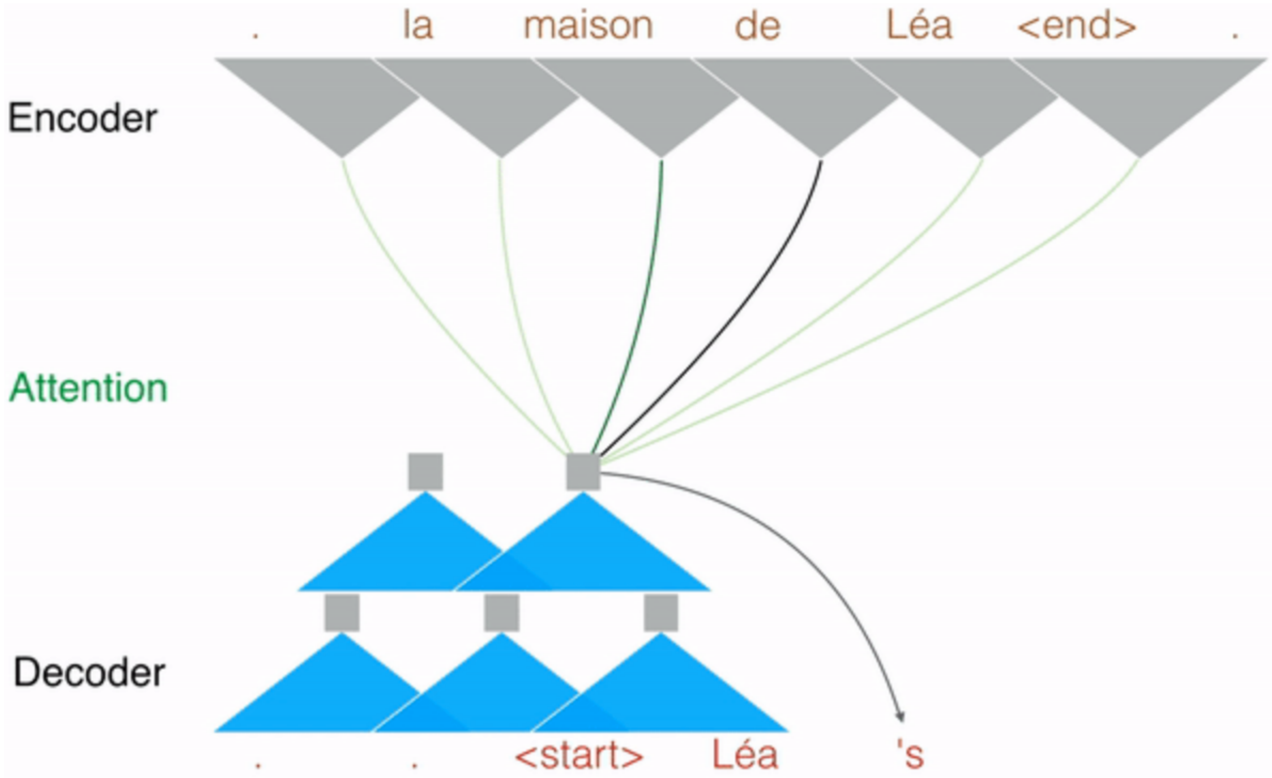
Generation


Input Word Embeddings
K2 Encoding Layer
K2 Encoding Layer
Transfer Layer
K3 Decoding Layer
K2 Decoding Layer
Selected Word
Output Word Embedding

Generation

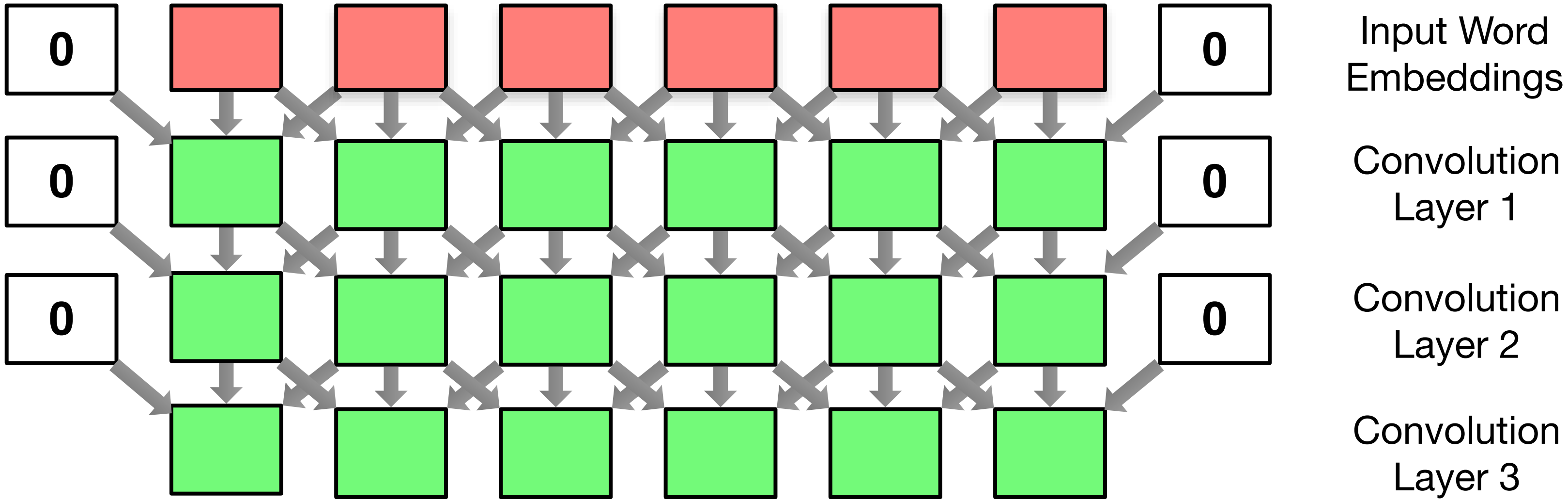
- Encode with convolutional neural network
- Decode with convolutional neural network
- Also include a linear recurrent neural network
- Important: predict length of output sentence
- Does it work?
  used successfully in re-ranking (Cho et al., 2014)

Convolutional Network with Attention



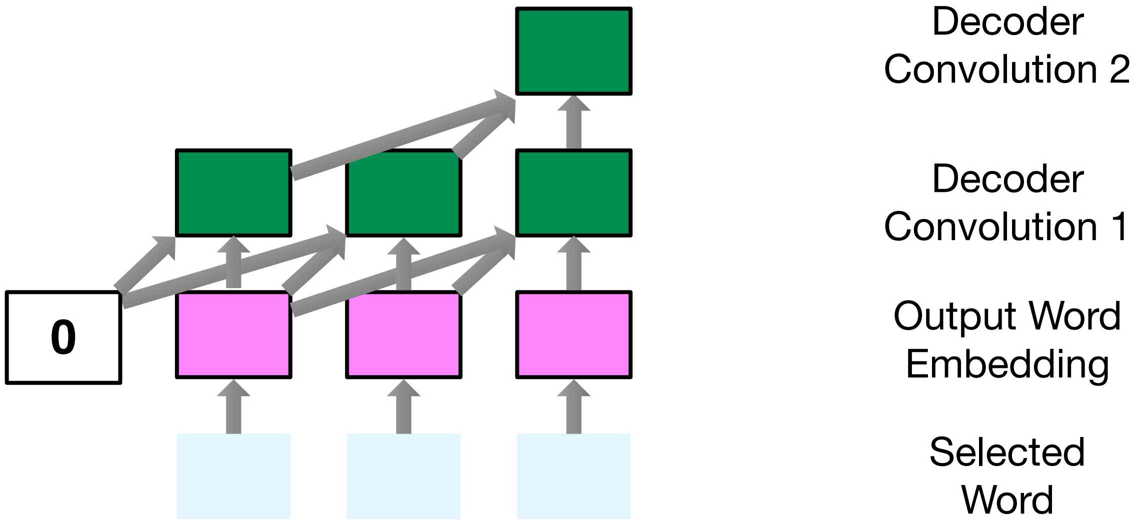
(Facebook, 2017)

Convolutional Encoder



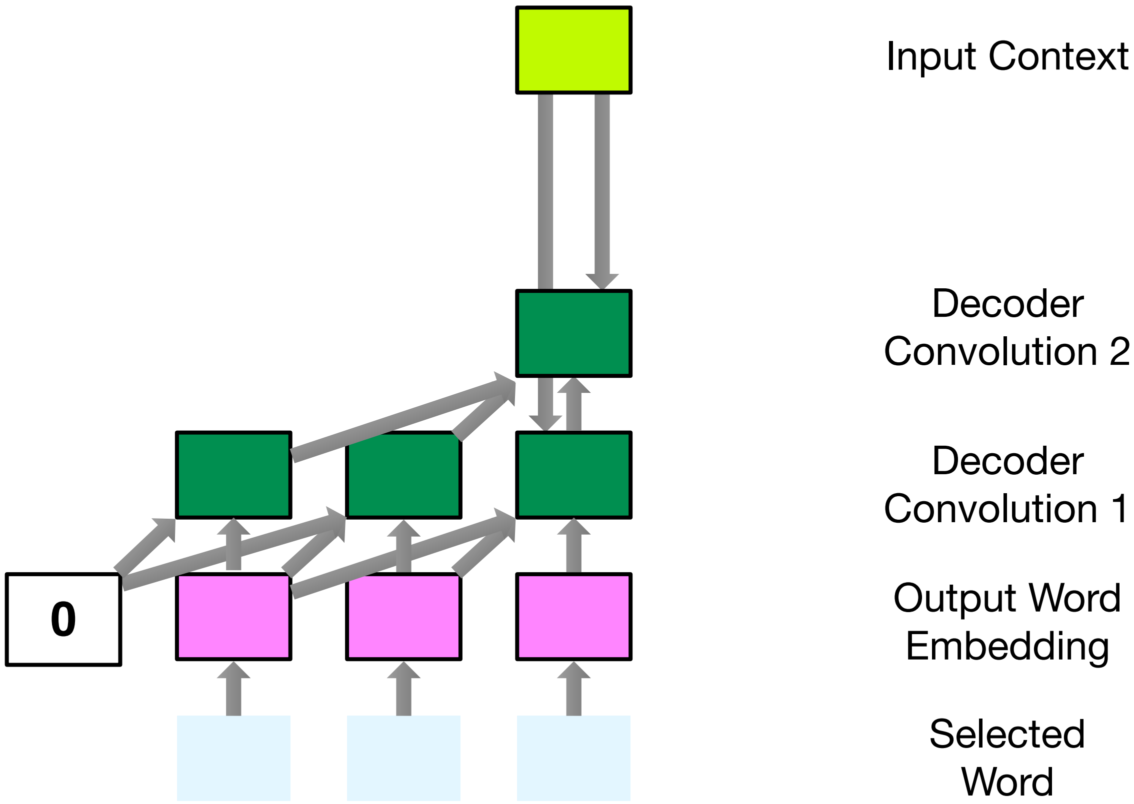
- Similar idea as deep recurrent neural networks
- Good: more parallelizable
- Bad: less context when refining representation of a word

Convolutional Decoder



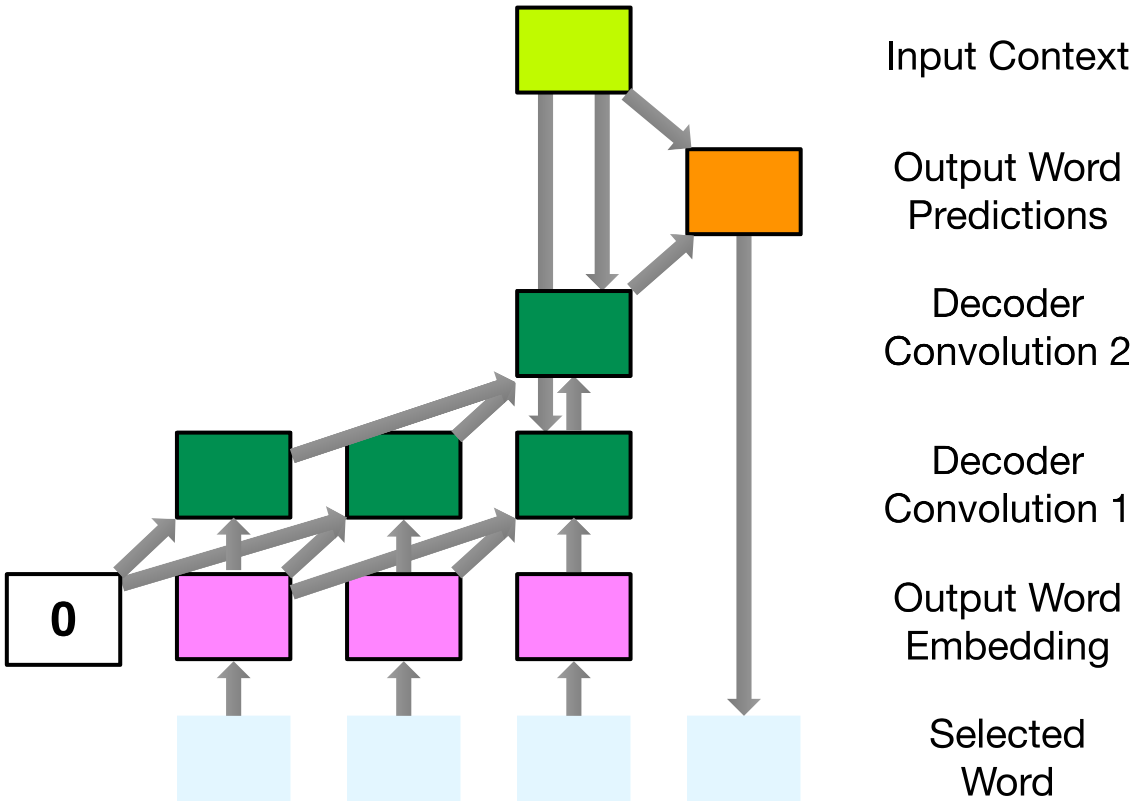
- Convolutions over output words
- Only previously produced output words
  (still left-to-right decoding)

Convolutional Decoder



- Inclusion of Input context
- Context result of attention mechanism (similar to previous)

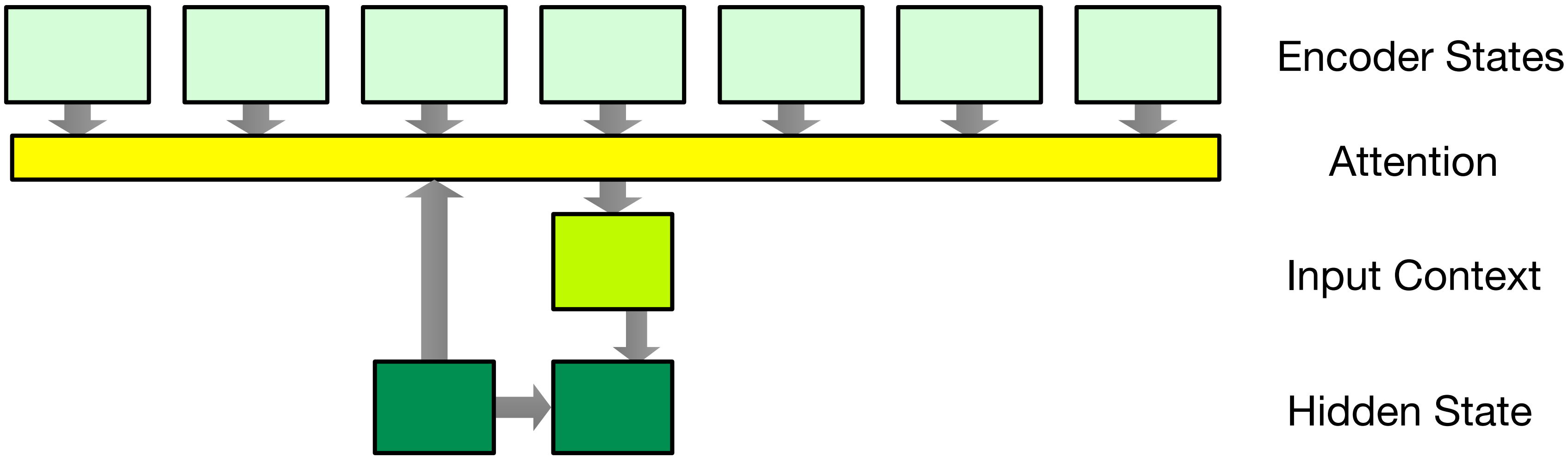
Convolutional Decoder



- Predict output word distribution
- Select output word

self-attention

Attention



- Compute association between last hidden state and encoder states

Attention Math

- Input word representation h_k
- Decoder state s_j
- Computations

$a_{jk} = \frac{1}{\|h\|} s_j h_k^T$	raw association
$\alpha_{jk} = \frac{\exp(a_{jk})}{\sum_\kappa \exp(a_{j\kappa})}$	normalized association (softmax)
$\text{self-attention}(h_j) = \sum_k \alpha_{j\kappa} h_k$	weighted sum

Self-Attention

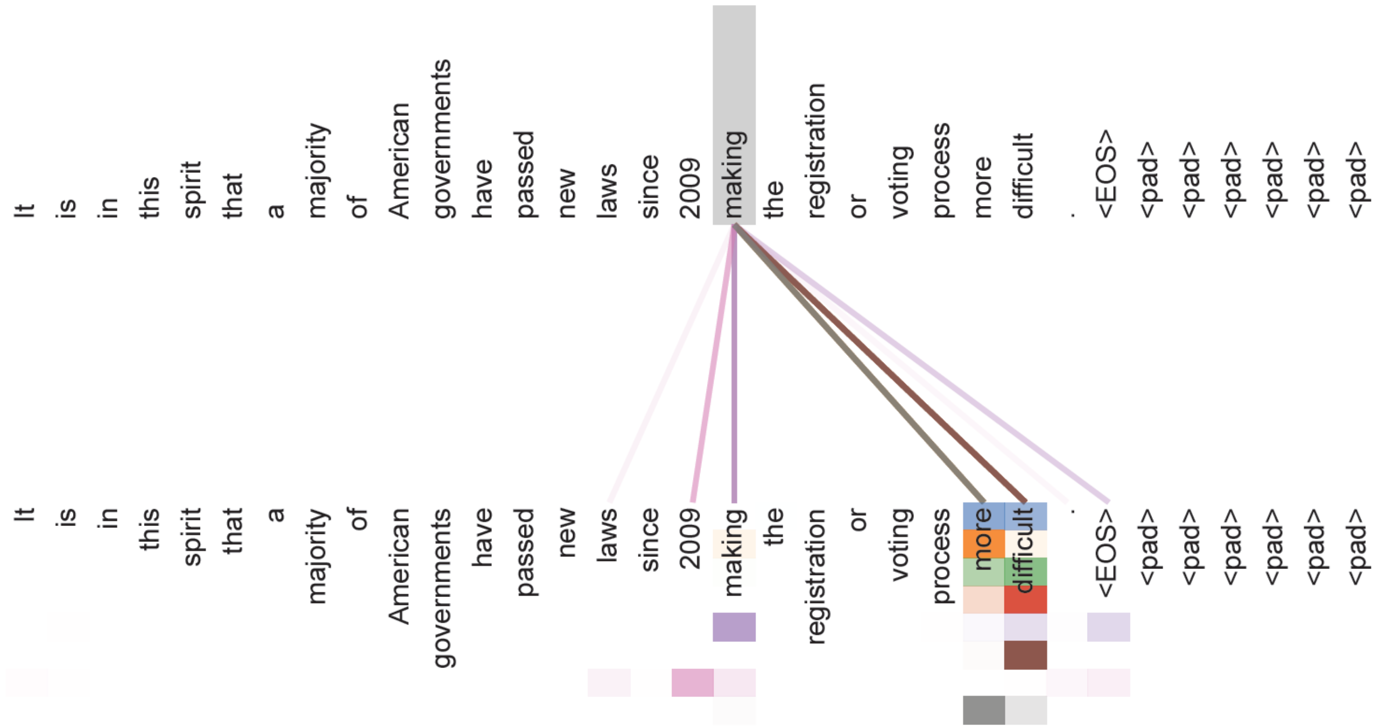
- Attention

$$a_{jk} = \frac{1}{|h|} s_j h_k^T$$

- Self-attention

$$a_{jk} = \frac{1}{|h|} h_j h_k^T$$

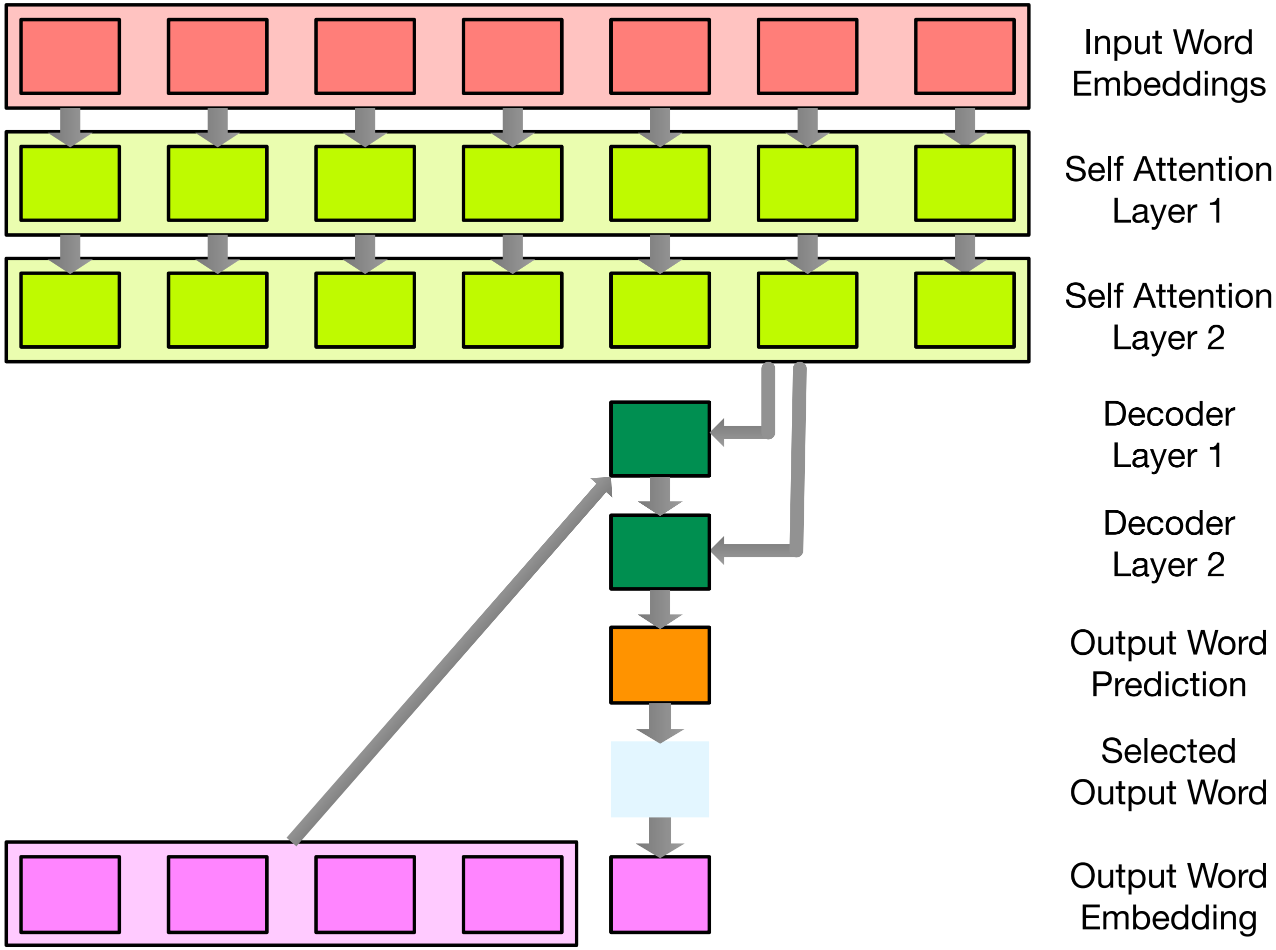
Why?



- Refine representation of word with related words

  making ... more difficult refines *making*
- Good: more parallelizable than recurrent neural network
- Good: wide context when refining representation of a word

Stacked Attention in Decoder



Where Are We Now?

- Recurrent neural network with attention currently dominant model
- Still many challenges
- New proposals in Spring 2017
  - convolutions (Facebook)
  - self-attention (Google)
- Too early to tell if either becomes the new paradigm
- Open source implementations are available

questions?