Natural Language Processing

Info 159/259

Lecture 20: Machine Translation (April 8, 2024)

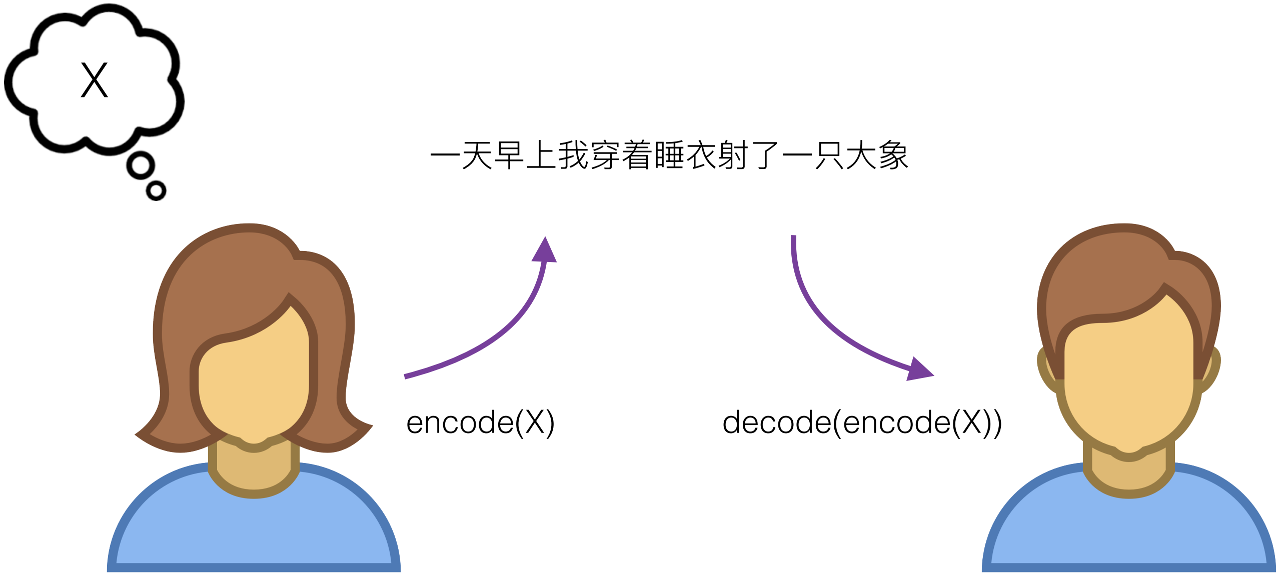
Many slides & instruction ideas borrowed from:

David Bamman, Greg Durrett, Kemal Oflazer, Graham Neubig & Maha Elbayad,

Logistics

- Quiz 8: Due tonight
- AP2: due tomorrow night (Tues April 9)
- 259 Mid-project report due this Thursday night (April 11)
- Homework 6 will be released this week

Machine Translation

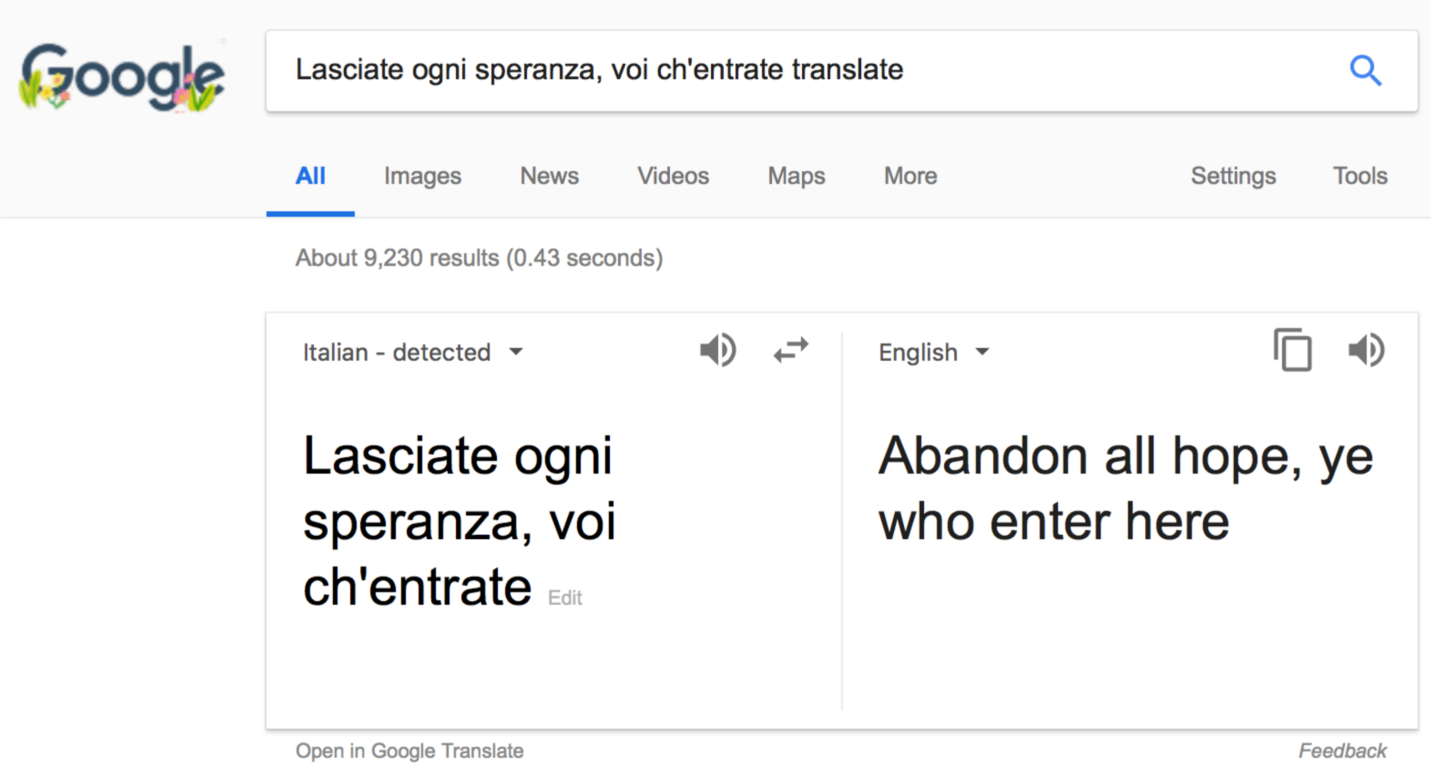


When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Weaver 1955

Machine Translation

Task	X	Y
Sentiment analysis	I hate this movie!	negative
POS tagging	I hate this movie!	PRP VB DT NN .
Parsing	I hate this movie!	[tree]
MT	Lasciate ogni speranza, voi ch'entrate	Abandon all hope, you who enter!
Conversational agent	How are you?	I’m great!



Lasciate ogni speranza, voi ch'entrate translate

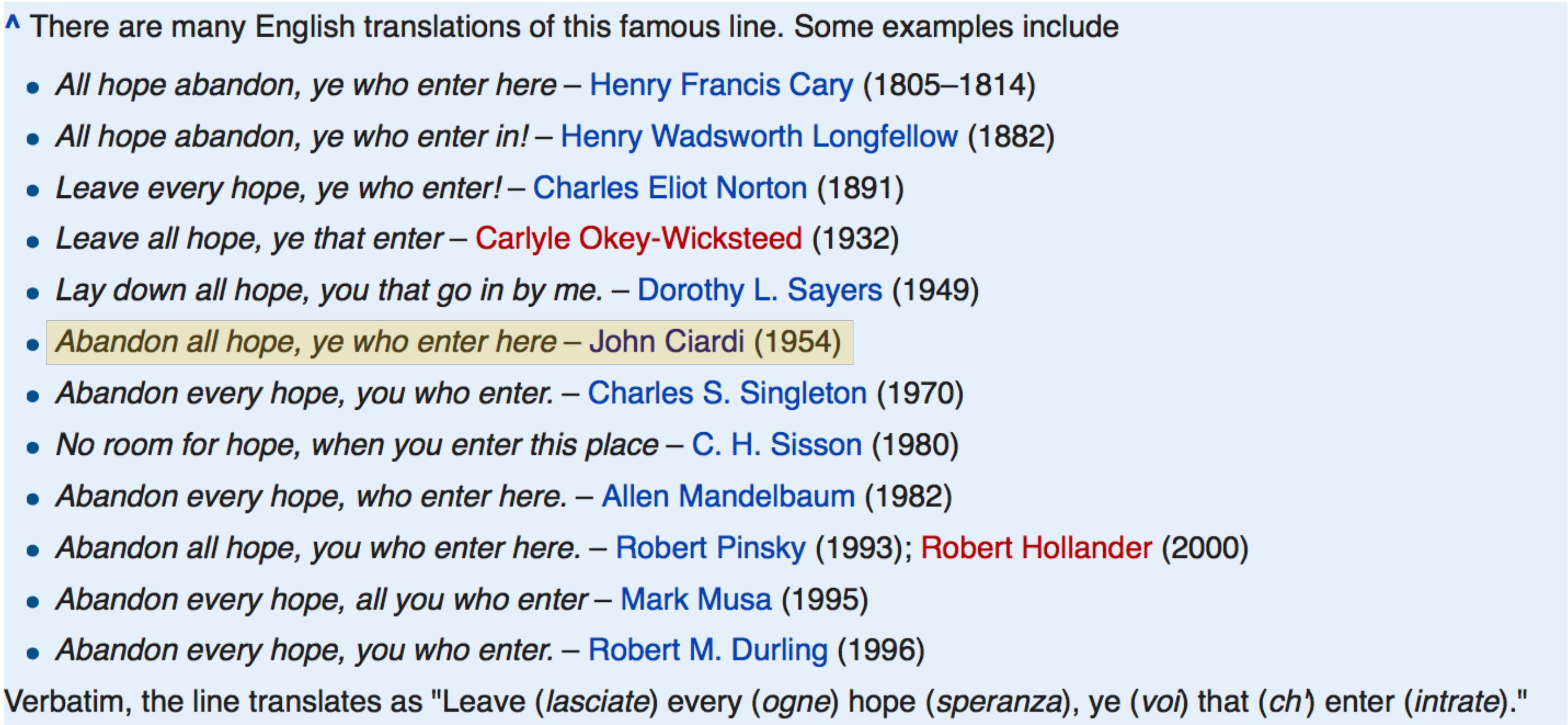
About 9,230 results (0.43 seconds)

Italian - detected

English

Lasciate ogni speranza, voi ch'entrate

Abandon all hope, ye who enter here



^ There are many English translations of this famous line. Some examples include

- *All hope abandon, ye who enter here* – Henry Francis Cary (1805–1814)
- *All hope abandon, ye who enter in!* – Henry Wadsworth Longfellow (1882)
- *Leave every hope, ye who enter!* – Charles Eliot Norton (1891)
- *Leave all hope, ye that enter* – Carlyle Okey-Wicksteed (1932)
- *Lay down all hope, you that go in by me.* – Dorothy L. Sayers (1949)
- *Abandon all hope, ye who enter here* – John Ciardi (1954)
- *Abandon every hope, you who enter.* – Charles S. Singleton (1970)
- *No room for hope, when you enter this place* – C. H. Sisson (1980)
- *Abandon every hope, who enter here.* – Allen Mandelbaum (1982)
- *Abandon all hope, you who enter here.* – Robert Pinsky (1993); Robert Hollander (2000)
- *Abandon every hope, all you who enter* – Mark Musa (1995)
- *Abandon every hope, you who enter.* – Robert M. Durling (1996)

Verbatim, the line translates as "Leave (*lasciate*) every (*ogne*) hope (*speranza*), ye (*voi*) that (*ch'*) enter (*intrate*)."

https://en.wikipedia.org/wiki/Inferno_(Dante)#cite_note-18

Data

- Modern machine translation systems are learned from parallel texts: pairs of documents in two languages that have been aligned at the sentence level.

Reprise de la session	Resumption of the session
Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.
Comme vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas produit. En revanche, les citoyens d'un certain nombre de nos pays ont été victimes de catastrophes naturelles qui ont vraiment été terribles.	Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.

European Parliament Proceedings Parallel Corpus 1996-2011

http://www.statmt.org/europarl/

Data

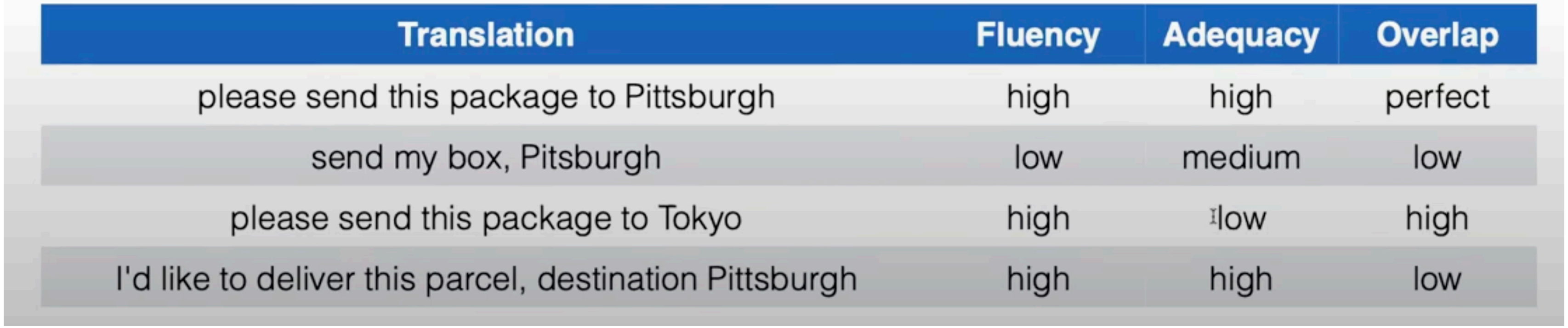
- Europarl (proceedings of European parliament, 50M words/language)
http://www.statmt.org/europarl/

- UN Corpus (United Nations documents, six languages, 300M words/langauge)
http://www.euromatrixplus.net/multi-un/

- Common crawl (Web documents, long tail of language pairs)

MT Evaluation

- **Adequacy**: conveying the meaning of the source sent in the target side.
- **Fluency**: Is the sentence grammatical in the target side?
- Automatic Evaluation: Measure the overlap between MT output and human references (BLEU score) or estimate semantic the semantic overlap (BERT).



Translation	Fluency	Adequacy	Overlap
please send this package to Pittsburgh	high	high	perfect
send my box, Pitsburgh	low	medium	low
please send this package to Tokyo	high	low	high
I'd like to deliver this parcel, destination Pittsburgh	high	high	low

Evaluation

ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον

• Tell me Muse, of the man of many ways	Lattimore 1965
• Sing to me of the man, Muse, the man of twists and turns	Fagles 1996
• Tell me about a complicated man	Wilson 2018

Evaluation

- BLEU (Papineni et al. 2002): what fraction of {1-4}-grams in the system translation appear in the reference translations?

$$p_n = \frac{\textbf{Number of ngram tokens in system and reference translations}}{\textbf{Number of ngram tokens in system translation}}$$

$$\textbf{BLEU} = BP \times \exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$$

Hypothesis translation
Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Reference translations
Orejuela appeared sedate as he was led to the American plane which will take him to Miami, Florida.
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Appeared
calm
when
he
was
taken
to
the
American

plane
,
which
will
to
Miami
,
Florida
.

$$p_1 = \frac{15}{18} = 0.833$$

Ngrams appearing >1 time in the hypothesis can match up to the max number of times they appear in a single reference — e.g., two commas in hypothesis but one max in any single reference.

Callison-Burch et al. (2006), Re-evaluating the Role of BLEU in Machine Translation Research

Hypothesis translation
Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Reference translations
Orejuela appeared sedate as he was led to the American plane which will take him to Miami, Florida.
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Appeared calm
calm when
when he
he was
was taken
taken to
to the
the American
American plane

plane ,
, which
which will
will to
to Miami
Miami ,
, Florida
Florida .

$$p_2 = \frac{10}{17} = 0.588$$

Callison-Burch et al. (2006), Re-evaluating the Role of BLEU in Machine Translation Research

$$\textbf{BLEU} = BP \times \exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$$

$$p_n = \frac{\textbf{Number of ngram tokens in system and reference translations}}{\textbf{Number of ngram tokens in system translation}}$$

- We could optimize the score by minimizing the denominator (the number of ngrams generated)
- Brevity penalty: $$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases}$$

- c = length of hypothesis translation
- r = length of the reference translation whose length is the closest to c

Evaluation

- BLEU (Papineni et al. 2002): what fraction of {1-4}-grams in the system translation appear in the reference translations?

Statistical MT

Noisy Channel

	X	Y
ASR	speech signal	transcription
MT	target text	source text
OCR	pixel densities	transcription

$$P(Y \mid X) \propto \underbrace{P(X \mid Y)}_{\text{channel model}} \underbrace{P(Y)}_{\text{source model}}$$

Noisy Channel

This the translation model

This is just a language model for the target language

$$P(Y \mid X) \propto \underbrace{P(X \mid Y)}_{\text{channel model}} \underbrace{P(Y)}_{\text{source model}}$$

- If we’re translating from English (X) into French (Y) we assume some true French sentence Y that was “corrupted” into English version X.

Noisy Channel

This the translation model

This is just a language model for the target language

$$P(Y \mid X) \propto \underbrace{P(X \mid Y)}_{\text{channel model}} \underbrace{P(Y)}_{\text{source model}}$$

Estimate this from parallel texts

Estimate this from monolingual data

Statistical MT

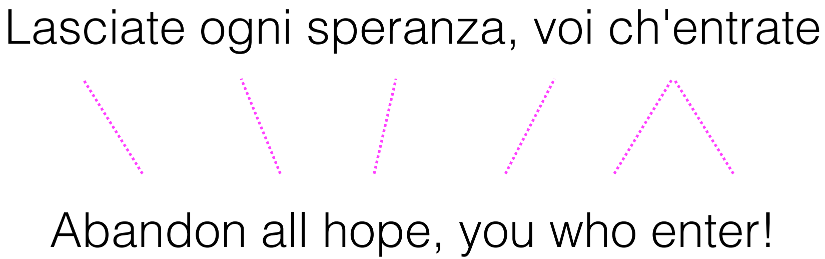
The statistical revolution in machine translation (1990) started by exploiting the structure of parallel sentences to learn the translation model.

Lasciate ogni speranza, voi ch'entrate

Abandon all hope, you who enter!

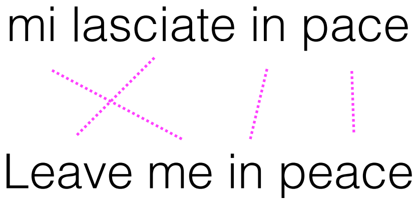
Brown et al. (1990), “A statistical approach to machine translation,” *Computational Linguistics*

Statistical MT



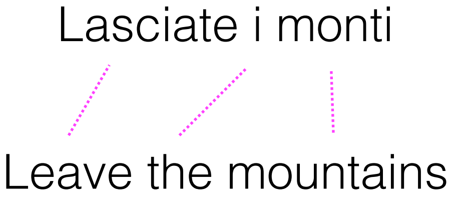
Lasciate ogni speranza, voi ch'entrate

Abandon all hope, you who enter!



mi lasciate in pace

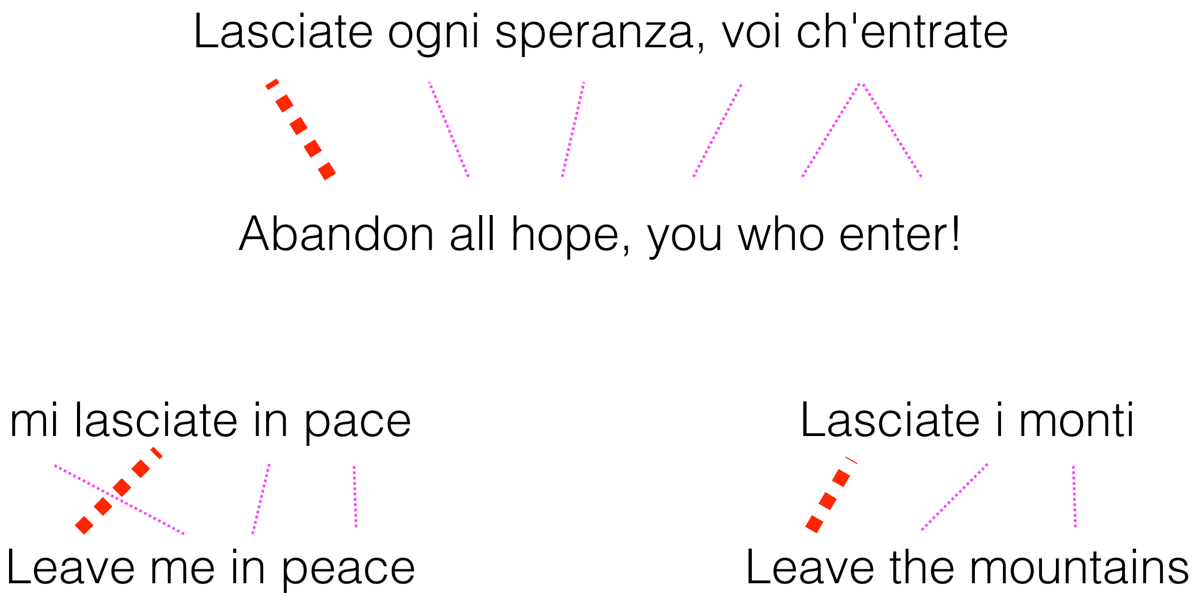
Leave me in peace



Lasciate i monti

Leave the mountains

Statistical MT



Lasciate ogni speranza, voi ch'entrate

Abandon all hope, you who enter!

mi lasciate in pace

Leave me in peace

Lasciate i monti

Leave the mountains

Statistical MT

Italian	English	P(English | Italian)
lasciate	leave	0.67
lasciate	abandon	0.33

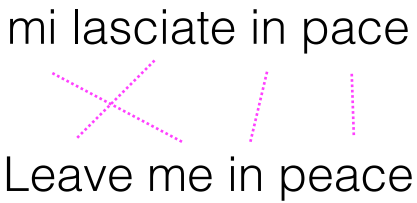
Translation table

Italian	English	P(English | Italian)
Voi ch’entrate	you who enter	0.91
Voi ch’entrate	you who are entering	0.09

Phrase translation table

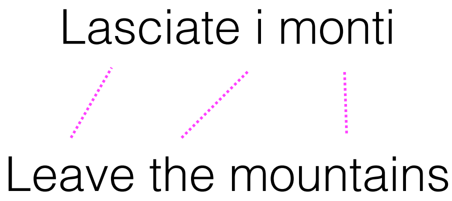
IBM Alignment models

If we had explicit word alignments we could estimate translation tables directly from them.



mi lasciate in pace

Leave me in peace



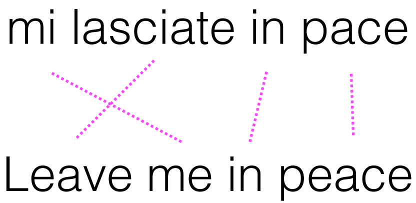
Lasciate i monti

Leave the mountains

But we don’t have word alignments — just sentence alignments!

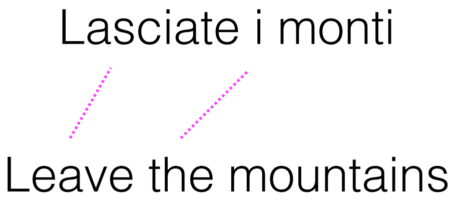
IBM Alignment models

Unsupervised models for aligning words and phrases in parallel sentences.



mi lasciate in pace

Leave me in peace



Lasciate i monti

Leave the mountains

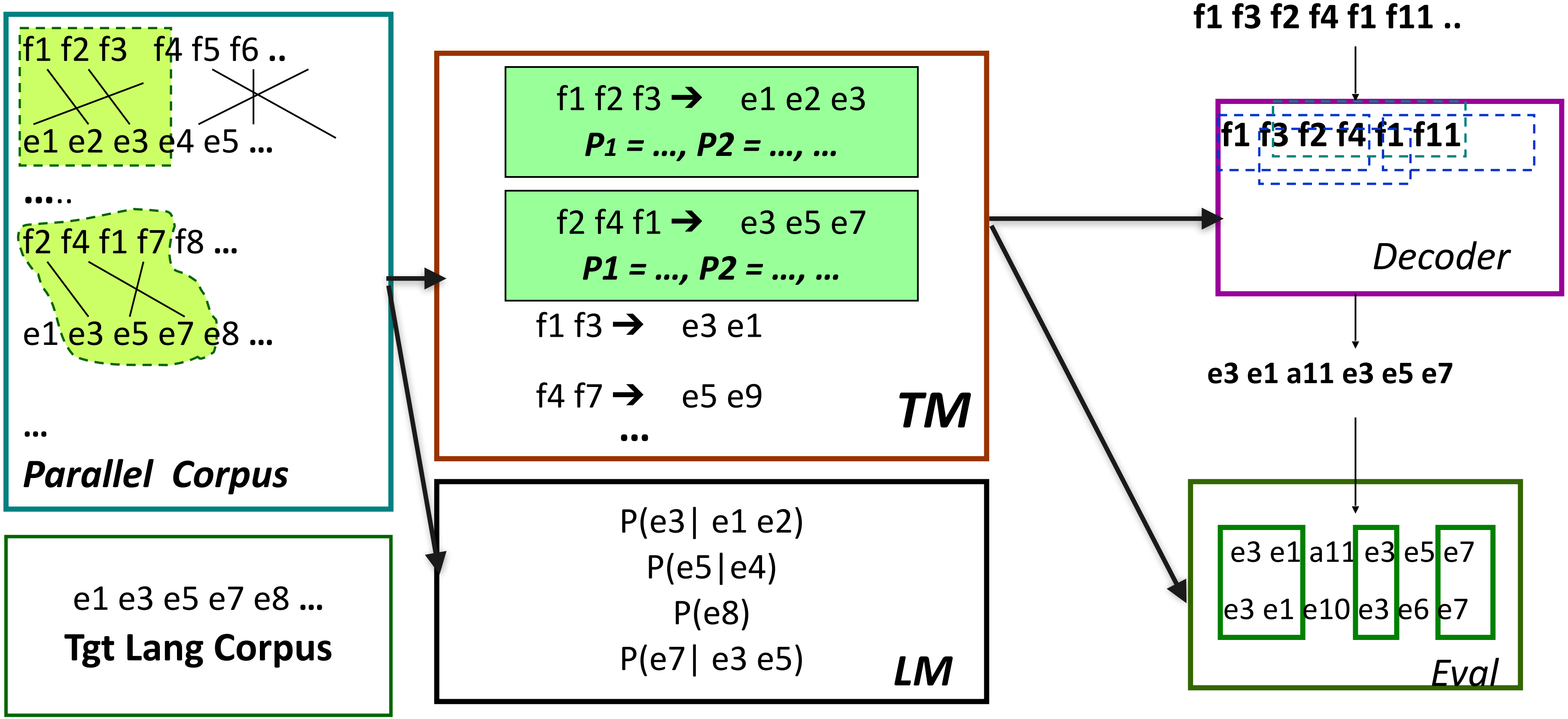
Brown, Peter F. (1993). “The mathematics of statistical machine translation: Parameter estimation,” Computational Linguistics

IBM Alignment models

Model 1	Independent word translation (order doesn’t matter)
Model 2	Word translation + distance between source and target position
Model 3	Word translation + fertility (how many target words a source word can align to)
Model 4	Word translation + relative ordering among target words of same source
Model 5	(Fixes deficiency of model 4)
HMM (Vogel et al. 1996)	Word translation plus relative ordering

Brown, Peter F. (1993). “The mathematics of statistical machine translation: Parameter estimation,” Computational Linguistics

Phrase-based MT



f1 f2 f3 f4 f5 f6 ..

e1 e2 e3 e4 e5 …

…..

f2 f4 f1 f7 f8 …

e1 e3 e5 e7 e8 …

…

Parallel Corpus

e1 e3 e5 e7 e8 …

Tgt Lang Corpus

f1 f2 f3 ➔ e1 e2 e3

P1 = …, P2 = …, …

f2 f4 f1 ➔ e3 e5 e7

P1 = …, P2 = …, …

f1 f3 ➔ e3 e1

f4 f7 ➔ e5 e9

…

TM

P(e3| e1 e2)

P(e5|e4)

P(e8)

P(e7| e3 e5)

LM

f1 f3 f2 f4 f1 f11 ..

f1 f3 f2 f4 f1 f11

Decoder

e3 e1 a11 e3 e5 e7

e3 e1 a11 e3 e5 e7

e3 e1 e10 e3 e6 e7

Eval

Phrase-based MT

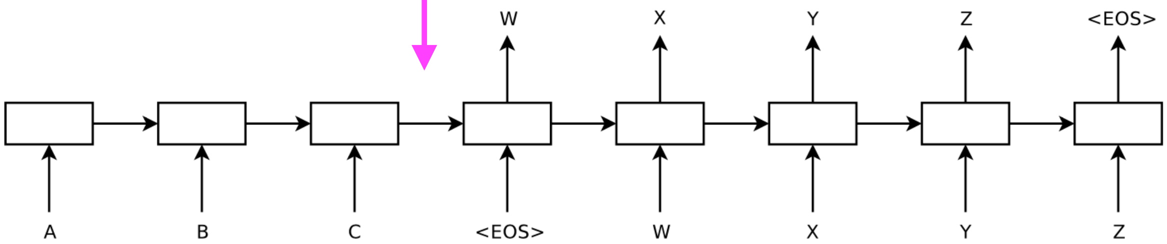
- Relatively light data requirement
    - Reasonable baseline with 1M+ (sent) parallel corpus
    - Competitive with Neural MT for many of low resource scenario.

Neural MT

- Encoder-decoder
- Encoder-decoder + attention
- Transformer (Vaswani et al. 2018)

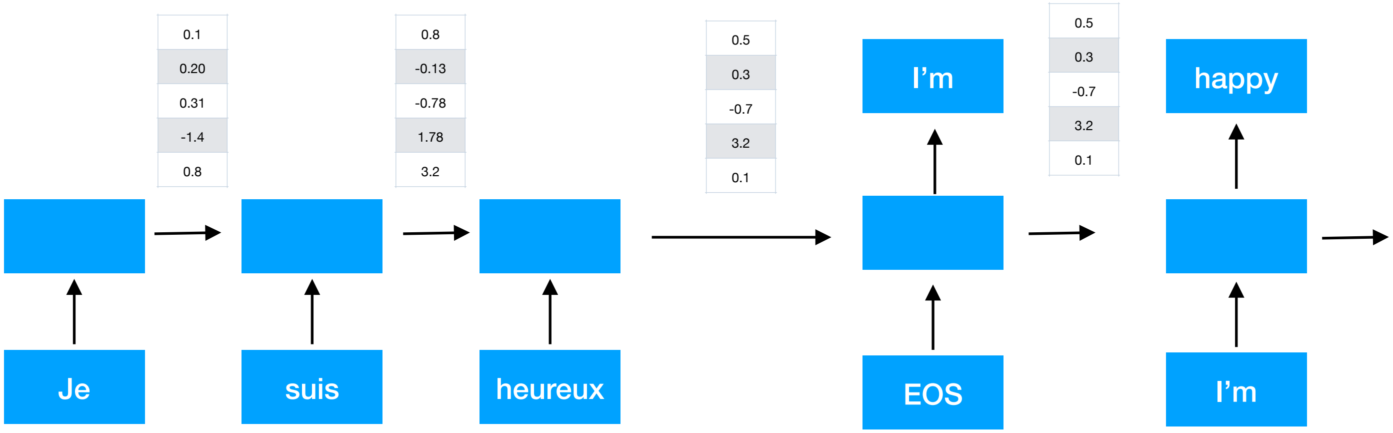
Encoder-decoder framework

K-dimensional vector representing entire context



Condition on word generated in translation

Sutskever et al. (2015);



0.1
0.20
0.31
-1.4
0.8

0.8
-0.13
-0.78
1.78
3.2

0.5
0.3
-0.7
3.2
0.1

I’m

0.5
0.3
-0.7
3.2
0.1

happy

Je

suis

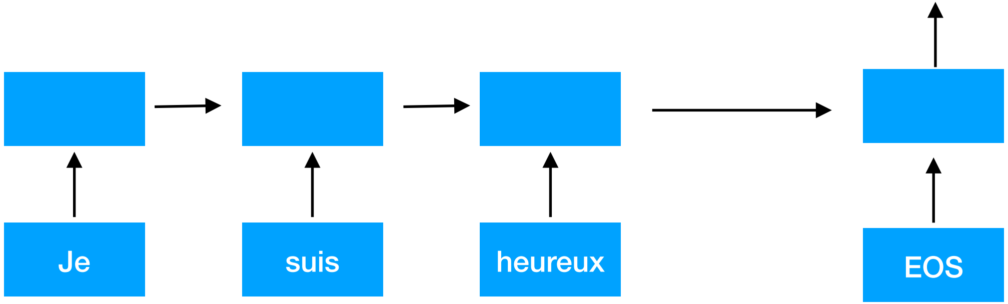
heureux

EOS

I’m

Training

- As in other RNNs, we can train by minimizing the loss between what we predict at each time step and the truth.



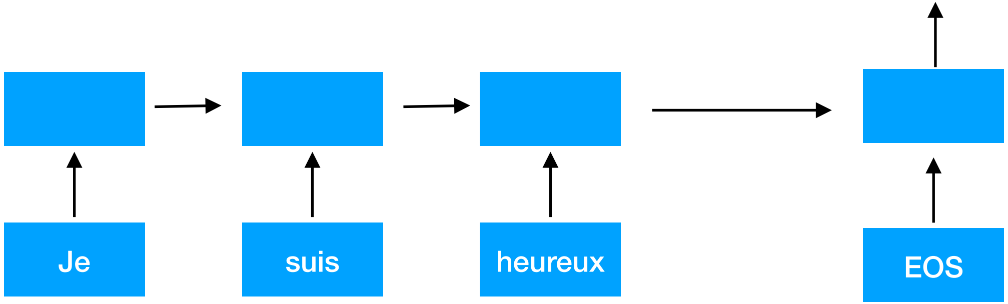
Training

truth

I’m	you	are	the	…
1	0	0	0	0

predicted

I’m	you	are	the	…
0.03	0.05	0.02	0.01	0.009

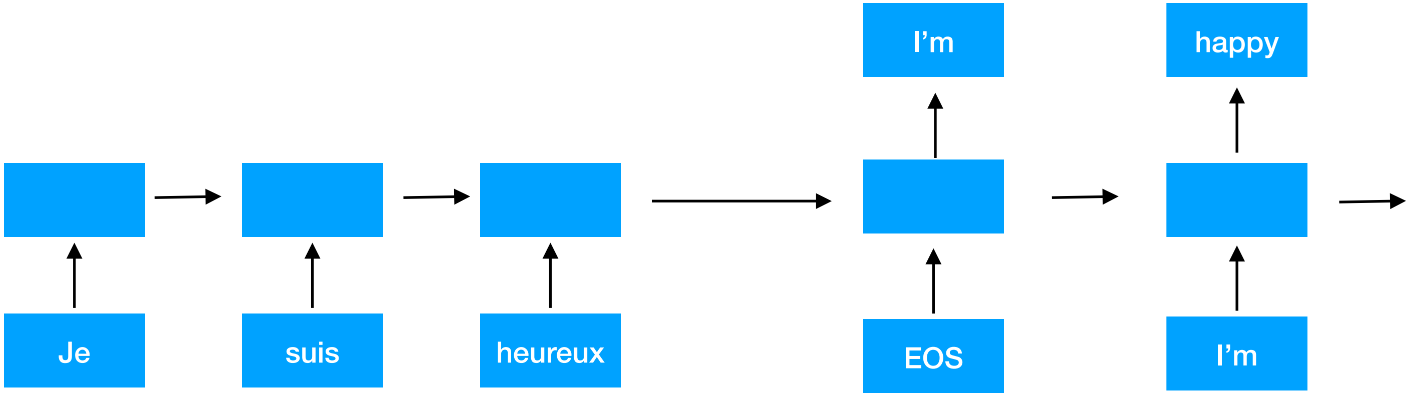


truth

happy	great	bad	ok	…
1	0	0	0	0

predicted

happy	great	bad	ok	…
0.13	0.08	0.01	0.03	0.009



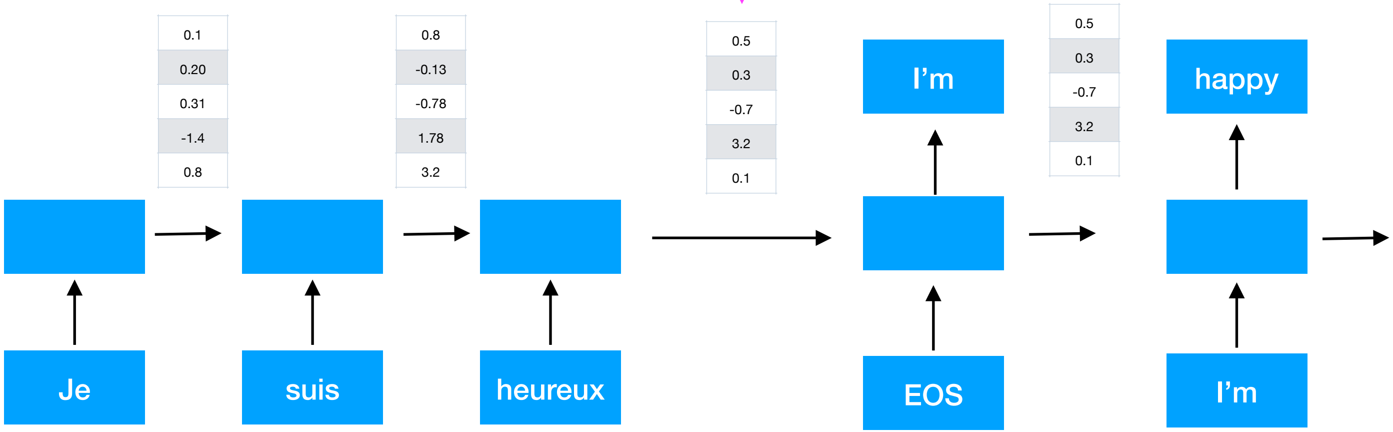
Encoder-decoder

- Sutskever et al. (2014) found better performance when the encoder reads the sentence in backwards, from right to left (increase in BLEU from 25.9 to 30.6)

Sutskever et al. (2014), “Sequence to Sequence Learning with Neural Networks”

Encoder-decoder

The entire source sentence is summarized in this one vector



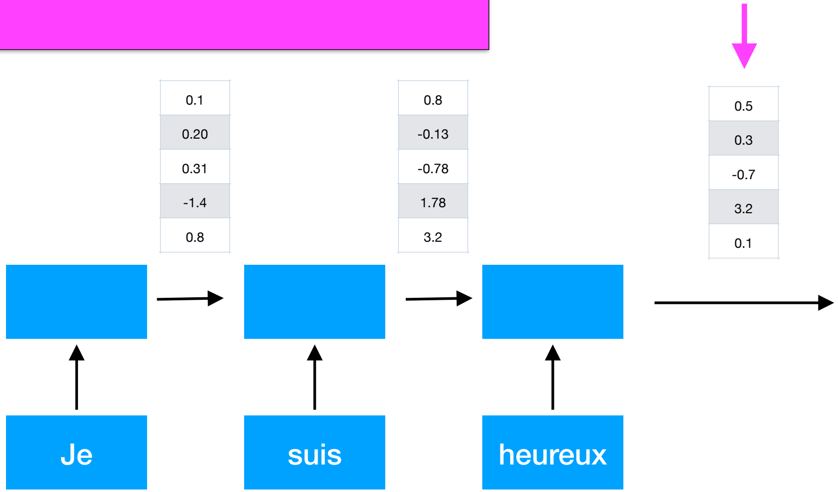
The decoder state depends just on the previous *state* and the previous *output*

$$s_i = f(s_{i-1}, y_{i-1})$$

Encoder-decoder with attention

The decoder state depends just the previous *state,* the previous *output,* and some *context*

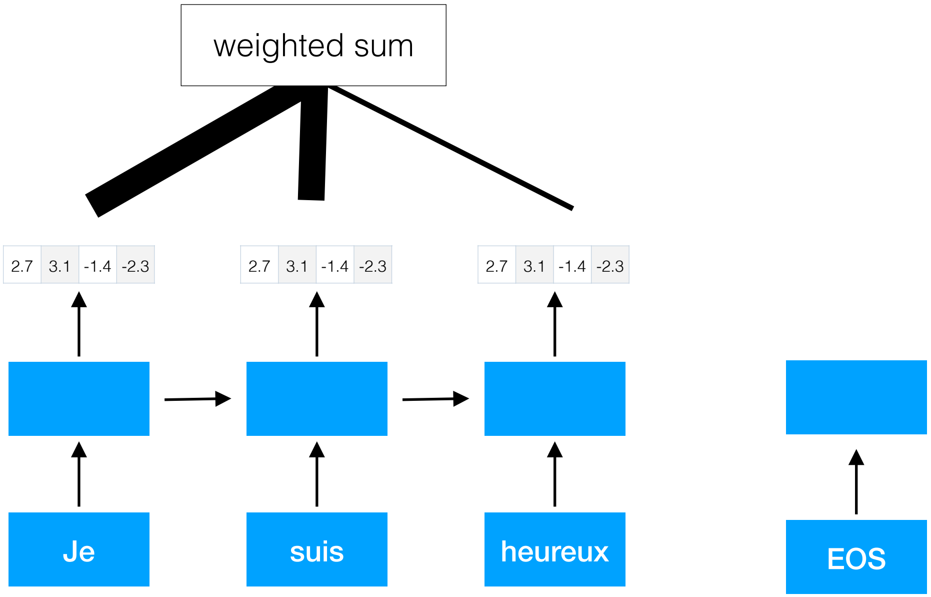
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



Encoder-decoder with attention

$$c = h_1a_1 + h_2a_2 + h_3a_3$$

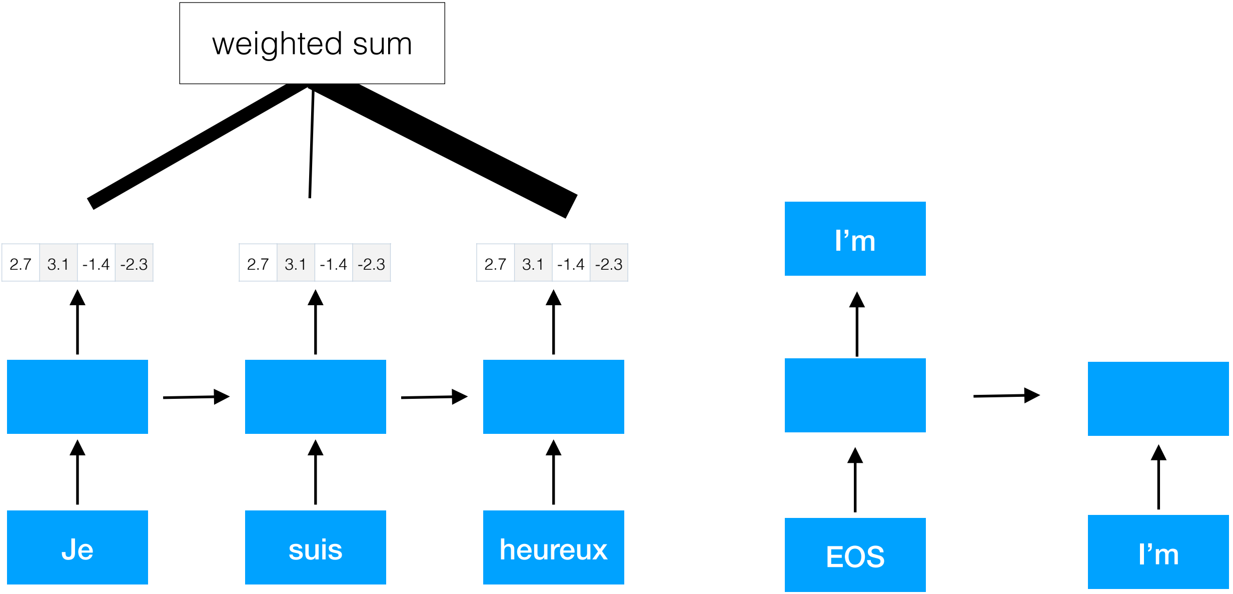
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



Encoder-decoder with attention

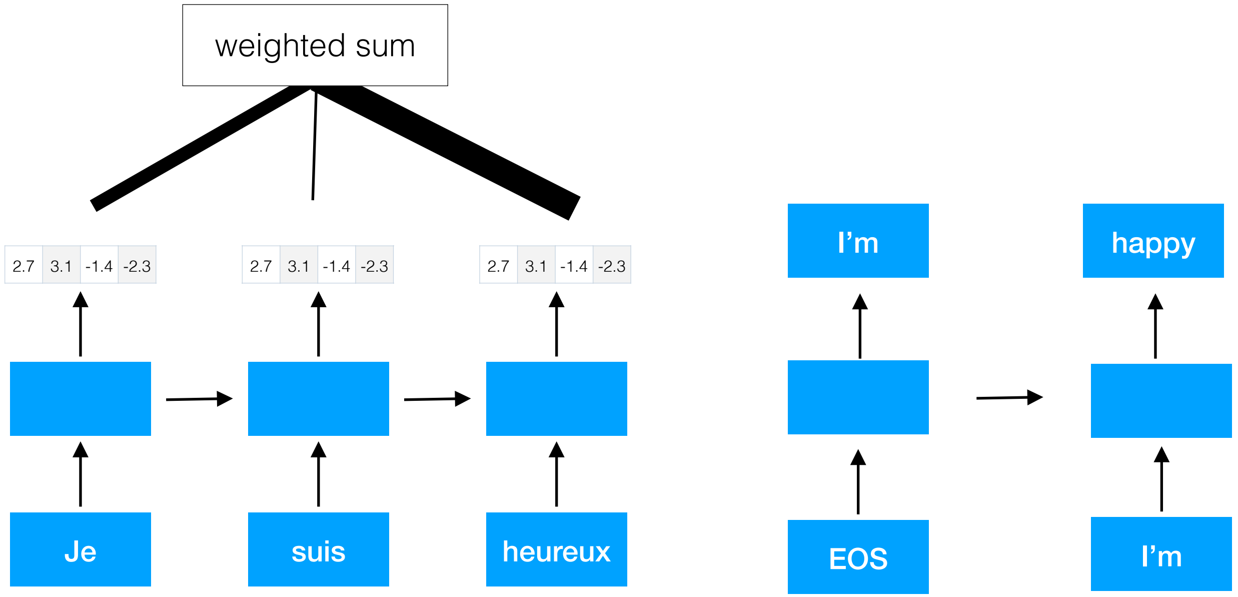
$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



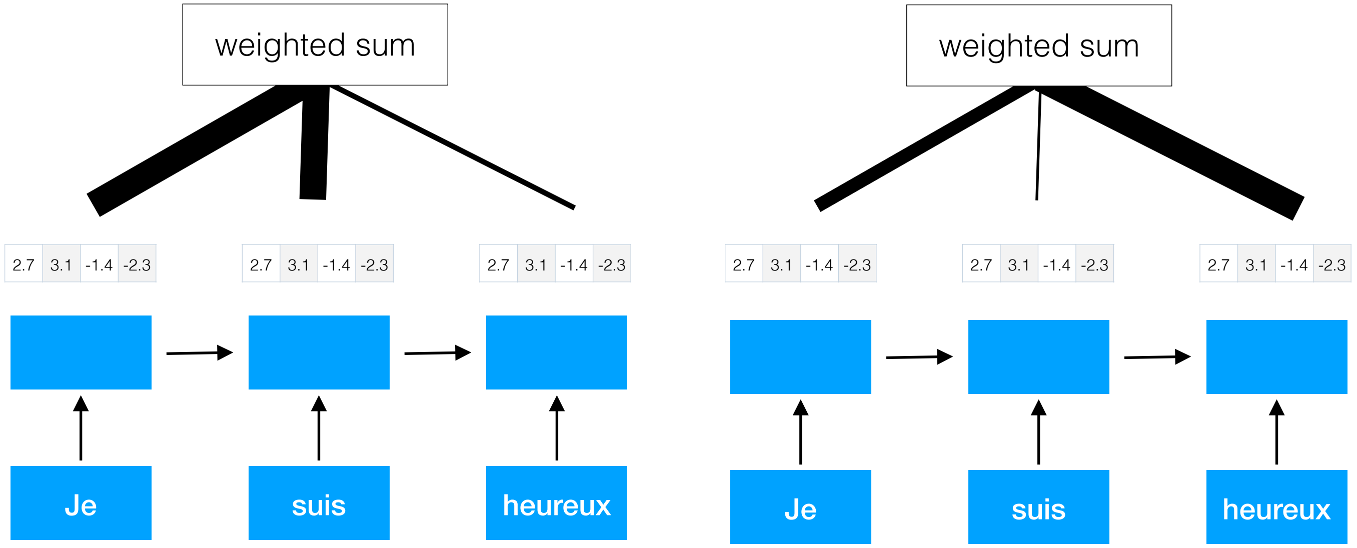
Encoder-decoder with attention

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$



Encoder-decoder with attention

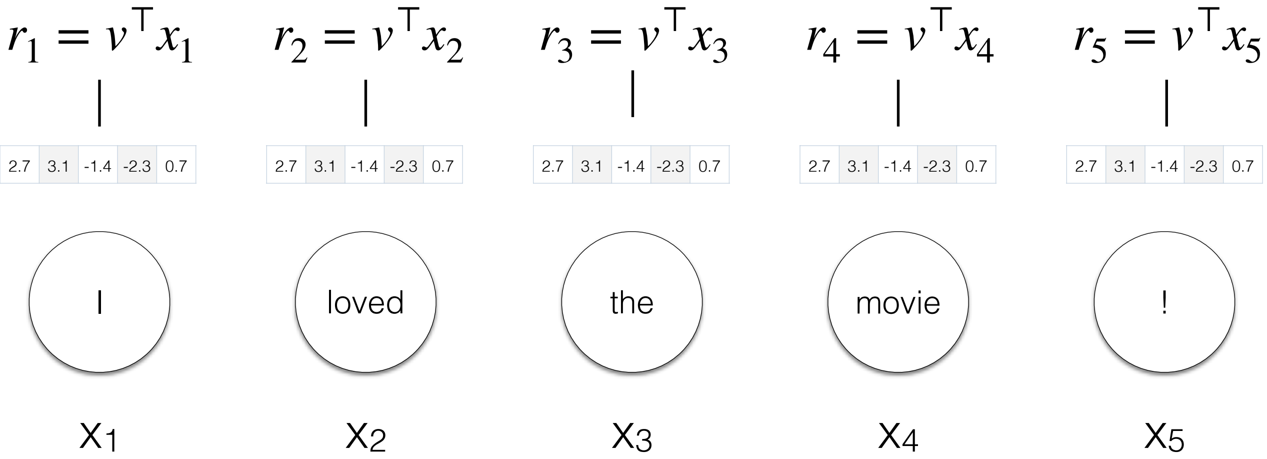
- Each time step in the decoder has its own weighted context vector



$$v \in \mathcal{R}^H$$

2.7	3.1	-1.4	-2.3	0.7

With document classification, we parameterized attention with a single vector v to be learned. Attention in an encoder-decoder network is a little different because we're comparing a pair of vectors.



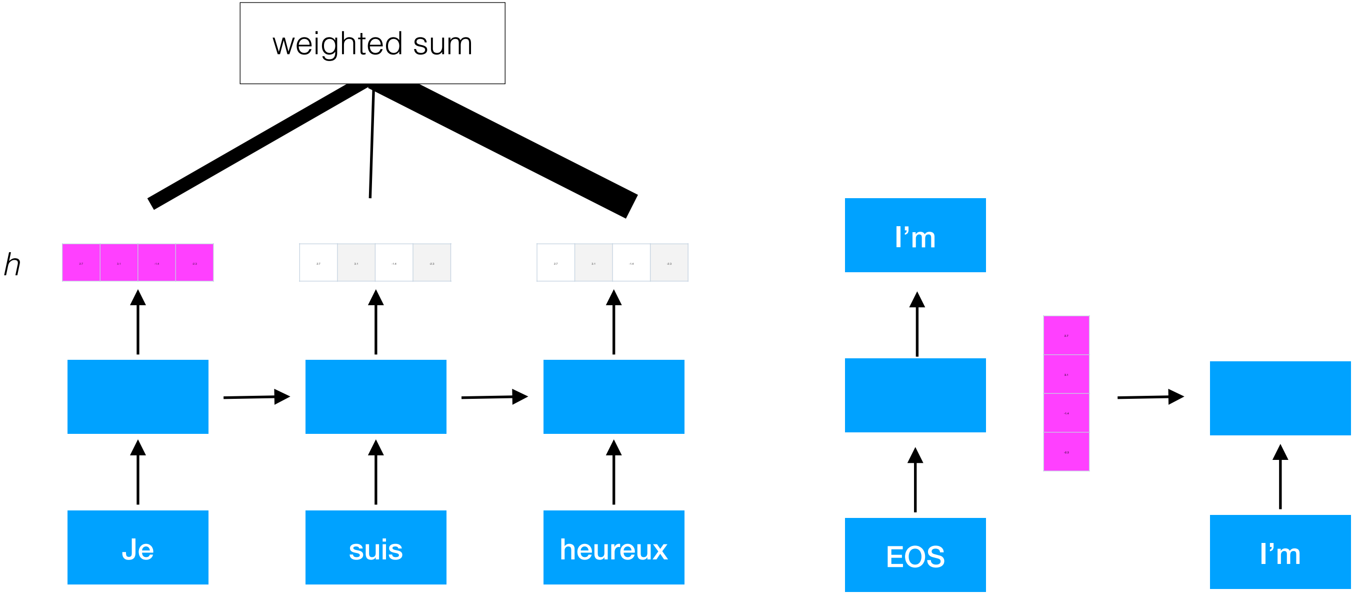
$r_1 = v^\top x_1$ $r_2 = v^\top x_2$ $r_3 = v^\top x_3$ $r_4 = v^\top x_4$ $r_5 = v^\top x_5$

2.7	3.1	-1.4	-2.3	0.7
2.7	3.1	-1.4	-2.3	0.7
2.7	3.1	-1.4	-2.3	0.7
2.7	3.1	-1.4	-2.3	0.7
2.7	3.1	-1.4	-2.3	0.7

Encoder-decoder with attention

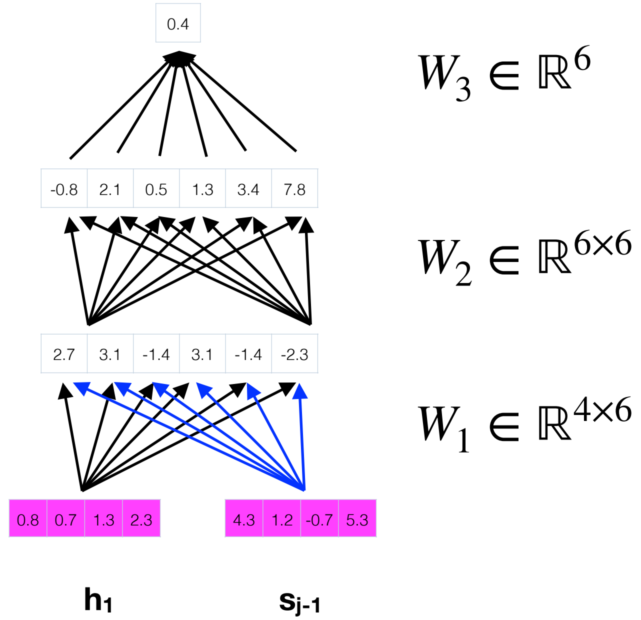
$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

$$r_{1,j} = FFNN(h_1, s_{j-1})$$



Feed-forward neural network

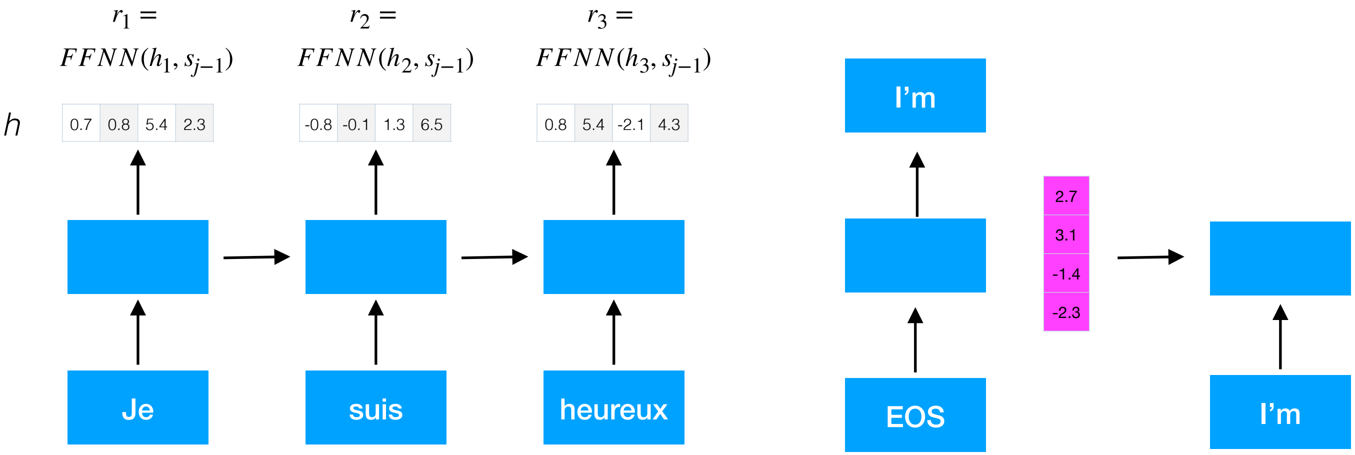
The feed-forward network here just takes the two vectors as input as outputs a single scalar. The parameters are all learned using backprop (just like every other parameter).



Encoder-decoder with attention

$$a = \text{softmax}(r)$$

$$r = [r_1, r_2, r_3]$$



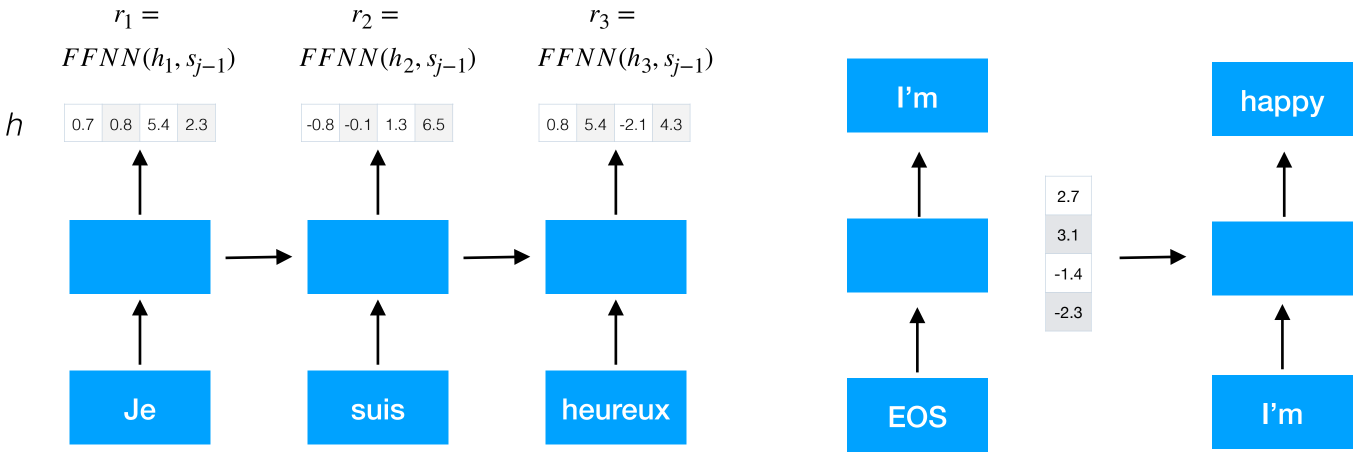
Encoder-decoder with attention

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

$$a = \text{softmax}(r)$$

$$r = [r_1, \ldots, r_5]$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



$r_1 = FFNN(h_1, s_{j-1})$

$r_2 = FFNN(h_2, s_{j-1})$

$r_3 = FFNN(h_3, s_{j-1})$

h

0.7	0.8	5.4	2.3

-0.8	-0.1	1.3	6.5

0.8	5.4	-2.1	4.3

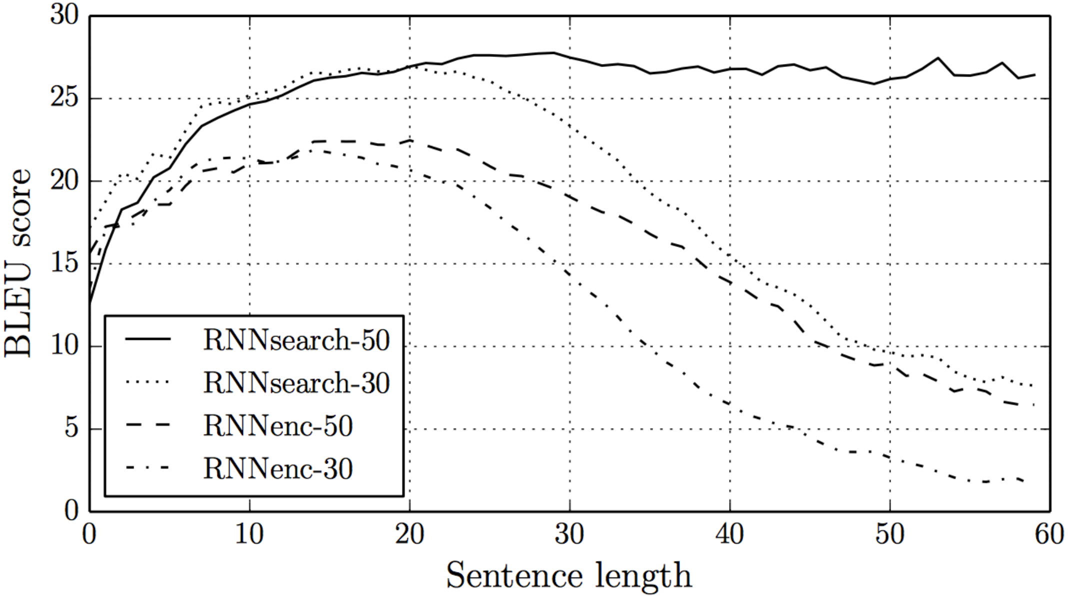
2.7
3.1
-1.4
-2.3

Attention

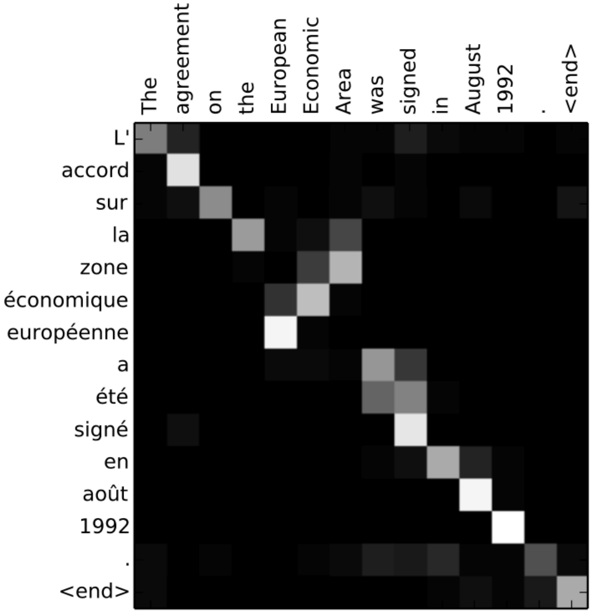
- For text classification, attention helps decide which words in the text are important for the label.

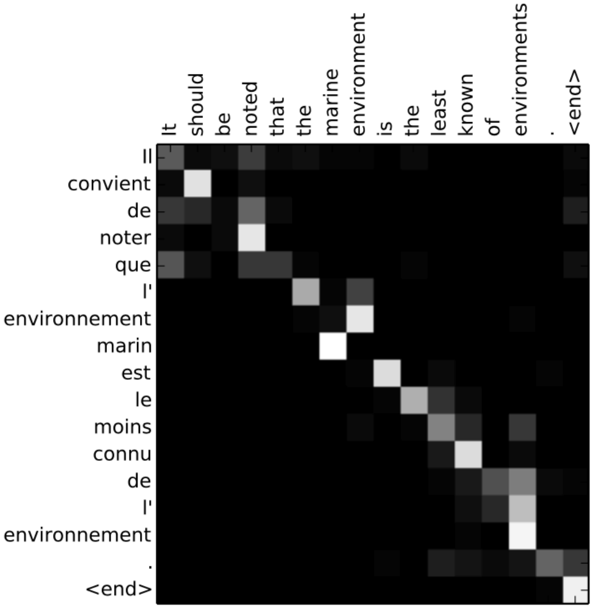
- For MT, attention changes with each word being generated during decoding. Each subsequent word pays attention to different parts of the input.

Better performance on long sentences



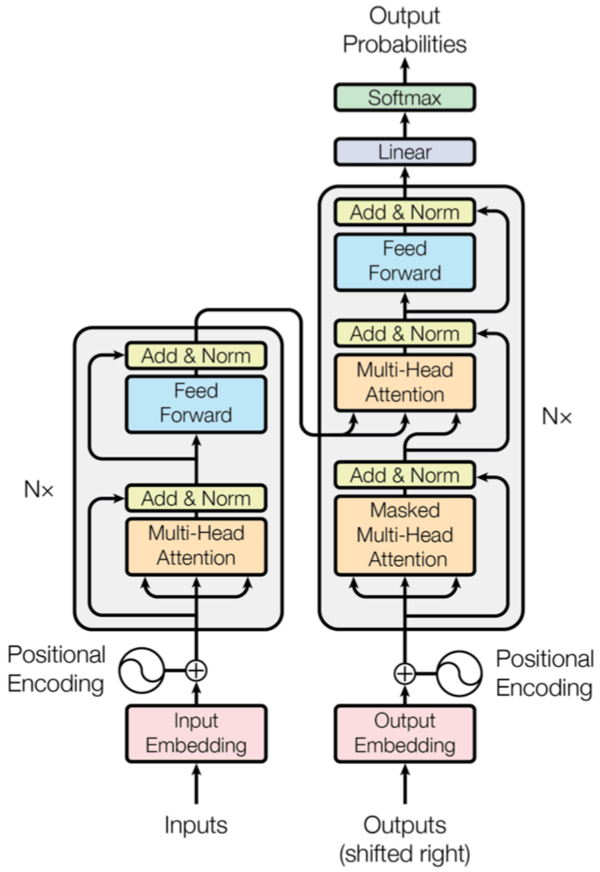
Bahdanau et al. (2016), “Neural Machine Translation by Jointly Learning to Align and Translate”





Bahdanau et al. (2016), “Neural Machine Translation by Jointly Learning to Align and Translate”

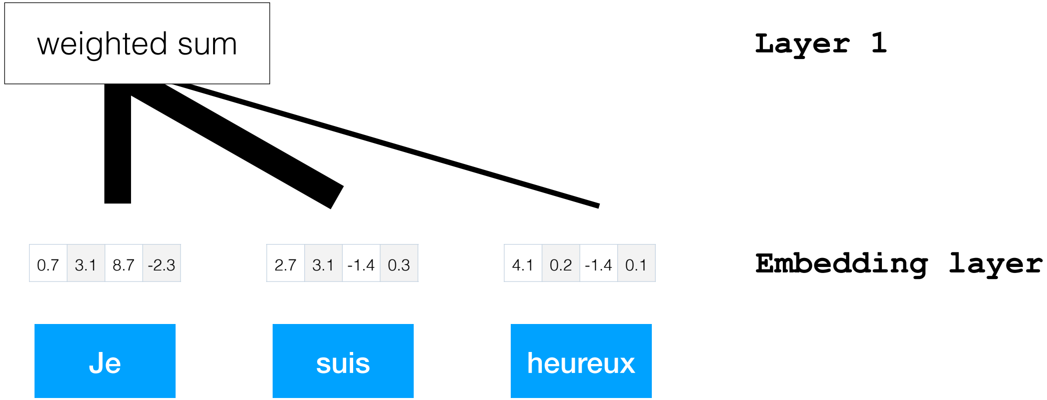
- Transformer network (Vaswani et al. 2017).



Self-attention

- Multiple layers of representations for an input sequence; each layer attends over the representations in the previous layer.

Self-attention



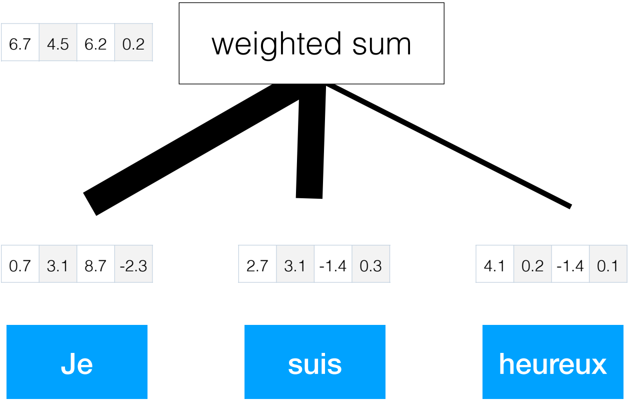
weighted sum

Layer 1

Embedding layer

Je	suis	heureux
0.7 3.1 8.7 -2.3	2.7 3.1 -1.4 0.3	4.1 0.2 -1.4 0.1

Self-attention



6.7 | 4.5 | 6.2 | 0.2

weighted sum

Layer 1

0.7 | 3.1 | 8.7 | -2.3

2.7 | 3.1 | -1.4 | 0.3

4.1 | 0.2 | -1.4 | 0.1

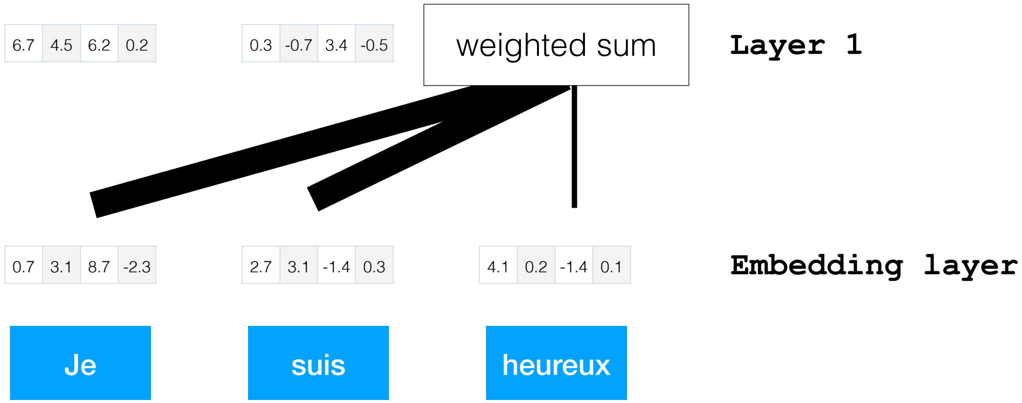
Embedding layer

Je

suis

heureux

Self-attention



6.7 | 4.5 | 6.2 | 0.2

0.3 | -0.7 | 3.4 | -0.5

weighted sum

Layer 1

0.7 | 3.1 | 8.7 | -2.3

2.7 | 3.1 | -1.4 | 0.3

4.1 | 0.2 | -1.4 | 0.1

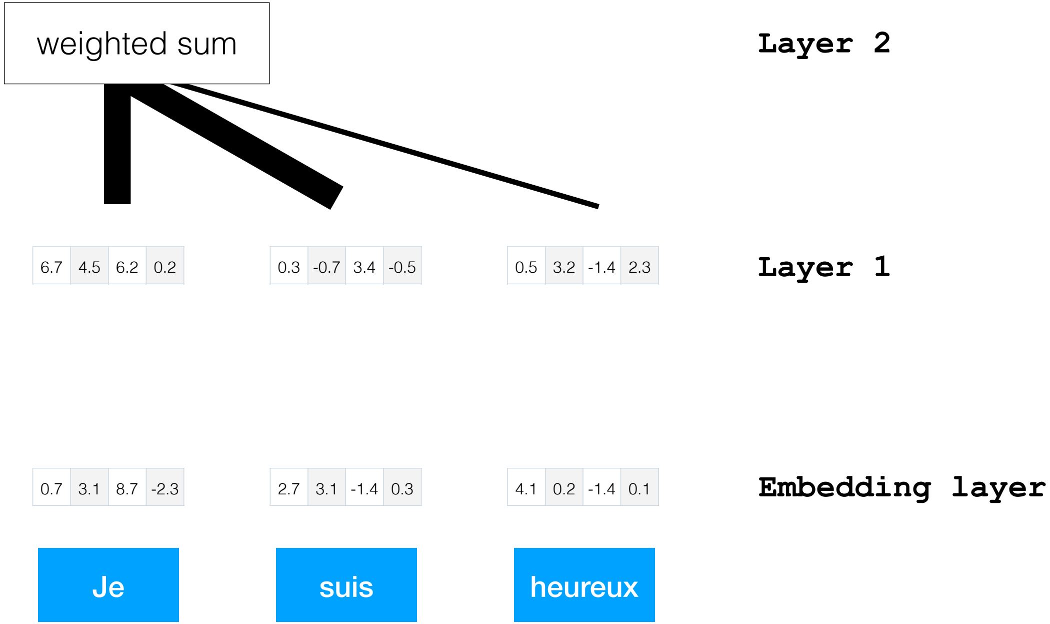
Embedding layer

Je

suis

heureux

Self-attention



weighted sum

Layer 2

6.7	4.5	6.2	0.2

0.3	-0.7	3.4	-0.5

0.5	3.2	-1.4	2.3

Layer 1

0.7	3.1	8.7	-2.3

2.7	3.1	-1.4	0.3

4.1	0.2	-1.4	0.1

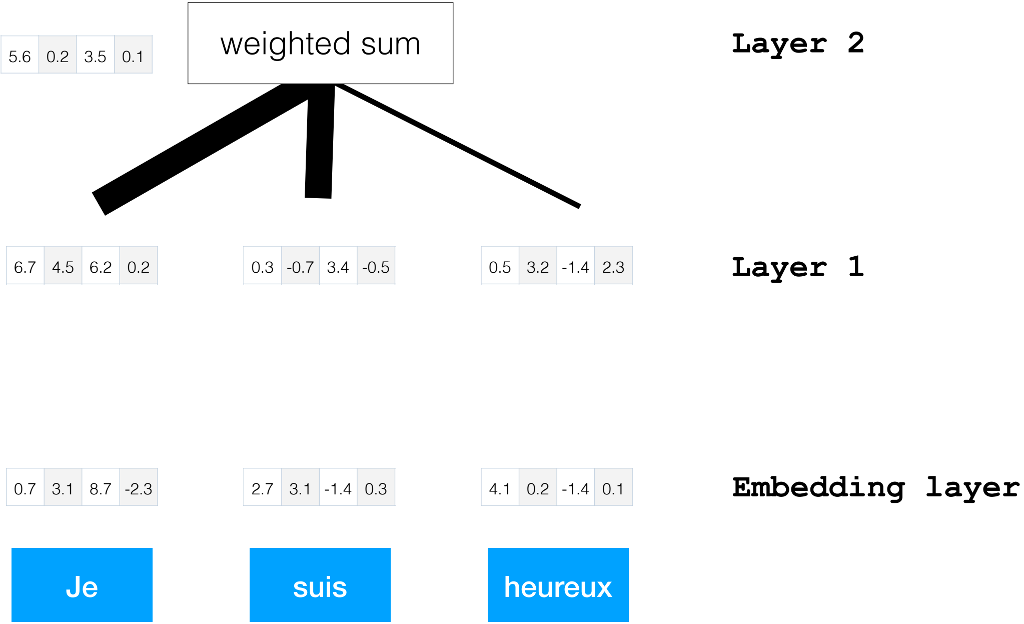
Embedding layer

Je

suis

heureux

Self-attention



5.6 0.2 3.5 0.1

Layer 2

6.7 4.5 6.2 0.2

0.3 -0.7 3.4 -0.5

0.5 3.2 -1.4 2.3

Layer 1

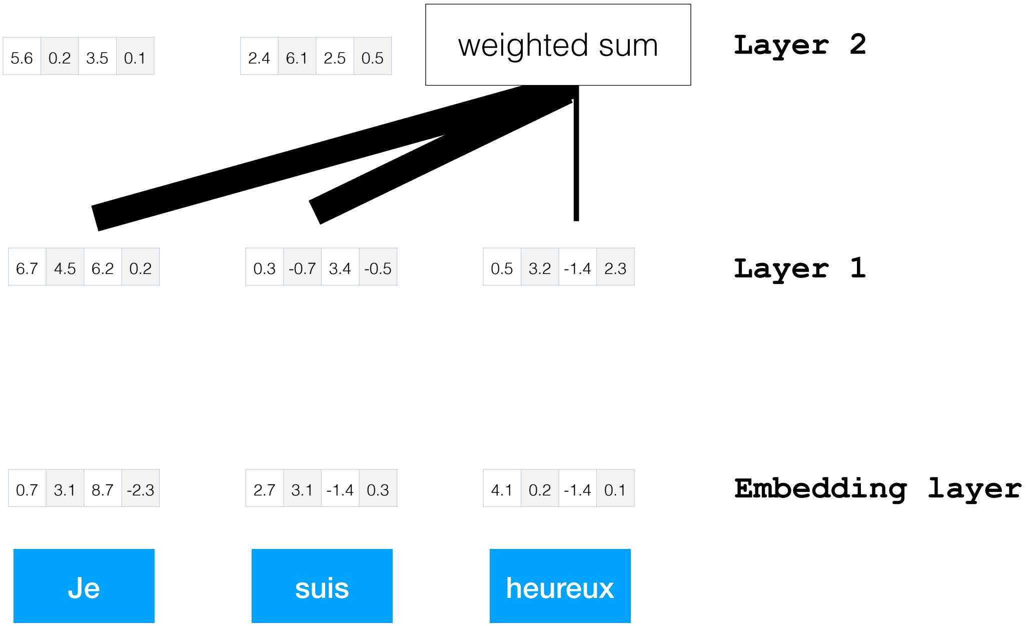
0.7 3.1 8.7 -2.3

2.7 3.1 -1.4 0.3

4.1 0.2 -1.4 0.1

Embedding layer

Self-attention



5.6	0.2	3.5	0.1

2.4	6.1	2.5	0.5

weighted sum

Layer 2

6.7	4.5	6.2	0.2

0.3	-0.7	3.4	-0.5

0.5	3.2	-1.4	2.3

Layer 1

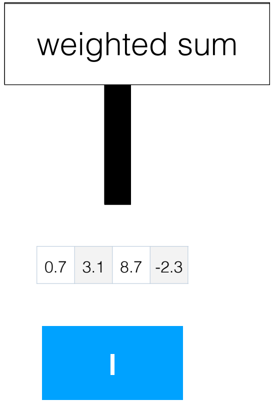
0.7	3.1	8.7	-2.3

2.7	3.1	-1.4	0.3

4.1	0.2	-1.4	0.1

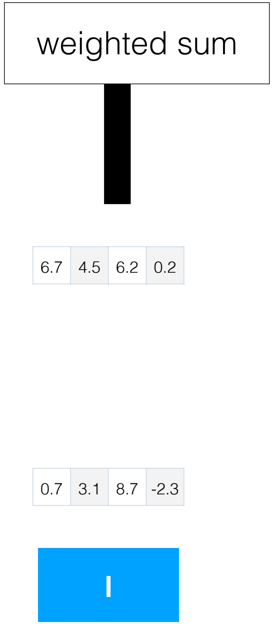
Embedding layer

Self-attention



- In the *decoder*, self-attention can only attend over words to the left of the position (since the right ones haven't been generated yet).

Self-attention



weighted sum

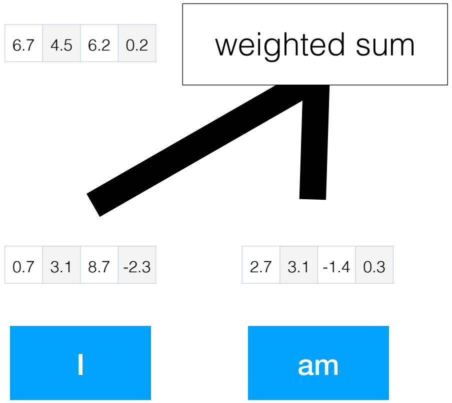
6.7	4.5	6.2	0.2

0.7	3.1	8.7	-2.3

I

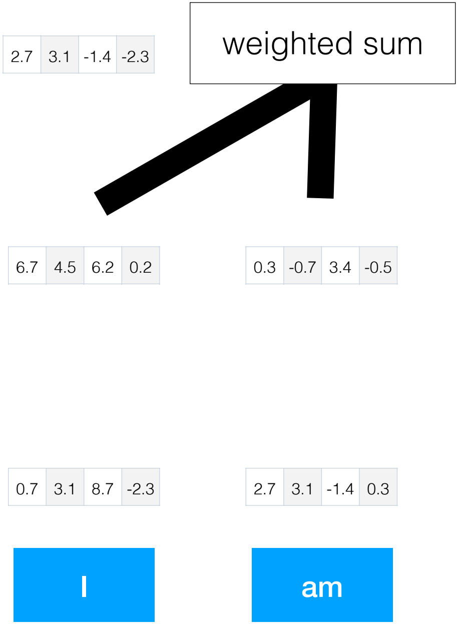
- In the decoder, self-attention can only attend over words to the left of the position (since the right ones haven't been generated yet).

Self-attention



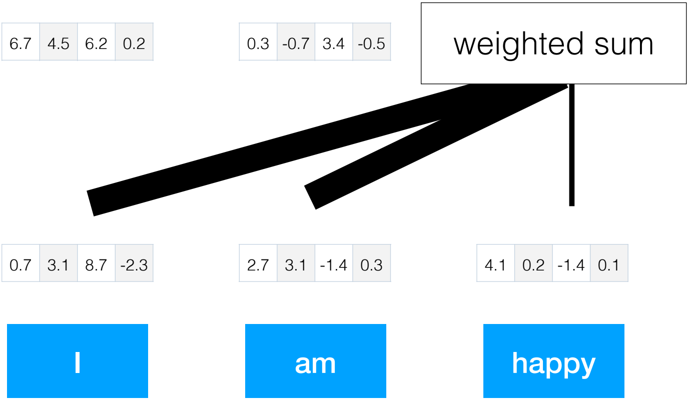
- In the decoder, self-attention can only attend over words to the left of the position (since the right ones haven’t been generated yet).

Self-attention



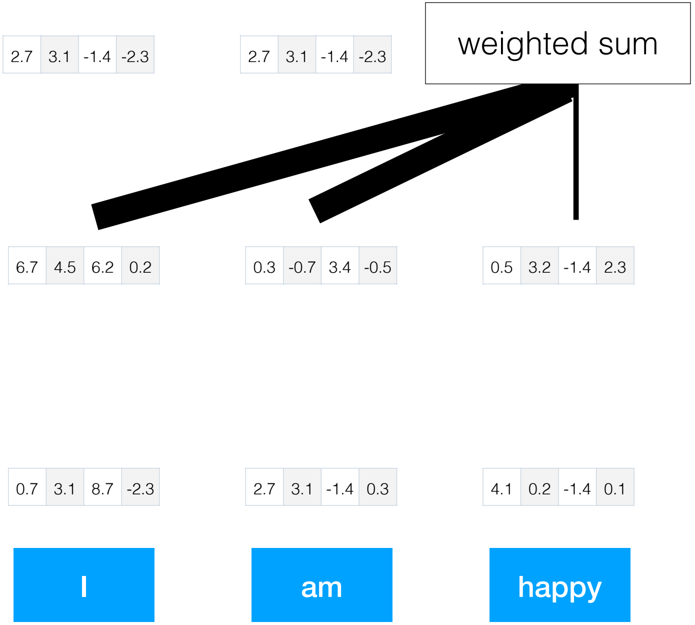
- In the decoder, self-attention can only attend over words to the left of the position (since the right ones haven't been generated yet).

Self-attention



- In the decoder, self-attention can only attend over words to the left of the position (since the right ones haven't been generated yet).

Self-attention



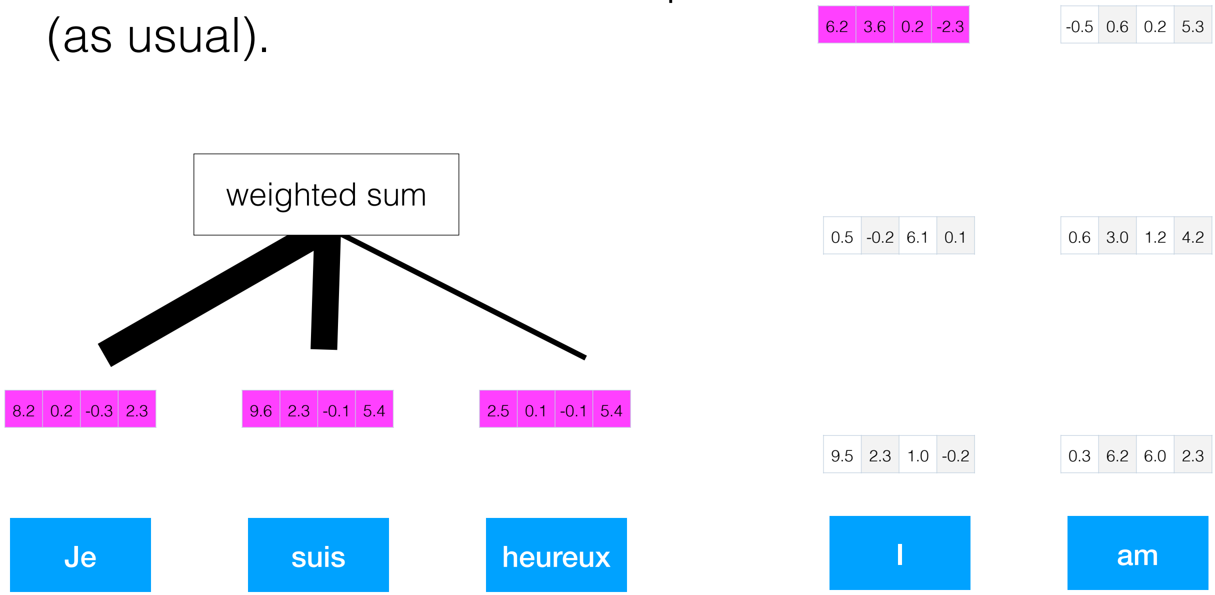
- In the decoder, self-attention can only attend over words to the left of the position (since the right ones haven't been generated yet).

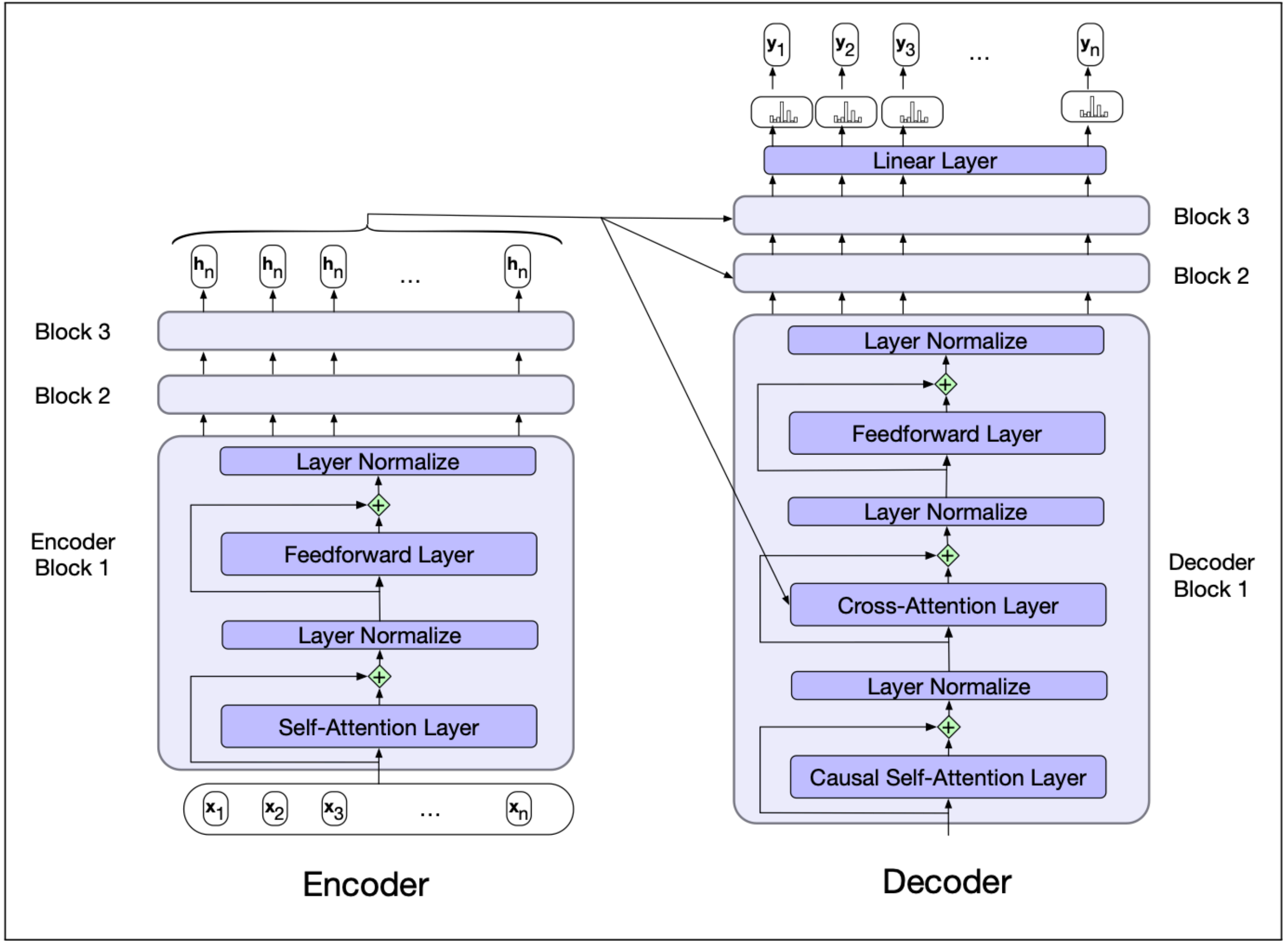
Encoder-decoder cross-attention

5.6 0.2 3.5 0.1	2.4 6.1 2.5 0.5	0.3 6.2 3.2 2.4	6.2 3.6 0.2 -2.3	-0.5 0.6 0.2 5.3
6.7 4.5 6.2 0.2	0.3 -0.7 3.4 -0.5	0.5 3.2 -1.4 2.3	0.5 -0.2 6.1 0.1	0.6 3.0 1.2 4.2
0.7 3.1 8.7 -2.3	2.7 3.1 -1.4 0.3	4.1 0.2 -1.4 0.1	9.5 2.3 1.0 -0.2	0.3 6.2 6.0 2.3
Je	suis	heureux	I	am

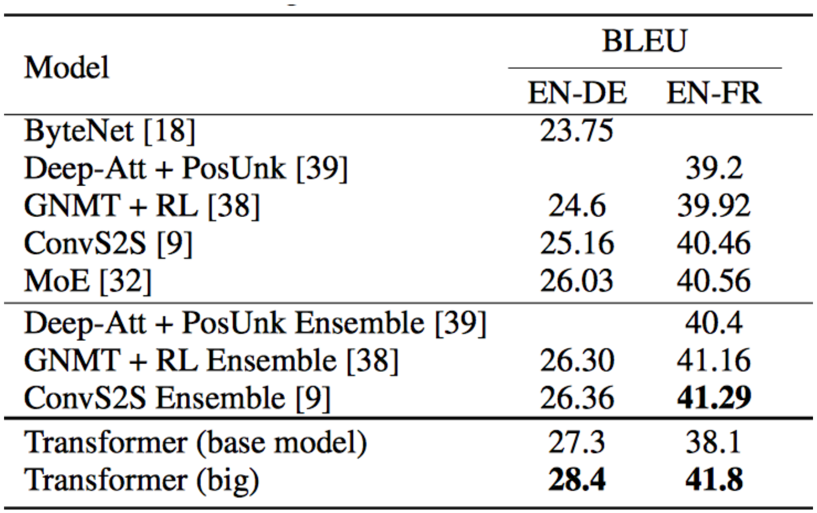
Encoder-decoder cross-attention

- Each layer in the decoder attends over the encoder output (as usual).





SLP3 fig. 10.6; https://web.stanford.edu/~jurafsky/slp3/13.pdf



Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	**41.29**
Transformer (base model)	27.3	38.1
Transformer (big)	**28.4**	**41.8**

MT Improvements

- Der Rolls-Royce Merlin ist ein 12-Zylinder-Flugmotor von Rolls-Royce in V-Bauweise, der vielen wichtigen britischen und US-amerikanischen Flugzeugmustern des Zweitenweltkriegs als Antrieb diente. Ab 1941 wurde der Motor in Lizenz von der Packard Motor Car Company in den USA als Packard V-1650 gebaut.
- Nach dem Krieg wurden diverse Passagier- und Frachtflugzeuge mit diesem Motor ausgestattet, so z. B. Avro Lancastrian, Avro Tudor und Avro York, später noch einmal die Canadair C-4 (umgebaute Douglas C-54). Der zivile Einsatz des Merlin hielt sich jedoch in Grenzen, da er als robust, aber zu laut galt.
- Die Bezeichnung des Motors ist gemäß damaliger Rolls-Royce Tradition von einer Vogelart, dem Merlinfalken, übernommen und nicht, wie oft vermutet, von dem Zauberer Merlin.

German Wikipedia

- The Rolls-Royce Merlin is a 12-cylinder V-shaped aircraft engine from Rolls-Royce that powered many important British and US aircraft of the Second World War. From 1941 the engine was built under license by the Packard Motor Car Company in the USA as the Packard V-1650.
- After the war, various passenger and cargo aircraft were equipped with this engine, e.g. B. Avro Lancastrian, Avro Tudor and Avro York, later again the Canadair C-4 (converted Douglas C-54). However, civilian use of the Merlin was limited because it was considered robust but too loud.
- In keeping with Rolls-Royce tradition at the time, the name of the engine was taken from a species of bird, the Merlin falcon, and not, as is often assumed, from the magician Merlin.

Google Translate 2023

MT Improvements

- The Rolls-Royce Merlin is a 12-cylinder aircraft engine from Rolls-Royce V-type, which served many important British and American aircraft designs of World War II as a drive. From 1941 the engine was built under license by the Packard Motor Car Company in the U.S. as a Packard V-1650th.
- After the war, several passenger and cargo aircraft have been equipped with this engine, such as Avro Lancastrian, Avro Tudor Avro York and, later, the Canadair C-4 (converted Douglas C-54). The civilian use of the Merlin was, however, limited as it remains robust, however, was too loud.
- The name of the motor is taken under the then Rolls-Royce tradition of one species, the Merlin falcon, and not, as often assumed, by the wizard Merlin.

Google Translate 2015

- The Rolls-Royce Merlin is a 12-cylinder V-shaped aircraft engine from Rolls-Royce that powered many important British and US aircraft of the Second World War. From 1941 the engine was built under license by the Packard Motor Car Company in the USA as the Packard V-1650.
- After the war, various passenger and cargo aircraft were equipped with this engine, e.g. B. Avro Lancastrian, Avro Tudor and Avro York, later again the Canadair C-4 (converted Douglas C-54). However, civilian use of the Merlin was limited because it was considered robust but too loud.
- In keeping with Rolls-Royce tradition at the time, the name of the engine was taken from a species of bird, the Merlin falcon, and not, as is often assumed, from the magician Merlin.

Google Translate 2023

Data Case Study: NLLB

- Meta’s No Language Left Behind Project
- Mass creation of parallel corpora for 200+ languages
  - 148 low resource languages
  - 135 million Wikipedia sentence