

INFERNO.

El mezzo del camin di nostra vita
n Mi ritrouai per una selua oscura;
Che la diritta uia era smarrita:
Et quanto a dir qual era, è cosa dura
Esta selua seluaggia et aspra et forte;
Che nel pensier rinuoua la paura.

Natural Language Processing

Info 159/259

Lecture 25: Machine Translation (Nov. 27, 2018)

David Bamman, UC Berkeley

Project presentations

- 3:30-5pm Tuesday 12/5, 202 South Hall
- We'll provide the posterboard + easels + wallspace; you bring a poster. (Any poster under 36" wide is fine).
- If you want to demo any systems (and want a table/monitor, email TAs and we can set one up).

Relation extraction

The Big Sleep is a 1946 film noir directed by Howard Hawks,^{[2][3]} the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."^[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

<i>subject</i>	<i>predicate</i>	<i>object</i>
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall
The Big Sleep	screenplay_by	William Faulkner
The Big Sleep	screenplay_by	Leigh Brackett
The Big Sleep	screenplay_by	Jules Furthman

from last time

Wikipedia Infoboxes

The Big Sleep is a 1946 film noir directed by Howard Hawks,^{[2][3]} the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."^[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

A remake starring Robert Mitchum as Philip Marlowe was released in 1978. This was the second film in three years featuring Mitchum as Marlowe. The remake was arguably more faithful to the novel, possibly due to fewer restrictions in 1978 on what could be portrayed on screen, however, it was far less successful than the original 1946 version. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.



Theatrical release lobby card

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	<i>The Big Sleep</i> by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby
Distributed by	Warner Bros.
Release date	August 23, 1946 (United States)
Running time	114 minutes (released cut) 116 minutes (re-released original cut)

Hearst patterns

<i>pattern</i>	<i>sentence</i>
NP {, NP}* {,} (and or) other NP _H	temples, treasures, and other important civic buildings
NP _H such as {NP,}* {(or and)} NP	red algae such as Gelidium
such NP _H as {NP,}* {(or and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP _H {,} including {NP,}* {(or and)} NP	common-law countries, including Canada and England
NP _H {,} especially {NP}* {(or and)} NP	European countries, especially France, England, and Spain

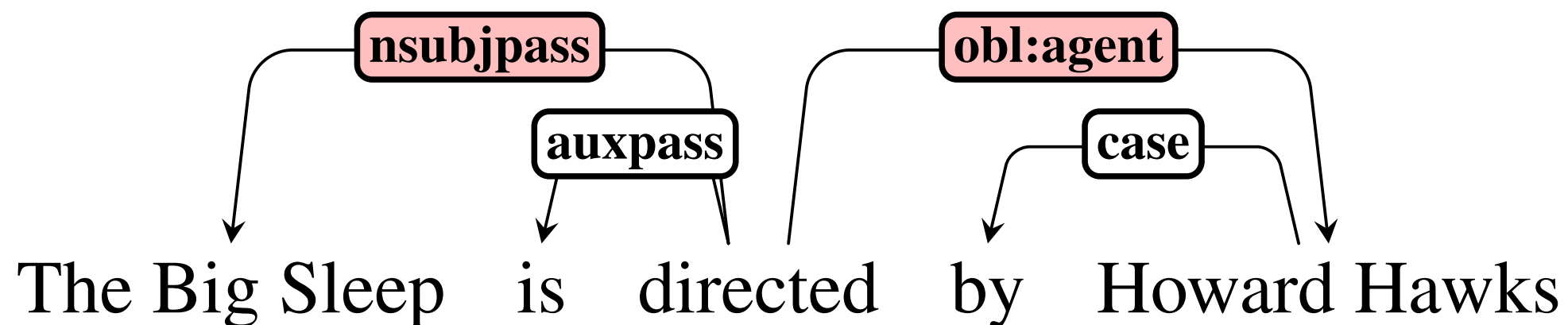
Supervised relation extraction

[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}, the first film version of Raymond Chandler's 1939 novel of the same name.

feature(m1, m2)
headwords of m1, m2
bag of words in m1, m2
bag of words between m1, m2
named entity types of m1, m2
syntactic path between m1, m2

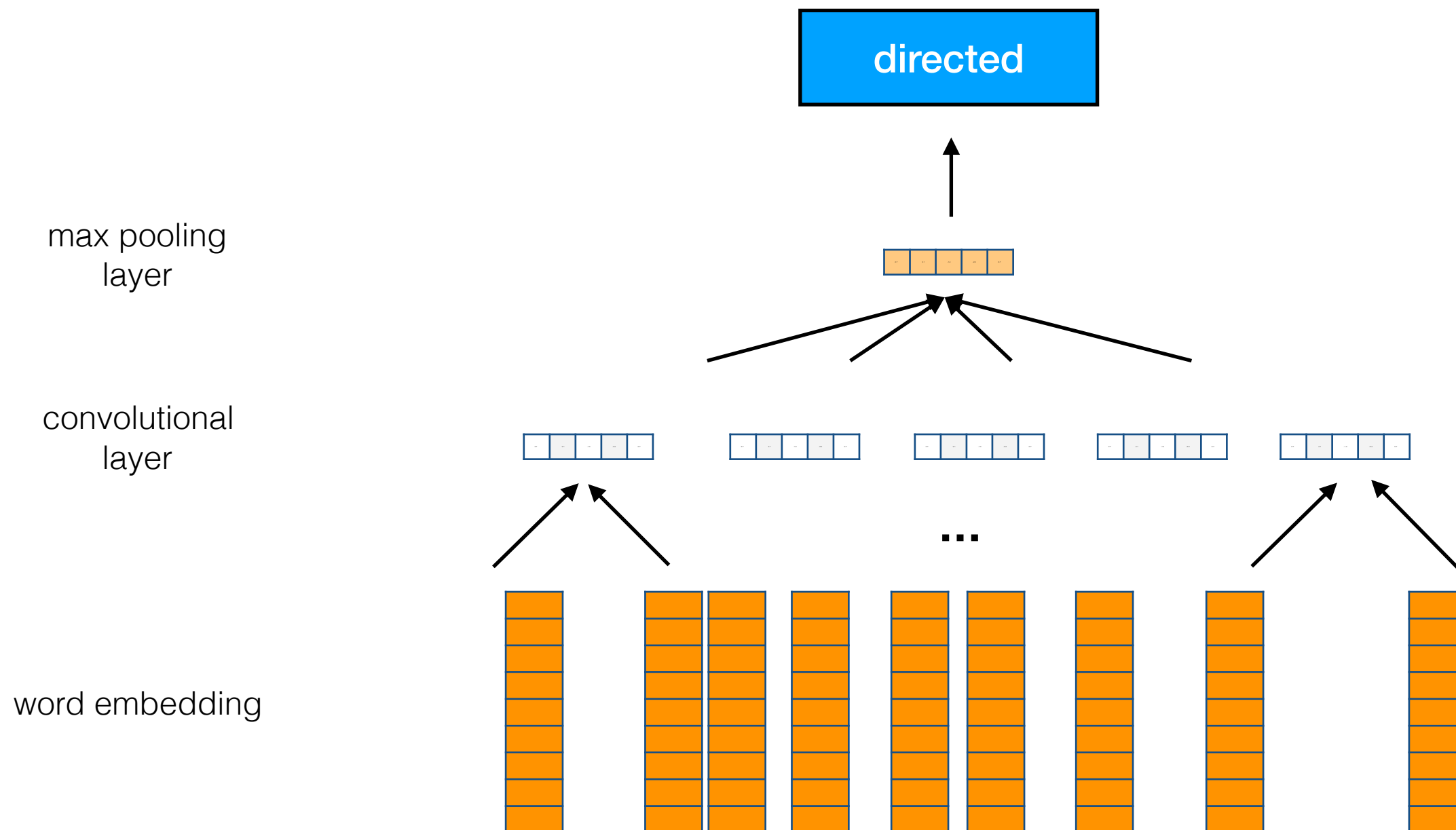
Supervised relation extraction

[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}, the first film version of Raymond Chandler's 1939 novel of the same name.



[The Big Sleep]_{m1} $\leftarrow$ *nsubjpass* directed $\rightarrow$ *obl:agent* [Howard Hawks]_{m2},

$m1 \leftarrow$ *nsubjpass* $\leftarrow$ directed $\rightarrow$ *obl:agent* $\rightarrow$ $m2$



[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}

We don't know which entities we're classifying!

directed(Howard Hawks, The Big Sleep)
genre(The Big Sleep, Film Noir)
year_of_release(The Big Sleep, 1946)

Neural RE

- To solve this, we'll add positional embeddings to our representation of each word — the distance from each word w in the sentence to $m1$ and $m2$

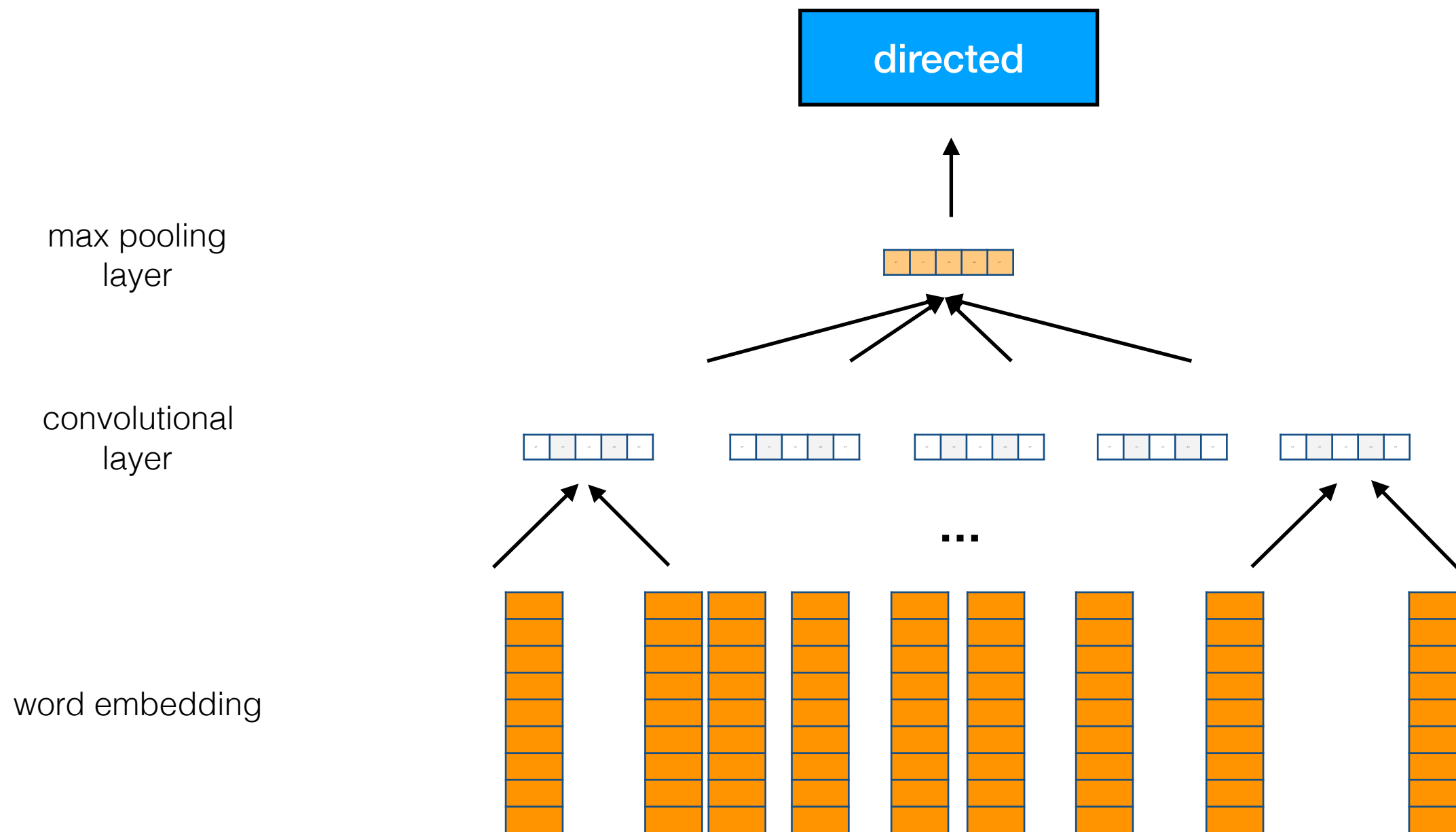
dist from m1	0	1	3	4	5	6	7	8	9
dist from m2	-8	-7	-6	-5	-4	-3	-2	-1	0
	[The Big Sleep]	is	a	1946	film	noir	directed	by	[Howard Hawks]

- 0 here uniquely identifies the head and tail of the relation; other position indicate how close the word is (maybe closer words matter more)

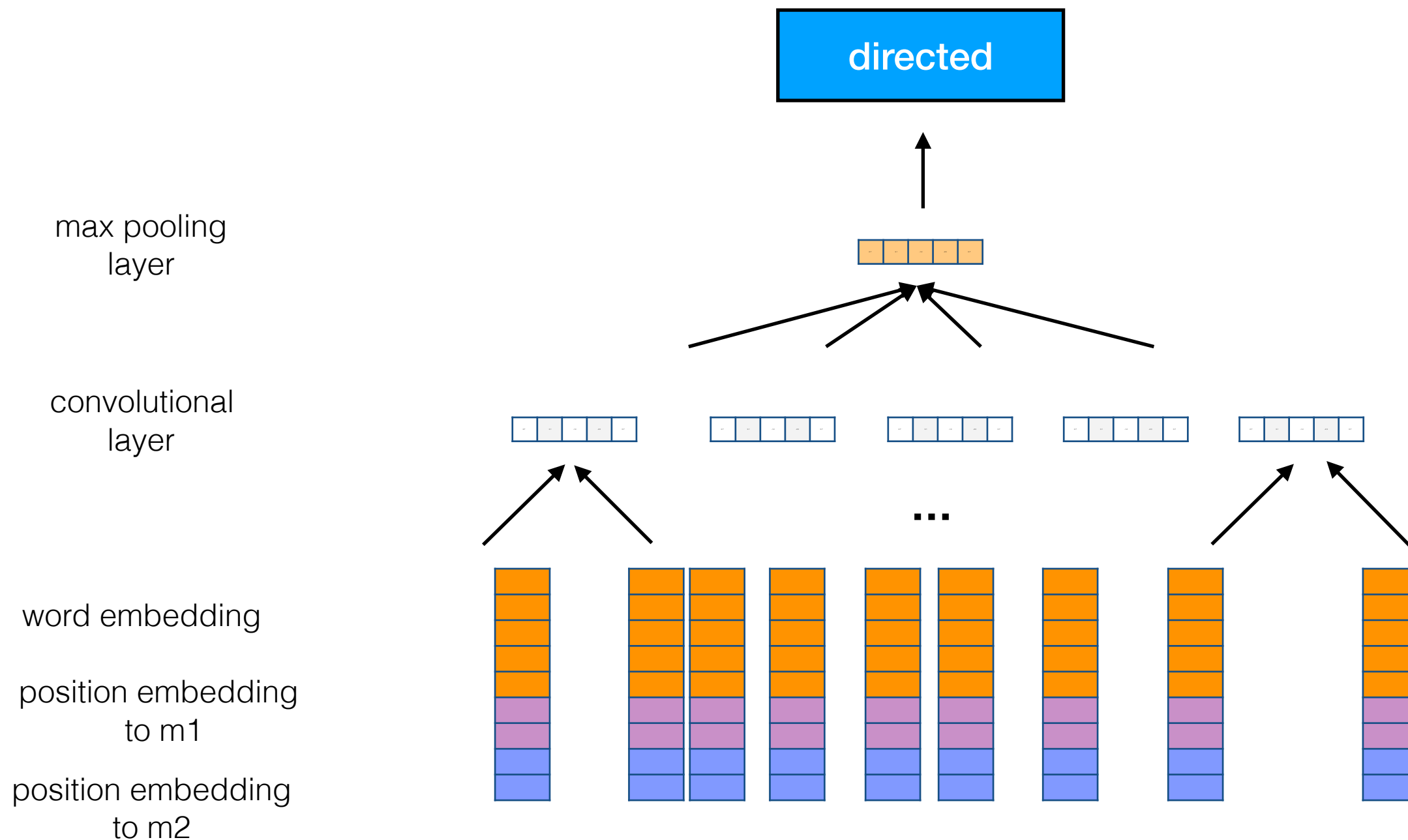
Neural RE

Each position then has an embedding

-4	2	-0.5	1.1	0.3	0.4	-0.5
-3	-1.4	0.4	-0.2	-0.9	0.5	0.9
-2	-1.1	-0.2	-0.5	0.2	-0.8	0
-1	0.7	-0.3	1.5	-0.3	-0.4	0.1
0	-0.8	1.2	1	-0.7	-1	-0.4
1	0	0.3	-0.3	-0.9	0.2	1.4
2	0.8	0.8	-0.4	-1.4	1.2	-0.9
3	1.6	0.4	-1.1	0.7	0.1	1.6
4	1.2	-0.2	1.3	-0.4	0.3	-1.0



[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}



[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}

Distant supervision

- It's uncommon to have labeled data in the form of $\langle \text{sentence}, \text{relation} \rangle$ pairs

<i>sentence</i>	<i>relations</i>
[The Big Sleep] _{m1} is a 1946 film noir directed by [Howard Hawks] _{m2} , the first film version of Raymond Chandler's 1939 novel of the same name.	directed_by(The Big Sleep, Howard Hawks)

Distant supervision

- More common to have knowledge base data about entities and their relations that's separate from text.
- We know the text likely expresses the relations somewhere, but not *exactly where*.

Wikipedia Infoboxes

The Big Sleep is a 1946 film noir directed by Howard Hawks,^{[2][3]} the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."^[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

A remake starring Robert Mitchum as Philip Marlowe was released in 1978. This was the second film in three years featuring Mitchum as Marlowe. The remake was arguably more faithful to the novel, possibly due to fewer restrictions in 1978 on what could be portrayed on screen, however, it was far less successful than the original 1946 version. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.



Theatrical release lobby card

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	<i>The Big Sleep</i> by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby
Distributed by	Warner Bros.
Release date	August 23, 1946 (United States)
Running time	114 minutes (released cut) 116 minutes (re-released original cut)

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Table 2: The 23 largest Freebase relations we use, with their size and an instance of each relation.

Distant supervision

mayor(Maynard Jackson, Atlanta)

Elected mayor of Atlanta in 1973, Maynard Jackson...

Atlanta's airport will be renamed to honor Maynard Jackson, the city's first Black mayor

Born in Dallas, Texas in 1938, Maynard Holbrook Jackson, Jr. moved to Atlanta when he was 8.

mayor(Fiorello LaGuardia, New York)

Fiorello LaGuardia was Mayor of New York for three terms...

Fiorello LaGuardia, then serving on the New York City Board of Aldermen...

Distant supervision

- For feature-based models, we can represent the tuple $\langle m1, m2 \rangle$ by aggregating together the representations from all the sentences they appear in

Distant supervision

[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}, the first film version of Raymond Chandler's 1939 novel of the same name.

[Howard Hawks]_{m2} directed the [The Big Sleep]_{m1}

feature(m1, m2)	value (e.g., normalized over all sentences)
“directed” between m1, m2	0.37
“by” between m1, m2	0.42
m1 ← <i>nsubjpass</i> ← directed → <i>obl:agent</i> → m2	0.13
m2 ← <i>nsubj</i> ← directed → <i>obj</i> → m2	0.08

Distant supervision

- Discovering Hearst patterns from distant supervision using WordNet (Snow et al. 2005)

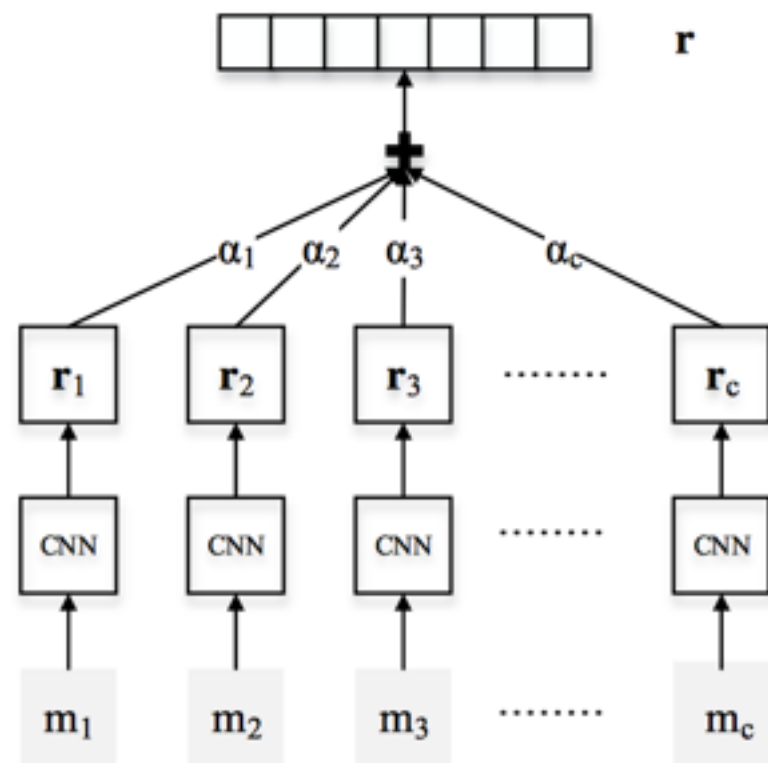
<i>pattern</i>	<i>sentence</i>
NP _H like NP	Many hormones like leptin...
NP _H called NP	a markup language called XHTML
NP is a NP _H	Ruby is a programming language...
NP, a NP _H	IBM, a company with a long...

Multiple Instance Learning

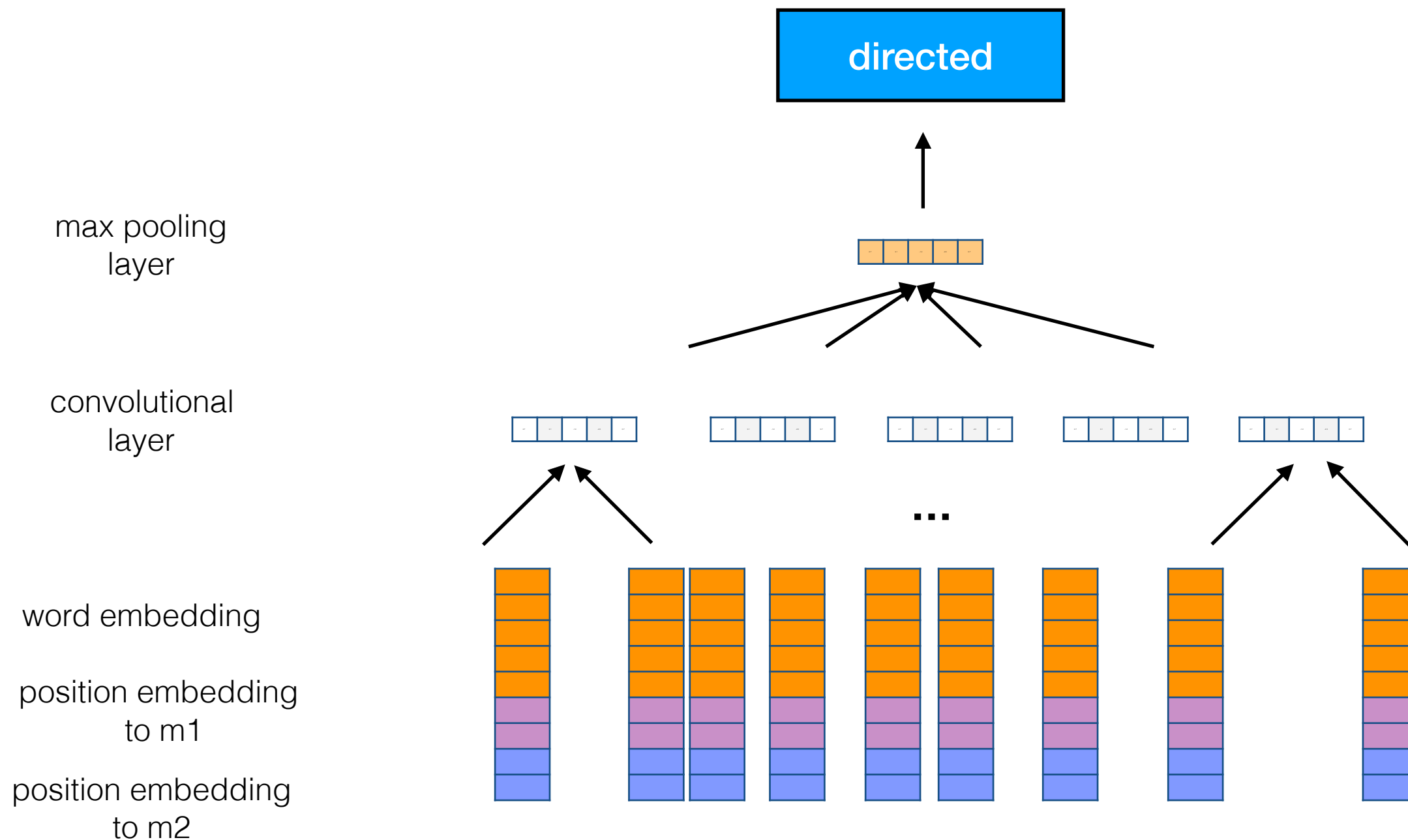
- Labels are assigned to a set of sentences, each containing the pair of entities m_1 and m_2 ; not all of those sentences express the relation between m_1 and m_2 .

Attention

- Let's incorporate structure (and parameters) into a network that captures which **sentences** in the input we should be **attending** to (and which we can ignore).

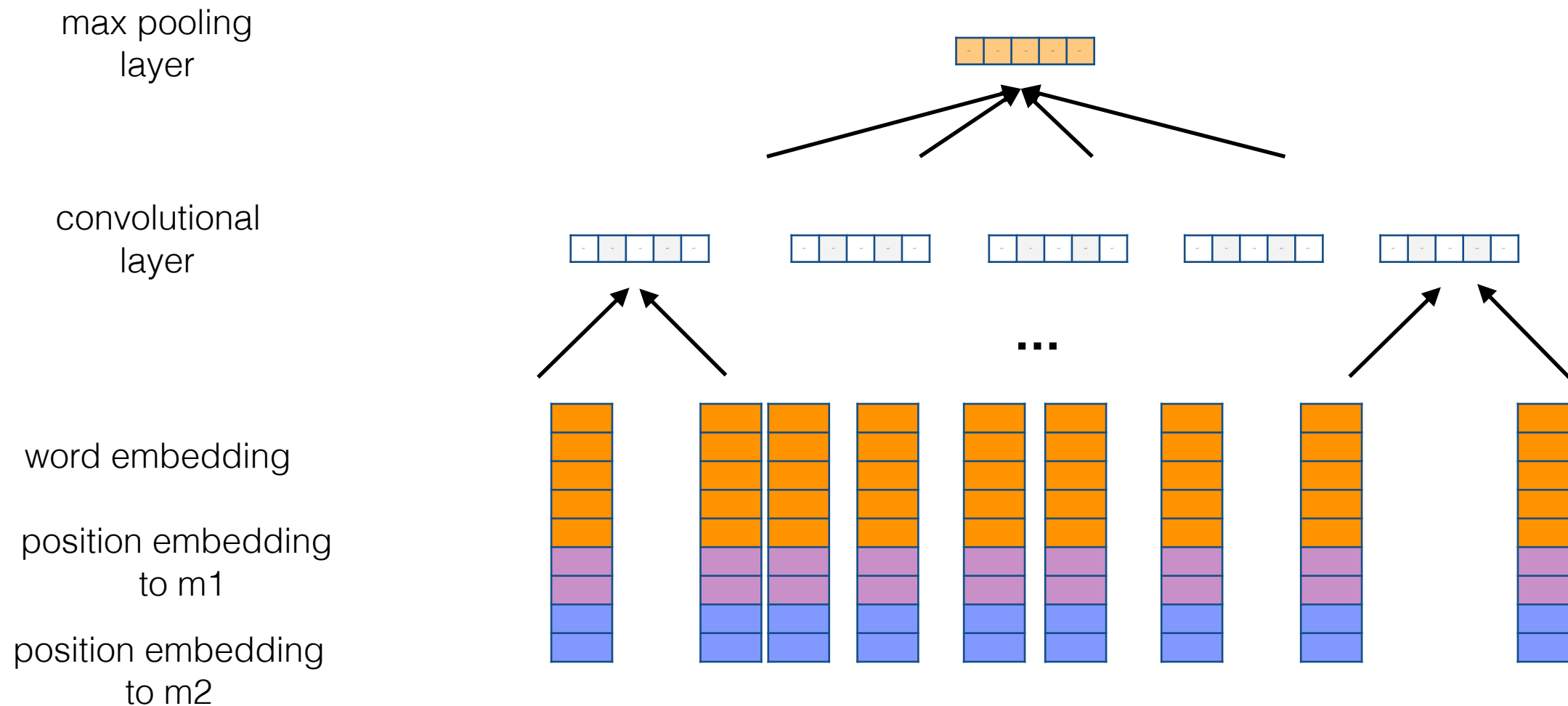


Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)

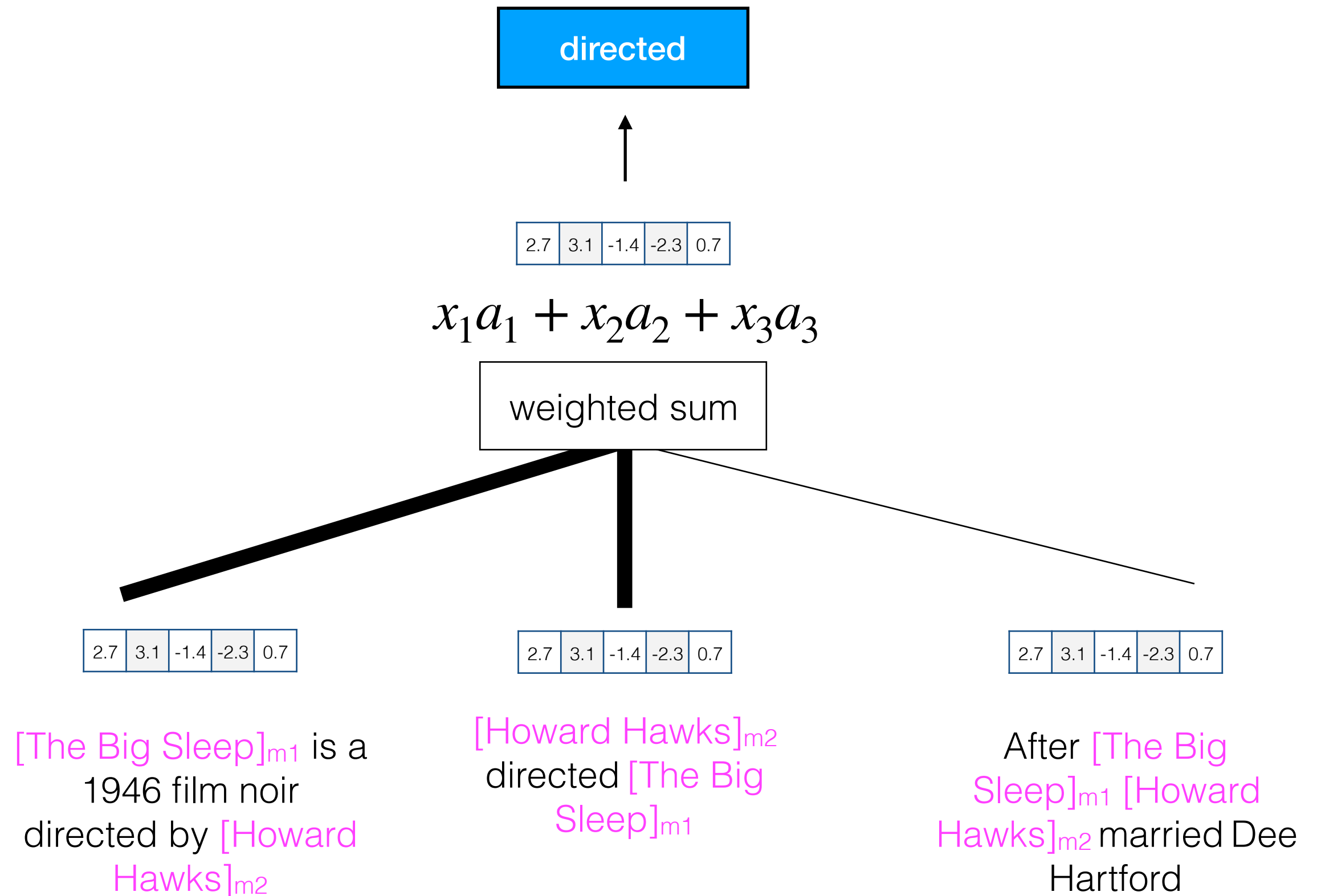


[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}

Now we just have an
encoding of a sentence



[The Big Sleep]_{m1} is a 1946 film noir directed by [Howard Hawks]_{m2}



sentence
encoding

Information Extraction

- Named entity recognition
- Entity linking
- Relation extraction
- Templated filling
- Event detection
- Event coreference
- Extra-propositional information (veridicality, hedging)

Machine Translation

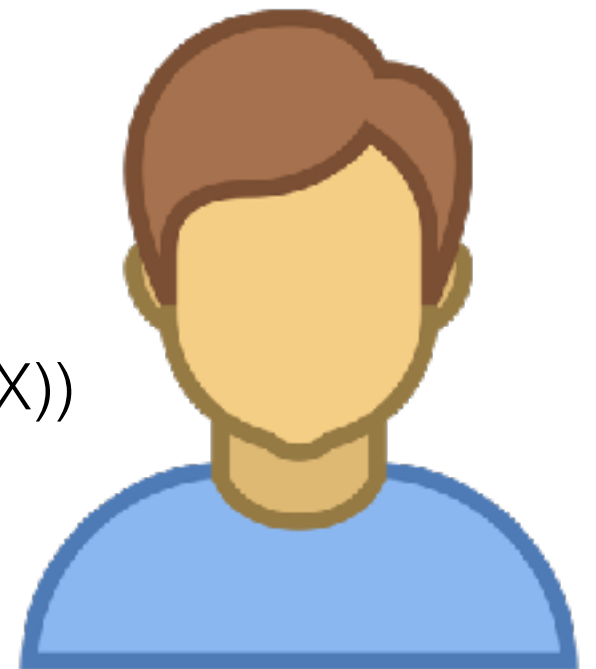


一天早上我穿着睡衣射了一只大象



encode(X)

decode(encode(X))



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



[All](#)

[Images](#)

[News](#)

[Videos](#)

[Maps](#)

[More](#)

[Settings](#)

[Tools](#)

About 9,230 results (0.43 seconds)

Italian - detected ▼



English ▼



Lasciate ogni
speranza, voi
ch'entrate [Edit](#)

Abandon all hope, ye
who enter here

[Open in Google Translate](#)

[Feedback](#)

^ 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*)."

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/euoparl/>

Data

- Europarl (proceedings of European parliament, 50M words/language)
<http://www.statmt.org/europarl/>
- UN Corpus (United Nations documents, six languages, 300M words/language)
<http://www.euromatrixplus.net/multi-un/>
- Common crawl (Web documents, long tail of language pairs)

Evaluation

ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον

- Tell me Muse, of the man of many ways
- Sing to me of the man, Muse, the man of twists and turns
- Tell me about a complicated man

Lattimore 1965

Eagles 1996

Wilson 2018

Evaluation

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

$$p_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

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

Hypothesis translation

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

Appeared
calm
when
he
was
taken
to
the
American

plane
,
which
will
to
Miami
,
Florida
.

Reference translations

Orejuela appeared calm 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.

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

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.

Hypothesis translation

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

Reference translations

Orejuela appeared calm 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.

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

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

$$p_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

- We could optimize the score by minimizing the denominator (the number of ngrams generated)

- Brevity penalty:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

- c = length of hypothesis translation
- r = length of closest reference translation

Statistical MT

Noisy Channel

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

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

Noisy Channel

This the
translation model

This is just a language
model for the target
language

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

- 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

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

Estimate this from
parallel texts

Estimate this from
monolingual data

Statistical MT

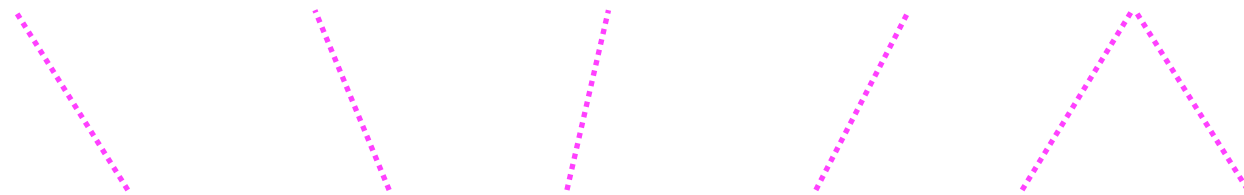
The statistical revolution in machine translation (1990) started by exploiting the structure of parallel sentences to learn the **translation model**.

Lasciate ogni speranza, voi ch'entrate

Abandon all hope, you who enter!

Statistical MT

Lasciate ogni speranza, voi ch'entrate



Abandon all hope, you who enter!

mi lasciate in pace



Leave me in peace

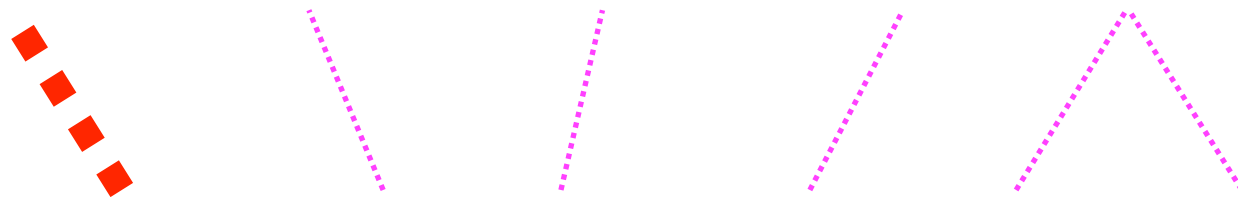
Lasciate i monti



Leave the mountains

Statistical MT

Lasciate ogni speranza, voi ch'entrate



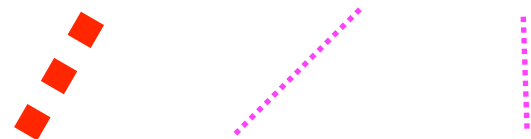
Abandon all hope, you who enter!

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



IBM Alignment models

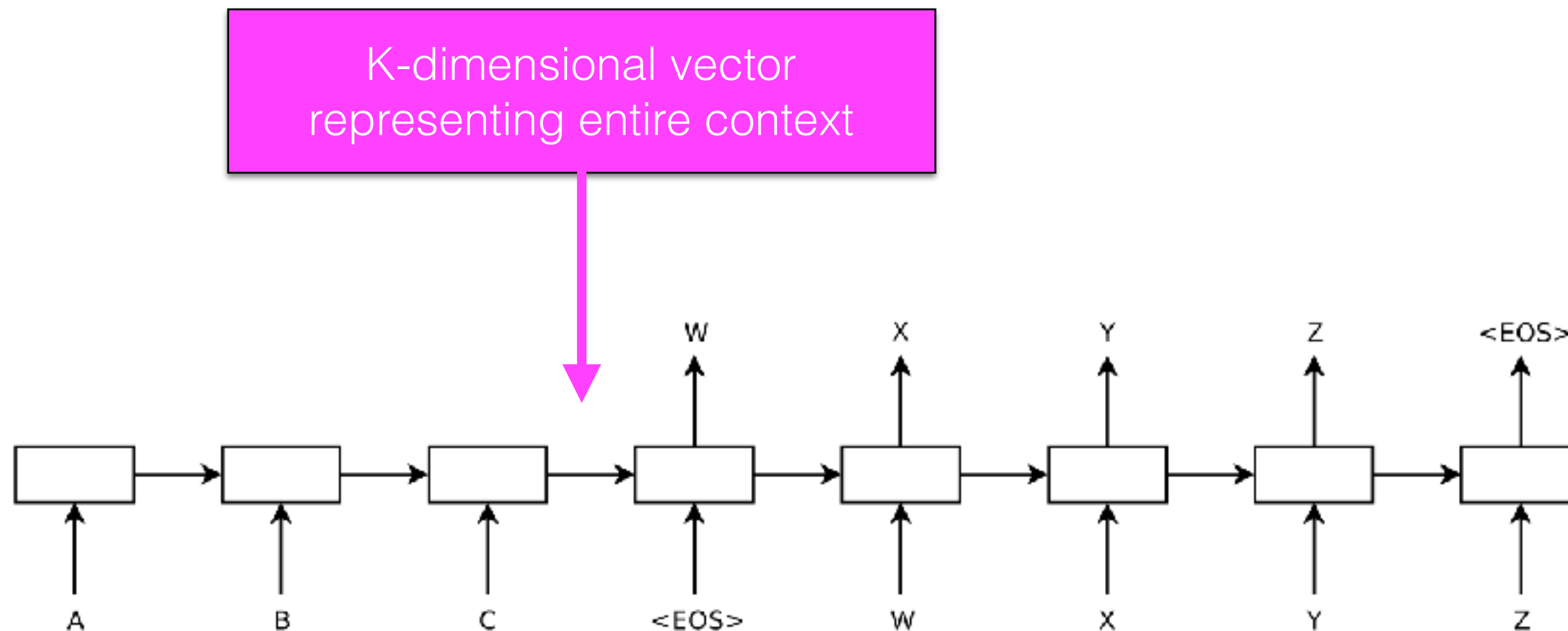
Model 1	Independent word translation (order doesn't matter)
Model 2	Word translation + distance between source and target position
Model 3	Word translation + fertility (how many target words a source word can align to)
Model 4	Word translation + relative ordering among target words of same source
Model 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

Neural MT

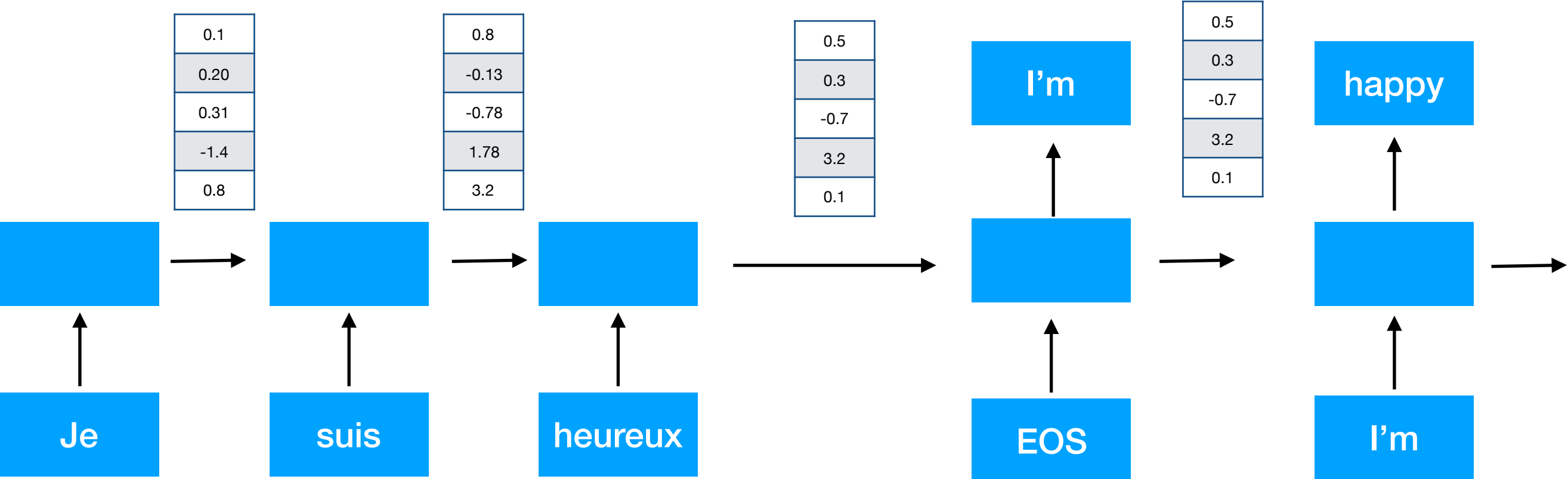
- Encoder-decoder
- Encoder-decoder + attention
- MT with *just* attention (Vaswani et al. 2018)

Encoder-decoder framework



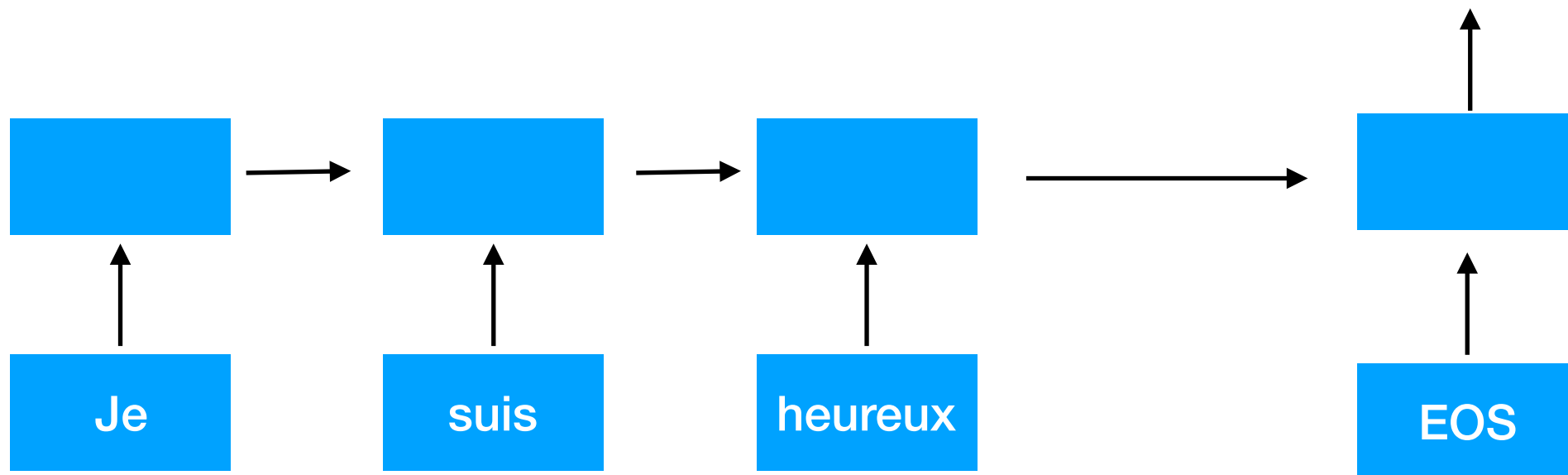
Sutskever et al. (2015);

Condition on word generated in translation



Training

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



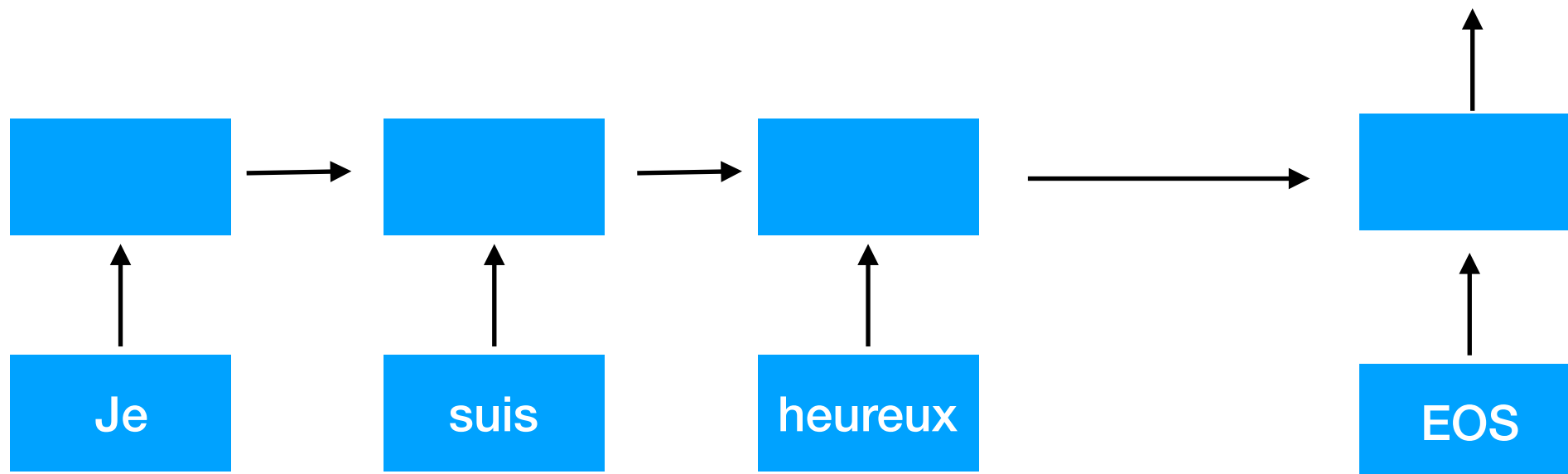
Training

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

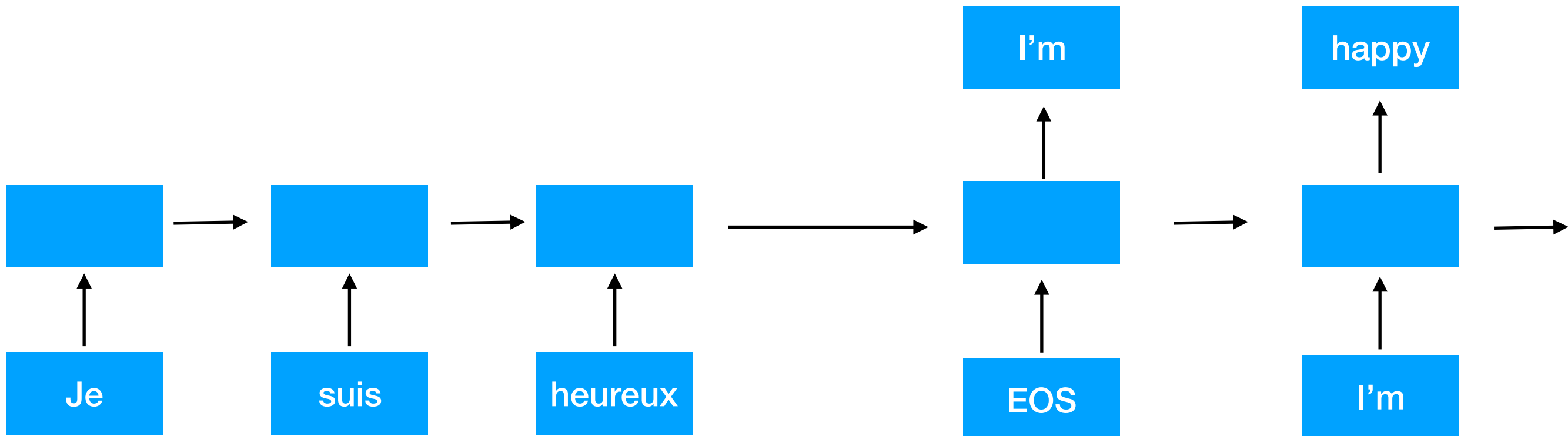


truth

happy	great	bad	ok	...
1	0	0	0	0

predicted

happy	great	bad	ok	...
0.13	0.08	0.01	0.03	0.009

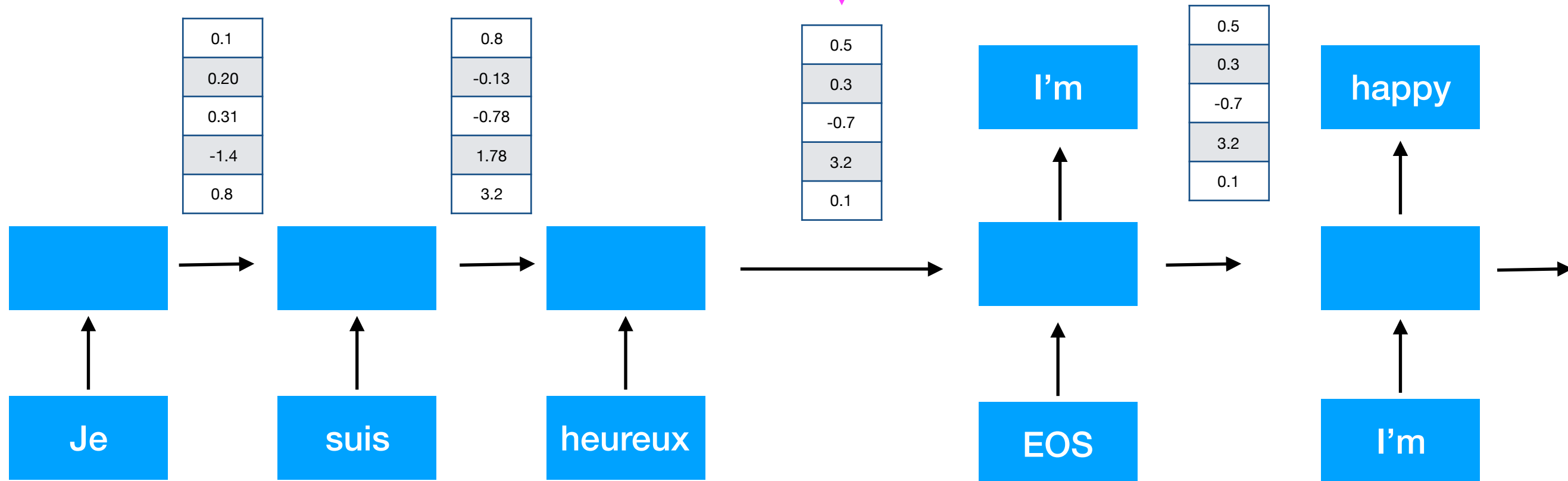


Encoder-decoder

- 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)

Encoder-decoder

The entire source sentence is summarized in this one vector



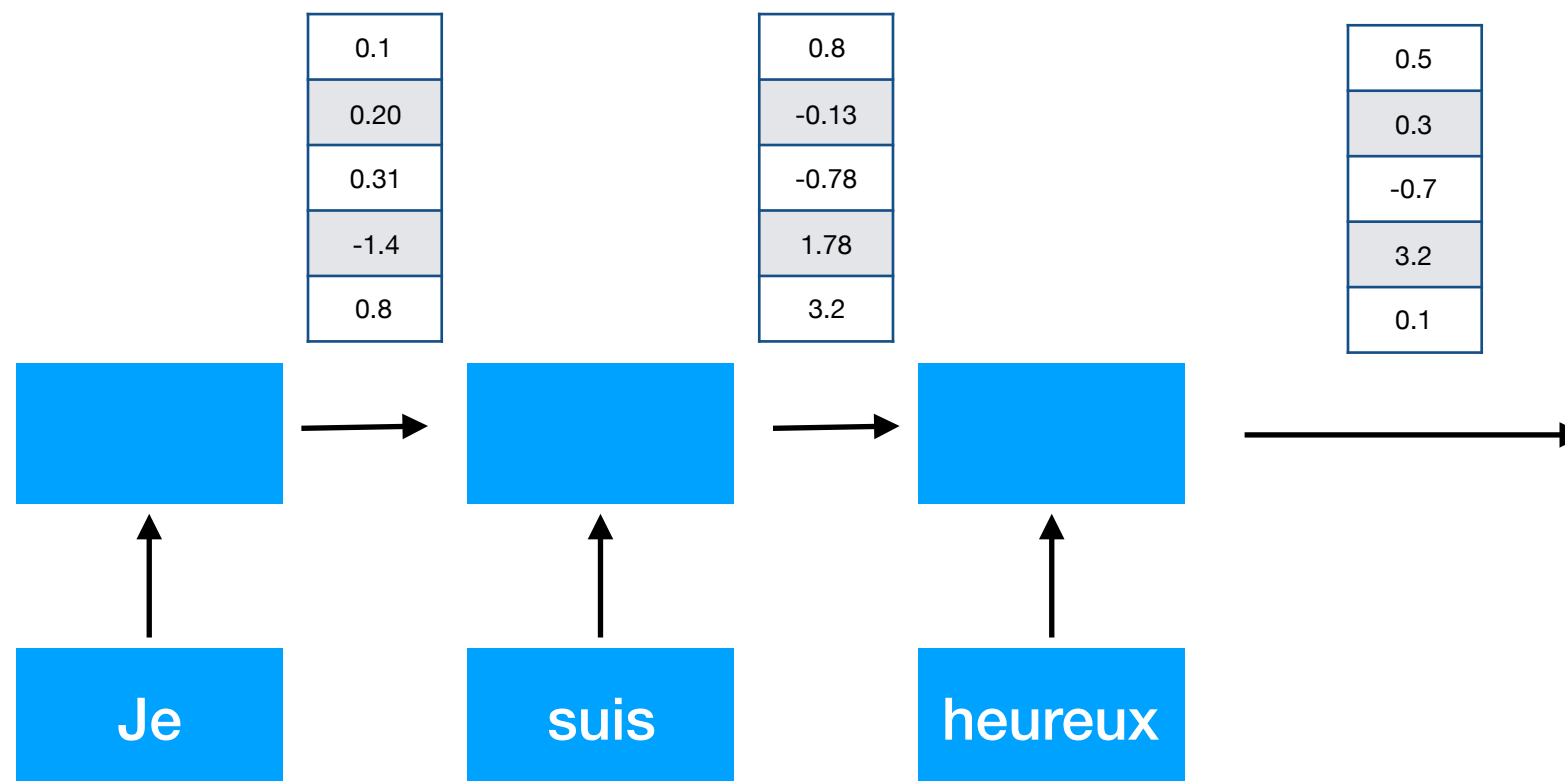
The decoder state depends just on the previous *state* and the previous *output*

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

Encoder-decoder with attention

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

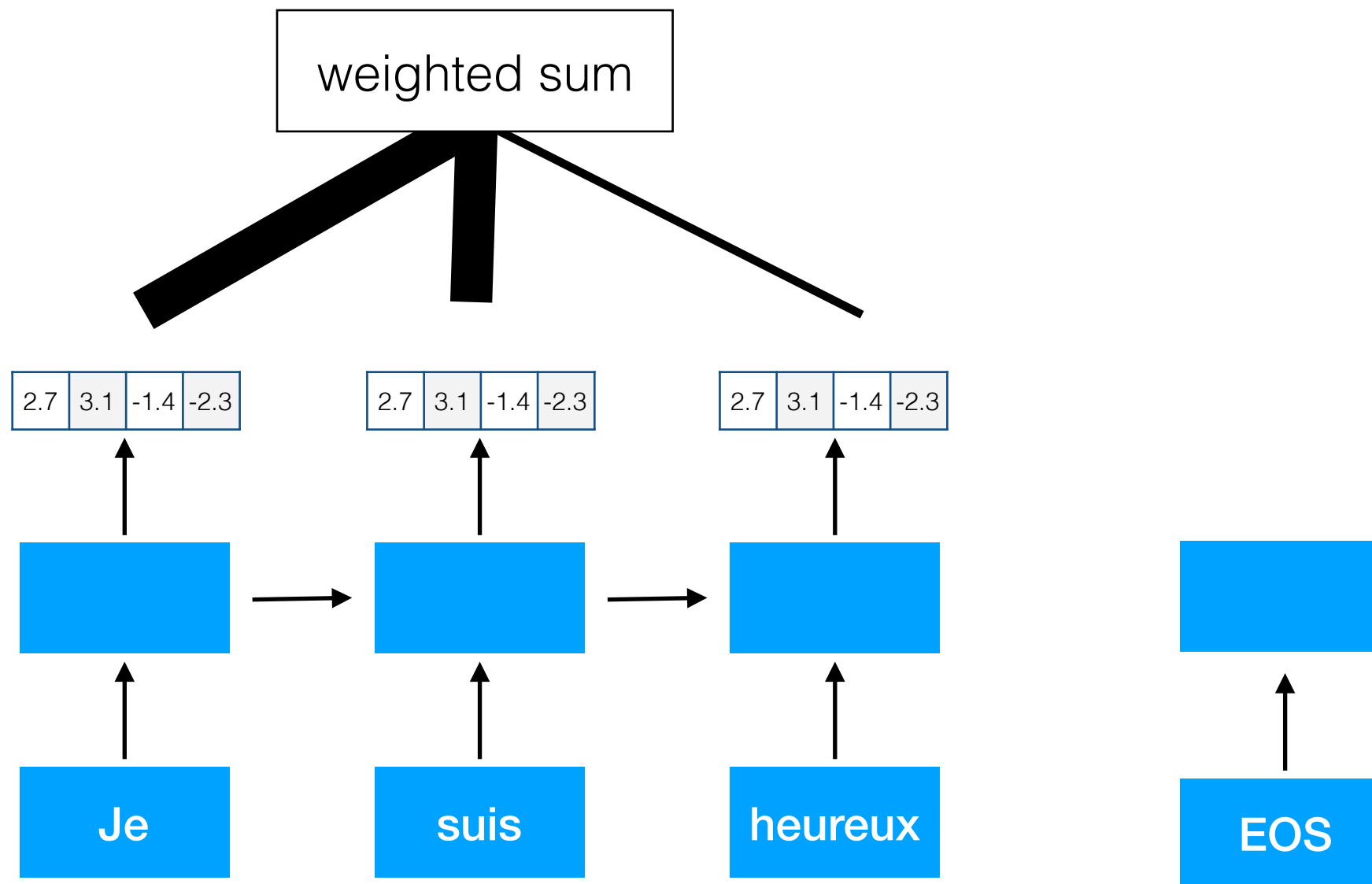
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



Encoder-decoder with attention

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

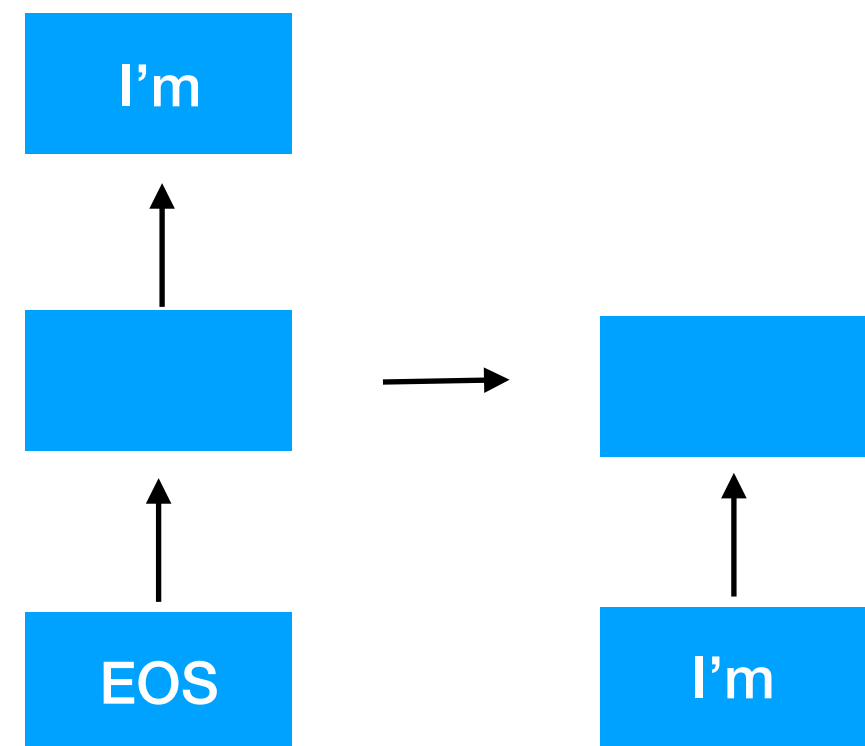
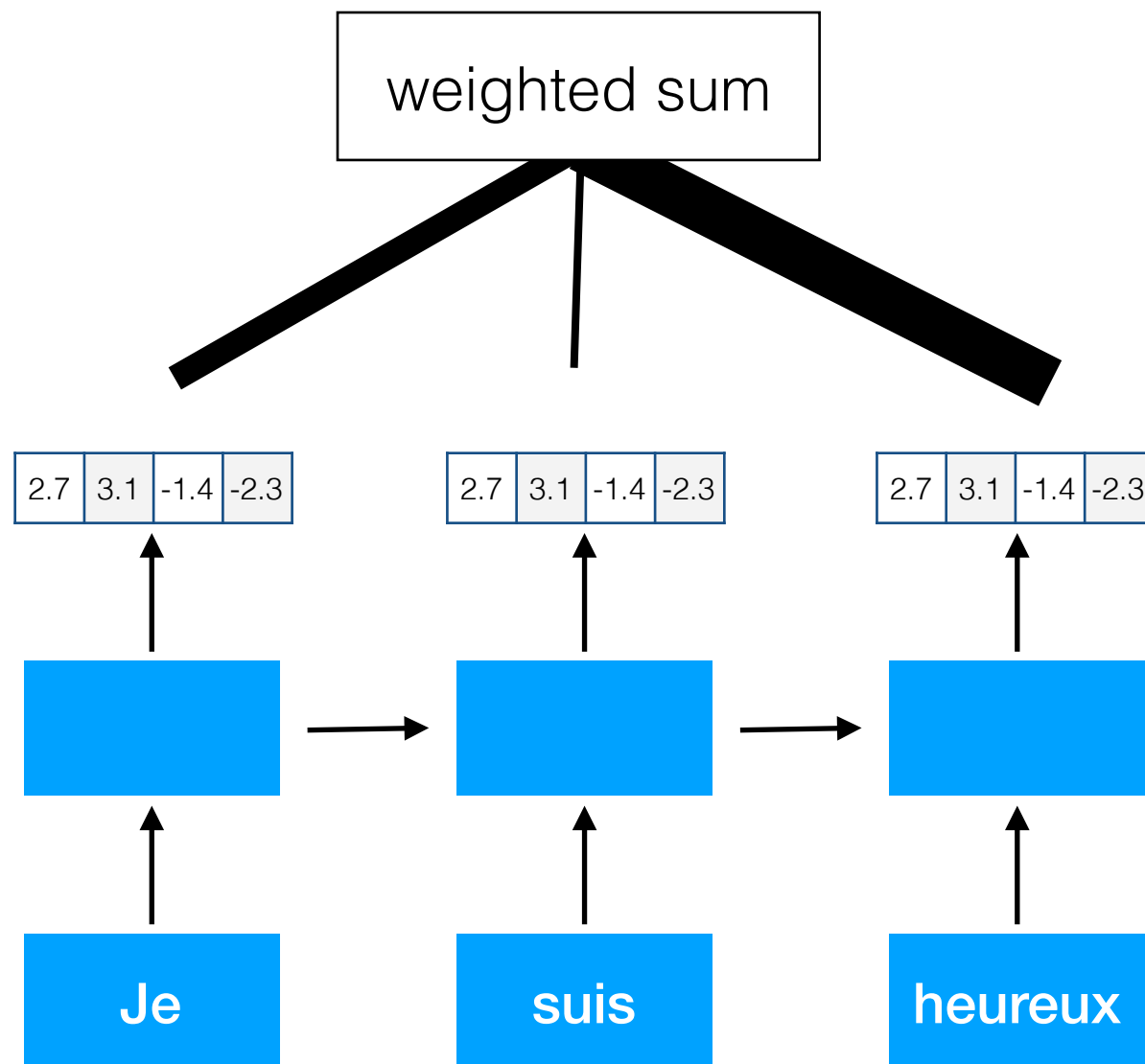
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$



Encoder-decoder with attention

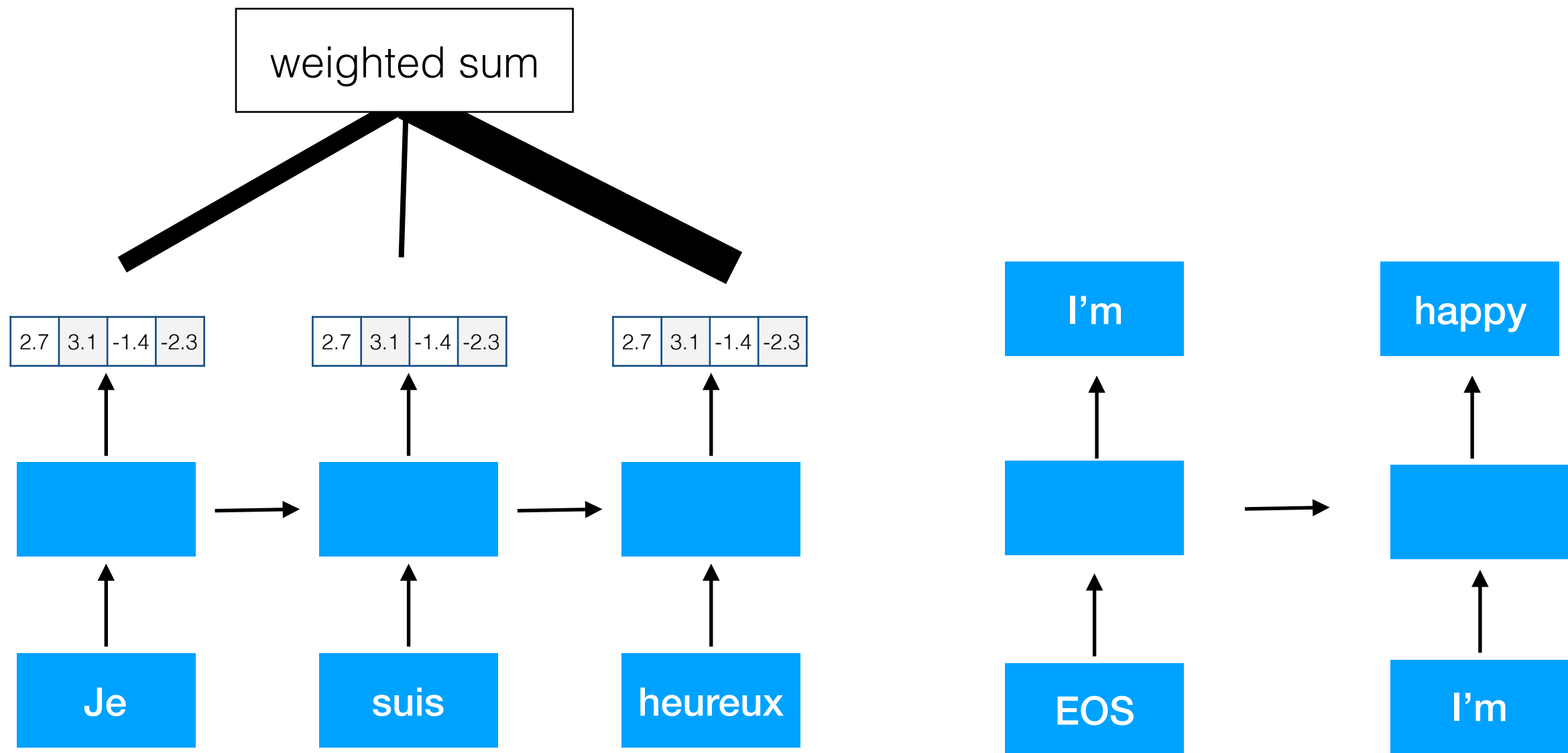
$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

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



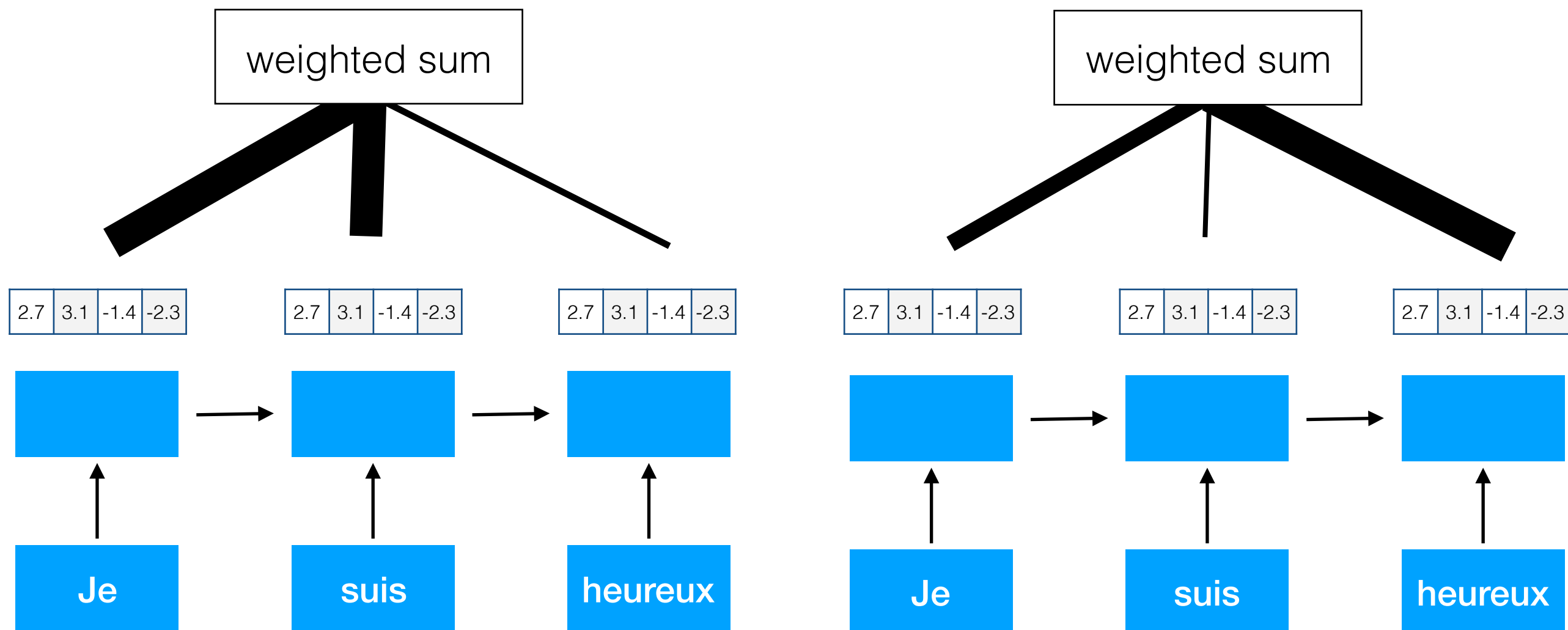
Encoder-decoder with attention

$$c = h_1a_1 + h_2a_2 + h_3a_3$$



Encoder-decoder with attention

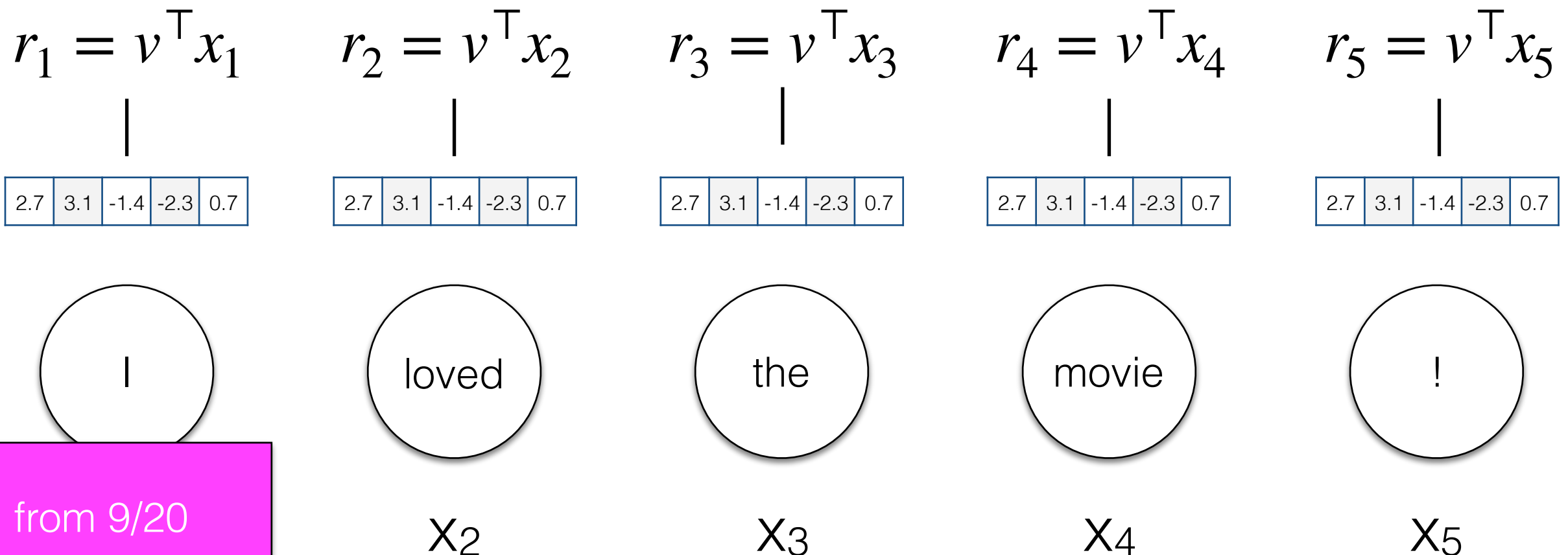
- Each time step in the decoder has its own weighted context vector



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

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

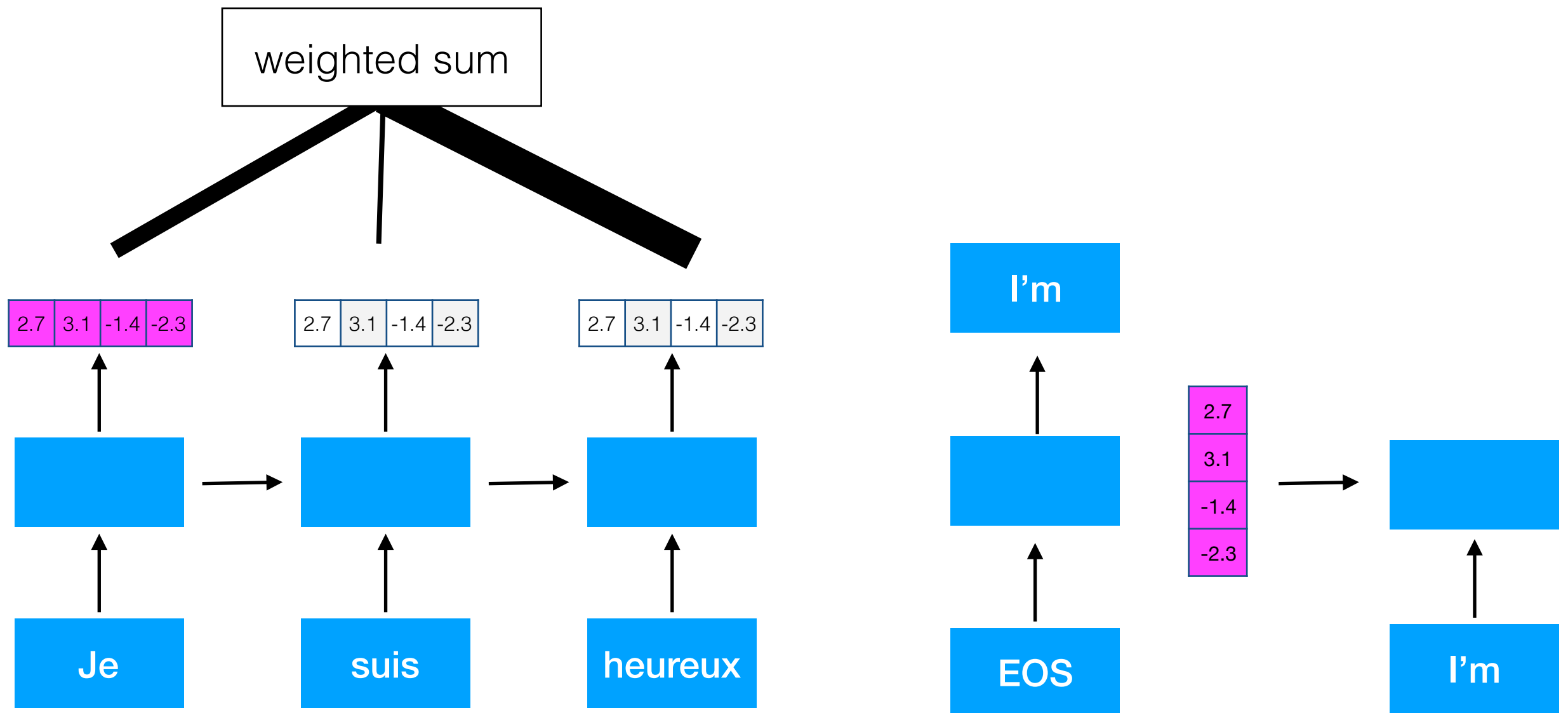
With document classification (lecture 9/20), we parameterized attention with a single vector v to be learned. Attention in an encoder-decoder network is a little different because we're comparing **a pair of vectors**.



Encoder-decoder with attention

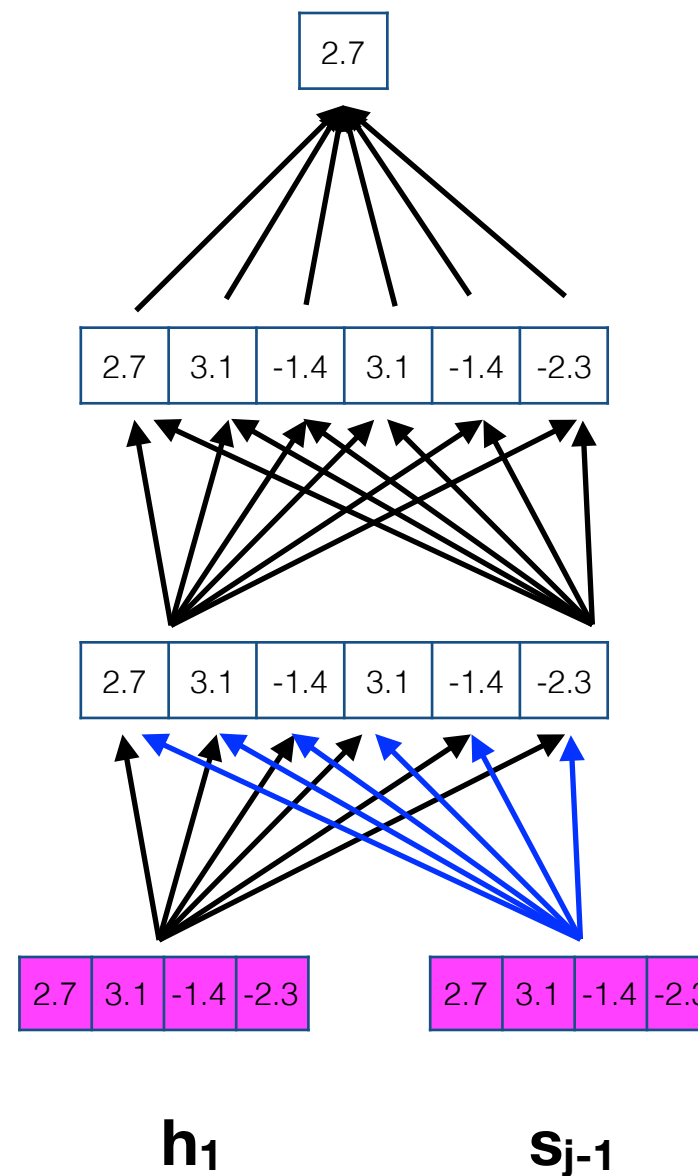
$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

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



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).



$$W_3 \in \mathbb{R}^6$$

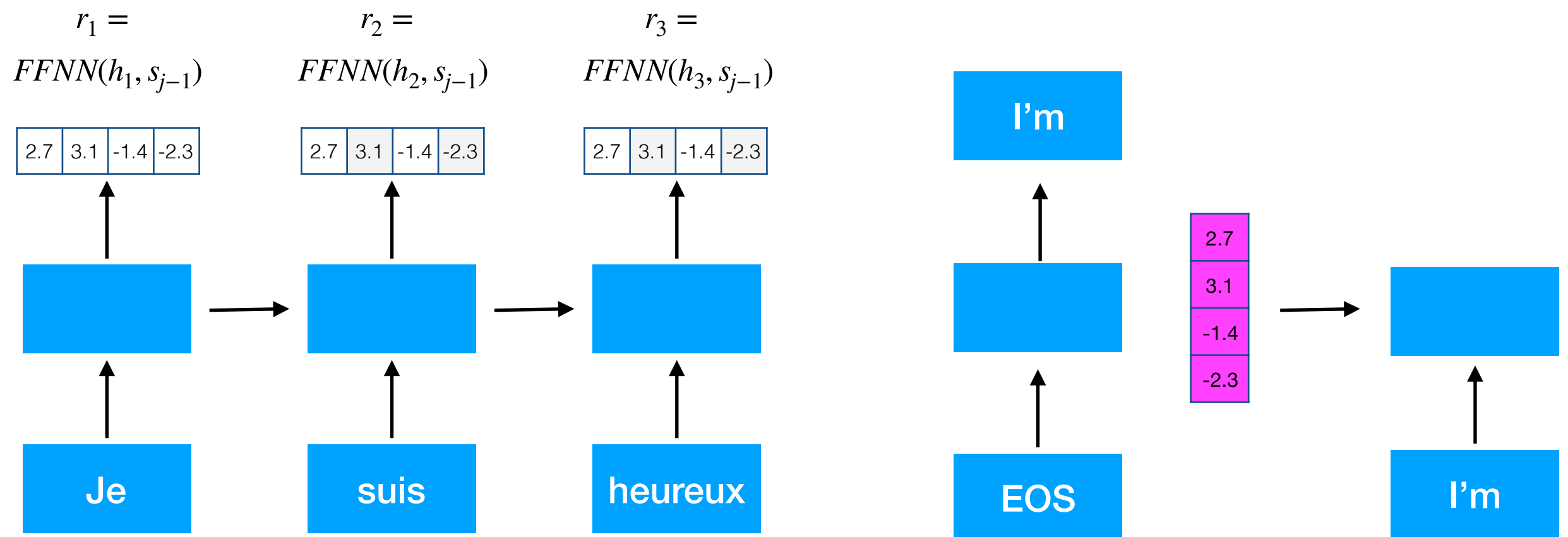
$$W_2 \in \mathbb{R}^{6 \times 6}$$

$$W_1 \in \mathbb{R}^{4 \times 6}$$

Encoder-decoder with attention

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

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



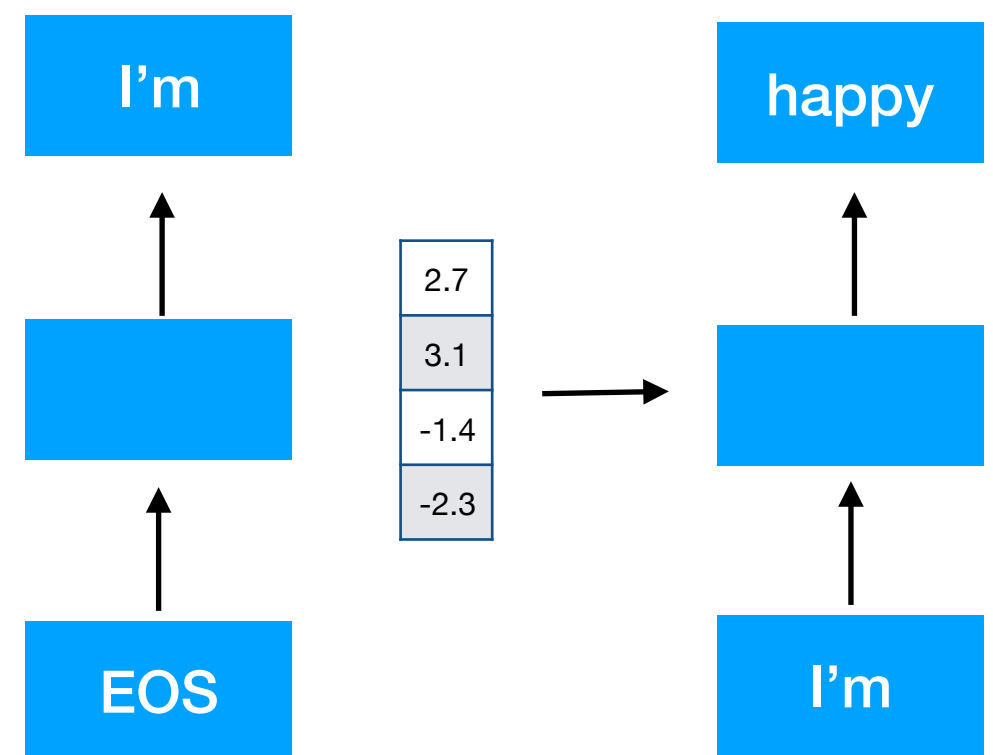
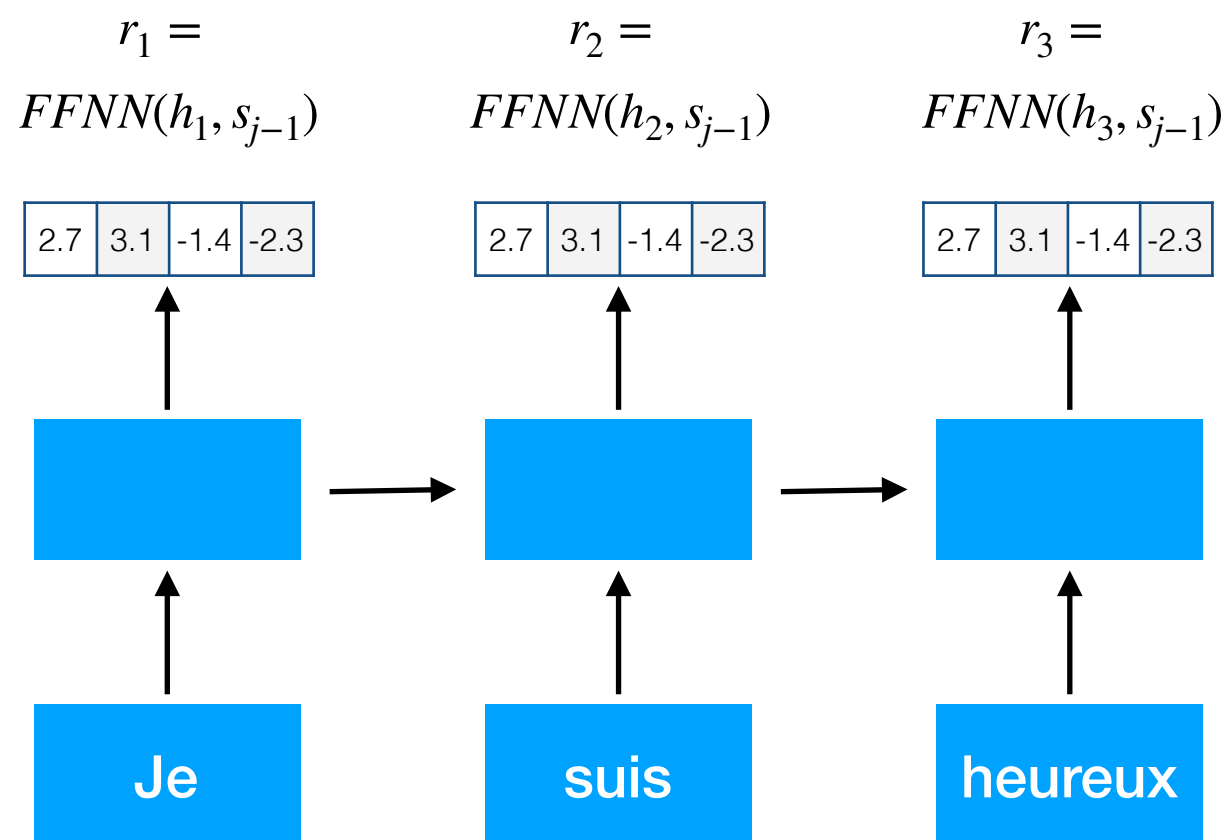
Encoder-decoder with attention

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

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

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

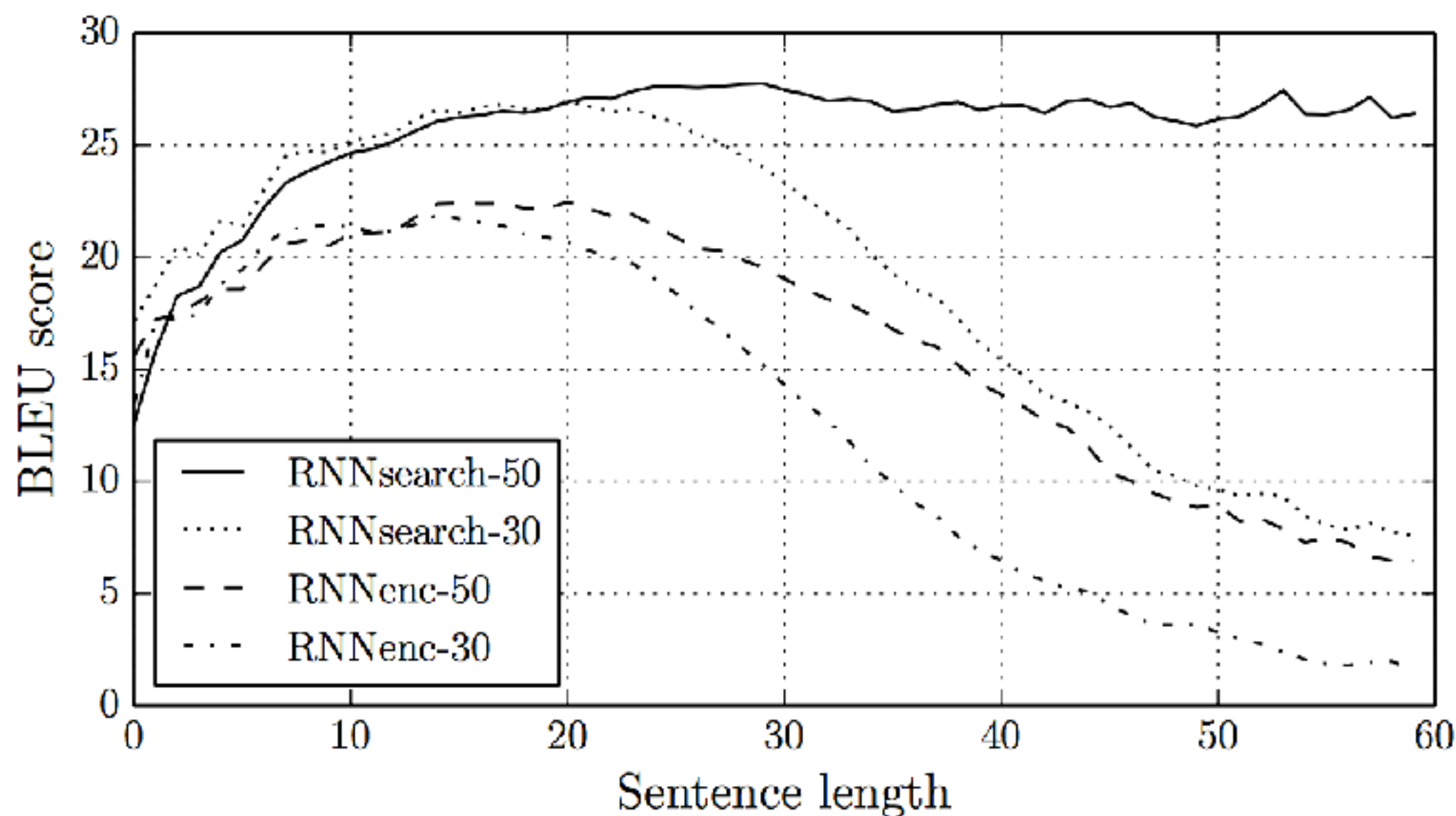
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

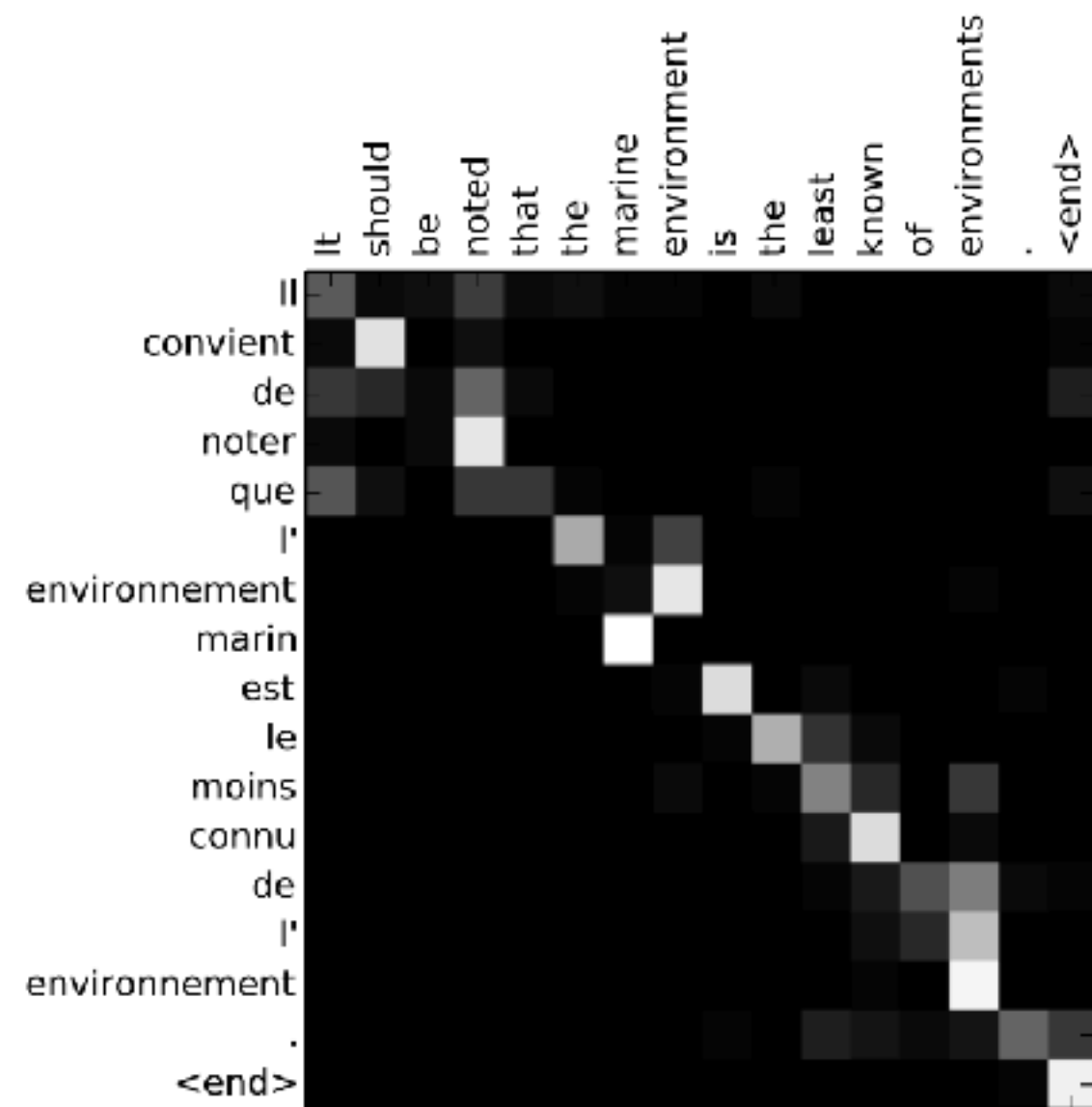
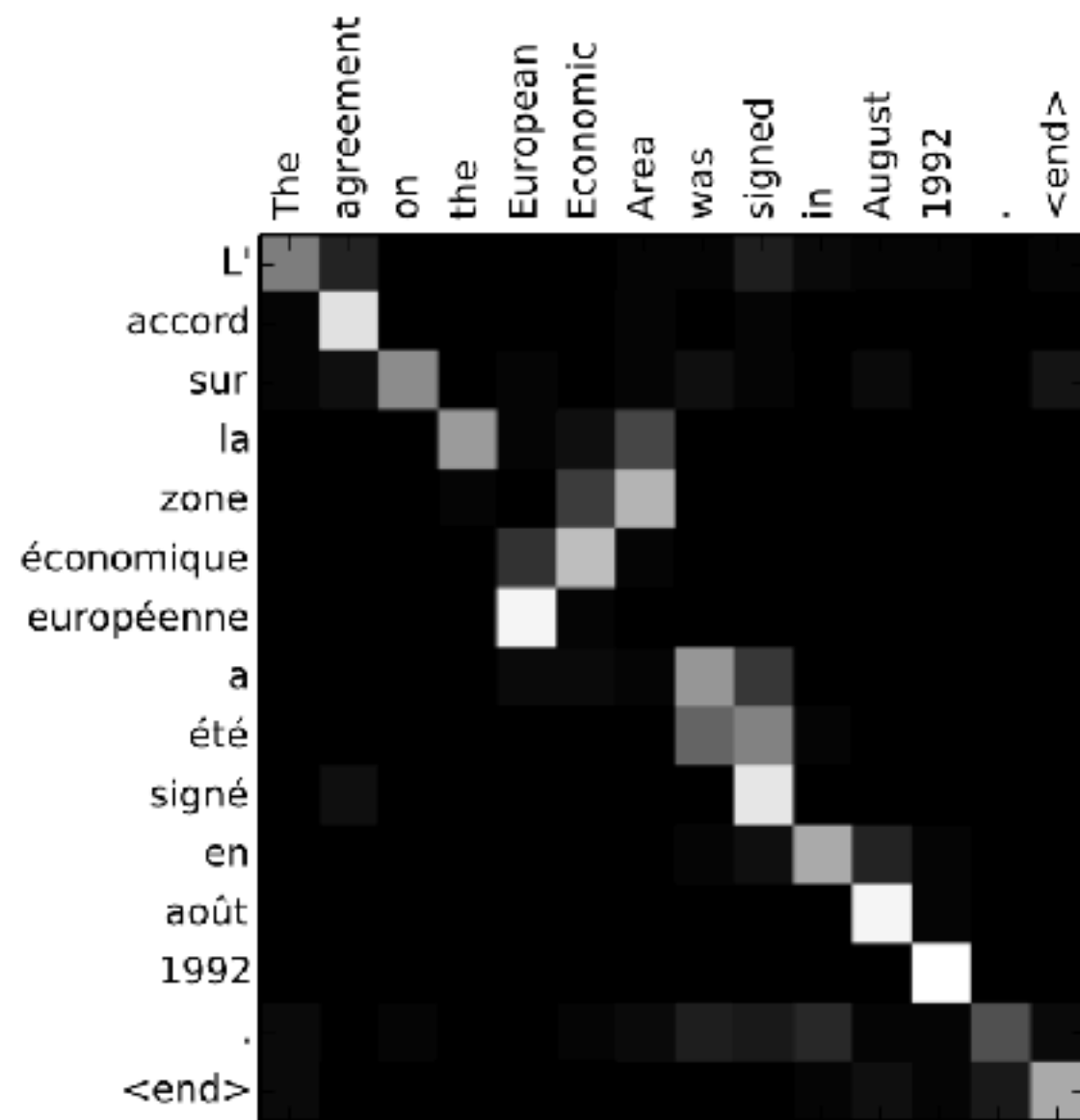


Attention

- For text classification, attention helps decide which words in the text are important for the label; a document has a single attention vector.
- 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

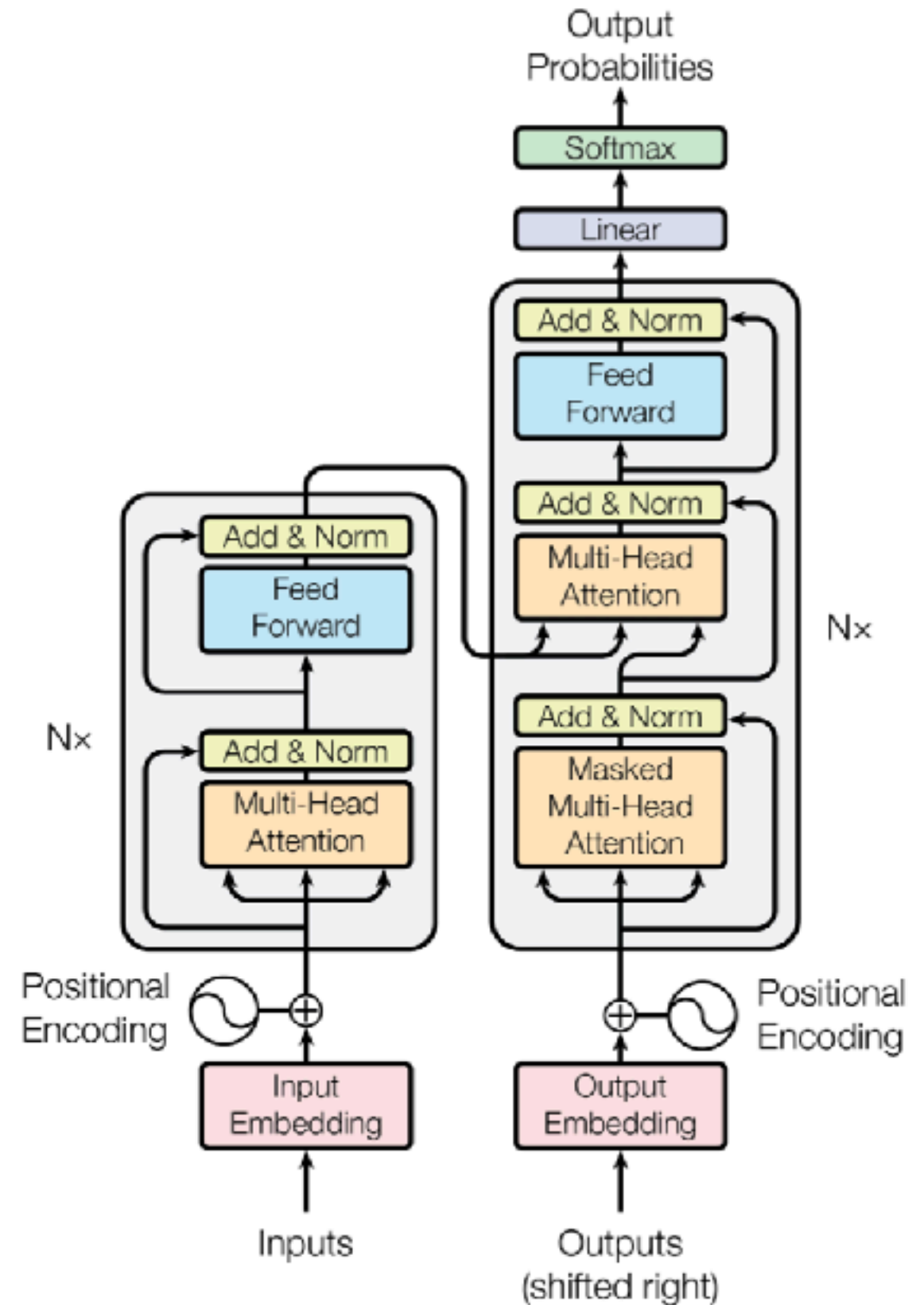




Attention

- RNNs are hard to parallelize; important factor for long sequence lengths.
- Attention gives us access to an entire input sequence. Why do we need recurrence at all?

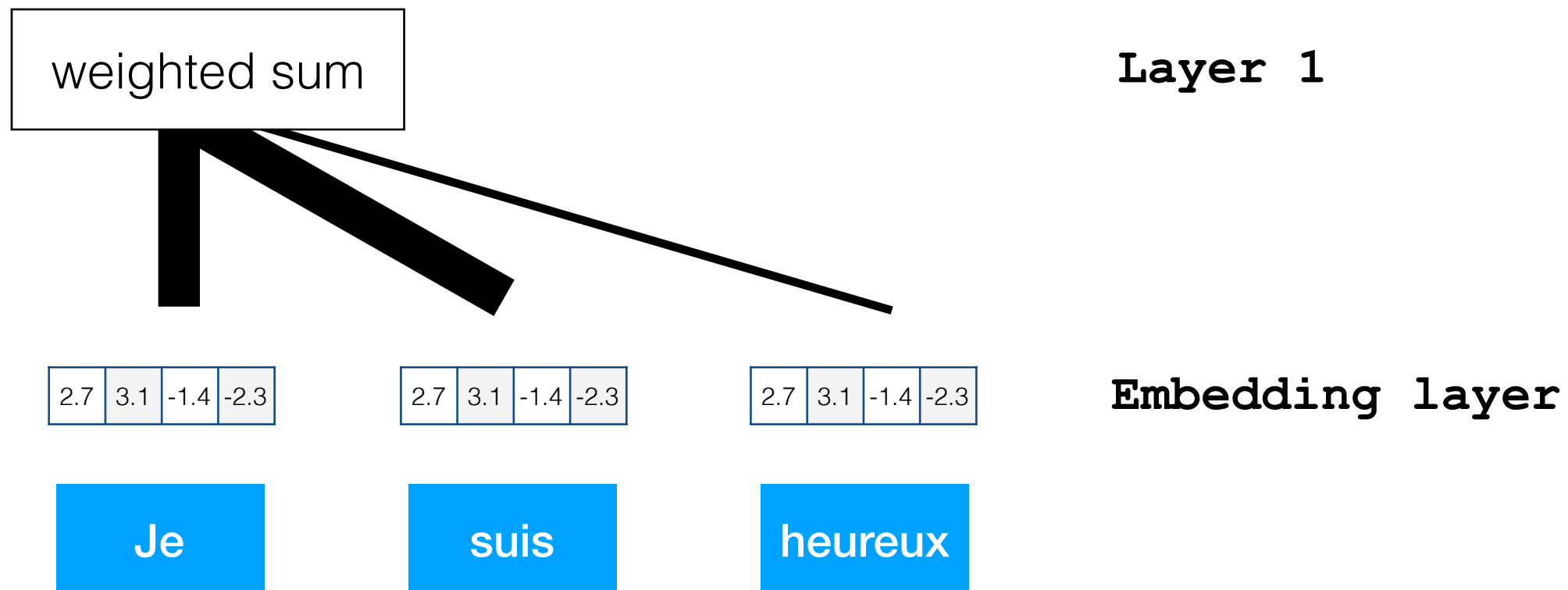
- Transformer network (Vaswani et al. 2017).



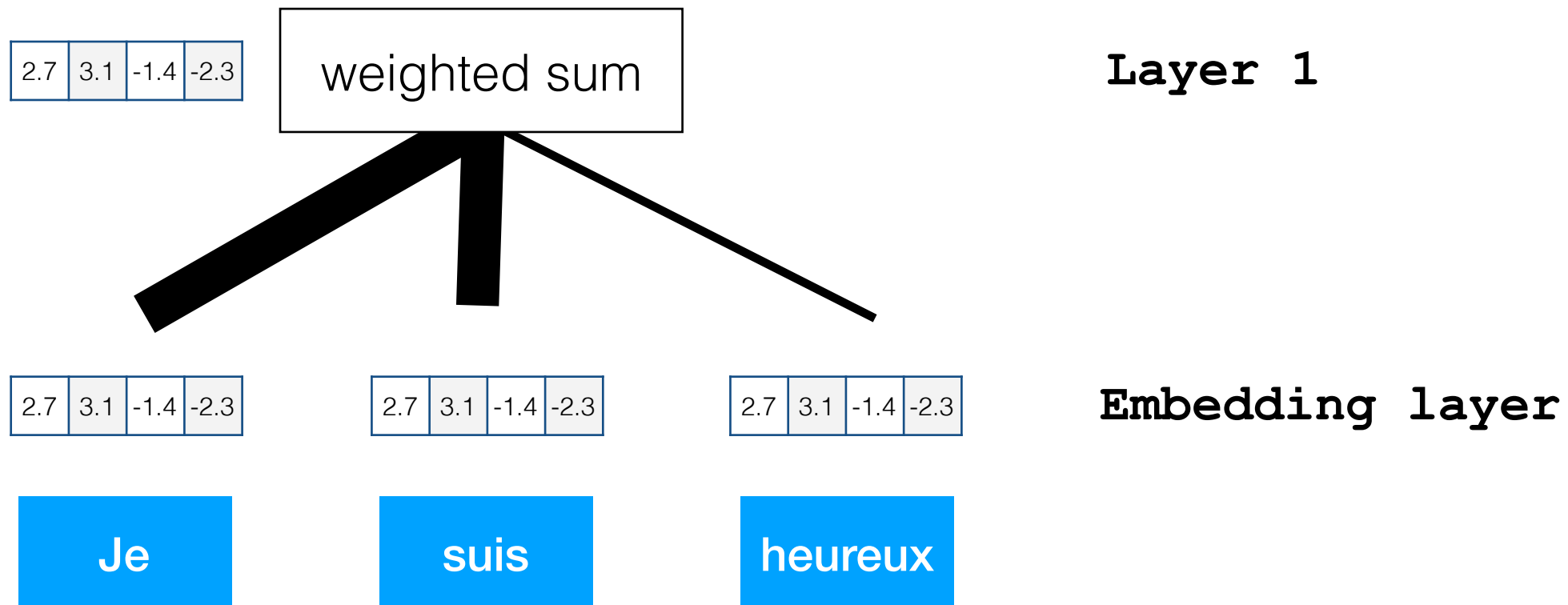
Self-attention

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

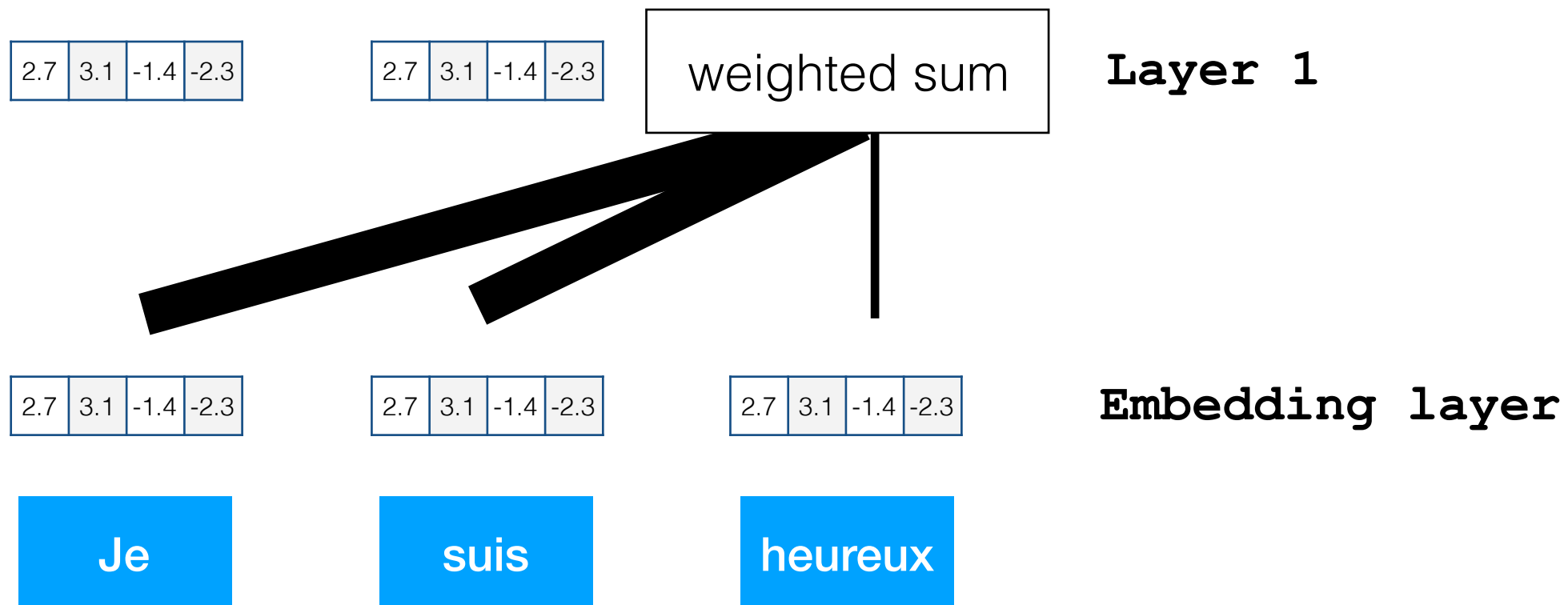
Self-attention



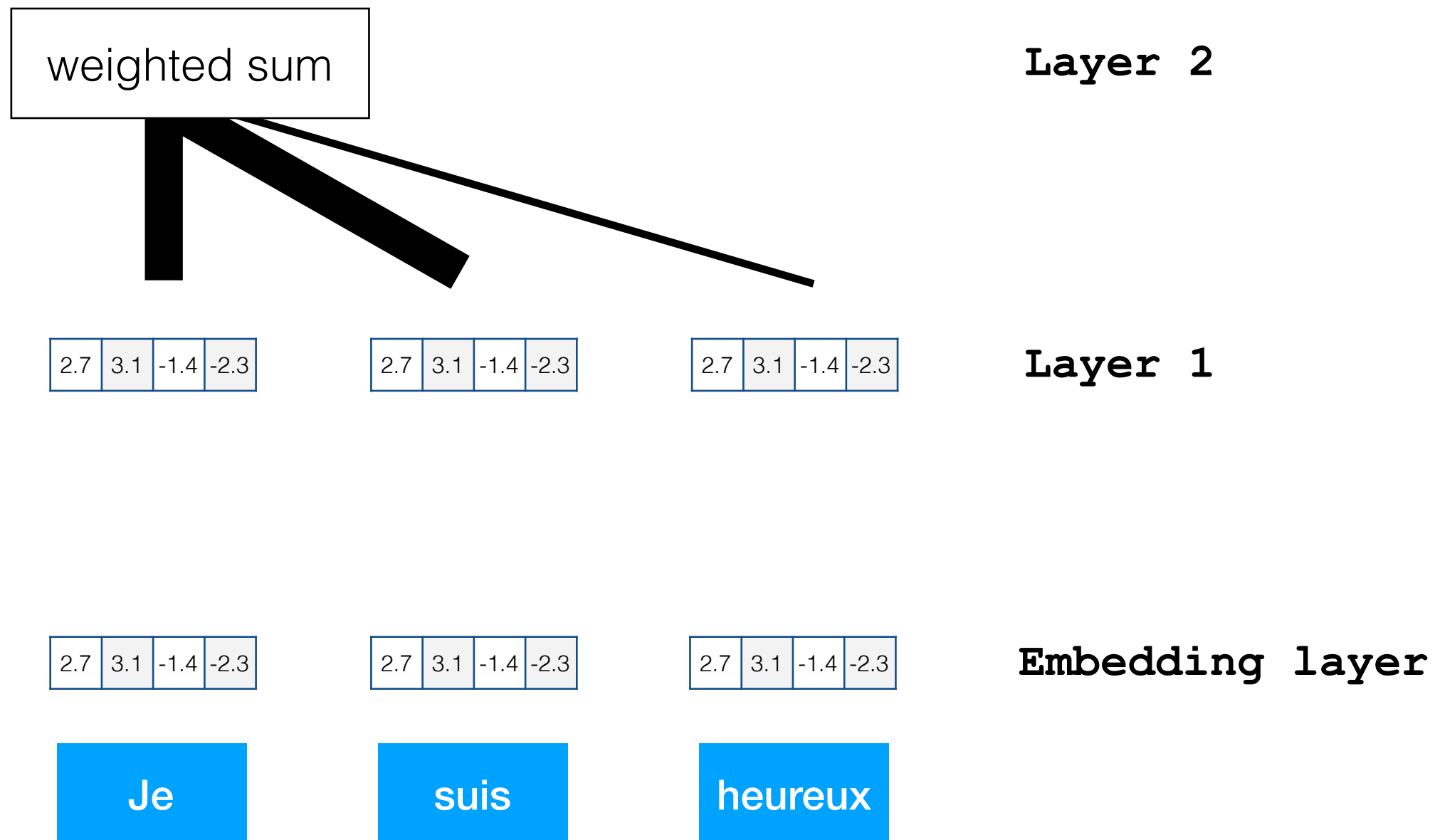
Self-attention



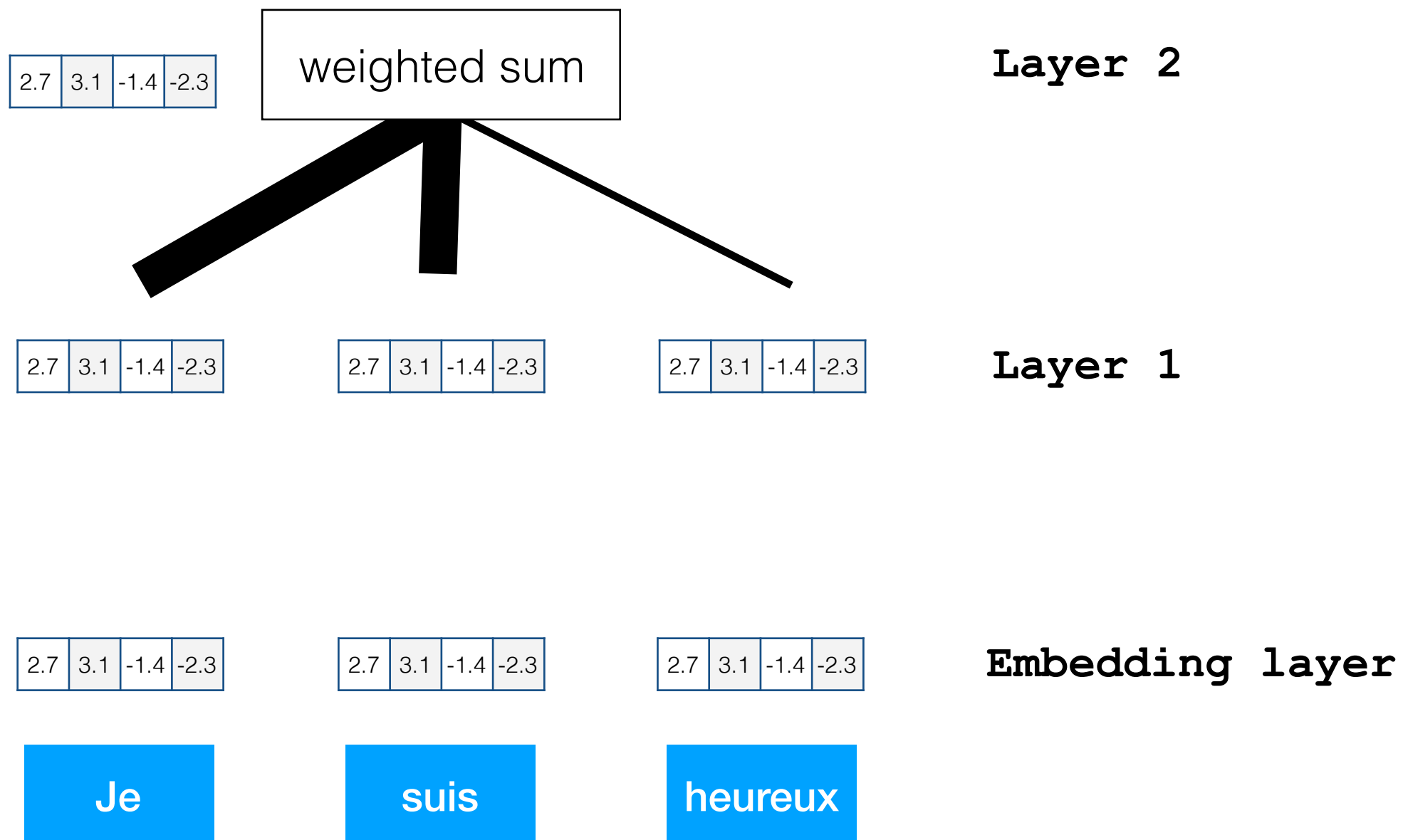
Self-attention



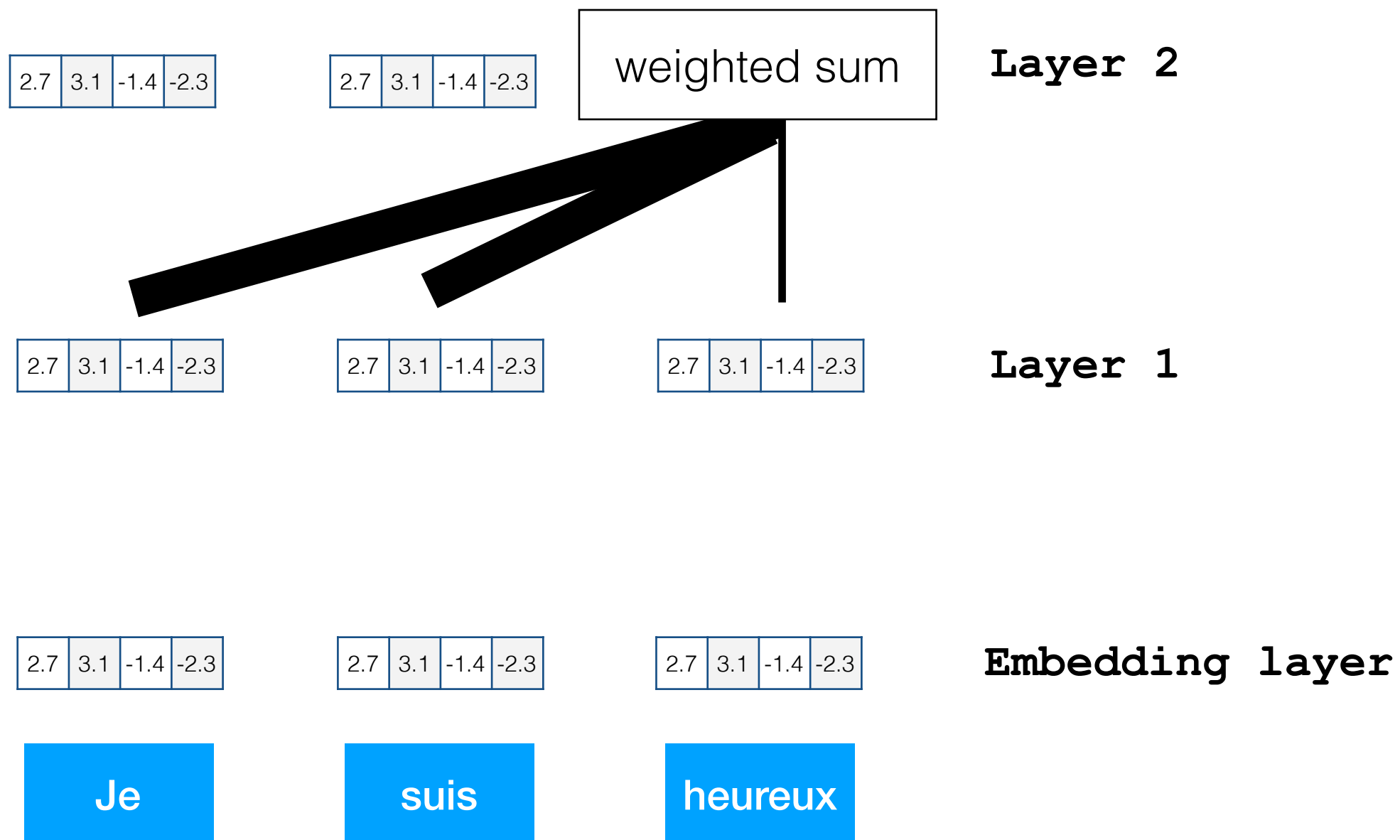
Self-attention



Self-attention

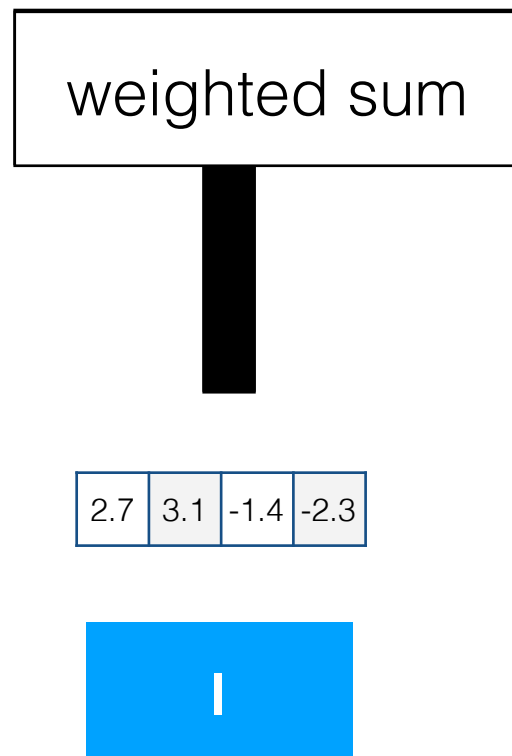


Self-attention



Self-attention

- In the decoder, self-attention attends only over words to the **left** of the position.



Self-attention

weighted sum



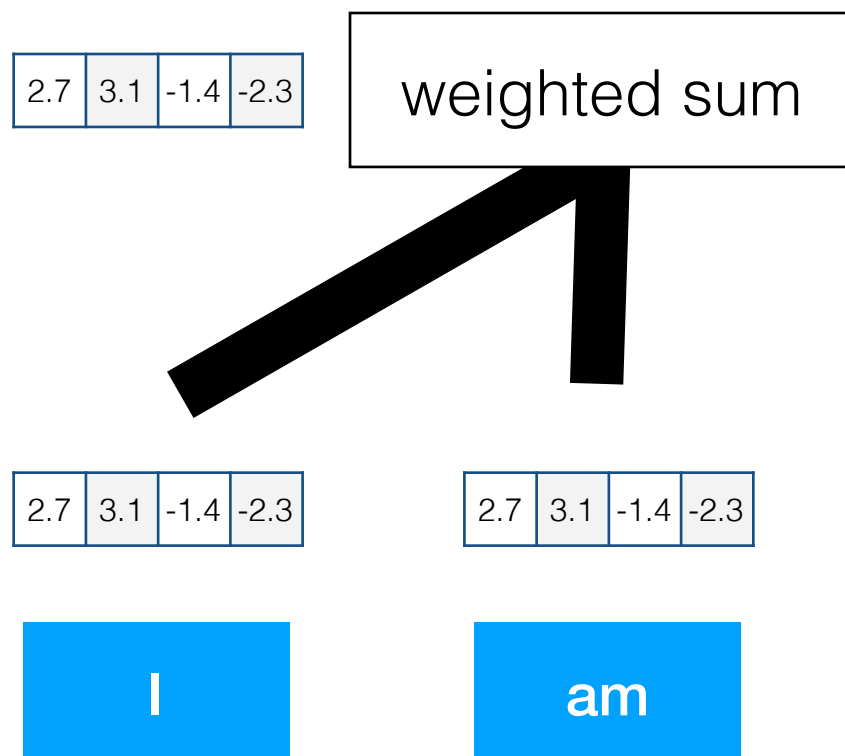
2.7	3.1	-1.4	-2.3
-----	-----	------	------

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



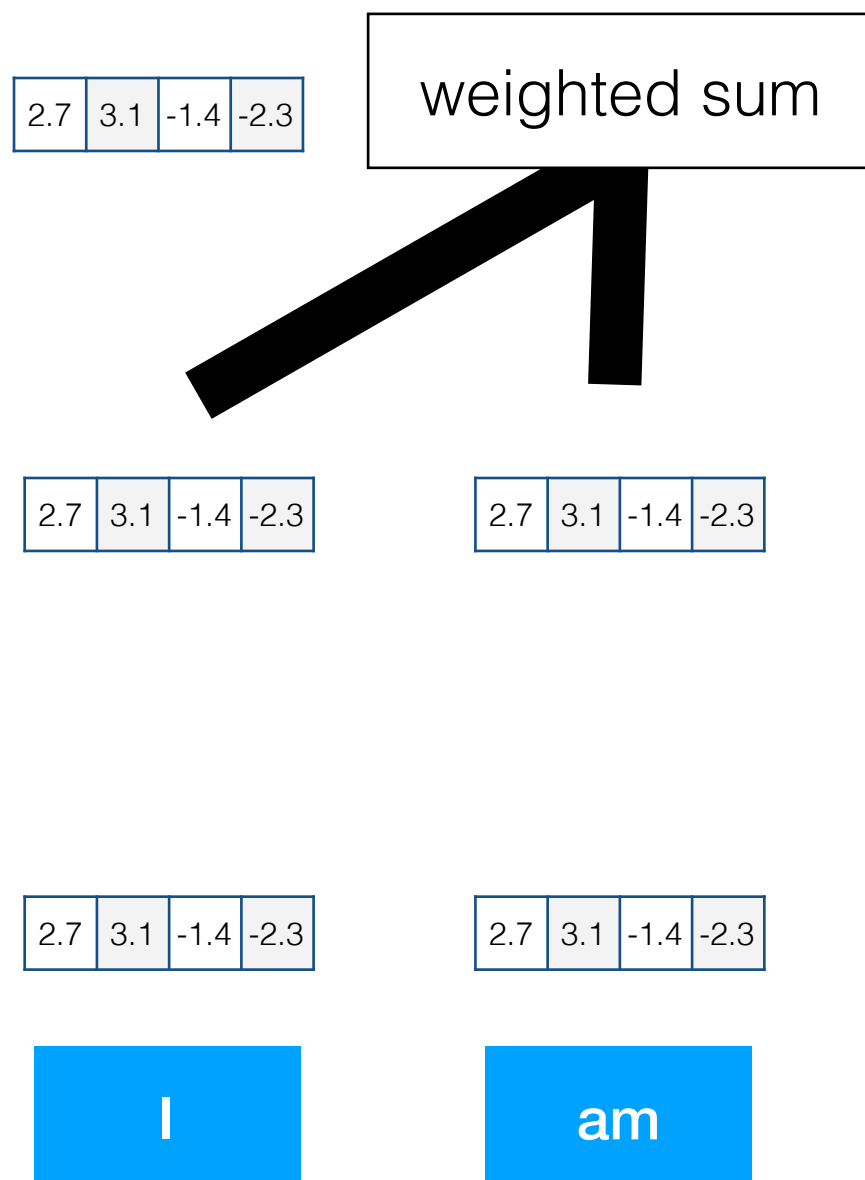
- In the decoder, self-attention attends only over words to the **left** of the position.

Self-attention



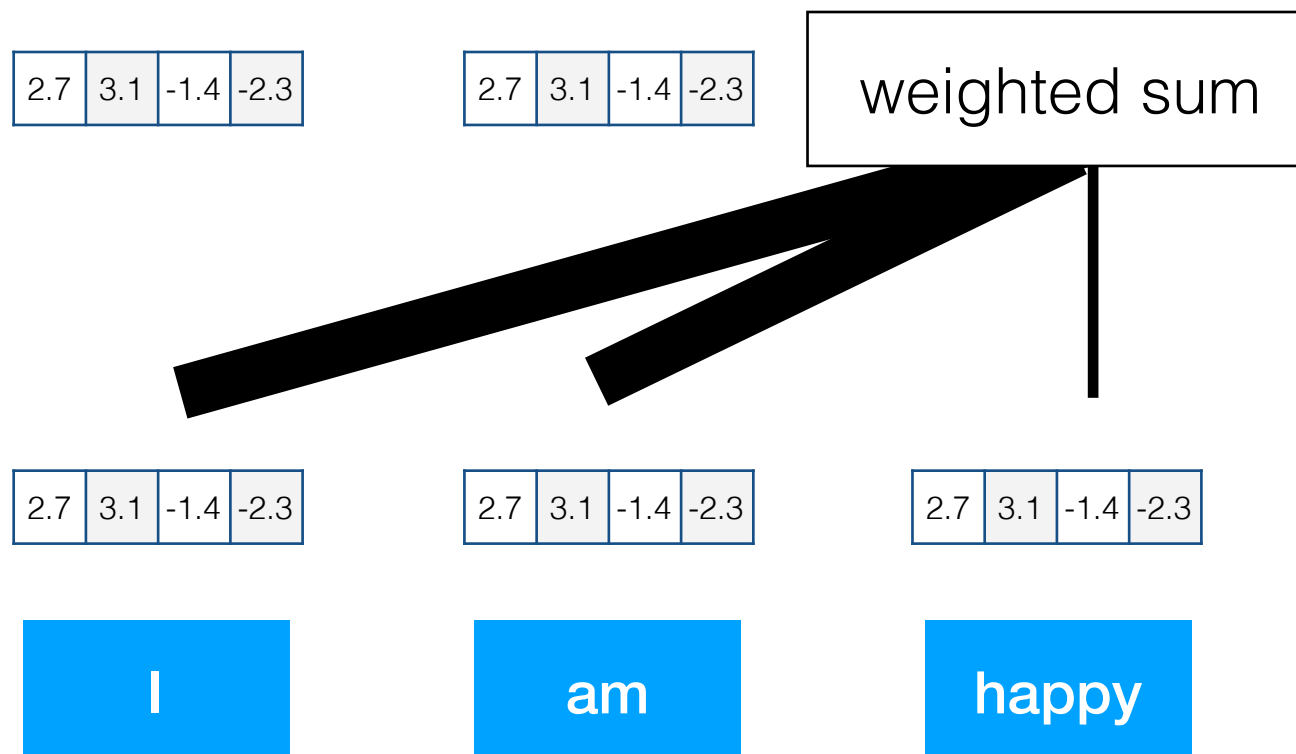
- In the decoder, self-attention attends only over words to the **left** of the position.

Self-attention



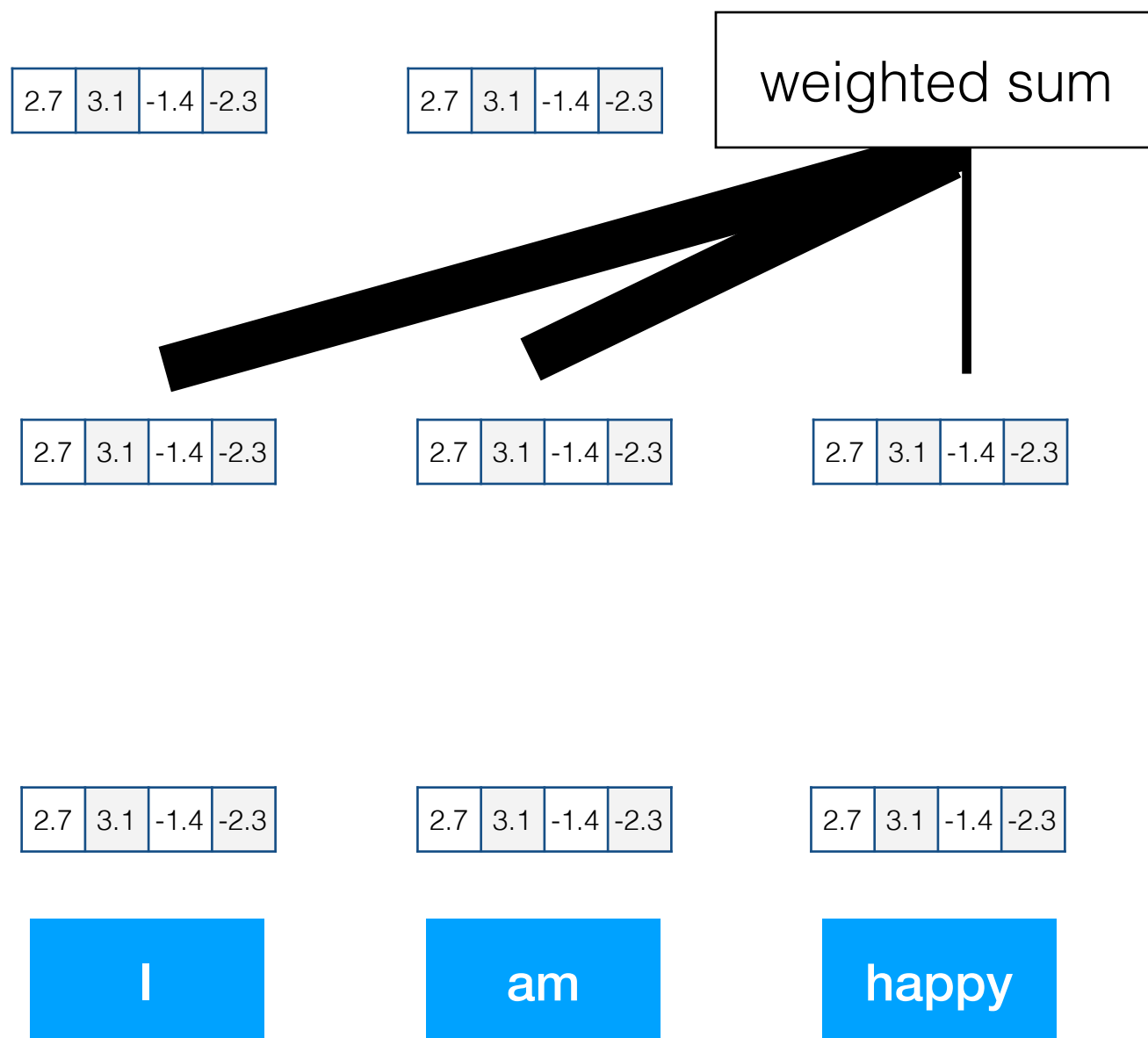
- In the decoder, self-attention attends only over words to the **left** of the position.

Self-attention



- In the decoder, self-attention attends only over words to the **left** of the position.

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 attention

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Je

suis

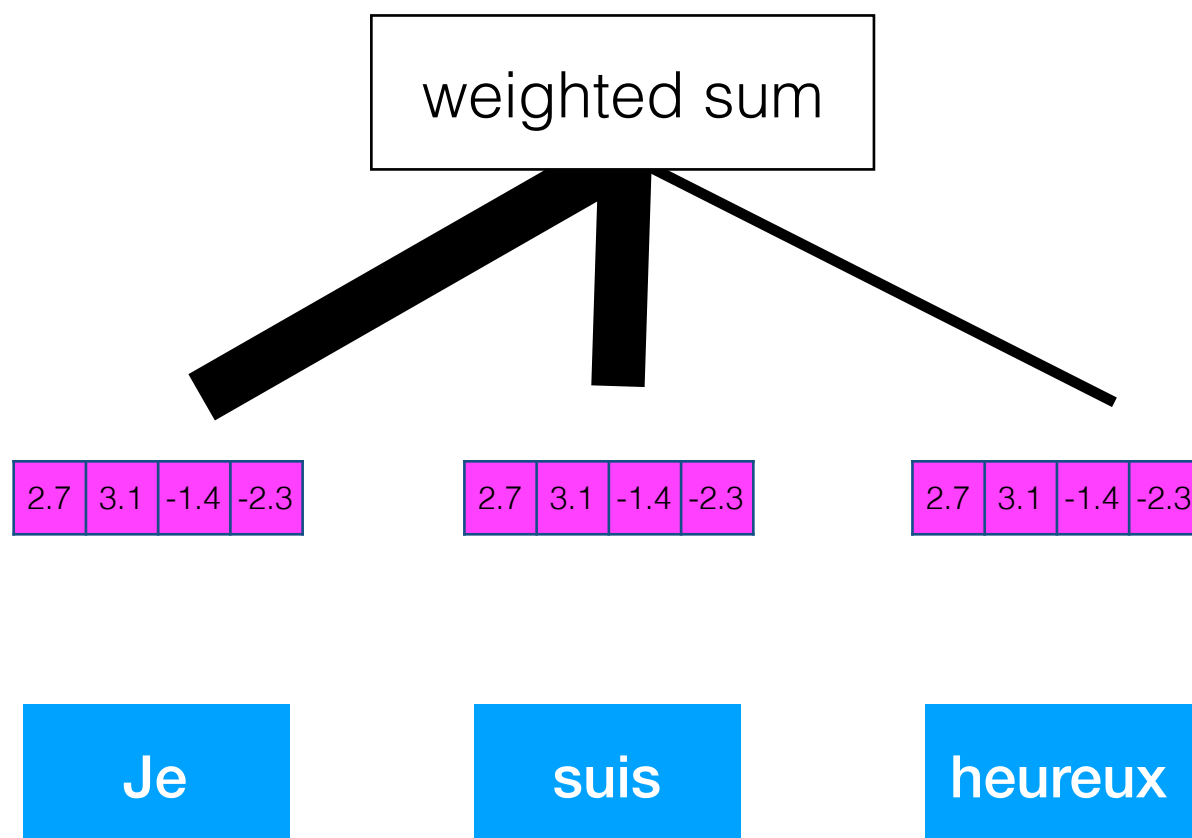
heureux

I

am

Encoder-decoder attention

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



2.7 3.1 -1.4 -2.3

2.7 3.1 -1.4 -2.3

2.7 3.1 -1.4 -2.3

2.7 3.1 -1.4 -2.3

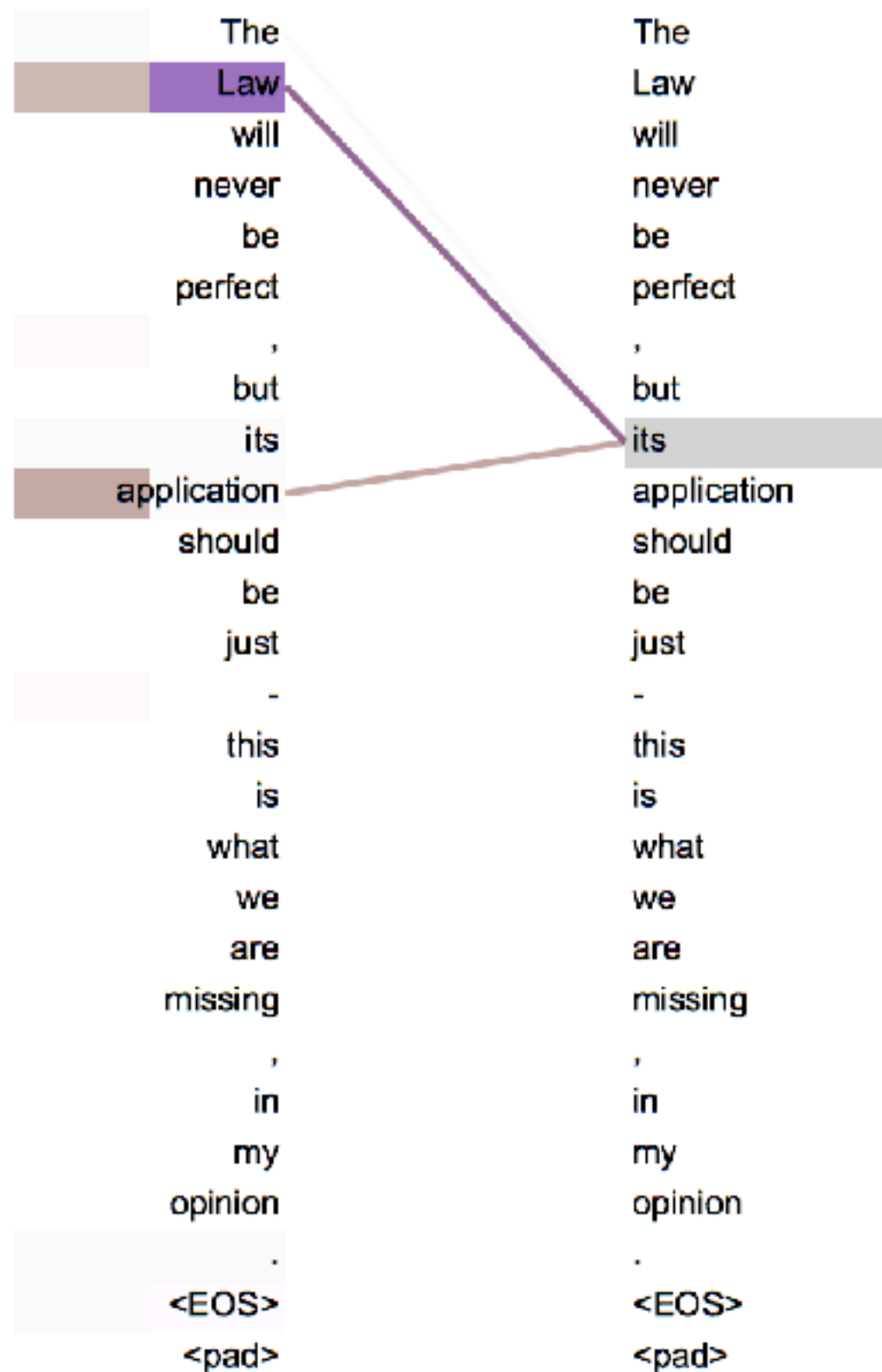
2.7 3.1 -1.4 -2.3

2.7 3.1 -1.4 -2.3

I

am

- Self-attention captures structure in the input (like coreference)



- Multiple heads capture different structure.



Thursday

- Wrap up and review!