Quiz instructions

Get a blue or black pen out now, before we hand out the quiz.

Do not talk or use electronic devices once we start handing out the quiz.

Leave the quiz facing down until we say it is time to start.

There are two versions of the quiz. The people either side of you must have a different coloured quiz to you.

When we say it is time to stop, hand your quiz to the end of your row.

Make sure you write your name and SID.

This is a closed-book, closed-note quiz. No electronic devices may be used in any way.

Note your responses by completely filling in the relevant circle(s) an square(s): ●

If you make a mistake, put an X over the filled in circle / square: ⊗

COMP 4446 / 5046 Lecture 6: Models – Encoder-Decoder

Jonathan K. Kummerfeld

Semester 1, 2025

menti.com 4843 3031

Simple

[Actually, I think if all higher math professors had to write for the Simple English Wikipedia for a year, we'd be in much better shape academically.]

Source: https://xkcd.com/547/

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Contextual Representations

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

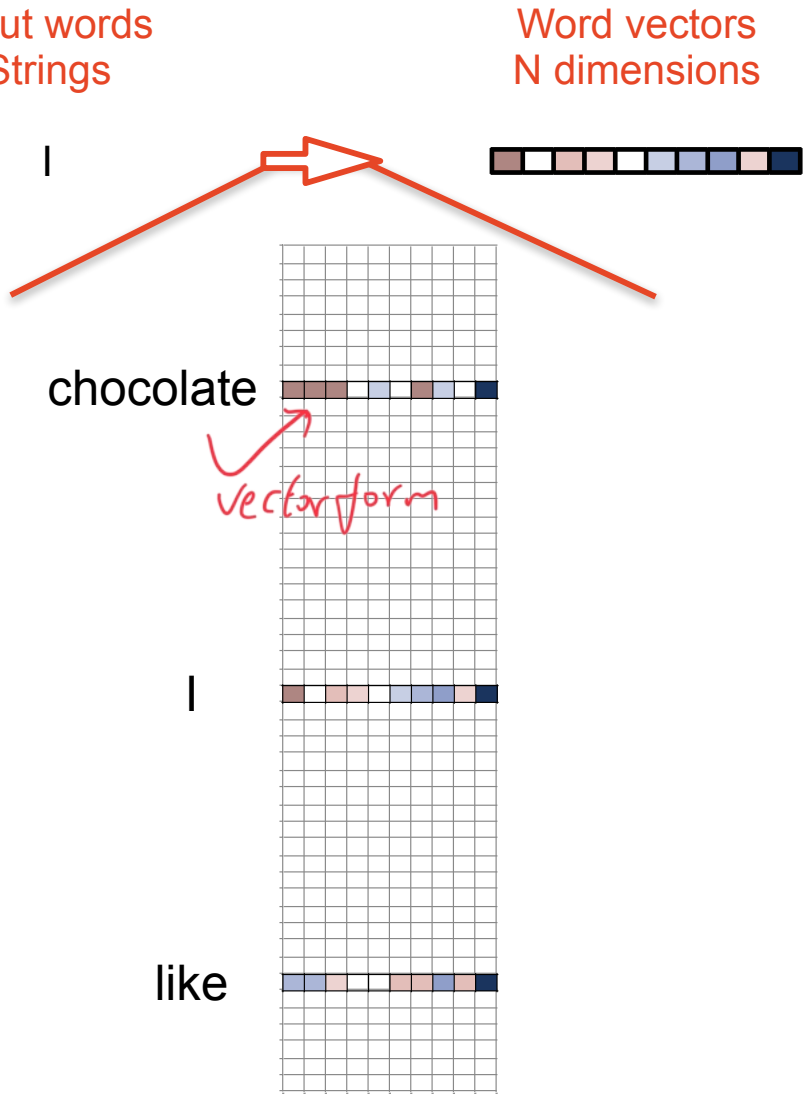
Attention

Workshop Preview

menti.com 4843 3031

Review

So far, our word embeddings have been looked up in a table



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

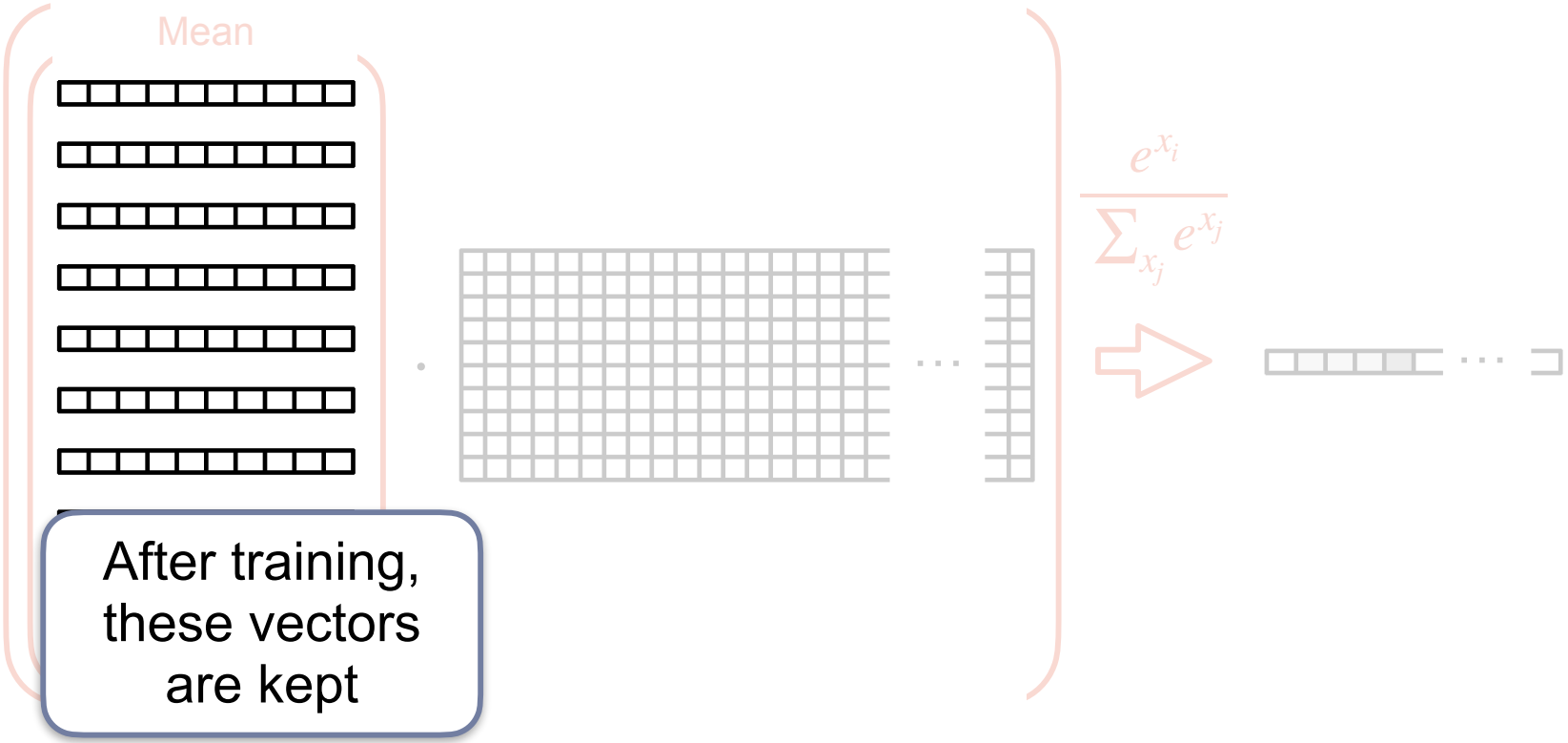
Where does the table come from?

Review

word2vec - Continuous Bag of Words ✓

Input: Context words

Output: One word



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Where does the table come from?

word2vec - SkipGram

Input: One word

Output: Set of context words

After training, these vectors are kept

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Where does the table come from? look for every word

GloVe

- Calculate co-occurrence statistics
- Form random vectors for each word
- Update vectors so the dot product of two vectors is approximately the co-occurence value

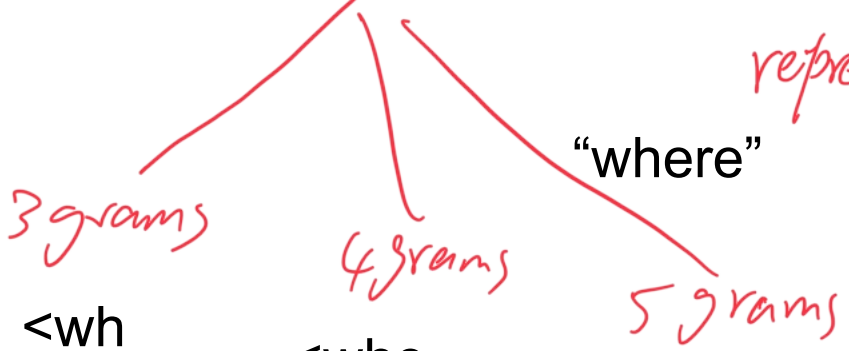
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Where do these come from?

FastText
- Represent words as a set of character ngrams
- Learn vectors for ngrams
- Words are the sum of vectors for their ngrams
- Use word2vec skipgram learning



representation of word is sum of those ngrams

Ex. maybe never see a word before, but see a gram before.

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

What if my data is not the same as the data used for training?

Standard sources of data:
- Websites
- Books
- News
- Research literature

Not:
- Medical records
- Internal company documents
- Email
- Instant messaging
- Text messages

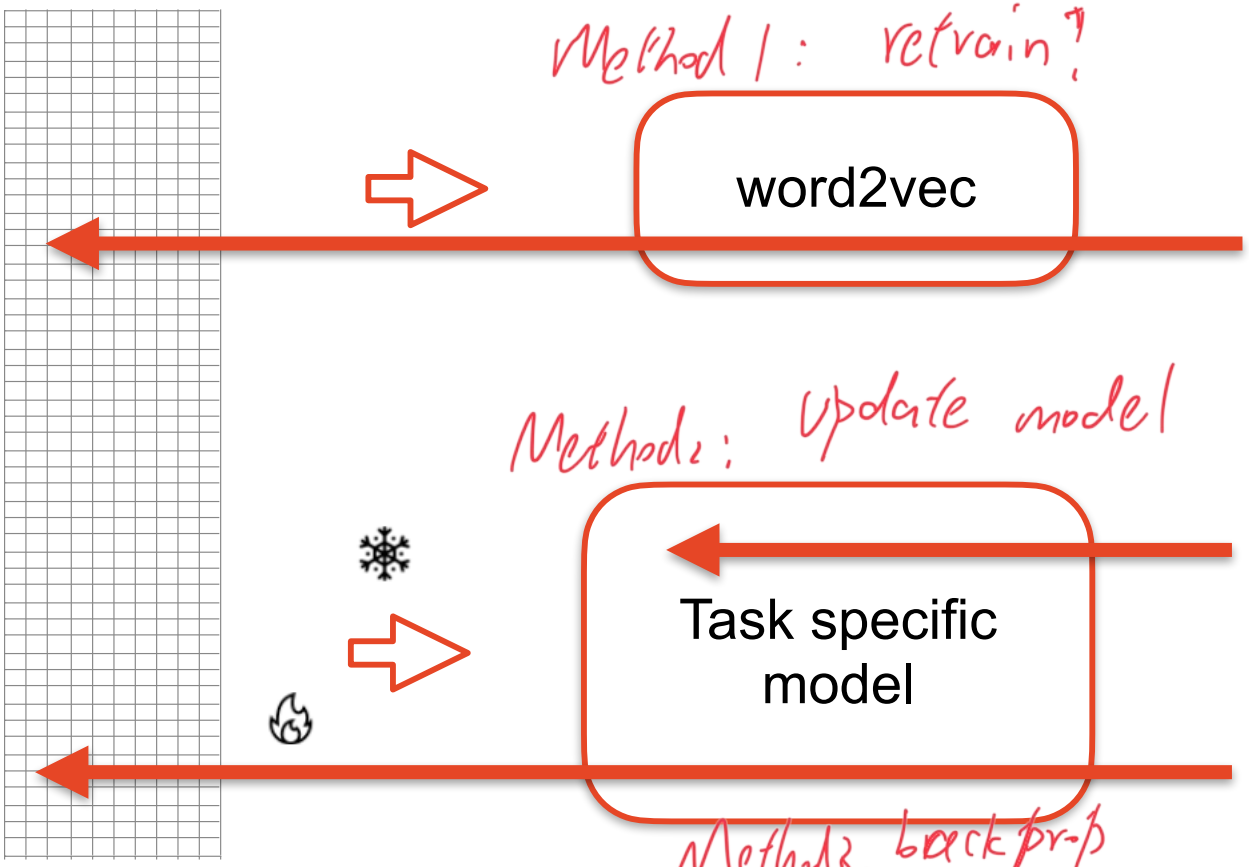
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

What if my data is not the same as the data used for training?

Fine-tuning - Update the embeddings for your task



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

What about word senses?


NN	NN	VB	DT	NN
Time	flies	like	an	arrow

NN	VBZ	IN	DT	NN
Time	flies	like	an	arrow

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

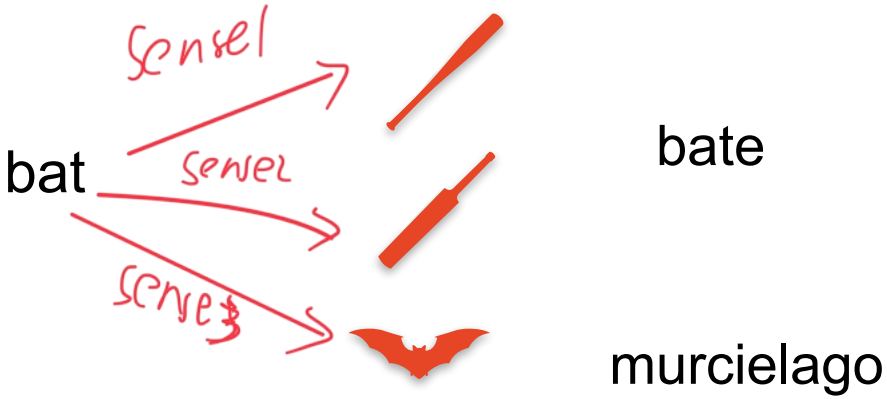
Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

What about word senses?



Which of those flights serve breakfast?

Does Air France serve Philadelphia?

?Does Air France serve breakfast and Philadelphia?

Jurafsky and Martin, Appendix G

COMP 4446 / 5046
Lecture 6, 2025

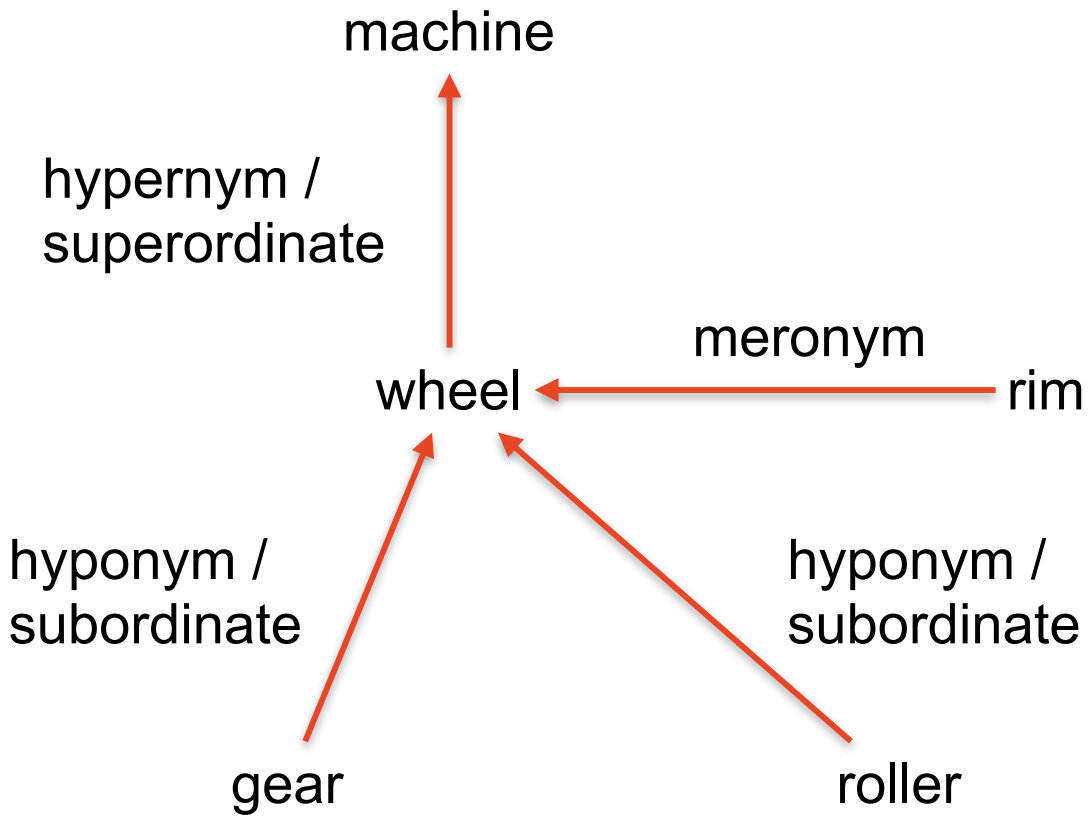
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

What about word senses?

WordNet

a simple machine consisting of a circular frame …



http://wordnetweb.princeton.edu/perl/webwn

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Train a model with multiple word vectors, one per sense?

Challenge: data

SemCor - 226,036 words
Others in the 1,000 - 10,000 range

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

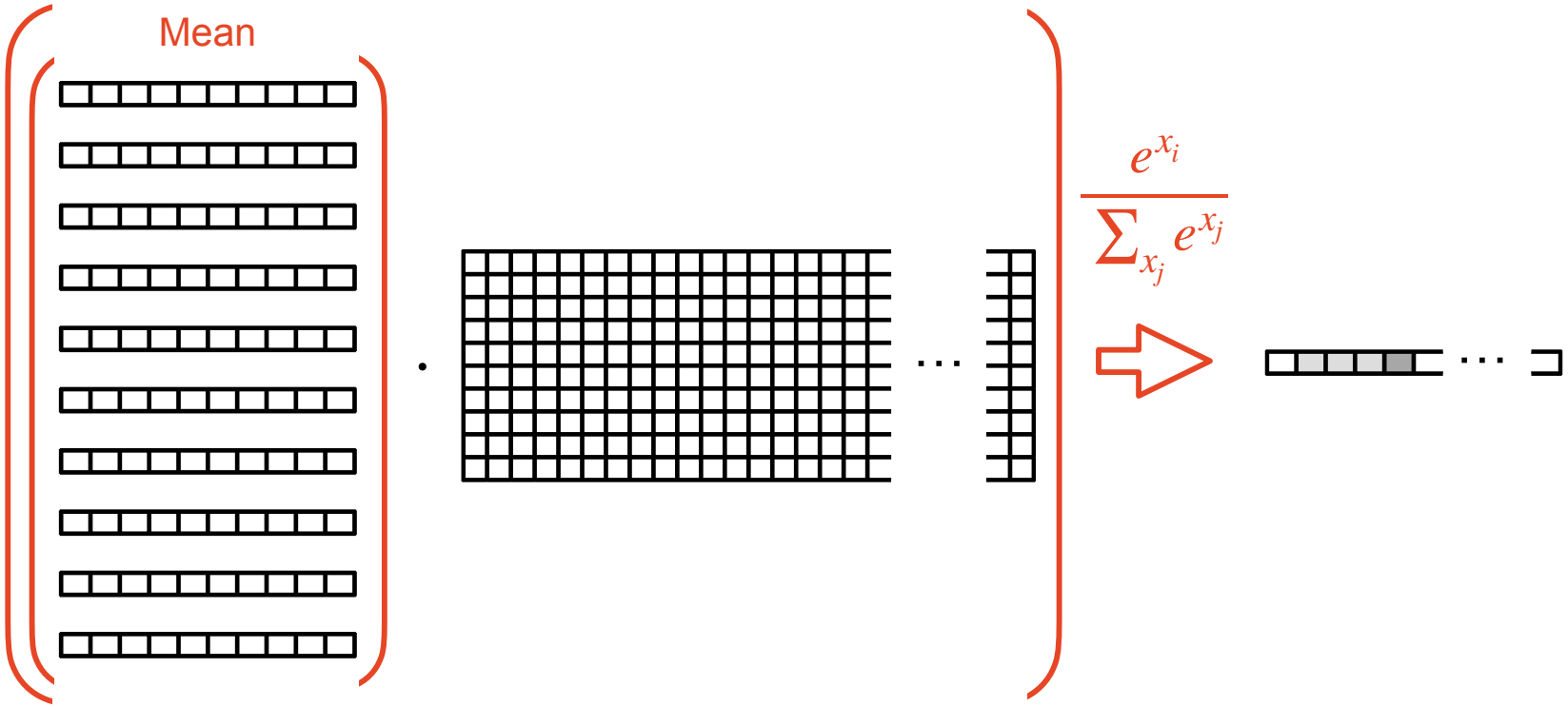
Attention

Workshop Preview

menti.com 4843 3031

Train a model with **contextual** representations

word2vec - Continuous Bag of Words
Input: Context words
Output: One word



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

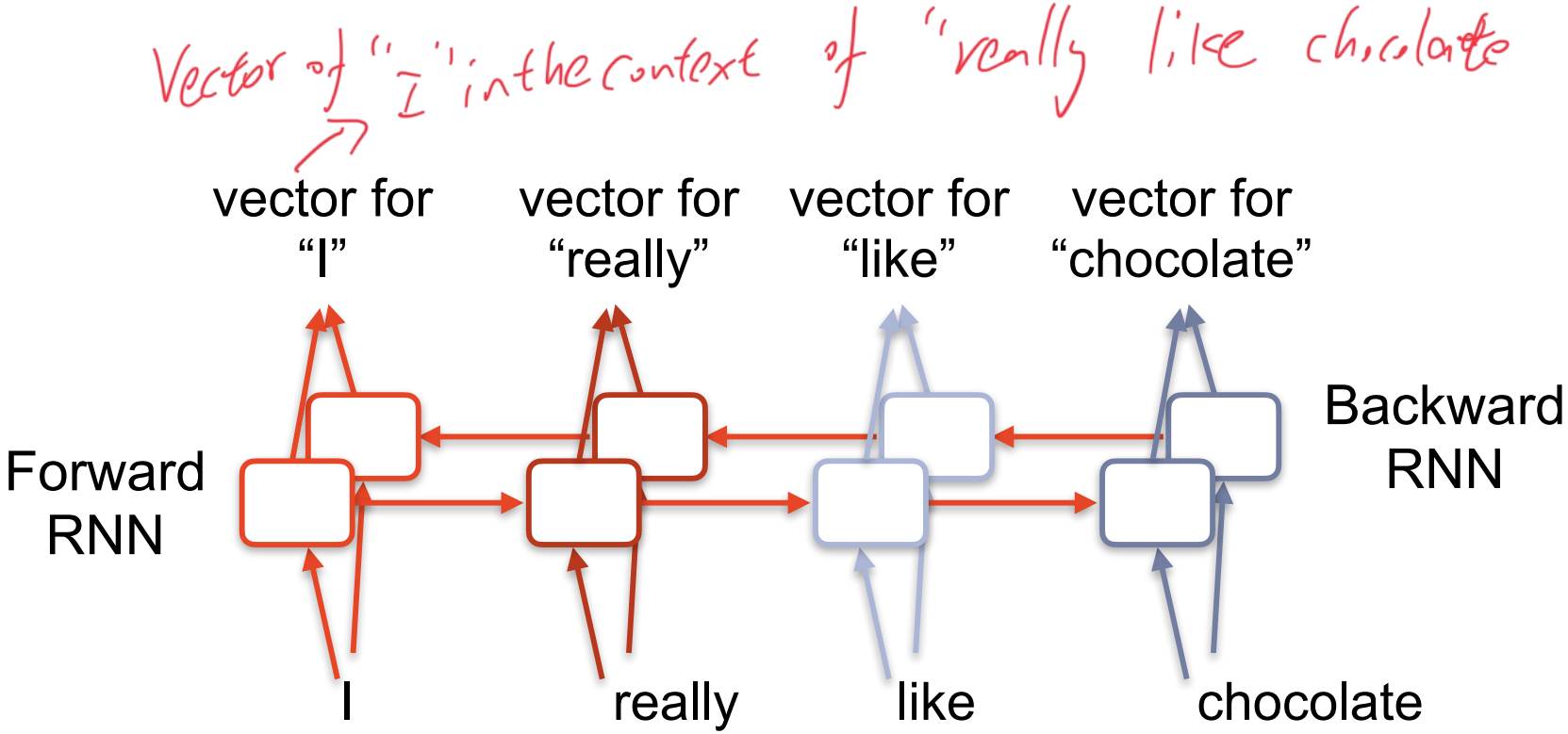
Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Form **contextual** representations using an RNN



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

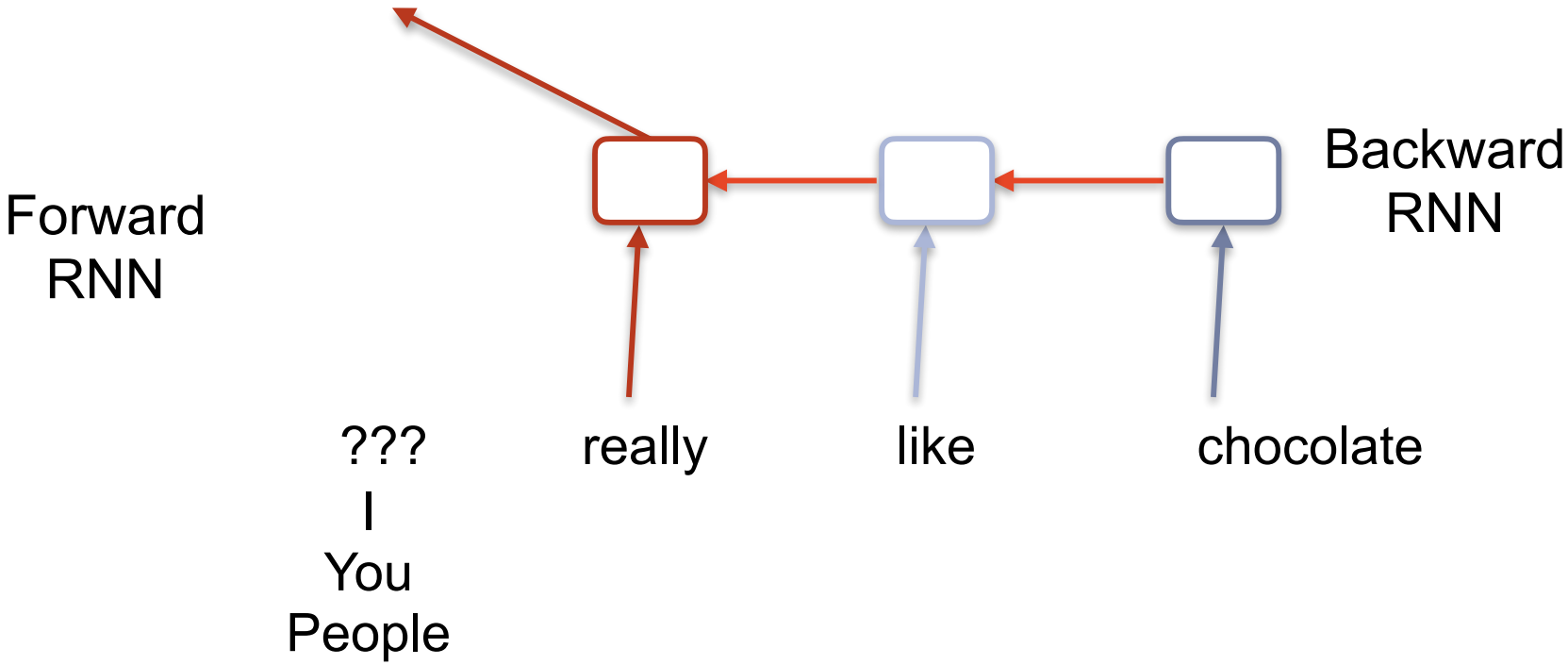
Attention

Workshop Preview

menti.com 4843 3031

How do we train the model?

Input: Context words
Output: One word



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

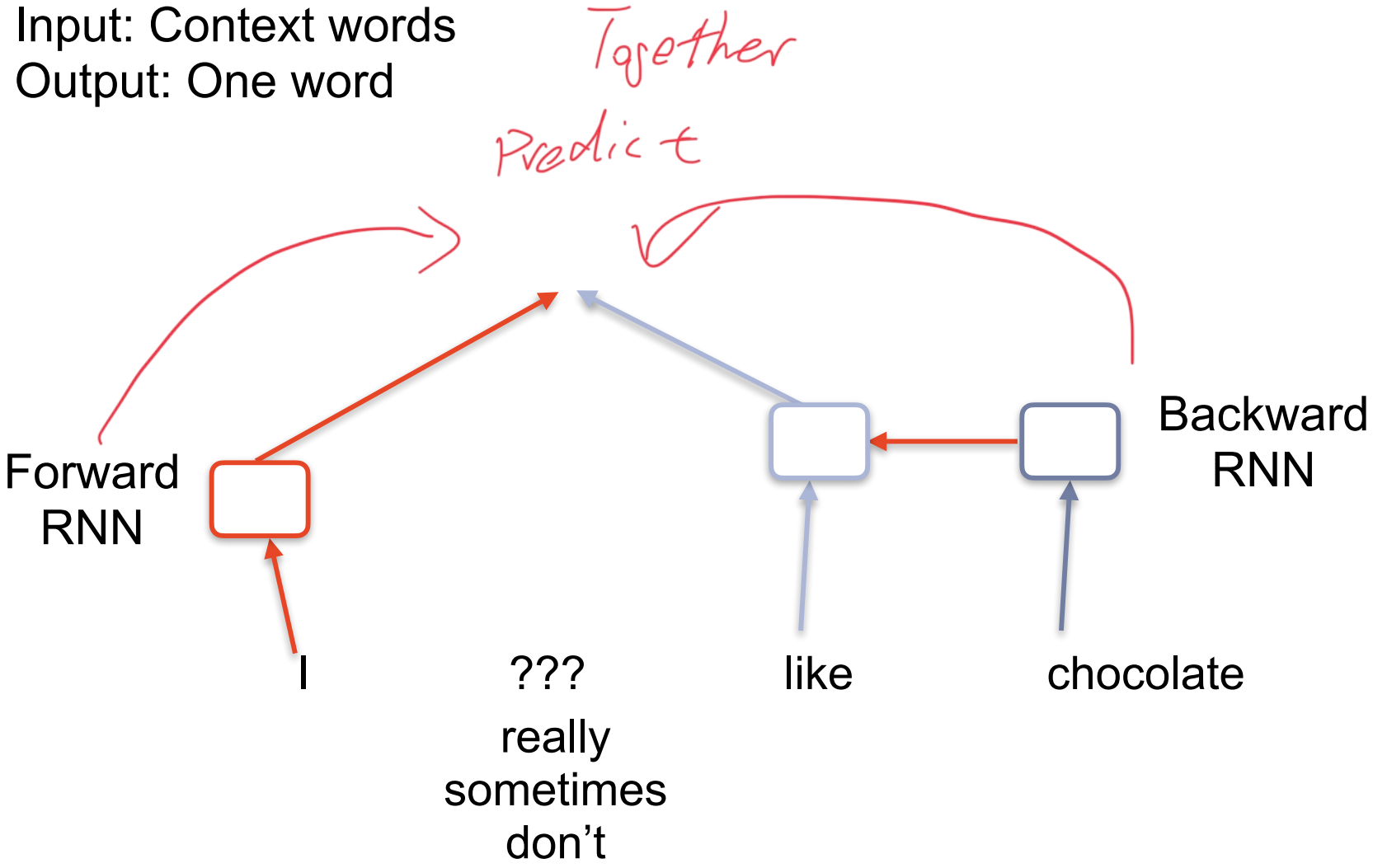
Attention

Workshop Preview

menti.com 4843 3031

How do we train the model?

Input: Context words
Output: One word



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

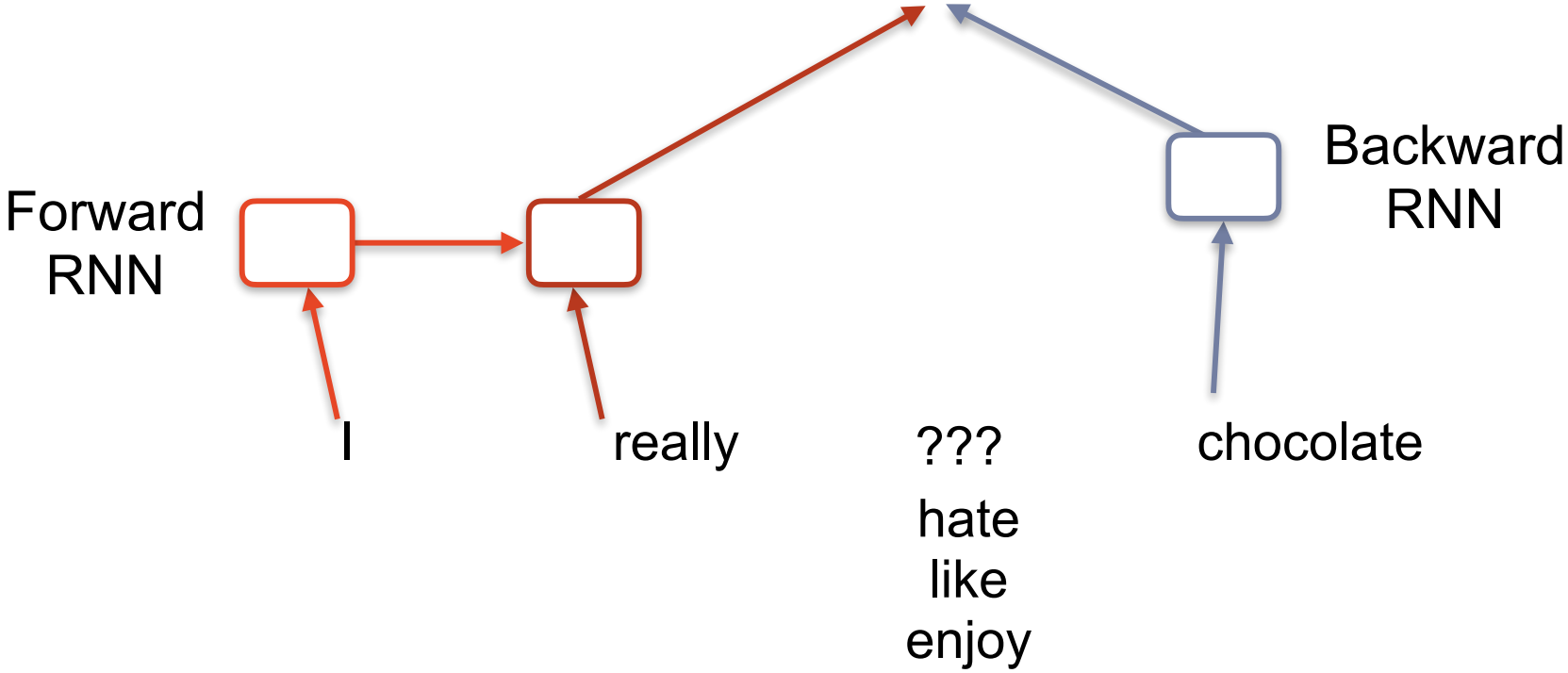
Attention

Workshop Preview

menti.com 4843 3031

How do we train the model?

Input: Context words
Output: One word



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

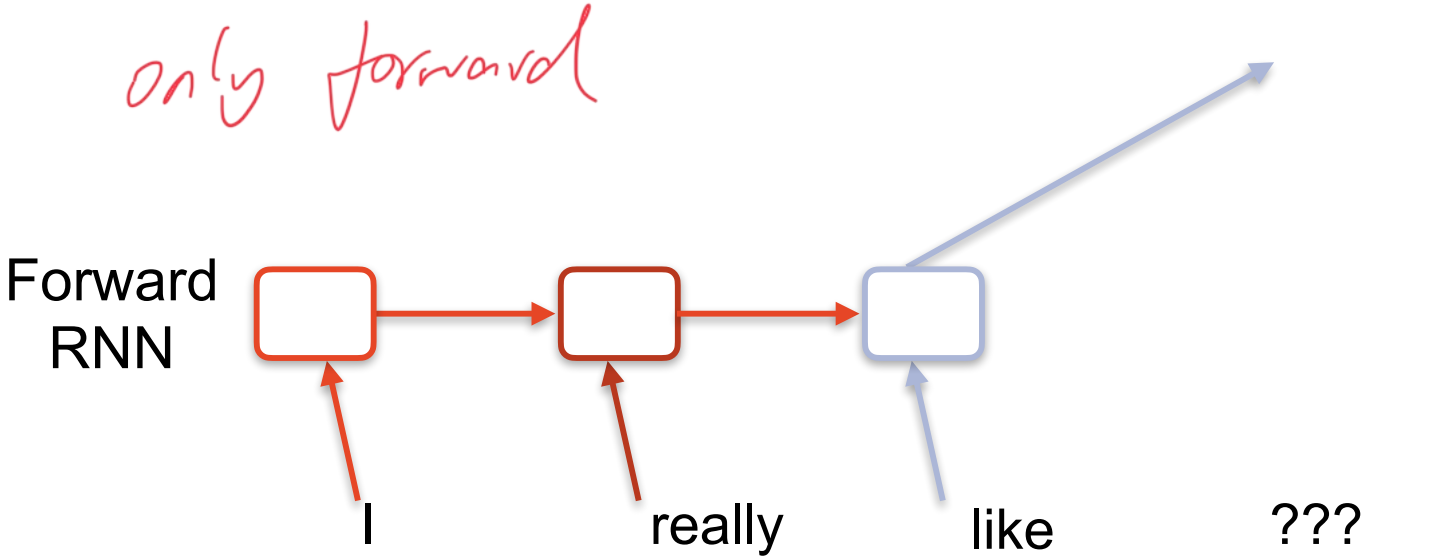
Workshop Preview

menti.com 4843 3031

How do we train the model?

Input: Context words
Output: One word

only forward



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

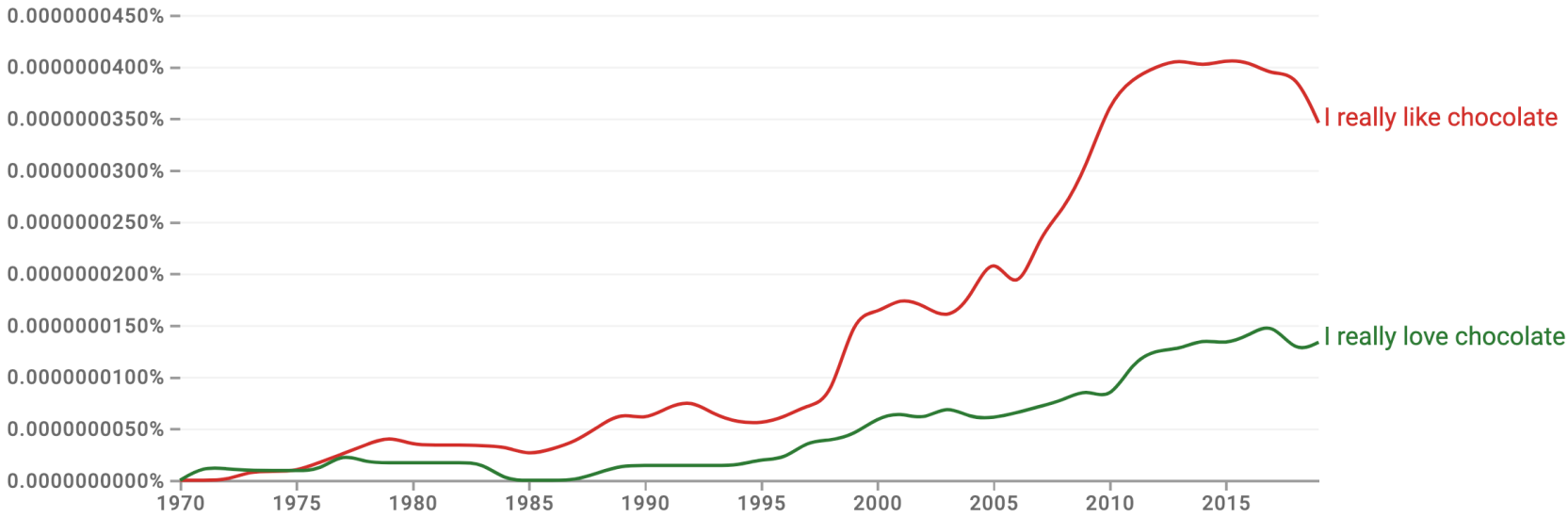
Workshop Preview

menti.com 4843 3031

[aside - Google shows that books agree!]

Counts in books since 1970 of:

"I really * chocolate"



https://books.google.com/ngrams

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

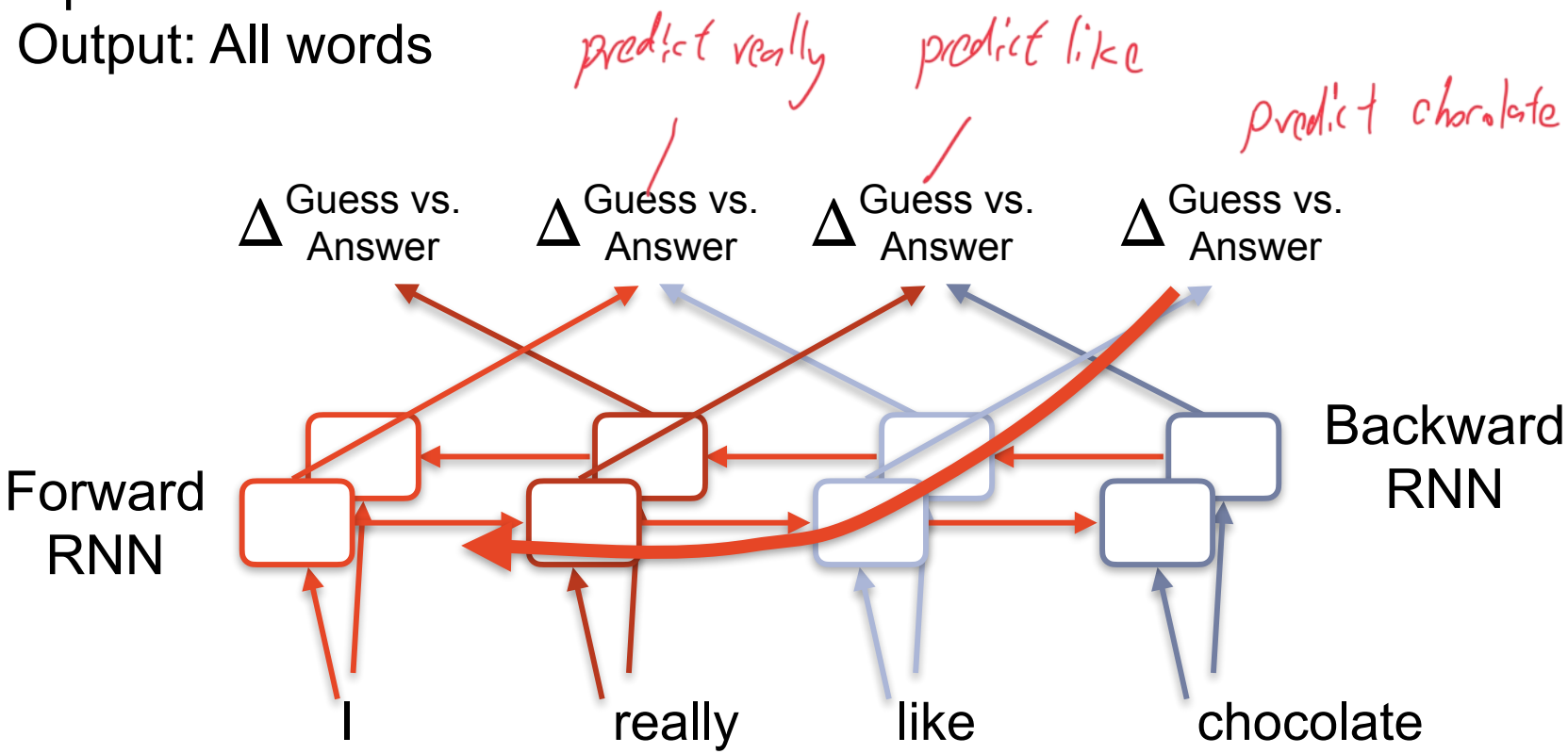
Attention

Workshop Preview

menti.com 4843 3031

We can train on multiple words at once

Input: All words
Output: All words



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Major turning point in NLP

Earliest use I know of - 2015

No major awareness at first (only slight improvements)

2018 - ELMo
"Deep contextualized word representations"

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Li		81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.0	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	91.93 ± 0.19	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Question Answering

Named Entity Recognition

Sentiment

https://sesameworkshop.org/our-work/shows/sesame-street/sesame-street-characters/

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

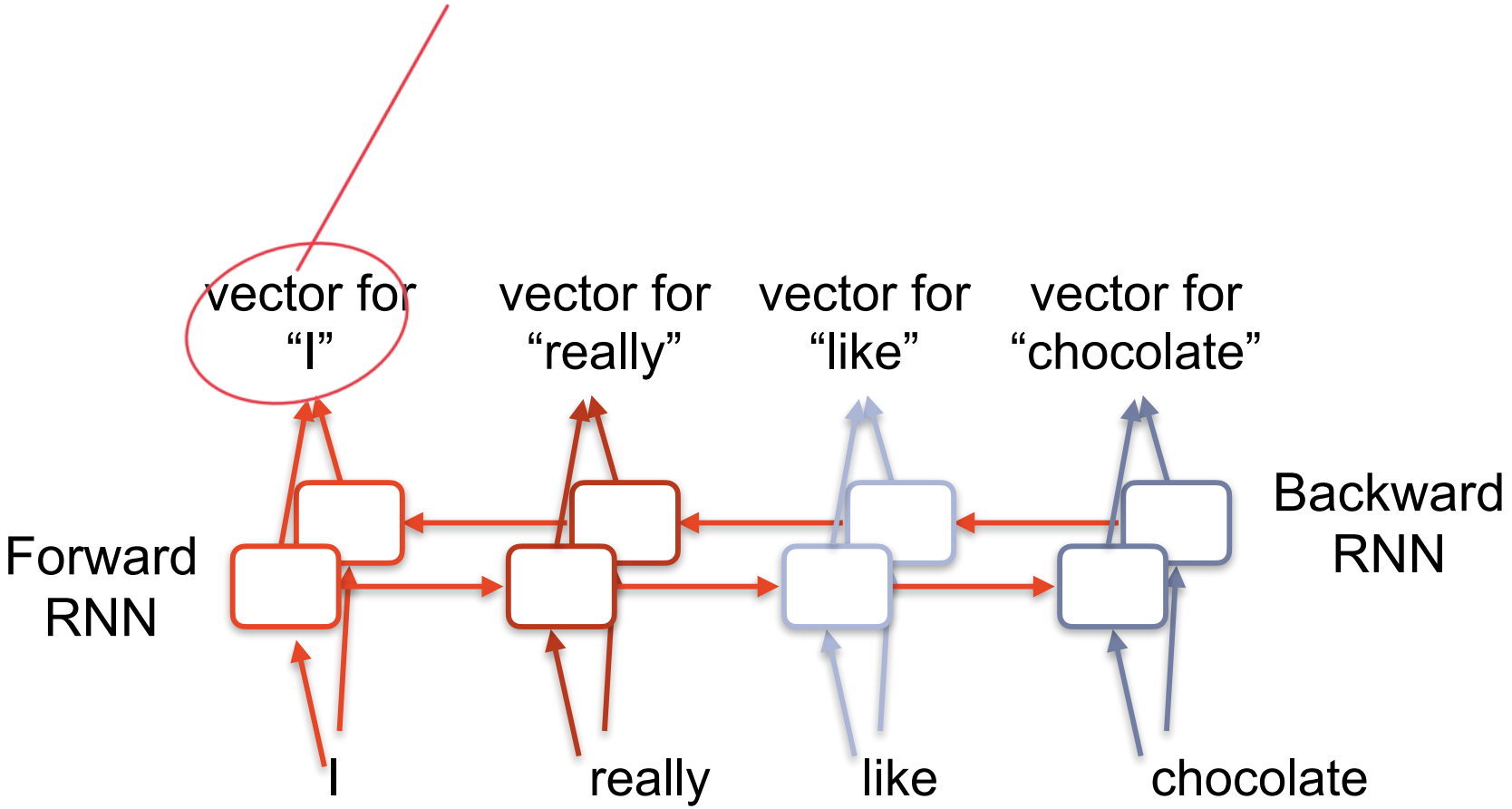
Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

How are these used?



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

What are they learning?

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

Tokenisation

Attention

Workshop Preview

menti.com 6165 8383

The LSTM is just one possible model…

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

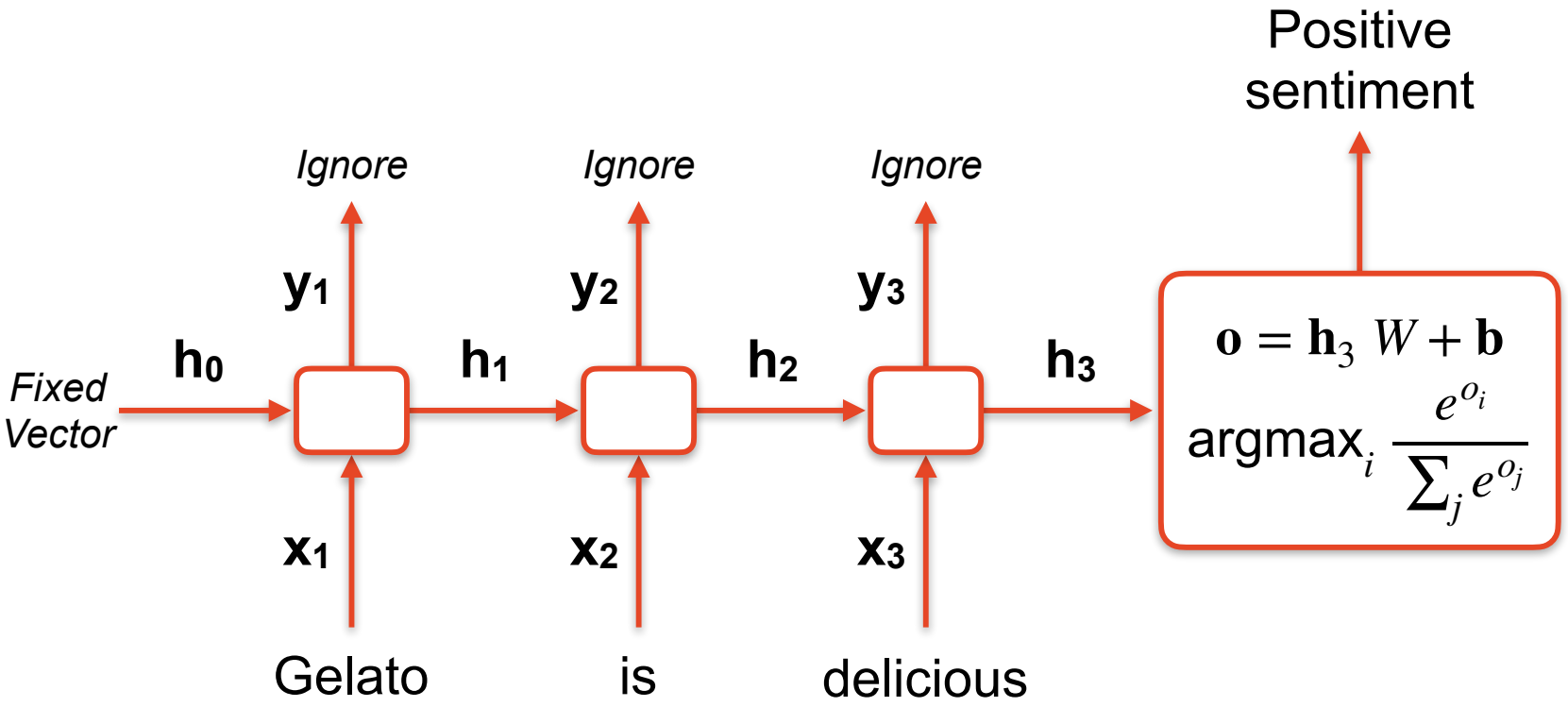
Encoder-Decoder

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

RNNs can be used as an accepter or encoder

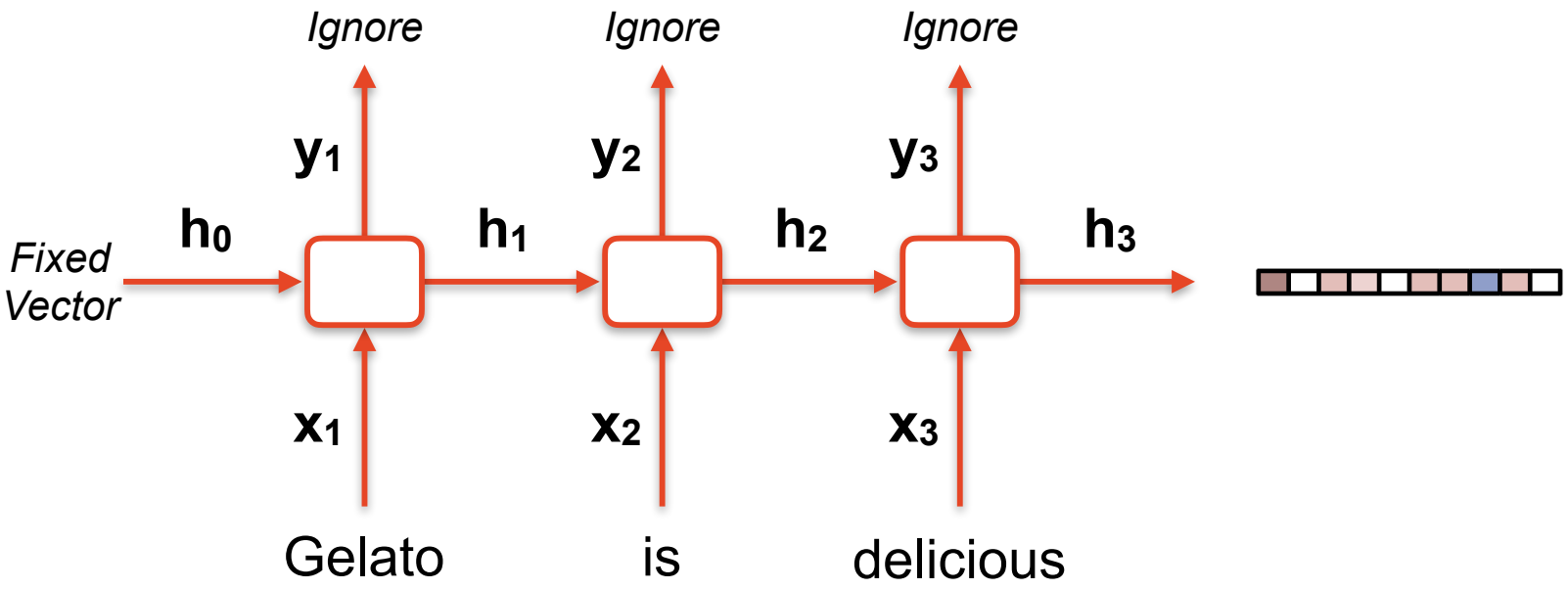


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

RNNs can be used as an accepter or encoder

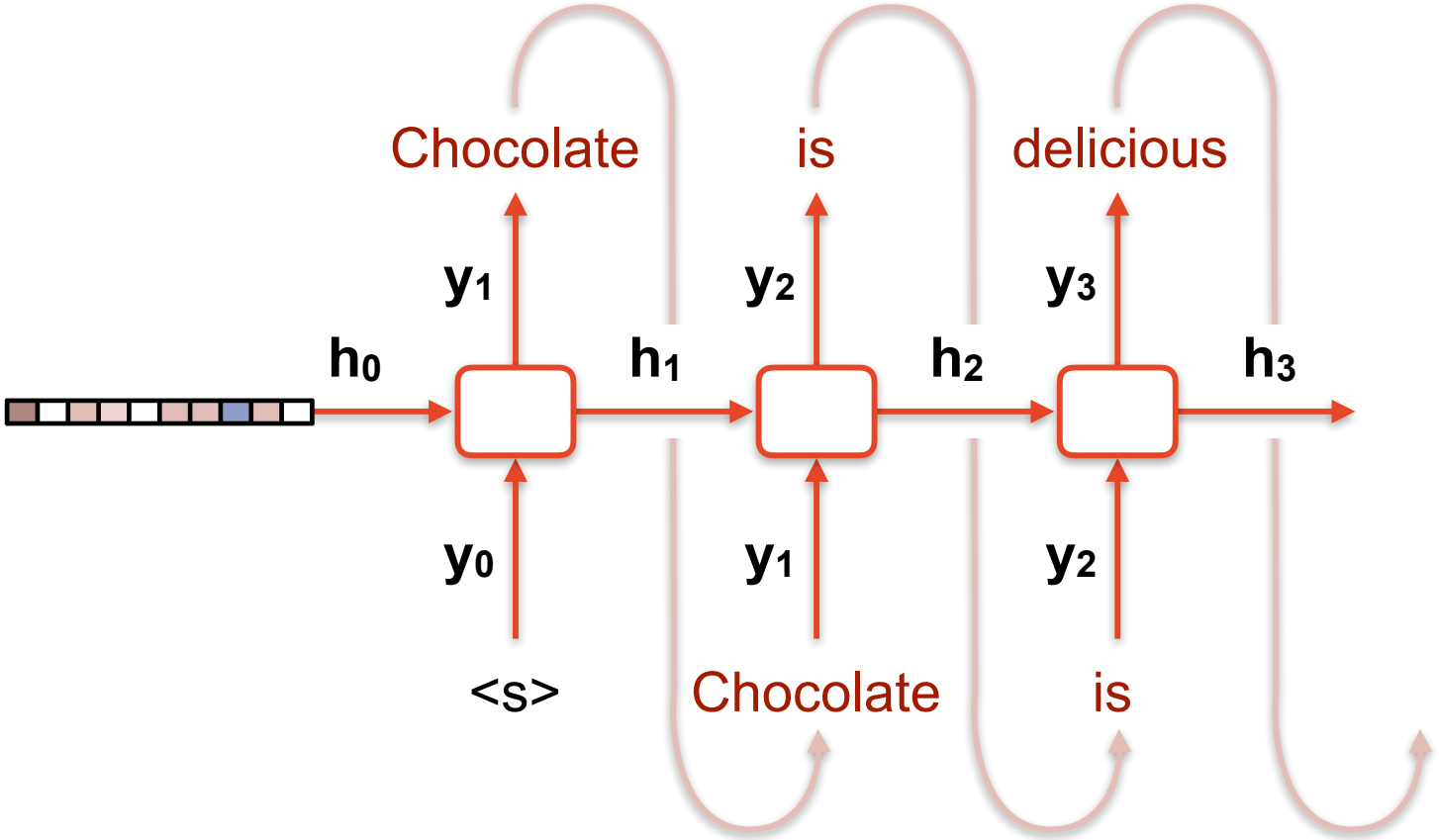


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

What is a decoder?

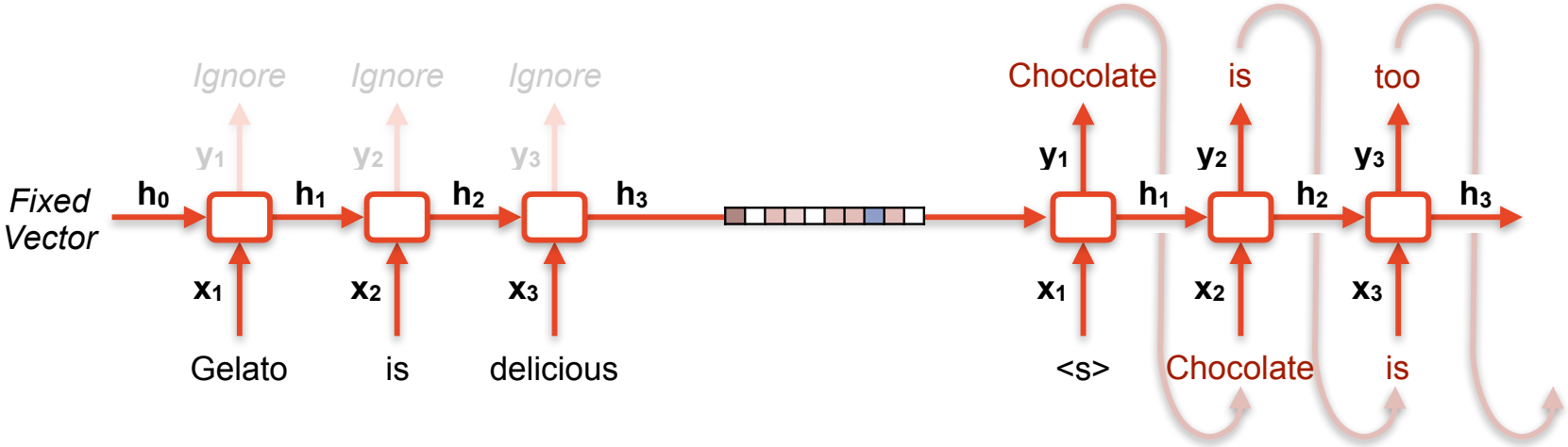


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We can put these two pieces together



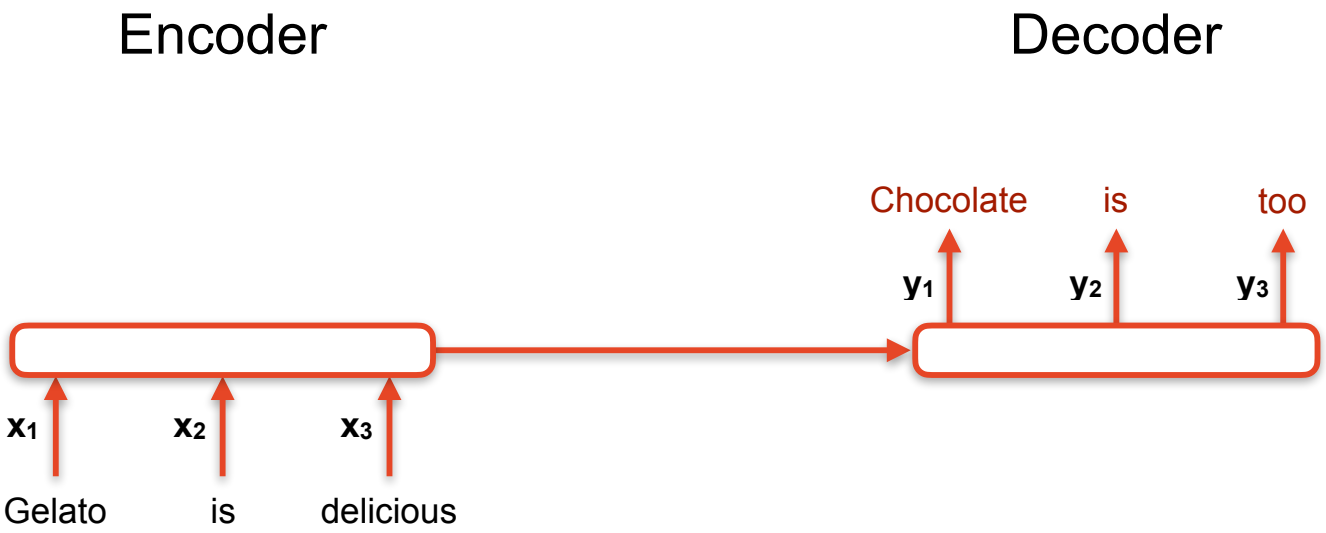
This is also called a 'sequence to sequence' model

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We can put these two pieces together

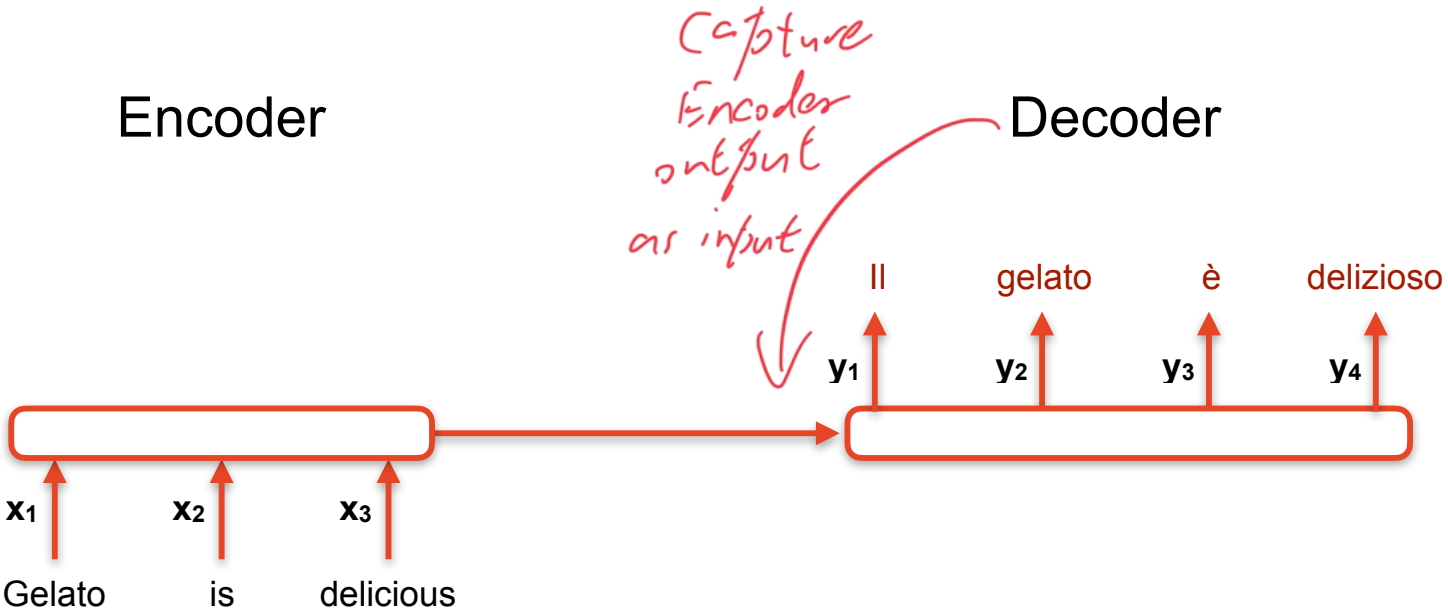


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

The first successful application was Machine Translation



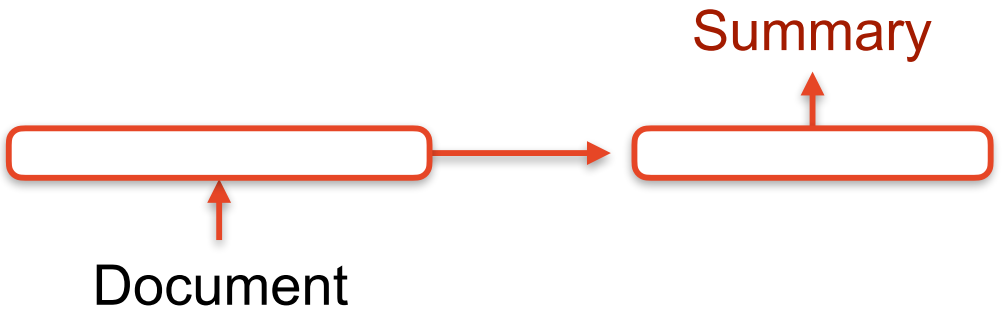
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

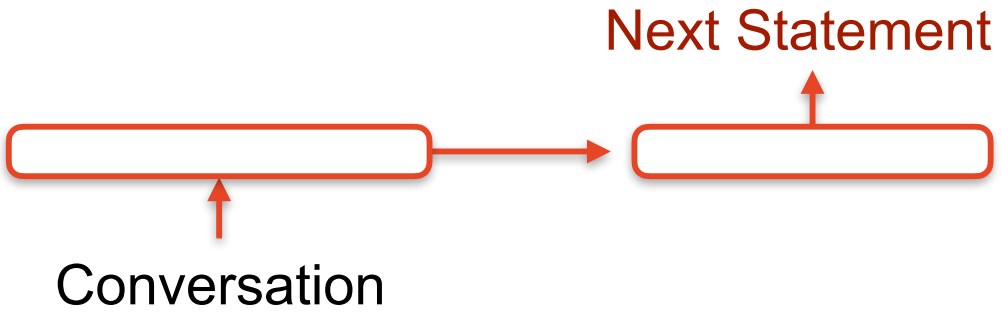
menti.com 4843 3031

But it can be used for a wide range of applications

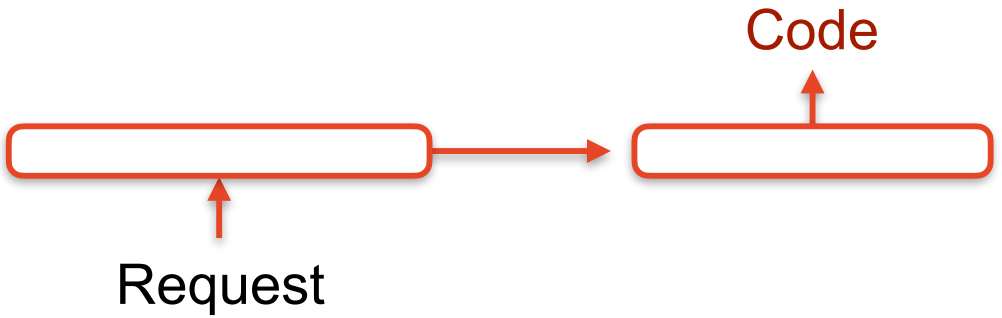
Summarisation



Dialogue



Code Generation



…

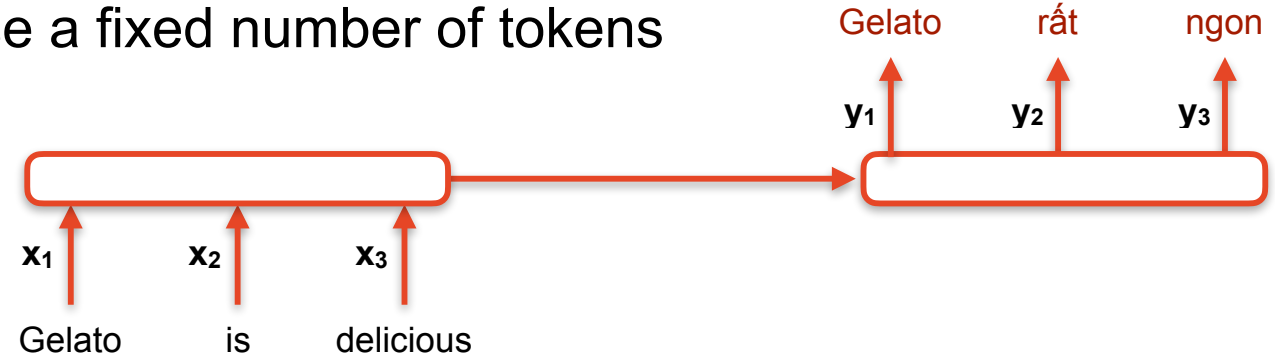
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

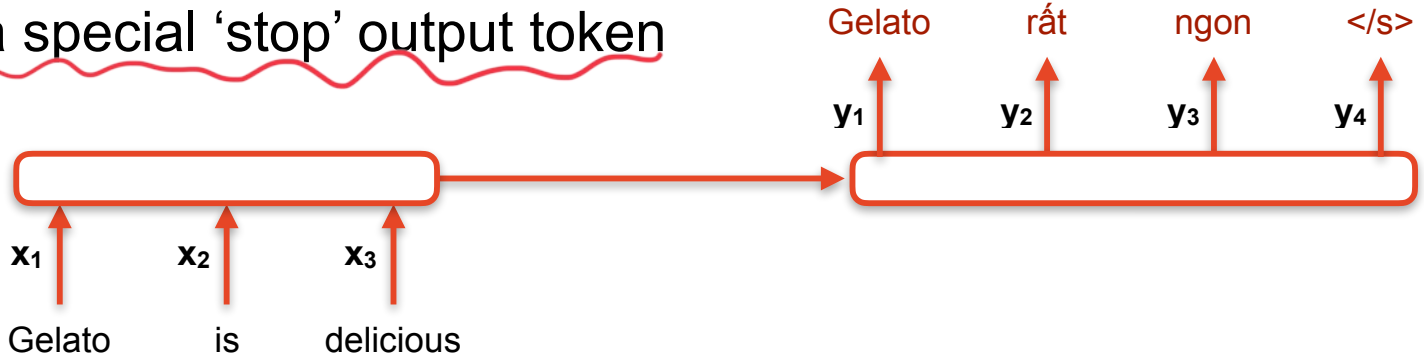
menti.com 4843 3031

When do you stop decoding?

Option 1:
Choose a fixed number of tokens



Option 2:
Have a special ‘stop’ output token



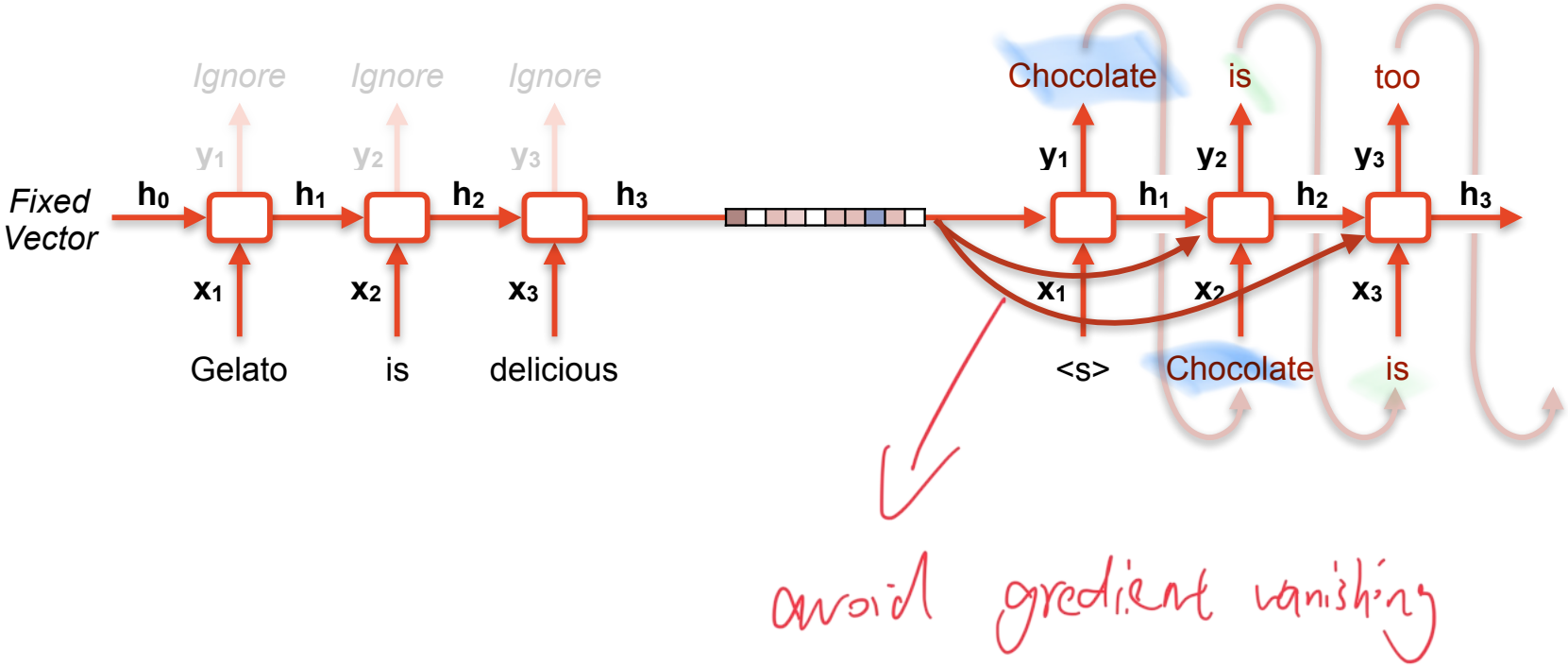
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

In practise, the structure is slightly more complex

Encoding / Context is passed in at each step



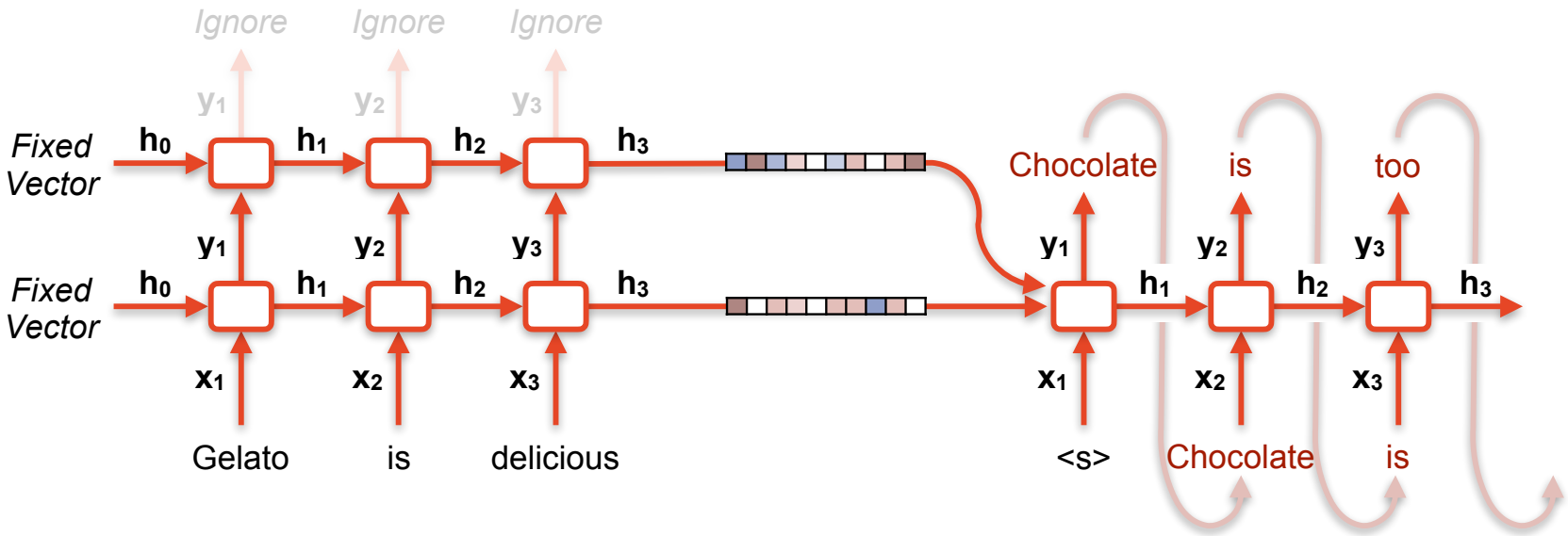
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

In practise, the structure is slightly more complex

Multi-layer models



Can also use bi-directional models
- Encoder, easy to set up
- Decoder, trickier, but also possible at the cost of computation

COMP 4446 / 5046
Lecture 6, 2025

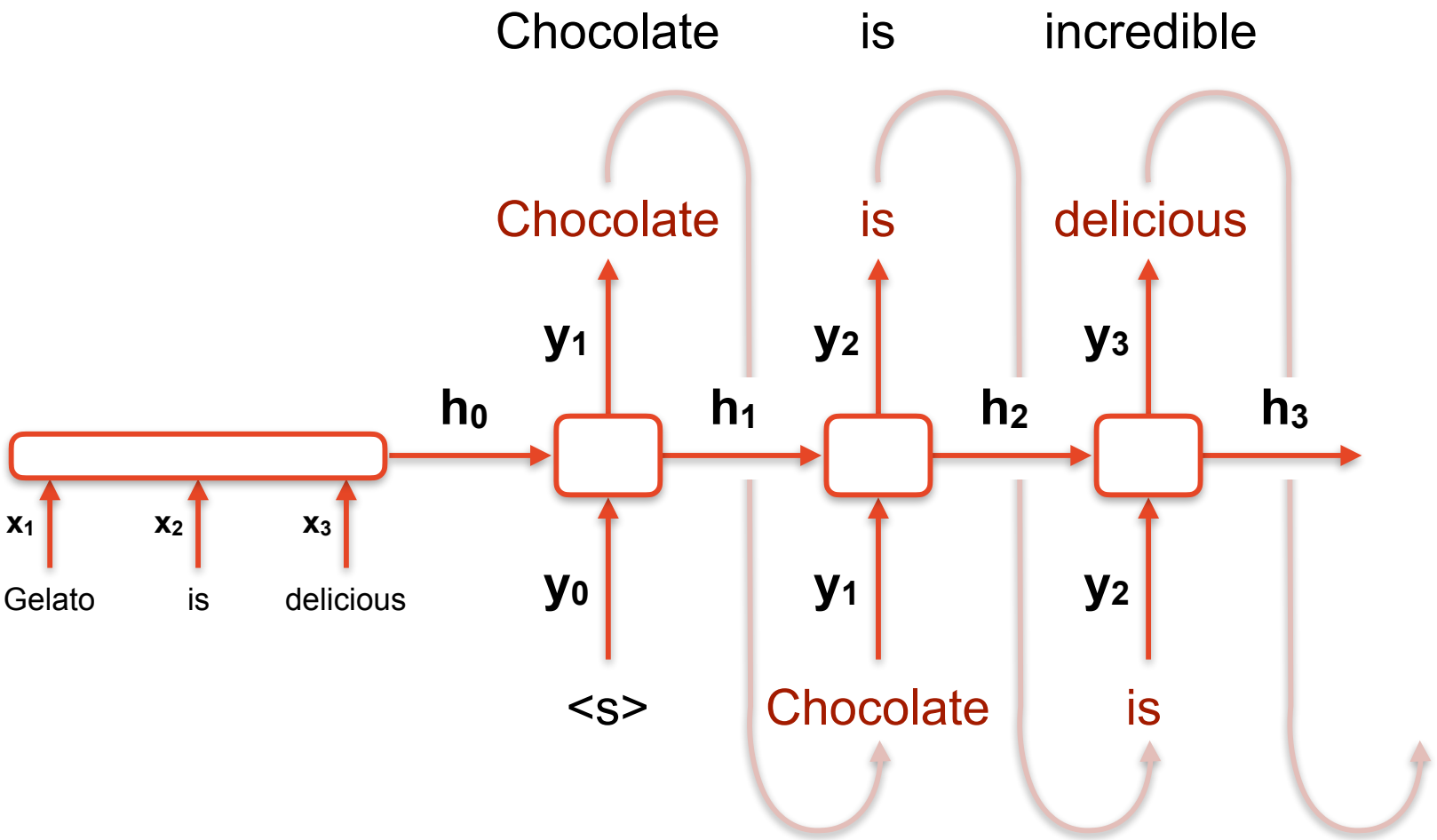
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do you train the decoder?

back prop

Δ Guess vs. Answer

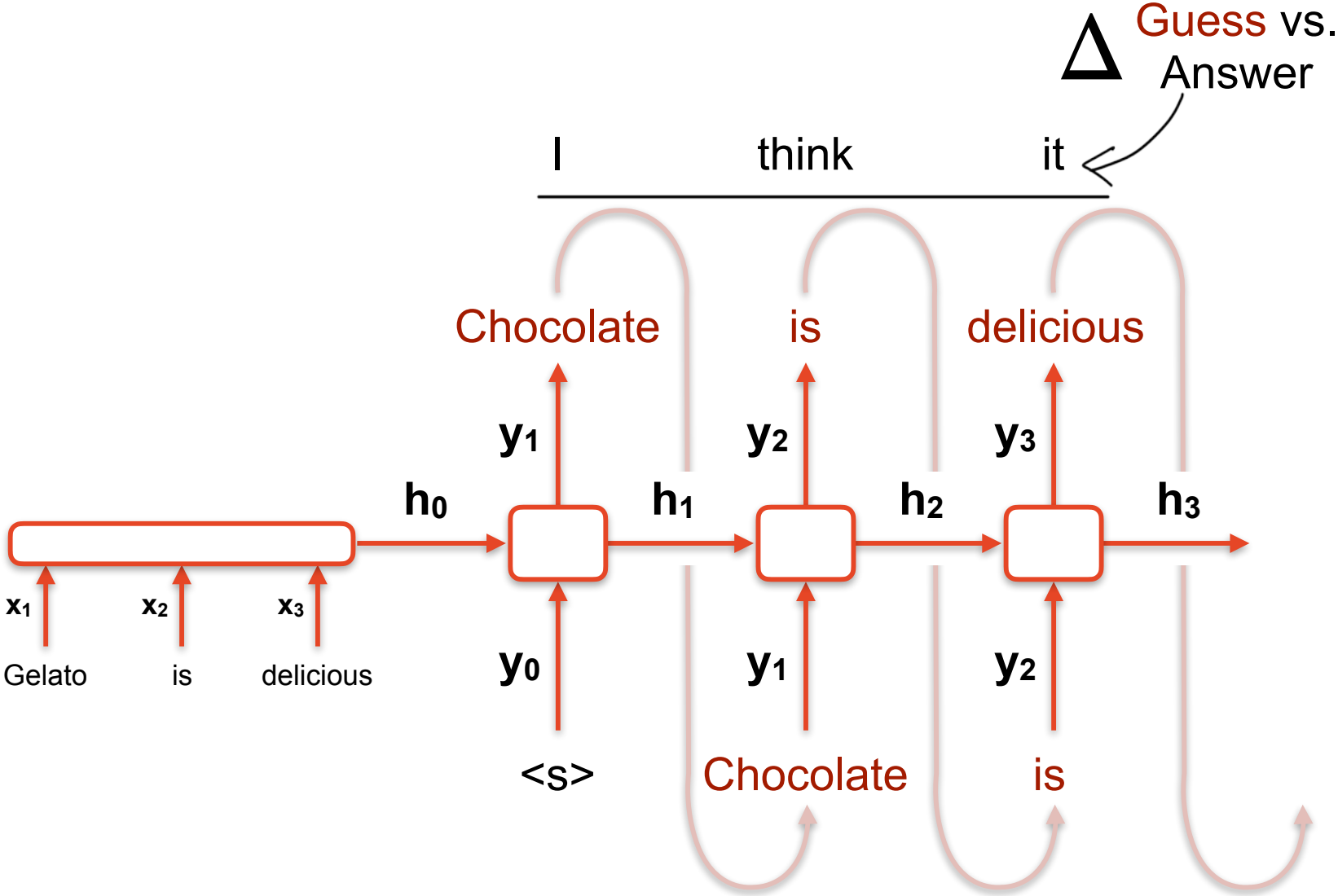


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do you train the decoder?



COMP 4446 / 5046
Lecture 6, 2025

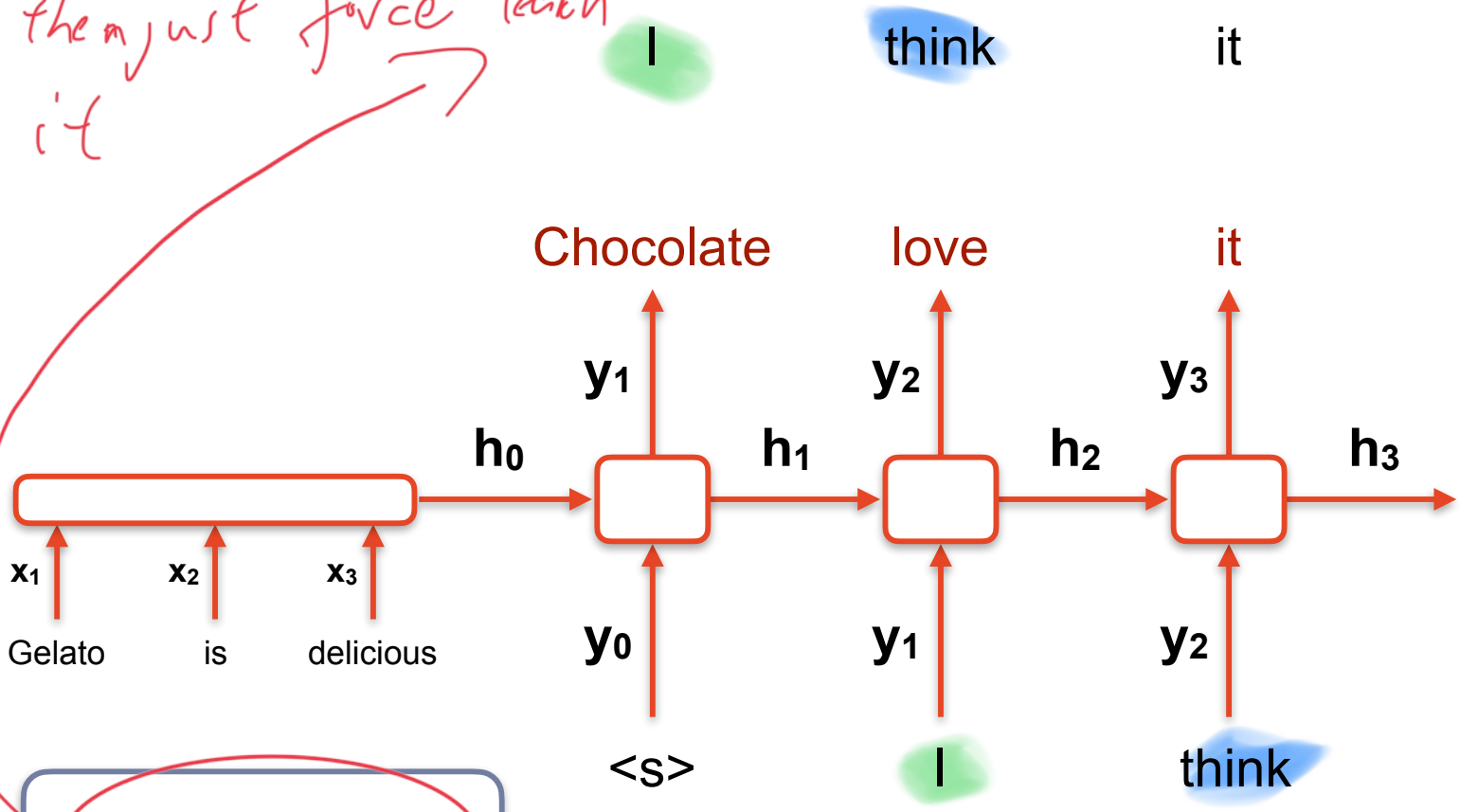
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do you train the decoder?

If totally wrong then just force teach it

Δ Guess vs. Answer



"Teacher forcing"

model really good then does not matter

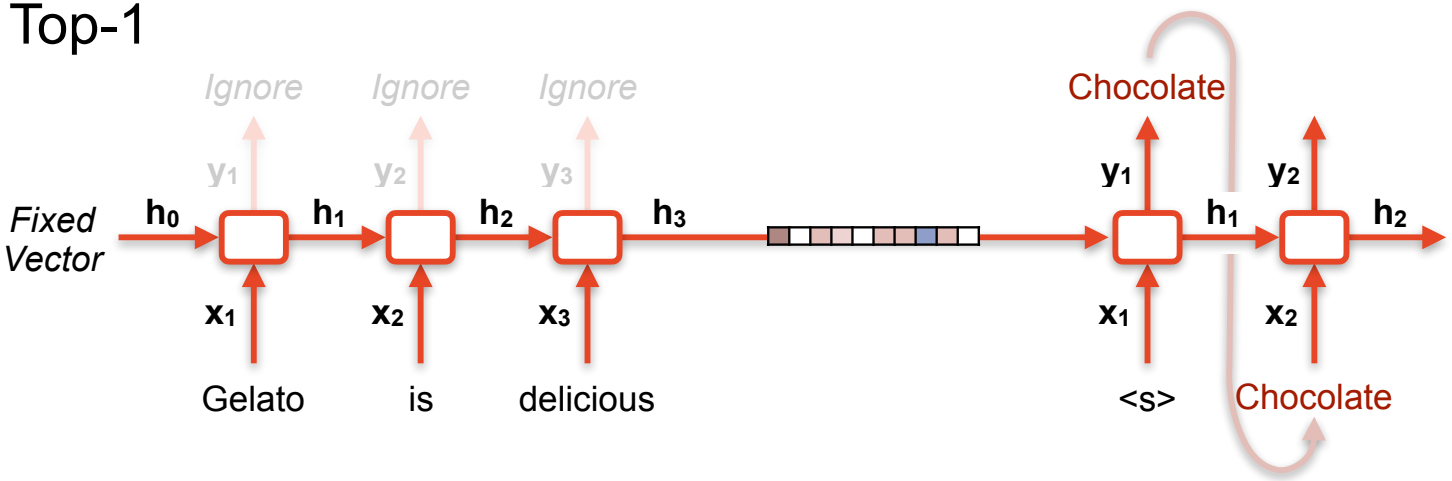
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

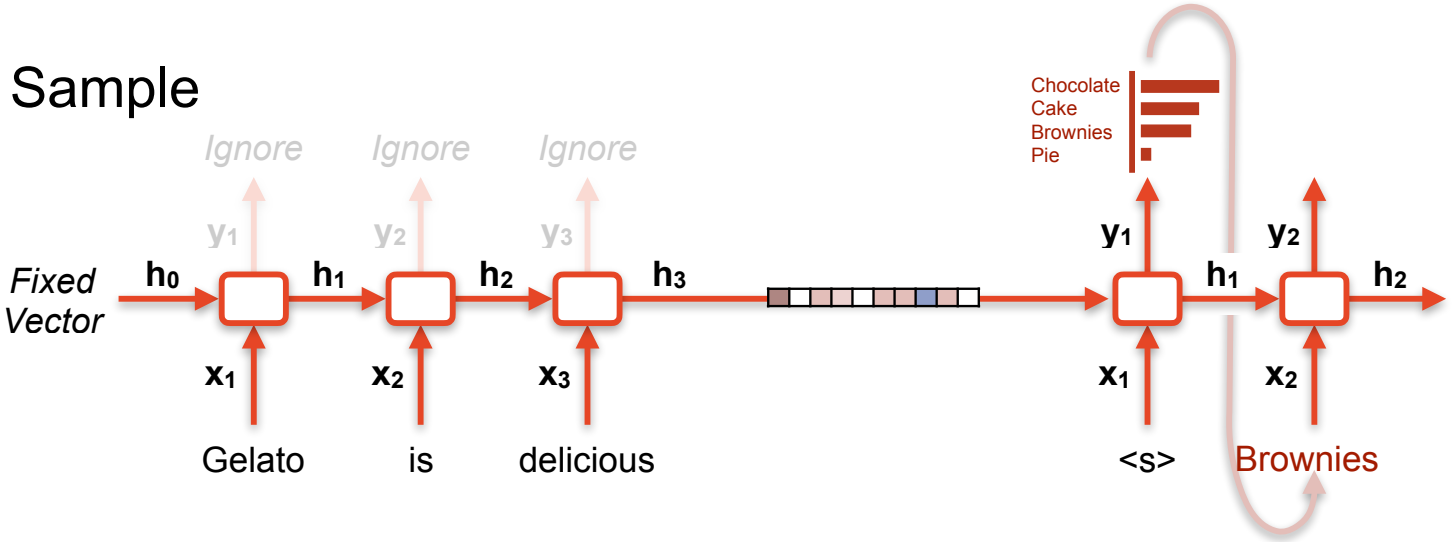
menti.com 4843 3031

Greedy inference with an encoder-decoder

Top-1



Sample



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

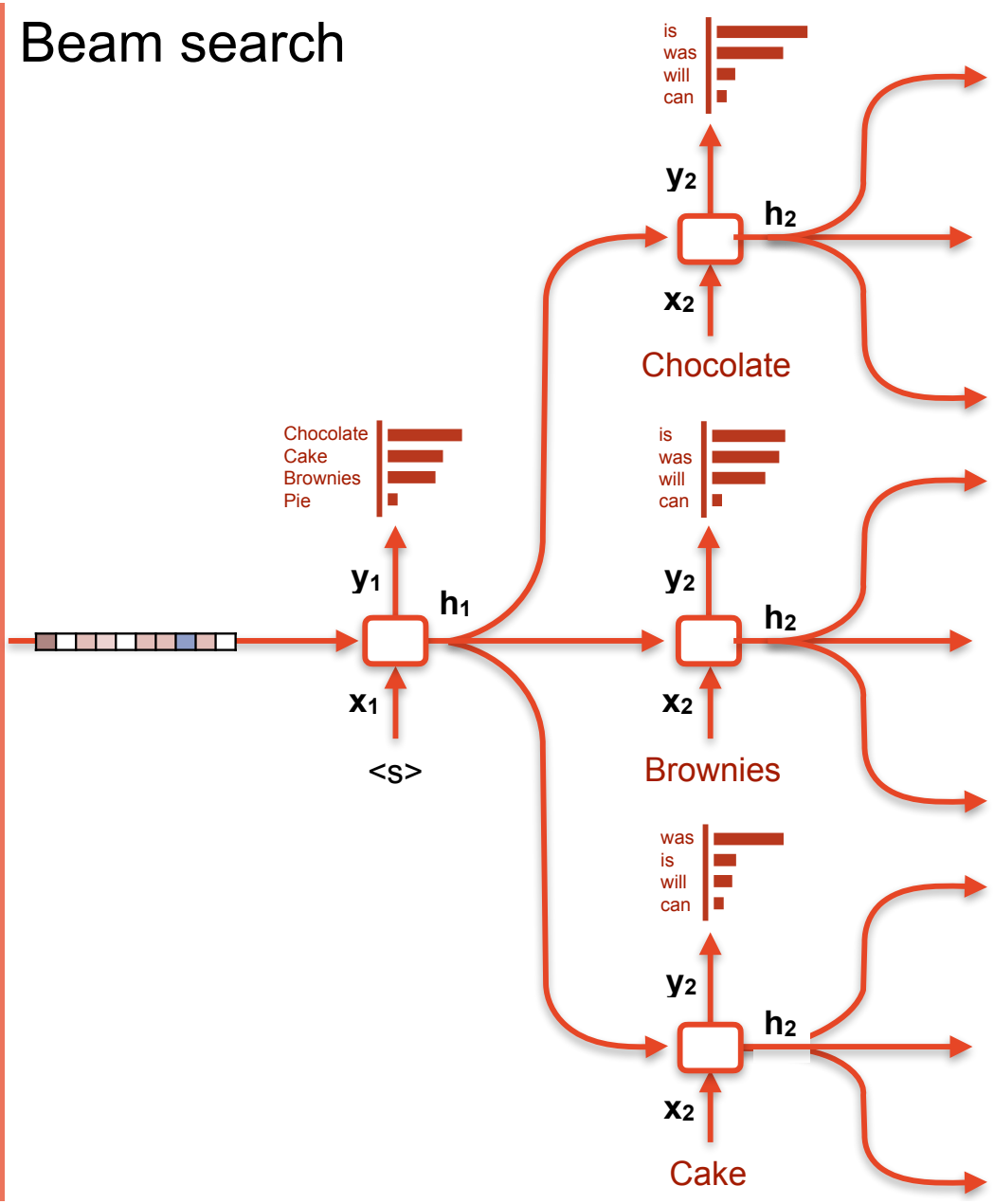
Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Beam search



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations

Encoder-Decoder

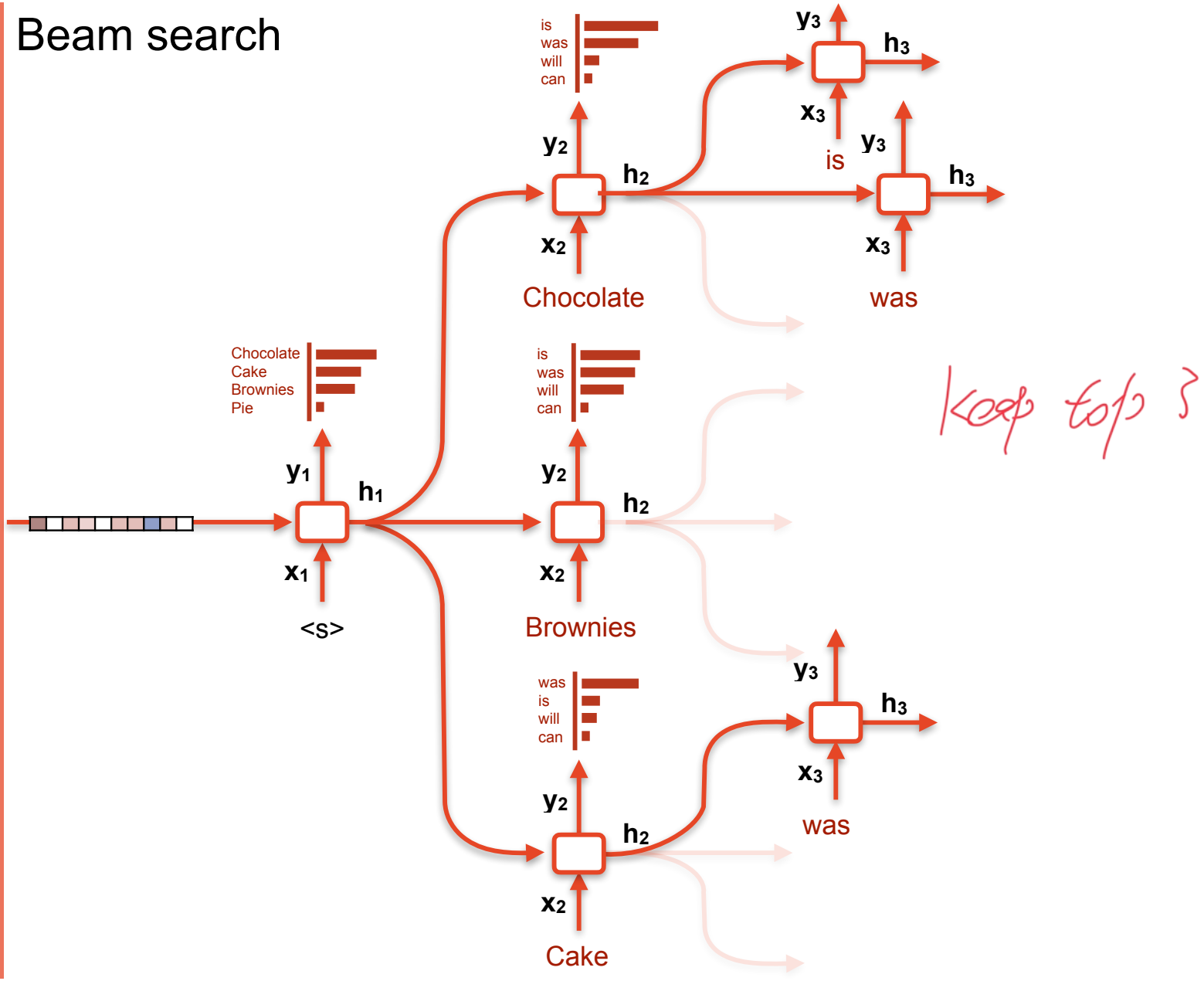
Tokenisation

Attention

Workshop Preview

menti.com 4843 3031

Beam search



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Beam search and stopping (at different lengths)

Chocolate is delicious .

Cake is delicious .

Chocolate is the most delicious food.

(1) Keep going until we have N outputs.

(2) Adjust the scores to get the average probability per word

$$\frac{1}{|\text{words}|}\sum \log P(\text{word}\,|\,\text{context})$$

deal with input with different length

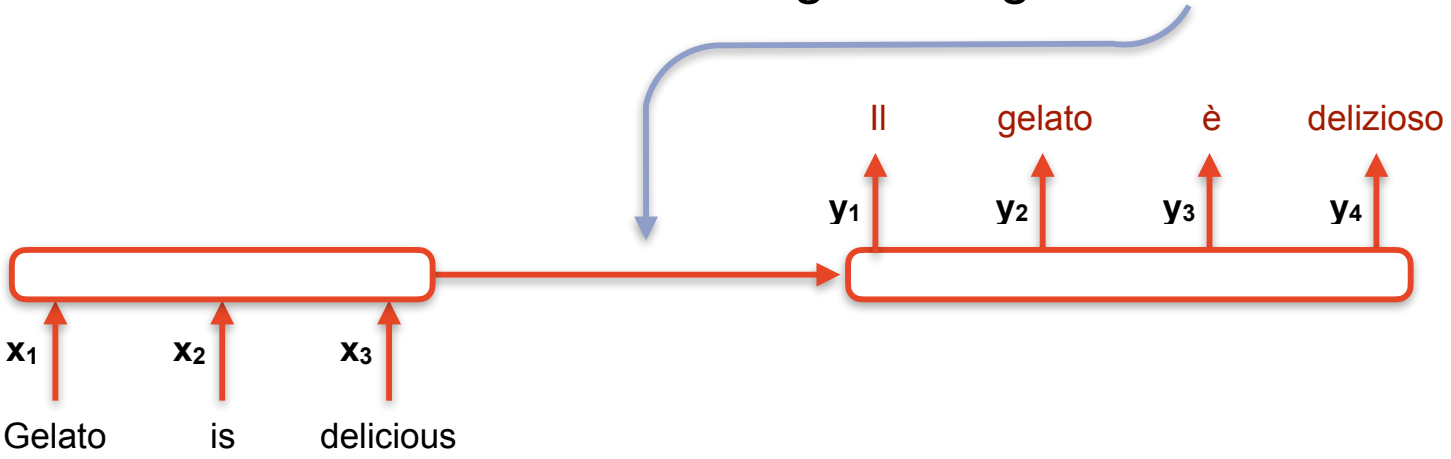
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

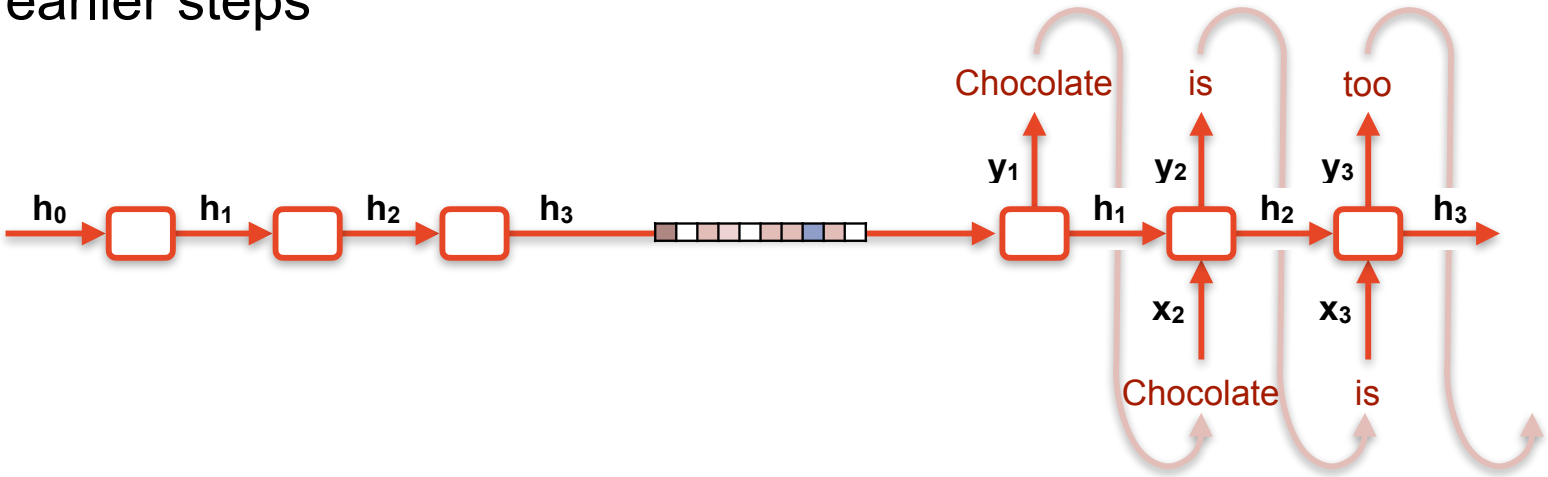
menti.com 4843 3031

Issues with the encoder-decoder as described so far

1. Bottleneck - all information has to go through one vector



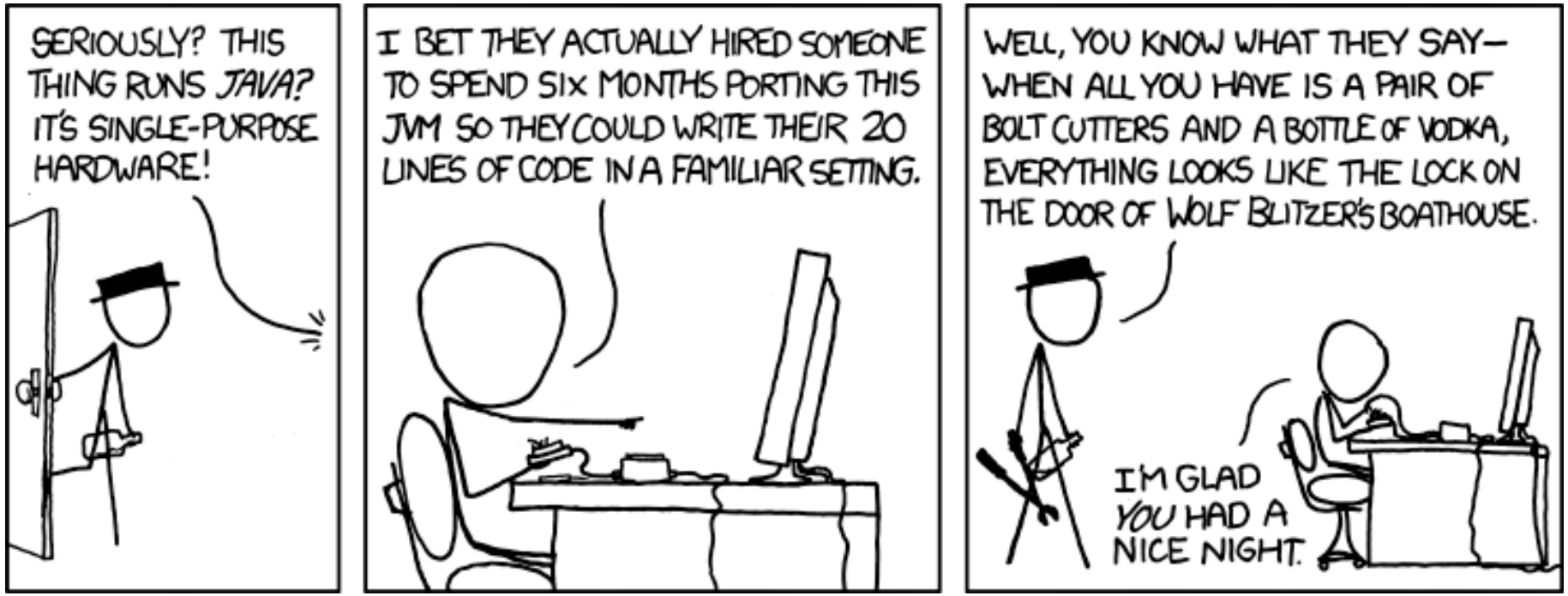
2. Hard to parallelise - every step of the calculation depends on the earlier steps



3 minute Break - stretch and visit Menti

menti.com
4843 3031

Golden Hammer



[Took me five tries to find the right one, but I managed to salvage our night out--if not the boat--in the end.]

Source: https://xkcd.com/801/

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Tokenisation

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

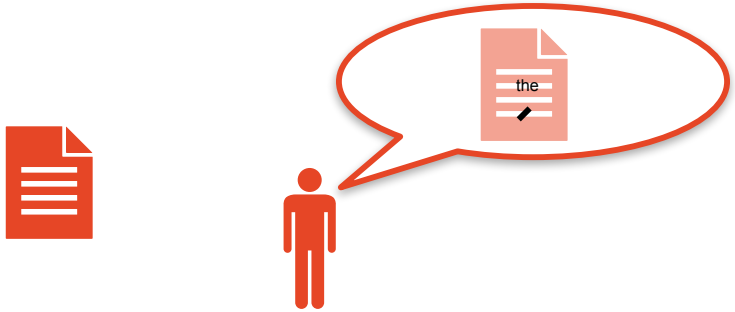
How do we measure performance of translation systems?

Human
- Fluency
- Adequacy

Rate

Rank

Edit & Compare



Bilingual? Not necessary If we have a sample translation

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Method 1

Human translation:
I like to read too

Machine translation:
Reading, I too like

chrF - compare character ngrams

I	Il	Ili	Ilik
l	li	lik	like
i	ik	ike	iket
k	ke	ket	keto
e - 2	et	eto	etor
t - 2	to - 2	tor	tore
o - 3	or	ore	orea
r	re	rea	read
a	ea	ead	eadt
d	ad	adt	adto
	dt	dto	dtoo
	oo	too	

R	Re	Rea	Read
e - 2	ea	ead	eadi
a	ad	adi	adin
d	di	din	ding
i - 2	in	ing	ing,
n	ng	ng,	ng,I
g	g,	g,I	g,It
,	,I	,It	,Ito
I	It	Ito	Itoo
t	to	too	tool
o - 2	oo	ool	ooli
l	ol	oli	olik
k	li	lik	like
	ik	ike	
	ke		

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Human translation:
I like to read too

Machine translation:
Reading, I too like

chrF - compare character ngrams

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{23}{23 + 35} = 0.40$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{23}{23 + 27} = 0.46$$

I	II	Ili	Ilik
l	li	lik	like
i	ik	ike	iket
k	ke	ket	keto
e - 2	et	eto	etor
t - 1,1	to - 1,1	tor	tore
o - 2,1	or	ore	orea
r	re	rea	read
a	ea	ead	eadt
d	ad	adt	adto
	dt	dto	dtoo
1	oo	too	

R	Re	Rea	Read
e - 2	ea	ead	eadi
a	ad	adi	adin
d	di	din	ding
i - 1,1	in	ing	ing,
n	ng	ng,	ng,I
g	g,	g,I	g,It
,	,I	,It	,Ito
I	It	Ito	Itoo
t	to	too	tool
o - 2	oo	ool	ooli
l	ol	oli	olik
k	li	lik	like
	ik	ike	
1	ke		

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Human translation:
I like to read too

Machine translation:
Reading, I too like

chrF - compare character ngrams

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{23}{23 + 35} = 0.40$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{23}{23 + 27} = 0.46$$

$$\text{F}_\beta\text{-score} = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}$$

$$\text{F}_2\text{-score} = \frac{5TP}{5TP + 4FN + FP} = \frac{5 \cdot 23}{5 \cdot 23 + 4 \cdot 27 + 35} = 0.45$$

The value of β leads the score to be closer to either P or R

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Human translatio
I like to read too

chrF - compare character ng

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{23}{23 + 35} = 0.40$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{23}{23 + 27} = 0.46$$

$\beta = 0$

$$\text{F}_0\text{-score} = \frac{(1+0)TP}{(1+0)TP + 0FN + FP} = \frac{TP}{TP + FP} = \text{Precision}$$

$$\text{F}_N\text{-score} = \frac{(1+N)TP}{(1+N)TP + N \cdot FN + FP}$$

If N is big

$$\approx \frac{N \cdot TP}{N \cdot TP + N \cdot FN + FP}$$

$\beta \to \infty$

$$= \frac{TP}{TP + FN + \frac{FP}{N}}$$

$$\approx \frac{TP}{TP + FN} = \text{Recall}$$

to either P or R

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Huma
I like t

Can also use word ngrams

Machine translation:
Reading, I too like

chrF - compare character ngrams

I	II	Ili	Ilik
l	li	lik	like
i	ik	ike	iket
k	ke	ket	keto
e - 2	et	eto	etor
t - 1,1	to - 2	tor	tore
o - 1,2	or	ore	orea
		rea	read
		ead	eadt
		adt	adto
		dto	dtoo
		too	

R	Re	Rea	Read
e - 2	ea	ead	eadi
a	ad	adi	adin
d	di	din	ding
i - 1,1	in	ing	ing,
n	ng	ng,	ng,I
g	g,	g,I	g,It
,	,I	,It	,Ito
I	It	Ito	Itoo
t	to	too	tool
o - 2	oo	ool	ooli
l	ol	oli	olik
k	li	lik	like
	ik	ike	
	ke		

Can move large chunks of the sentence and not impact the score. Maybe good? Maybe not?

move two part of sentence, only middle will be effect

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Method 2

Human translation:
I like to read too

Machine translation:
Reading, I too like

BLEU - compare word ngrams

I	Reading
like	,
to	I
read	too
too	like
I like	Reading ,
like to	, I
to read	I too
read too	too like
I like to	Reading , I
like to read	, I too
to read too	I too like
I like to read	Reading , I too
like to read too	, I too like

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do we measure performance of translation systems?

Automatic

Human translation:
I like to read too

Machine translation:
Reading, I too like

Tokenisation?

BLEU - compare word ngrams

- Calculate precision for unigrams, bigrams, trigrams, 4grams separately
- Take the geometric mean
- Appl
  is sh

Geometric mean only works well with multiple sentences - no 4grams matches means a score of 0!

I	Reading
like	
to	
read	
too	
I like	Reading ,
	, I
ad	I too
too	too like
to	Reading , I
read	, I too
ad too	I too like
I like to read	Reading , I too
like to read too	, I too like

No credit for read vs. Reading

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

So far, we have always split on whitespace to get tokens

"Chocolate is delicious"

`text.split()`

"Chocolate" "is" "delicious"

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

What about these cases?

Contractions: won’t

Punctuation: Great!

e.g.,

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

And can tokenisation help with rare words?

Variations	taaaaaaasty
Misspellings	laern
New words	transformerify

Previously we used method's like Fasttext's representations to address this

Examples from Stanford's cs224n course

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Split into sub-tokens

Variations	taaaaaaasty	→	ta## aaaaa## asty
Misseplings	laern	→	la## ern
New words	transformerify	→	transformer## ify

How do we determine where to split?

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

Vocab:

_, d, e, i, l, n, o, r, s, t, w

Pairs and counts:

9 r _	2 t _
9 e r	2 s t
8 w e	2 e s
8 n e	
8 e w	
7 w _	
7 o w	
7 l o	
3 w i	
3 i d	
3 d e	

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l o w _
2 l o w e s t _
6 n e w e r_
3 w i d e r_
2 n e w _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_

Pairs and counts:

9 e r_
8 w e
8 n e
8 e w
7 w _
7 o w
7 l o
3 w i
3 i d
3 d e
2 t _
2 s t
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_

new

Pairs and counts:

8 n e
8 e w
7 w _
7 o w
7 l o
6 w er_
3 w i
3 i d
3 d er_
2 w e
2 t _
2 s t
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l o w _
2 l o w e s t _
6 ne w er_
3 w i d er_
2 ne w _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne

Pairs and counts:

8 ne w 2 e s
7 w _
7 o w
7 l o
6 w er_
3 w i
3 i d
3 d er_
2 w e
2 t _
2 s t

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l o w _
2 l o w e s t _
6 new er_
3 w i d er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new

Pairs and counts:

7 o w
2 e s
7 l o
6 new er_
5 w _
3 w i
3 i d
3 d er_
2 w e
2 t _
2 s t
2 new _

Examples from Jurafsky and Martin textbook

breaking tie in various ways

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 l ow _
2 l ow e s t _
6 new er_
3 w i d er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new, ow

Pairs and counts:

7 l ow
6 new er_
5 ow _
3 w i
3 i d
3 d er_
2 t _
2 s t
2 ow e
2 new _
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low _
2 low e s t _
6 new er_
3 w i d er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new, ow, low

Pairs and counts:

6 new er_
5 low _
3 w i
3 i d
3 d er_
2 t _
2 s t
2 new _
2 low e
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low _
2 low e s t _
6 newer_
3 w i d er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new, ow, low, newer_

Pairs and counts:

5 low _
3 w i
3 i d
3 d er_
2 t _
2 s t
2 new _
2 low e
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:
5 low_
2 low e s t _
6 newer_
3 w i d er_
2 new _

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new, ow, low, newer_, low_

Pairs and counts:
3 w i
3 i d
3 d er_
2 t _
2 s t
2 new _
2 low e
2 e s

Examples from Jurafsky and Martin textbook

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low_
2 low e s t _
6 newer_
3 wi d er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi

Pairs and counts:

3 wi d
3 d er_
2 t _
2 s t
2 new _
2 low e
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low_
2 low e s t _
6 newer_
3 wid er_
2 new _

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid

Pairs and counts:

3 wid er_
2 t _
2 s t
2 new _
2 low e
2 e s

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5	low_
2	low e s t _
6	newer_
3	wider_
2	new _

Pairs and counts:

2	t _
2	s t
2	new _
2	low e
2	e s

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5	low_
2	low e s t_
6	newer_
3	wider_
2	new _

Pairs and counts:

2 s t_
2 new _
2 low e
2 e s

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_, t_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:

1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low_
2 low e st_
6 newer_
3 wider_
2 new _

Pairs and counts:

2 new _
2 low e
2 e st_

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_, t_, st_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:
5 low_
2 low e st_
6 newer_
3 wider_
2 new_

Pairs and counts:
2 low e
2 e st_

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_, t_, st_,
new_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:
5 low_
2 lowe st_
6 newer_
3 wider_
2 new_

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_, t_, st_,
new_, lowe

Pairs and counts:
2 lowe st_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:

5 low_
2 lowest_
6 newer_
3 wider_
2 new_

Vocab:

_, d, e, i, l, n, o, r, s, t, w, r_,
er_, ne, new, ow, low, newer_,
low_, wi, wid, wider_, t_, st_,
new_, lowe, lowest_

Pairs and counts:

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Byte-Pair Encoding (BPE)

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

Data:
5 low_
2 lowest_
6 newer_
3 wider_
2 new_

Pairs and counts:

Vocab:
_, d, e, i, l, n, o, r, s, t, w, r_, er_, ne, new, ow, low, newer_, low_, wi, wid, wider_, t_, st_, new_, lowe, lowest_

make new words

⇒ low er_
new e st_
wid e st_

Examples from Jurafsky and Martin textbook

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Other similar methods vary step 2

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items using some method
3. Create a new vocabulary item for that pair
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

1) BPE

Step 2 'some method' = adjacent most frequently

2) WordPiece

Step 2 'some method' =
- Train an n-gram language model
- Consider all possible vocabulary pairs
- Choose the one that will decrease perplexity most if added to the model

Step 2 is different for those methods

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Other similar methods vary step 2

Algorithm:
1. Set vocabulary to be all characters
2. Find the two vocabulary items using some method
3. Create a new vocabulary item for that pair
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

HuggingFace
WordPiece

Step 2 'some method' =
Pair that maximises

$$\frac{|\text{combined}|}{|\text{first symbol}| \cdot |\text{second symbol}|}$$

https://huggingface.co/learn/nlp-course/en/chapter6/6

COMP 4446 / 5046
Lecture 6, 2025

menti.com 4843 3031

Unigram / SentencePiece takes a different approach

So far: start with small units, then combine units

Alternative: start with big and small units, then delete units

Algorithm:
1. Set vocabulary to be all characters and all frequent character sequences, up to full words.
2. Find vocabulary items to remove using a complex method.
3. Repeat until the vocabulary gets down to the desired size.

How? I don't think it's done size

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

How do they compare?

Original:	furiously	**Original:**	tricycles	**Original:**	nanotechnology
BPE:	_fur ious ly	**BPE:**	_t ric y cles	**BPE:**	_n an ote chn ology
Uni. LM:	_fur ious ly	**Uni. LM:**	_tri cycle s	**Uni. LM:**	_nano technology

Original:	Completely preposterous suggestions
BPE:	_Comple t ely _prep ost erous _suggest ions
Unigram LM:	_Complete ly _pre post er ous _suggestion s

Original:	corrupted	**Original:**	1848 and 1852,
BPE:	_cor rupted	**BPE:**	_184 8 _and _185 2,
Unigram LM:	_corrupt ed	**Unigram LM:**	_1848 _and _1852 ,

Original	磁性は様々に分類がなされている。							
BPE	磁	性は	様々	に分類	がなされている	。		
Unigram LM	磁	性	は	様々	に	分類	がなされている	。
Gloss	magnetism	(top.)	various ways	in	classification	is done	.	
Translation	Magnetism is classified in various ways.							

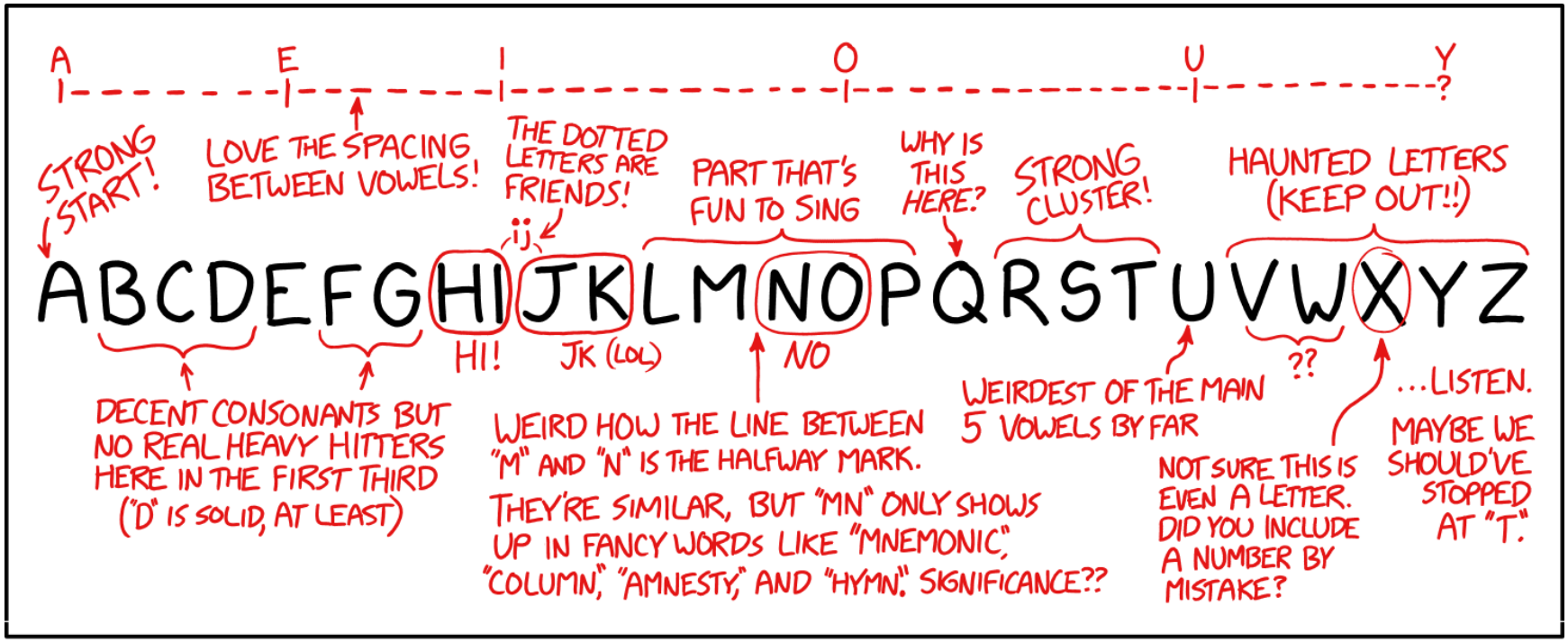
More frequent in BPE					**Unigram LM**				
_H	_L	_M	_T	_B	s	.	,	ed	d
_P	_C	_K	_D	_R	ing	e	ly	t	_a

Bostrom and Durrett (2021)

3 minute Break - stretch and visit Menti

menti.com
4843 3031

Alphabet Notes



DESIGN NOTES ON THE ALPHABET

[Listen, you're very cute, but if you rearrange the alphabet to put U and I together it will RUIN the spacing!]

Source: https://xkcd.com/2794/

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Attention

COMP 4446 / 5046
Lecture 6, 2025

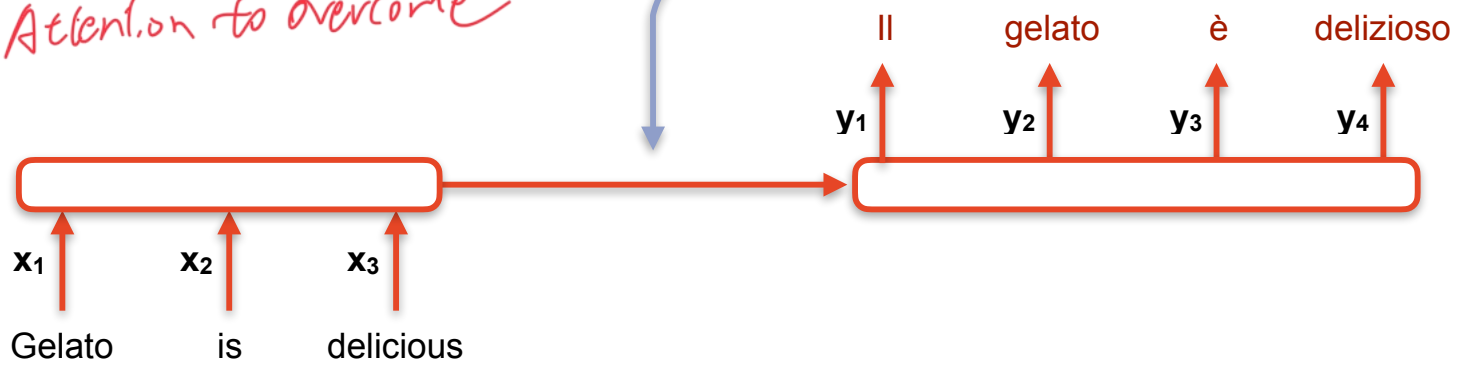
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

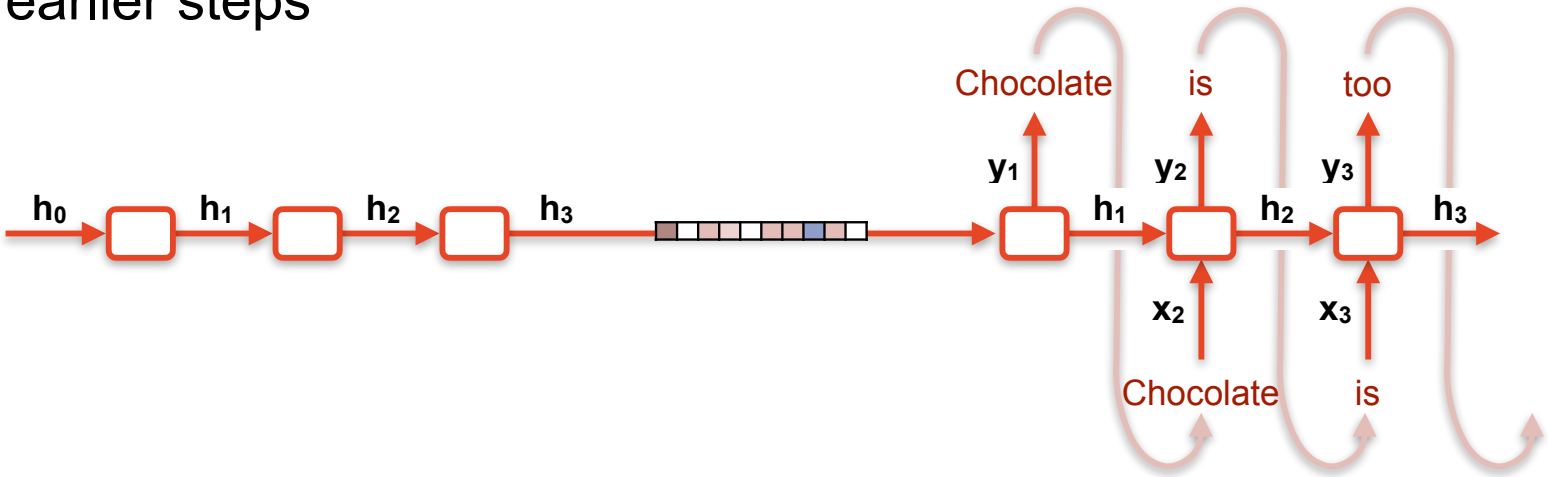
Reminder: Issues with the encoder-decoder

Bottleneck - all information has to go through one vector

Use Attention to overcome



Hard to parallelise - every step of the calculation depends on the earlier steps



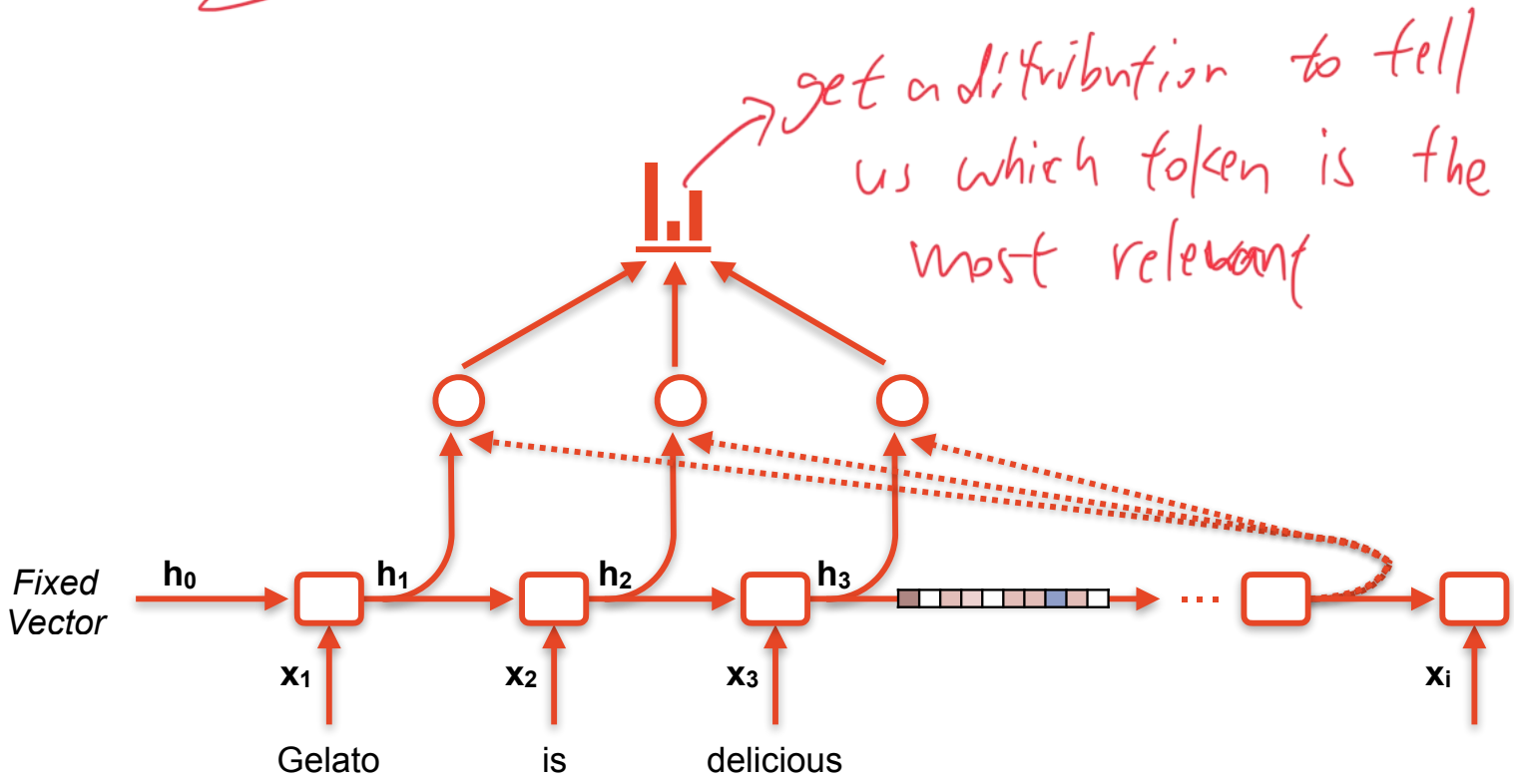
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We will solve this with ‘attention’

Give each output step
a representation of the input
that is most useful for that output decision



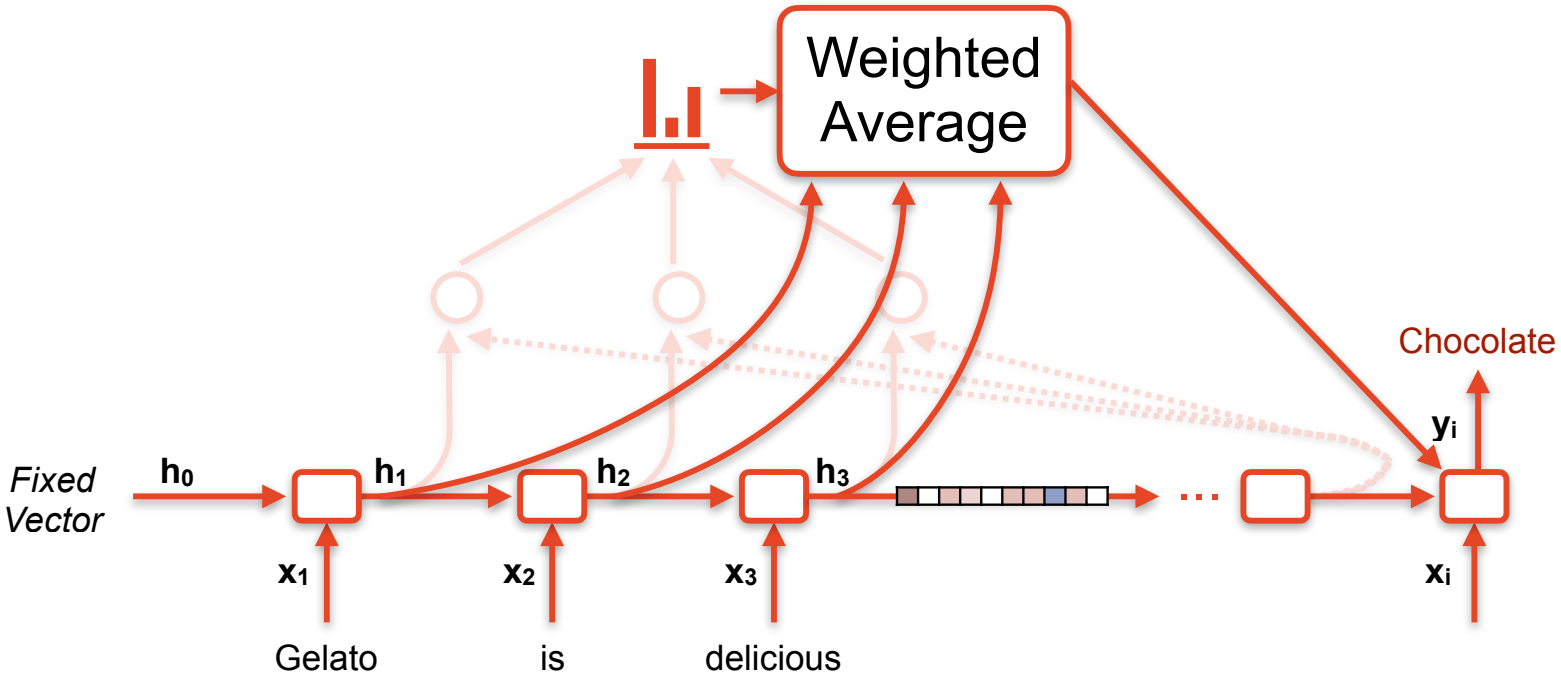
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We will solve this with ‘attention’

Give each output step
a representation of the input
that is most useful for that output decision



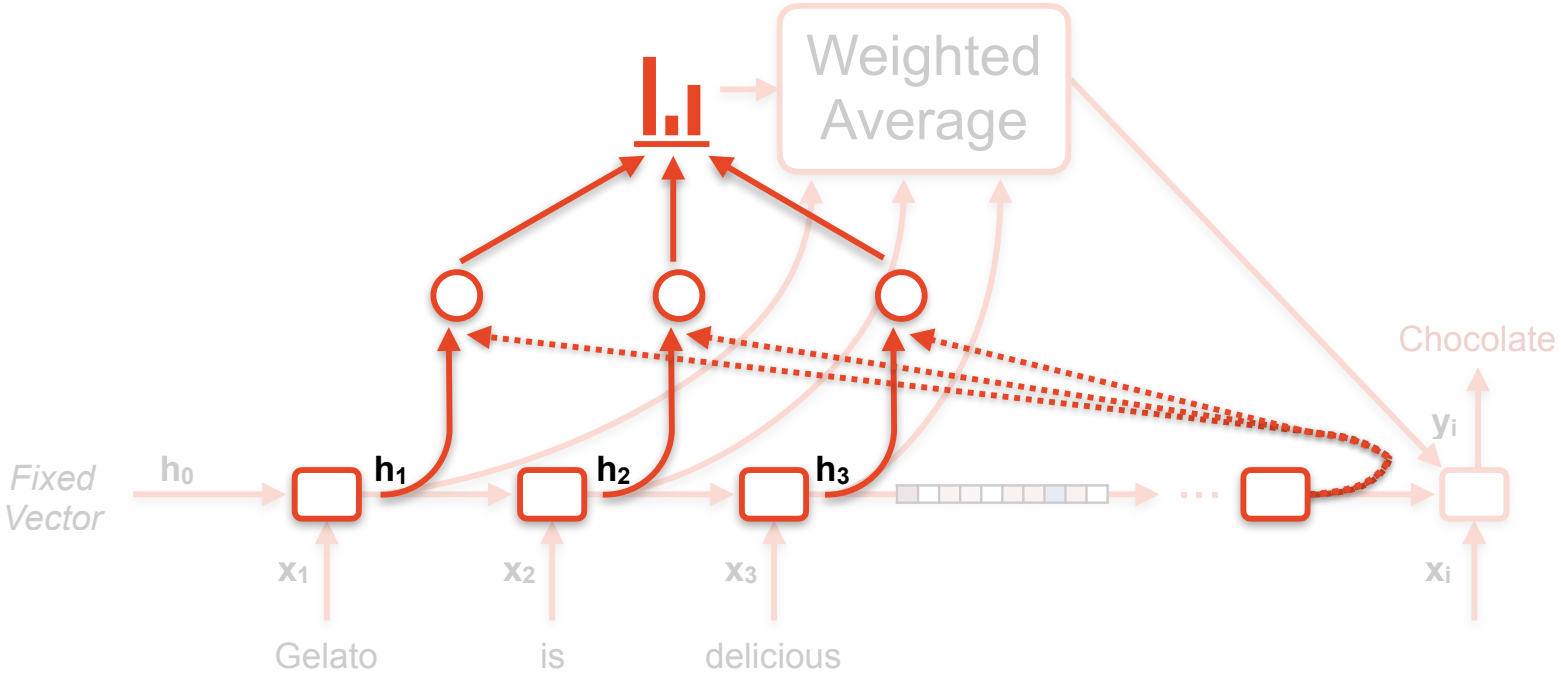
COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We will solve this with ‘attention’

Give each output step
a representation of the input
that is most useful for that output decision



COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We will solve this with ‘attention’

Give each output step
a representation of the input
that is most useful for that output decision

Hidden vectors from input $\mathbf{h_1}, \mathbf{h_2}, \mathbf{h_3} \ldots \mathbf{h_N} \in \mathbb{R}^{d_h}$

Hidden vector for current output step $\mathbf{s_i} \in \mathbb{R}^{d_h}$

Calculate attention scores $\mathbf{e_i} = [\mathbf{s_i}^\top \mathbf{h_1}, \mathbf{s_i}^\top \mathbf{h_2}, \mathbf{s_i}^\top \mathbf{h_3}, \ldots \mathbf{s_i}^\top \mathbf{h_N}] \in \mathbb{R}^N$

This is dot product attention

Normalise s … ax($\mathbf{e_i}$) $\in \mathbb{R}^N$

Calculate weighted average $\mathbf{a_i} = \sum_{j=1}^{N} \alpha_{i,j} \mathbf{h_j} \in \mathbb{R}^{d_h}$

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

There are many forms of attention

Dot product attention

$$\mathbf{e} = \mathbf{s}^\top \mathbf{h}$$

Scaled dot product attention

$$\mathbf{e} = \frac{\mathbf{s}^\top \mathbf{h}}{\sqrt{d_h}}$$

Motivated by statistical properties of dot products

Multiplicative attention / Bilinear attention

$$\mathbf{e} = \mathbf{s}^\top \mathbf{W} \mathbf{h}$$

s and h can be different sizes

Reduced-rank multiplicative attention

$$\mathbf{e} = \mathbf{s}^\top (\mathbf{U}^\top \mathbf{V}) \mathbf{h}$$
$$= (\mathbf{s}\mathbf{U})^\top (\mathbf{V}\mathbf{h})$$

Can improve efficiency

Additive attention / Feedforward attention

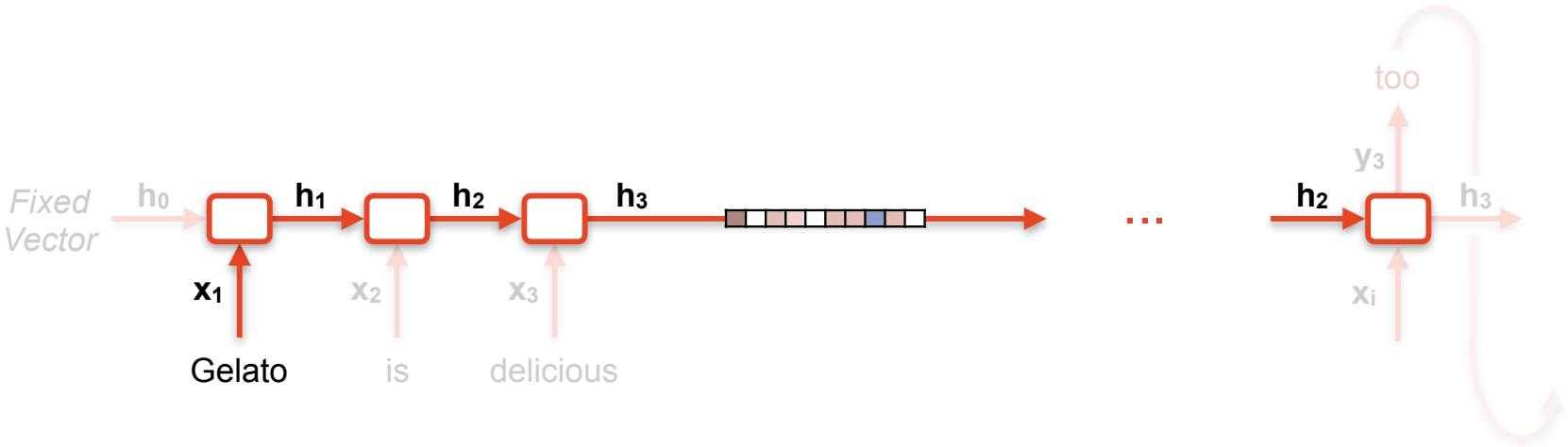
$$\mathbf{e} = \mathbf{b} \tanh(\mathbf{W}_1 \mathbf{h} + \mathbf{W}_2 \mathbf{s})$$

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

This resolves the bottleneck problem without introducing many extra parameters

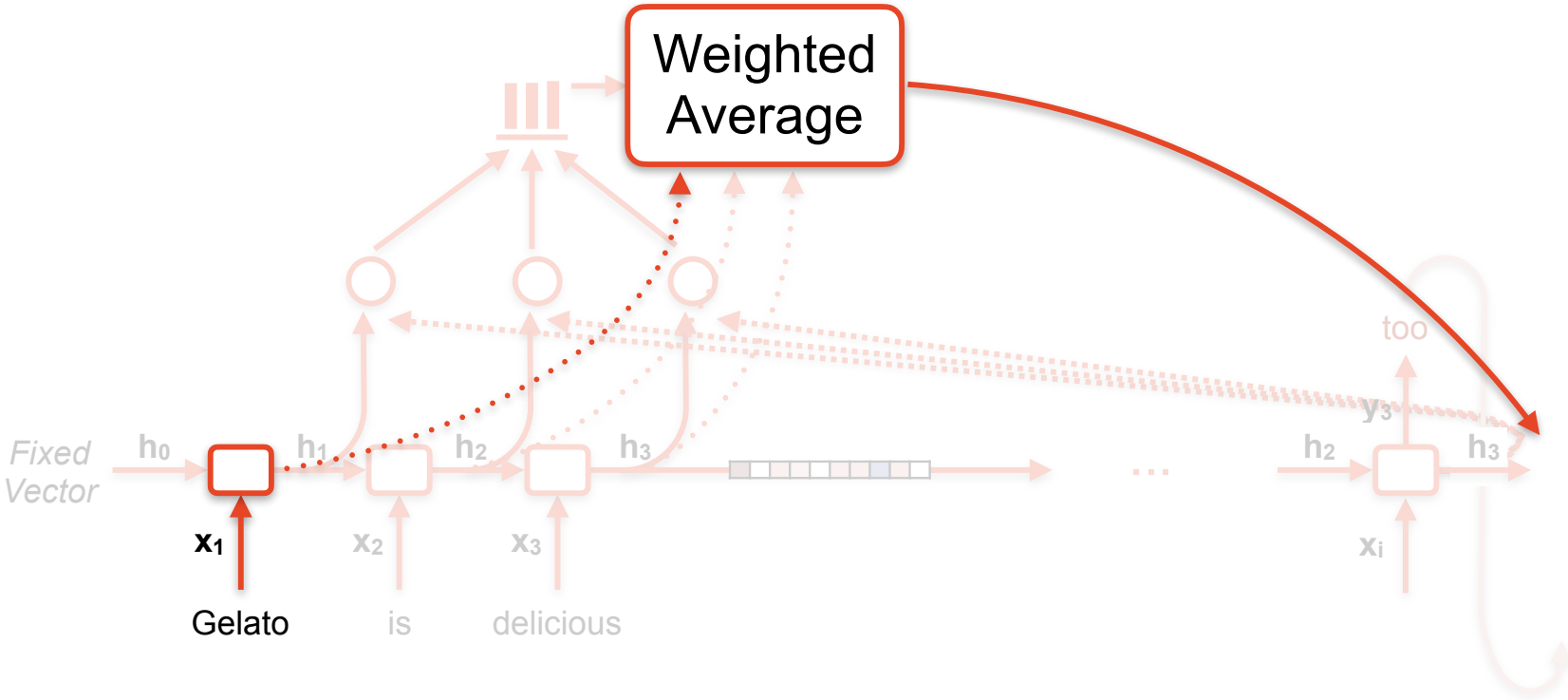


COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

This resolves the bottleneck problem without introducing many extra parameters



COMP 4446 / 5046
Lecture 6, 2025

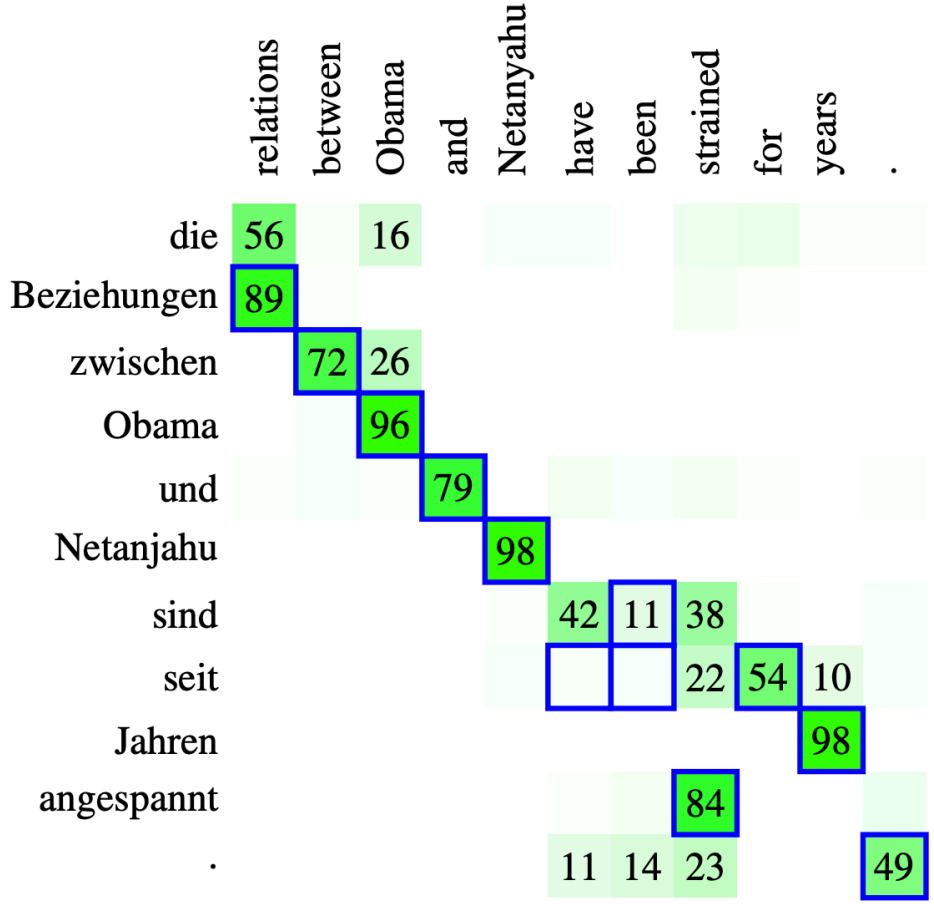
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Attention in machine translation seems to capture word alignment!

Blue boxes:
Alignment
algorithm output

Numbers:
Attention scores



Koehn and Knowles (2017)

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

We need to be careful not to read too much into these

Understanding Neural Networks through Representation Erasure

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Attention is not not Explanation

Is Attention Explanation? An Introduction to the Debate

Adrien Bibal, **Rémi Cardon**, **David Alfter**, **Rodrigo Wilkens**, **Xiaoou Wang**, **Thomas François*** and **Patrick Watrin***
CENTAL, IL&C, University of Louvain, Belgium
{adrien.bibal,remi.cardon,david.alfter,xiaoou.wang,
thomas.francois,patrick.watrin}@uclouvain.be

Abstract

The performance of deep learning models in NLP and other fields of machine learning has led to a rise in their popularity, and so the need for explanations of these models becomes paramount. Attention has been seen as a solution to increase performance, while providing some explanations. However, a debate has started to cast doubt on the explana-

(global vs. local attention, according to Luong et al. (2015)) and where the query is generated (cross vs. self-attention as in the works of Bahdanau et al. (2015) and Vaswani et al. (2017)). In this paper, we focus on attention regardless of these technical differences. There are mainly two ways of computing the attention weights $\hat{\alpha}$: Bahdanau et al. (2015) introduced additive attention $\hat{\alpha} = \text{softmax}(\mathbf{w_3}^{\text{T}}\tanh(\mathbf{W_1}K + \mathbf{W_2}Q))$, where

Yuval Pinter*
ol of Interactive Computing
rgia Institute of Technology
uvp@gatech.edu

ns for, e.g., model debugging or
ction. A recent paper (Jain and
oints to possible pitfalls that may
s to misapply attention scores as
model behavior, based on a pren
ble attention distributions should
with other feature-importance mea
exclusive given a prediction.[1] Its
which we elaborate in §2, is th
attention distributions exist that
results to those obtained by the
hen the original model's attentio

COMP 4446 / 5046
Lecture 6, 2025

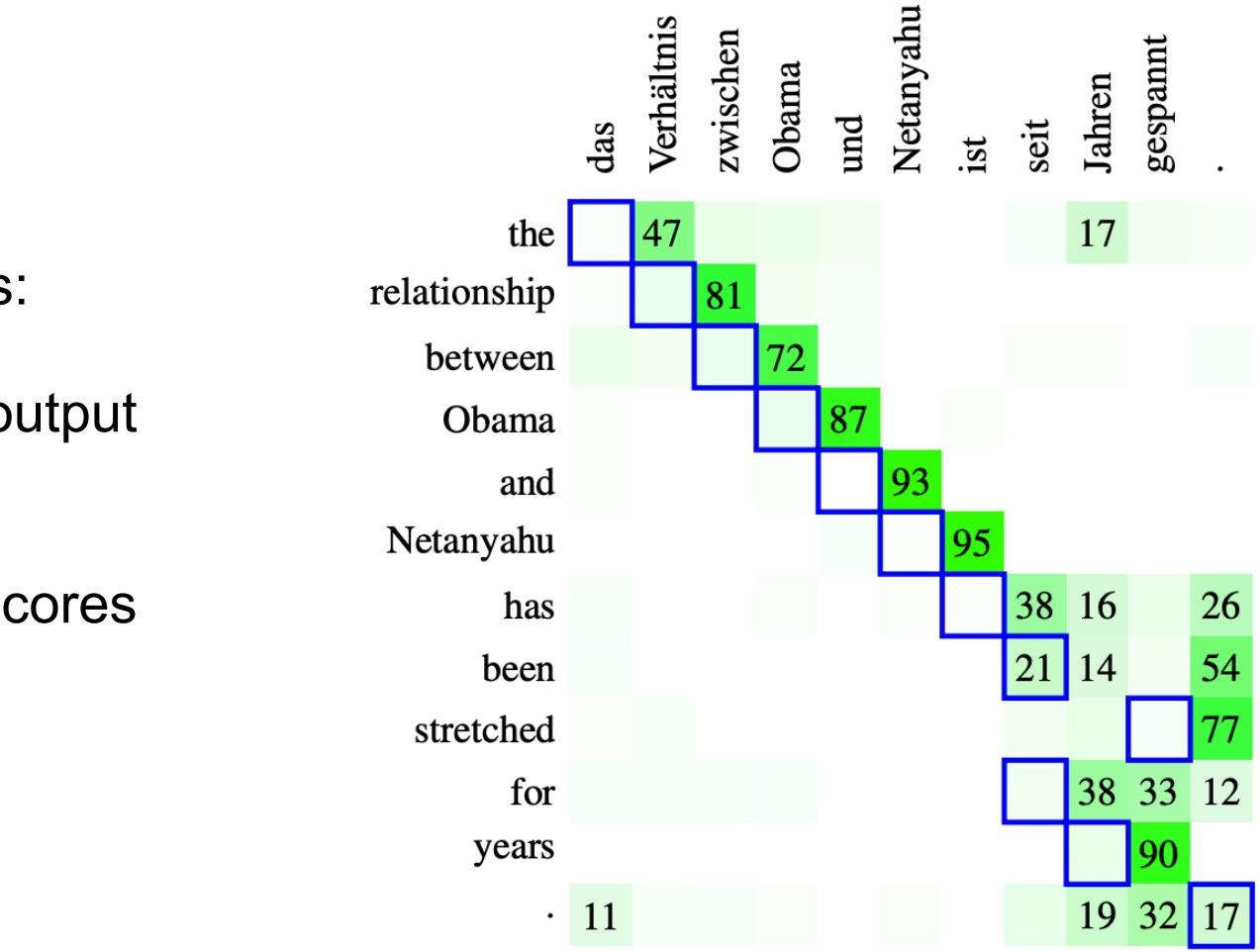
Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Attention in machine translation seems to capture word alignment! Or does it?

Blue boxes:
Alignment
algorithm output

Numbers:
Attention scores



Koehn and Knowles (2017)

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Workshop Preview

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Pre-work: None this week

In-class: spaCy

COMP 4446 / 5046
Lecture 6, 2025

Contextual Representations
Encoder-Decoder
Tokenisation
Attention
Workshop Preview

menti.com 4843 3031

Muddy Card

Open shortly, closes at 7:05pm

https://saipll.shinyapps.io/student-interface/

If you do not wish to participate in the study, use the Ed form instead

Go to Ed → Lessons → Muddy Cards Lecture 6