

Word Embeddings and A Very Simple Word Embedding Based Model

Matthew Engelhard



Problem: our model counts words,
but has no understanding of their meaning

happy	content	joyful	satisfied	merry	ecstatic	gleeful	euphoric	sad	unhappy	melancholy	depressed	upset	down	miserable	sorrowful

Goal: predict sentiment (positive/negative)



To effectively predict sentiment, it would be helpful to understand which words have similar meaning

I am sad
I am miserable
I am sorrowful
I am upset
I am down
I am content
I am joyful
I am merry
I am satisfied
I am euphoric



training set

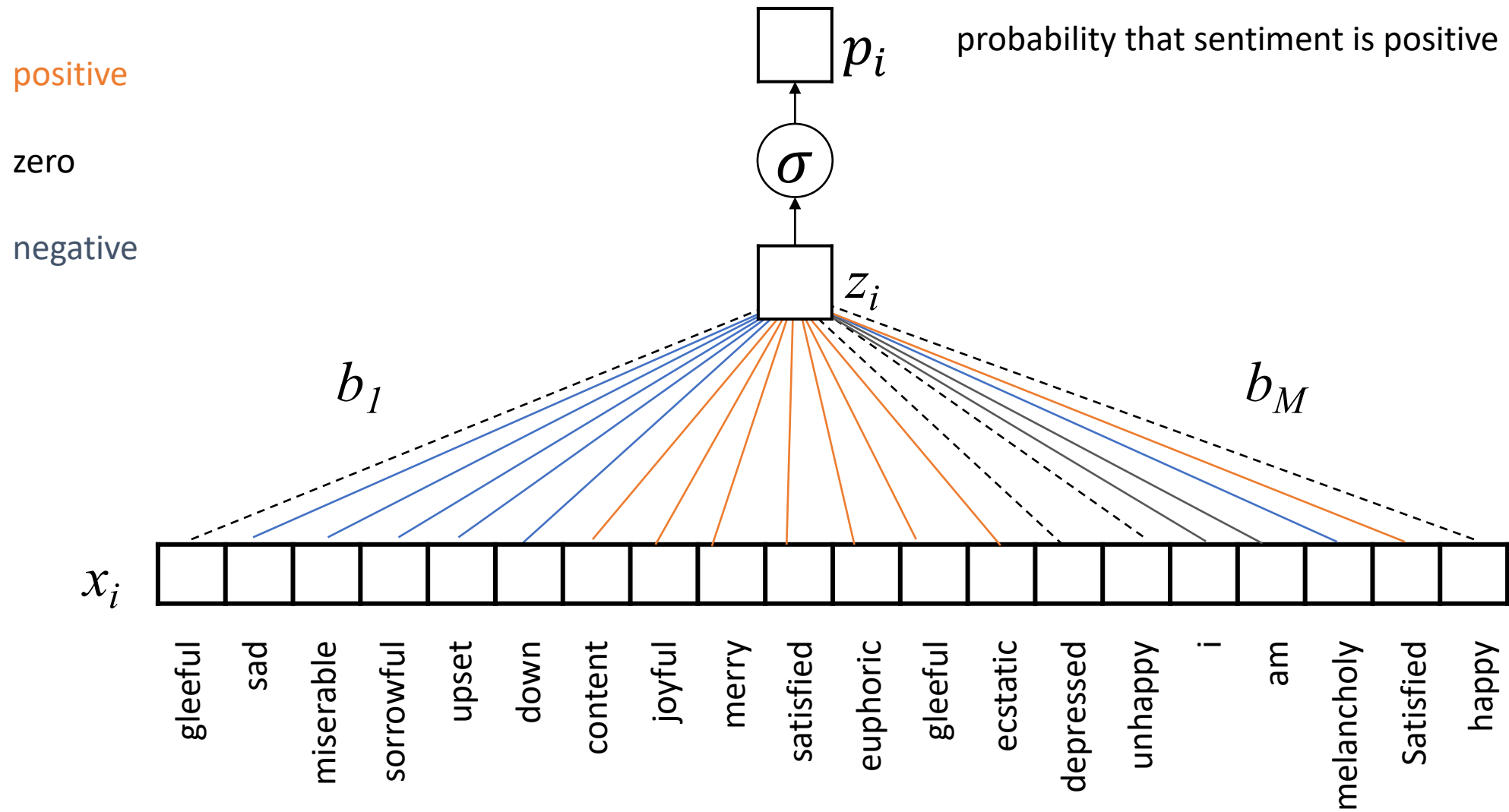
I am depressed
I am unhappy
I am happy
I am gleeful



test set



logistic regression: positive / negative sentiment

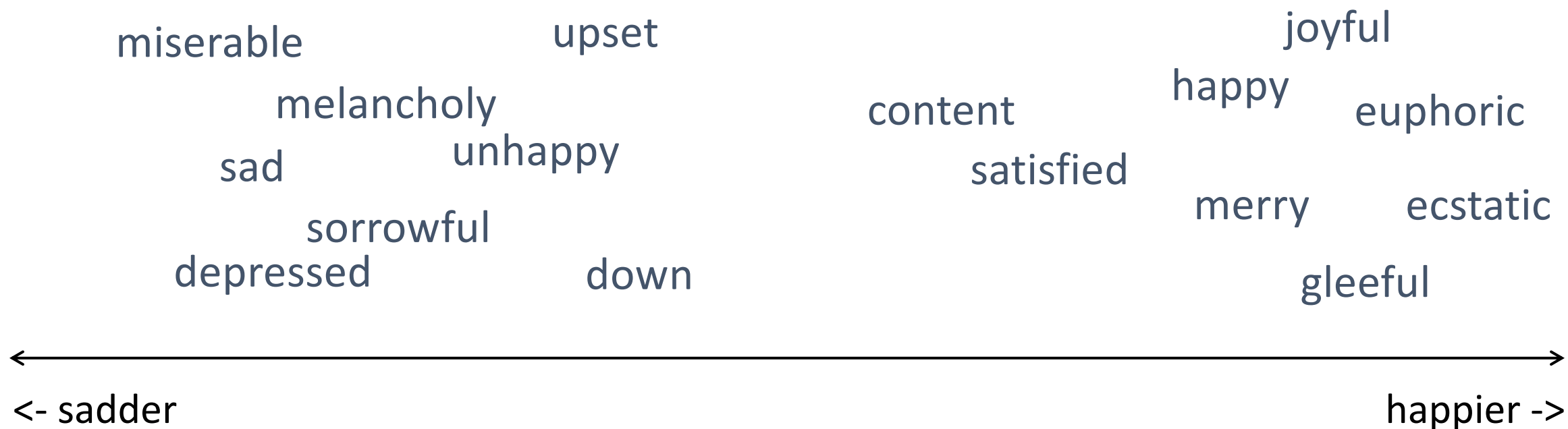


"I'm so depressed coach K is retiring..."

"I'm so happy coach K is finally retiring..."



We'd like a numeric representation of words that encodes their meaning



Numeric value indicating whether the word is happy or sad



Training a robot to buy groceries



Example from Anand Chowdhury, MMCI 2019



Grocery List

- ☐ granulated sugar
- ☐ vanilla extract
- ☐ dark brown sugar
- ☐ carrots
- ☐ table salt
- ☐ eggs



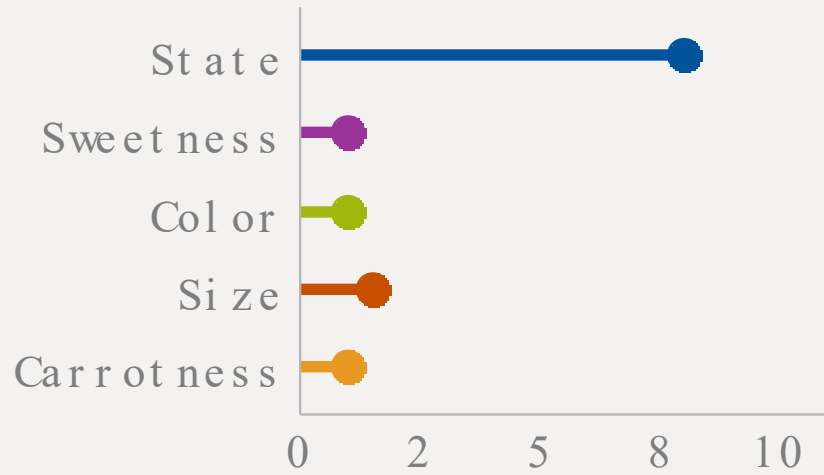
Characteristics/Dimensions

Dimension	1	10
State	Liquid	Solid
Sweetness	Bland	Sweet
Color	Light	Dark
Size	Small	Large
Carrottness	Not really	

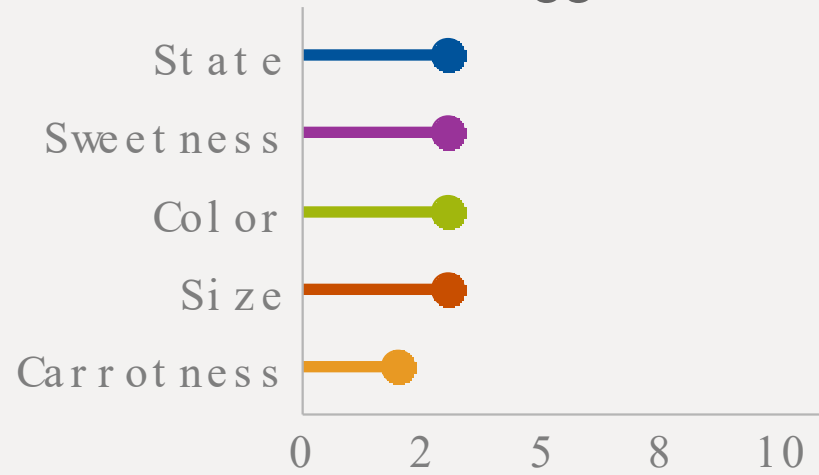


Five dimensions

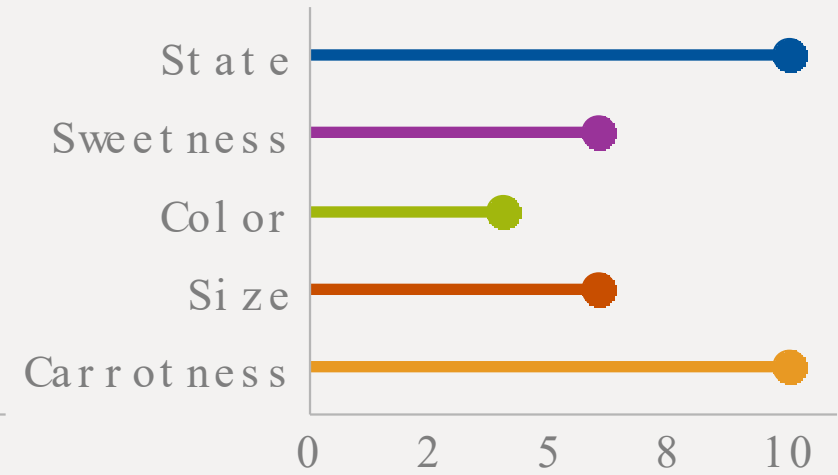
Table Salt



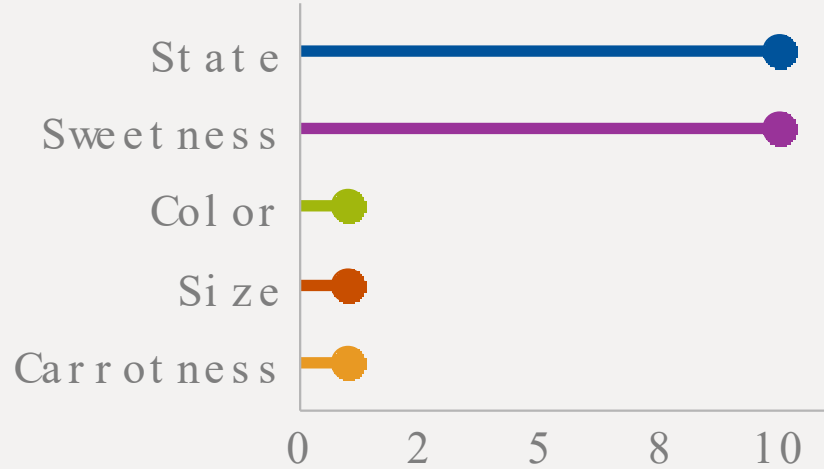
Eggs



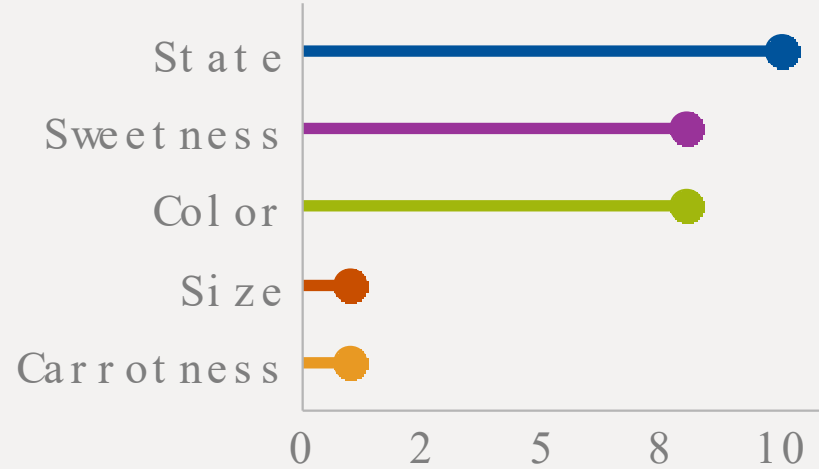
Carrots



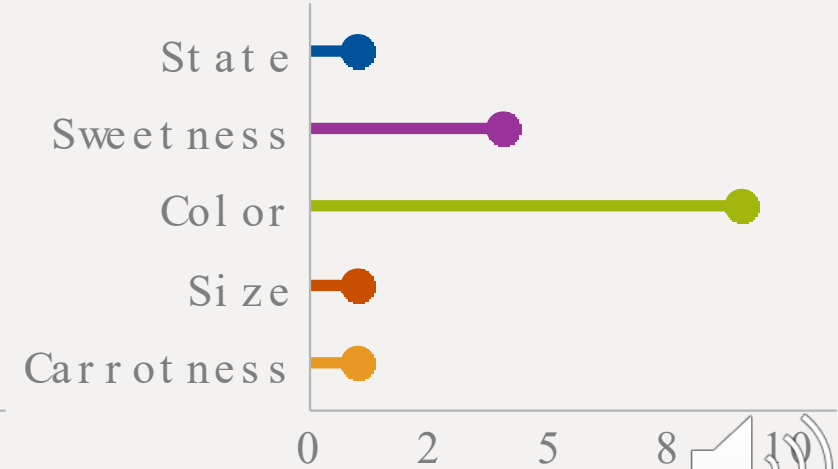
Granulated Sugar



Dark Brown Sugar



Vanilla Extract



Make Sense of Items not Seen Before

Item	State	Sweetness	Color	Size	Carrottness
???	0	8	7	6	0
???	0	0	10	6	0
???	8	9	8	3	0
???	0	5	3	4	10



Make Sense of Items not Seen Before

Item	State	Sweetness	Color	Size	Carrotiness
Soda / Sweet Tea	0	8	7	6	0
Black Coffee	0	0	10	6	0
Chocolate	8	9	8	3	0
Carrot Juice	0	5	3	4	10



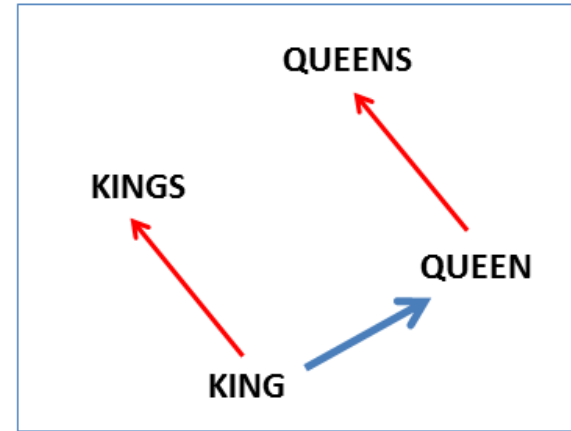
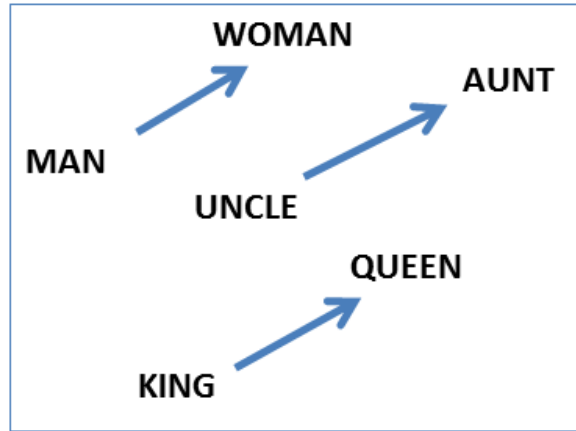
Recipe

Dark Brown Sugar – Granulated Sugar + Carrots

	Item	State	Sweetness	Color	Size	Carrotness
	Dark Brown Sugar	10	8	8	1	1
-	Granulated Sugar	10	10	1	1	1
+	Carrots	10	6	4	6	10
=	???	10	4	11	6	10



Word Embeddings: Assign Each Word in our Vocabulary to a Numeric Vector (of characteristics)



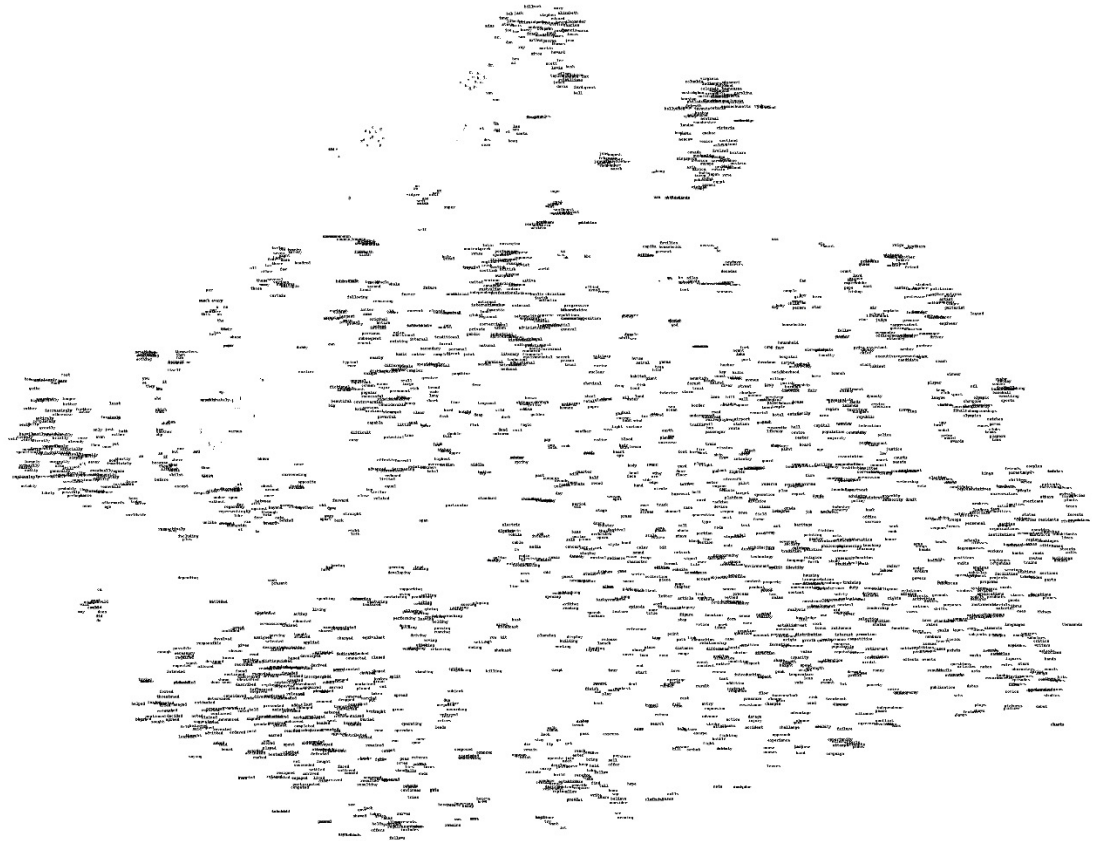
Dimension	1	10
Gender	Male	Female
Class	Commoner	Royalty
Plural	One	Many



Visualizing Word Embeddings

Here we show the learned numeric representations (limited here to 2 dimensions) of many different vocabulary words

Too many words here to see! Let's zoom in on a smaller section.

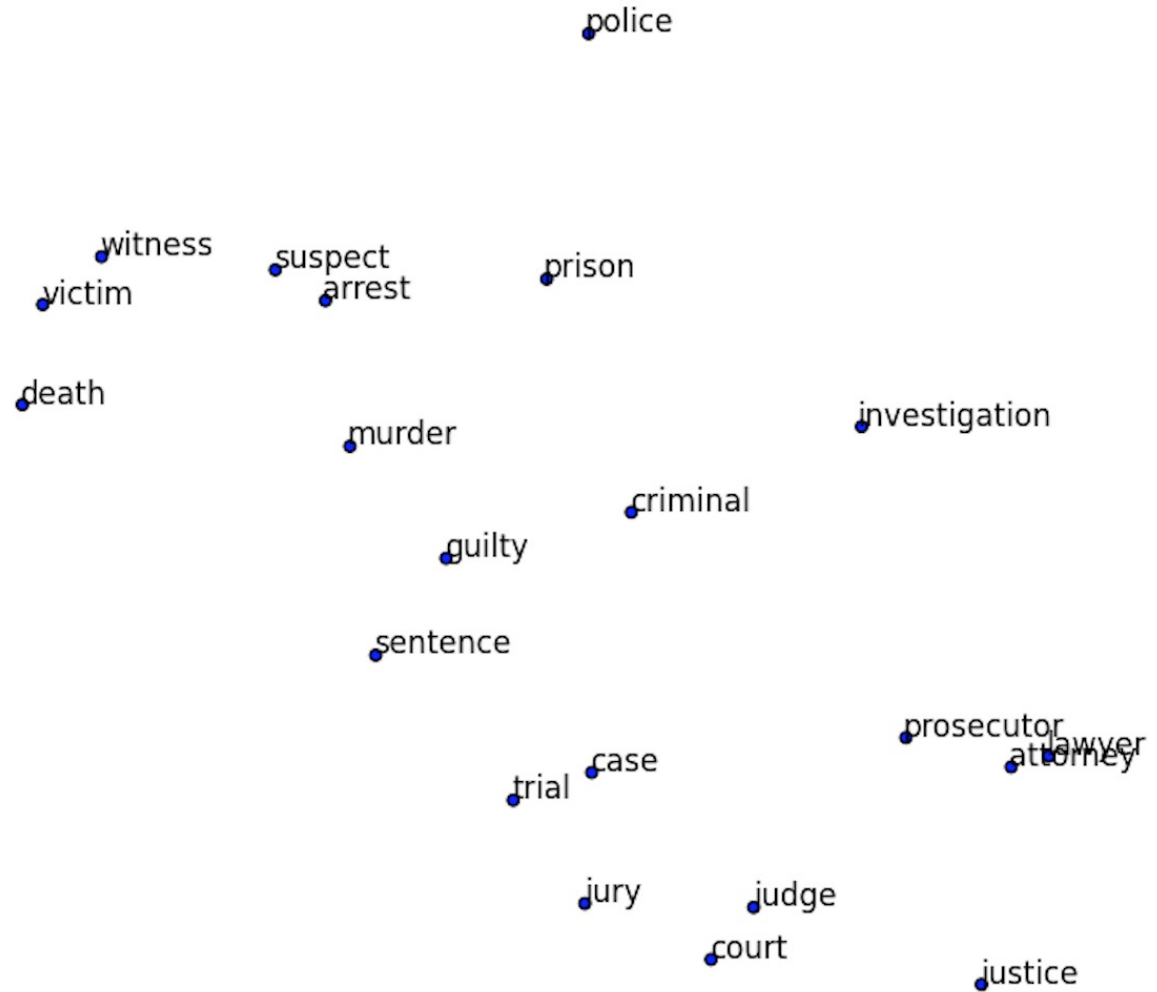


Visualizing Word Embeddings

If we zoom in on a small region of our word map, it's all related words.

Note the similarity of all the words as a whole, but also of the individual neighbors.

“Lawyer” and “attorney” are right next to each other – they have almost identical characteristics!



A brief note on how word embeddings are learned...

KEY IDEA: words are *defined* by the context in which they appear

A **woman** strolls down the street

A **man** strolls down the street

A **child** strolls down the street

A **crocodile** strolls down the street

A **banana** strolls down the street

A **concept** strolls down the street



KEY IDEA: words are *defined* by the context in which they appear

-> if words are always exchangeable, they must have very similar meaning



learn word meaning like an adult:
explicit definitions

<https://www.parenting.com/activities/baby/teach-baby-to-talk/>



learn word meaning like an child:
implicit definitions from context

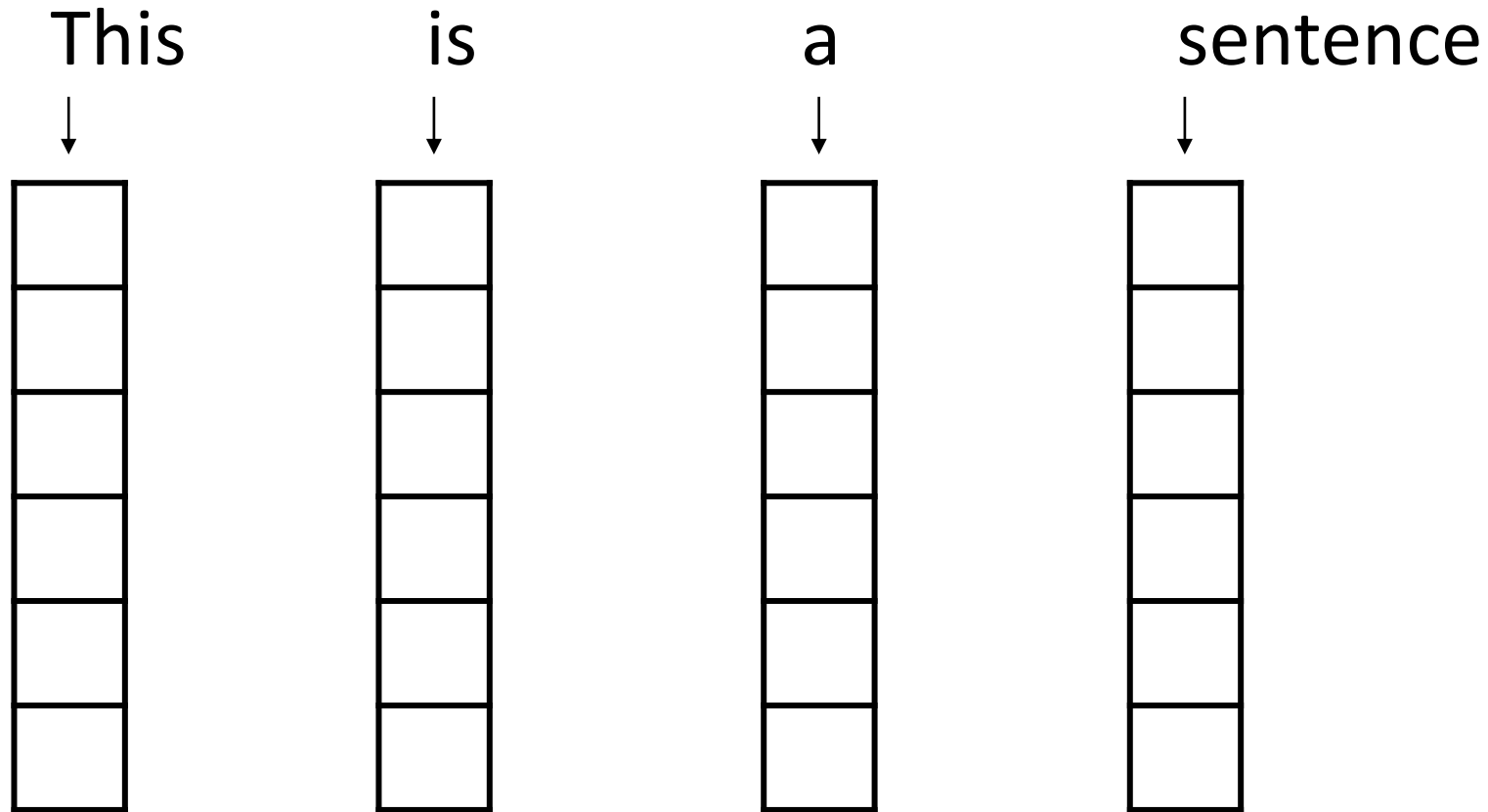


A Very Simple Word Embedding- Based Model

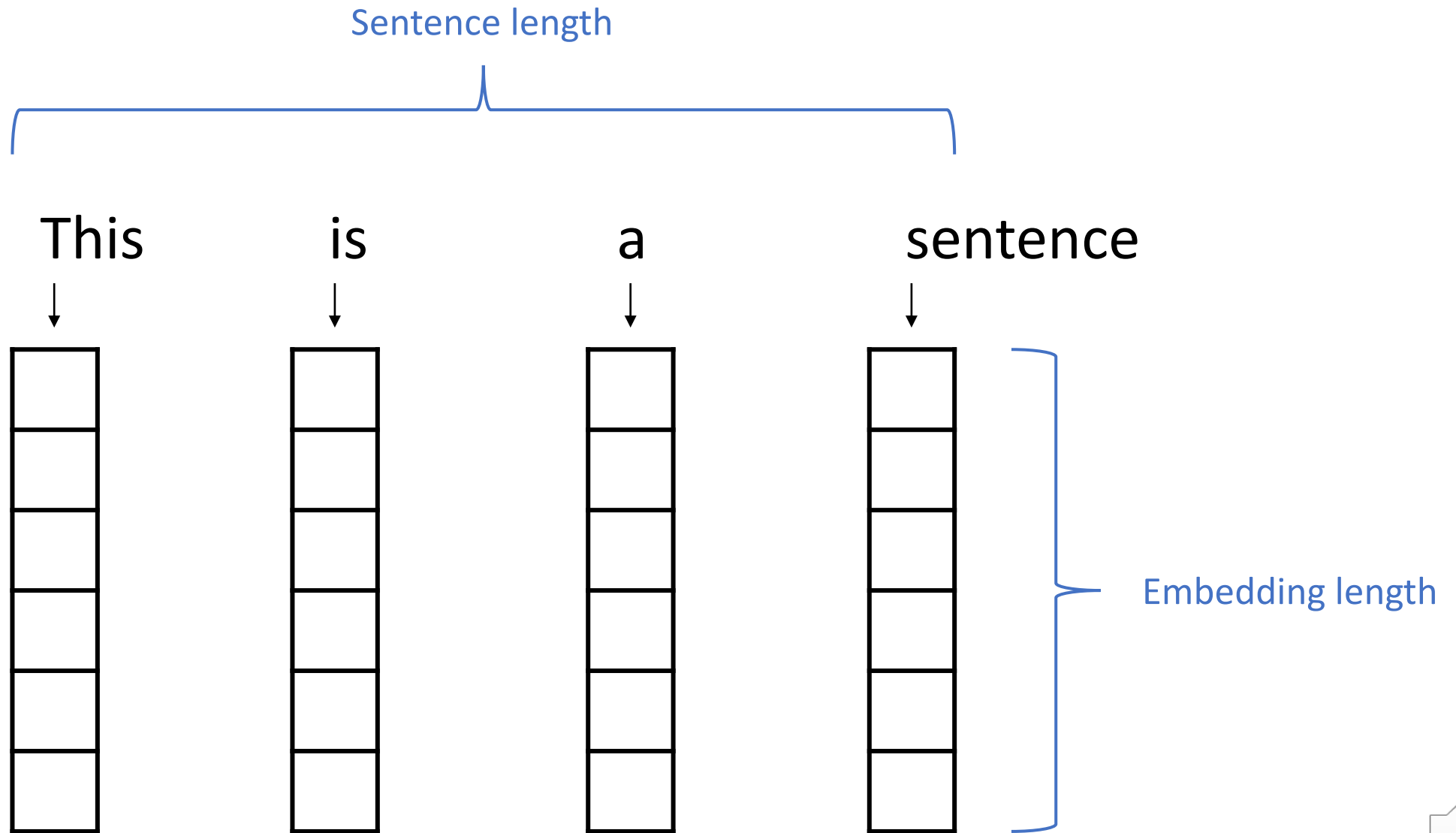


VSWEM Step 1: Convert sentence to vectors

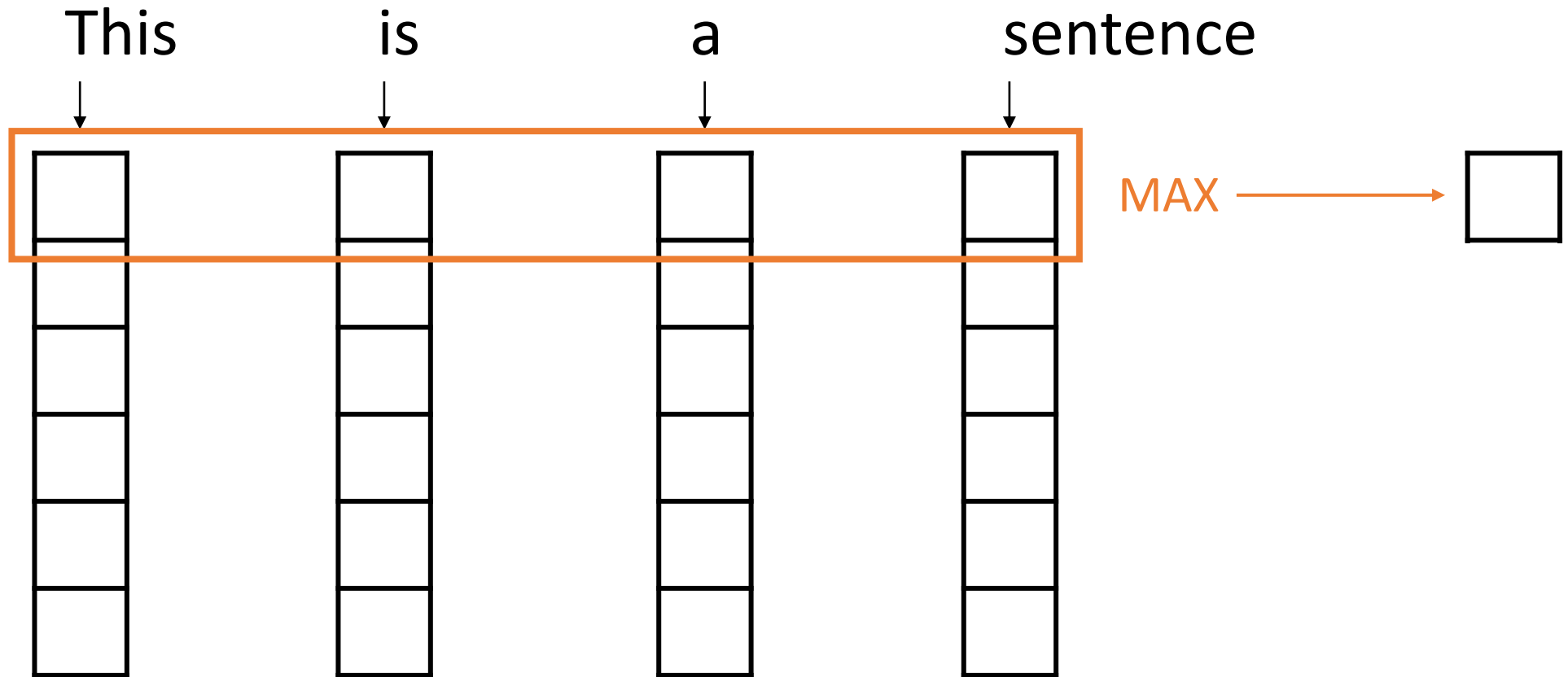
- Look up words individually to obtain their vectors
- Construct a sequence of vectors



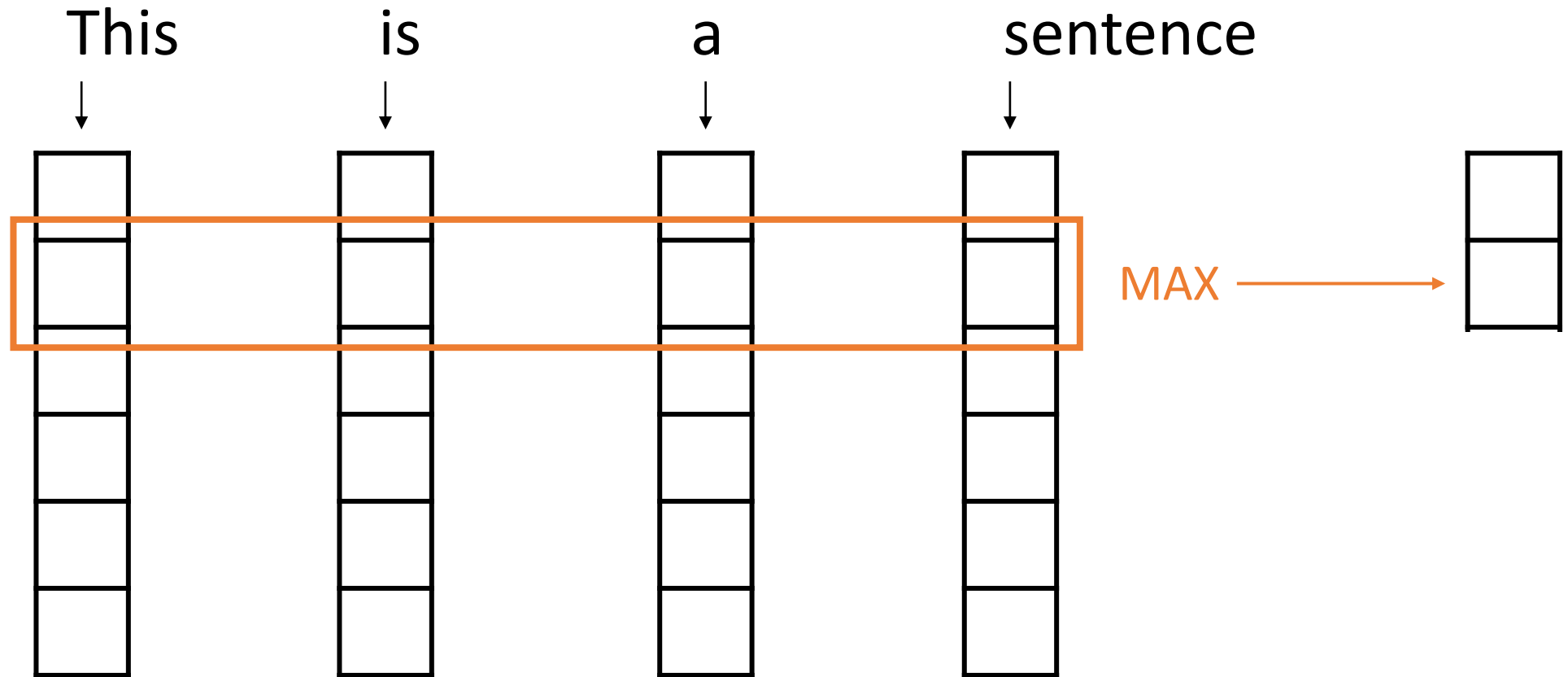
VSWEM Step 1: Convert sentence to vectors



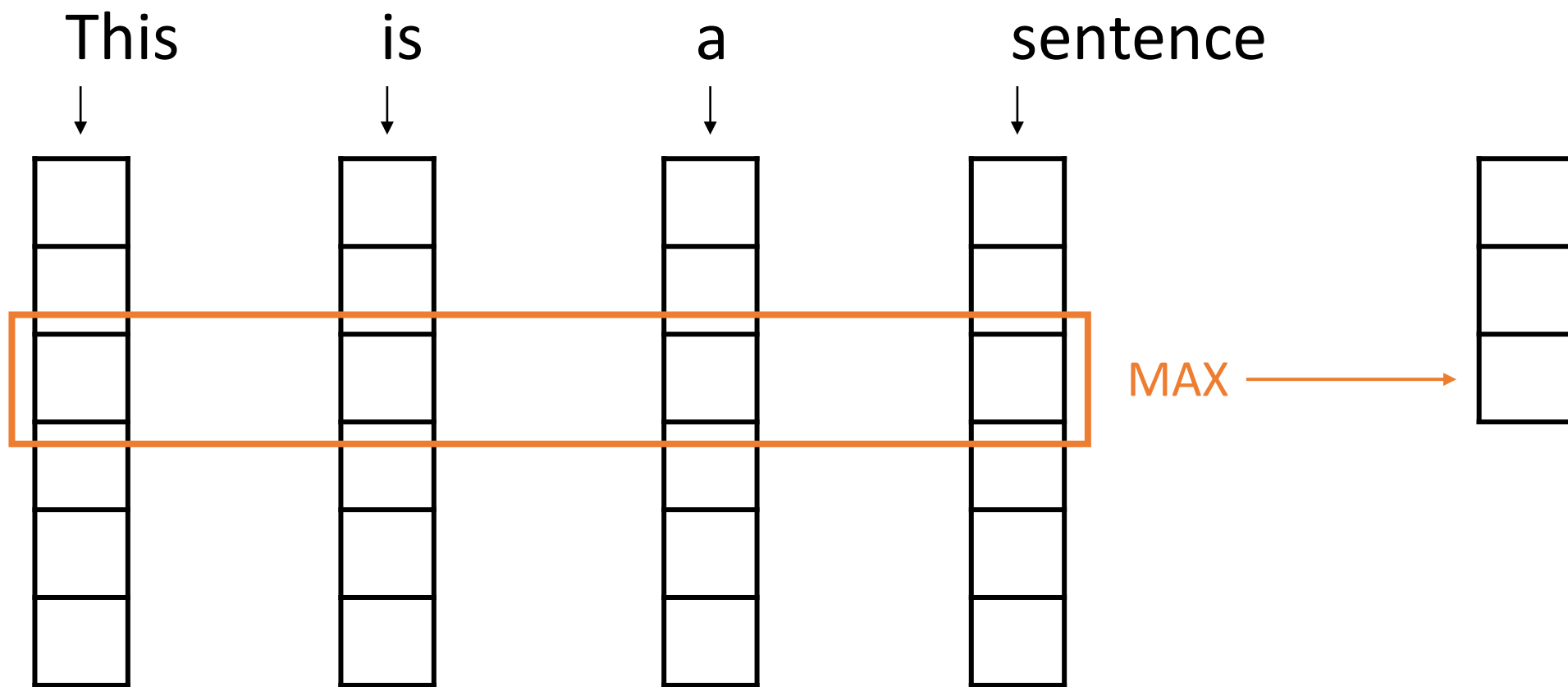
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



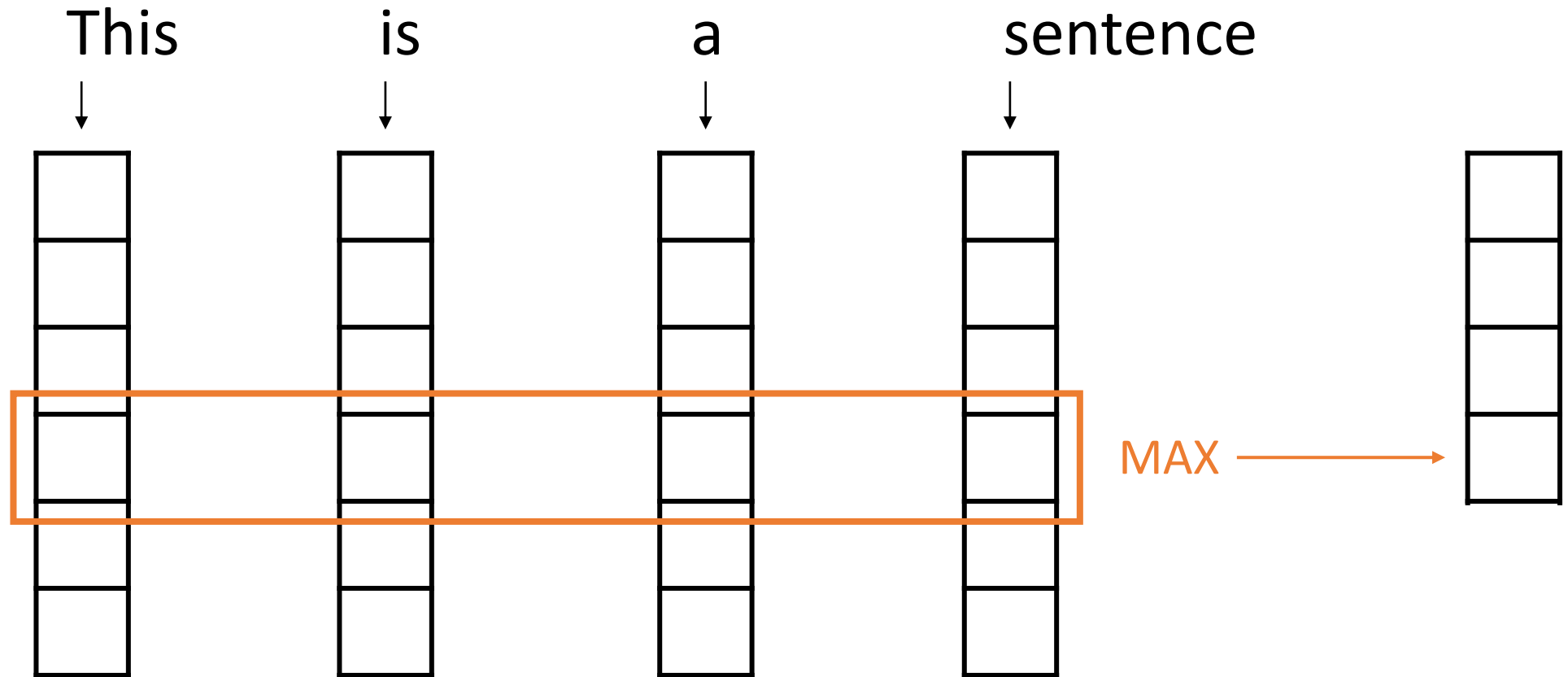
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



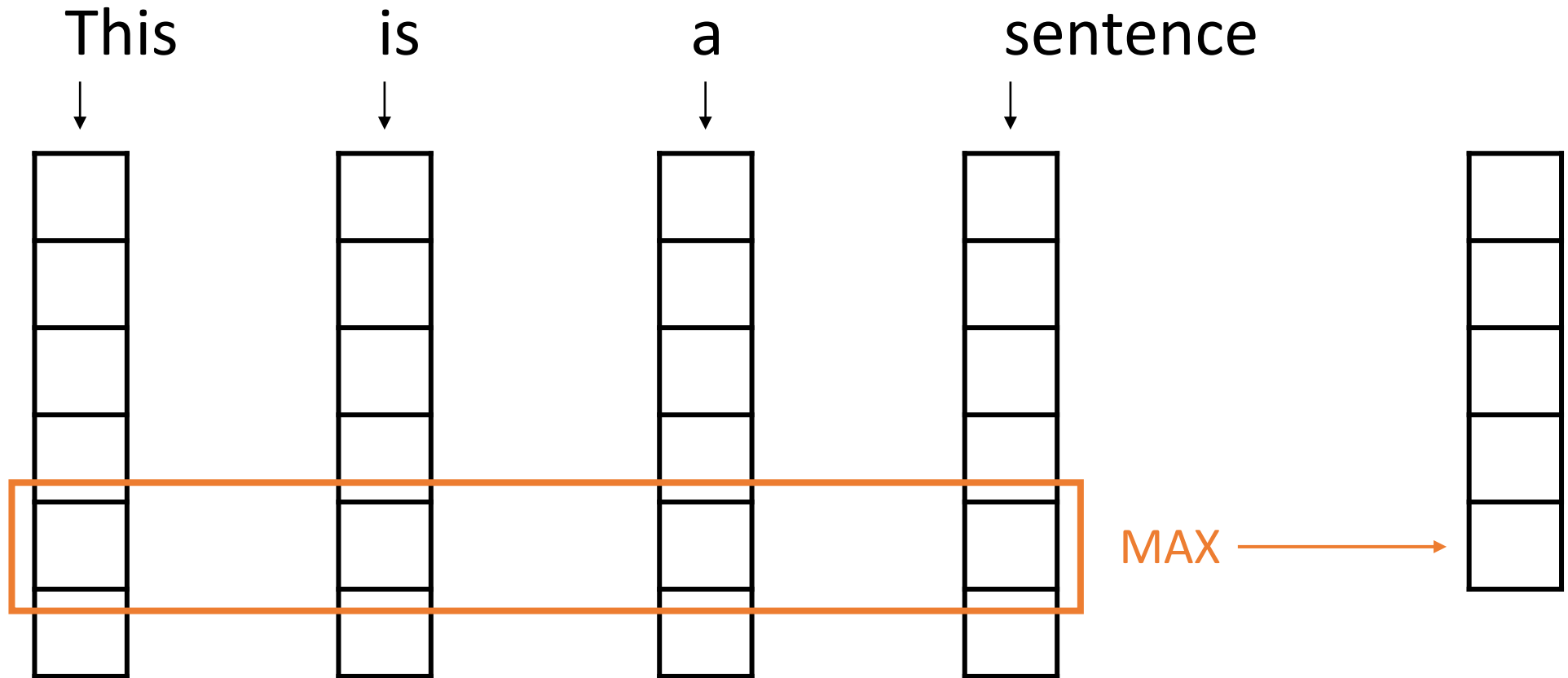
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



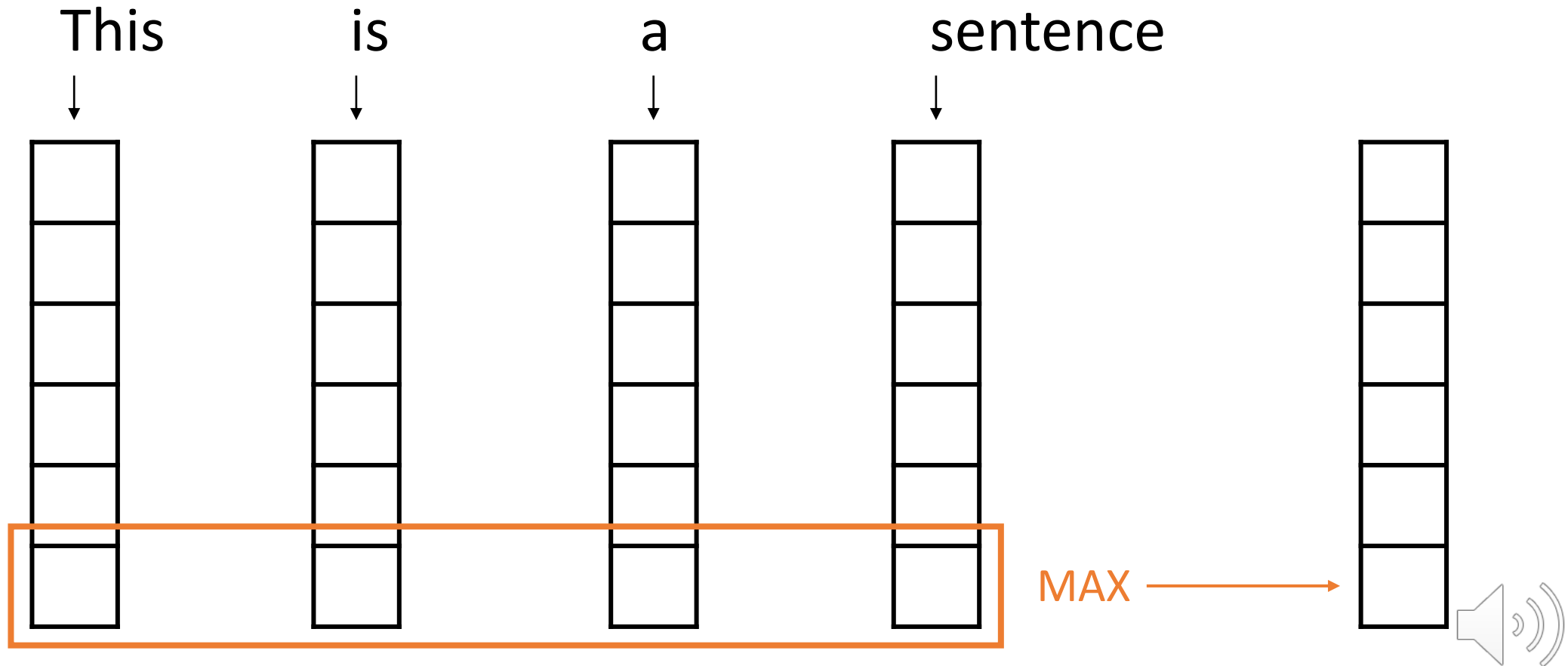
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



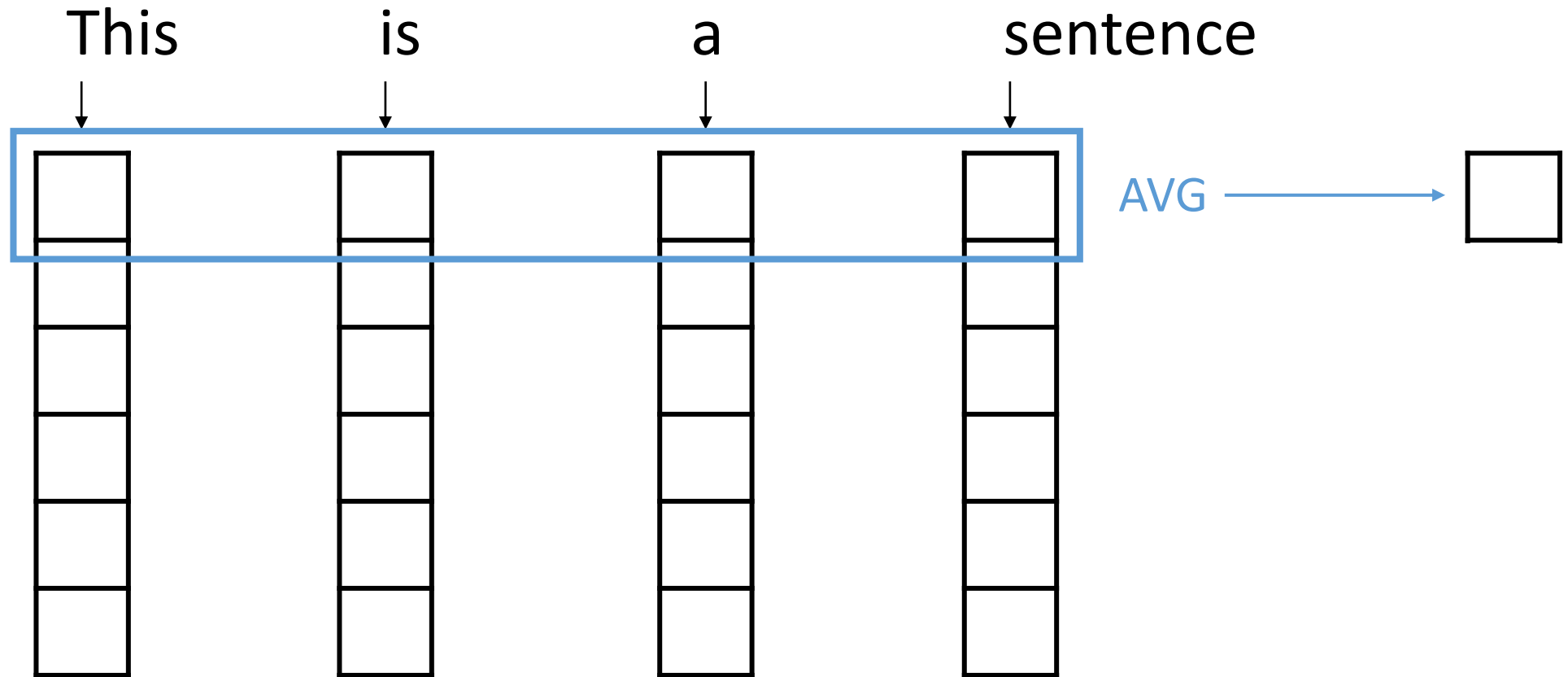
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



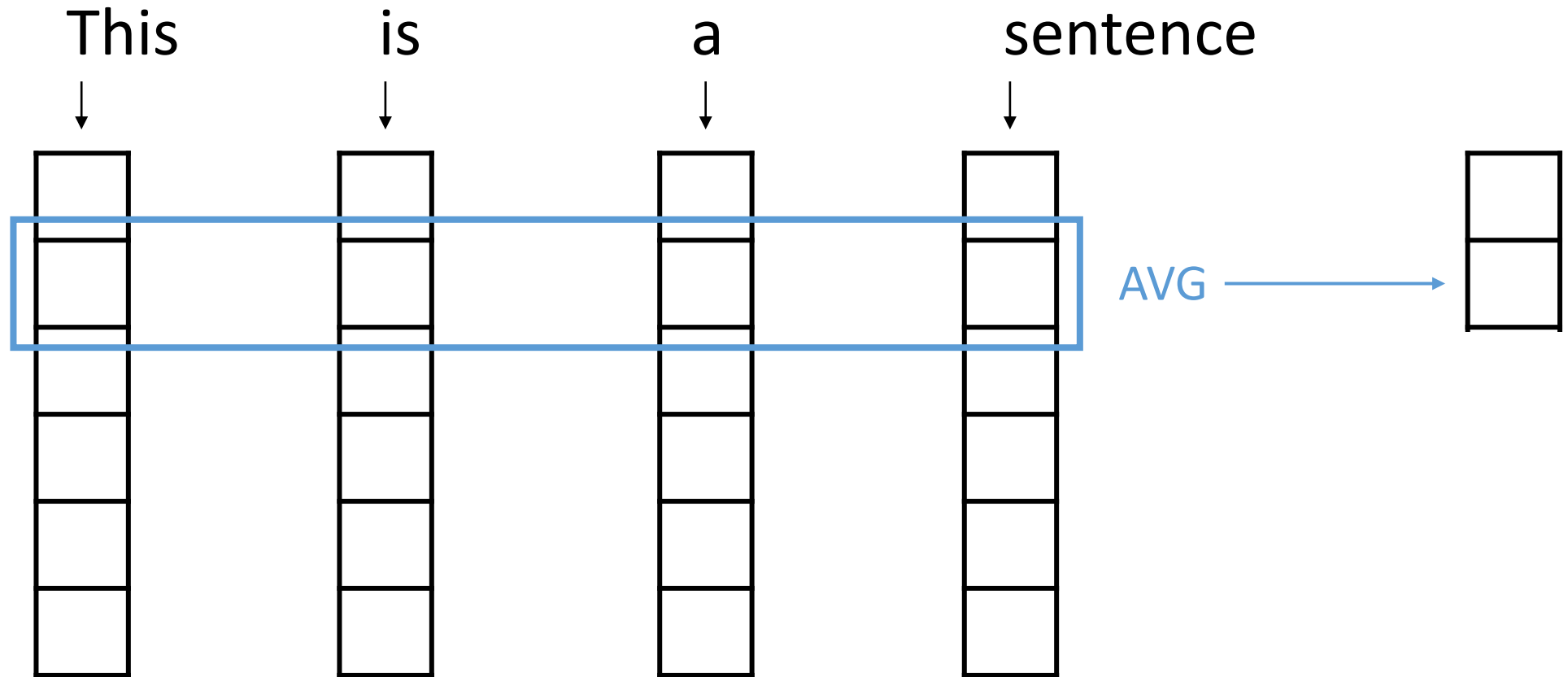
VSWEM Step 2: Take the MAX over the sentence for each embedding dimension



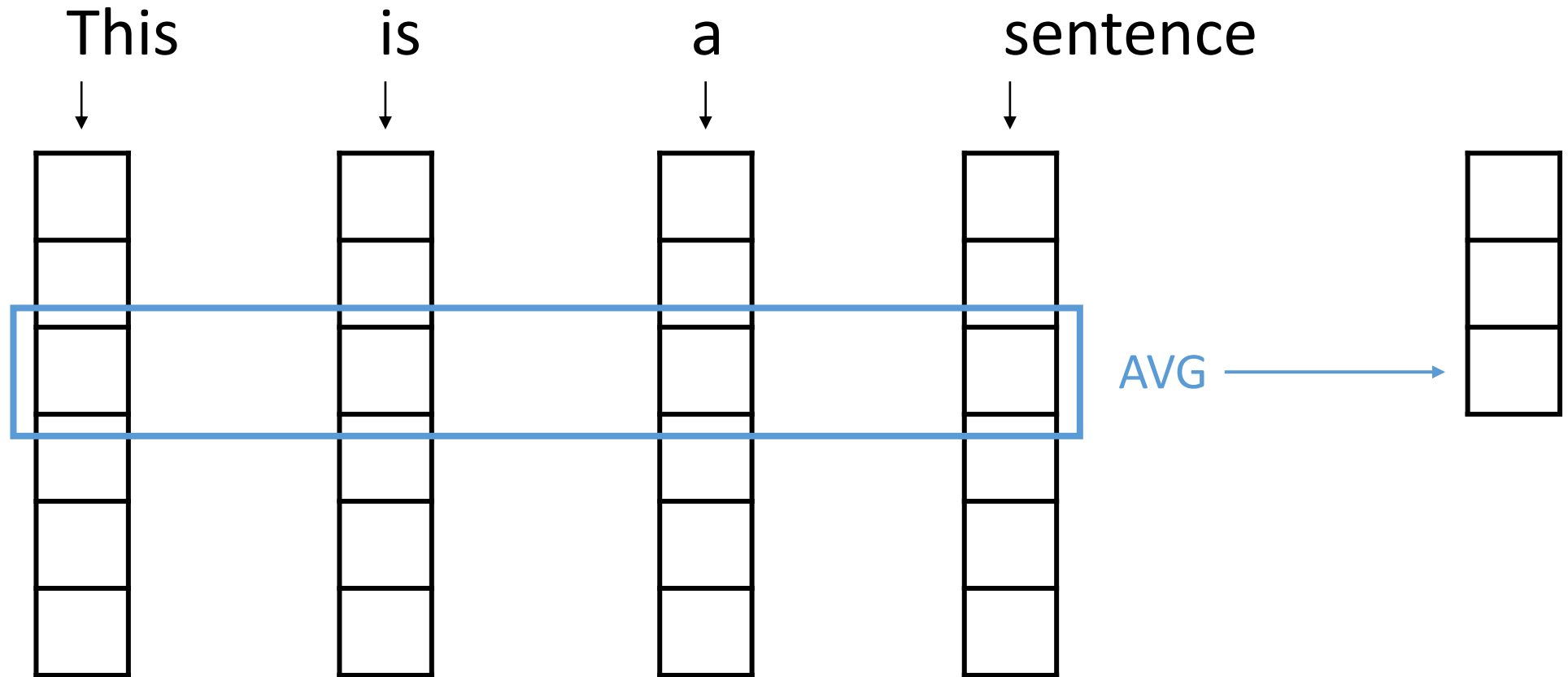
VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



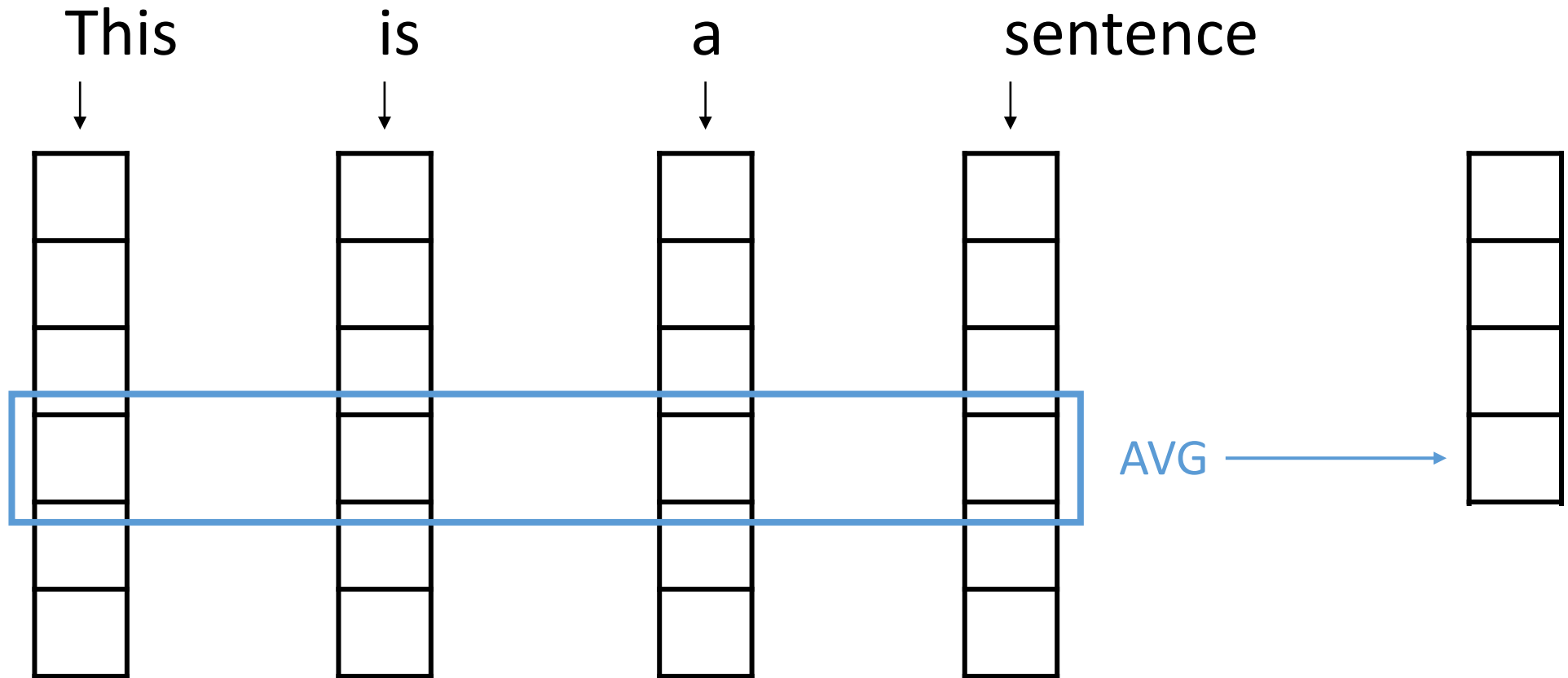
VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



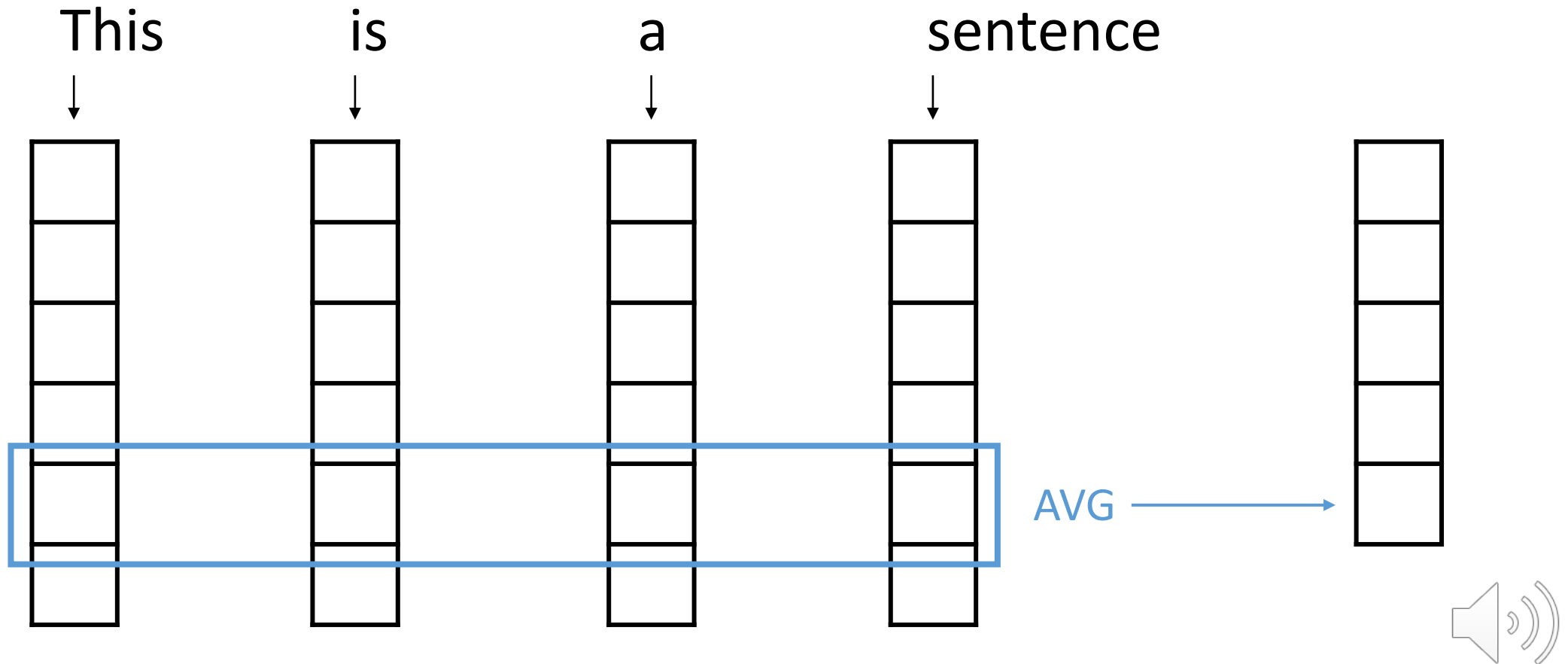
VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



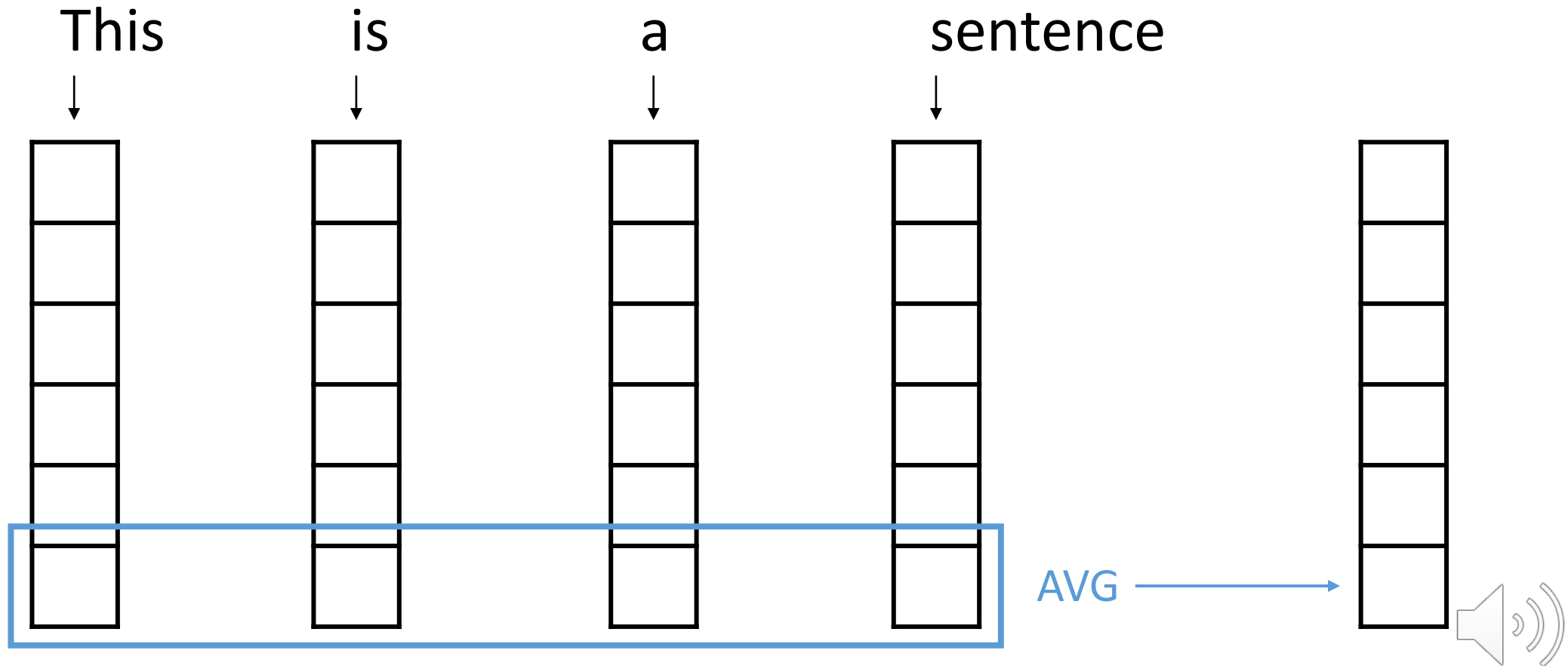
VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



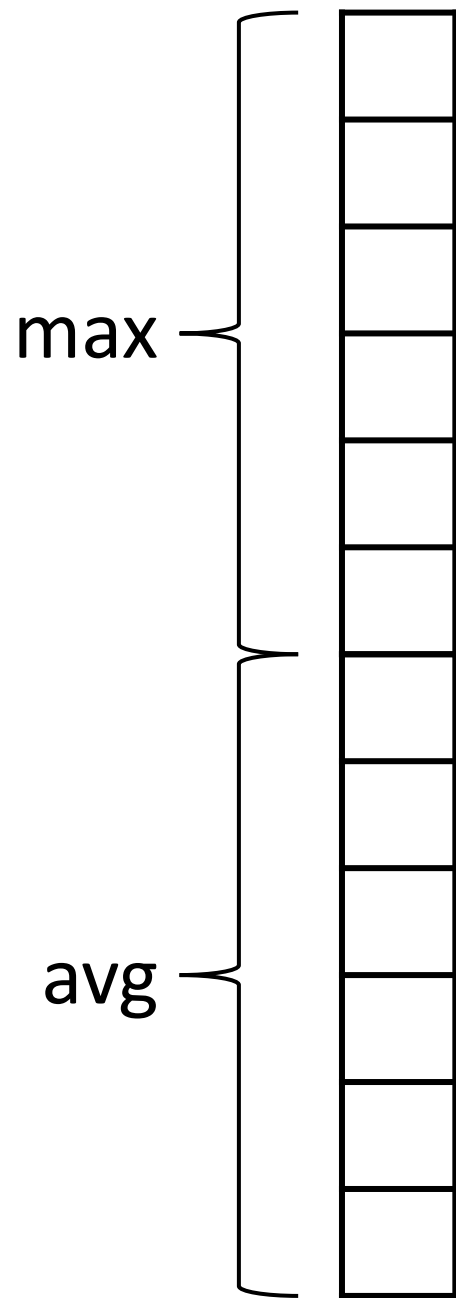
VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



VSWEM Step 3: Take the AVERAGE over the sentence for each embedding dimension



VSWEM Step 4: Concatenate MAX and AVG



Sentence i

This is a sentence

x_i

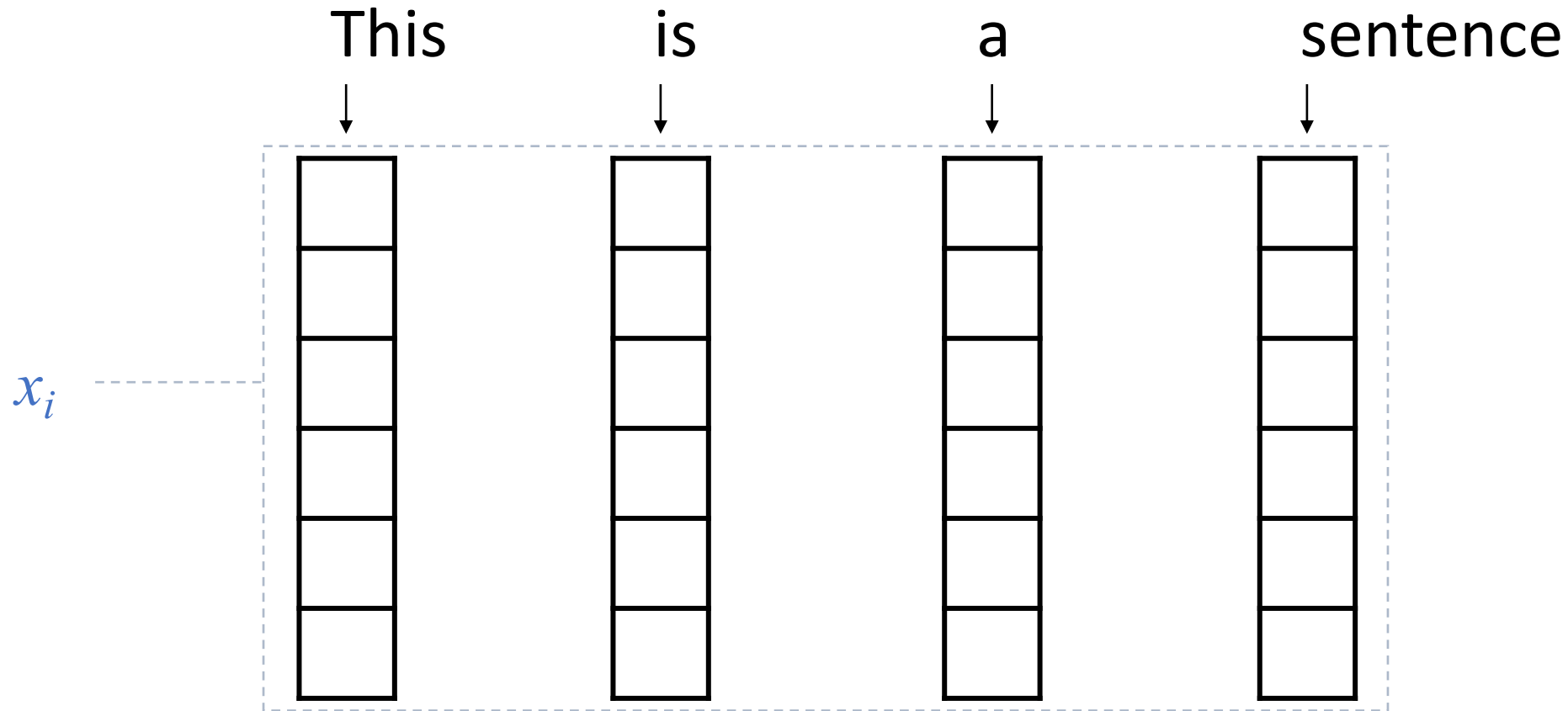
Question:

What information are we
losing when we do this?



Much more complex models...

- Look up words individually to obtain their vectors
- Construct a sequence of vectors
- Then, apply an NN-based model designed for sequences (e.g. transformer, RNN)



State of the art NLP models have *billions* of parameters (up to 1T).

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

“Better Language Models and Their Implications”

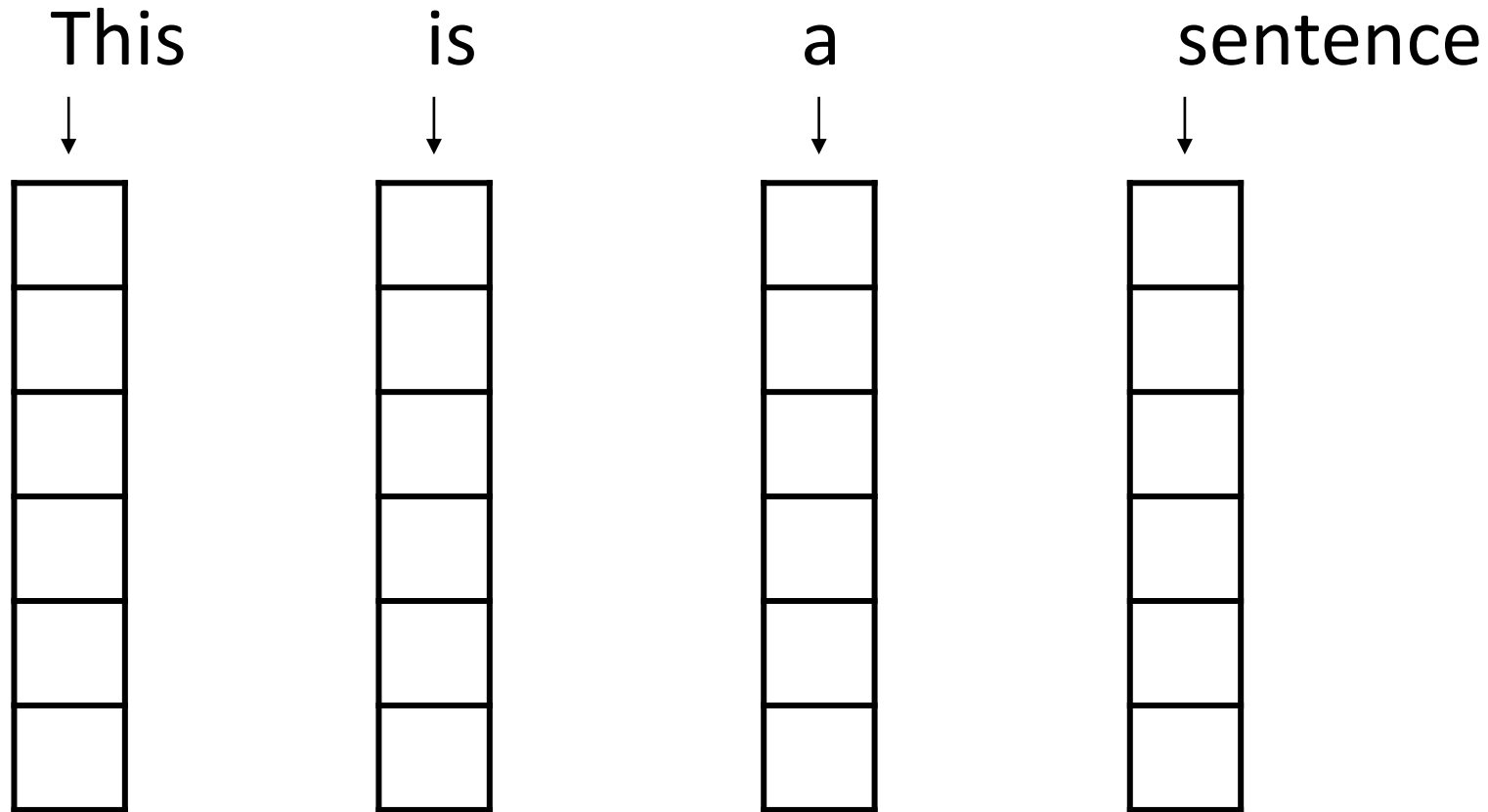
2/14/19

OPENAI

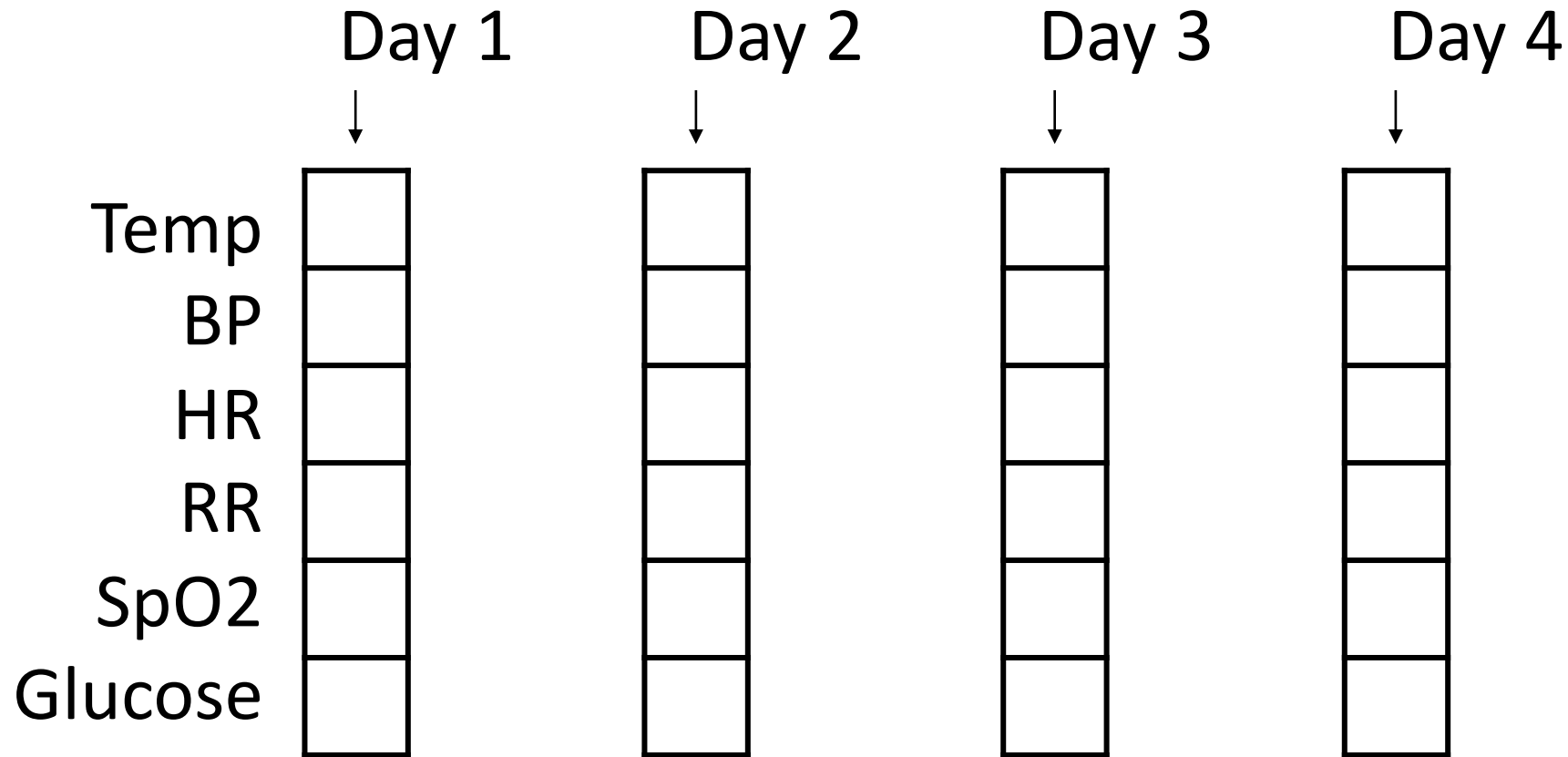


With word vectors, methods for text vs sequences are similar:

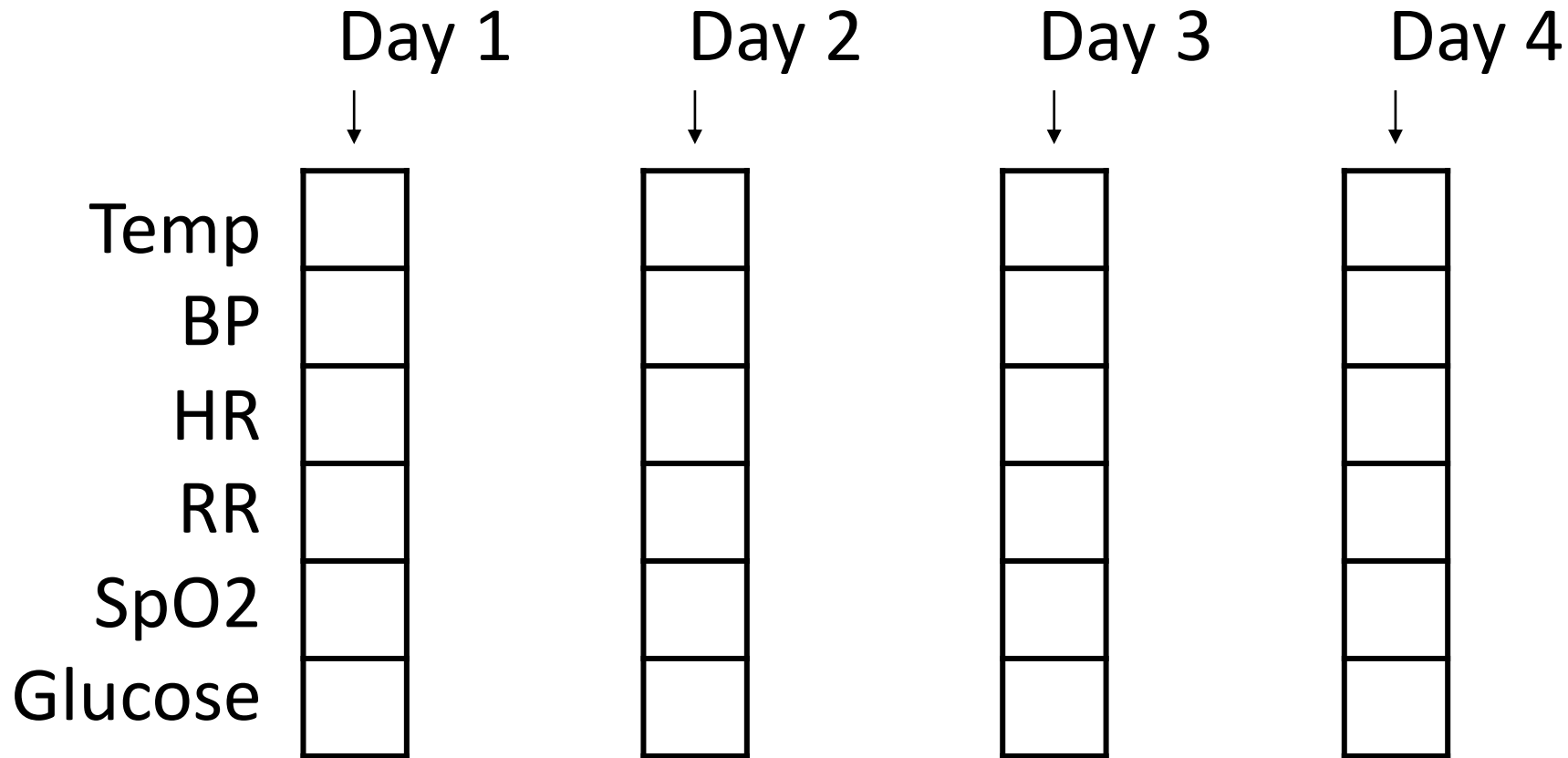
A sequence of word vectors...



With word vectors, methods for text vs sequences are similar:
...now looks just like a sequence of measurements.

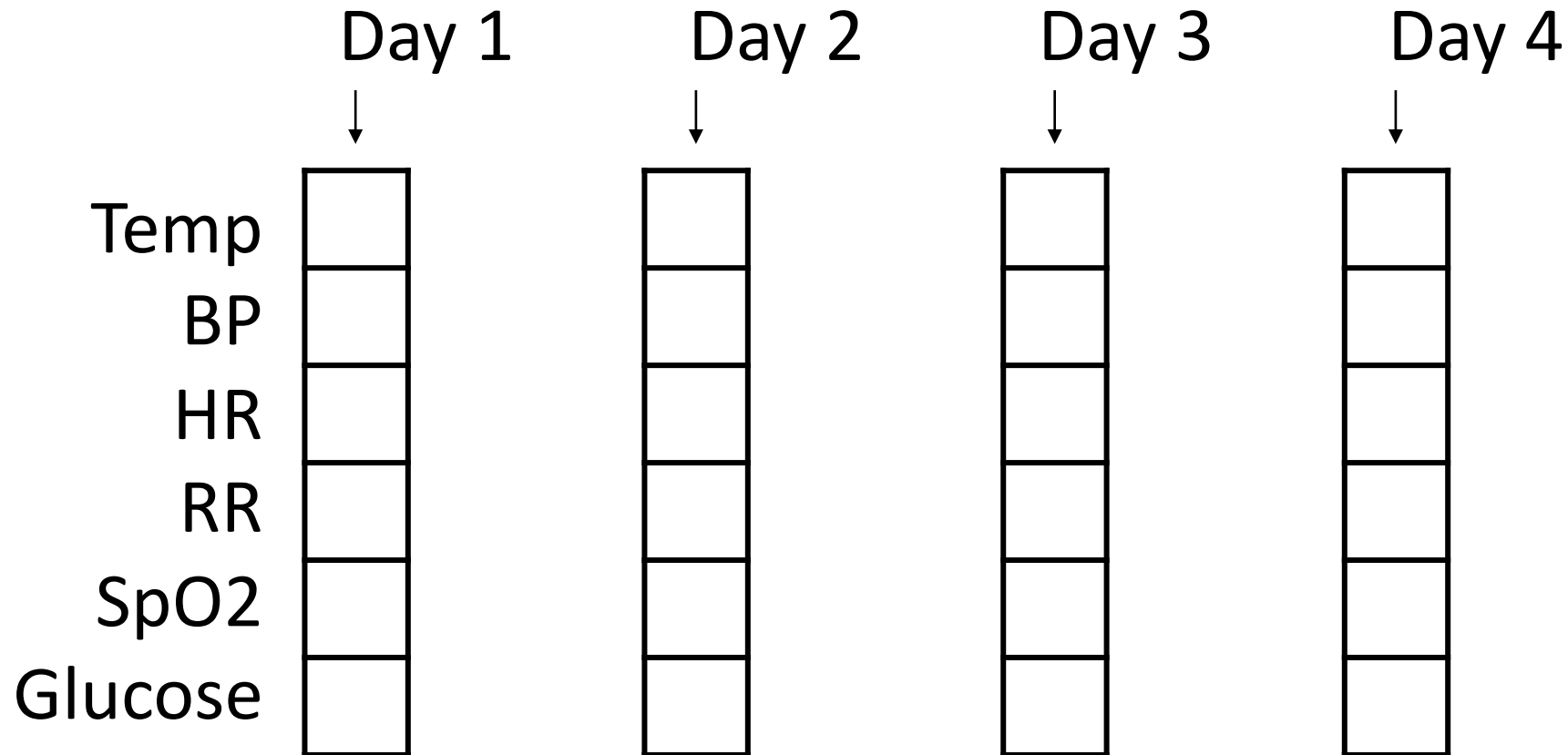


In this case, too, we can get a single numeric vector for our predictive models by taking a max and average (or any other summary statistics we'd like)



But when we do this, we lose information about *order*.

- Next time, we'll talk about ways to overcome this limitation



Conclusions


- Word vectors capture semantic attributes of words, allowing NLP models to make similar predictions for words with similar attributes
- The key idea behind the learning of word vectors is that words are defined by the various contexts in which they appear
- We explored a ‘very simple word embedding based model’, which makes predictions based on a summary of word attributes across an entire sentence of interest
- A sequence of word vectors is very similar to a time series of numeric measurements, so we can use similar models (e.g. RNNs) for the two



Bonus: we can also do this with categorical variables!

- Locations (city/state)
 - Dx and procedure codes
 - Medical concepts
-
- *What attributes could be used to encode the meaning of medical concepts?*

Proceedings — AMIA Joint Summits
on Translational Science


INFORMATICS PROFESSIONALS. LEADING THE WAY.

[AMIA Jt Summits Transl Sci Proc.](#) 2016; 2016: 41–50.
Published online 2016 Jul 20.

PMCID: PMC5001761
PMID: [27570647](#)


Learning Low-Dimensional Representations of Medical Concepts

[Youngduck Choi](#),¹ [Chill Yi-I Chiu](#), MS,¹ and [David Sontag](#), PhD¹

▸ [Author information](#) ▸ [Copyright and License information](#) [Disclaimer](#)

This article has been [cited by](#) other articles in PMC.

Abstract

Go to: 

We show how to learn low-dimensional representations (embeddings) of a wide range of concepts in medicine, including diseases (e.g., ICD9 codes), medications, procedures, and laboratory tests. We expect that these embeddings will be useful across medical informatics for tasks such as cohort selection and patient summarization. These embeddings are learned using a technique called neural language modeling from the natural language processing community. However, rather than learning the embeddings solely from text, we show how to learn the embeddings from claims data, which is widely available both to providers and to payers. We also show that with a simple algorithmic adjustment, it is possible to learn medical concept embeddings in a privacy preserving manner from co-occurrence counts derived from clinical narratives. Finally, we establish a methodological framework, arising from standard medical ontologies such as UMLS, NDF-RT, and CCS, to further investigate the embeddings and precisely characterize their quantitative properties.

