COMP 4446 / 5046
Lecture 5: Inference – Greedy and Search

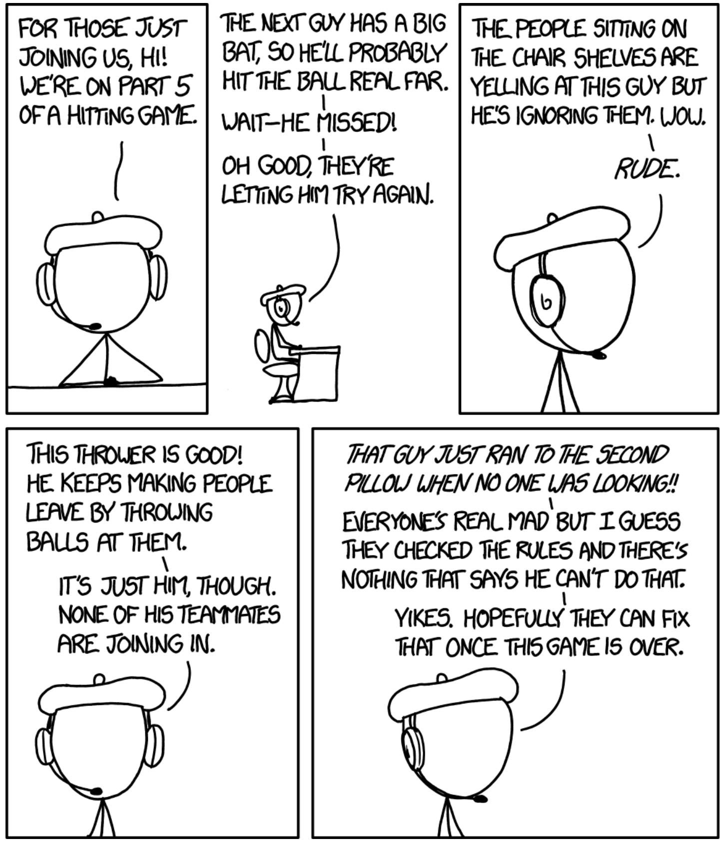
Jonathan K. Kummerfeld

Semester 1, 2024

THE UNIVERSITY OF SYDNEY

menti.com 6165 8383

Play-By-Play



[The thrower started hitting the bats too much, so the king of the game told him to leave and brought out another thrower from thrower jail.]

Source: https://xkcd.com/1593/

ChatGPT - Are you using it in a way that helps you learn?

Remember that the exam will include:
- Content from labs
- Content from assignments
- Programming questions

COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- Contextual Embeddings

Inference
Lab Preview

menti.com 6165 8383

Representations

Representations

Static Embeddings

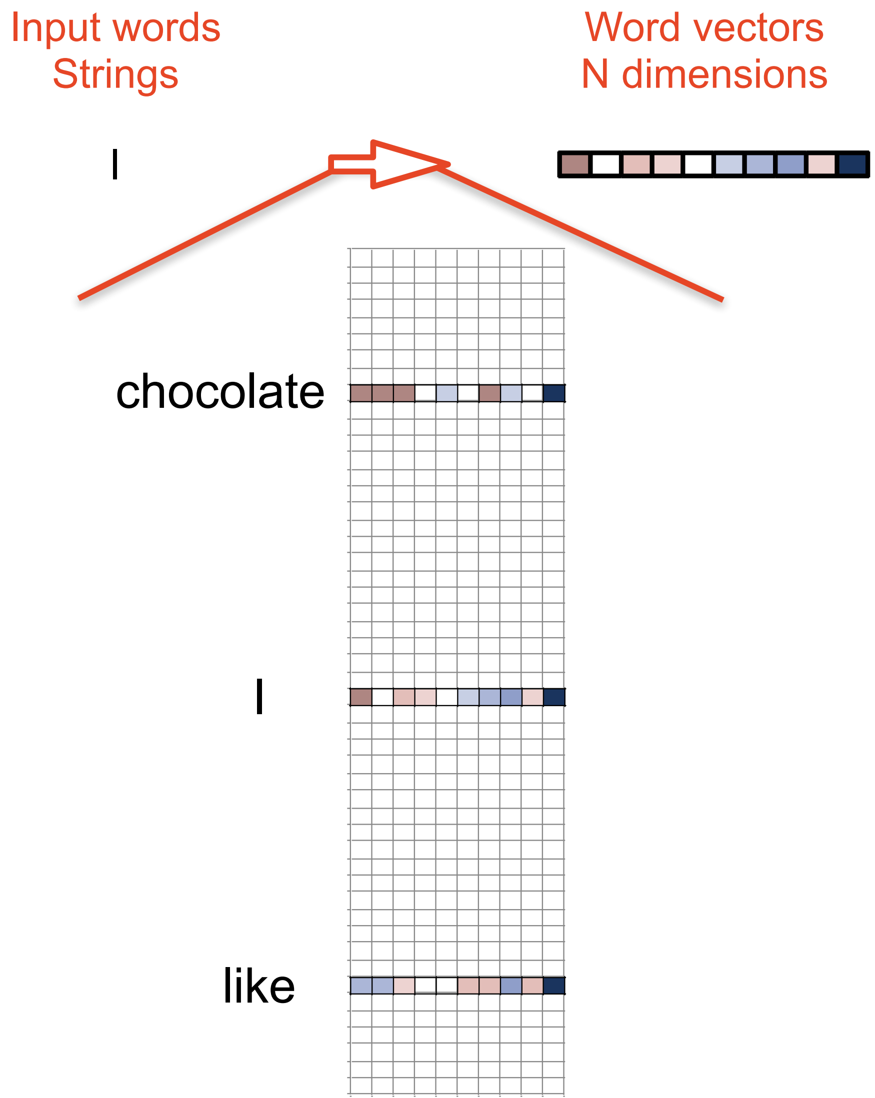
Contextual Embeddings

Inference

Lab Preview

menti.com 6165 8383

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



COMP 4446 / 5046
Lecture 2, 2024

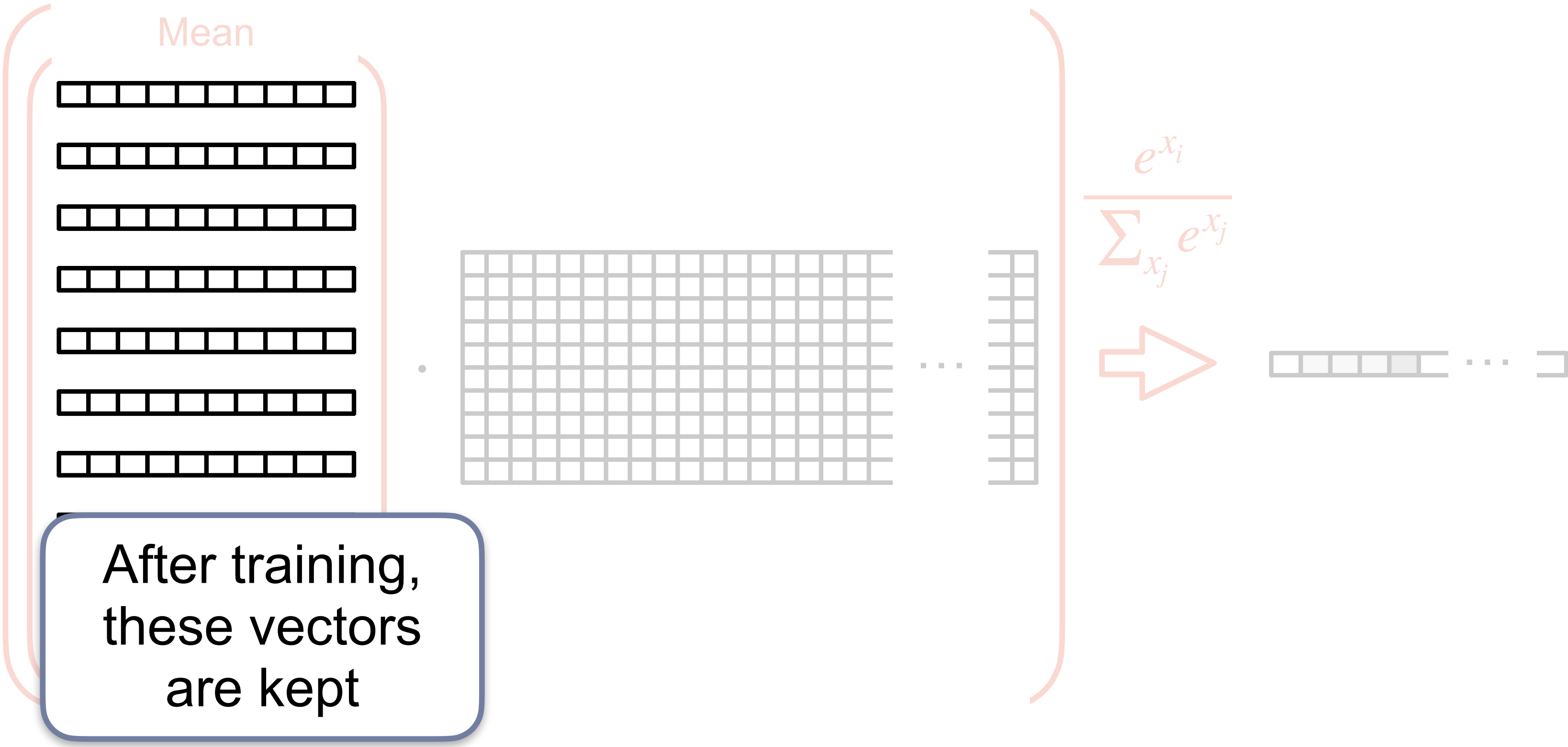
Representations
- **Static Embeddings**
- Contextual Embeddings

Inference
Lab Preview

menti.com 6165 8383

Where does the table come from?

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



Where does the table come from?

word2vec - SkipGram
Input: One word
Output: Set of context words

After training, these vectors are kept

Where does the table come from?

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 2, 2024

Representations
Static Embeddings
Contextual Embeddings
Inference
Lab Preview

menti.com 6165 8383

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

"where"

<wh whe her ere re>	<whe wher here ere>	<wher where here>	<where where>	<where>

COMP 4446 / 5046
Lecture 2, 2024

Representations
Static Embeddings
Contextual Embeddings
Inference
Lab Preview

menti.com 6165 8383

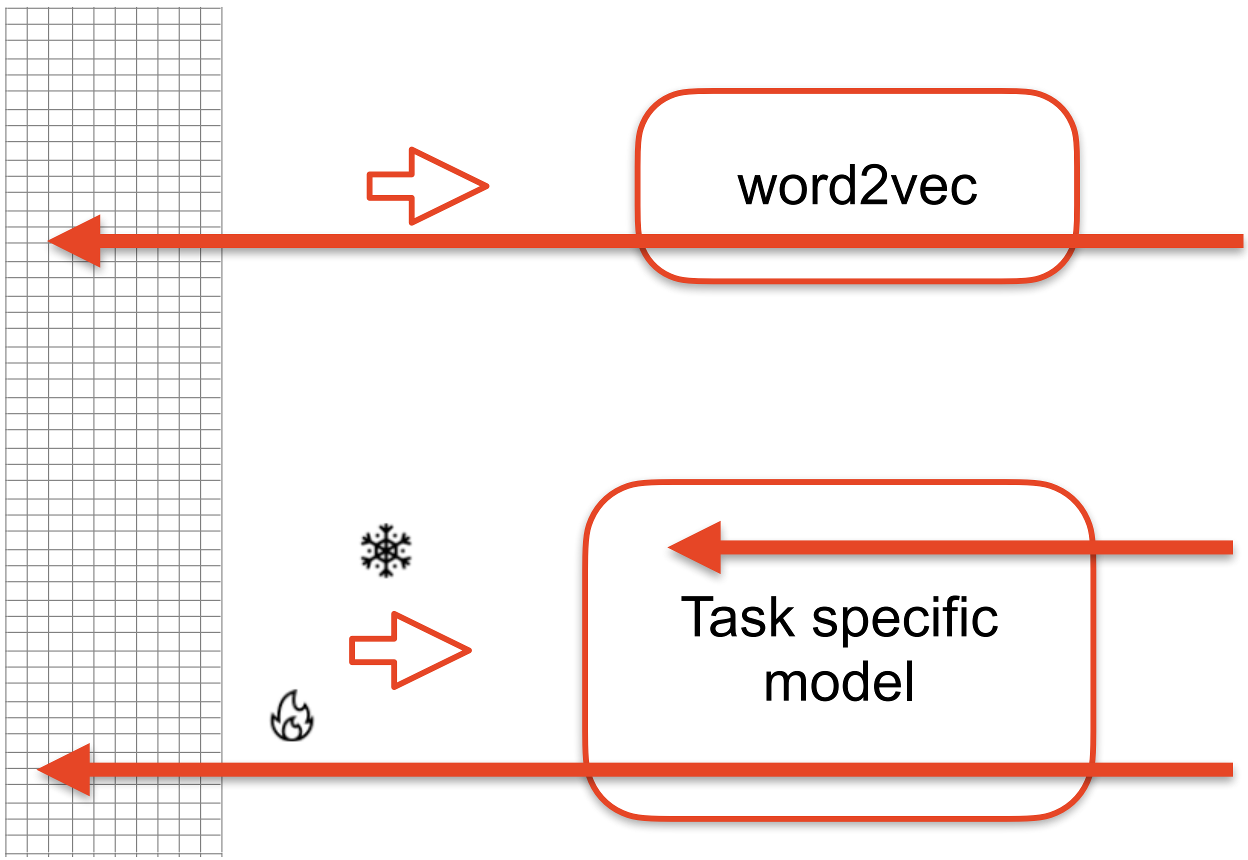
What if my data is not the same as the data used for training?

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

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

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 2, 2024

Representations
Static Embeddings
Contextual Embeddings
Inference
Lab Preview

menti.com 6165 8383

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

Another view - training your own word embeddings you can:
- Randomly initialise
- Use some initial values someone else gives you

COMP 4446 / 5046
Lecture 2, 2024

- Representations
  - **Static Embeddings**
  - Contextual Embeddings
- Inference
- Lab Preview

menti.com 6165 8383

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

	Tied	Embeddings Input		Embeddings Output		Dev PPL Std	Dev PPL Rare
(a)	Tied	Frozen	Random	Frozen	Random	680	1120
	Untied	Frozen	Random	Frozen	Random	680	1120
	Untied	Unfrozen	Random	Frozen	Random	680	431
	Untied	Unfrozen	Pretrained	Frozen	Random	220	372
	Untied	Frozen	Pretrained	Frozen	Random	218	360
(b)	Untied	Frozen	Random	Frozen	Pretrained	121	202
	Untied	Unfrozen	Random	Frozen	Pretrained	95.0	170
	Untied	Unfrozen	Pretrained	Frozen	Pretrained	91.3	147
	Tied	Frozen	Pretrained	Frozen	Pretrained	90.7	136
	Untied	Frozen	Pretrained	Frozen	Pretrained	90.7	136
(c)	Untied	Frozen	Random	Unfrozen	Random	82.2	143
	Untied	Frozen	Random	Unfrozen	Pretrained	81.4	142
(d)	Untied	Unfrozen	Random	Unfrozen	Random	65.3	120
	Untied	Unfrozen	Random	Unfrozen	Pretrained	64.1	113
	Untied	Unfrozen	Pretrained	Unfrozen	Random	62.5	105
	Untied	Unfrozen	Pretrained	Unfrozen	Pretrained	61.7	98.5
	Untied	Frozen	Pretrained	Unfrozen	Pretrained	61.6	**97.1**
	Tied	Unfrozen	Random	Unfrozen	Random	61.3	112
	Untied	Frozen	Pretrained	Unfrozen	Random	61.1	98.1
	Tied	Unfrozen	Pretrained	Unfrozen	Pretrained	**59.8**	98.7

= Tied parameters = Untied parameters
= Frozen in training = Unfrozen in training
= Random init. = Pretrained init.

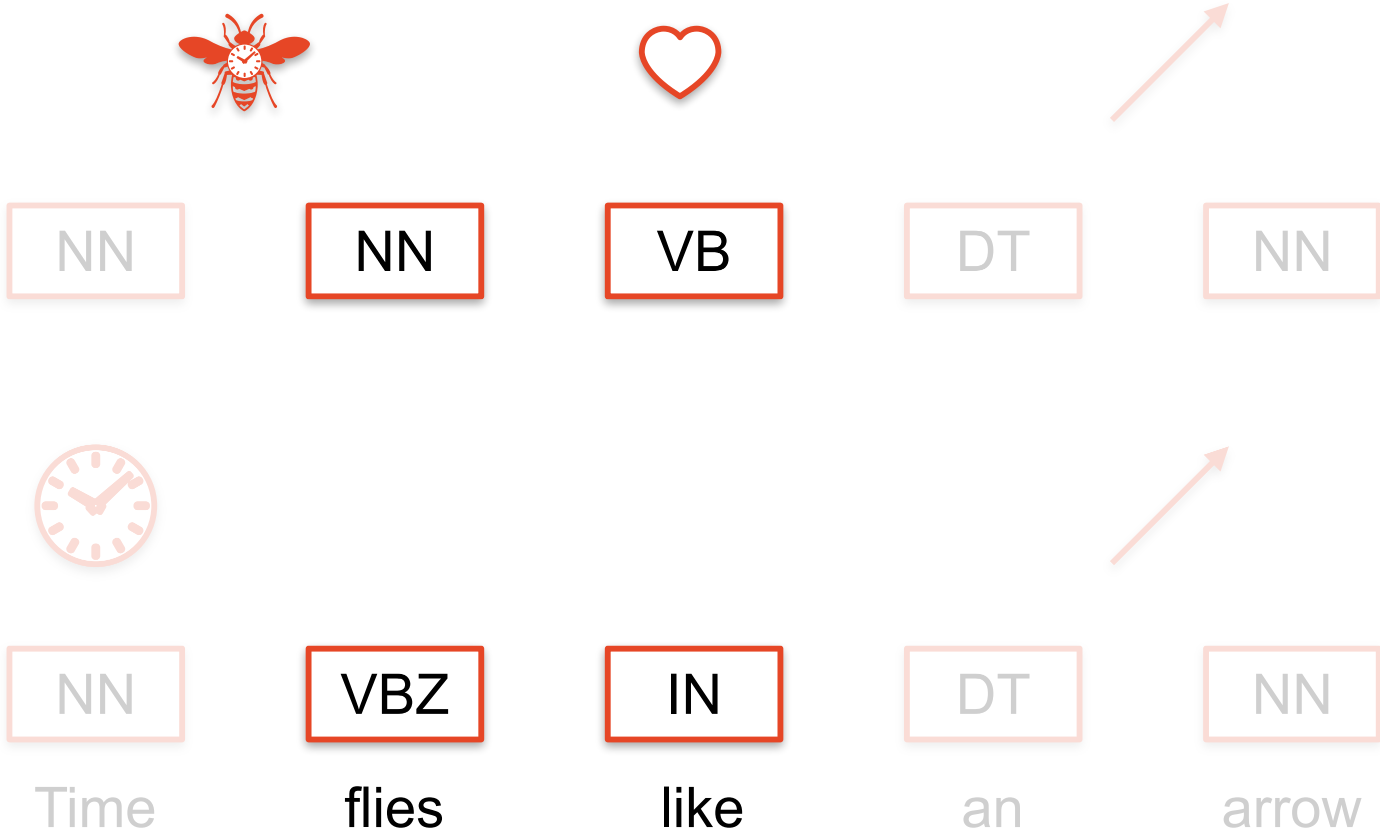
Allowing the embeddings to change *may* be unwise.

Why? Only some get changed, so you get inconsistency.

Welch, Mihalcea, Kummerfeld (EMNLP 2020)

menti.com 6165 8383

What about word senses?



COMP 4446 / 5046
Lecture 2, 2024

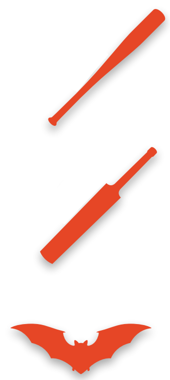
Representations
- Static Embeddings
- **Contextual Embeddings**

Inference
Lab Preview

menti.com 6165 8383

What about word senses?

bat



bate

murcielago

Which of those flights serve breakfast?
Does Air France serve Philadelphia?
?Does Air France serve breakfast and Philadelphia?

Jurafsky and Martin, Appendix G

menti.com 6165 8383

What about word senses?

Dictionary

right adj. located nearer the right hand esp. being on the right when facing the same direction as the observer.

left adj. located nearer to this side of the body than the right.

red n. the color of blood or a ruby.

blood n. the red liquid that circulates in the heart, arteries and veins of animals.

Jurafsky and Martin, Appendix G

COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- **Contextual Embeddings**

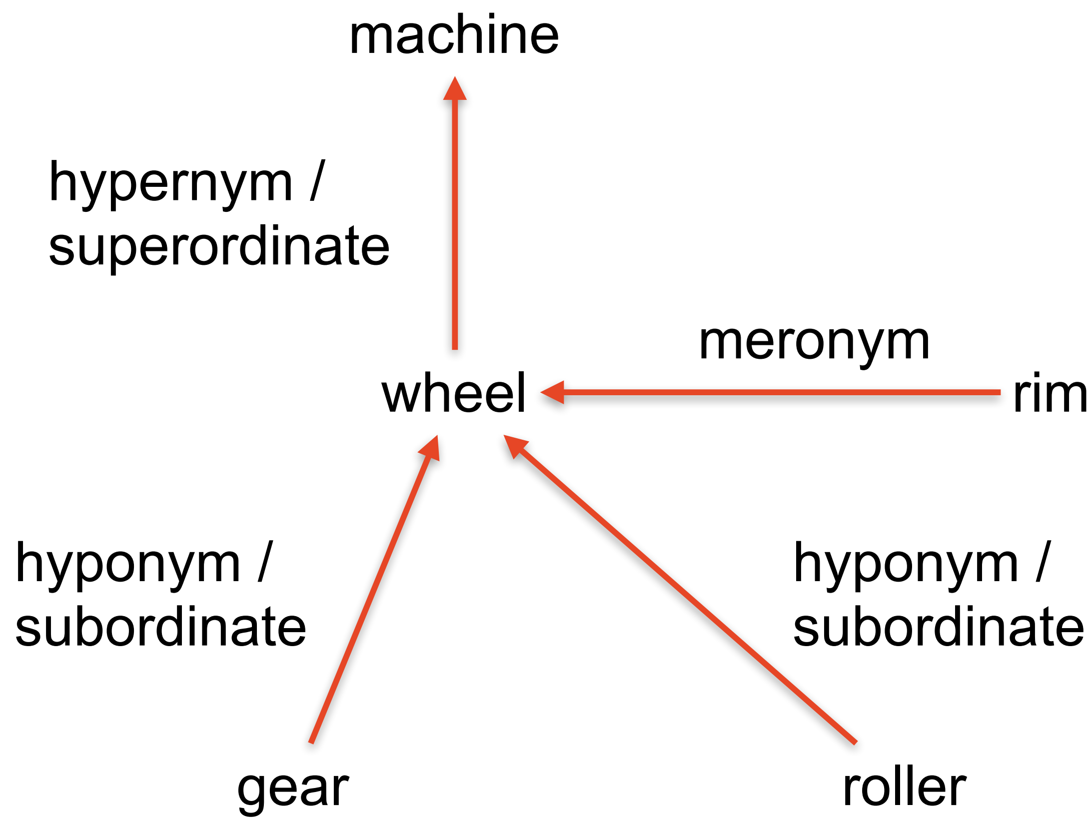
Inference
Lab Preview

menti.com 6165 8383

What about word senses?

WordNet

a simple machine consisting of a circular frame …



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

What about word senses?

WordNet

Synonym set for ‘run’

scat, run, scarper, turn tail, lam, run away, hightail it, bunk, head for the hills, take to the woods, escape, fly the coop, break away

What about word senses?

WordNet

English:
- 117,798 nouns
- 11,529 verbs
- 22,479 adjectives
- 4,481 adverbs.

Also in 200+ other languages! But… smaller

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

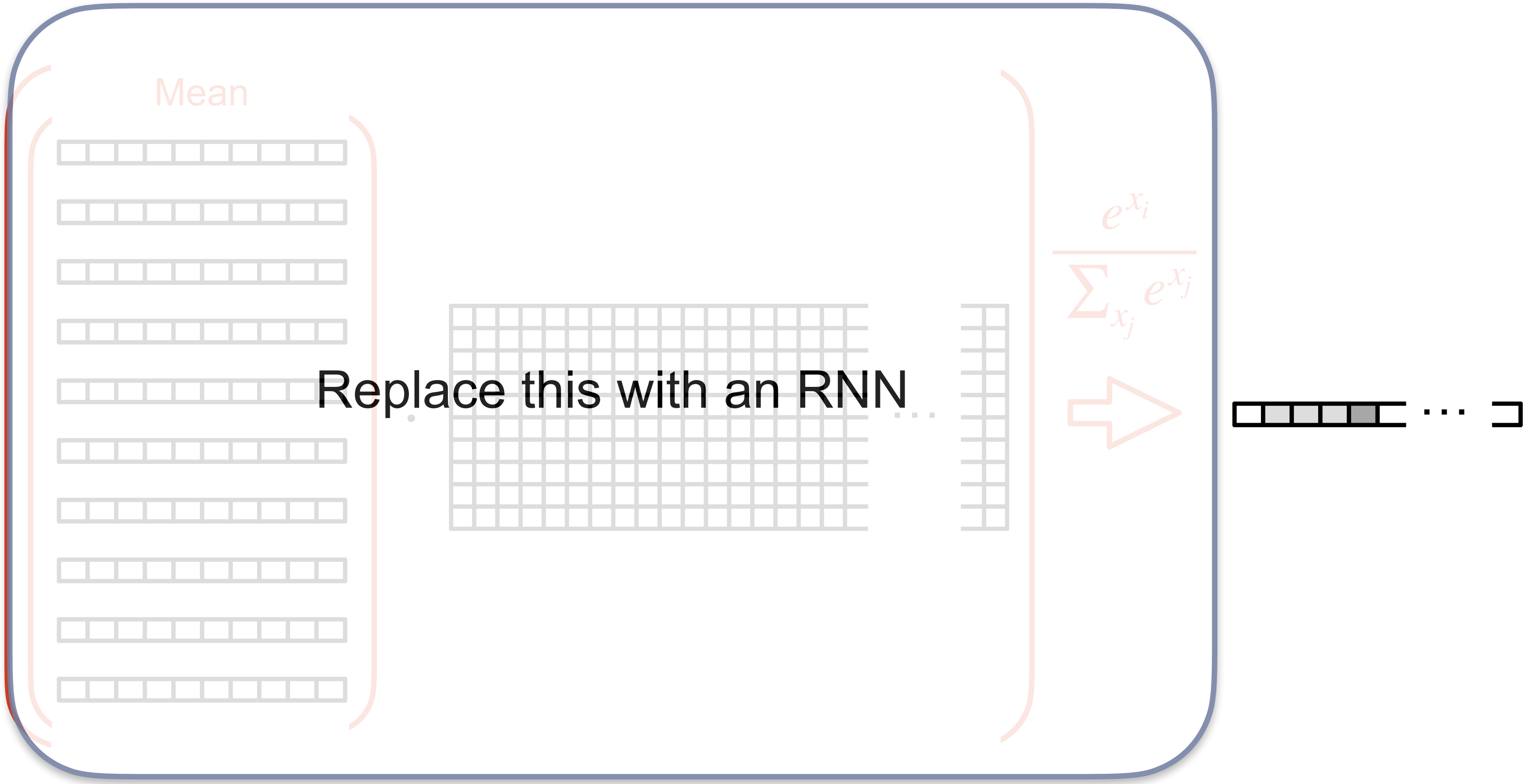
Challenge: data

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

menti.com 6165 8383

Train a model with **contextual** representations

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



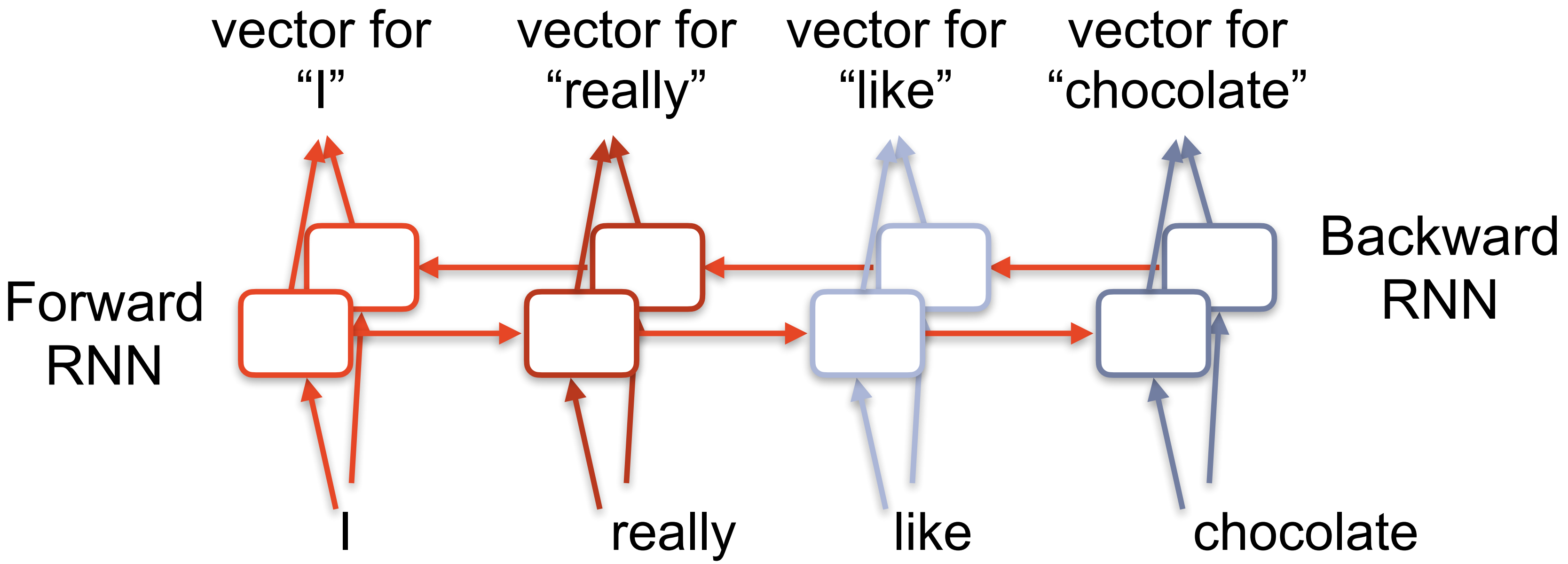
COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- **Contextual Embeddings**

Inference
Lab Preview

menti.com 6165 8383

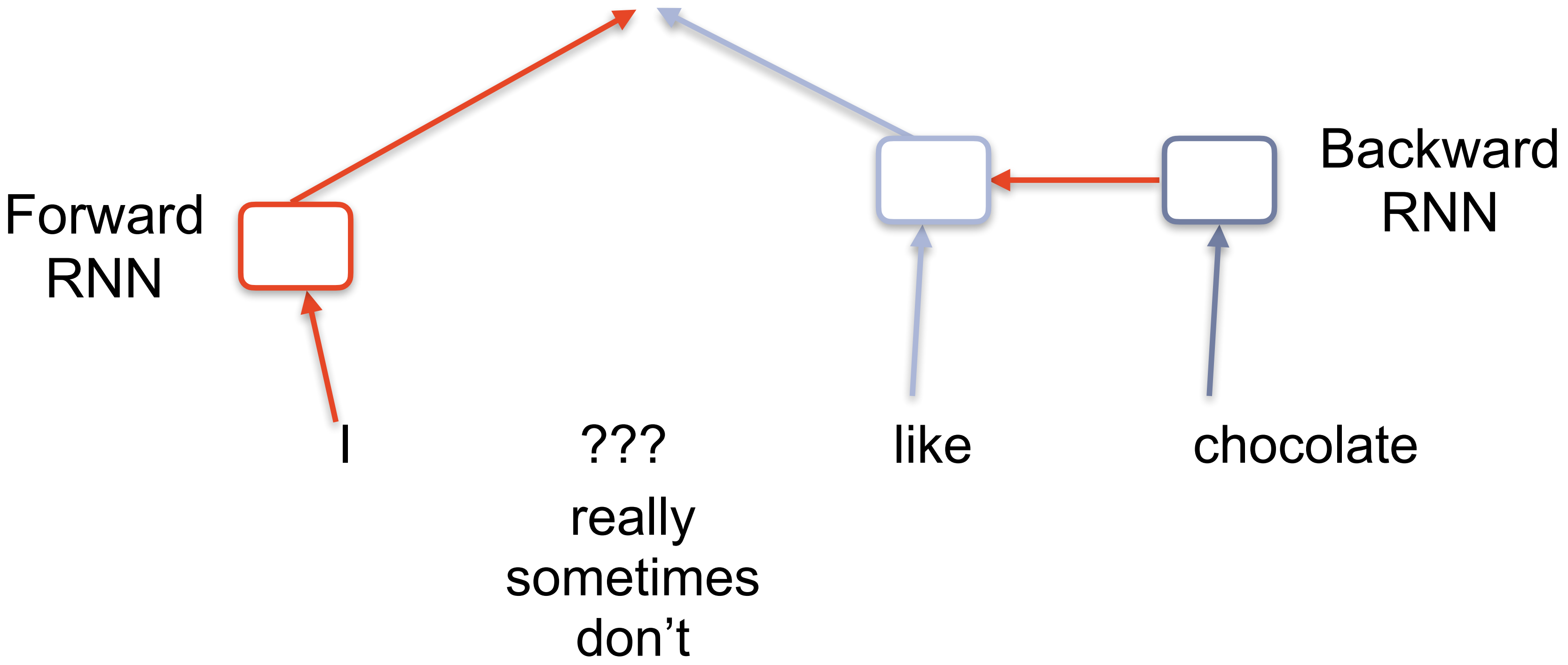
From **contextual** representations using an RNN



menti.com 6165 8383

How do we train the model?

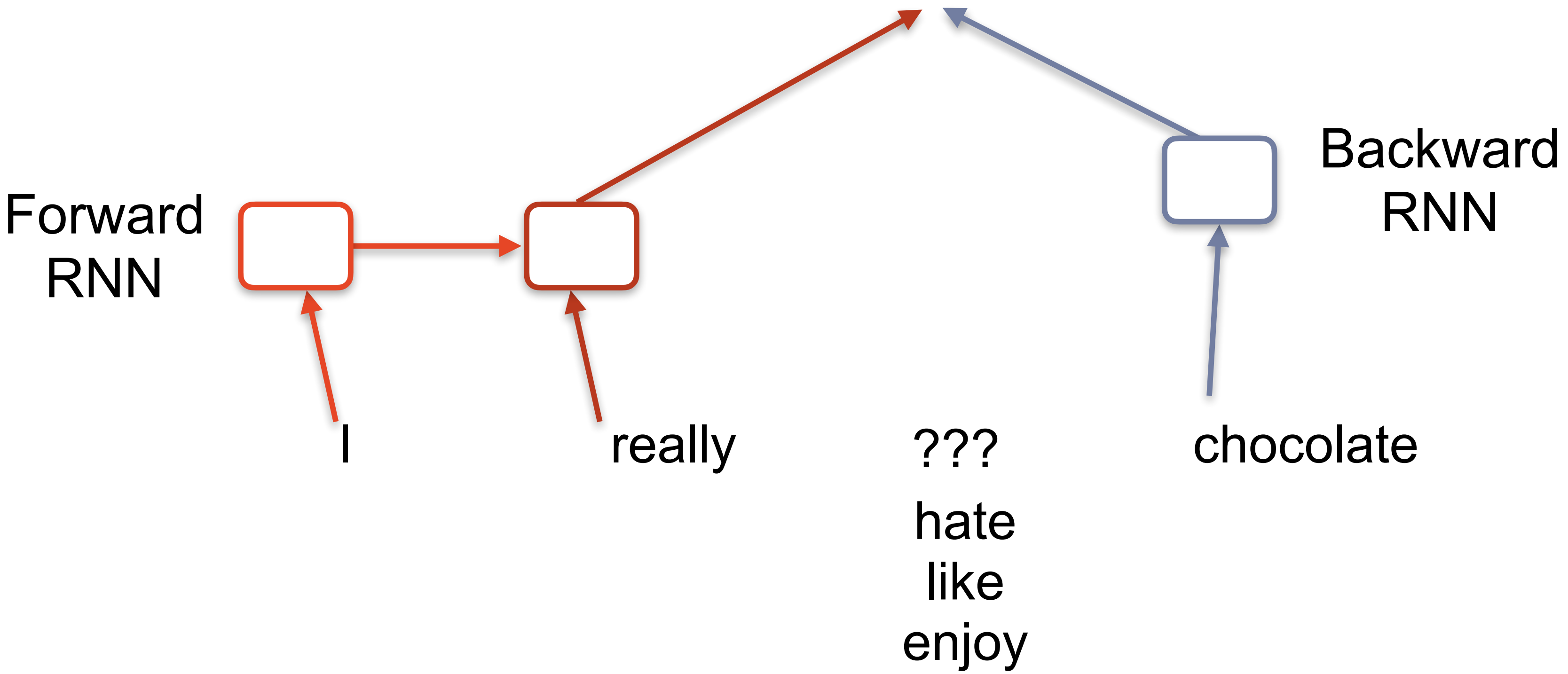
Input: Context words
Output: One word



menti.com 6165 8383

How do we train the model?

Input: Context words
Output: One word



COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- **Contextual Embeddings**

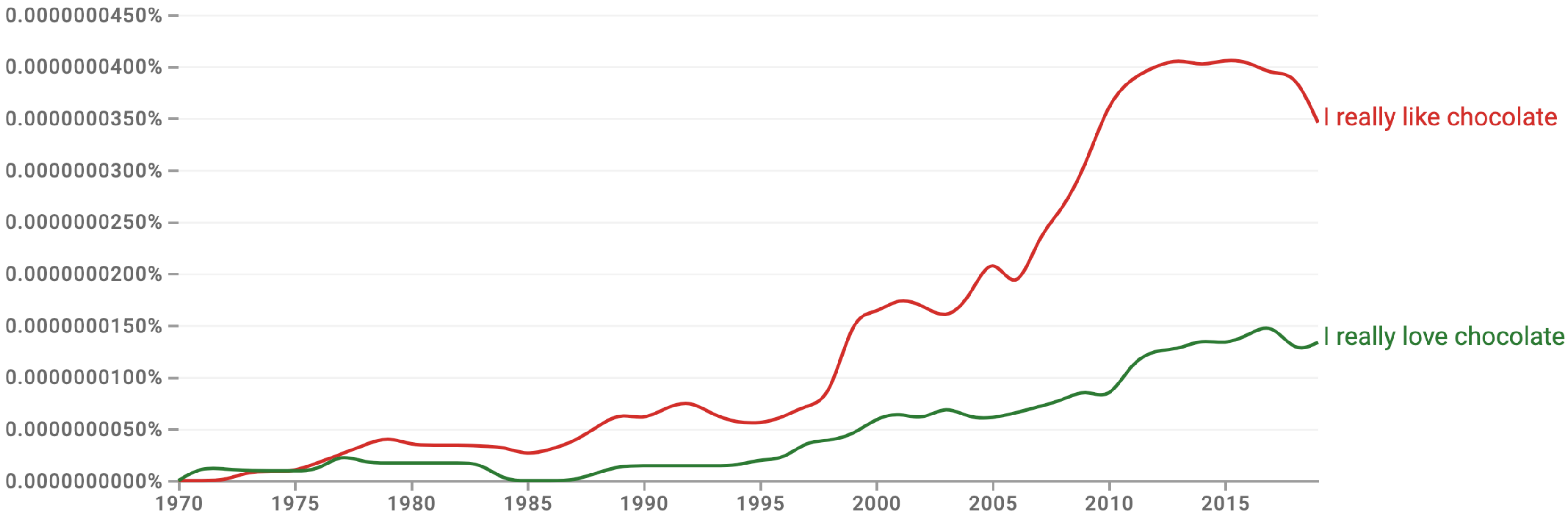
Inference
Lab Preview

menti.com 6165 8383

[aside - Google shows that books agree!]

Counts in books since 1970 of:

“I really * chocolate”

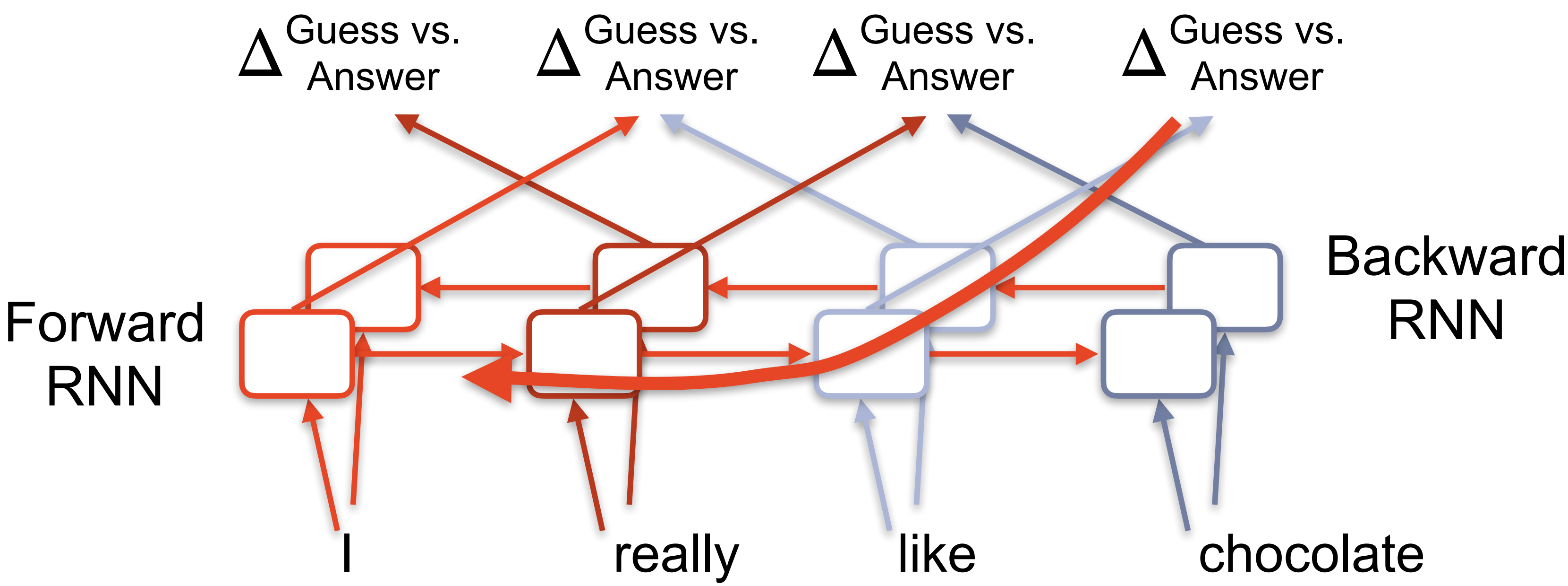


https://books.google.com/ngrams

menti.com 6165 8383

We can train on multiple words at once

Input: All words
Output: All words



COMP 4446 / 5046
Lecture 2, 2024

- Representations
  - Static Embeddings
  - **Contextual Embeddings**
- Inference
- Lab Preview

menti.com 6165 8383

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[illegible]	[illegible]	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	[illegible]	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)	[illegible]	[illegible]	[illegible]	3.2 / 9.8%
NER	Peters et al. (2017)	[illegible]	[illegible]	[illegible]	2.06 / 21%
SST-5	[illegible] (2017)	[illegible]	[illegible]	[illegible]	3.3 / 6.8%

Question Answering

Named Entity Recognition

Sentiment

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

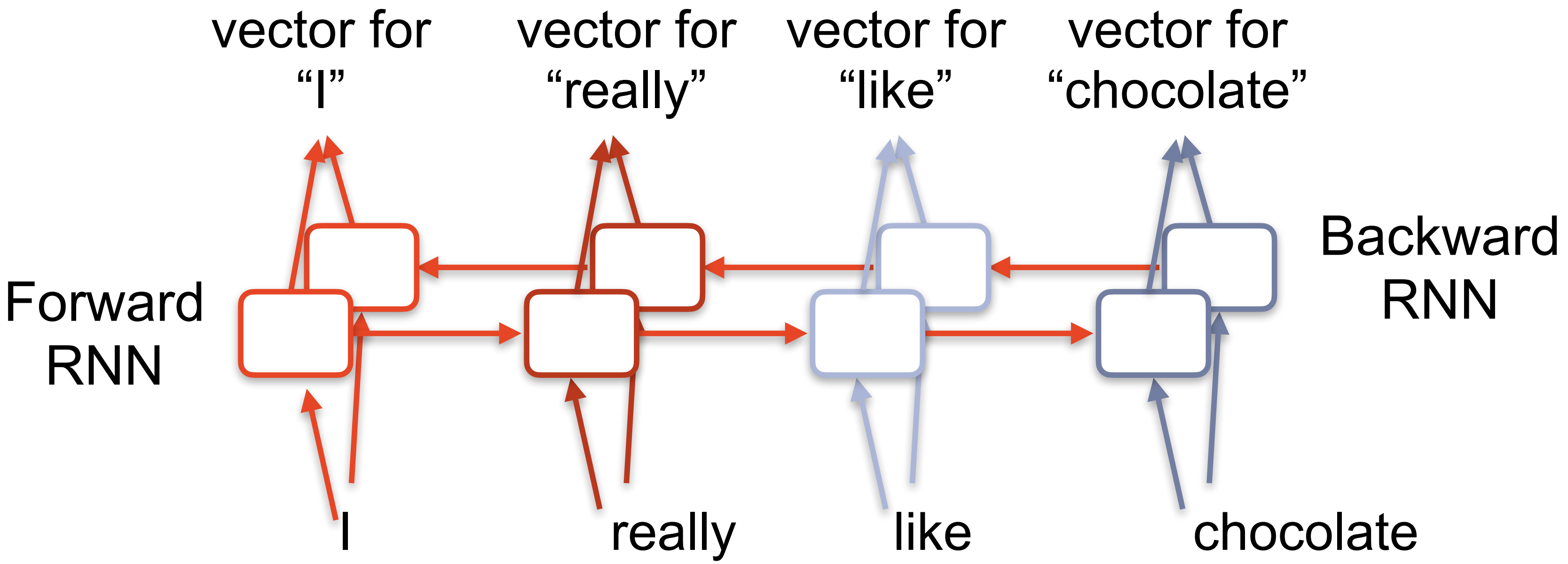
COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- **Contextual Embeddings**

Inference
Lab Preview

menti.com 6165 8383

How are these used?



COMP 4446 / 5046
Lecture 2, 2024

Representations
- Static Embeddings
- **Contextual Embeddings**

Inference
Lab Preview

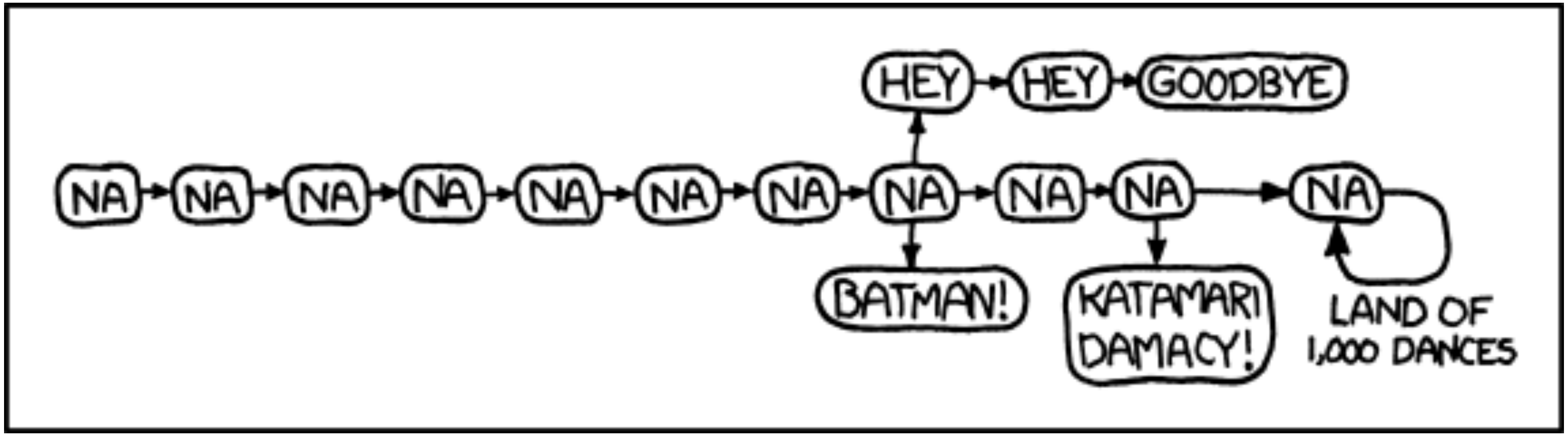
menti.com 6165 8383

The LSTM is just one possible model…

3 minute Break - stretch and visit Menti

menti.com
6165 8383

Na



[I hear that there are actual lyrics later on in Land of 1,000 Dances, but other than the occasional "I said," I've never listened long enough to hear any of them.]

Source: https://xkcd.com/851/

menti.com 6165 8383

Inference

COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- Beam search
- Graph Search

Lab Preview

menti.com 6165 8383

Now, let's explore different inference methods

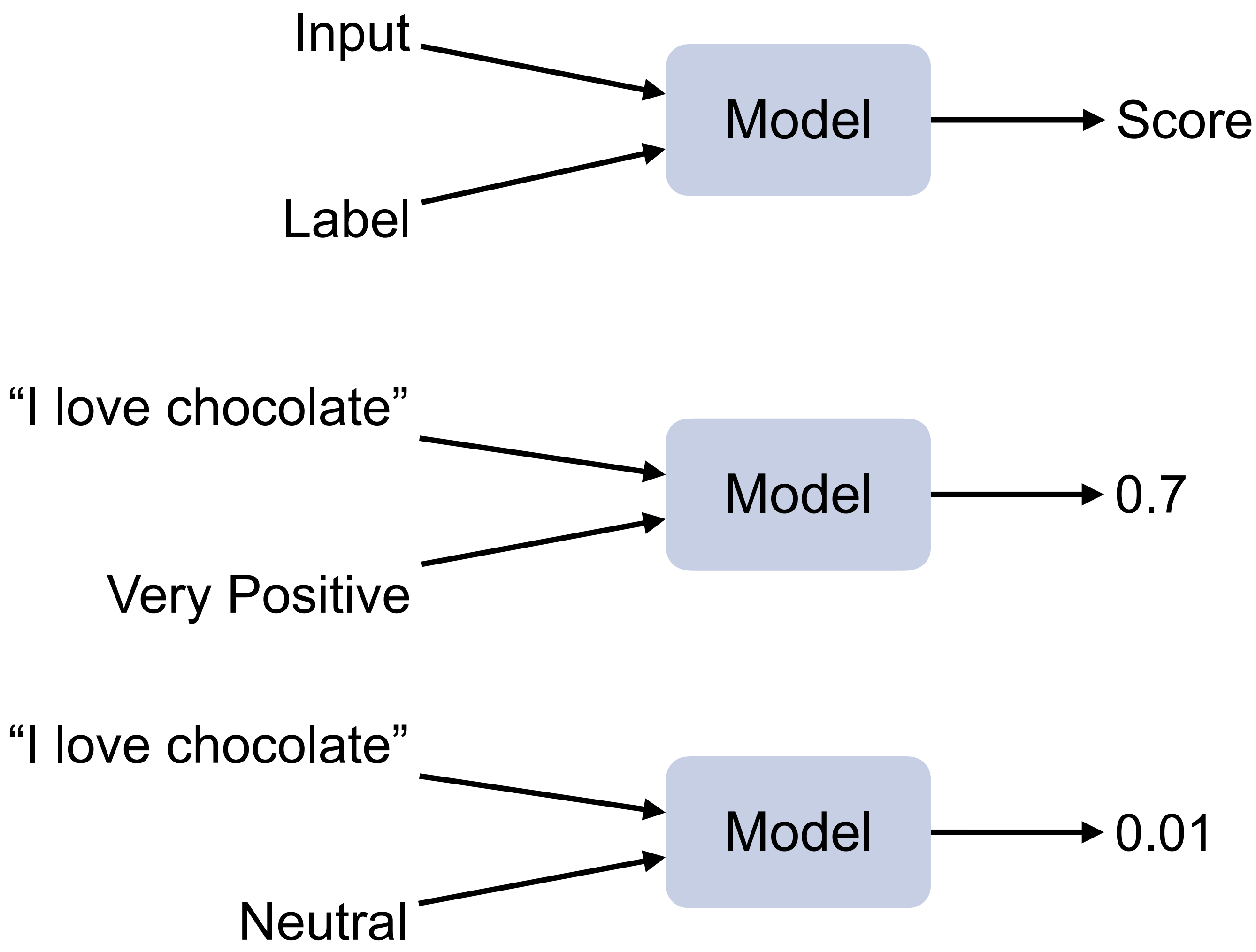
Data — Examples of the language phenomena we want our system to handle

Model — A function that maps (input, output) pairs to scores

Inference Method — A way to make a prediction for an example given a **Model**

Learning Method — A way to update a **Model** given **Data** and an **Inference Method**

We'll treat the model as a black box



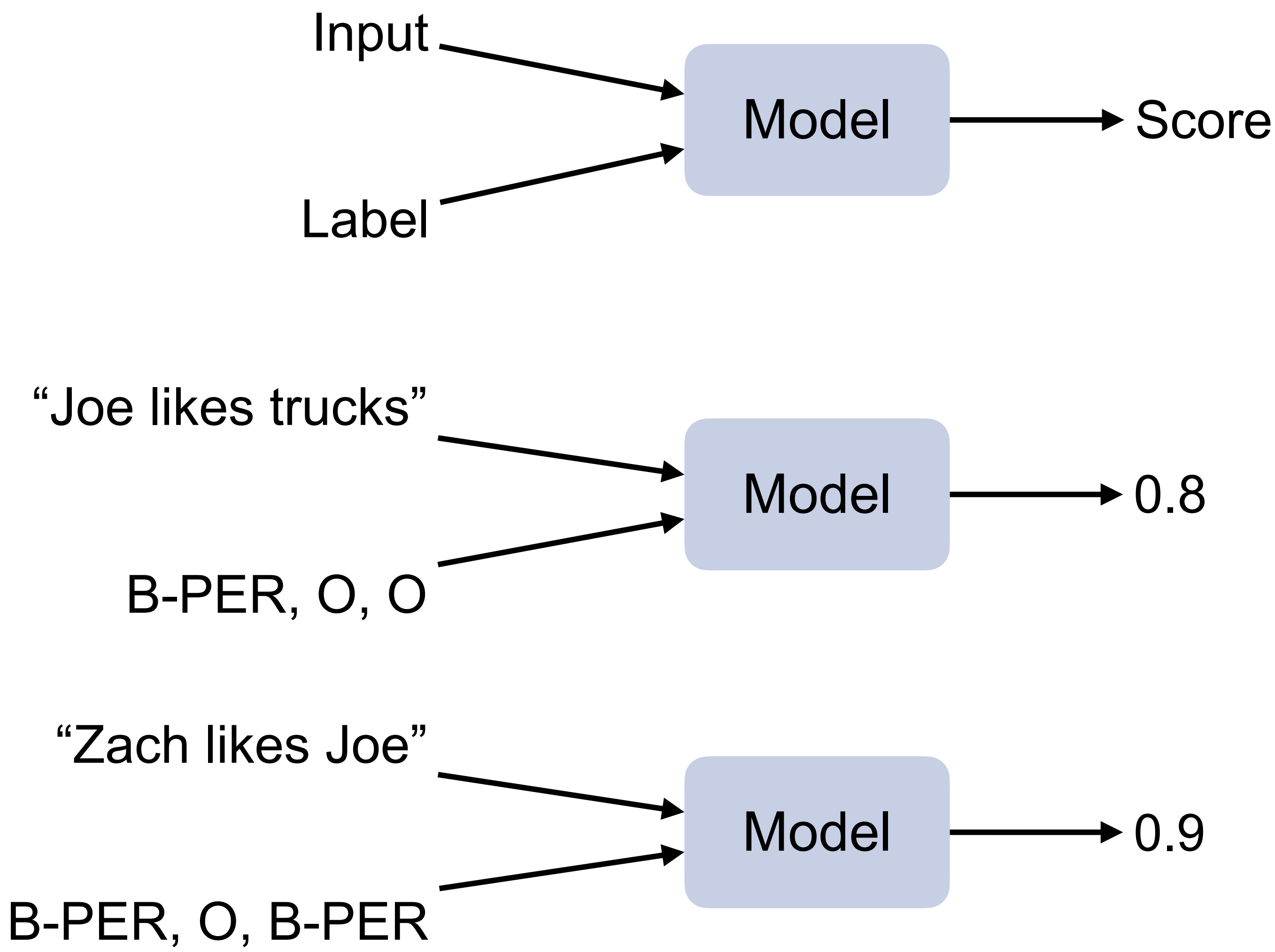
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- Beam search
- Graph Search

Lab Preview

menti.com 6165 8383

The model could score whole sequences

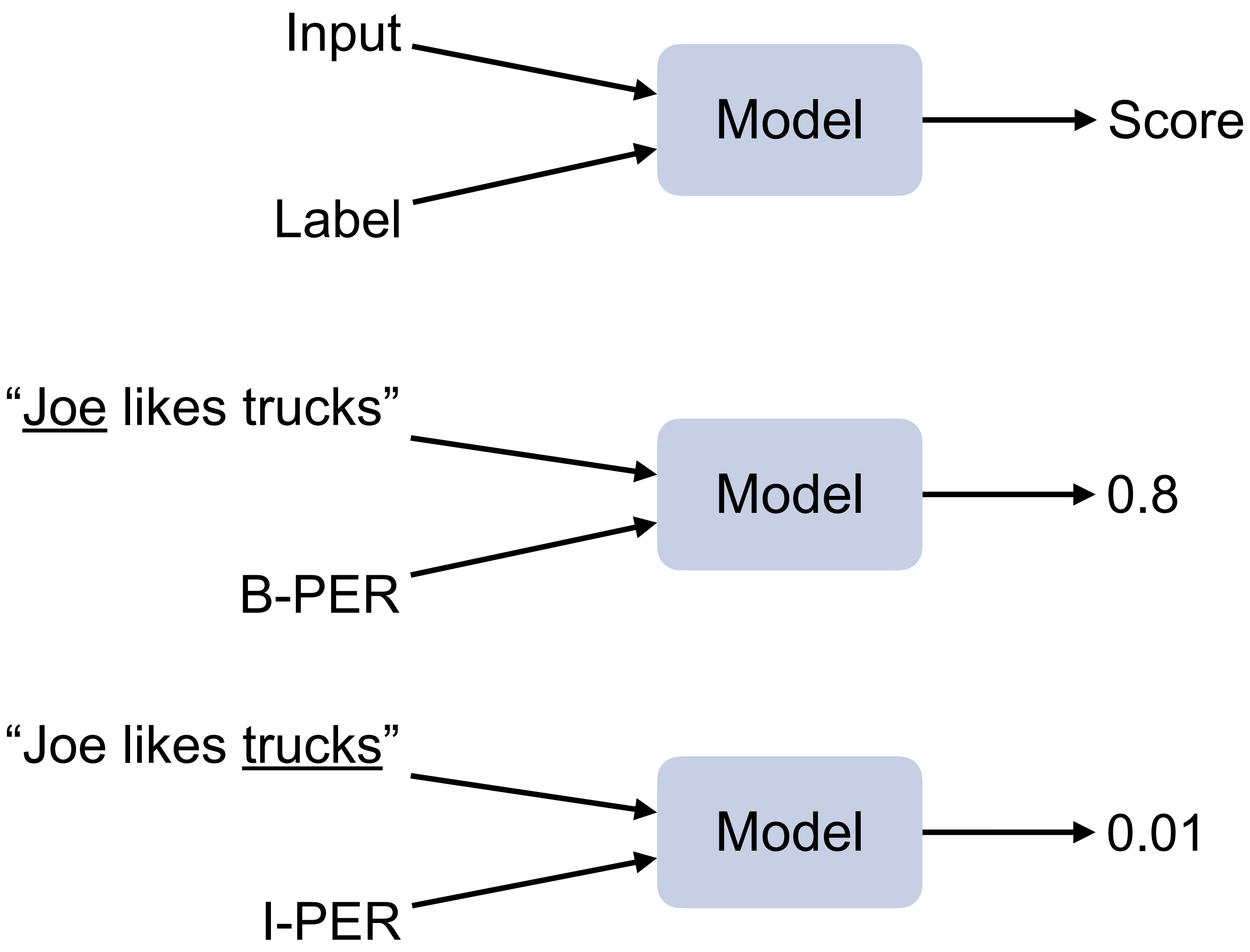


COMP 4446 / 5046
Lecture 2, 2024

- Representations
- Inference
    - Exhaustive
    - Greedy
    - Beam search
    - Graph Search
- Lab Preview

menti.com 6165 8383

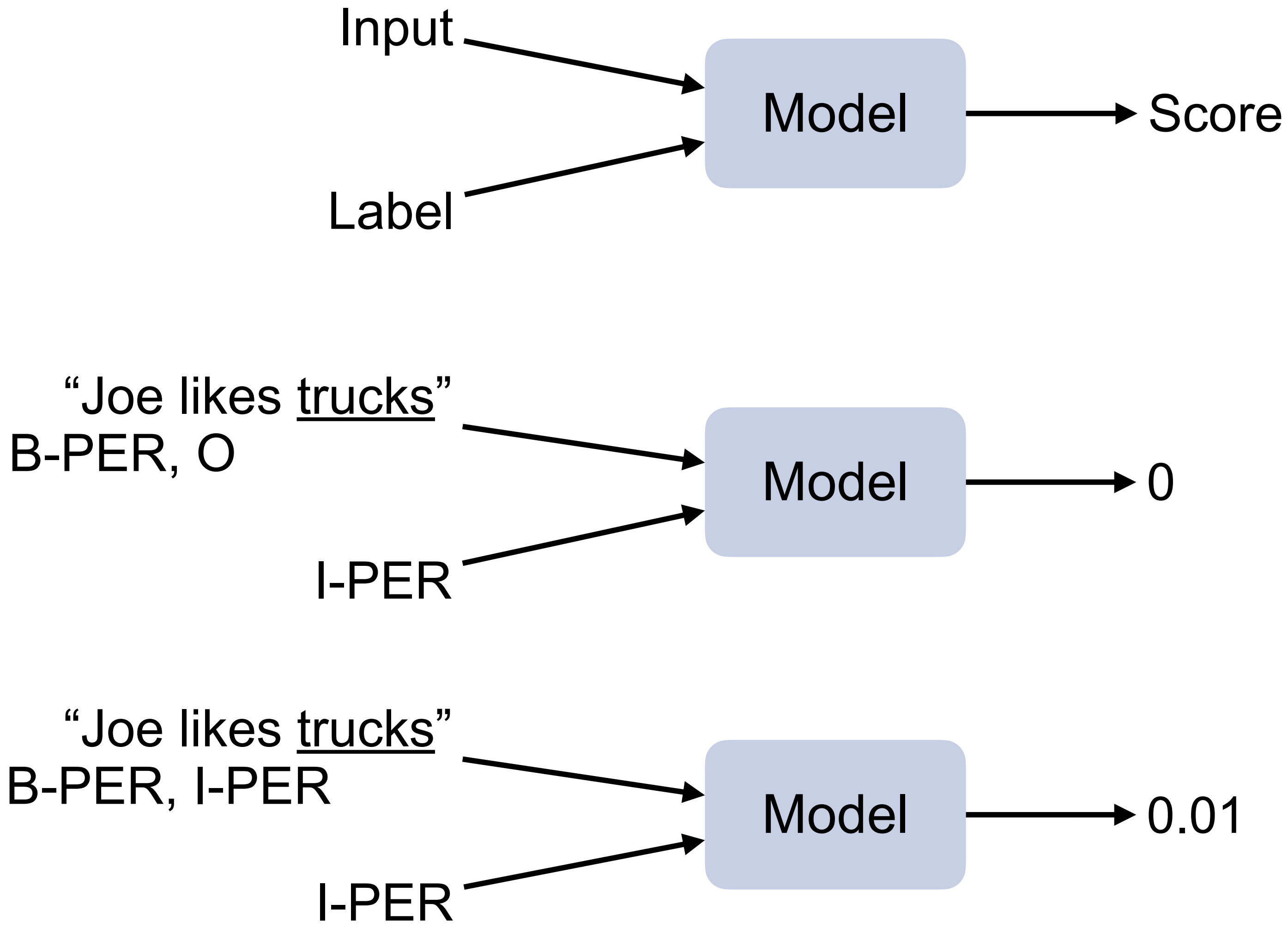
The model could score whole sequences, or just one part



Representations
Inference
- Exhaustive
- Greedy
- Beam search
- Graph Search

Lab Preview

The model could score whole sequences, or just one part, possibly with context



menti.com 6165 8383

COMP 4446 / 5046
Lecture 2, 2024

menti.com 6165 8383

We'll use several running examples in this section

Noun	Verb	Prep	Noun
Adj	Noun	Verb	Noun
Fruit	flies	like	bananas

We'll use several running examples in this section

Ingredients:

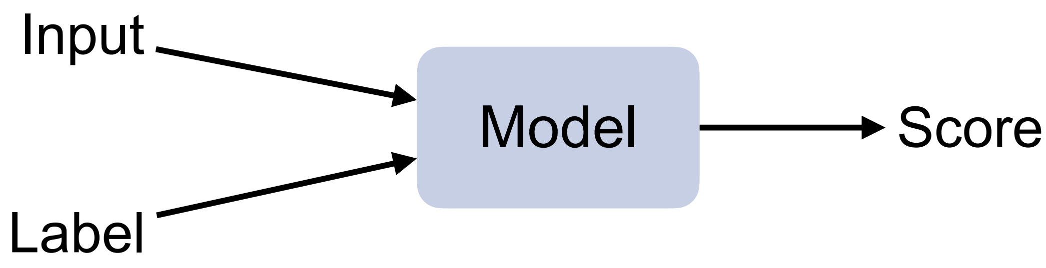
4oz Chocolate, 70% cocoa
1cup Milk
Chocolate shavings

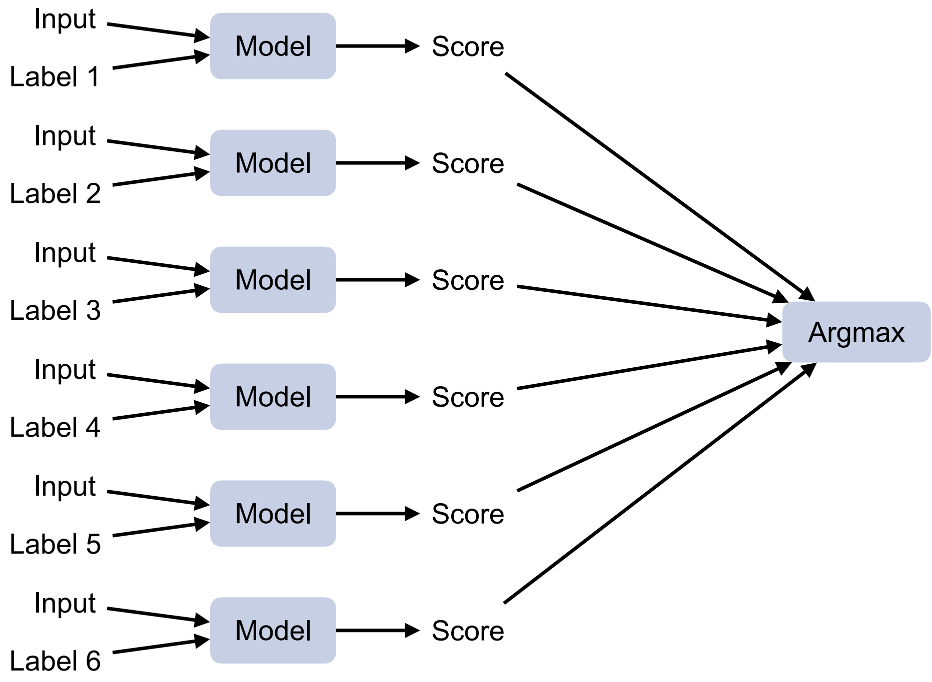
Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering
4. Take off heat and let cool completely (~20 min)
5. Return to stove and heat to desired temperature
6. Top with chocolate shavings

menti.com 6165 8383

In assignment 2, we are using an exhaustive method





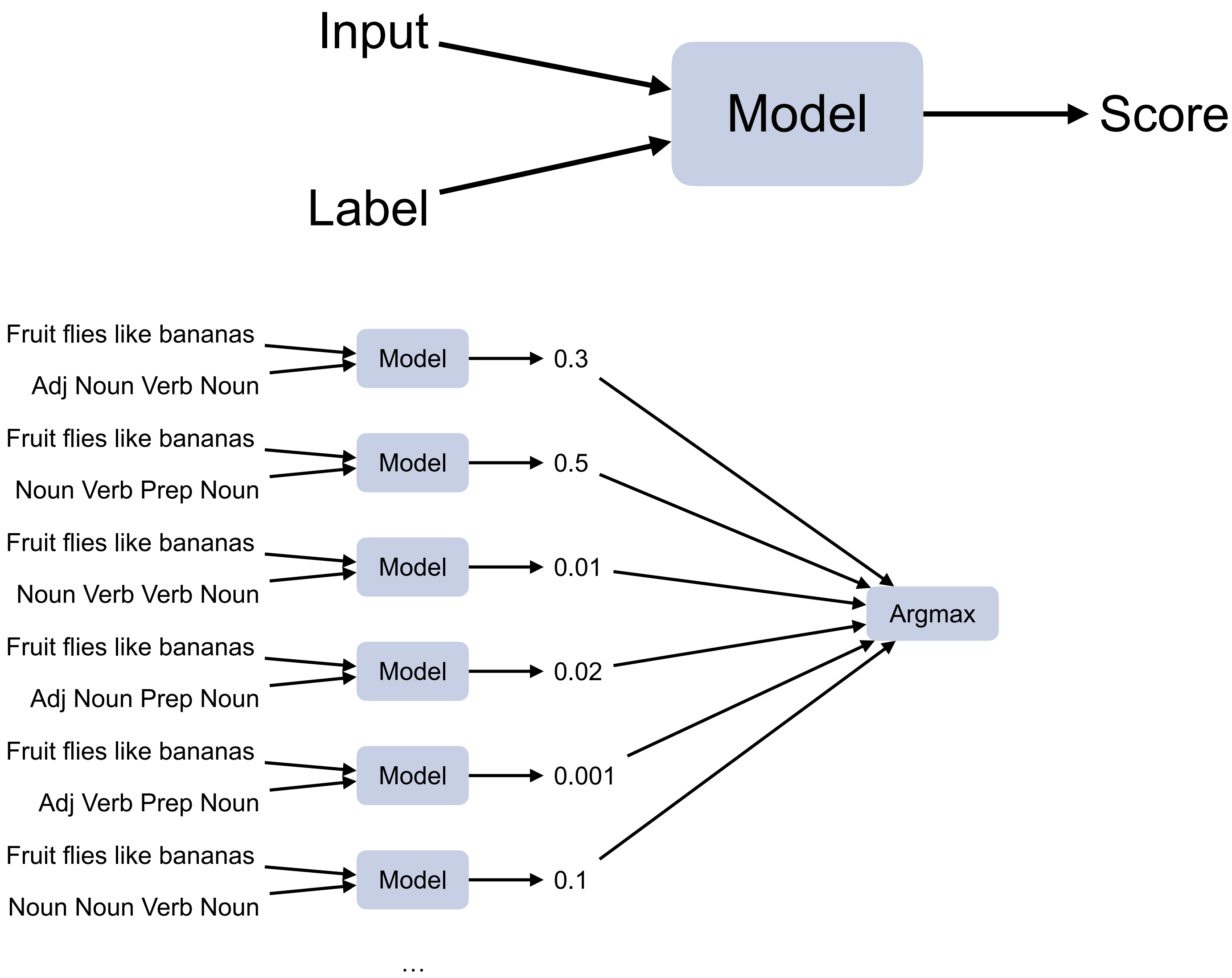
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- **Exhaustive**
- Greedy
- Beam search
- Graph Search

Lab Preview

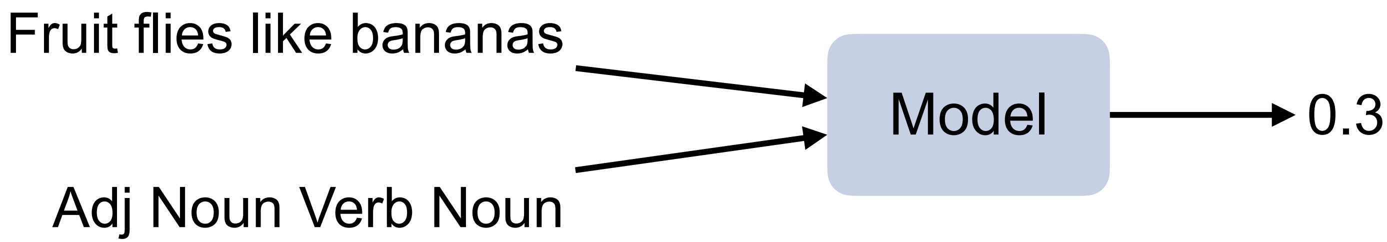
menti.com 6165 8383

In assignment 2, we are using an exhaustive method



Exhaustive search is flexible, but not scalable

Benefit - The model can look at the entire structure



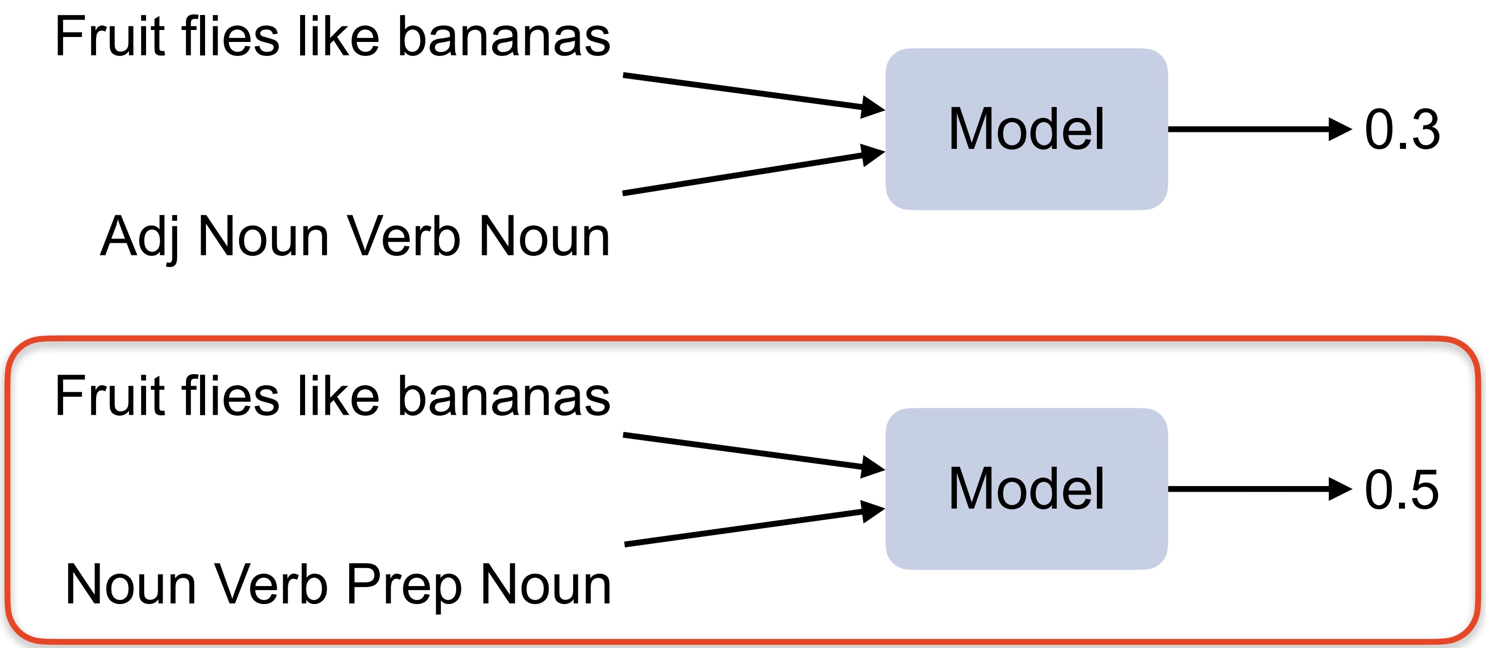
Problem - for many tasks, the search space is huge

17 tags in Universal Dependencies

Options = $|\text{tags}|^{|\text{words}|}$

For this example, 17^4 = 83,521 options!

Note: we are looking for the highest scoring output, which might be wrong



This suggests another way of describing these two components of an NLP system

Model — Tries to assign the highest score to the true output

Inference Method — Tries to find the output that gets the highest score from the **Model**

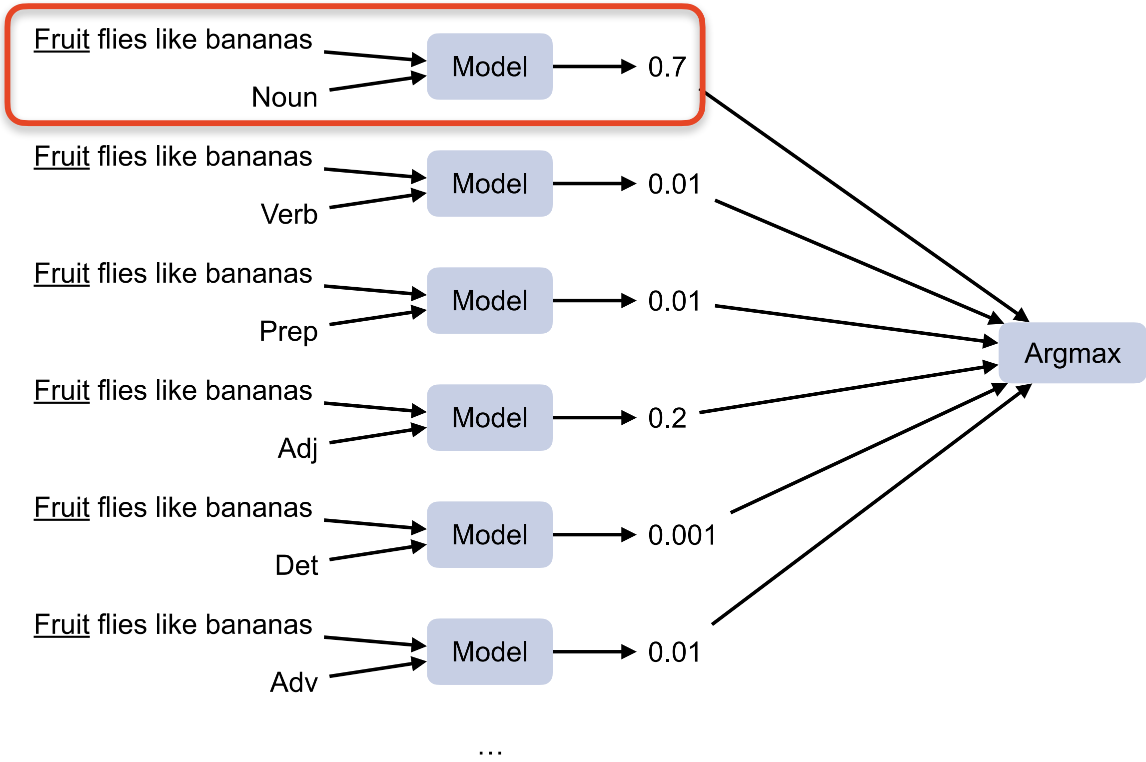
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- **Greedy**
- Beam search
- Graph Search

Lab Preview

menti.com 6165 8383

Core greedy idea: Make choices one at a time



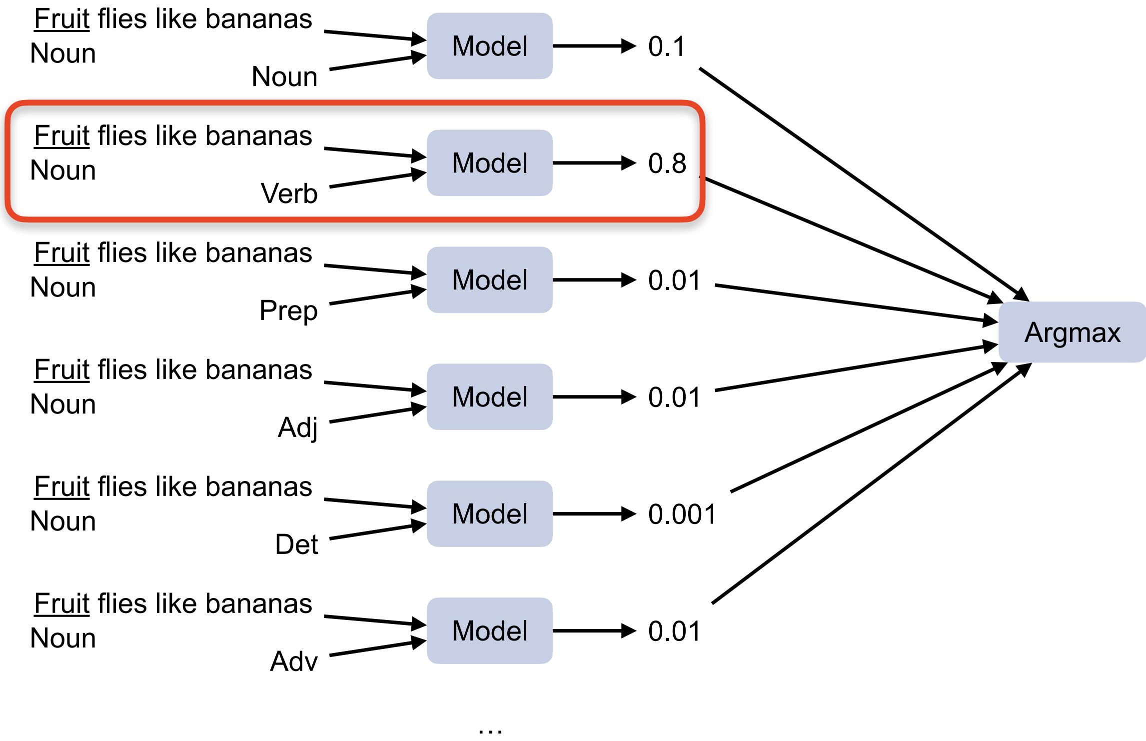
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- **Greedy**
- Beam search
- Graph Search

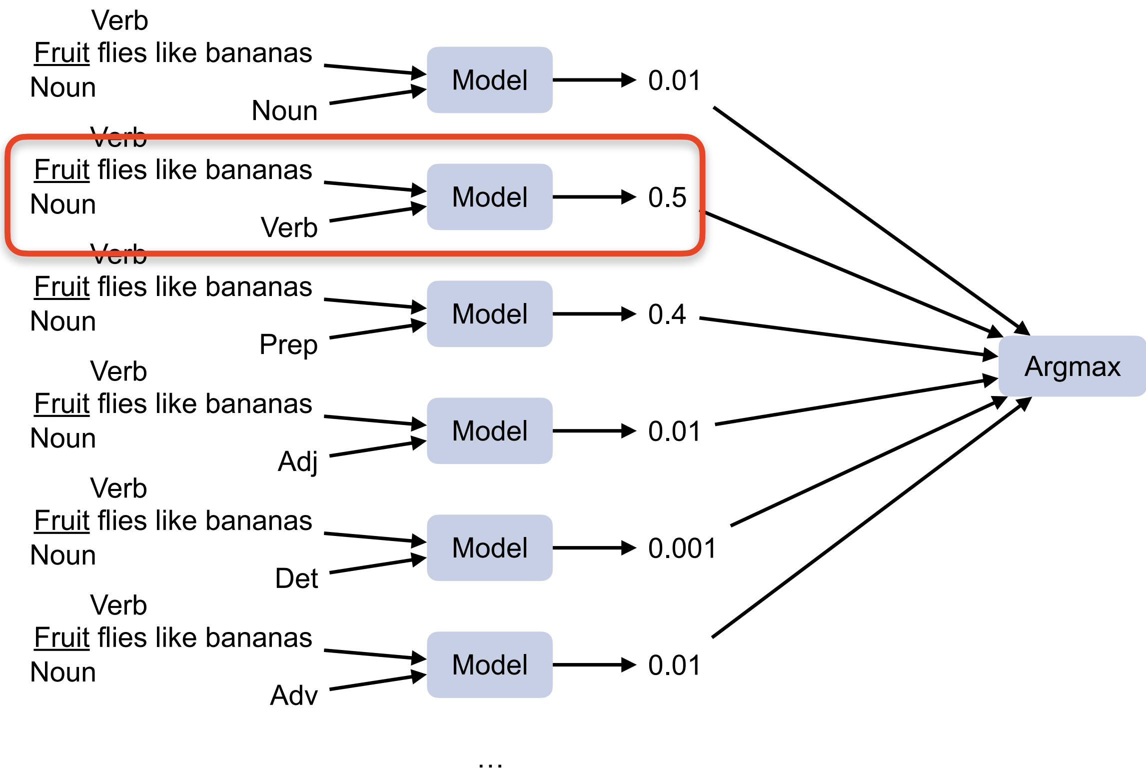
Lab Preview

menti.com 6165 8383

Core greedy idea: Make choices one at a time



Core greedy idea: Make choices one at a time



menti.com 6165 8383

menti.com 6165 8383

The complexity problem is fixed!

17 tags in Universal Dependencies

Options considered in each step = |tags|

Step = |words|

Complexity = |tags| * |words|
= 17 * 4
= 68

menti.com 6165 8383

But… the answer was different

Exhaustive:	Noun	Verb	Prep	Noun
Greedy:	Noun	Verb	Verb	Noun
	Fruit	flies	like	bananas

Note - both of these are wrong, but one is the highest scoring according to the model and the other is not.

menti.com 6165 8383

A few notes on comparing exhaustive and greedy

Sometimes the answer can match. For example, if every part of the output is independent.

Greedy has less information about the output, but can still use a lot of context to make the decision

First variant: **Top-1**

Method: At each step, choose the highest scoring option

This is what we just saw!

First variant: **Top-1**

Method: At each step, choose the highest scoring option

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering
4. Melt chocolate on stove
5. Add milk
6. Heat
7. Melt chocolate on stove

…

Second variant: **Random sampling**

Method: At each step, choose using random sampling from the probability distribution

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

Melt chocolate on stove

Take off heat and let cool completely (~20 min)

0.3 - Melt chocolate on stove
0.29 - Take off heat and let cool completely (~20 min)
0.2 - Pour into mug
0.1 - Simmer for 5 minutes
…

Third variant: **Top-K** sampling

Method: At each step, filter to the top K options, then choose using random sampling from the probability distribution

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

0.3 - Melt chocolate on stove
0.29 - Take off heat and let cool completely (~20 min)
0.2 - Pour into mug
0.1 - Simmer for 5 minutes
…

K = 3

Third variant: **Top-K** sampling

Method: At each step, filter to the top K options, then choose using random sampling from the probability distribution

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

Pour into mug

Take off heat and let cool completely (~20 min)

0.38 - Melt chocolate on stove
0.37 - Take off heat and let cool completely (~20 min)
0.25 - Pour into mug
0.1 - Simmer for 5 minutes
…

K = 3

Fourth variant: **Top-P** sampling

Method: At each step, filter to the options that cover P% of the probability distribution, then choose using random sampling

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

0.3 - Melt chocolate on stove
0.29 - Take off heat and let cool completely (~20 min)
0.2 - Pour into mug
0.1 - Simmer for 5 minutes
...

P = 80%

Fourth variant: **Top-P** sampling

Method: At each step, filter to the options that cover P% of the probability distribution, then choose using random sampling

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

Pour into mug

Simmer for 5 minutes

0.34 - Melt chocolate on stove
0.33 - Take off heat and let cool completely (~20 min)
0.22 - Pour into mug
0.11 - Simmer for 5 minutes
...

P = 80%

Fifth variant: **Contrastive** sampling

Method: At each step, adjust scores based on similarity with recent outputs, then choose the highest scoring option

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

0.3 - * 0.1 ocolate on stove
0.29 * 1.0 off heat and let cool completely (~20 min)
0.2 - * 1.0 to mug
0.1 - * 1.0 r for 5 minutes
… …

menti.com 6165 8383

Fifth variant: **Contrastive** sampling

Method: At each step, adjust scores based on similarity with recent outputs, then choose the highest scoring option

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

0.03 - Melt chocolate on stove
0.29 - Take off heat and let cool completely (~20 min)
0.2 - Pour into mug
0.1 - Simmer for 5 minutes
…

Fifth variant: **Contrastive** sampling

Method: At each step, adjust scores based on similarity with recent outputs, then choose the highest scoring option

Ingredients:

4oz Chocolate, 70% cocoa
1cup Milk

Steps:

1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering

This is starting to mix modelling with inference

0.04 - Melt chocolate on stove
0.40 - Take off heat and let cool completely (~20 min)
0.27 - Pour into mug
0.14 - Simmer for 5 minutes
…

This can be combined with previous approaches

Greedy variant comparison

Top-1	Argmax
Random	Sample - full distribution
Top-K	Sample - partial list, fixed length
Top-P	Sample - partial list, variable length
Contrastive	Adjust scores, then argmax

Core beam idea: Keep track of multiple options

Input at each step:

Fruit flies like bananas	Fruit flies like bananas	Fruit flies like bananas	Fruit flies like bananas
-	Noun	Noun Verb	Noun Verb Verb

Output so far [Greedy]:

Noun	Noun Verb	Noun Verb Prep	Noun Verb Prep Noun

COMP 4446 / 5046
Lecture 2, 2024

- Representations
- Inference
  - Exhaustive
  - Greedy
  - **Beam search**
  - Graph Search
- Lab Preview

menti.com 6165 8383

Core beam idea: Keep track of multiple options

Input at each step:

Fruit flies like bananas	Fruit flies like bananas	Fruit flies like bananas	Fruit flies like bananas
-	Noun	Noun Verb	Adj Noun Verb
	Verb	Adj Noun	Noun Verb Prep
	Adj	Noun Noun	Noun Verb Verb

Output so far [Beam]:

Noun	Noun Verb	Adj Noun Verb	Adj Noun Verb Noun
Verb	Adj Noun	Noun Verb Prep	Noun Verb Prep Noun
Adj	Noun Noun	Noun Verb Verb	Adj Noun Verb Adj

COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- **Beam search**
- Graph Search

Lab Preview

menti.com 6165 8383

Implement with a list of options at each step that you consider extending

K best options so far
(K = 5 here)

A
B
C
D
E

...

New K best options after doing this step

A, q
A, p
A, g
A, u

COMP 4446 / 5046
Lecture 2, 2024

menti.com 6165 8383

Implement with a list of options at each step that you consider extending

K best options so far
(K = 5 here)

New K best options after doing this step

Implement with a list of options at each step that you consider extending

K best options so far
(K = 5 here)

A
B
C
D
E

New K best options after doing this step

E, q
A, x
A, q
B, x

Depending on your scoring method, sometimes you can stop early - if you know that none of the remaining options will be better

Representations
Inference
Exhaustive
Greedy
Beam search
Graph Search
Lab Preview

For variable length outputs, there are multiple possible beam definitions

Beams based on number of output lines:

Ingredients:
4oz Chocolate, 70% cocoa
1cup Milk
Chocolate shavings
Steps:
1. Melt chocolate on stove

Ingredients:
4oz Chocolate, 70% cocoa
1cup Milk
Steps:
1. Melt chocolate on stove
2. Slowly add milk

Beams based on step number:

Ingredients:
4oz Chocolate, 70% cocoa
1cup Milk
Chocolate shavings
Steps:
1. Melt chocolate on stove

Ingredients:
4oz Chocolate, 70% cocoa
1cup Milk
Steps:
1. Melt chocolate on stove

Implementation note - how to select the top K?

A
B
C
D
E

Simple approach, O(nk)
For each option, go through the list to find where it goes. Once found, insert and update.

Heap approach, O(n log k)
Use a min-heap. Update it with each new item.

Quickselect approach, O(n)
Record all options. Use quickselect to find the Kth best. Make one pass through the list to get the other K-1.

Usually k is small enough that any of these are fine and the computation of different options dominates anyway

Sometimes beam search does not capture useful variation

Ingredients:
4oz Chocolate, 70% cocoa

Ingredients:
4oz Chocolate, 75% cocoa

Ingredients:
3.5oz Chocolate, 75% cocoa

Ingredients:
3.5oz Chocolate, 75% cocoa

Ingredients:
4oz Cocoa Powder

We can also treat the problem as a search task and use graph theory

-

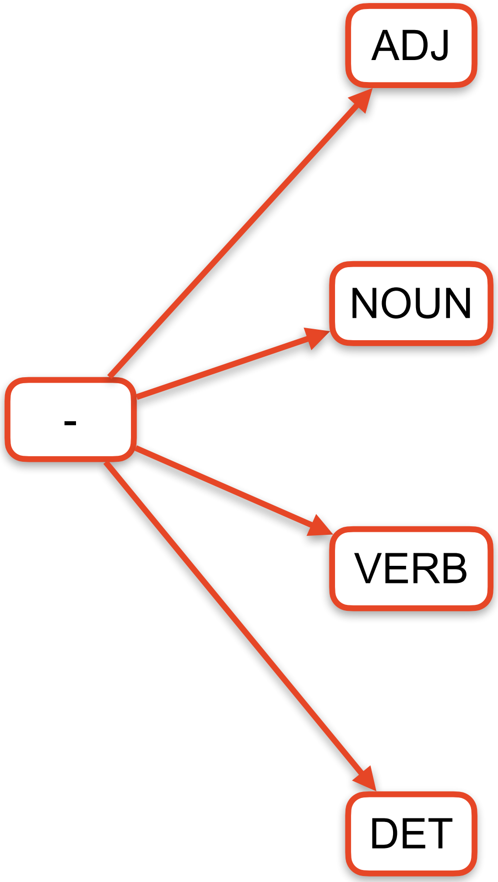
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- Beam search
- **Graph Search**

Lab Preview

menti.com 6165 8383

We can also treat the problem as a search task and use graph theory

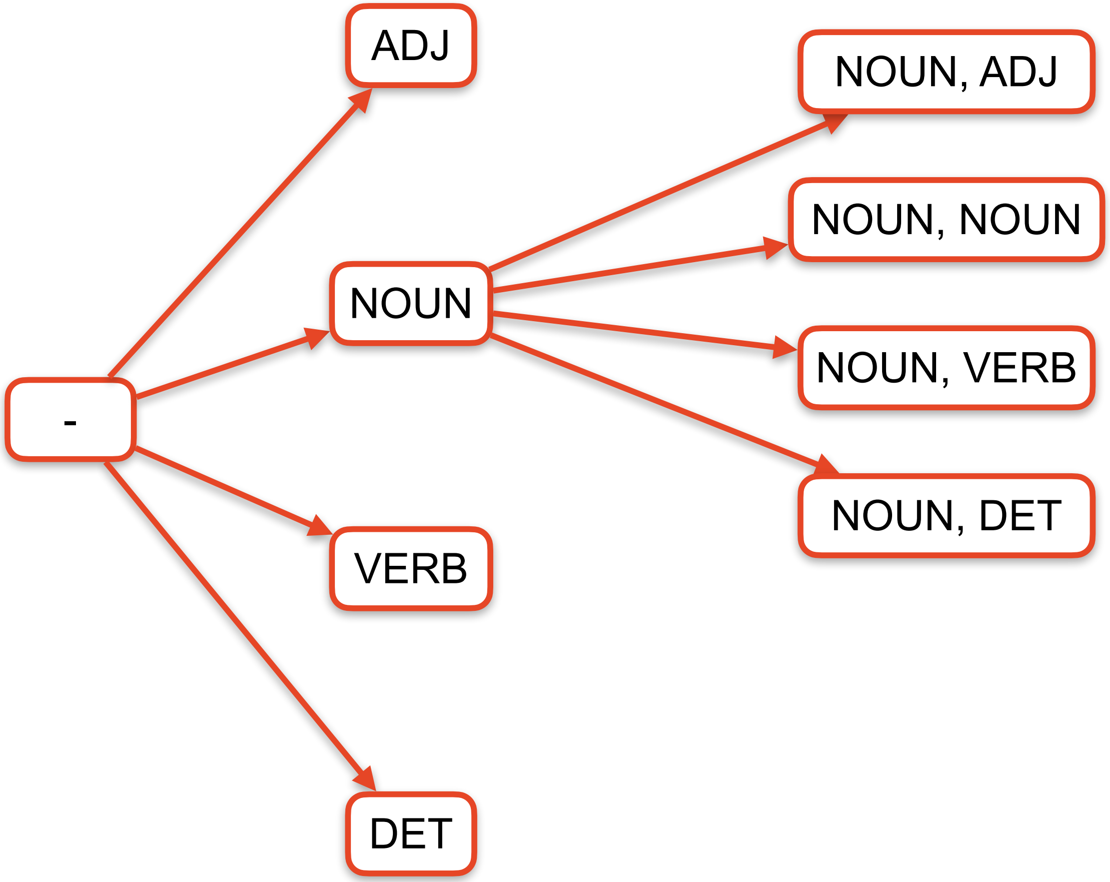


Representations
Inference
- Exhaustive
- Greedy
- Beam search
- **Graph Search**

Lab Preview

menti.com 6165 8383

We can also treat the problem as a search task and use graph theory



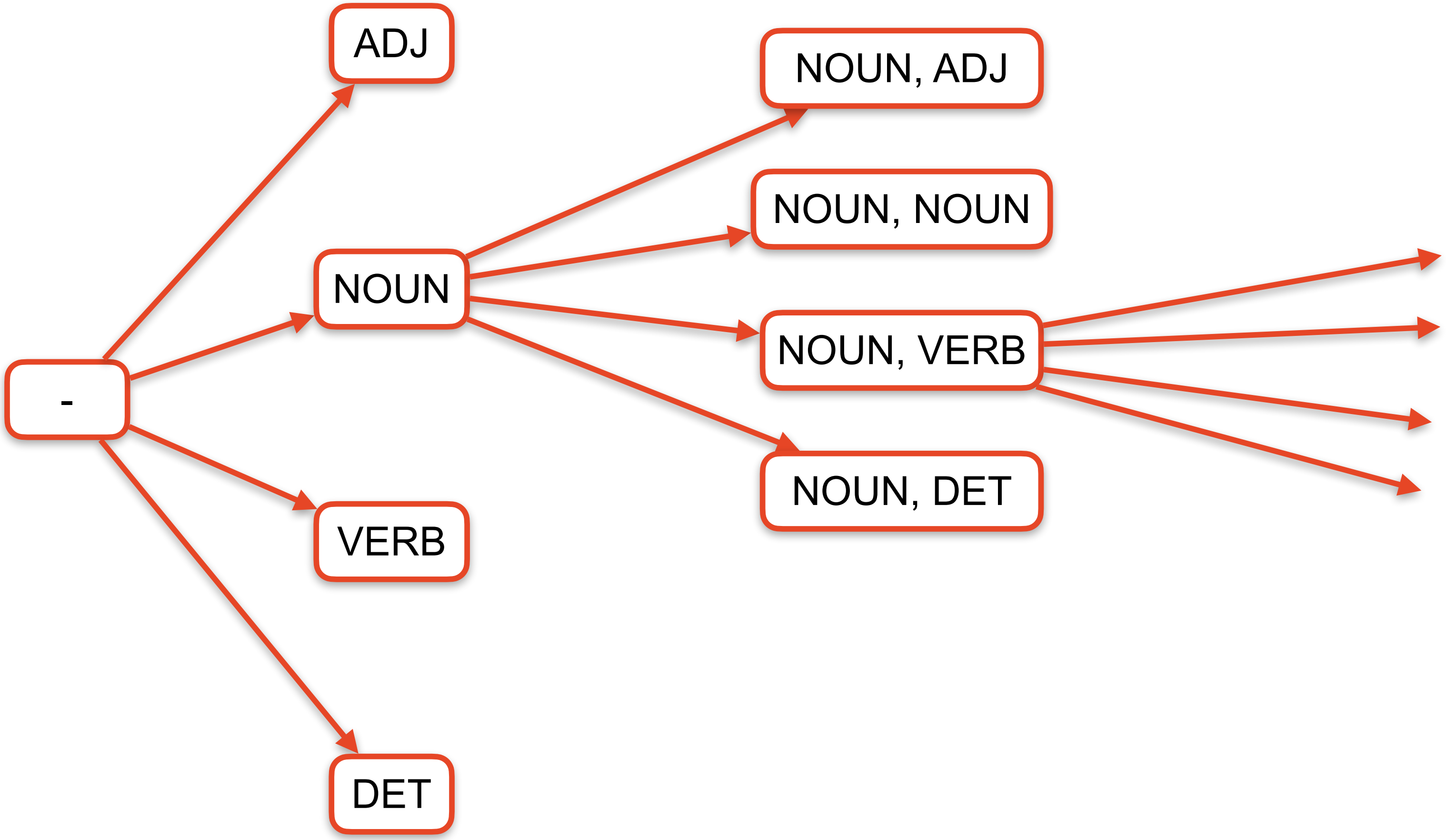
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- Beam search
- **Graph Search**

Lab Preview

menti.com 6165 8383

We can also treat the problem as a search task and use graph theory



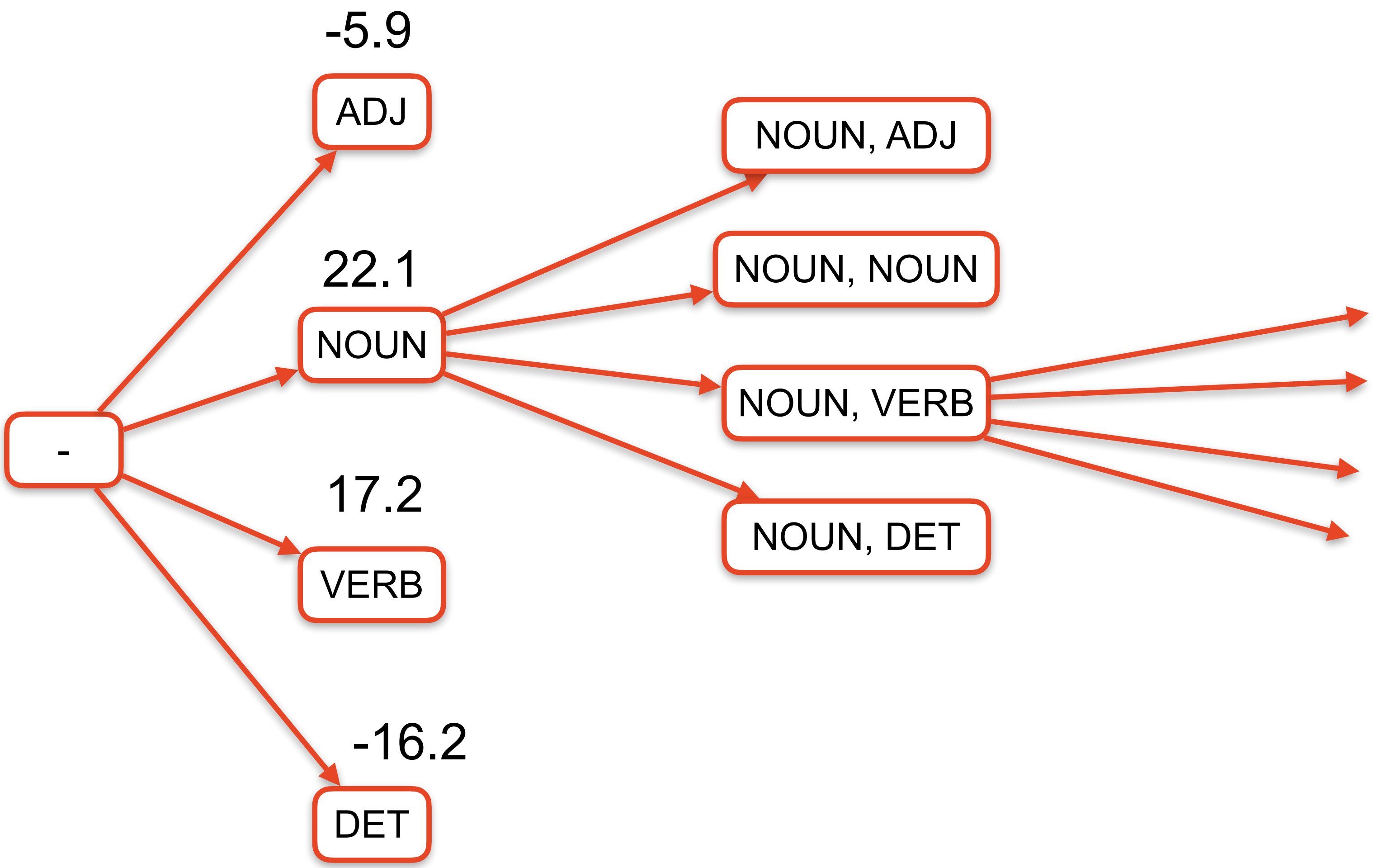
COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
- Exhaustive
- Greedy
- Beam search
- **Graph Search**

Lab Preview

menti.com 6165 8383

Core idea: Estimate final score and use that to guide generation



COMP 4446 / 5046
Lecture 2, 2024

- Representations
- Inference
  - Exhaustive
  - Greedy
  - Beam search
  - **Graph Search**
- Lab Preview

menti.com 6165 8383

Core idea: Estimate final score and use that to guide generation

Ingredients:
4oz Chocolate, 70% cocoa
1cup Milk
Chocolate shavings

Steps:
1. Melt chocolate on stove
2. Slowly add milk
3. Heat until simmering
4. Take off heat and let cool

Score of generation so far

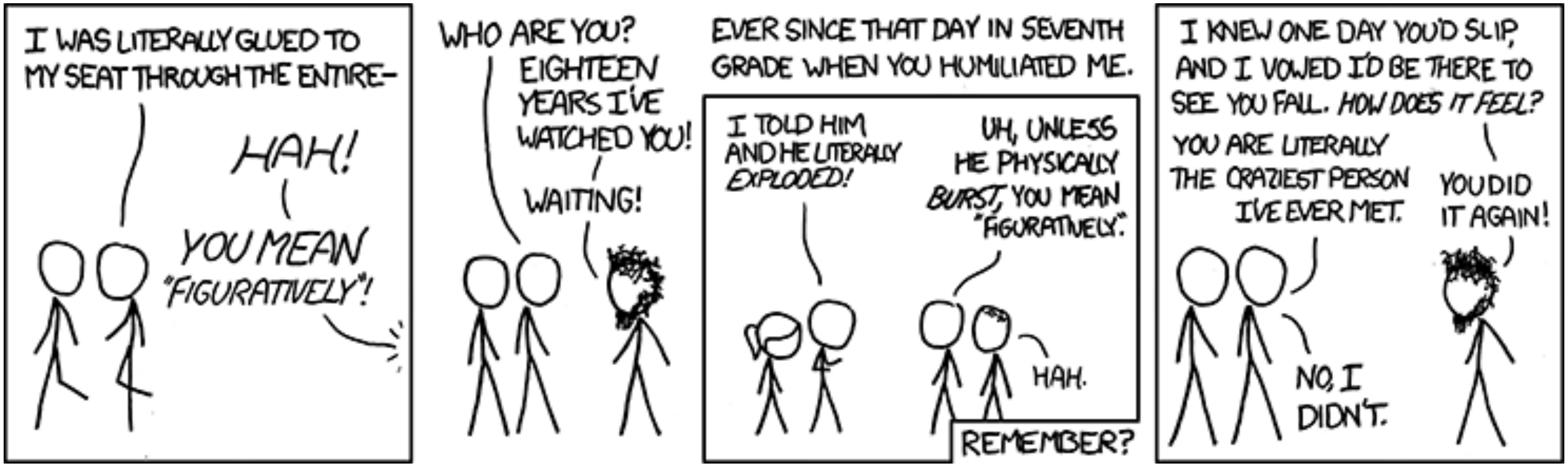
Estimate of score to finish

Note - if you learned about A* search, this sounds similar, but there we are finding the shortest path, here we want the largest score.

3 minute Break - stretch and visit Menti

menti.com
6165 8383

Literally



[The chemistry experiment had me figuratively -- and then shortly thereafter literally -- glued to my seat.]

Source: https://xkcd.com/725/

COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
Lab Preview

menti.com 6165 8383

Lab Preview

COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
Lab Preview

menti.com 6165 8383

New lecture section!

So far - PyTorch

This week - Tensorflow! (Kind of)

Keras - A library on top of Tensorflow

1. Walkthrough of tasks similar to those in past labs
2. Making a multi-layer bidirectional LSTM
3. Trying different sampling methods for text generation

COMP 4446 / 5046
Lecture 2, 2024

Representations
Inference
Lab Preview

menti.com 6165 8383

Muddy Card

Open now, closes at 7:05pm

Go to Ed → Lessons → Lecture 5

https://edstem.org/au/courses/14541/lessons/50402/