THE OHIO STATE UNIVERSITY

CSE 5525: Foundations of Speech and Language Processing

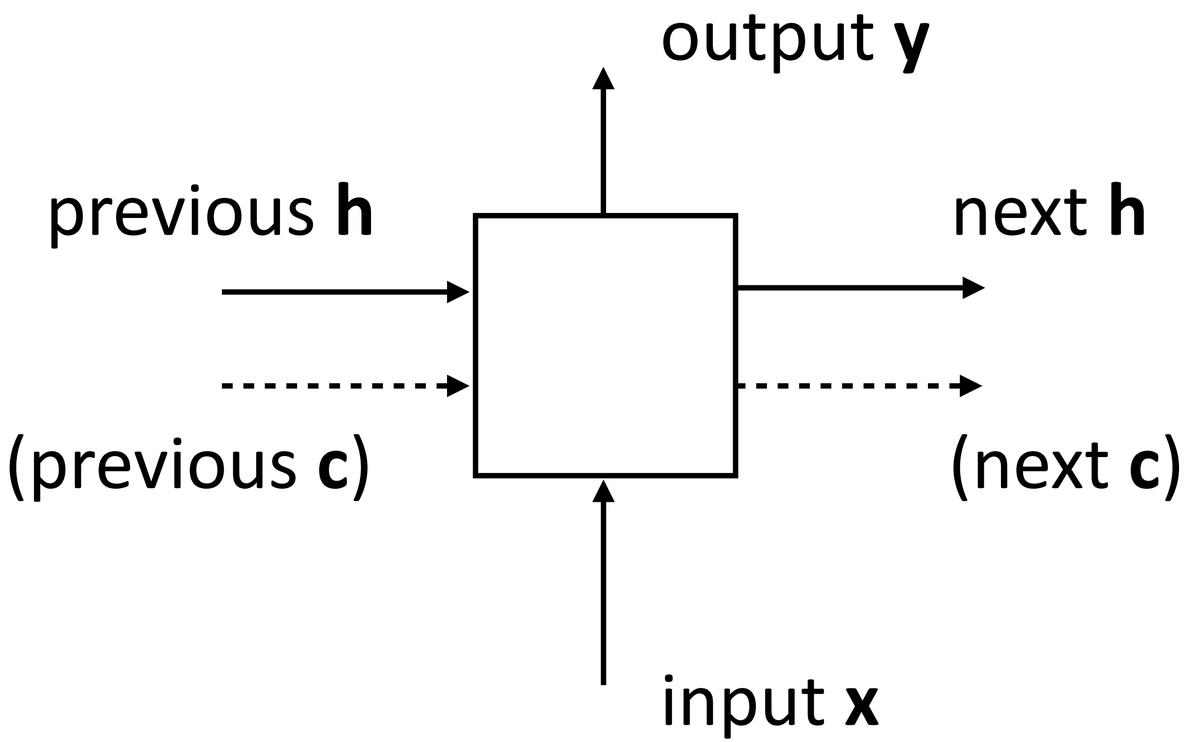
Language Modeling + Pretraining

Huan Sun, Xiang Deng (CSE@OSU)

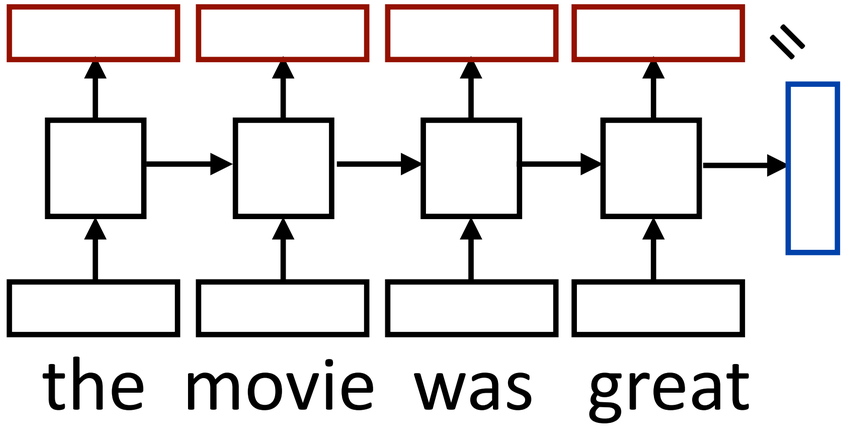
Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides.
Some content adopted from Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd Edition).

Recall: RNNs

- Cell that takes some input **x**, has some hidden state **h**, and updates that hidden state and produces output **y** (all vector-valued)



Recall: RNN Abstraction



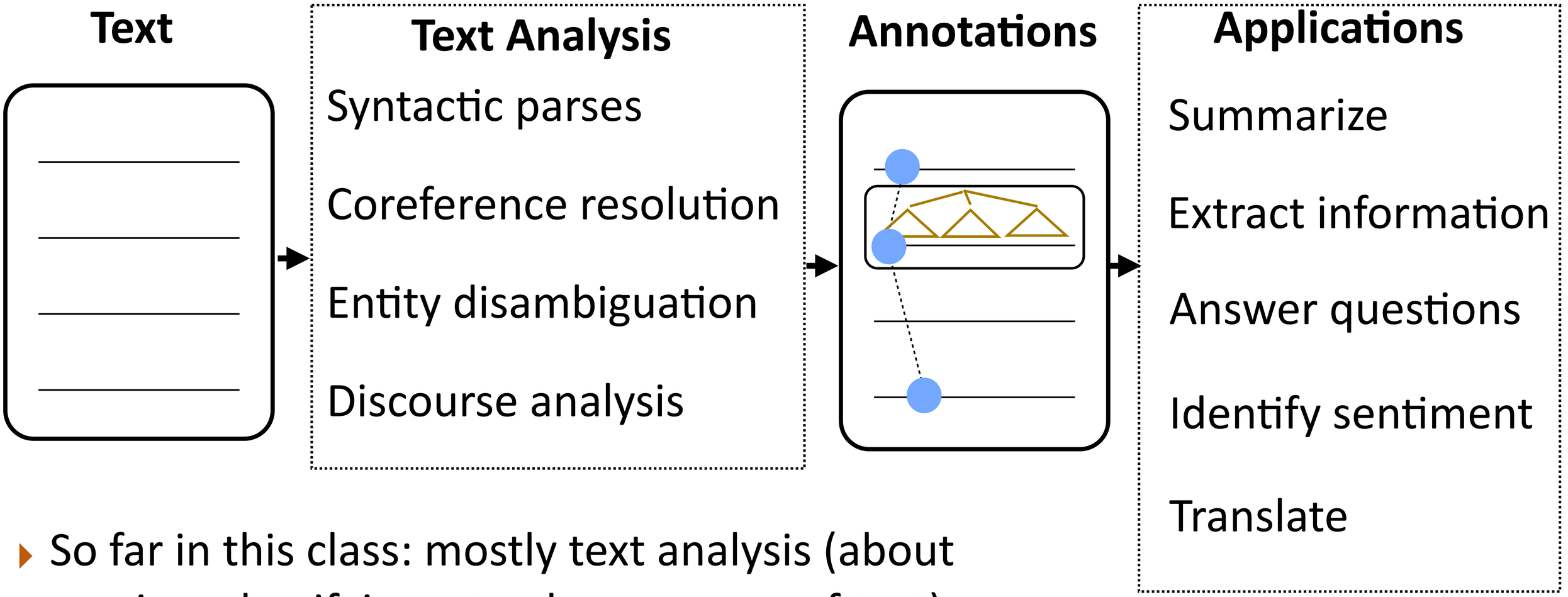
- Encoding of the sentence — can pass this a decoder or make a classification decision about the sentence
- Encoding of each word — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding)
- RNN can be viewed as a transformation of a sequence of vectors into a sequence of context-dependent vectors

This Lecture

- Language modeling
  - N-gram models
  - Neural LMs
- LM-based pretraining: ELMo

Language Modeling

NLP Analysis Pipeline



- So far in this class: mostly text analysis (about tagging, classifying, etc. the structure of text)
- Haven't talked about text *generation* tasks

Challenges in Text Generation

- Dialogue, machine translation, summarization, etc.
  - What to say (content selection + content planning) and how to say it
- Template-based generation systems always generate fluent output
  - A example from the ELIZA system, a chatbot from Weizenbaum (1966). (See 26.2.1 in JM for more info)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
::
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?
My father
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Challenges in Text Generation

- Dialogue, machine translation, summarization, etc.
  - What to say (content selection + content planning) and how to say it
- Template-based generation systems always generate fluent output
- For learned systems, how do we make sure language is plausible?
- Language models: place a distribution P(***w***) over strings ***w*** in a language
  - Next week: P(T,***w***) modeled by probabilistic context-free grammars
  - Today: autoregressive models $P(\mathbf{w}) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots$

Challenges in Text Generation

- Language models: place a distribution P(***w***) over strings ***w*** in a language
  - Today: autoregressive models $P(\mathbf{w}) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots$
    - Machine Translation:
      - P(**high** winds tonight) > P(**large** winds tonight)
    - Spell Correction
      - P(drive for fifteen **minutes**) > P(drive for fifteen **minuets**)
    - Speech Recognition
      - P(I saw a van) >> P(eyes awe of an)

N-gram Language Models

$$P(\mathbf{w}) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots$$

- n-gram models: distribution of next word is a multinomial conditioned on previous n-1 words $P(w_i|w_1, \ldots, w_{i-1}) = P(w_i|w_{i-n+1}, \ldots, w_{i-1})$

  I visited San _____ put a distribution over the next word

$$P(w|\text{visited San}) = \frac{\text{count}(\text{visited San}, w)}{\text{count}(\text{visited San})}$$

  Maximum likelihood estimate of this 3-gram probability from a corpus

- Just relies on counts, even in 2008 could scale up to 1.3M word types, 4B n-grams (all 5-grams occurring >40 times on the Web)

Smoothing N-gram Language Models

- What happens when we scale to longer contexts?

$P(w|\text{to})$ *to* occurs 10M times in corpus

$P(w|\text{go to})$ *go to* occurs 100,000 times in corpus

$P(w|\text{to go to})$ *go to* occurs 10,000 times in corpus

$P(w|\text{want to go to})$ *want to go to*: only 100 occurrences

- Probability counts get very sparse, and we often want information from 5+ words away

Smoothing N-gram Language Models

I visited San _____ put a distribution over the next word

- Add-one estimation, or Laplace smoothing
  - Pretend we saw each word one more time than we did

$$P(w \mid \text{visited San}) = \frac{\text{count}(\text{visited San}, w) + 1}{\text{count}(\text{visited San}) + V}$$

Smoothing N-gram Language Models

I visited San _____ put a distribution over the next word

- Backoff and Interpolation
  - Longer context leads to sparse probability counts, how about condition on less context for contexts you haven't learned much about:
  - Backoff: revert to N-1 if do not have good estimation using N-gram
  - Interpolation: Use a mix of unigram, bigram, …

$$P(w \mid \text{visited San}) = \lambda_1 \frac{\text{count}(\text{visited San}, w)}{\text{count}(\text{visited San})} + \lambda_2 \frac{\text{count}(\text{San}, w)}{\text{count}(\text{San})}$$

Smoothing N-gram Language Models

I visited San _____ put a distribution over the next word

$$P(w \mid \text{visited San}) = \lambda_1 \frac{\text{count}(\text{visited San}, w)}{\text{count}(\text{visited San})} + \lambda_2 \frac{\text{count}(\text{San}, w)}{\text{count}(\text{San})}$$

- How to set the lambdas
  - Use a held-out corpus, train/dev/test
  - Choose lambdas to maximize the probability of held-out data
    - Fix the N-gram probabilities on the training data
    - Search for lambdas using the held-out data

Smoothing N-gram Language Models

I visited San _____ put a distribution over the next word

- One technique is “absolute discounting:” subtract off constant *k* from numerator, set lambda to make this normalize (*k*=1 is like leave-one-out)

$$P(w|\text{visited San}) = \frac{\text{count}(\text{visited San}, w) - k}{\text{count}(\text{visited San})} + \lambda \frac{\text{count}(\text{San}, w)}{\text{count}(\text{San})}$$

- Use held-out data to find a good k
  - Church and Gale (1991) divide up 22M words of AP Newswire
  - 0.75 seems a good estimation

Bigram count in training	Bigram count in heldout set
0	.0000270
1	0.448
2	1.25
3	2.24
4	3.23
5	4.21
6	5.23
7	6.21
8	7.21

Kneser-Ney Smoothing

- Suppose context = *John visited Madagascar* and *Madagascar* has count 0
- Absolute discounting:

$$P(w|\text{John visited Madagascar}) = \lambda_1 \frac{\text{count}(\text{John visited Madagascar}, w)}{\text{count}(\text{John visited Madagascar})} + \ldots + \lambda_4 \frac{\text{count}(w)}{\text{count}(\cdot)}$$

  - First terms all end up being 0/0, so we back off to unigram probability
  - What will the highest probability word be here?
- Instead: use probability based on number of unique contexts *w* occurs in (type counts):

$$P(w) = \frac{|\{p : \text{count}(p, w) > 0\}|}{\sum_w |\{p : \text{count}(p, w) > 0\}|}$$

- Kneser-Ney (1994): absolute discounting w/ type counts for lower-order probs

Engineering N-gram Models

- For 5+-gram models, need to store between 100M and 10B context-word-count triples

(a) Context-Encoding

w	c	val
1933	15176585	3
1933	15176587	2
1933	15176593	1
1933	15176613	8
1933	15179801	1
1935	15176585	298
1935	15176589	1

(b) Context Deltas

Δw	Δc	val
1933	15176585	3
+0	+2	1
+0	+5	1
+0	+40	8
+0	+188	1
+2	15176585	298
+0	+4	1

(c) Bits Required

$\lvert\Delta w\rvert$	$\lvert\Delta c\rvert$	$\lvert val\rvert$
24	40	3
2	3	3
2	3	3
2	9	6
2	12	3
4	36	15
2	6	3

- Make it fit in memory by *delta encoding* scheme: store deltas instead of values and use variable-length encoding

Pauls and Klein (2011), Heafield (2011)

Engineering N-gram Models

- Language Modeling Toolkits
  - SRILM http://www.speech.sri.com/projects/srilm/
  - KenLM https://kheafield.com/code/kenlm/
- N-gram
  - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

```
serve as the incoming 92
serve as the incubator 99
serve as the independent 794
serve as the index 223
serve as the indication 72
serve as the indicator 120
serve as the indicators 45
```

LM Evaluation

- Put the model in task, and evaluate with task accuracy - what task to use?
- Check the accuracy on test text corpus — predicting the next word is generally impossible so accuracy values would be very low

LM Evaluation

- Evaluate LMs on the likelihood of held-out data (averaged to normalize for length)

$$\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_1,\ldots,w_{i-1})$$

- Perplexity: exp(average negative log likelihood). Lower is better
  - Suppose we have probs 1/4, 1/3, 1/4, 1/3 for 4 predictions
  - Avg NLL (base e) = 1.242 Perplexity = 3.464
- What is the perplexity of a random guessing model, with vocab of N

LM Evaluation

- Evaluate LMs on the likelihood of held-out data (averaged to normalize for length)

$$\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_1,\ldots,w_{i-1})$$

- Perplexity: exp(average negative log likelihood). Lower is better
  - Suppose we have probs 1/4, 1/3, 1/4, 1/3 for 4 predictions
  - Avg NLL (base e) = 1.242 Perplexity = 3.464
- What is the perplexity of a random guessing model, with vocab of N

$$PP = exp(-\frac{1}{n}\sum_{i=1}^{n} logP(w_i \mid w_i, \ldots, w_{i-1})) = exp(-\frac{1}{n}\sum_{i=1}^{n} log\frac{1}{N}) = N$$

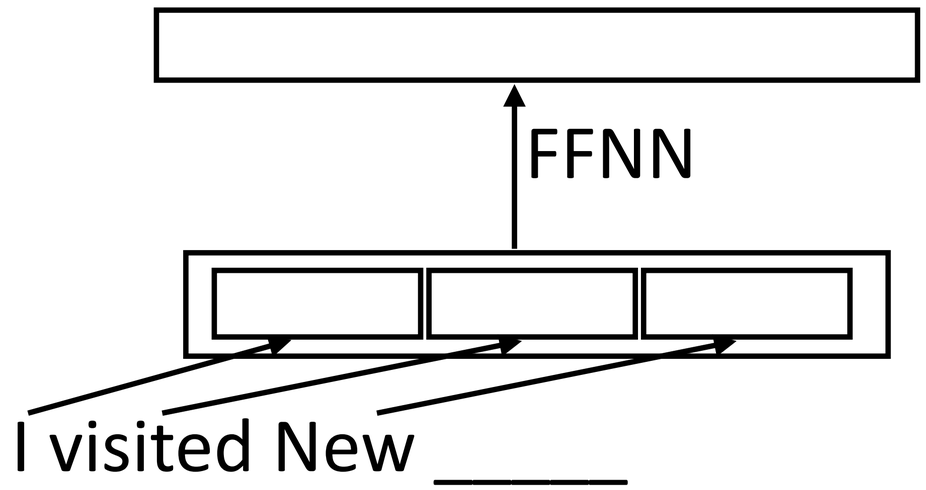
Results

- Evaluate on Penn Treebank: small dataset (1M words) compared to what's used in MT, but common benchmark
- Kneser-Ney 5-gram model with cache: PPL = 125.7
- LSTM: PPL ~ 60-80 (depending on how much you optimize it)
- Melis et al.: many neural LM improvements from 2014-2017 are subsumed by just using the right regularization (right dropout settings). So LSTMs are pretty good
  - Main tricks: changing some tricky dropout settings + how params are structured (sizes, etc.)

Merity et al. (2017), Melis et al. (2017)

Neural Language Models

- Early work: feedforward neural networks looking at context

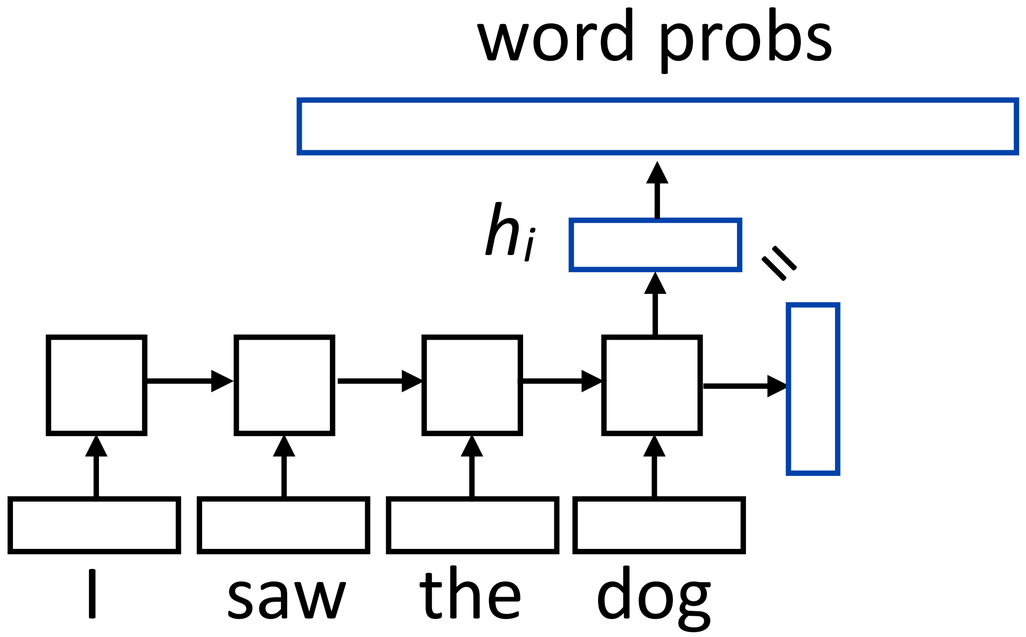


$$P(w_i|w_{i-n},\ldots,w_{i-1})$$

- Slow to train over lots of data!
- Still only look at a fixed window of information…can we use more?

Mnih and Hinton (2003)

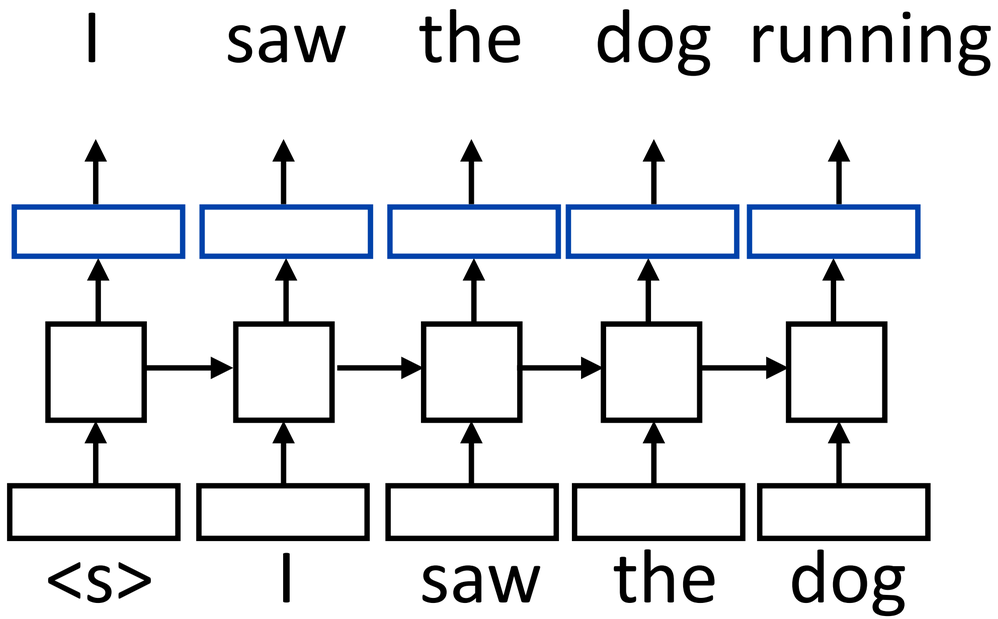
RNN Language Modeling



$$P(w|\text{context}) = \text{softmax}(W\mathbf{h}_i)$$

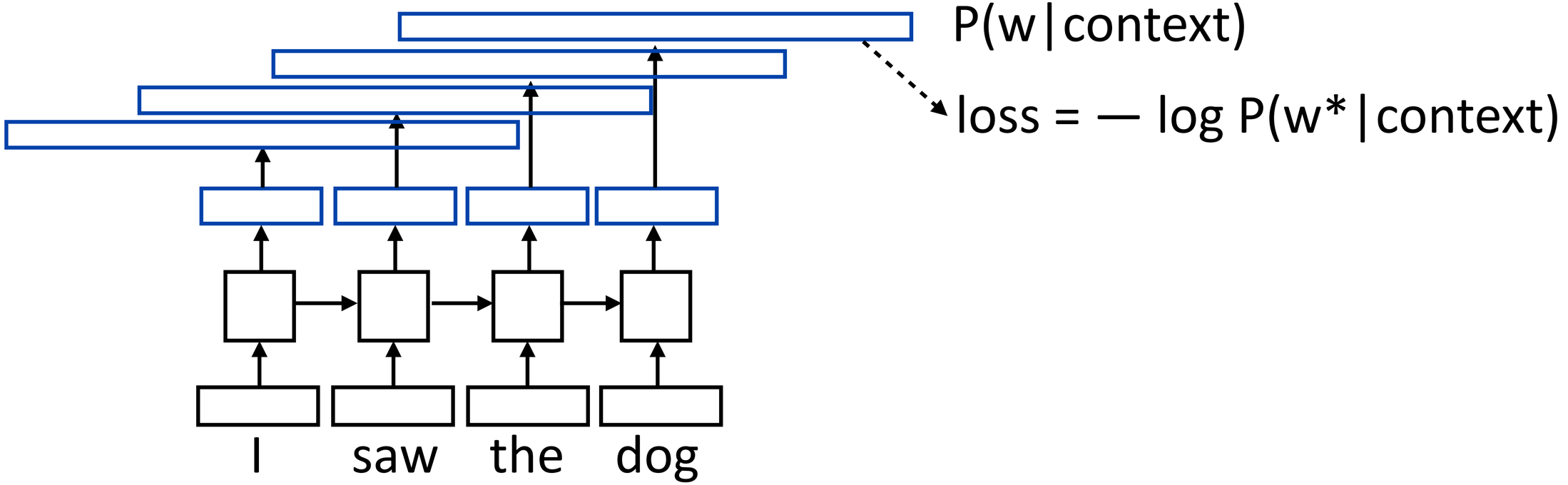
- W is a (vocab size) x (hidden size) matrix

Training RNNLMs



- Input is a sequence of words, output is those words shifted by one,
- Allows us to efficiently batch up training across time (one run of the RNN)

Training RNNLMs

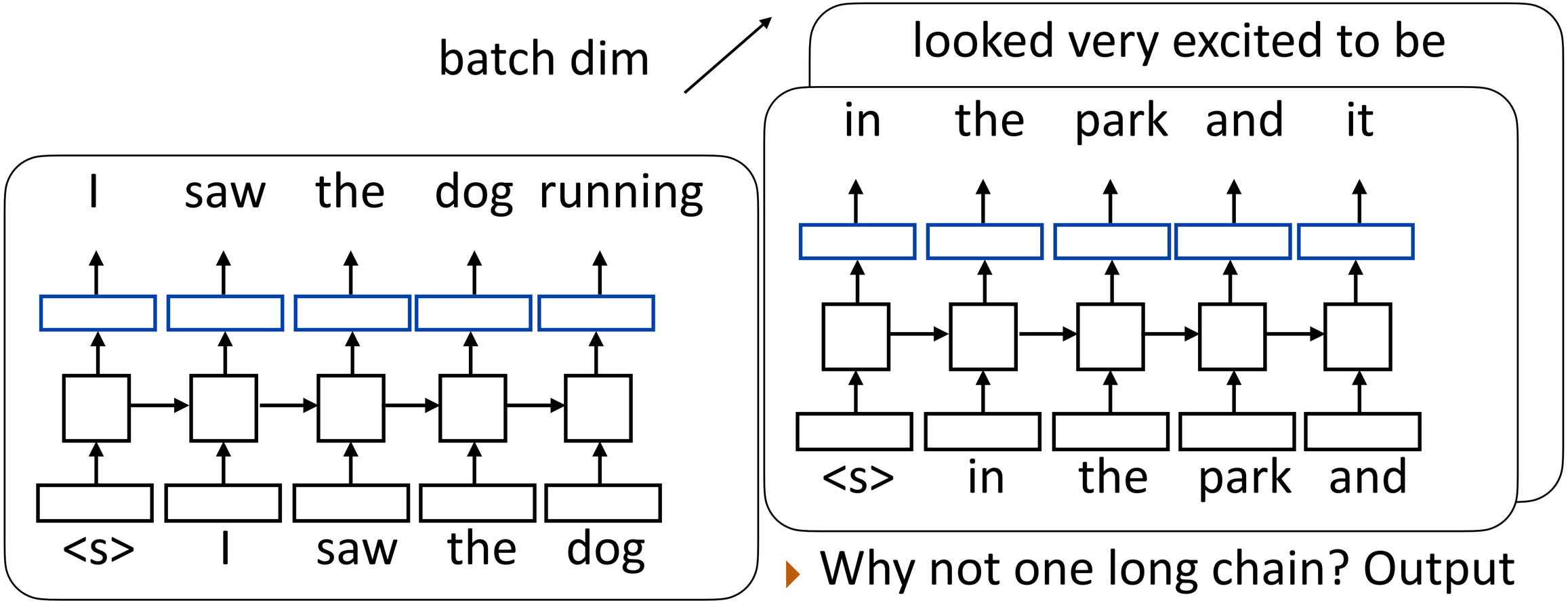


- Total loss = sum of negative log likelihoods at each position
- Backpropagate through the network to simultaneously learn to predict next word given previous words at all positions

Batched LM Training

I saw the dog running | in the park and it | looked very excited to be | there

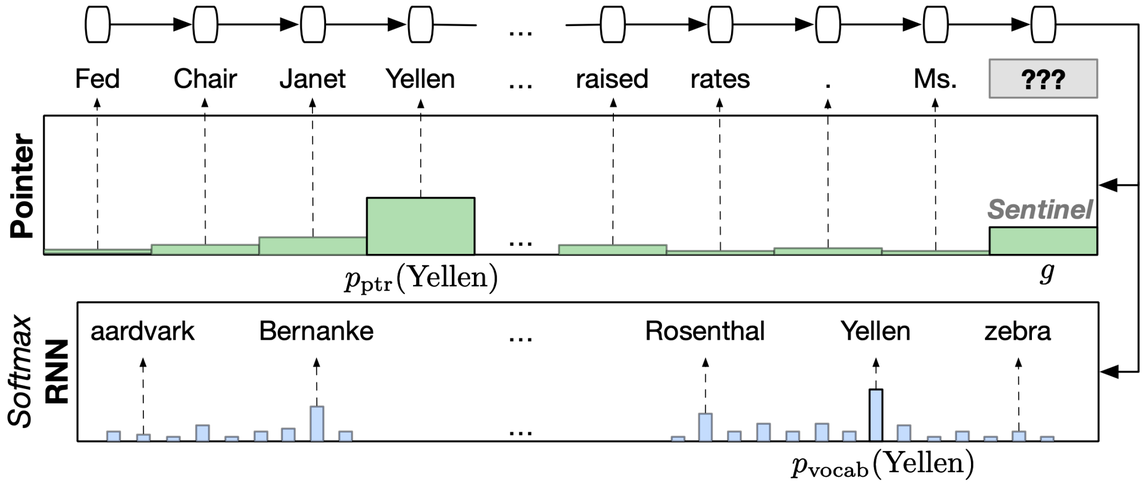
batch dim



- Why not one long chain? Output depends on previous timesteps

Limitations of LSTM LMs

- Need some kind of pointing mechanism to repeat recent words



$$p(\text{Yellen}) = g\, p_{\text{vocab}}(\text{Yellen}) + (1-g)\, p_{\text{ptr}}(\text{Yellen})$$

Merity et al. (2016)

- Transformers can do this (will discuss later in the course)
- SOTA: GPT-3

  PTB (dataset) perplexity = 65.85 with 117M params
  - => 35.76 W/ 1542M params
  - => 20.5 W/ 175B params

Applications of Language Modeling

- All generation tasks: translation, dialogue, text simplification, paraphrasing, etc.
- Grammatical error correction
- Predictive text
- Pretraining!

Pretraining / ELMo

Recall: Context-dependent Embeddings

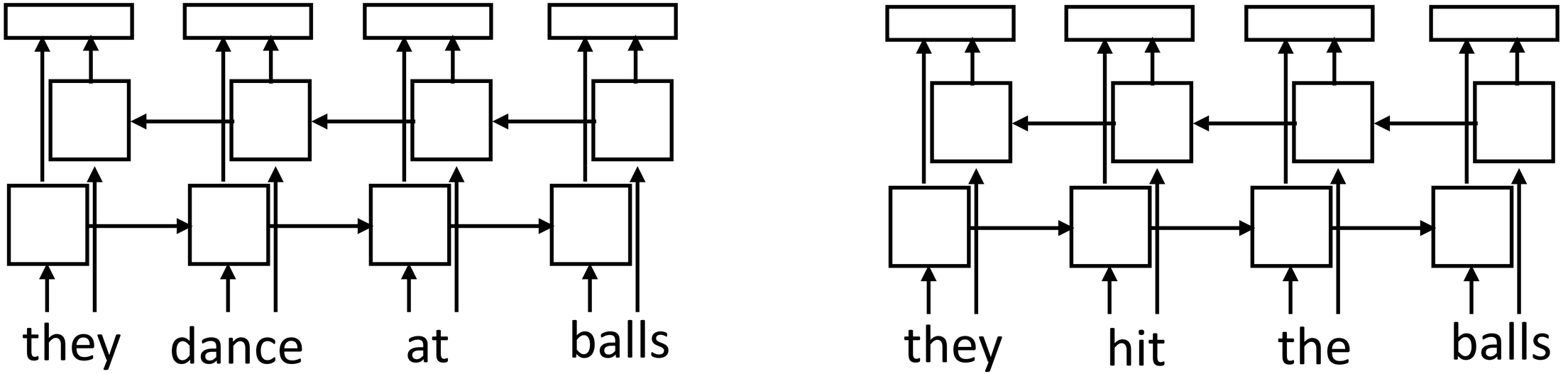
- How to handle different word senses? One vector for *balls*

they dance at balls

they hit the balls

Recall: Context-dependent Embeddings

- How to handle different word senses? One vector for *balls*



- Train a neural language model (from each direction) to predict the next word given previous words in the sentence, and use its internal representations as word vectors

ELMo: Deep contextualized word representations Peters et al. (2018)

ELMo

- Key idea: language models can allow us to form useful word representations in the same way word2vec did
- Take a powerful language model, train it on large amounts of data, then use those representations in downstream tasks
  - Data: Wikipedia and Toronto Books Corpus

ELMo: Deep contextualized word representations Peters et al. (2018)

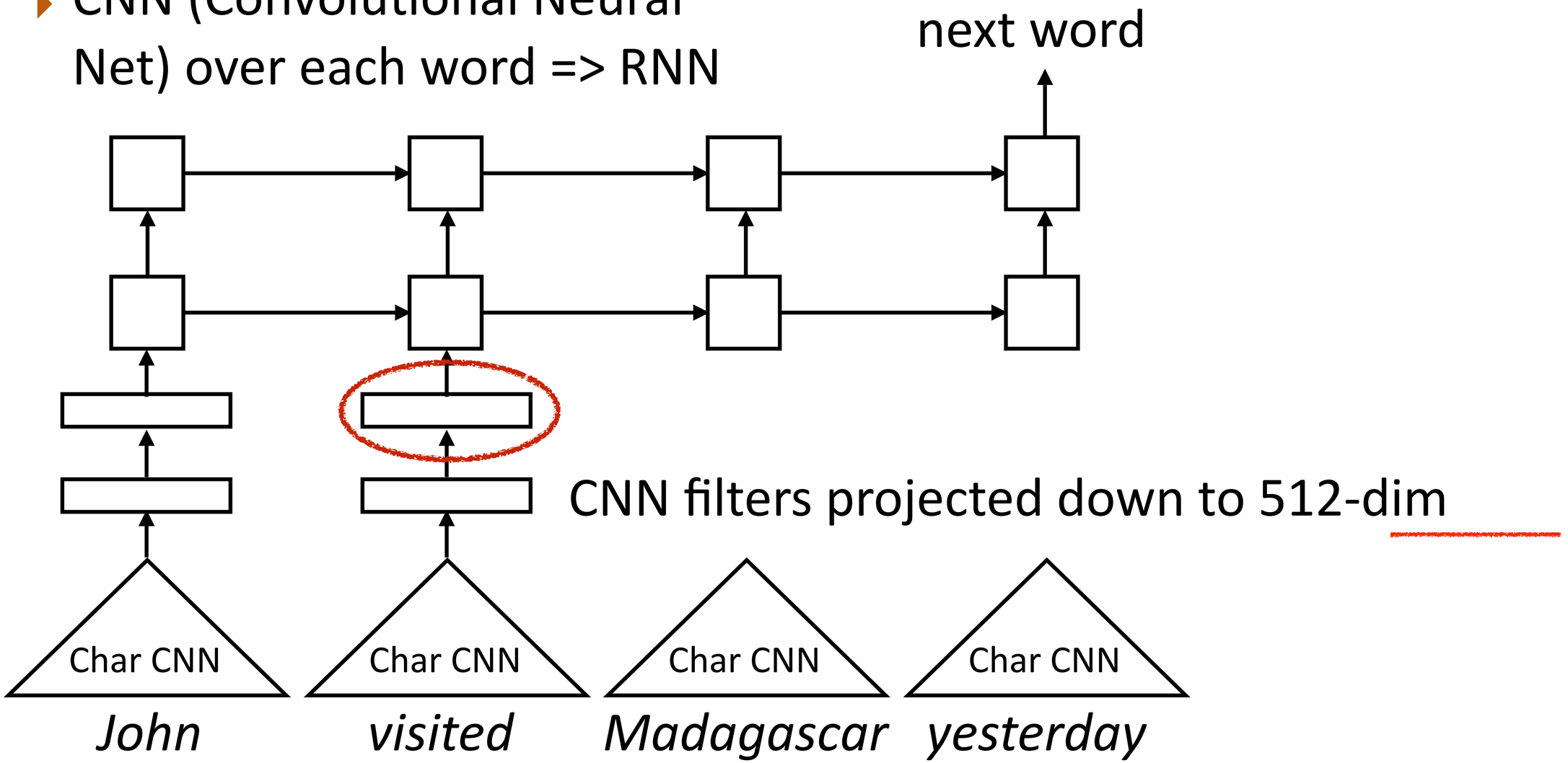
ELMo

- Key idea: language models can allow us to form useful word representations in the same way word2vec did
- Take a powerful language model, train it on large amounts of data, then use those representations in downstream tasks
  - Data: Wikipedia and Toronto Books Corpus
- What do we want our LM to look like?

ELMo: Deep contextualized word representations Peters et al. (2018)

ELMo

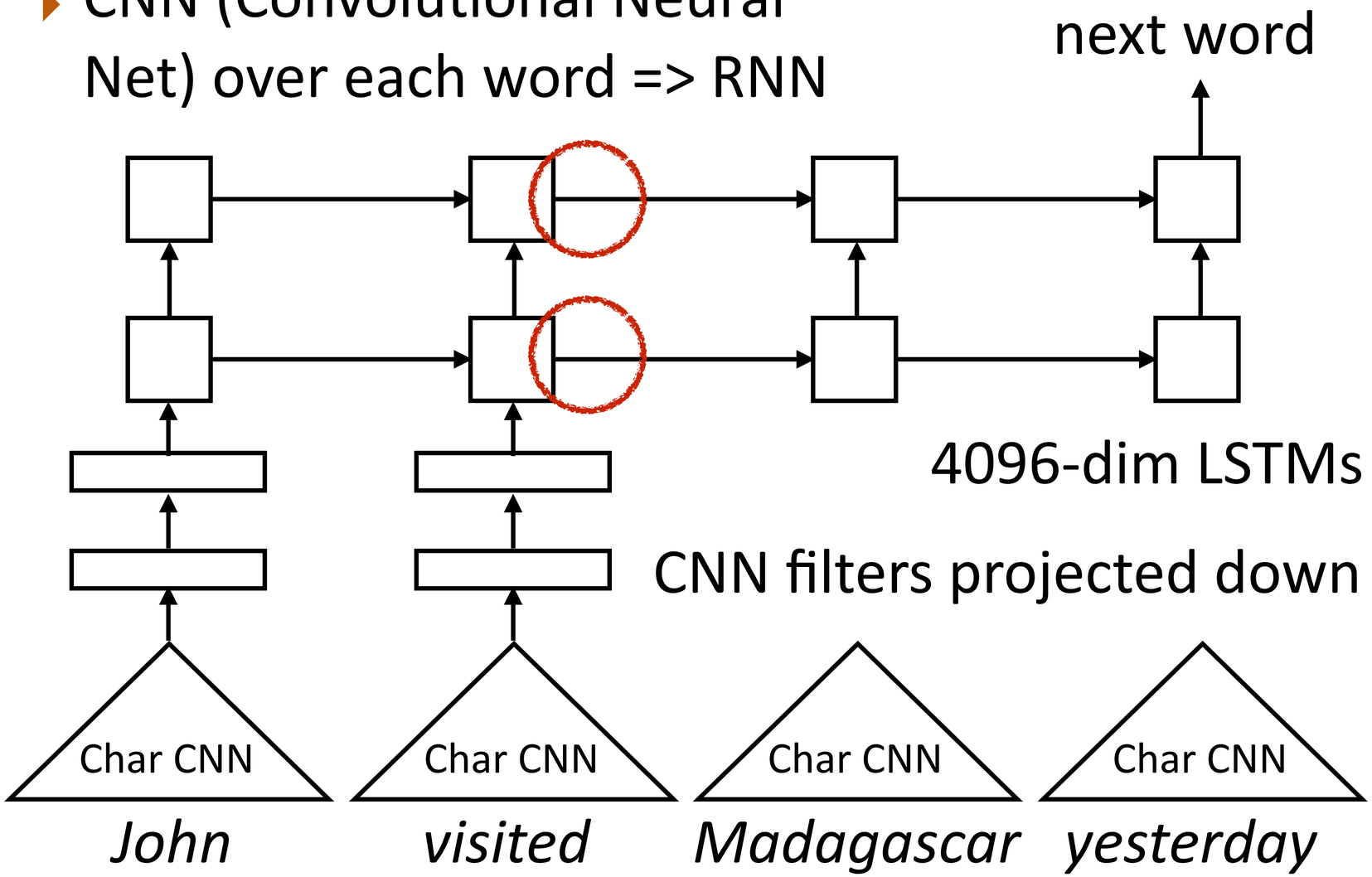
- CNN (Convolutional Neural Net) over each word => RNN



Peters et al. (2018)

ELMo

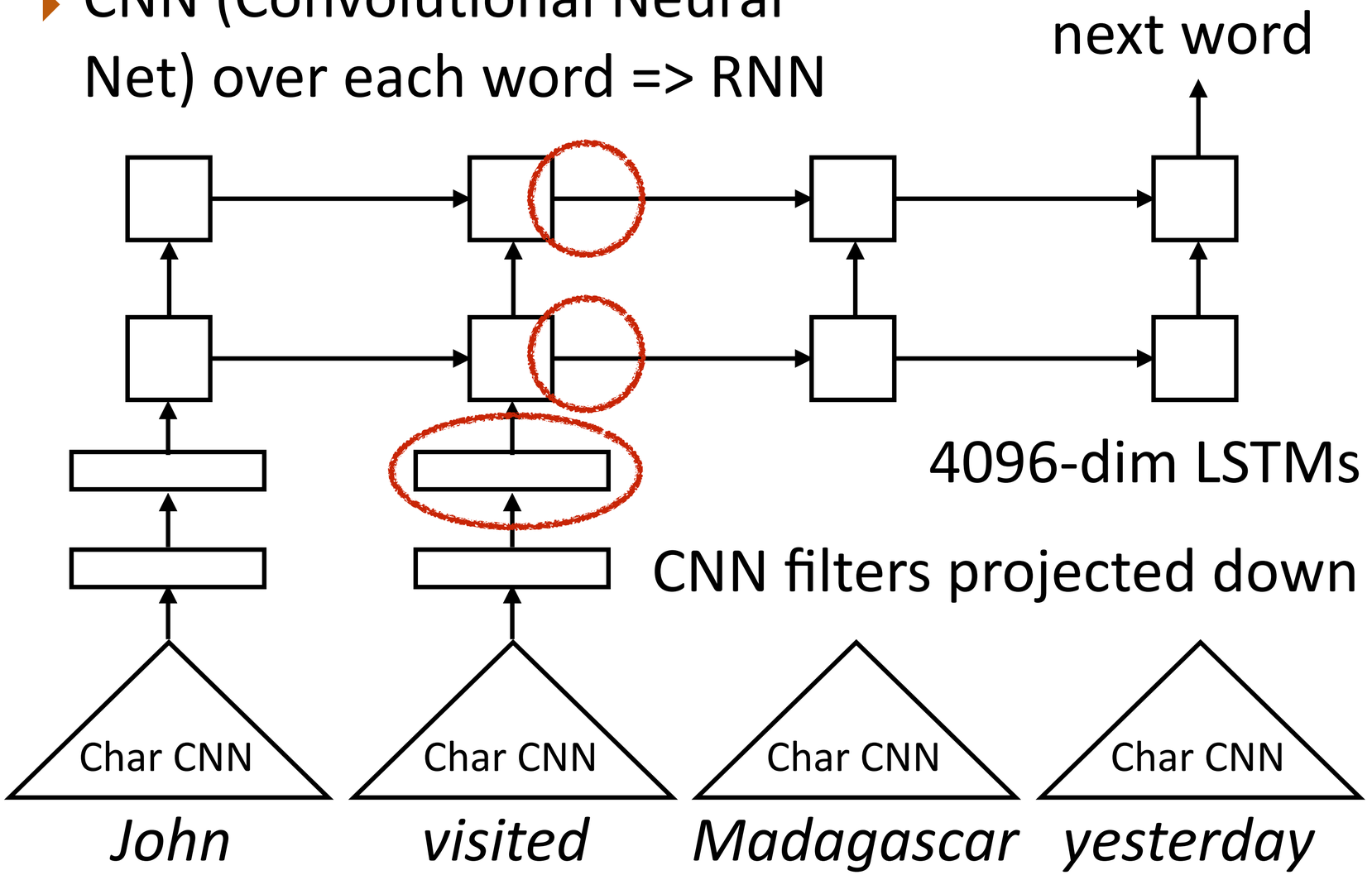
- CNN (Convolutional Neural Net) over each word => RNN



Peters et al. (2018)

ELMo

- CNN (Convolutional Neural Net) over each word => RNN



Representation of *visited* (plus vectors from backward LM)

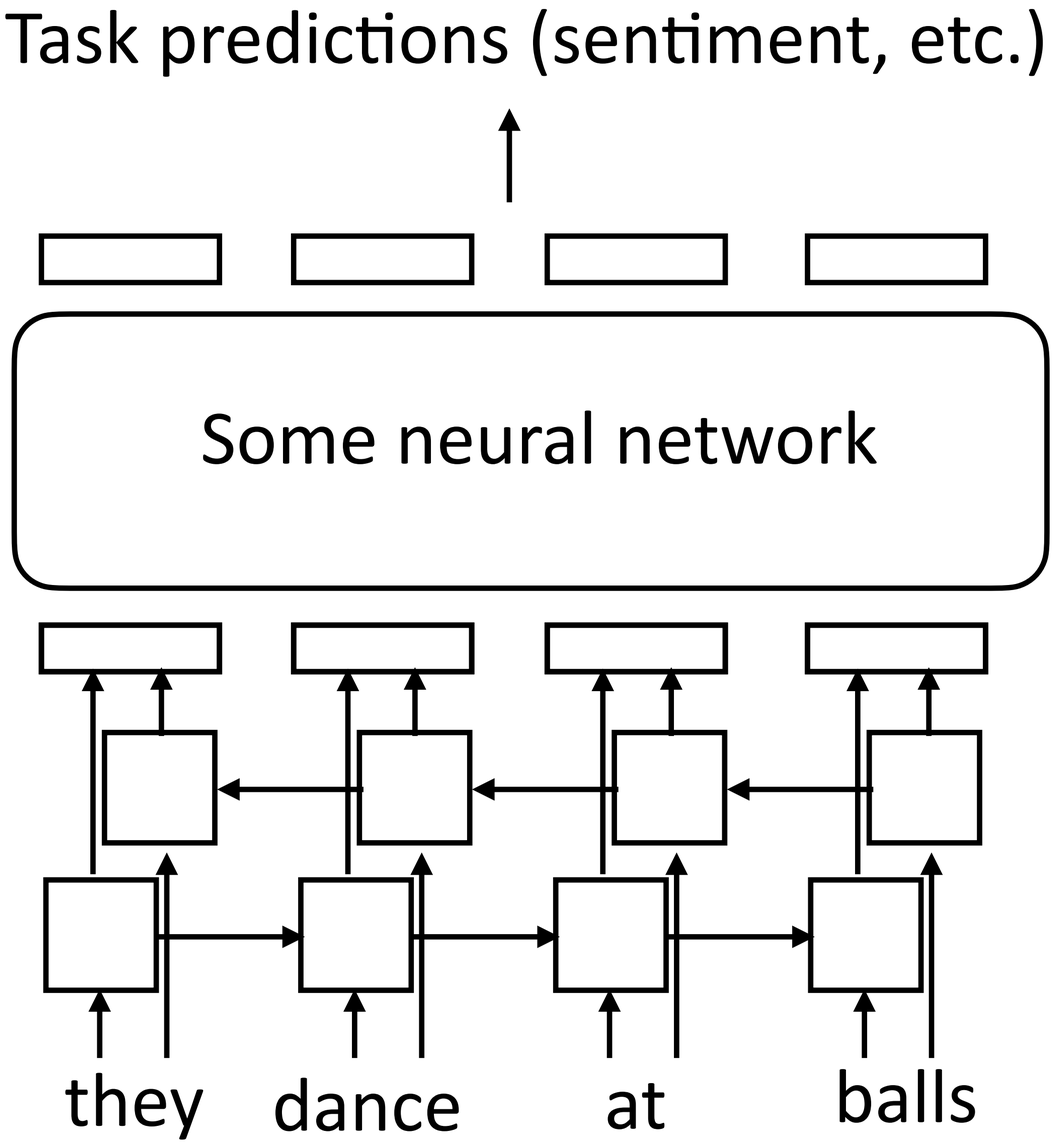
4096-dim LSTMs w/ 512-dim projections

CNN filters projected down to 512-dim

Peters et al. (2018)

How to apply ELMo?

- Take those embeddings and feed them into whatever architecture you want to use for your task
- *Frozen* embeddings: update the weights of your network but keep ELMo's parameters frozen
- *Fine-tuning*: backpropagate all the way into ELMo when training your model



Peters, Ruder, Smith (2019)

Results: Frozen ELMo

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	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	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

- Massive improvements across 5 benchmark datasets: question answering, natural language inference, semantic role labeling (discussed later in the course), coreference resolution, named entity recognition, and sentiment analysis

Peters et al. (2018)

Why is language modeling a good objective?

- "Impossible" problem but bigger models seem to do better and better at distributional modeling (no upper limit yet)
- Successfully predicting next words requires modeling lots of different aspects in text

Context: My wife refused to allow me to come to Hong Kong when the plague was at its height and –" "Your wife, Johanne? You are married at last ?" Johanne grinned. "Well, when a man gets to my age, he starts to need a few home comforts.
Target sentence: After my dear mother passed away ten years ago now, I became _____.
Target word: lonely

- LAMBADA dataset (Papernot et al., 2016): explicitly targets world knowledge and very challenging LM examples
- Coreference, Winograd schema, and much more

Why did this take time to catch on?

- Earlier version of ELMo by the same authors in 2017, but it was only evaluated on tagging tasks, gains were 1% or less
- Required: training on lots of data, having the right architecture, significant hyperparameter tuning

Probing ELMo

- From each layer of the ELMo model, attempt to predict something: POS tags, word senses, etc.
- Higher accuracy => ELMo is capturing that thing more strongly

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Table 5: All-words fine grained WSD F_1. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

Table 6: Test set POS tagging accuracies for PTB. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Peters et al. (2018)

Probing ELMo

- From each layer of the ELMo model, attempt to predict something: POS tags, word senses, etc.
- Higher accuracy => ELMo is capturing that thing more strongly

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Table 5: All-words fine grained WSD F_1. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Model	Acc.
Collobert et al. (2011)	97.3
biLM, Second Layer	96.8

Table 6: Test set POS tagging accuracies for PTB. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

These experiments confirm different layers in the biLM represent different types of information and explain why including all biLM layers is important for the highest performance in downstream tasks

Peters et al. (2018)

Takeaways

- Language modeling involves predicting the next word given context. Several techniques to do this, more later in the course
- Learning a neural network to do this induces useful representations for other tasks, similar to word2vec/GloVe
- Next class: seq2seq