Recurrent Neural Networks for Text Data

July 18, 2019

Block 4, Lecture 1
Applied Data Science
MMCi Term 4, 2019

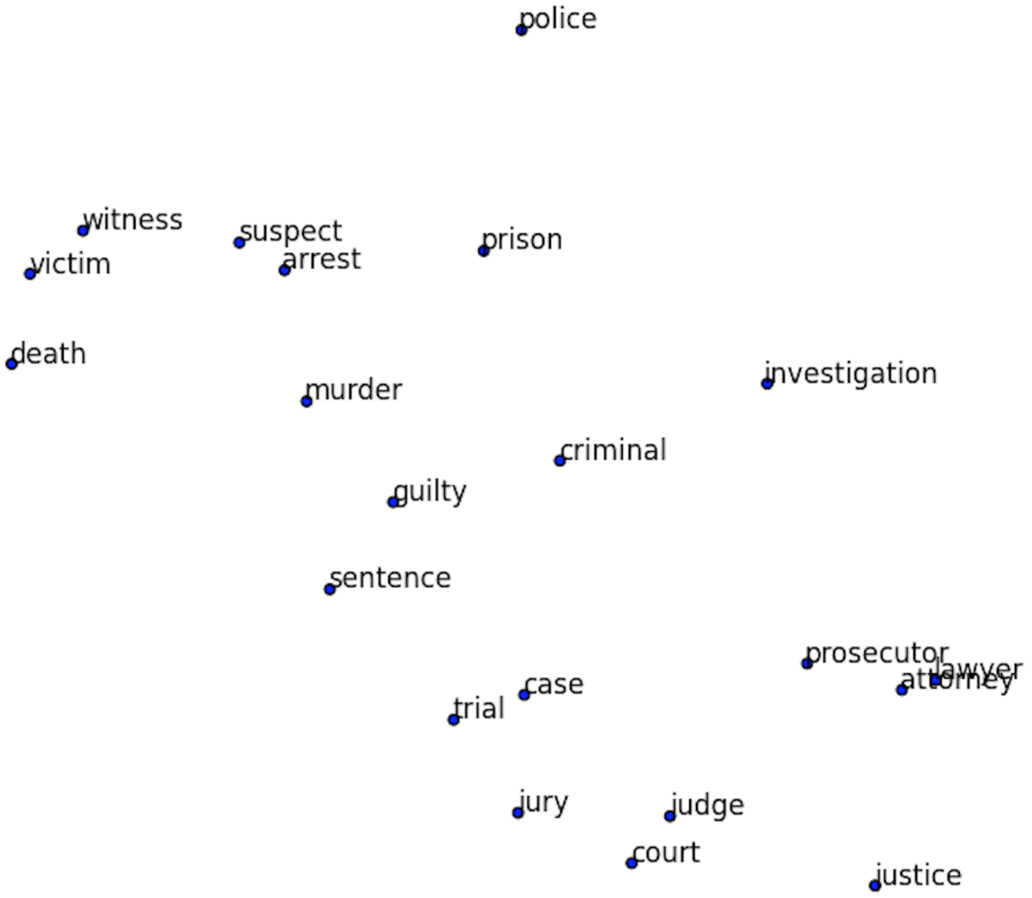
Matthew Engelhard

Recall: Word embeddings allow us to quantify word meaning

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

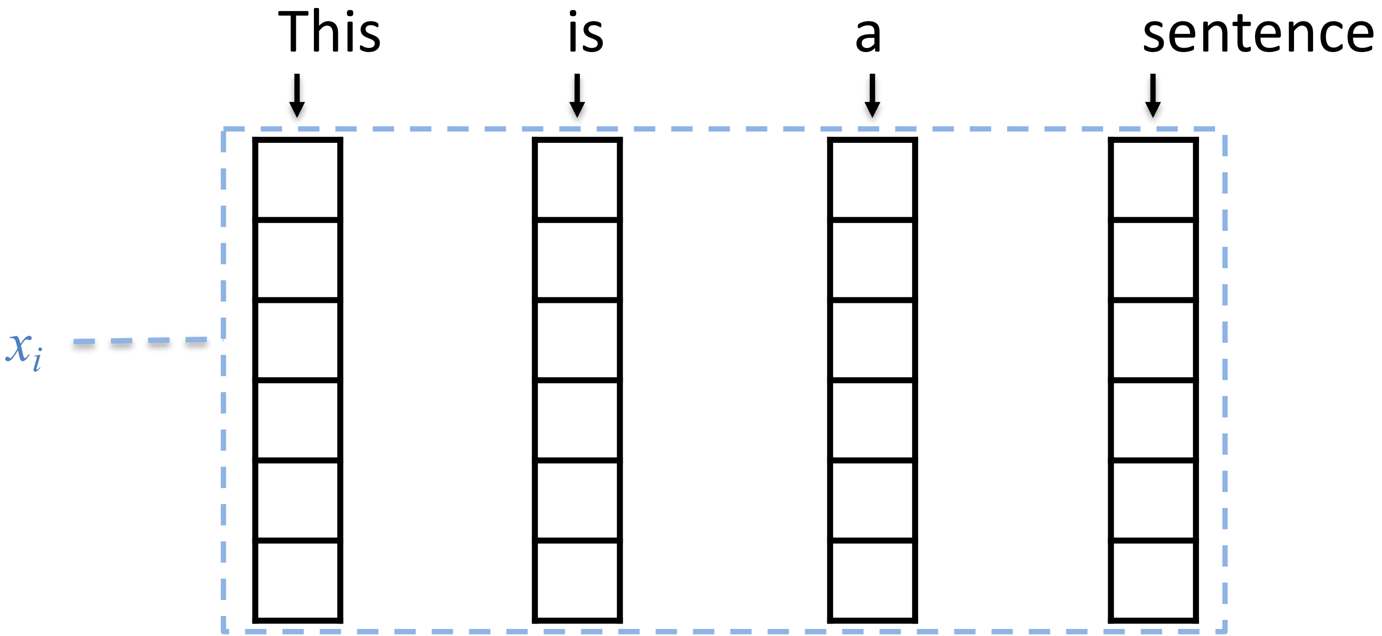
Note the similarity of all the words as a whole, but also of the individual neighbors.

"Lawyer" and "attorney" are nearly identical in space!

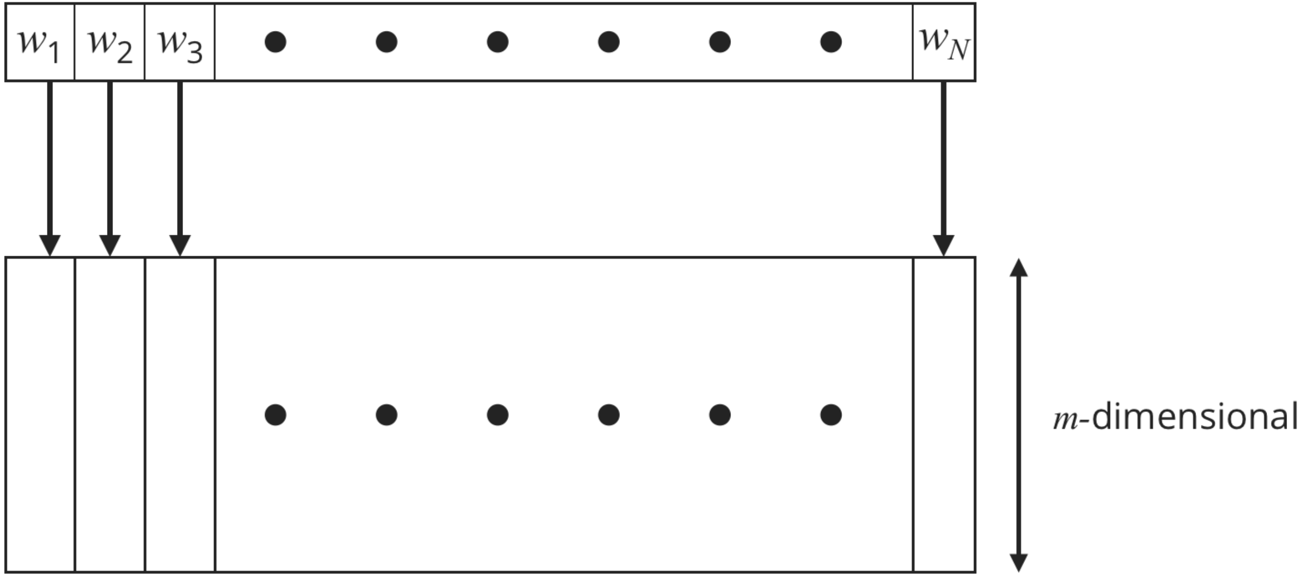


Applying Word Embeddings to a Sentence

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

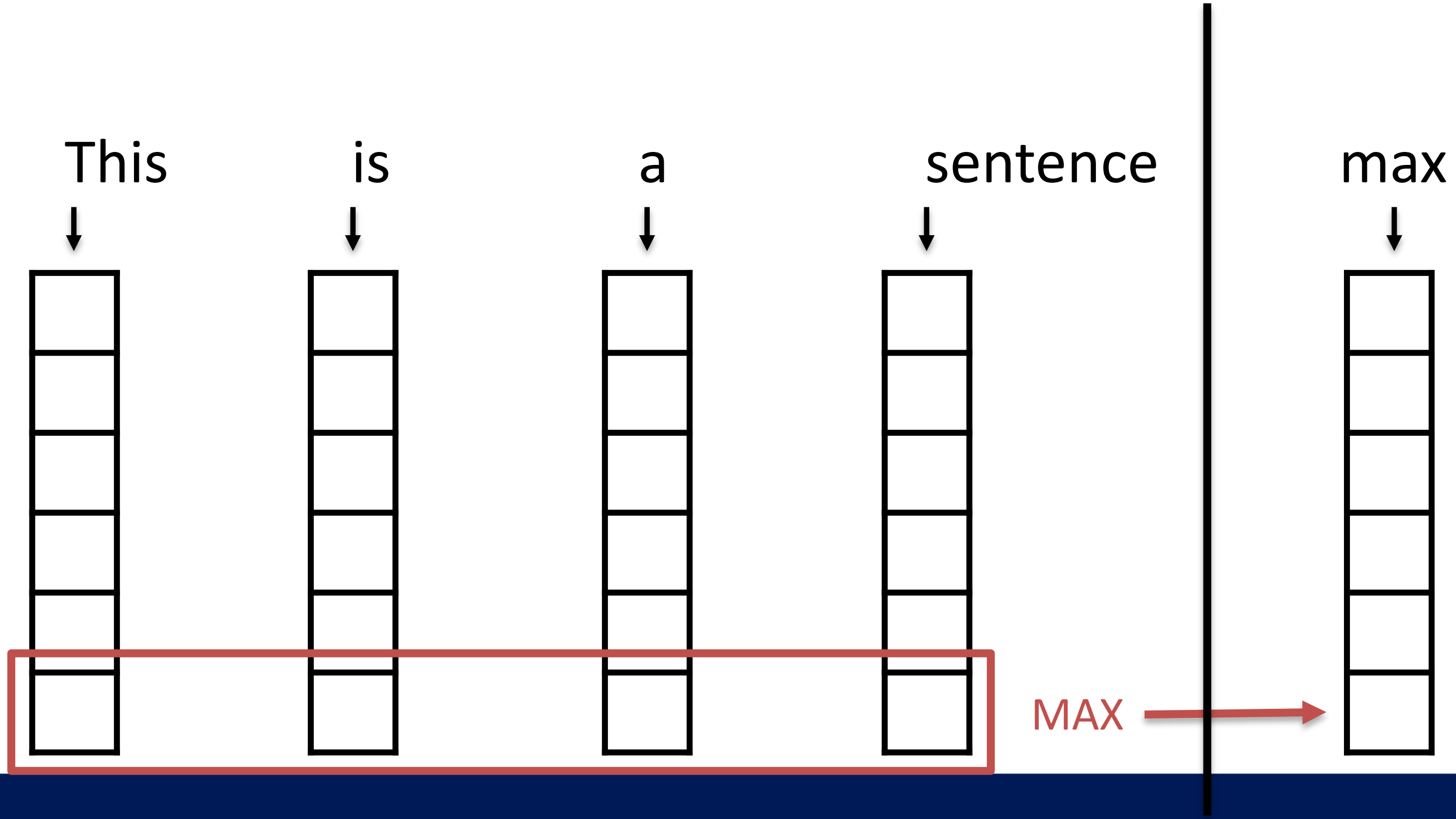


Using Word Embeddings

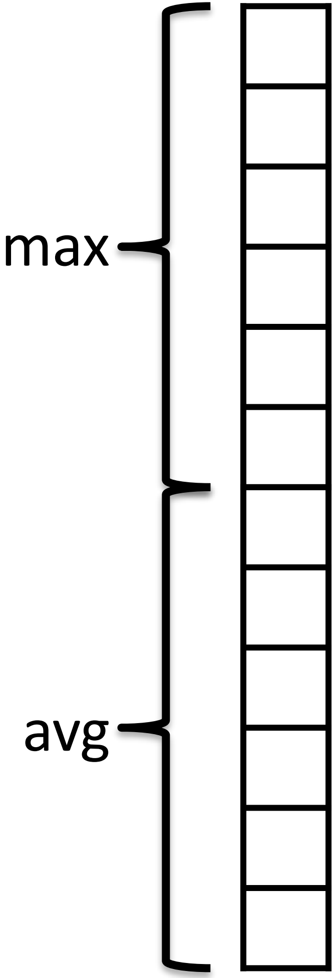


- Our representation depends on the number of words
  - Not a constant number of features!

VSWEM allows us to convert a variable-length sentence to a fixed-length feature vector



Concatenate MAX and AVG



Sentence i → This is a sentence

x_i

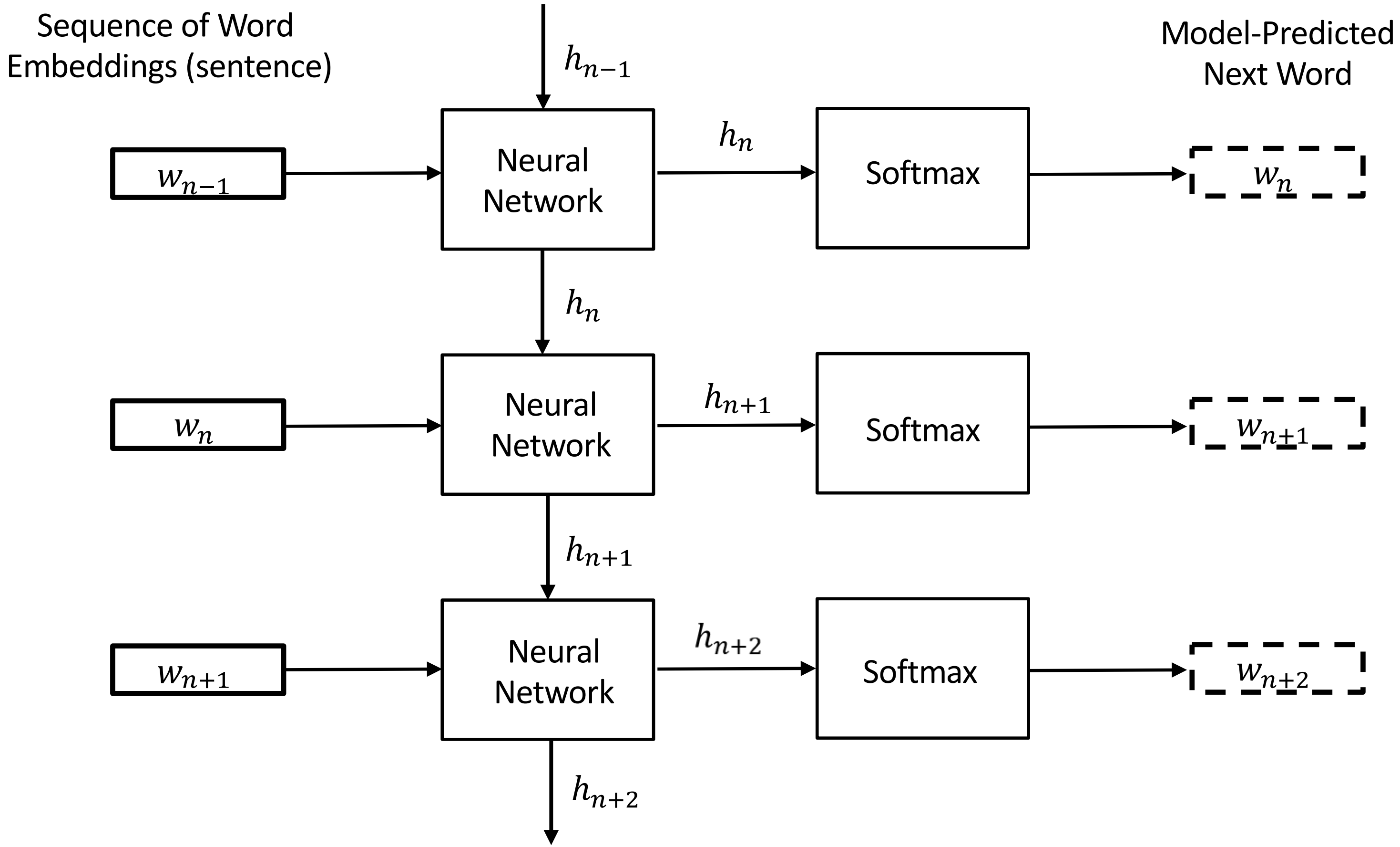
fixed-length feature vector

We’d like a more flexible model…

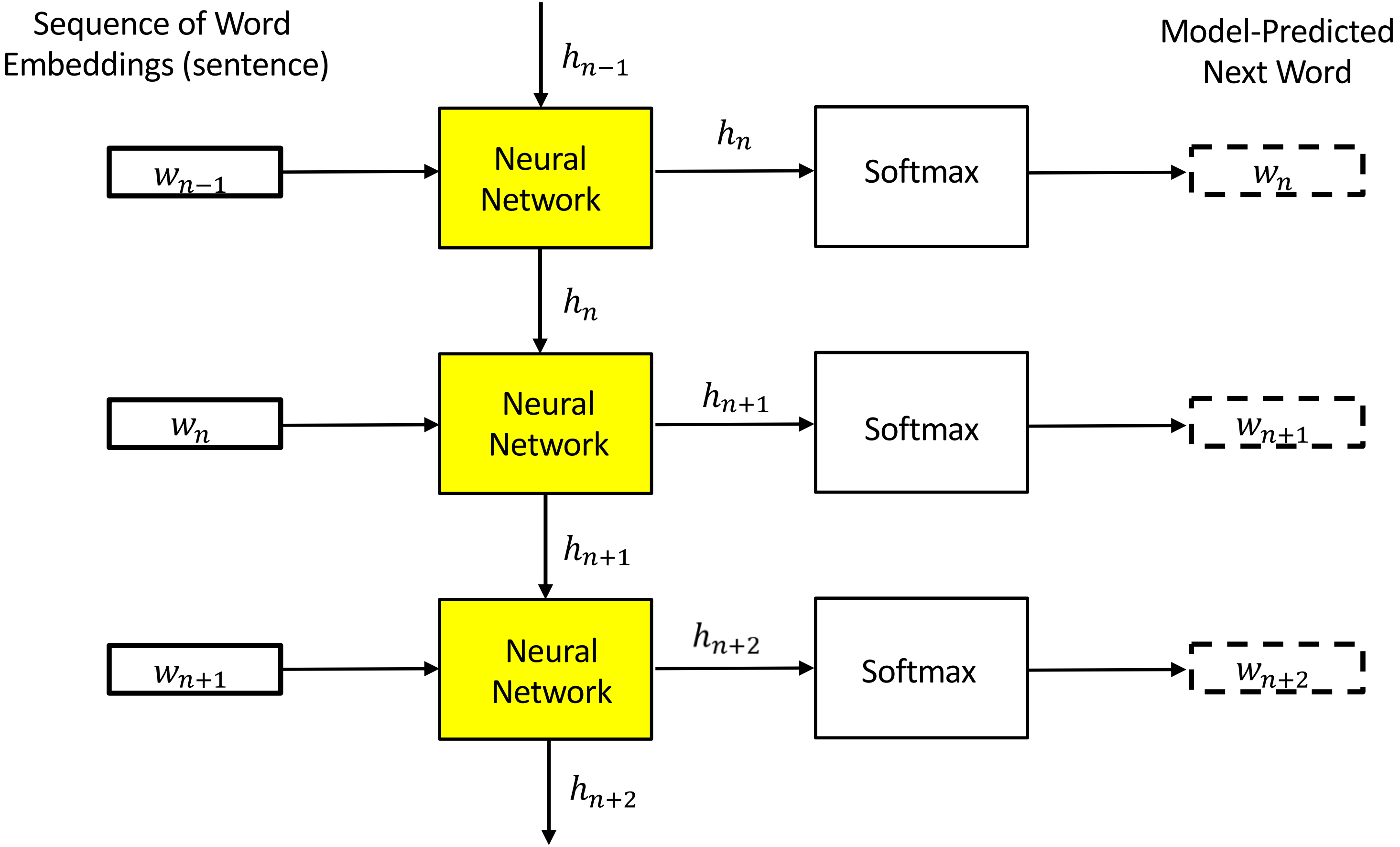
- Interpret words in context (i.e. allow word meaning to be modified by earlier words)
- Make predictions for each word rather than the sentence as a whole (e.g. part of speech tagging, PHI identification)
- Generate text

RECURRENT NEURAL NETWORKS

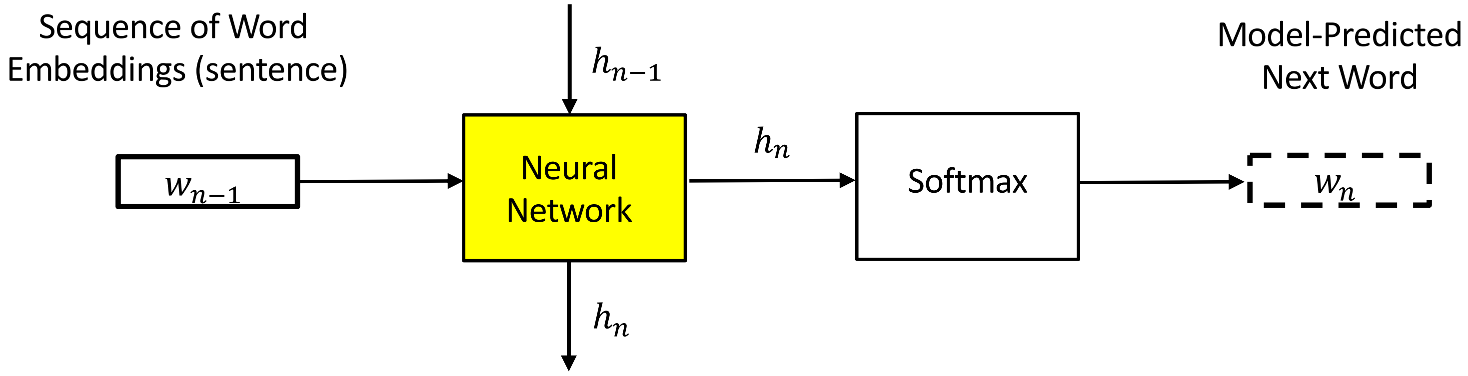
Recurrent Neural Network



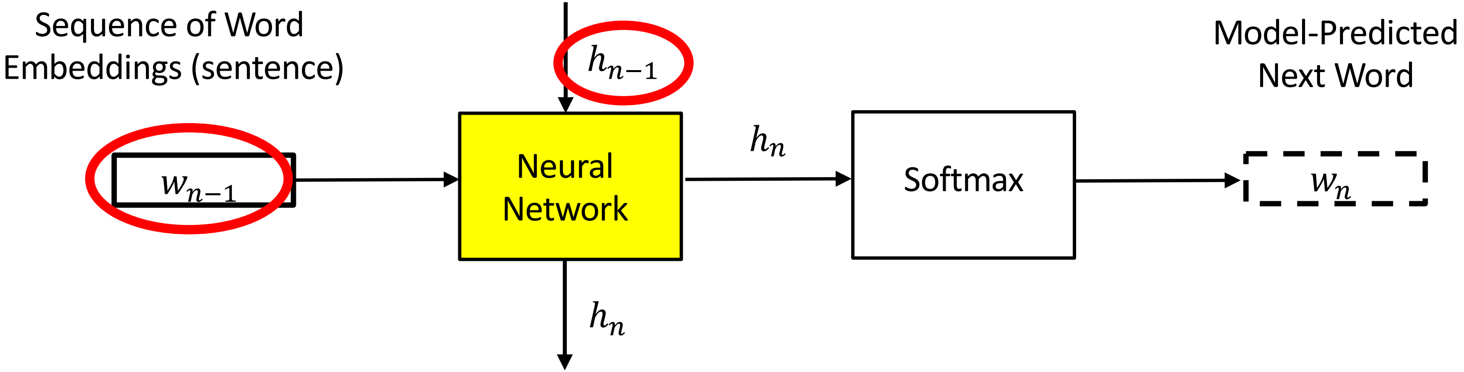
Recurrent Neural Network



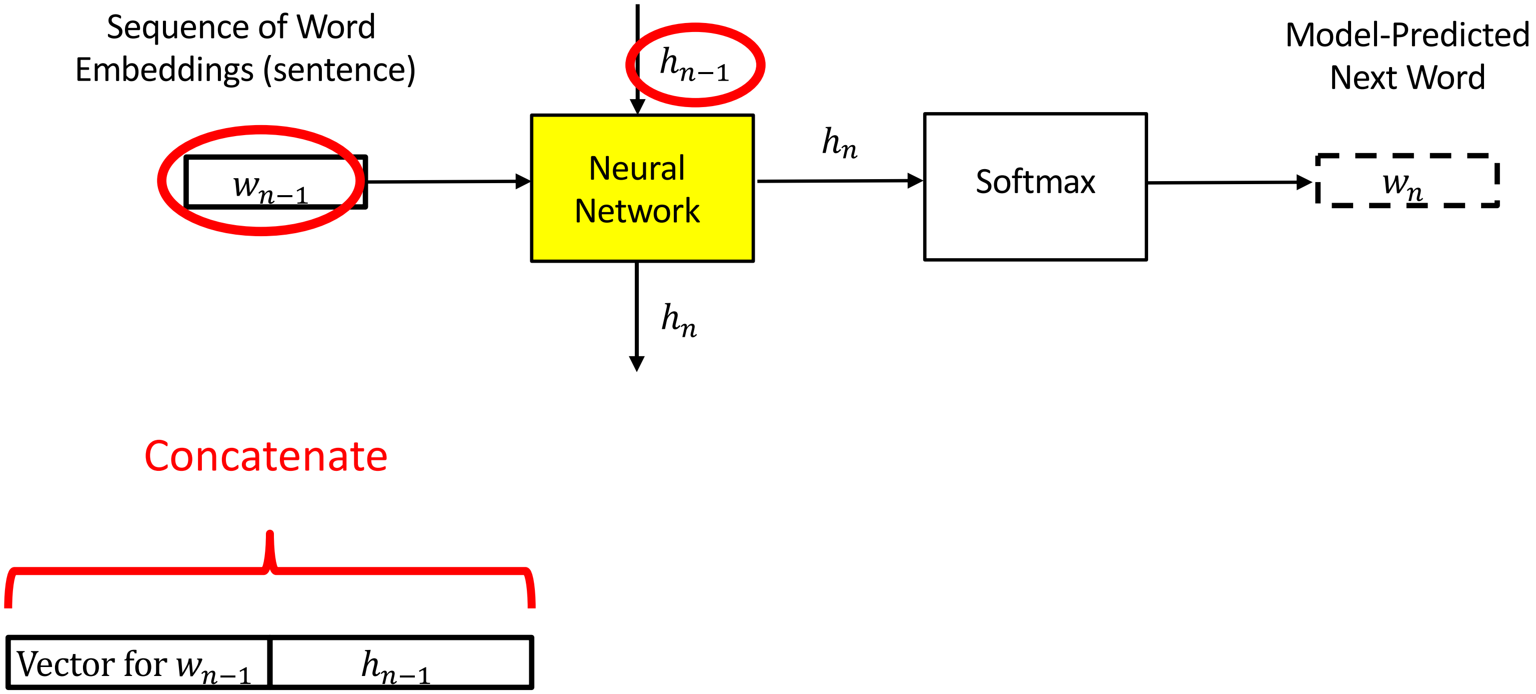
Recurrent Neural Network



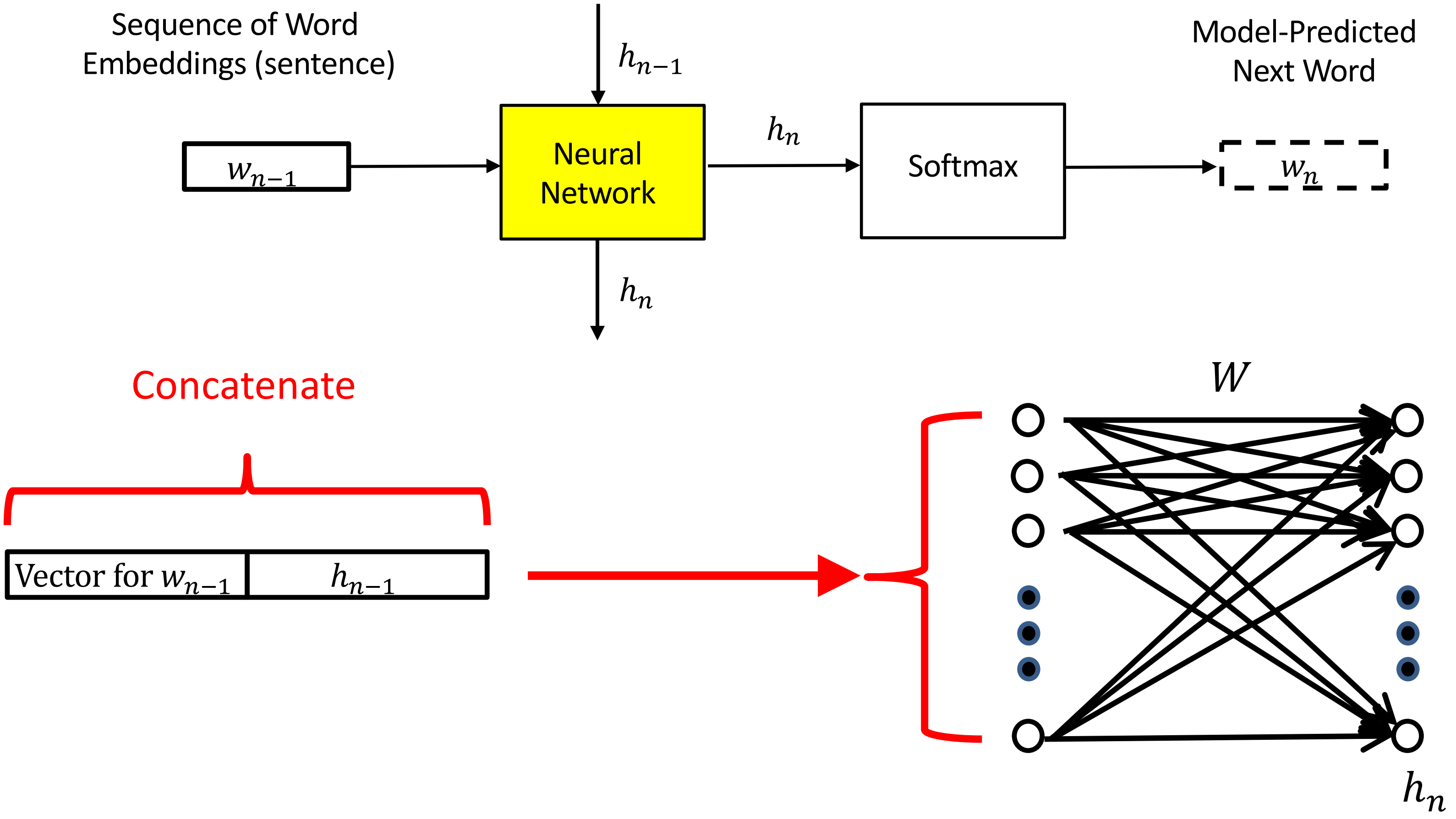
Recurrent Neural Network



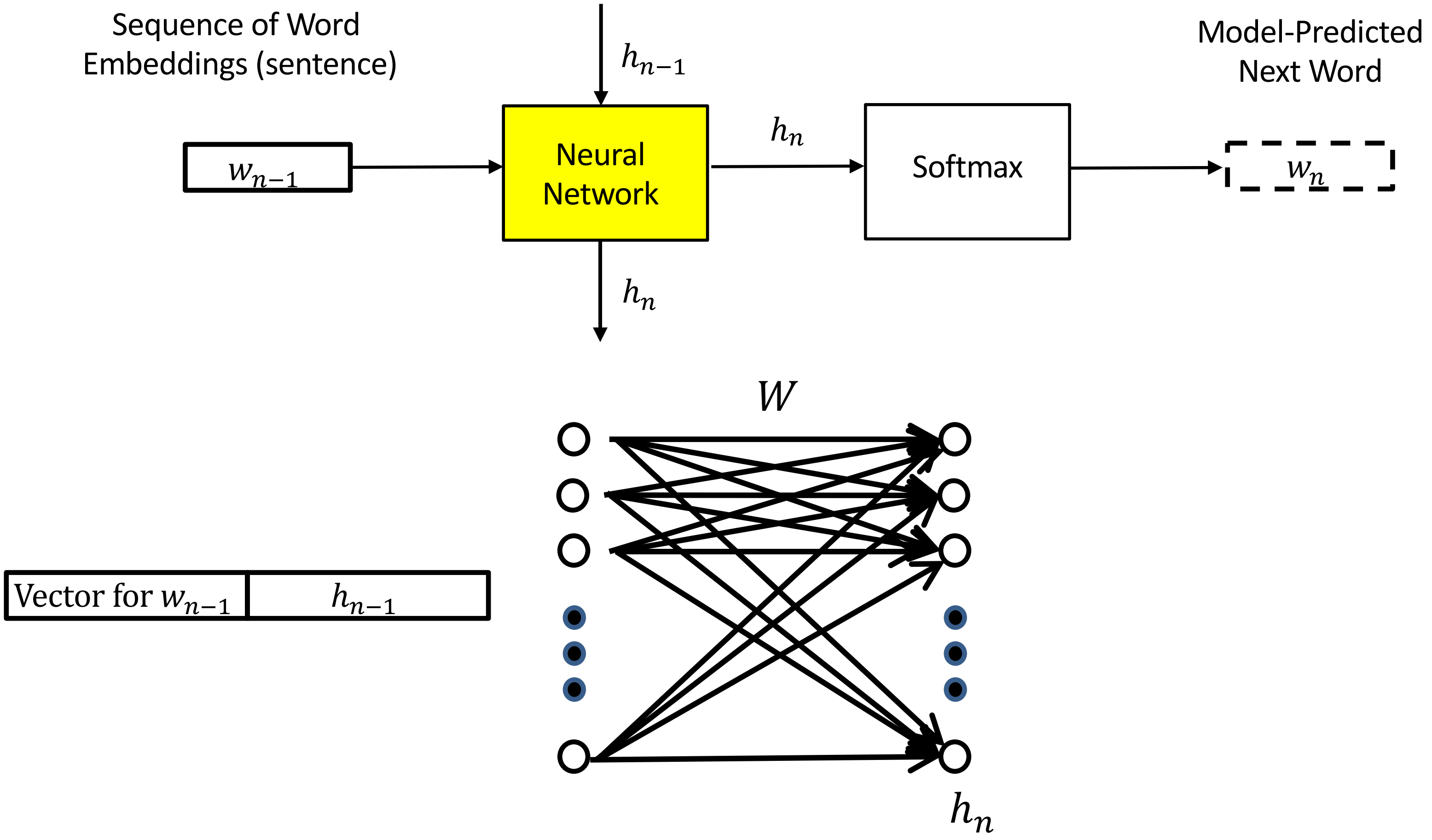
Recurrent Neural Network



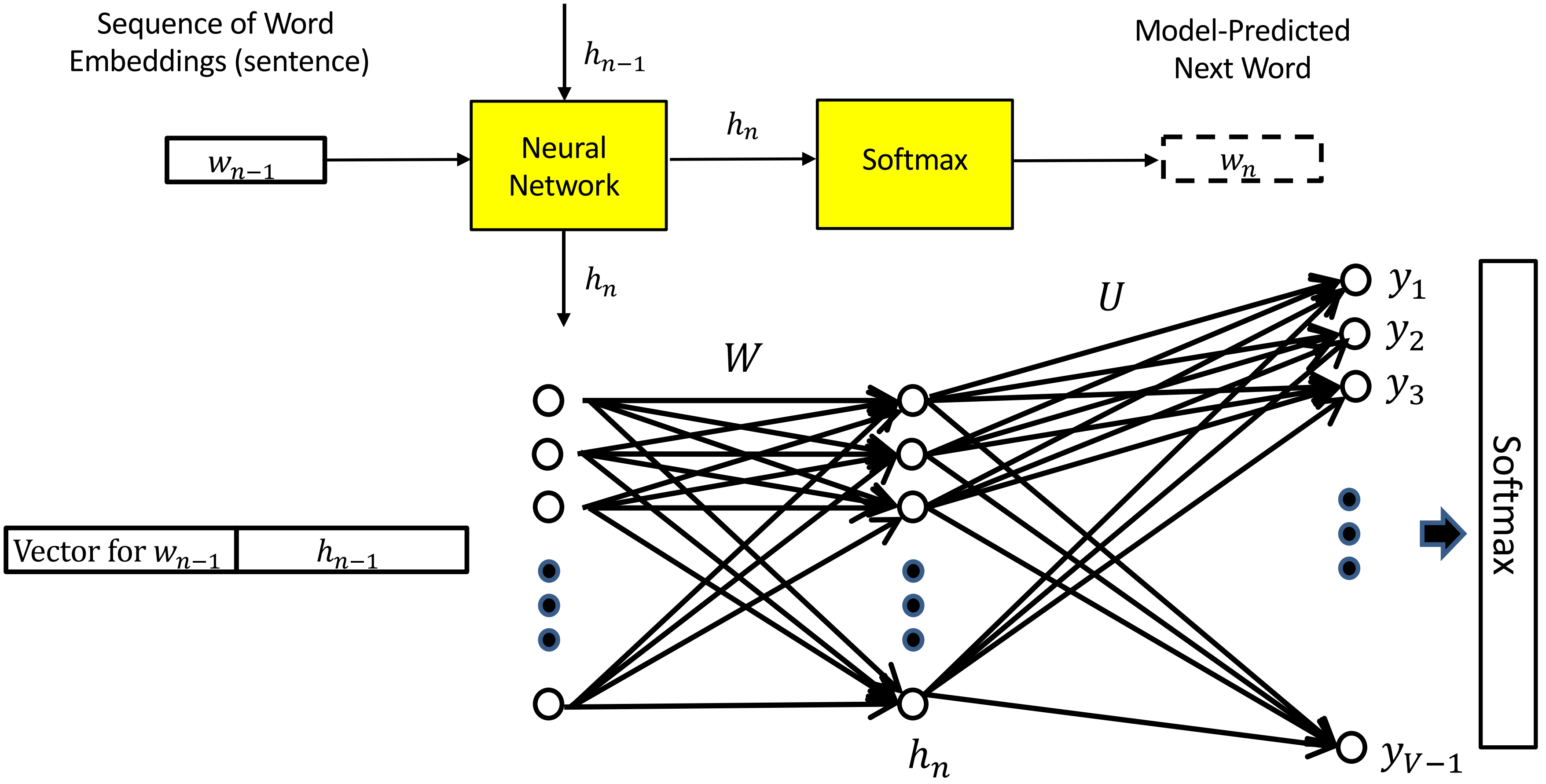
Recurrent Neural Network



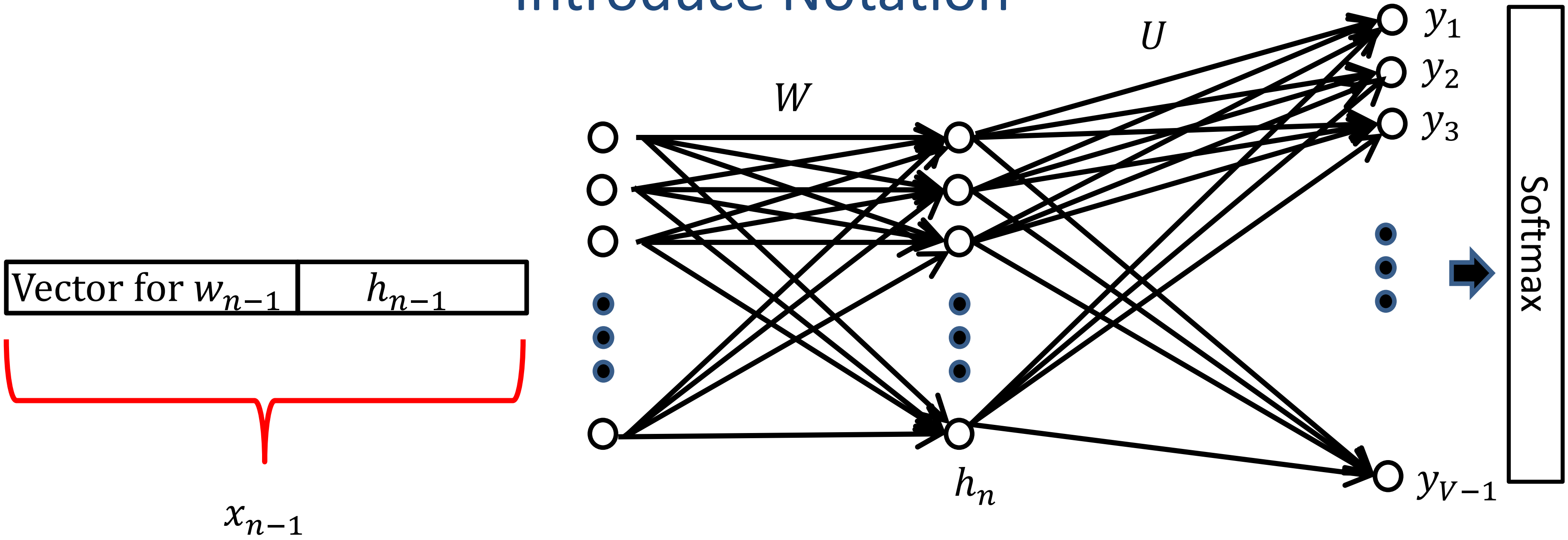
Recurrent Neural Network



Recurrent Neural Network



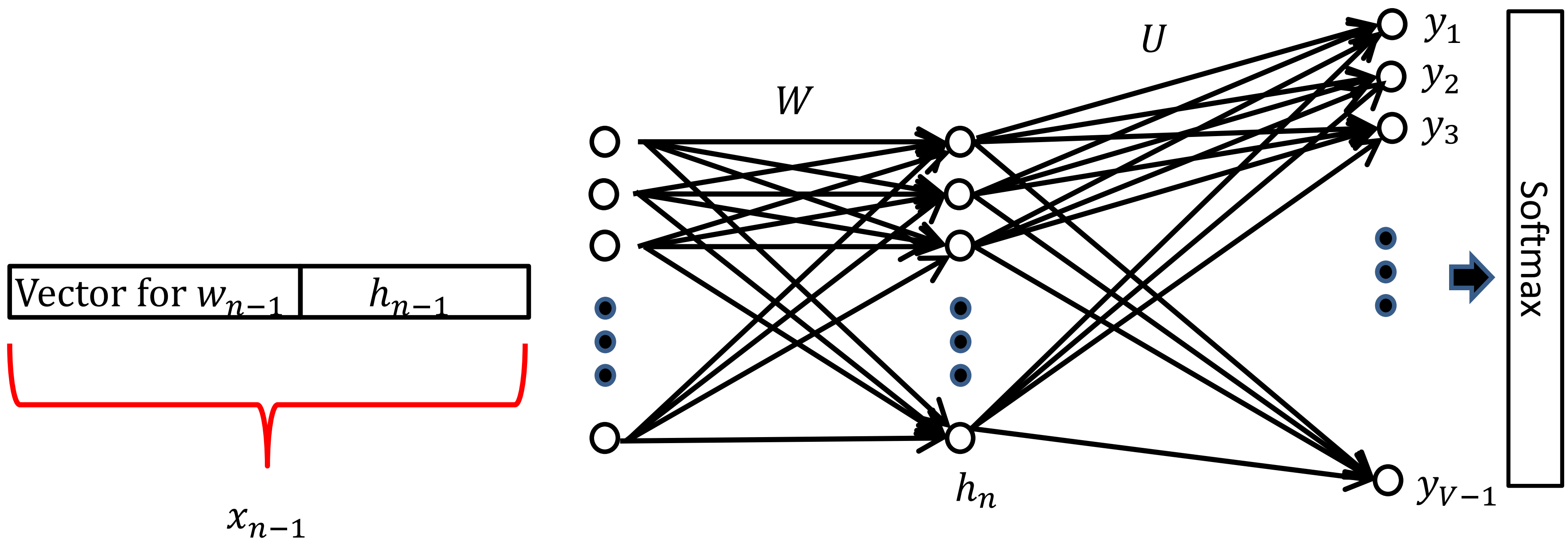
Introduce Notation



$$h_n = \tanh(W \cdot x_{n-1} + b)$$

$$p(w_n | w_{n-1}, h_{n-1}) = \text{softmax}(U \cdot h_n + \beta)$$

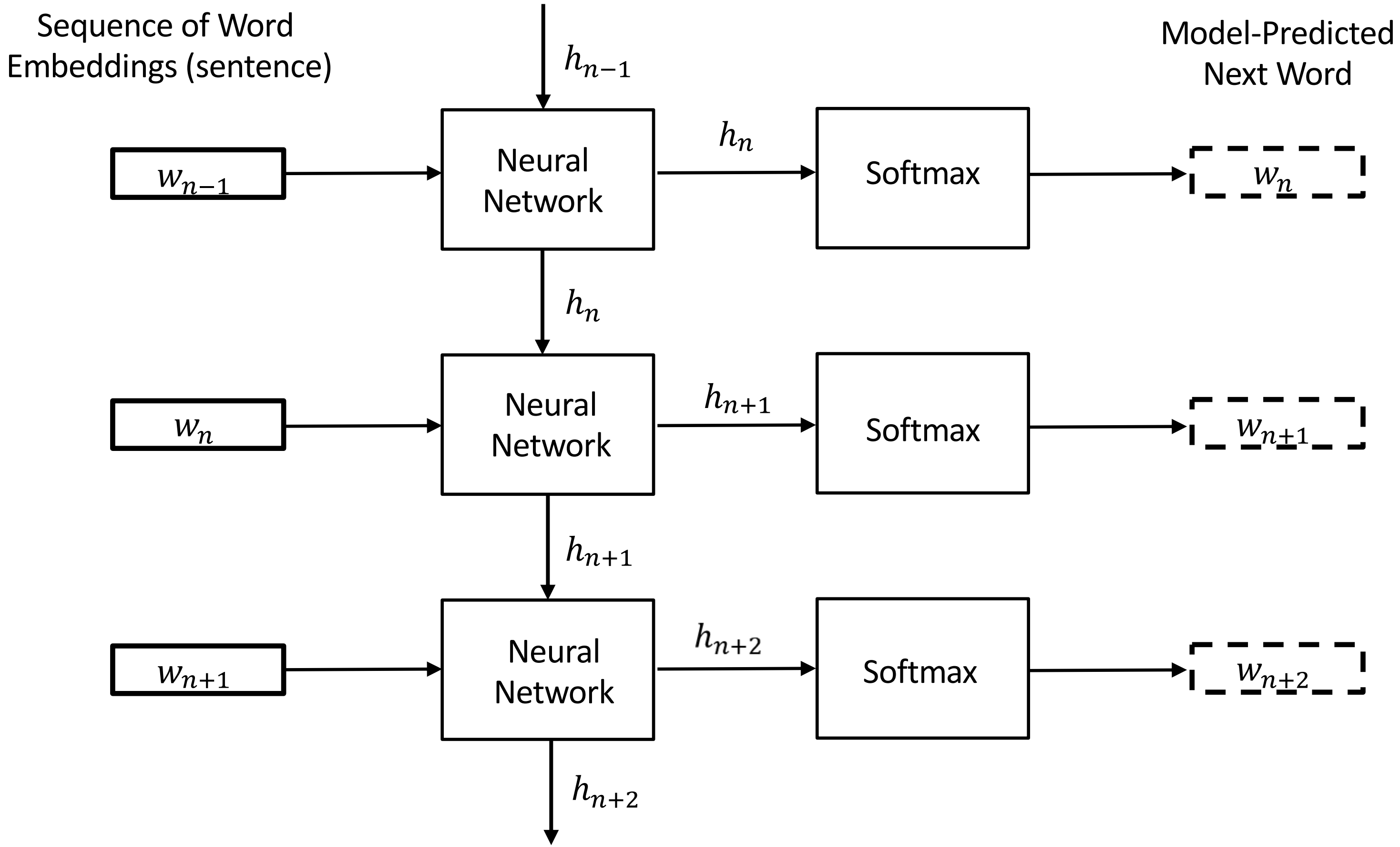
Intuition on Model for Predicting nth Word



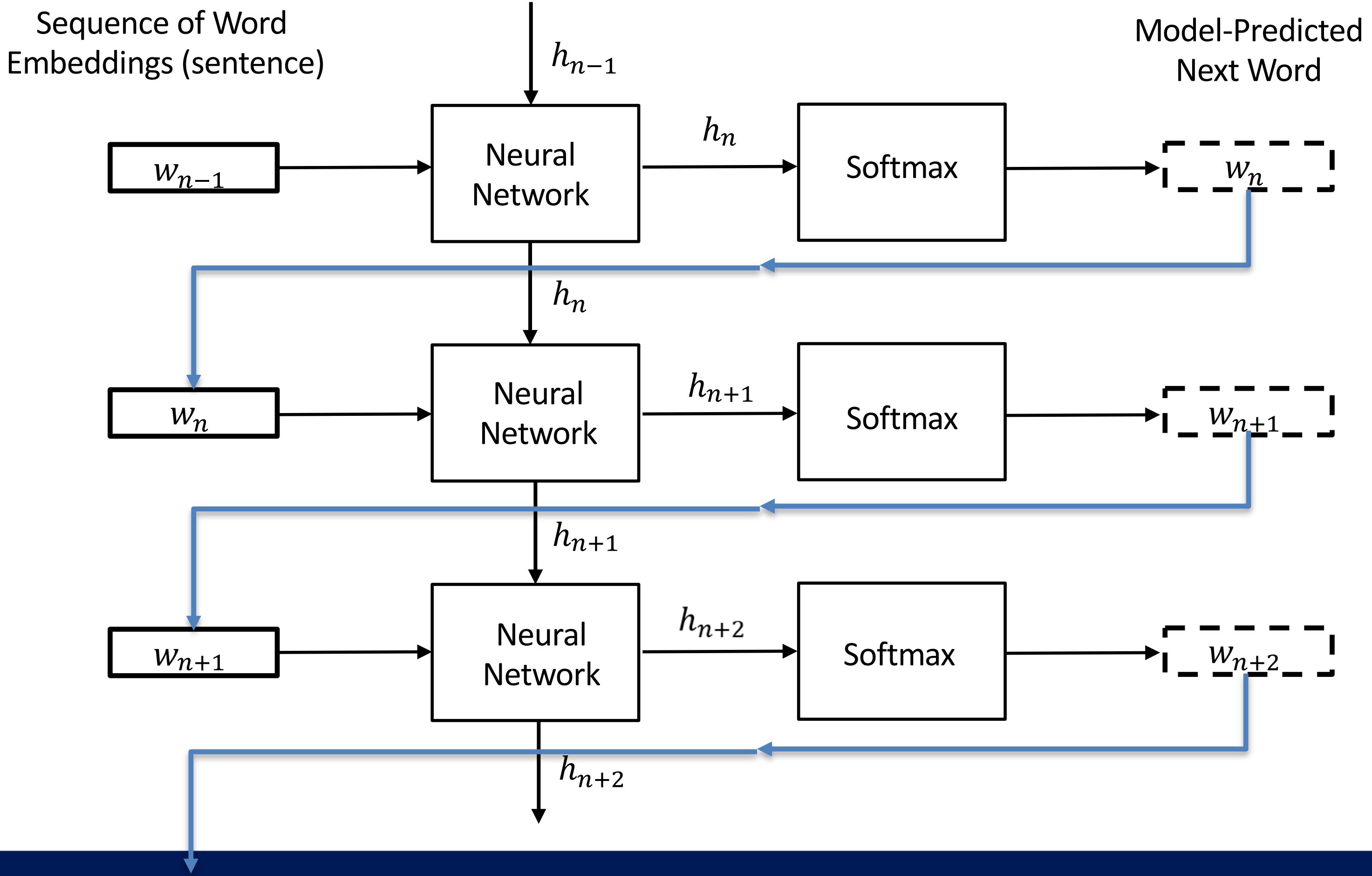
h_{n-1}: Tells us which words were likely prior to selection of previous word (context)

w_{n-1}: Tells us which word was used/selected at point $n-1$ in text, as we predict the nth word

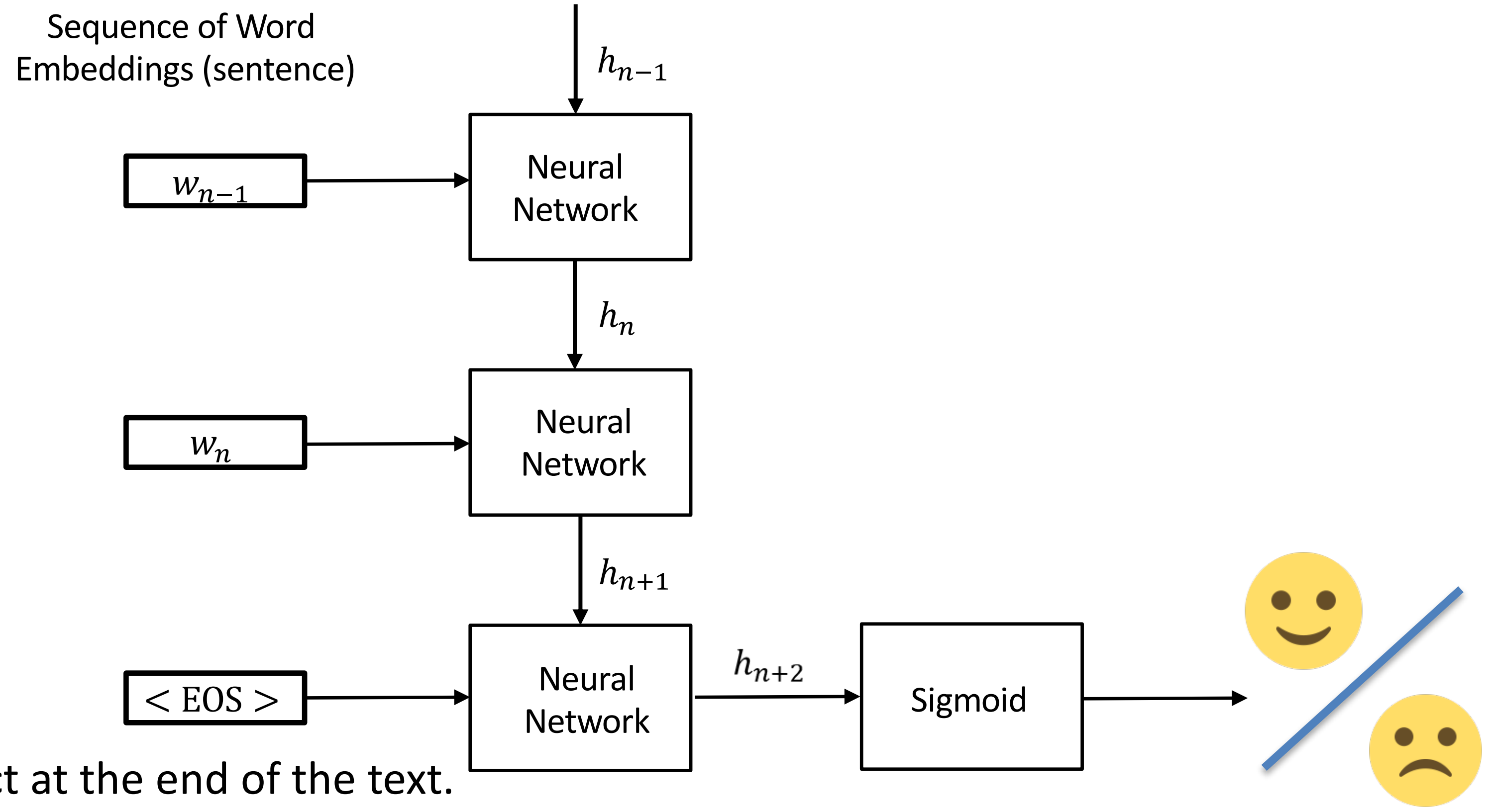
Recurrent Neural Network



Generating Text



Predicting a Single Output



Predict at the end of the text.

Deidentification of Patient Notes

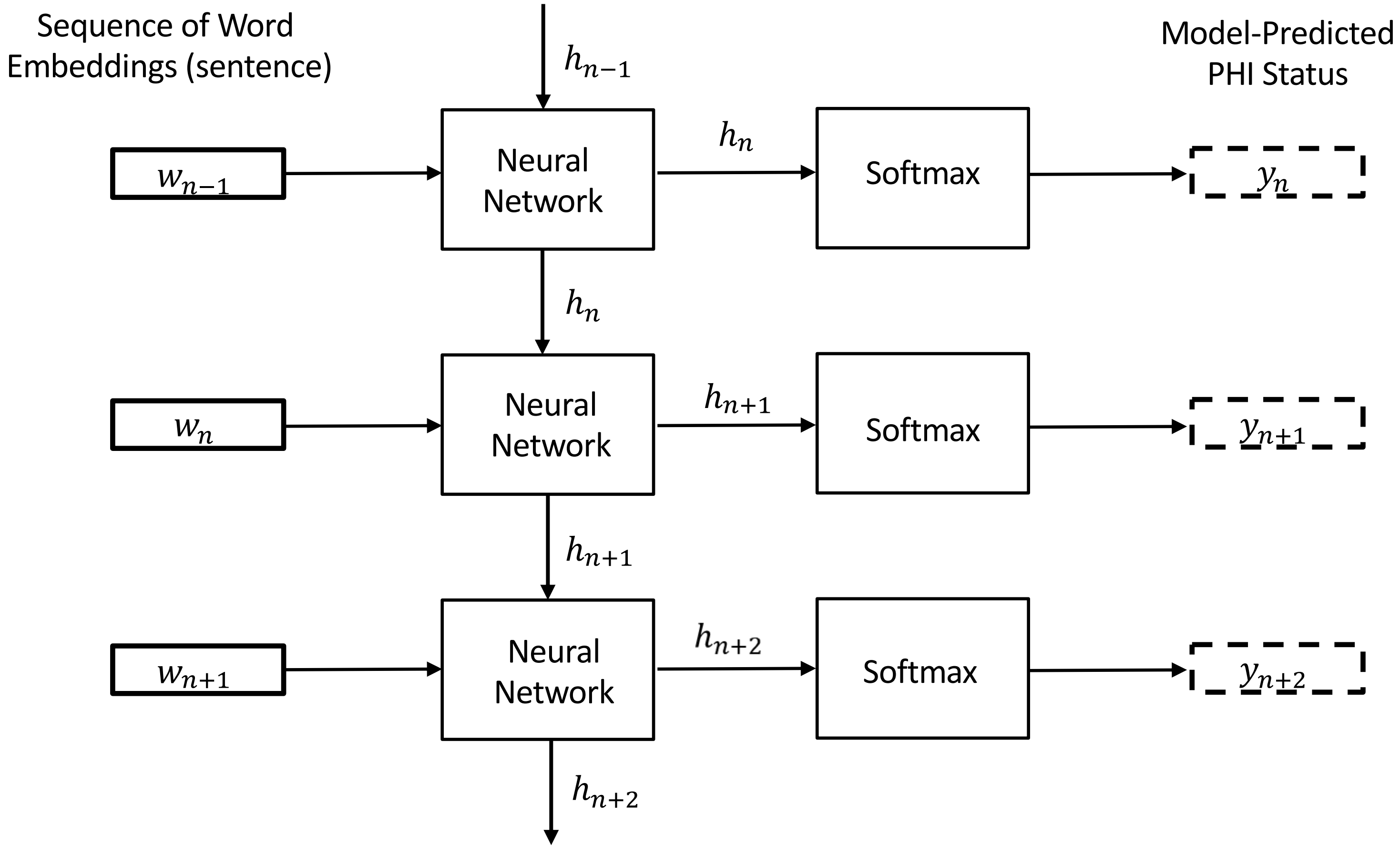
Table 5. Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness **86F** reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert, H/O paroxysmal afib VNA **171-311-7974** ======= Medications
DATE	During his **May** hospitalization he had dysphagia Social history: divorced, quit smoking in **08**, sober x 10 yrs, She is to see him on the **29th** of this month at 1:00 p.m. He did have a renal biopsy in teh late **60s** adn thus will look for results, Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

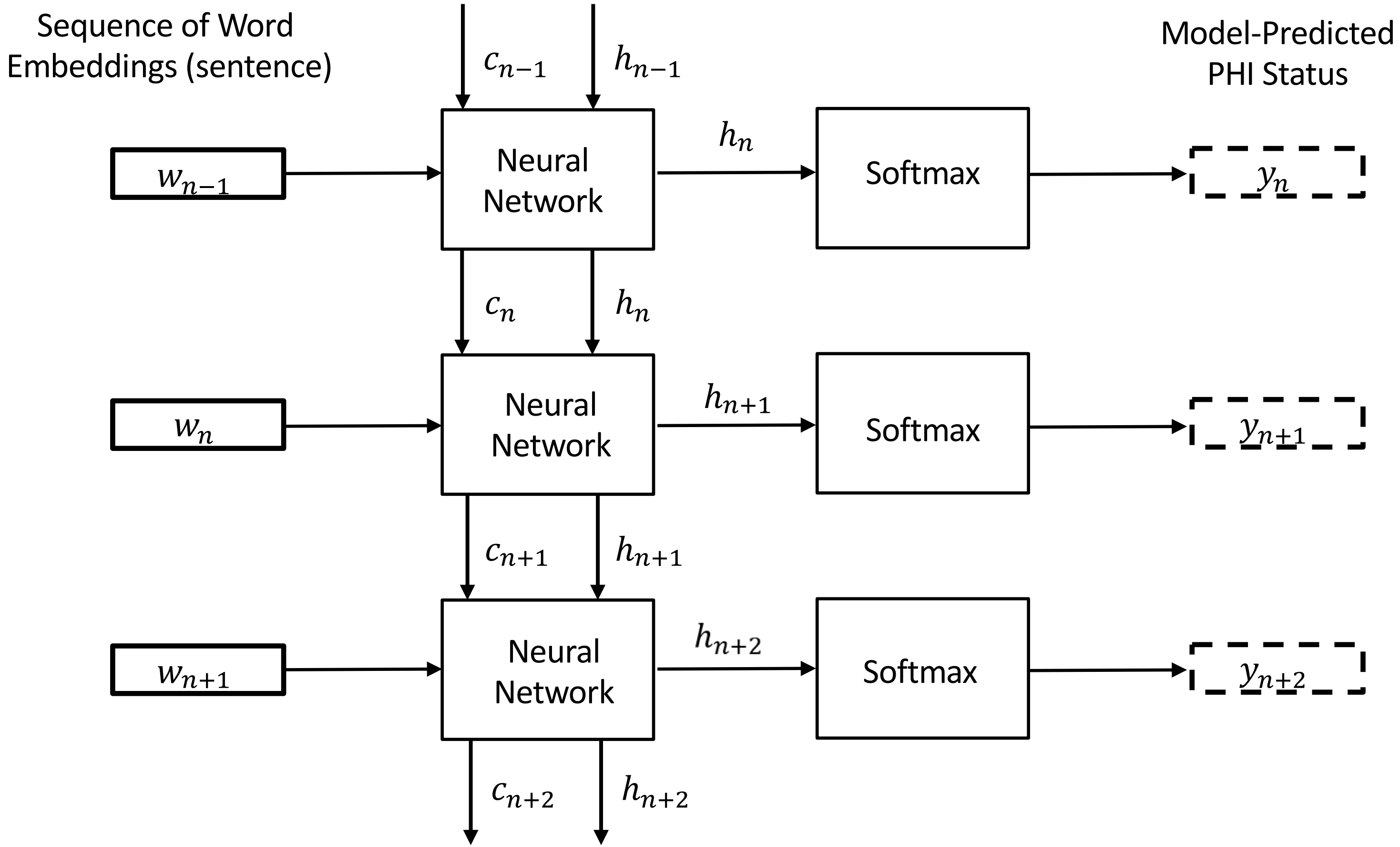
De-identification of patient notes with recurrent neural networks
Dernoncourt F, Lee JY, Uzuner O, Szolovits P
JAMIA 24(3), 2017, 596–606

- A bidirectional RNN is used to identify PHI (18 HIPAA fields)
- *i2b2*: 889 discharge summaries, >28k PHI tokens
- *MIMIC*: 1635 discharge summaries, >60k PHI tokens
- State of the art sensitivity and F1 metric on both datasets

Deidentification via Recurrent Neural Network…

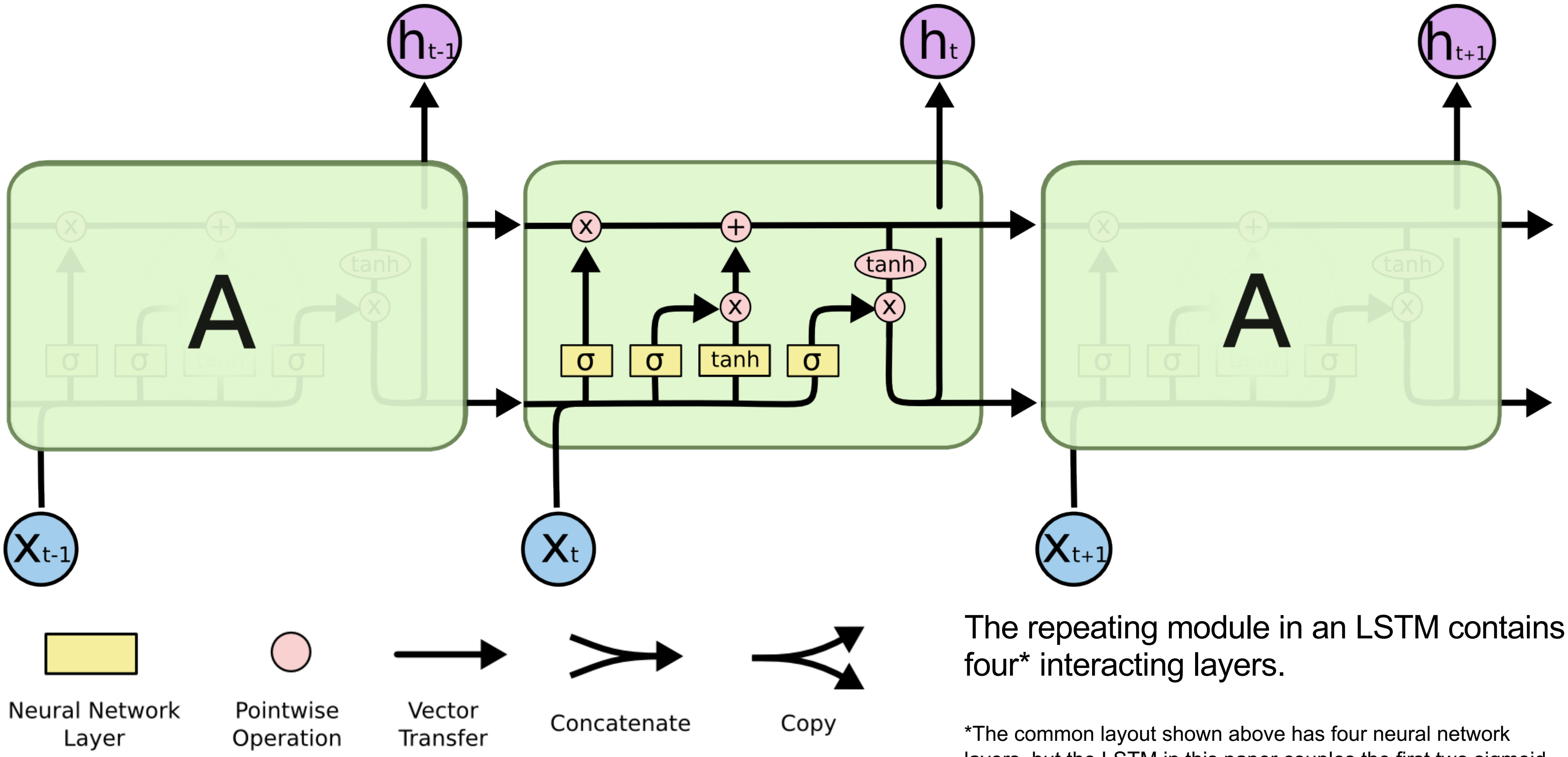


…Specifically, an LSTM



Inside an LSTM

http://colah.github.io/posts/2015-08-Understanding-LSTMs/



The repeating module in an LSTM contains four* interacting layers.

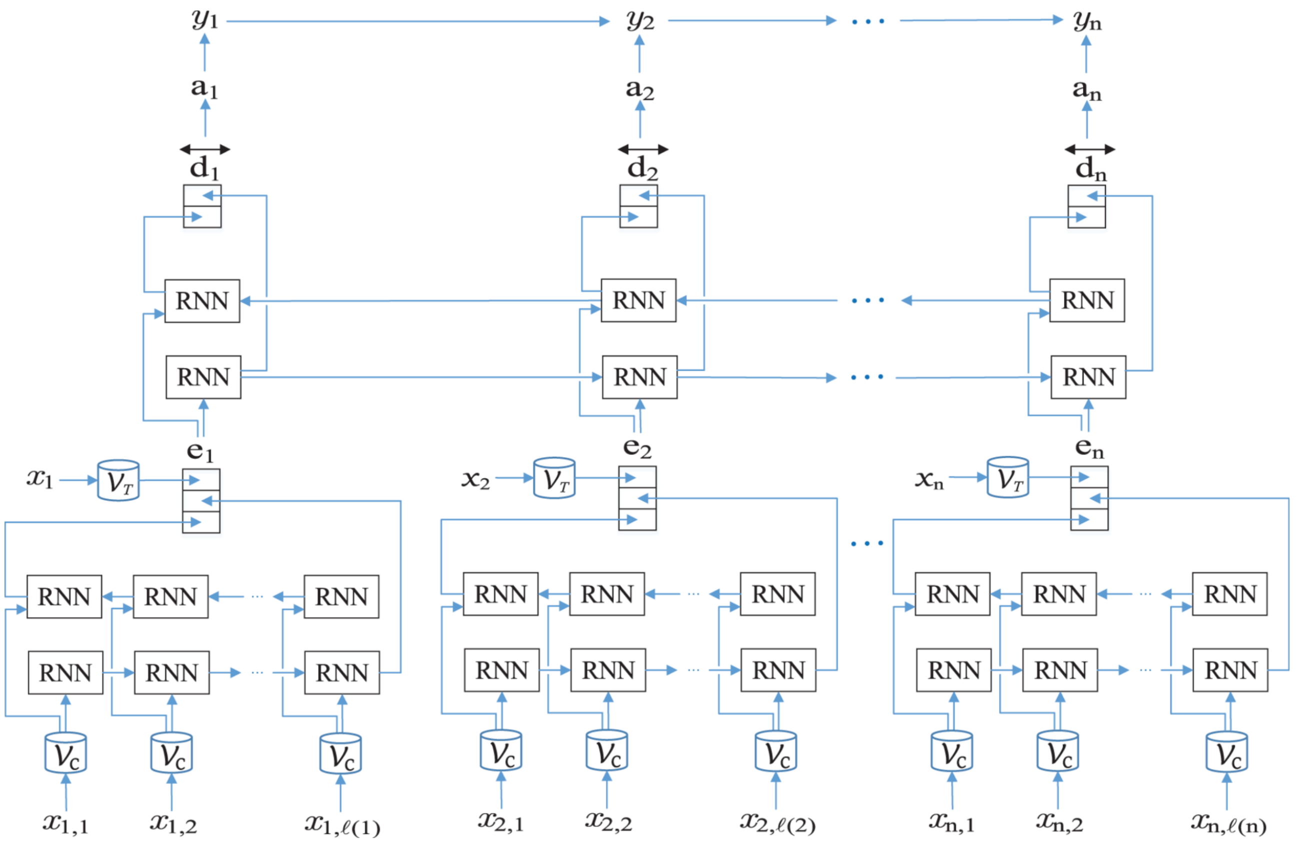
*The common layout shown above has four neural network layers, but the LSTM in this paper couples the first two sigmoid layers.

Modifications to RNN for Deidentification

1. LSTM provides a more flexible representation of previous context
   - output and cell state are both passed to the next block

2. Bidirectional LSTM provides context from subsequent words as well as previous words

3. Character-level RNN allows a non-trivial representation of out-of-dictionary tokens



From: De-identification of patient notes with recurrent neural networks

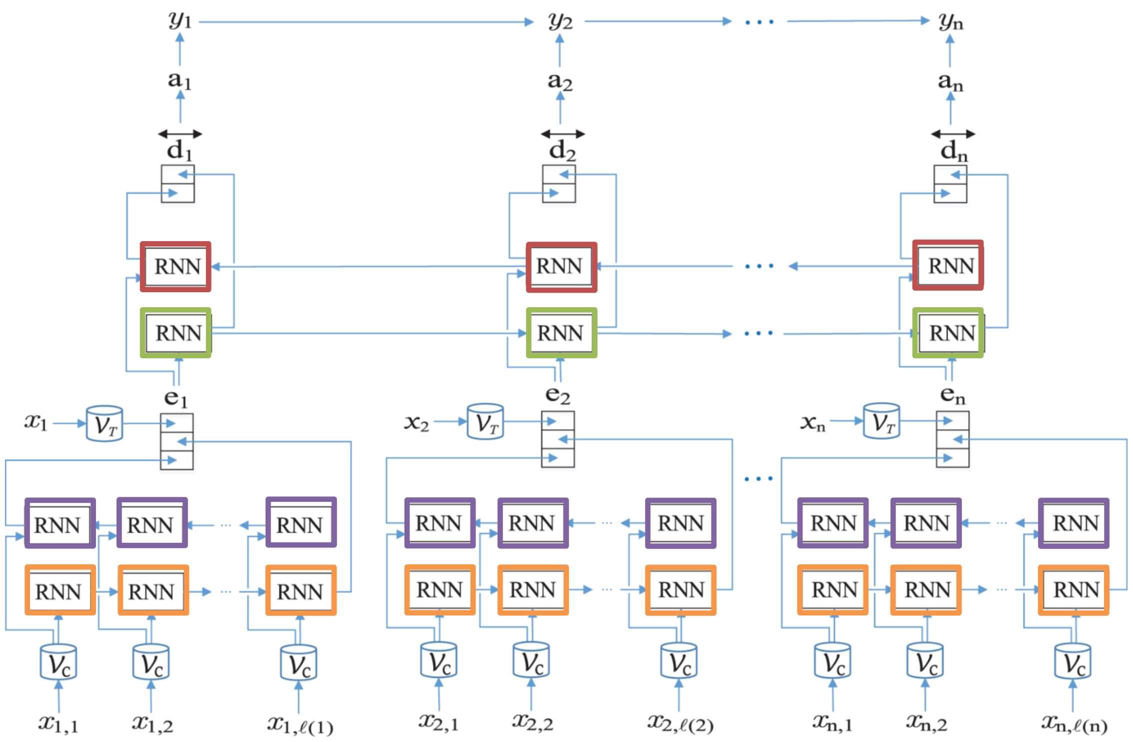
J Am Med Inform Assoc. 2016;24(3):596-606. doi:10.1093/jamia/ocw156

backward token LSTM

forward token LSTM

backward character LSTM

forward character LSTM



Character embedding dimension: **25**

Character-based token-embedding LSTM dimension: **25**

Token embedding dimension: **100**

Label prediction LSTM dimension: **100**

Deidentification of Patient Notes

RESULTS

Train, Validation, Test

MIMIC:

80% train/validation

20% test

i2b2:

60% train/validation

40% test

"All results were computed using the official evaluation script from the i2b2 2014 de-identification challenge."

Table 3. Overview of the i2b2 and MIMIC datasets

Statistics	i2b2	MIMIC
Vocabulary size	46 803	69 525
Number of notes	1304	1635
Number of tokens	984 723	2 945 228
Number of PHI instances	28 867	60 725
Number of PHI tokens	41 355	78 633

Examples of PHI Identified by the RNN

AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness **86F** reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert, H/O paroxysmal afib VNA **171-311-7974** ======== Medications
DATE	During his **May** hospitalization he had dysphagia Social history: divorced, quit smoking in **08**, sober x 10 yrs, She is to see him on the **29th** of this month at 1:00 p.m. He did have a renal biopsy in teh late **60s** adn thus will look for results, Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

Evaluation Metrics

Precision, or positive predictive value:

$$\frac{\text{true positives}}{\text{all positive predictions}}$$

Recall, or sensitivity:

$$\frac{\text{true positives}}{\text{all condition positives}}$$

F1-score:

$$\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

	Condition Positive	**Condition Negative**
Prediction Positive	True Positive	False Positive
Prediction Negative	False Negative	True Negative

RNN Model Outperforms Previous Benchmarks

Table 4. Performance (%) on the PHI as defined in HIPAA

	i2b2			MIMIC		
Model	Precision	Recall	F1	Precision	Recall	F1
Nottingham	**99.000**	96.400	97.680	–	–	–
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	**99.208**	99.251	**99.229**
CRF + ANN	97.920	**97.835**	**97.877**	98.820	**99.398**	99.108