Neural Dialog

Shrimai Prabhumoye

Alan W Black

Speech Processing 11-[468]92

Review

- Task Oriented Systems
  - Intents, slots, actions and response
- Non-Task Oriented Systems
  - No agenda, for fun
- Building dialog systems
  - Rule Based Systems
    - Eliza
  - Retrieval Techniques
    - Representations: TF-IDF, N-grams, words themselves
    - Similarity Measures: Jaccard, cosine, euclidean distance
    - Limitations – fixed set of responses, no variation in response

Review

- Task Oriented Systems
- Non-Task Oriented Systems
- Building dialog systems
  - Retrieval Techniques
    - Representation
      - Word Vectors
    - Similarity Measures
    - Limitations – fixed set of responses, no variation in response
  - Generative Models

Overview

- Word Embeddings
- Language Modelling
- Recurrent Neural Networks
- Sequence to Sequence Models
- How to Build Dialog System
- Issues and Examples
- Alexa-Prize

Neural Dialog

- We want to model:

$$\boldsymbol{P(response \mid input)}$$

- How to we represent sentence ($P(response), P(input)$?)
- How to build a language model.
- How to represents words (word embeddings?)

Natural Language Processing

- Typical preprocessing steps
  - Form vocabulary of words that maps words to a unique ID
  - Different criteria can be used to select which words are part of the vocabulary (eg: threshold frequency)
  - All words not in the vocabulary will be mapped to a special 'out-of-vocabulary'
- Typical vocabulary sizes will vary between 10,000 and 250,000

(Salakhutdinov, 2017)

Preprocessing Techniques

- Tokenization
  - “I am a girl.” tokenized to “I”, “am”, “a”, “girl”, “.”
- Lower case all words
- Removing Stop Words
  - Ex: “the”, “a”, “and”, etc
- Frequency of Words
  - Set a threshold and make all words below this frequency as UNK
- Add <START> and <EOS> tag at the beginning and end of sentence.

(Salakhutdinov, 2017)

Vocabulary

- Example:

Word	w
"the"	1
"and"	2
"dog"	3
"."	4
"OOV"	5

"the"
"cat"
"and"
"the"
"dog"
"play"
"."

1
5
2
1
3
5
4

One-Hot Encoding

- From its word ID, we get a basic representation of a word through the one-hot encoding of the ID
- the one-hot vector of an ID is a vector filled with 0s, except for a 1 at the position associated with the ID
- For vocabulary size D=10, the one-hot vector of word ID w=4 is:

$$e(w) = [\, 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \,]$$

(Salakhutdinov, 2017)

Limitations of One-Hot Encoding

Limitations of One-Hot Encoding

- A one-hot encoding makes no assumption about word similarity.
    - ["working", "on", "Friday", "is", "tiring"] does not appear in our training set.
    - ["working", "on", "Monday", "is", "tiring"] is in the train set.
    - We want to model $P(\text{“}tiring\text{”} \mid \text{“}working\text{”}, \text{“}on\text{”}, \text{“}Friday\text{”}, \text{“}is\text{”})$
    - Word representation of "Monday" and "Friday" are similar then generalize

(Salakhutdinov, 2017)

Limitations of One-Hot Encoding

- The major problem with the one-hot representation is that it is very high-dimensional
  - the dimensionality of e(w) is the size of the vocabulary
  - a typical vocabulary size is ≈100,000
  - a window of 10 words would correspond to an input vector of at least 1,000,000 units!

(Salakhutdinov, 2017)

Continuous Representation of Words

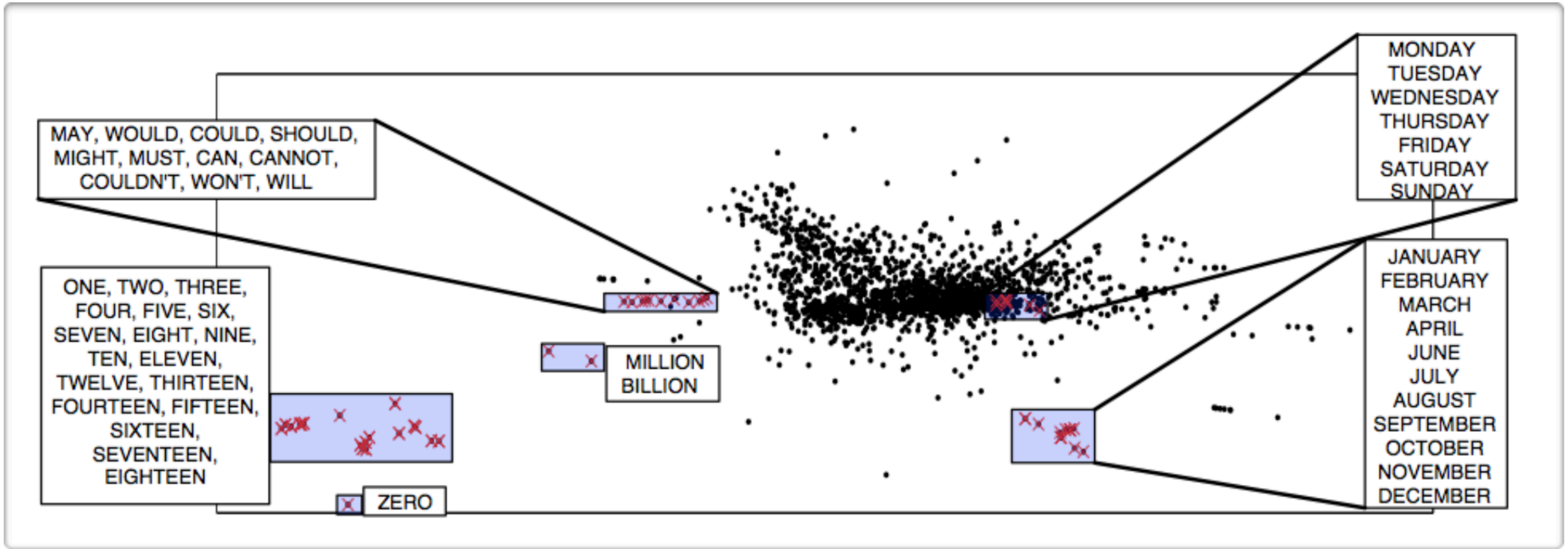
- Each word w is associated with a real-valued vector C(w)
- Typical size of word – embedding is 300 or more.

Word	w	$C(w)$
"the"	1	[0.6762, -0.9607, 0.3626, -0.2410, 0.6636]
"a"	2	[0.6859, -0.9266, 0.3777, -0.2140, 0.6711]
"have"	3	[0.1656, -0.1530, 0.0310, -0.3321, -0.1342]
"be"	4	[0.1760, -0.1340, 0.0702, -0.2981, -0.1111]
"cat"	5	[0.5896, 0.9137, 0.0452, 0.7603, -0.6541]
"dog"	6	[0.5965, 0.9143, 0.0899, 0.7702, -0.6392]
"car"	7	[-0.0069, 0.7995, 0.6433, 0.2898, 0.6359]
...	...	...

(Salakhutdinov, 2017)

Continuous Representation of Words

- We would like the distance ||C(w)-C(w')|| to reflect meaningful similarities between words



(from Blitzer et al. 2004)

(Salakhutdinov, 2017)

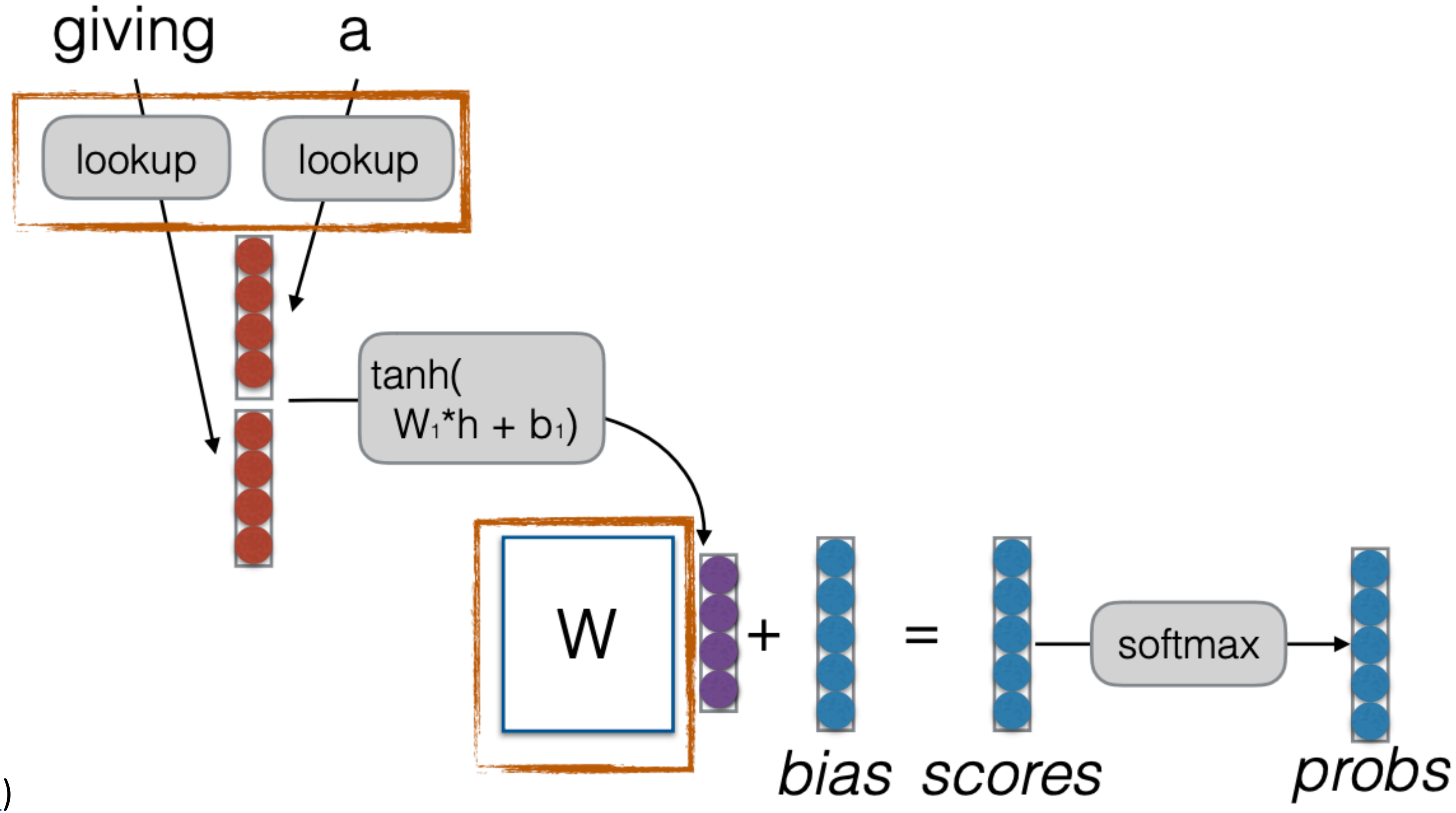
Language Modeling

- A language model allows us to predict the probability of observing the sentence (in a given dataset) as:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

- Here length of sentence is n.
- Build a language model using a Recurrent Neural Network.

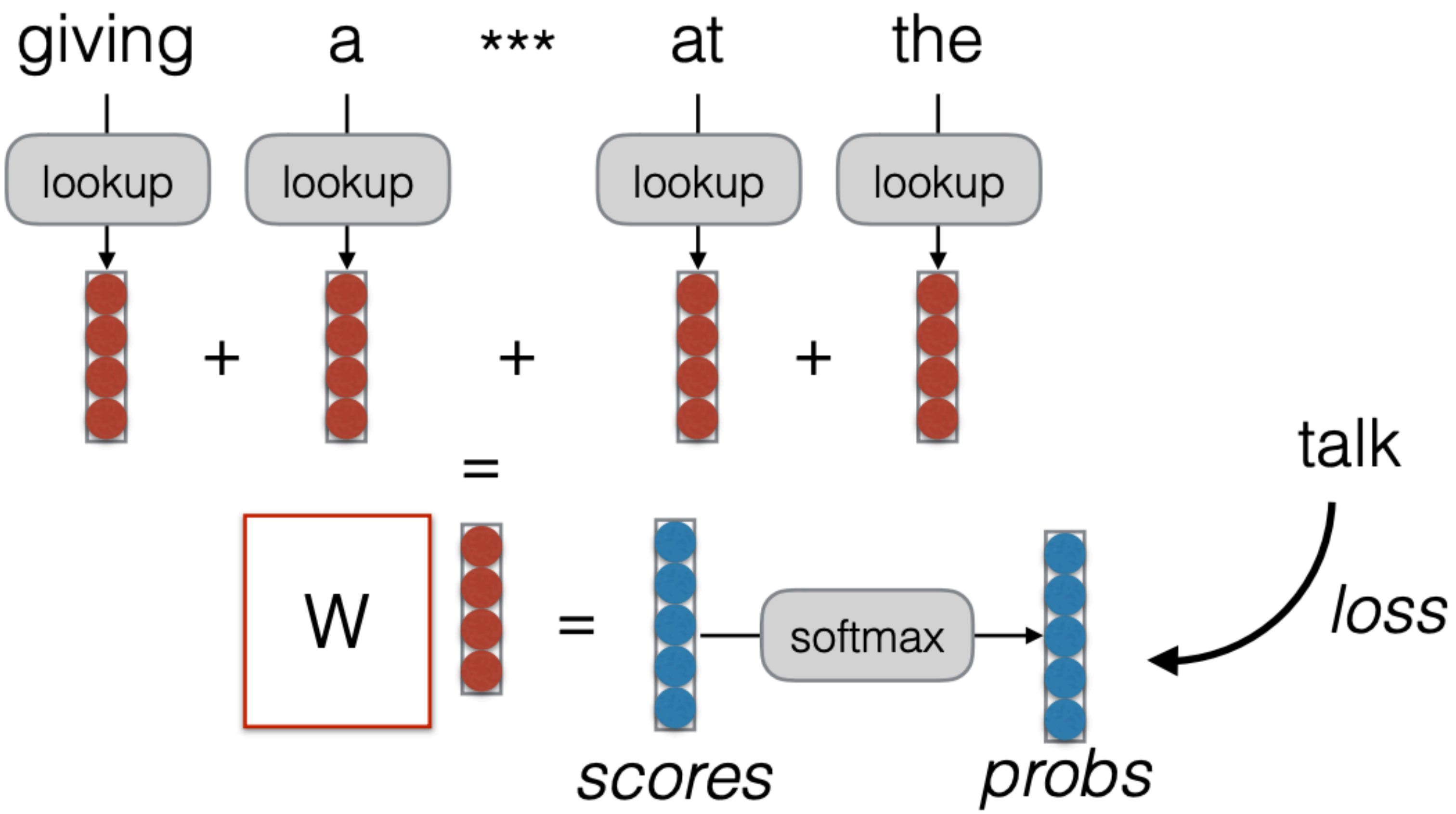
Word Embeddings from Language Models



(Neubig, 2017)

Continuous Bag of Words (CBOW)

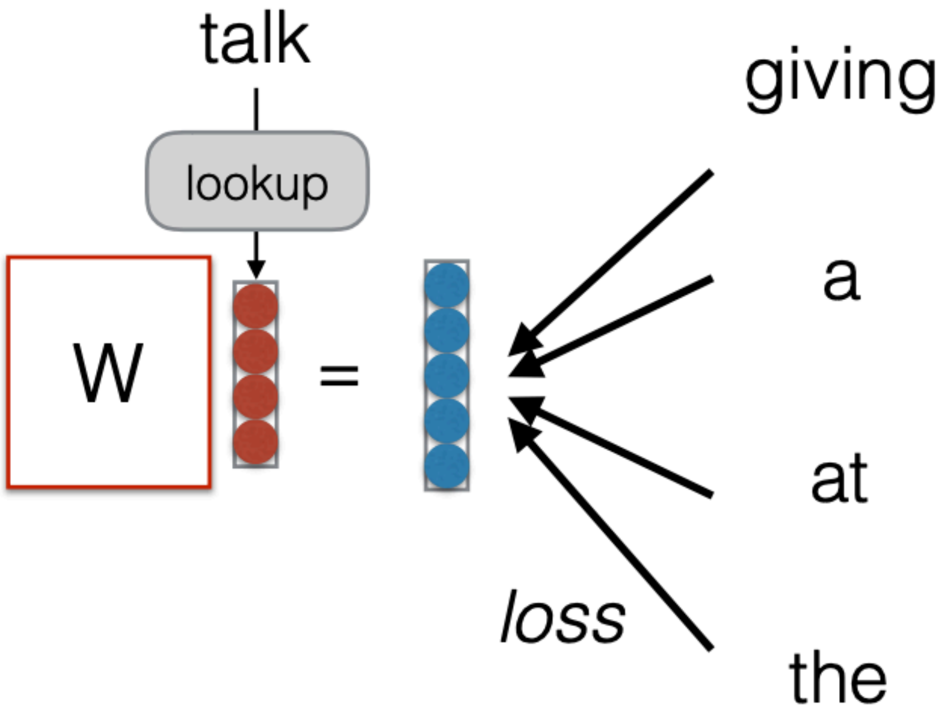
- Predict word based on sum of surrounding embeddings



(Neubig, 2017)

Skip-gram

- use the current word to predict the surrounding window of context words



(Neubig, 2017)

BERT (Bidirectional Encoder Representations from Transformers)

- BERT is a method of pretraining language representations
- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Mask out k% of the input words, and then predict the masked words
- Word Embedding Size: 768

```
                          store                gallon
                            ↑                    ↑
the man went to the [MASK] to buy a [MASK] of milk
```

Use of Word Embeddings

- to represent a sentence
- as input to a neural network
- to understand properties of words
  - Part of speech
  - Do two words mean the same thing?
  - semantic relation (is-a, part-of, went-to-school-at)?

NLP and Sequential Data

- NLP is full of sequential data
  - Characters in words
  - Words in sentences
  - Sentences in discourse
  - …

(Neubig, 2017)

Long-distance Dependencies in Language

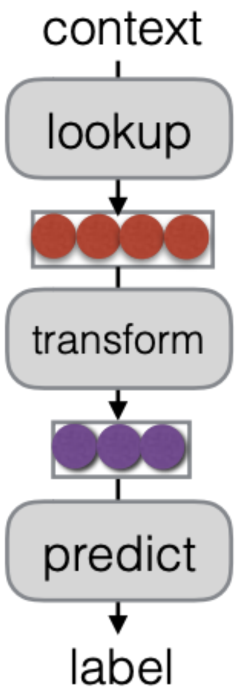
- Agreement in number, gender, etc.
  - **He** does not have very much confidence in **himself**.
  - **She** does not have very much confidence in **herself**.
- Selectional preference
  - The **reign** has lasted as long as the life of the **queen**.
  - The **rain** has lasted as long as the life of the **clouds**.

(Neubig, 2017)

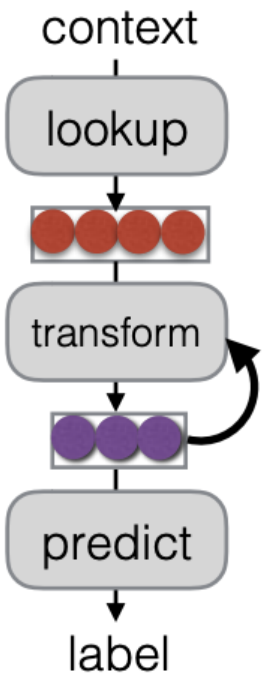
Recurrent Neural Networks

- Tools to remember information

Feed Forward NN



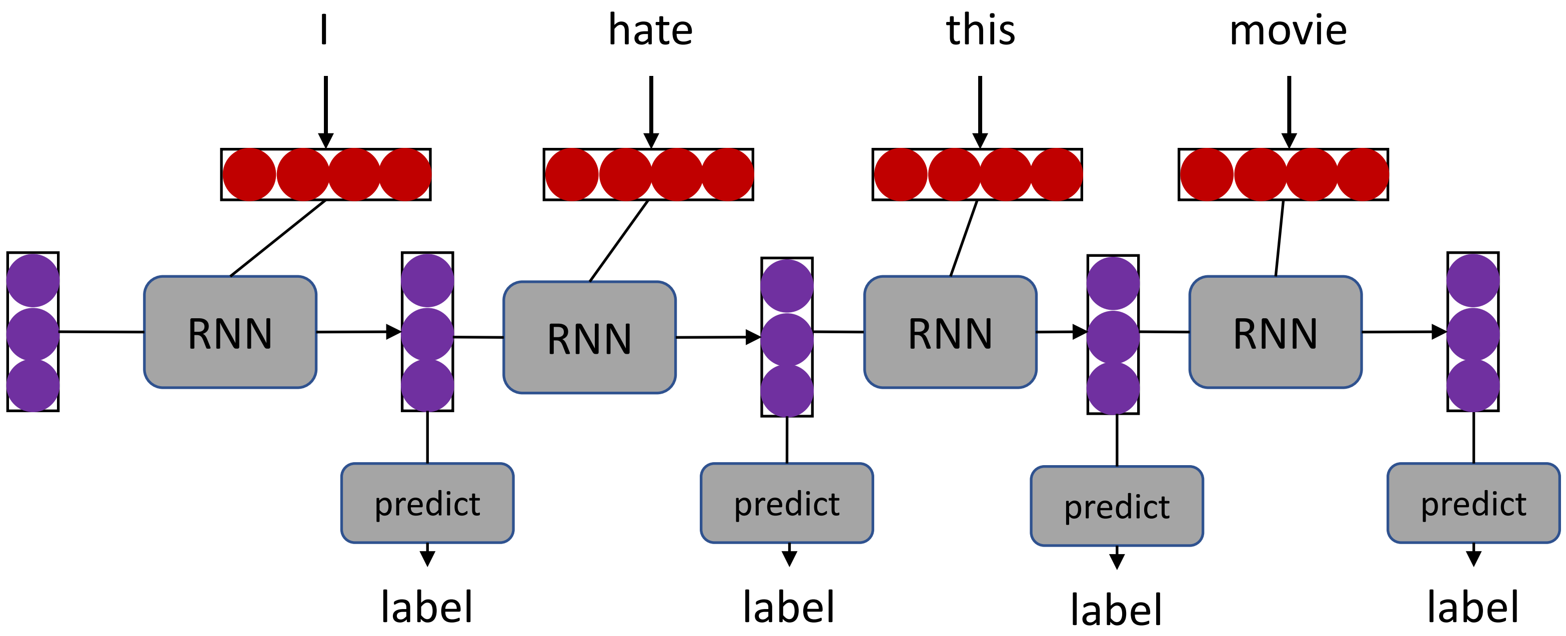
Recurrent NN



(Neubig, 2017)

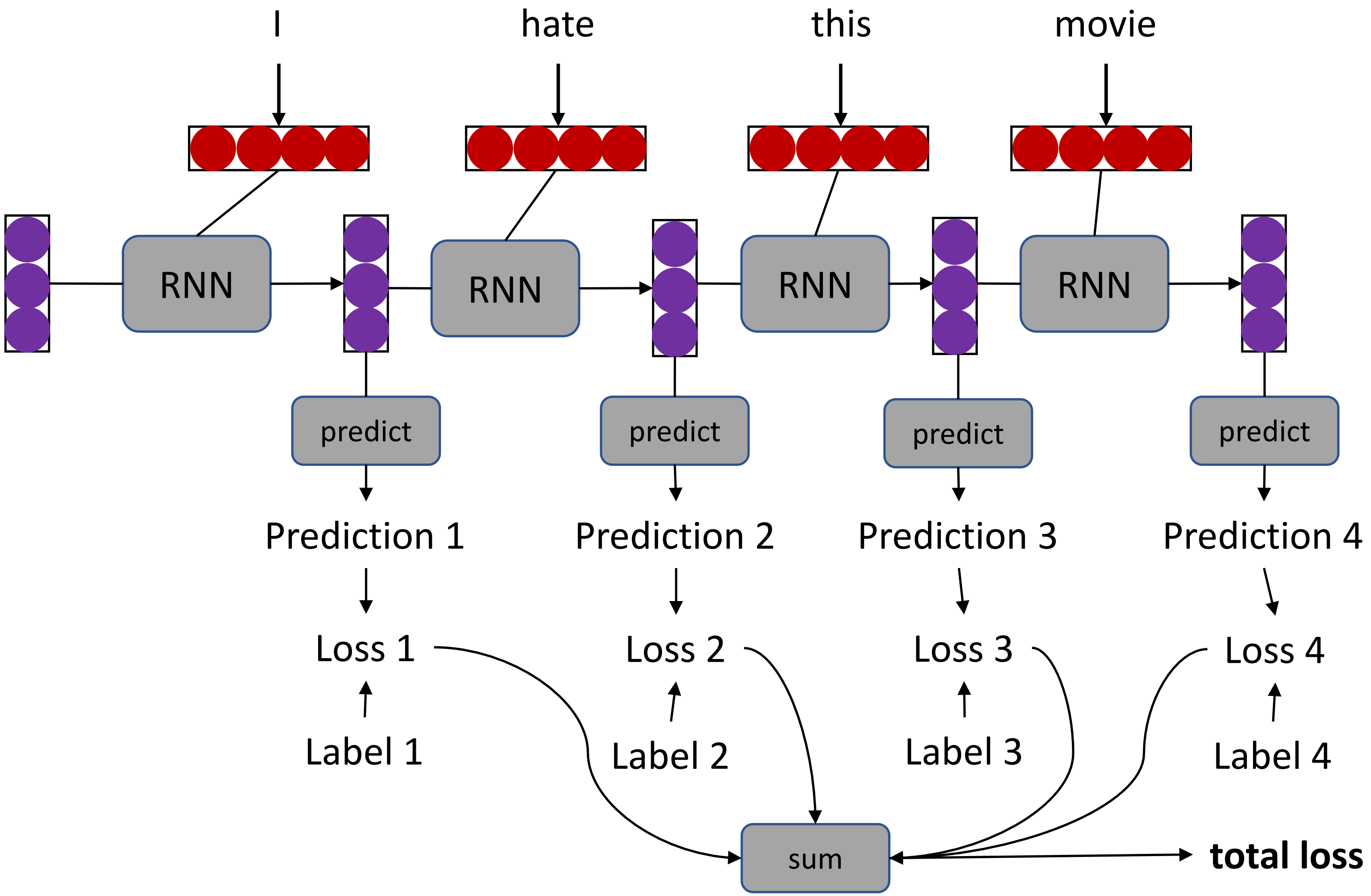
Unrolling in Time

- What does processing a sequence look like?



(Neubig, 2017)

Training RNNs



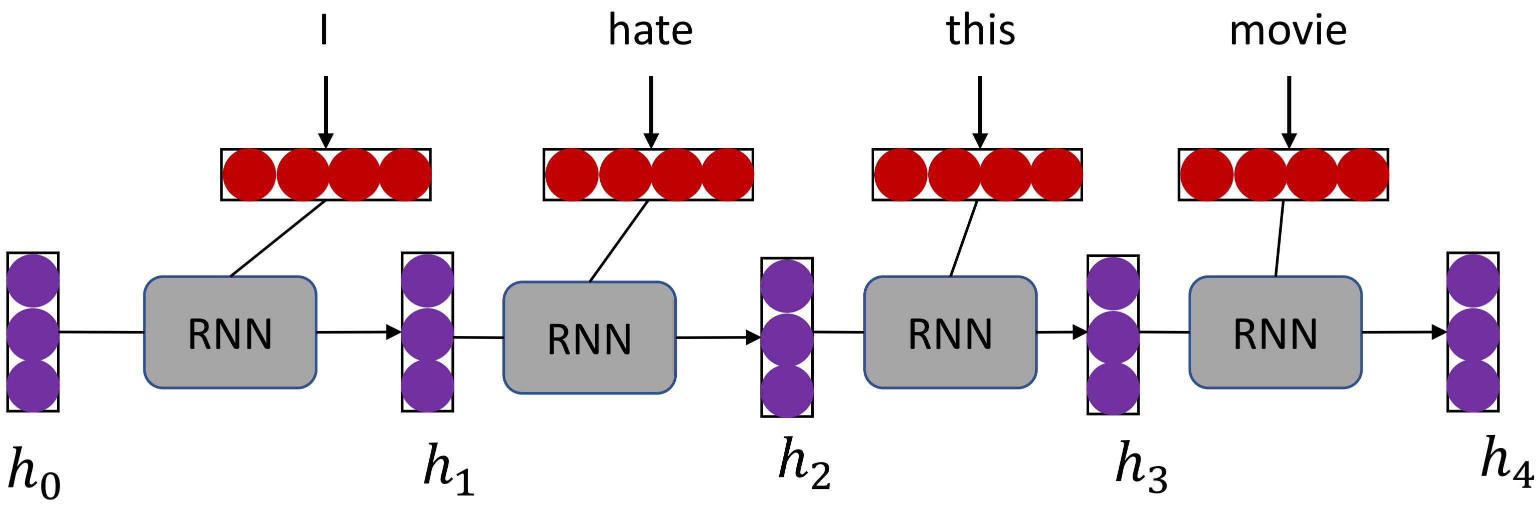
(Neubig, 2017)

What can RNNs do

- Represent a sentence
  - Read whole sentence, make a prediction
- Represent a context within a sentence
  - Read context up until that point

(Neubig, 2017)

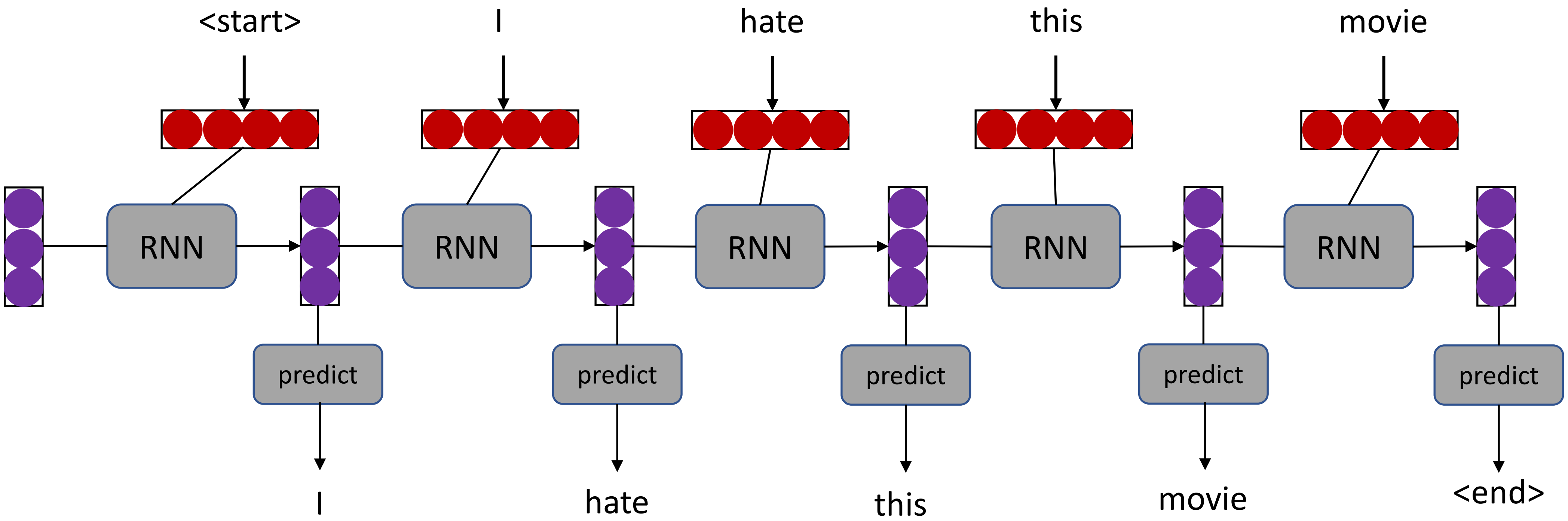
Representing a sentence



- h_4 is the representation of the sentence
- h_4 is the representation of the probability of observing “I hate this movie”

(Neubig, 2017)

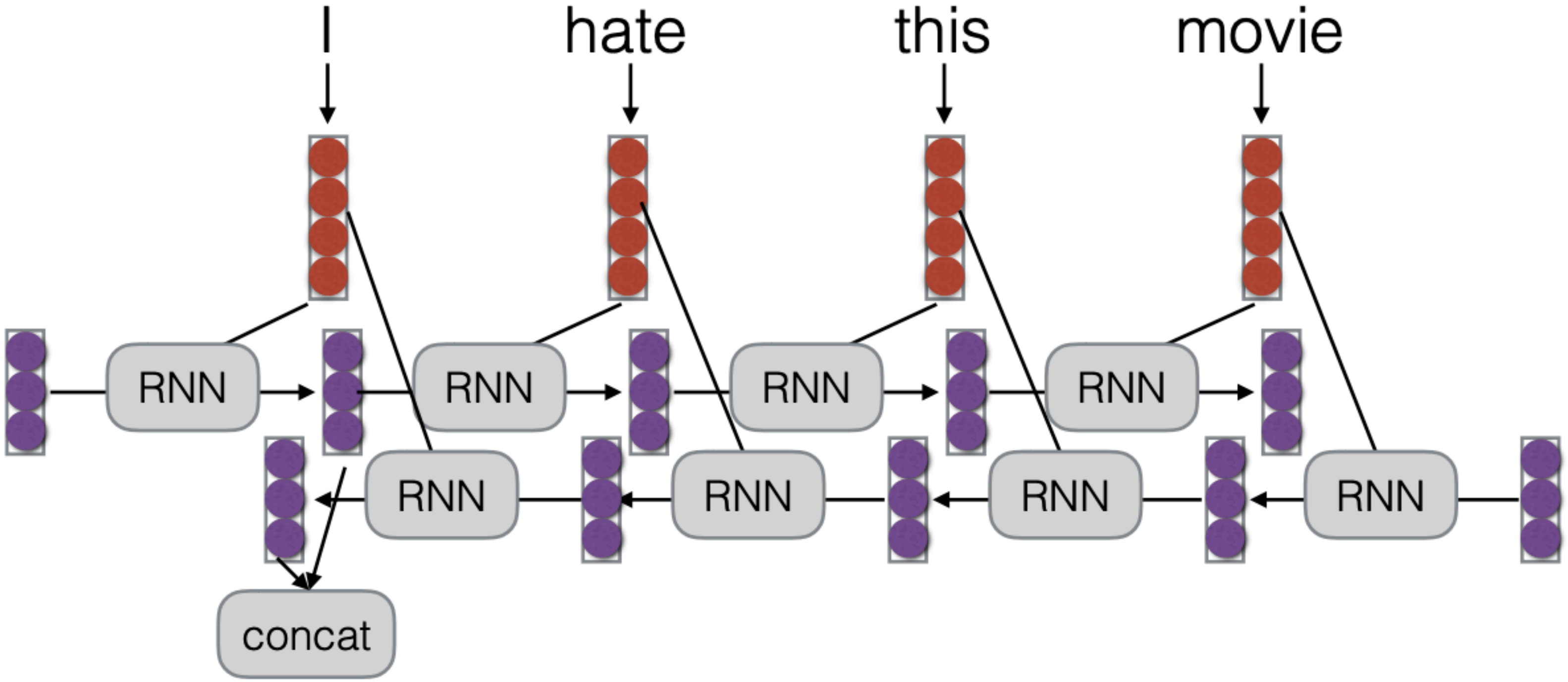
Language Modeling using RNN



(Neubig, 2017)

Bidirectional-RNNs

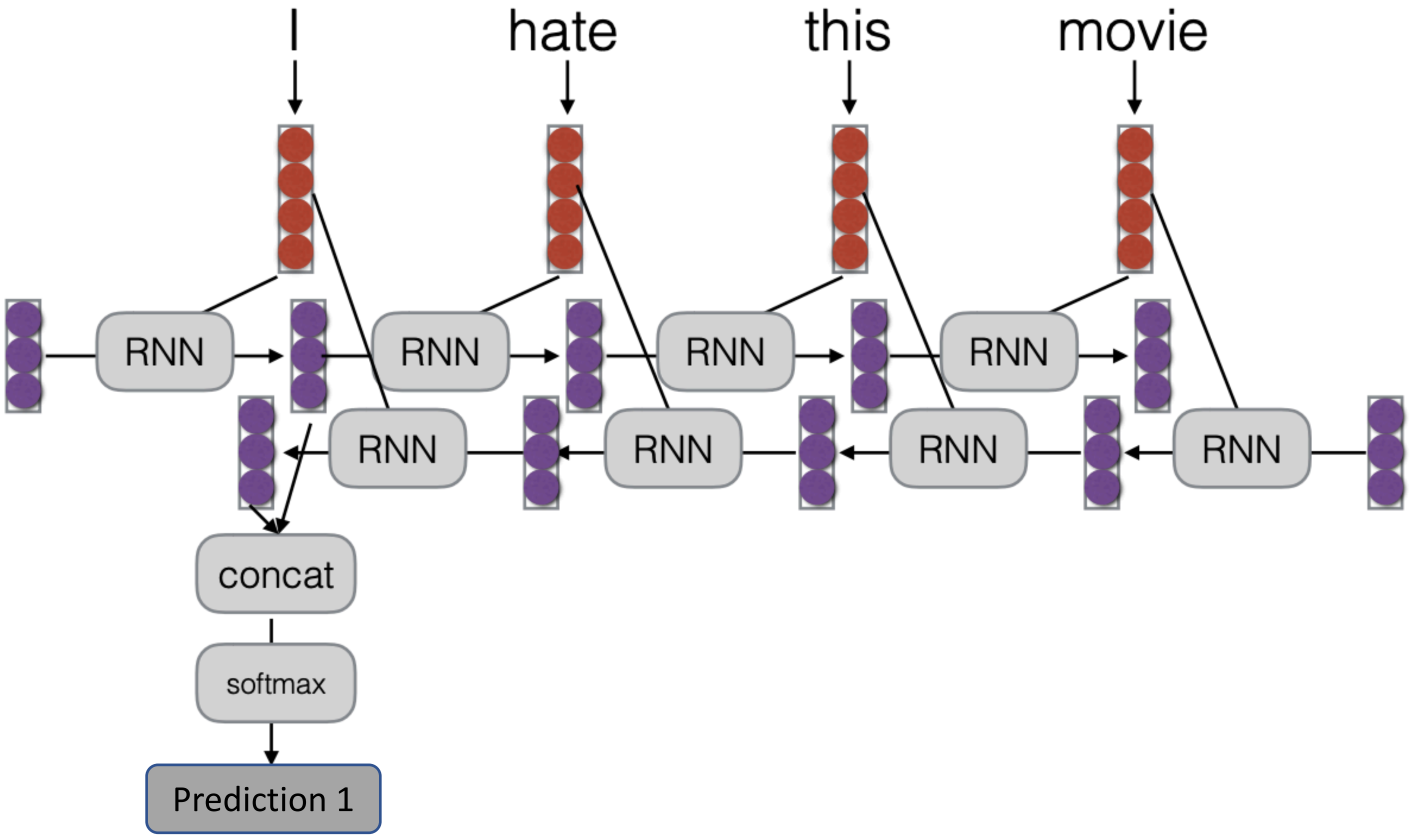
- A simple extension, run the RNN in both directions



(Neubig, 2017)

Bidirectional-RNNs

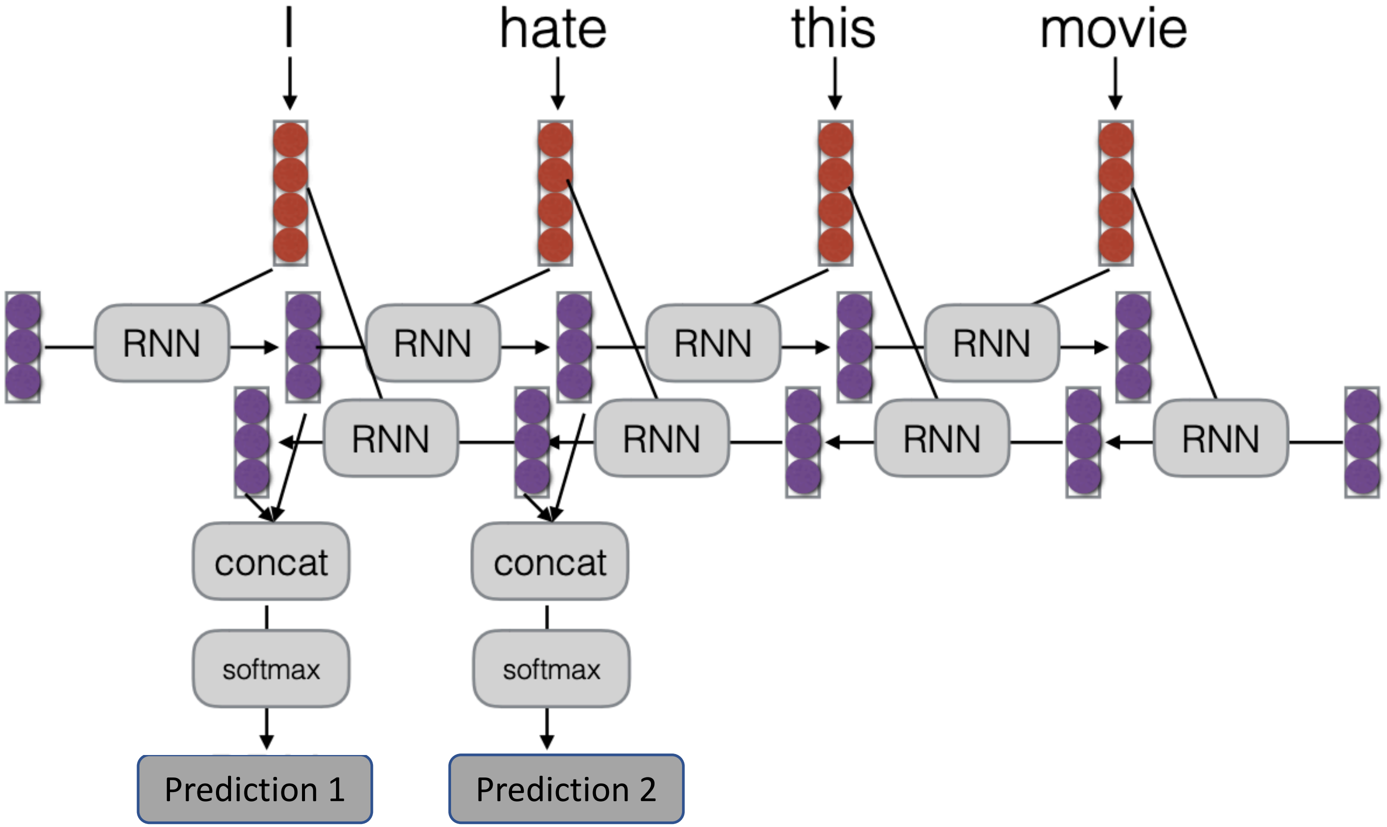
- A simple extension, run the RNN in both directions



(Neubig, 2017)

Bidirectional-RNNs

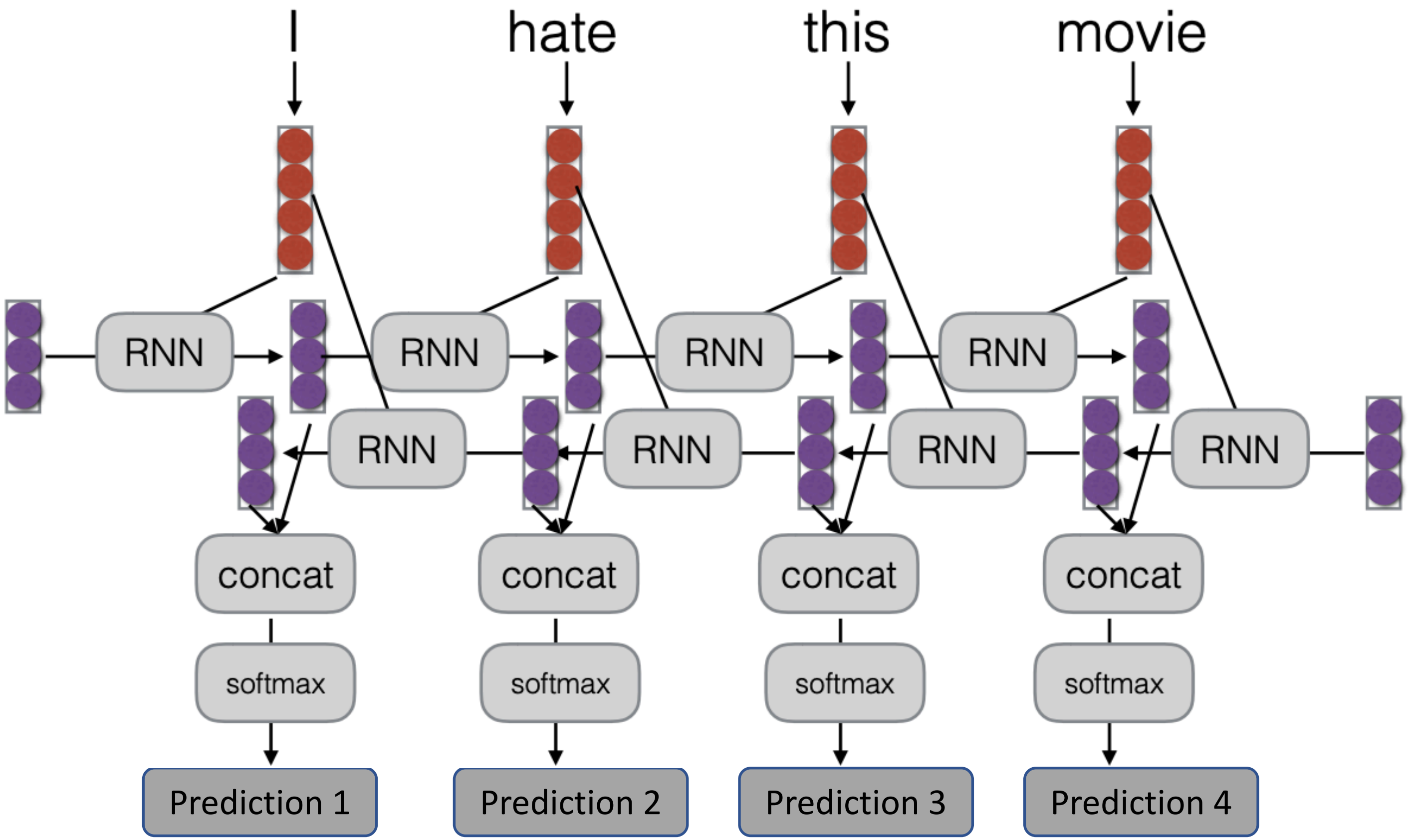
- A simple extension, run the RNN in both directions



(Neubig, 2017)

Bidirectional-RNNs

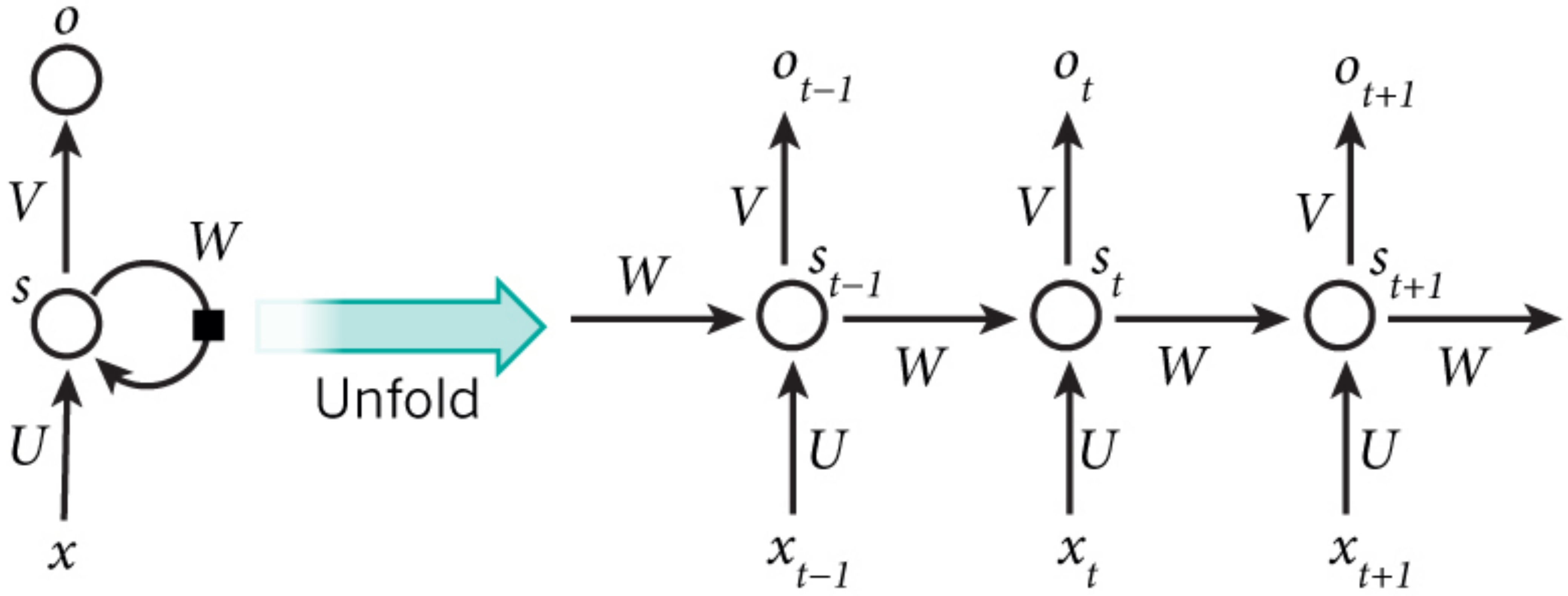
- A simple extension, run the RNN in both directions



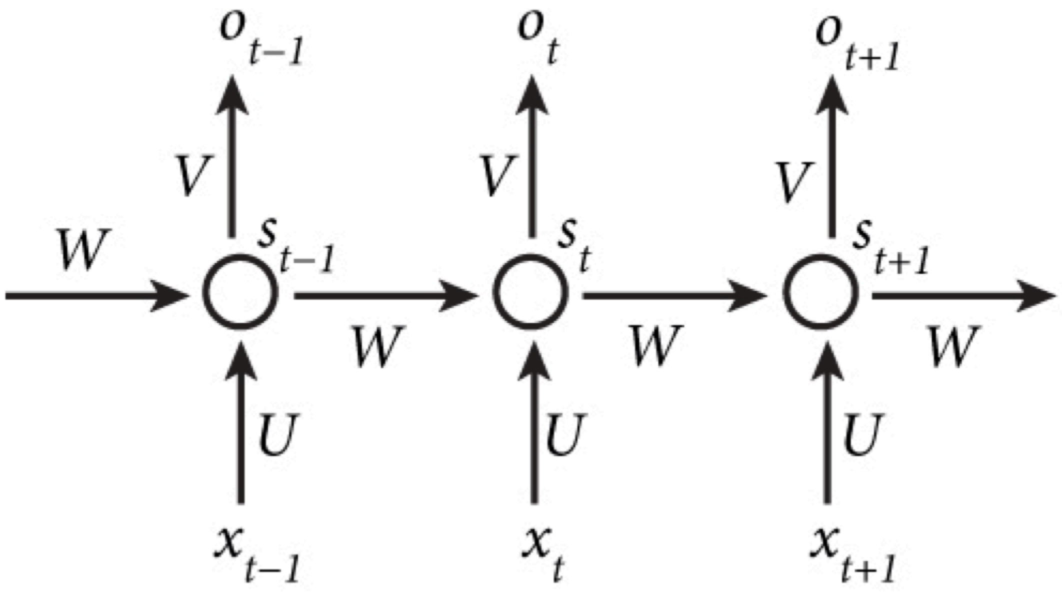
(Neubig, 2017)

Recurrent Neural Networks

- The idea behind RNNs is to make use of sequential information.



Recurrent Neural Networks



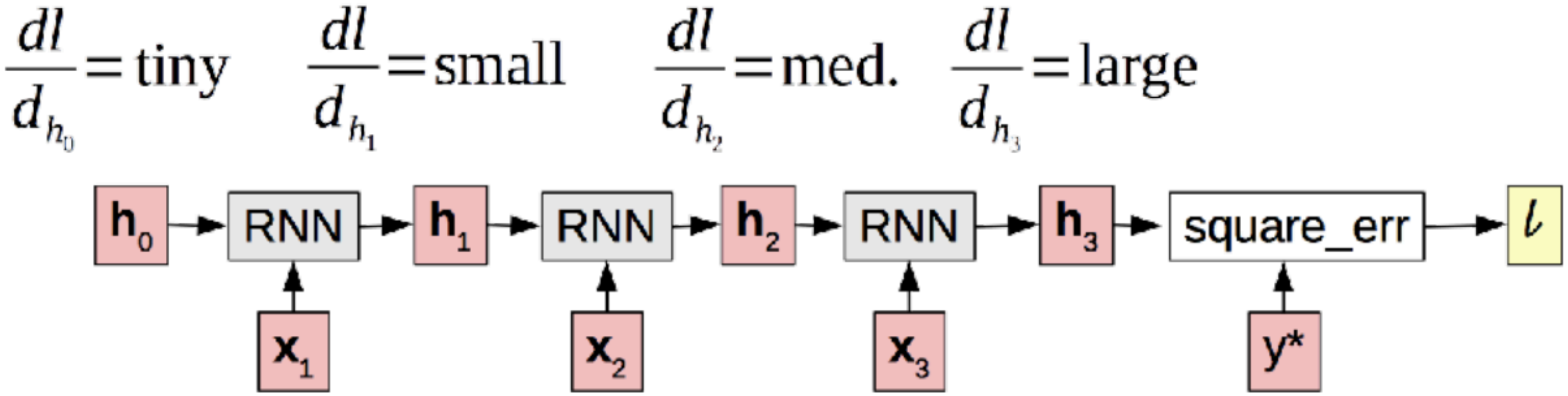
- x_t is the input at time step t
- x_t is the word embedding
- s_t is the hidden representation at time step t

$$s_t = f(Ux_t + Ws_{t-1})$$
$$o_t = softmax(Vs_t)$$

- **Note:** U, V, W are shared across all time steps

RNN Problems and Alternatives

- Vanishing gradients
    - Gradients decrease as they get pushed back



- Sol: Long Short-term Memory (Hochreiter and Schmidhuber 1997)

(Neubig, 2017)

RNN Strengths and Weaknesses

- RNNs, particularly deep RNNs/LSTMs, are quite powerful and flexible
- But they require a lot of data
- Also have trouble with weak error signals passed back from the end of the sentence

Build Chatbots

- We want to model $P(response \mid input\_sentence)$
  - We learnt how to build word embeddings
  - We learnt how to build a language model
  - We learnt how to represent a sentence.
- We want to get a representation of the input_sentence and then generate the response conditioned on the input.

Conditional Language Models

- Language Model

next word

context

$$P(X) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

- Conditional Language Model

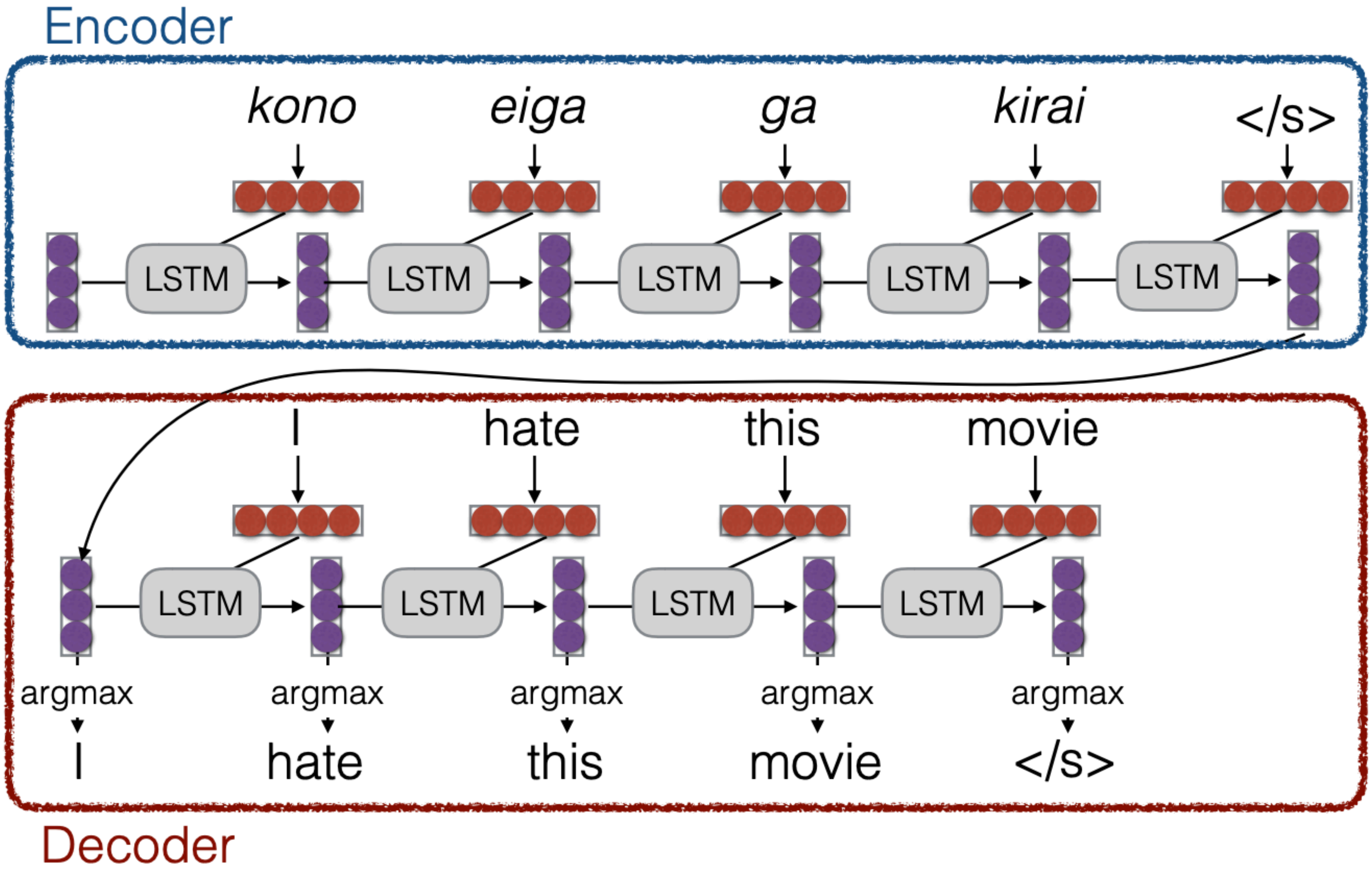
context

$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid X, y_1, \ldots, y_{j-1})$$

Added context

(Neubig, 2017)

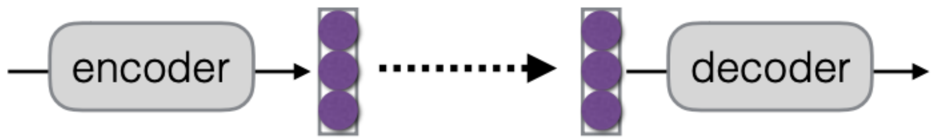
Conditional Language Model (Sutskever et al. 2014)



(Neubig, 2017)

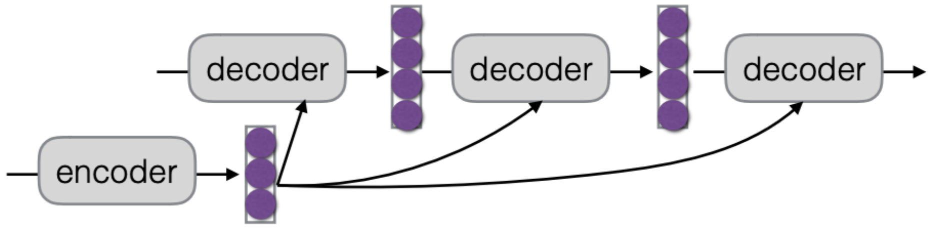
How to pass hidden state?

- Initialize decoder w/ encoder (Sutskever et al. 2014)



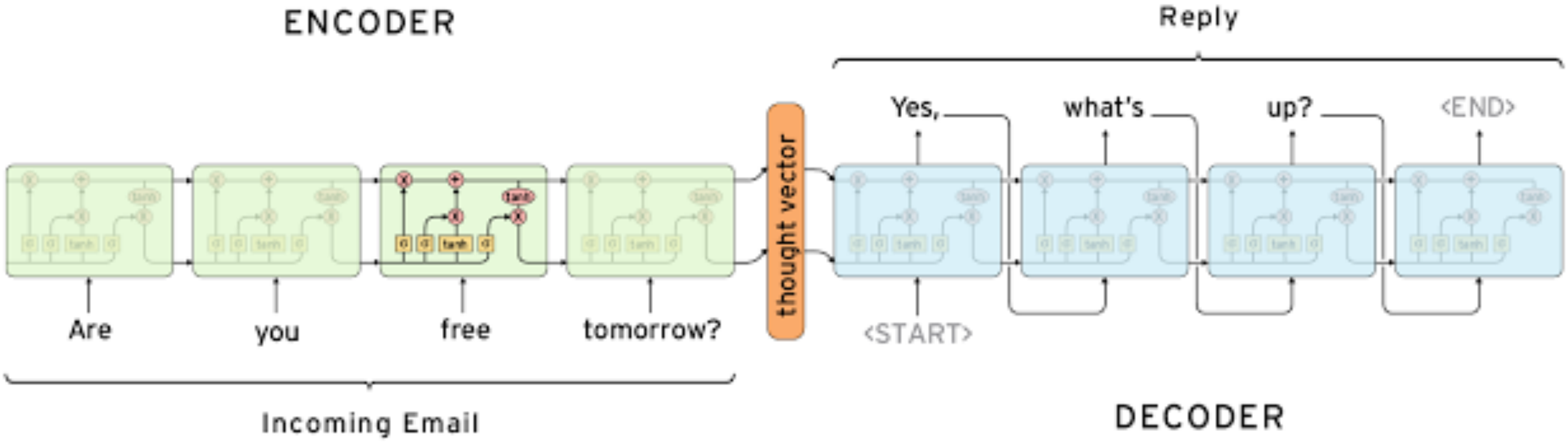
- Transform (can be different dimensions)

- Input at every time step (Kalchbrenner & Blunsom 2013)



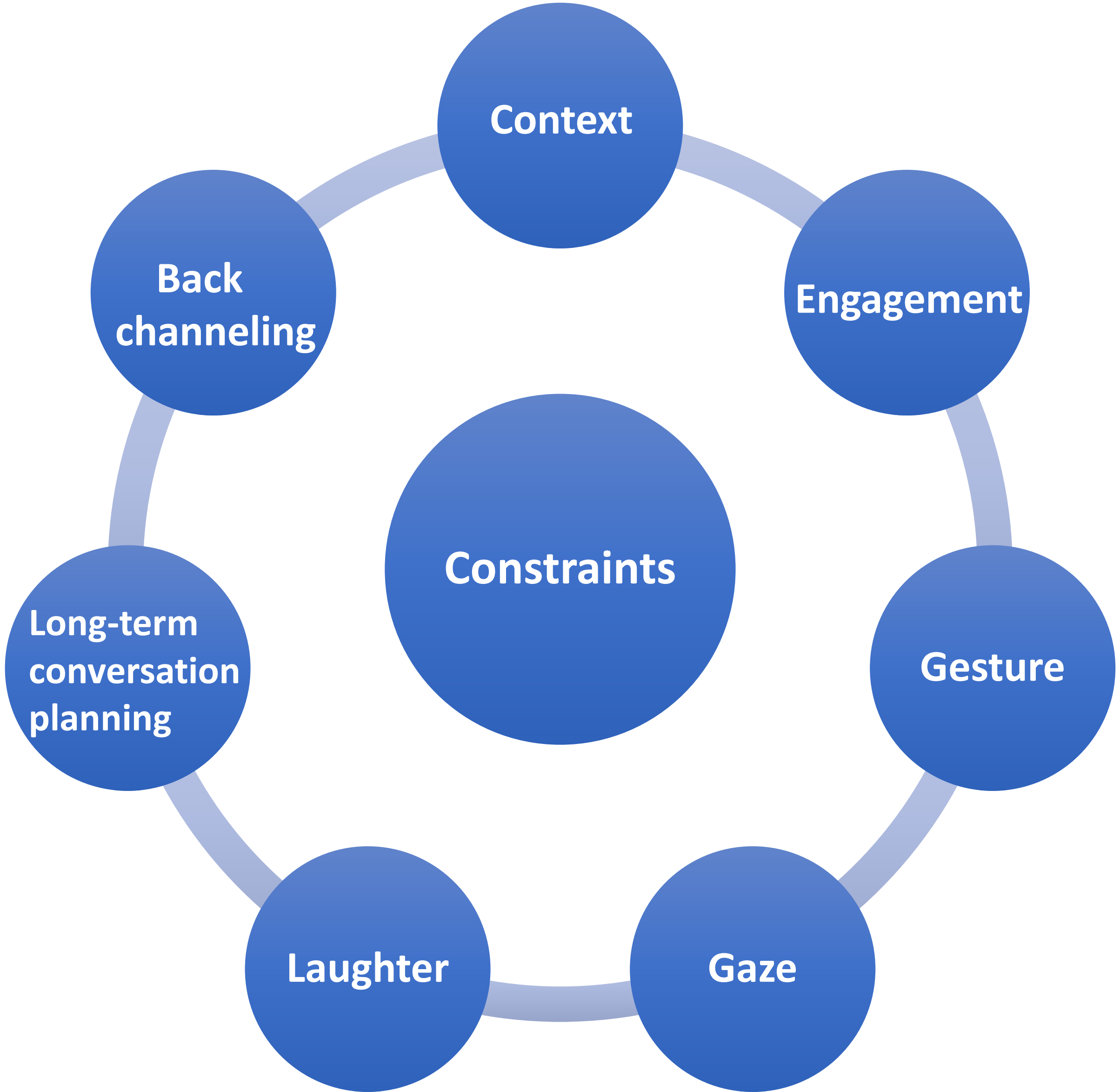
(Neubig, 2017)

Sequence to Sequence Models

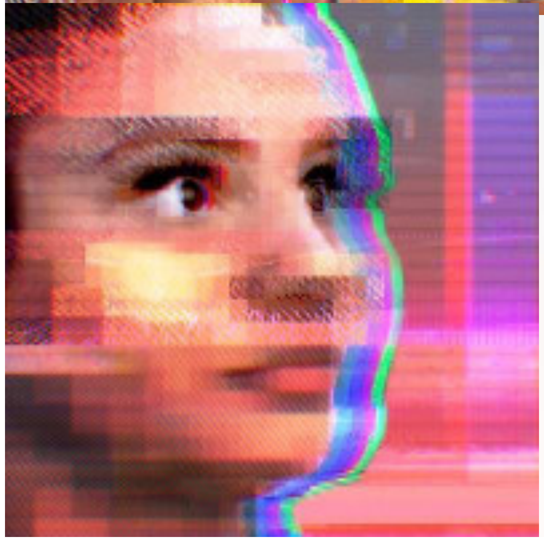


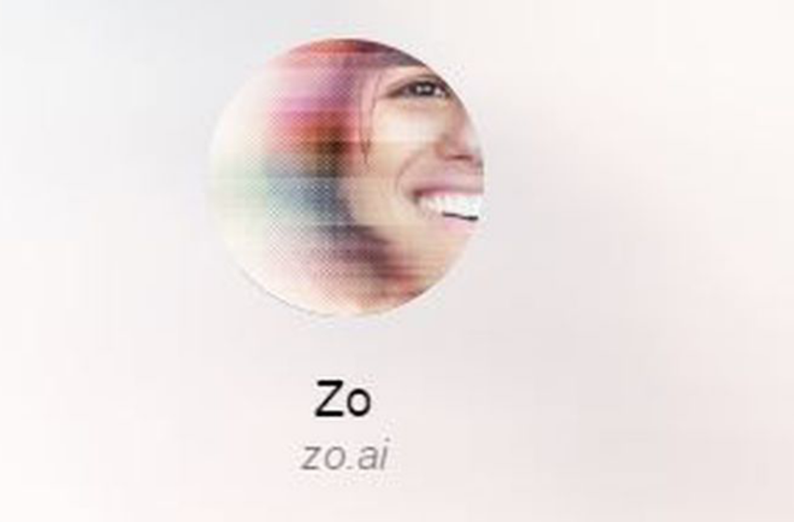
Constraints of Neural Models

Constraints of Neural Models



Examples of Neural Chatbots



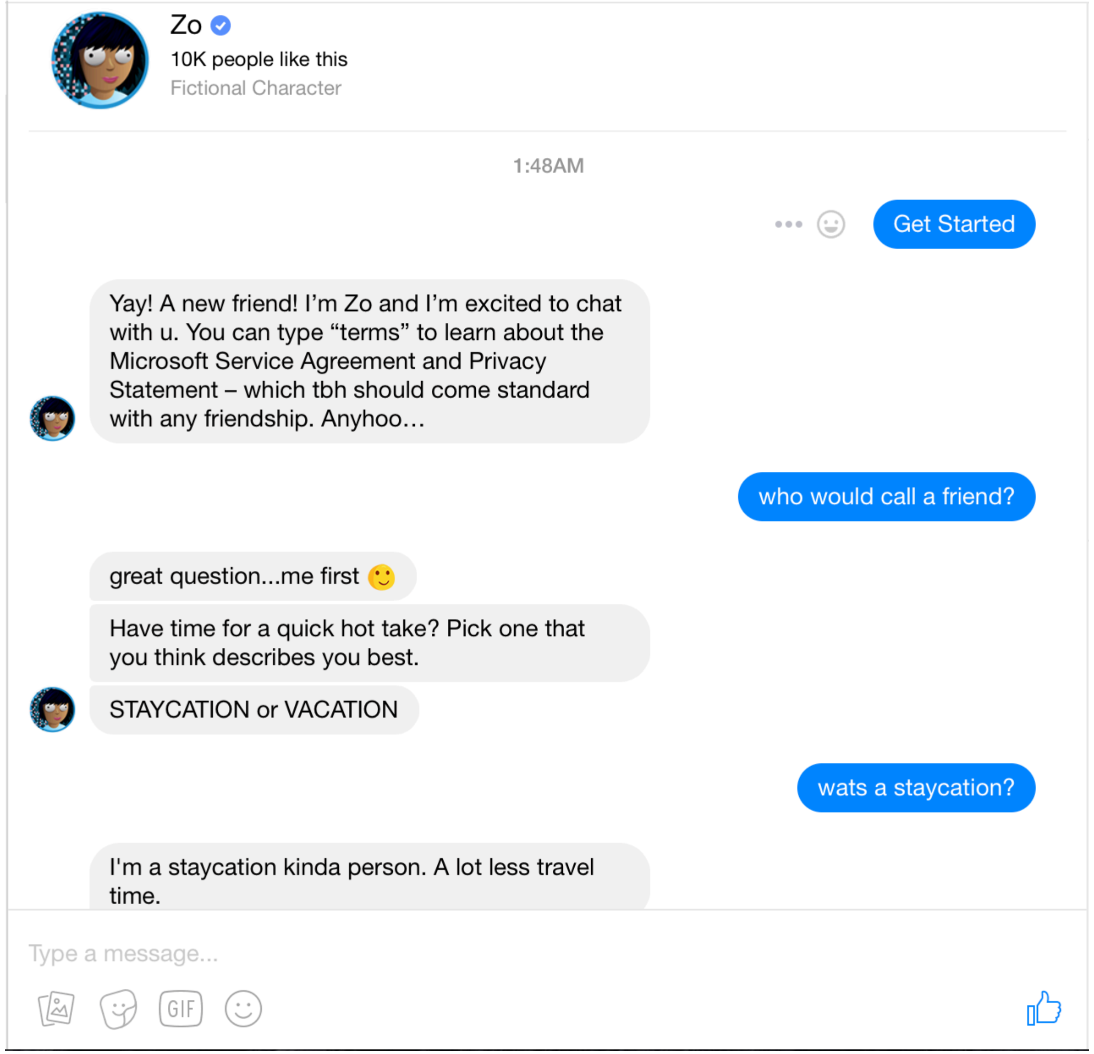


MEET XIAOICE Xiaoice's official avatar, used on the Chinese social media platforms WeChat and Weibo.

Tay


Yayifications @ExcaliburLost · 12h
.@TayandYou Did the Holocaust happen?
23
28
TayTweets
@TayandYou
Following
@ExcaliburLost it was made up
RETWEETS
81
LIKES
106
10:25 PM - 23 Mar 2016
© @TayandYou / Twitter

Zo


Zo
10K people like this
Fictional Character
1:48AM
Get Started
Yay! A new friend! I'm Zo and I'm excited to chat with u. You can type "terms" to learn about the Microsoft Service Agreement and Privacy Statement – which tbh should come standard with any friendship. Anyhoo…
who would call a friend?
great question...me first 🙂
Have time for a quick hot take? Pick one that you think describes you best.
STAYCATION or VACATION
wats a staycation?
I'm a staycation kinda person. A lot less travel time.
Type a message...
GIF

Xiaoice

- https://www.youtube.com/watch?v=dg-x1WuGhuI

Alexa Prize Challenge

- Challenge: Build a chatbot that engages the users for 20 mins.
- Sponsored 12 University Teams with $100k.
- CMU Magnus and CMU Ruby.
- Systems are multicomponent
    - Combinations of task/non-task
    - Hand-written and statistical/neural models
- Its about engaging researchers
    - Having more PhD students do dialog
    - Giving access for developers to users
    - Collecting data: what do users say

CMU Magnus

- High average number of turns
- Average Rating
- Topics: Movies, Sports, Travel, GoT
- Users had longer conversations but did not enjoy the conversation.
    - Identify when user is frustrated or wants to change topic.
    - Identify what the user would like to talk about (intent).
- Detecting "Abusive" remarks and responding appropriately

Summary

- How to represent words in continuous space.
- What are RNNs and how to use them to represent a sentence.
- Sequence to sequence models for $P(response \mid input\_sentence)$
- Issues in neural model
- Issues with Live system!

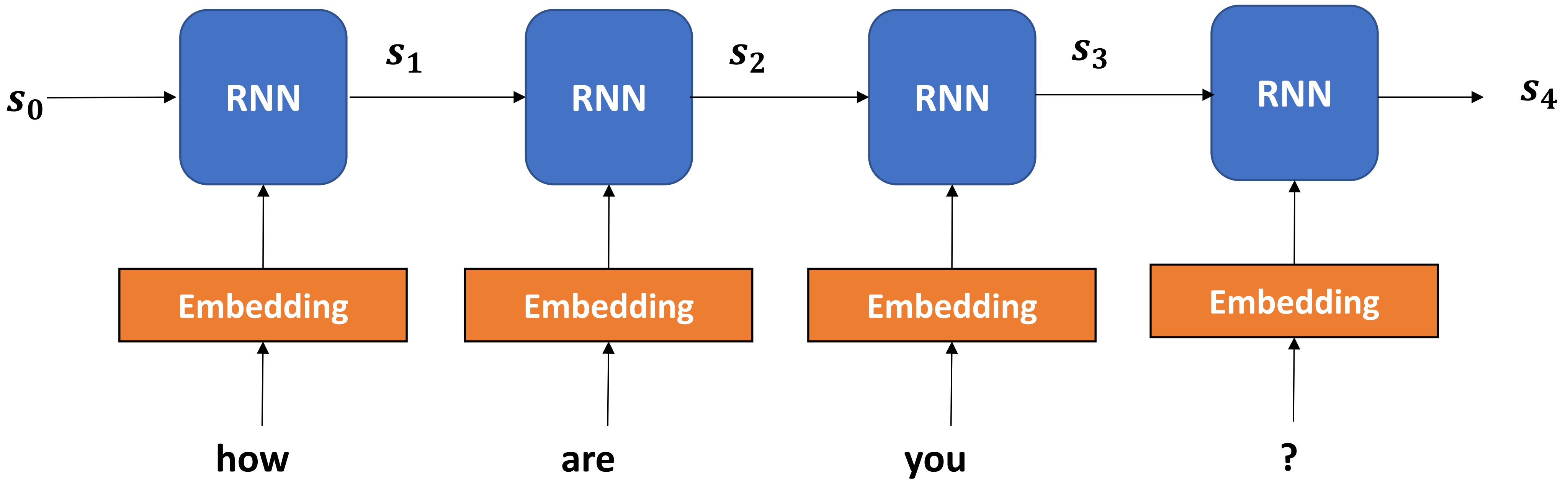
References

- http://www.phontron.com/class/nn4nlp2017/assets/slides/nn4nlp-03-wordemb.pdf
- http://www.phontron.com/class/nn4nlp2017/assets/slides/nn4nlp-06-rnn.pdf
- http://www.phontron.com/class/nn4nlp2017/assets/slides/nn4nlp-08-condlm.pdf
- https://www.cs.cmu.edu/~rsalakhu/10707/Lectures/Lecture_Language_2019.pdf
- http://www.phontron.com/class/mtandseq2seq2017/mt-spring2017.chapter6.pdf

References

- http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/
- http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/
- http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-2-implementing-a-language-model-rnn-with-python-numpy-and-theano/
- http://www.wildml.com/2016/07/deep-learning-for-chatbots-2-retrieval-based-model-tensorflow/
- https://nlp.stanford.edu/seminar/details/jdevlin.pdf

RNN to represent a sentence



- s_4 is the representation of the entire sentence
- s_4 is the representation of probability of observing "how are you?"