NLP Basics

Shangbin Feng

LUD Lab, Xi’an Jiaotong University

wind_binteng@stu.xjtu.edu.cn

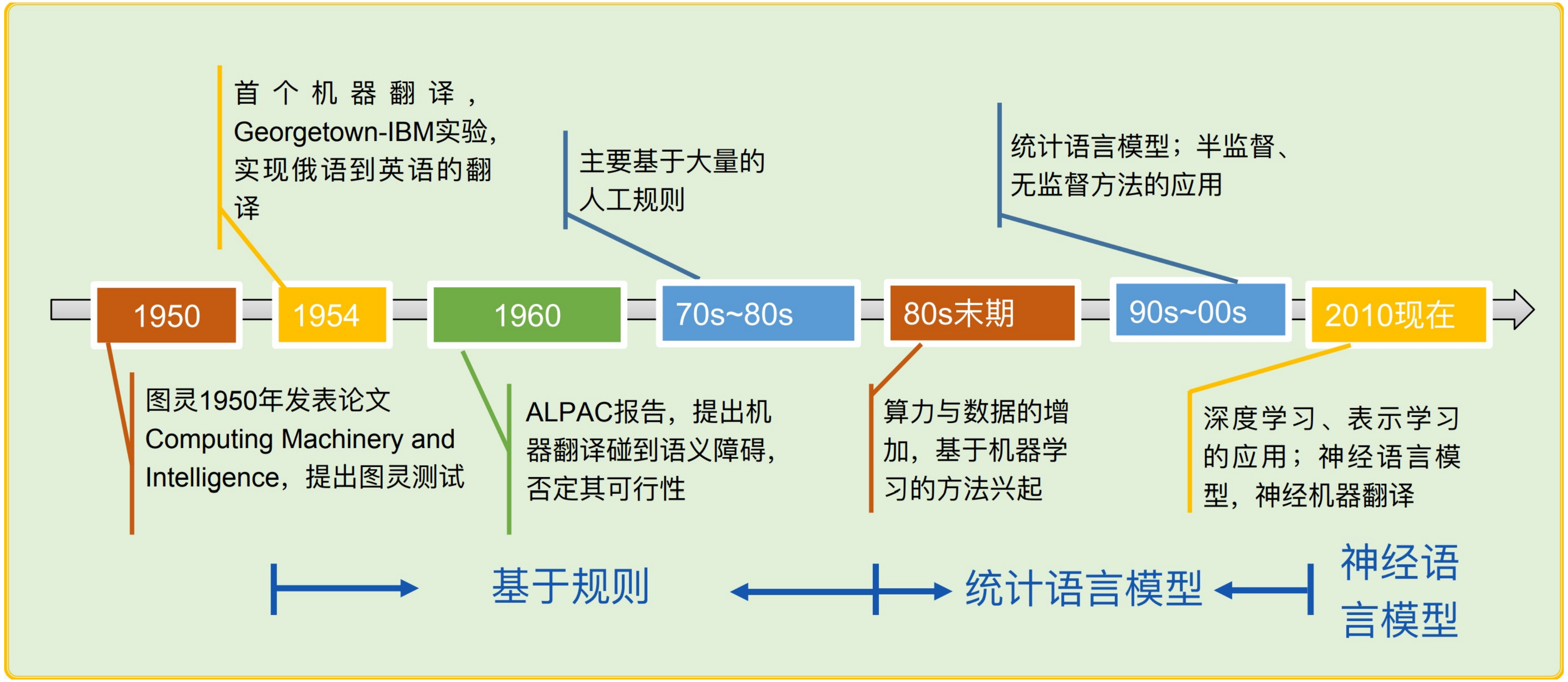
February 6, 2022

Contents

- Introduction to NLP
- Word Embedding and RNNs
- Attention and Transformers
- Self-supervised Learning
- Pre-trained Language Models

What is NLP?

- Natural Language Processing



1950

图灵1950年发表论文Computing Machinery and Intelligence，提出图灵测试

1954

首个机器翻译，Georgetown-IBM实验，实现俄语到英语的翻译

1960

ALPAC报告，提出机器翻译碰到语义障碍，否定其可行性

70s~80s

主要基于大量的人工规则

80s末期

算力与数据的增加，基于机器学习的方法兴起

90s~00s

统计语言模型；半监督、无监督方法的应用

2010现在

深度学习、表示学习的应用；神经语言模型，神经机器翻译

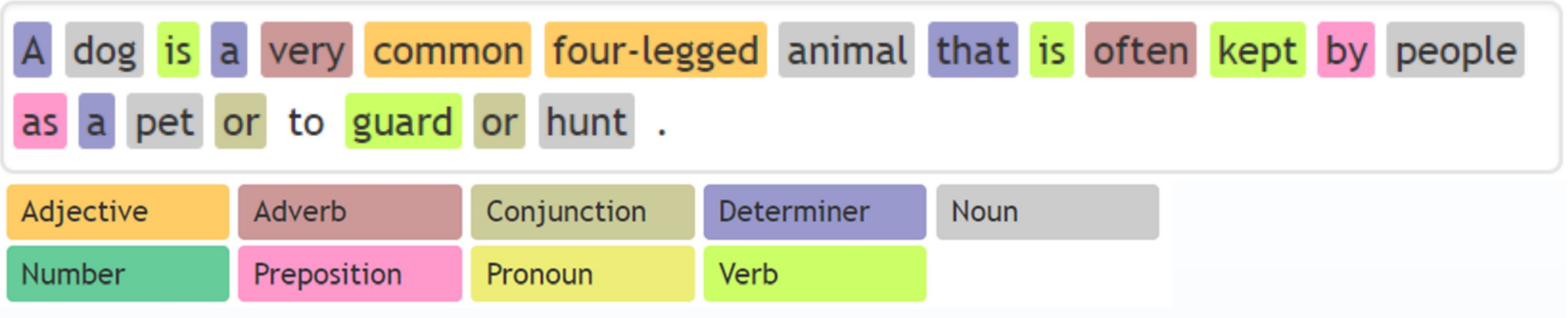
基于规则

统计语言模型

神经语言模型

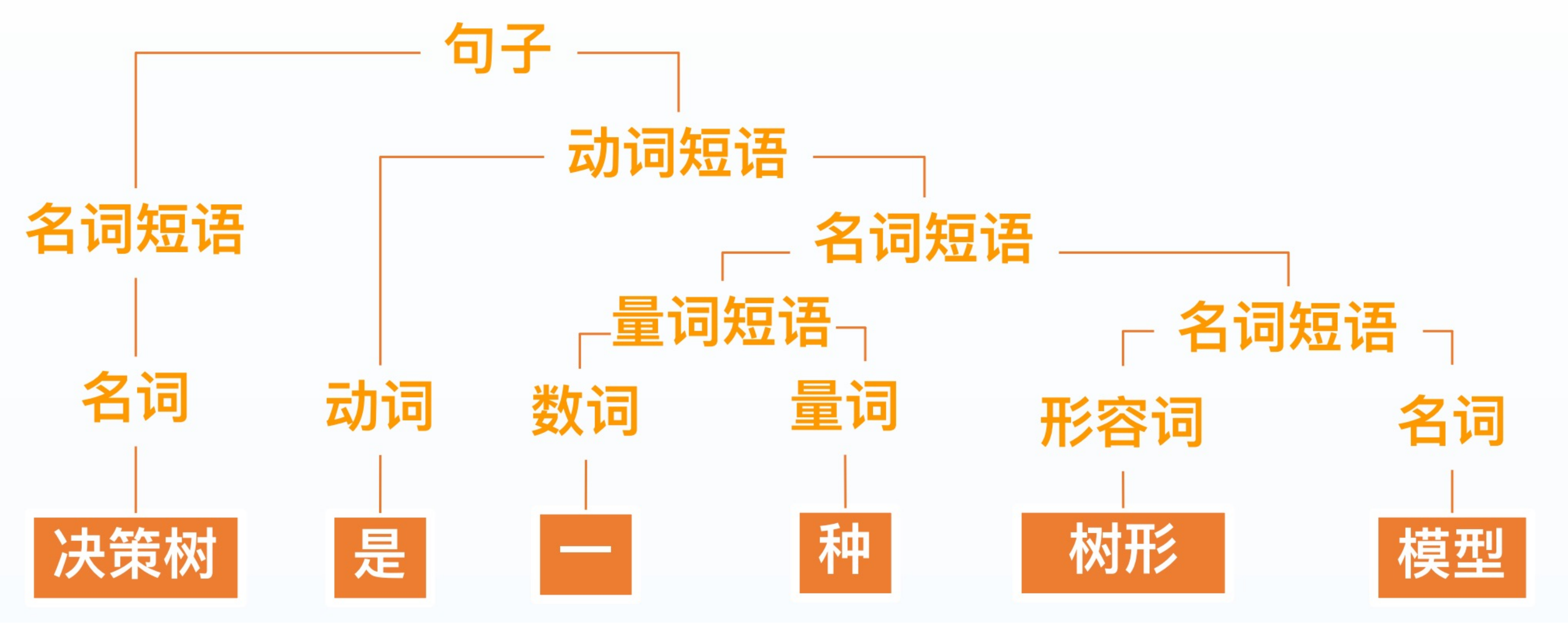
https://en.wikipedia.org/wiki/Natural_language_processing#History

Task 1: PoS Tagging

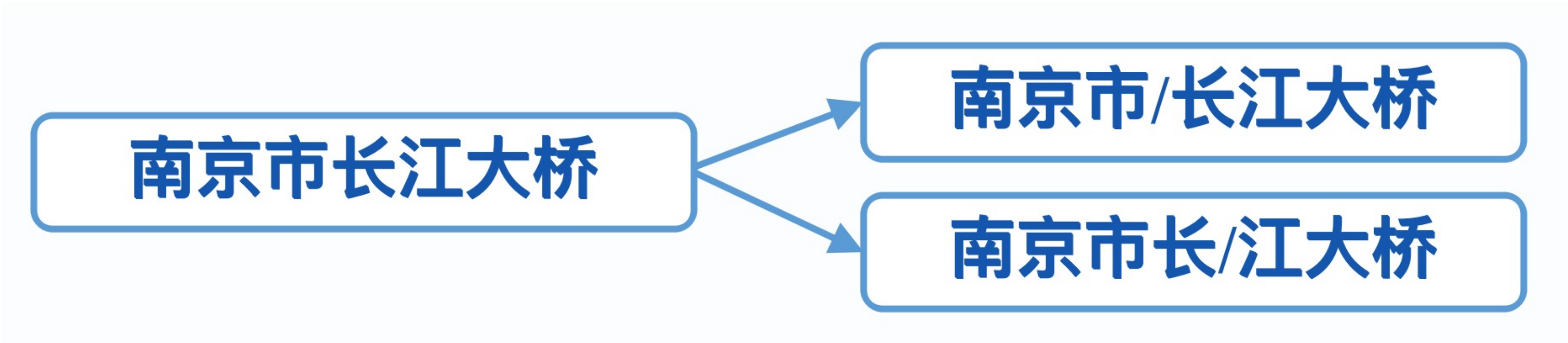


A dog is a very common four-legged animal that is often kept by people as a pet or to guard or hunt .

Task 2: Parsing



Task 3: Tokenization

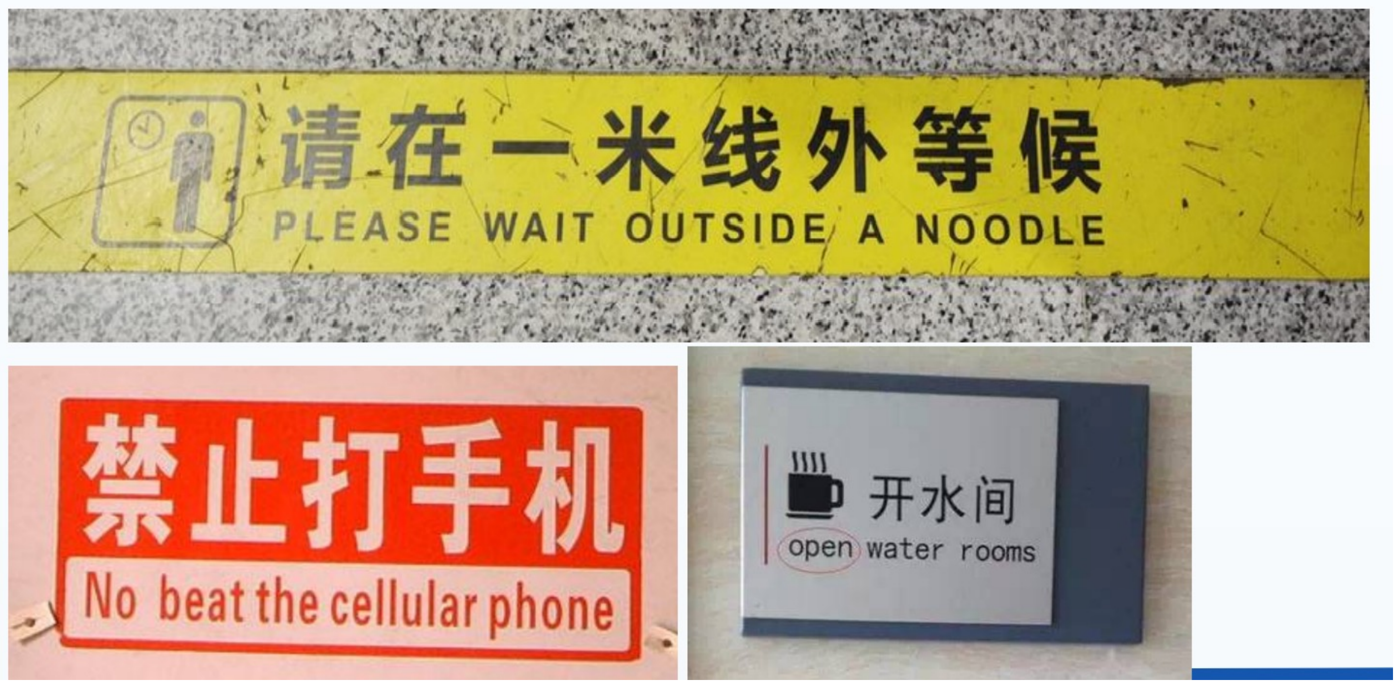


南京市长江大桥

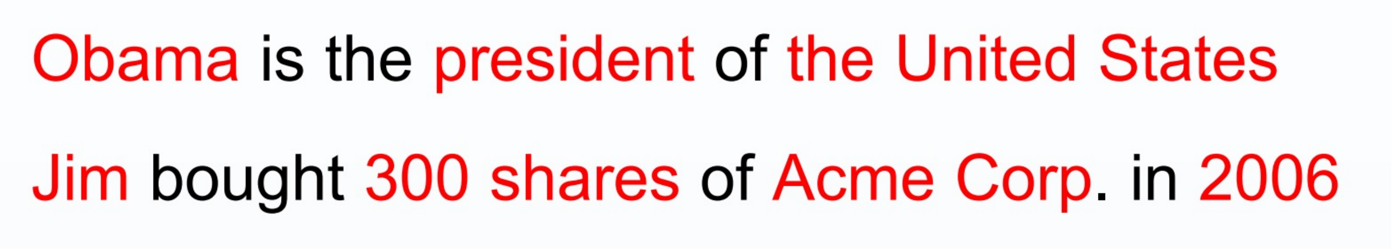
南京市/长江大桥

南京市长/江大桥

Task 4: Machine Translation



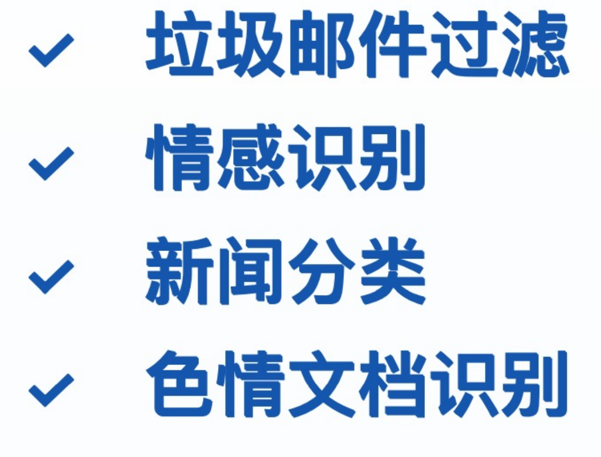
Task 5: Named Entity Recognition



Obama is the president of the United States

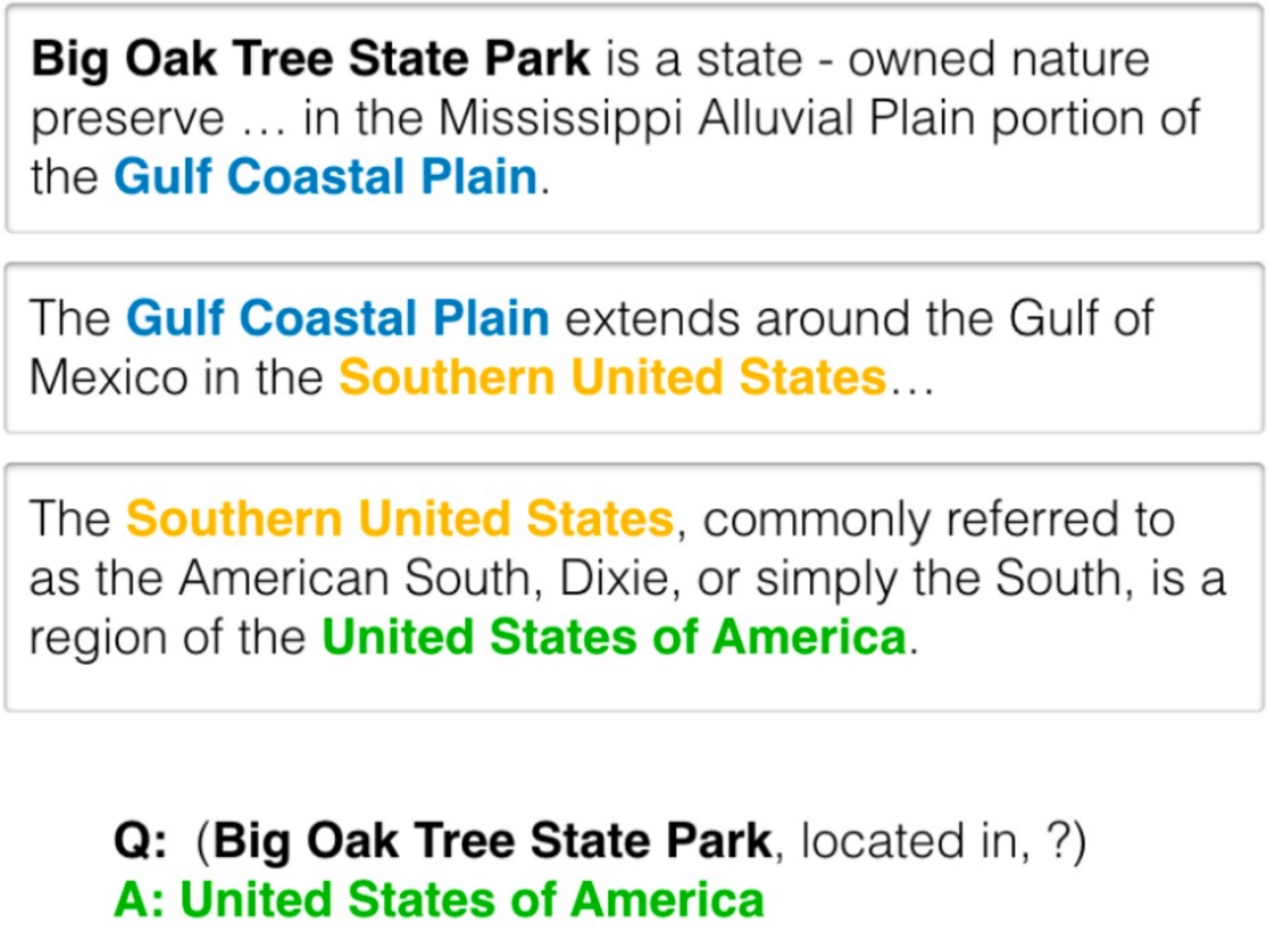
Jim bought 300 shares of Acme Corp. in 2006

Task 6: Text Classification



- ✓ 垃圾邮件过滤
- ✓ 情感识别
- ✓ 新闻分类
- ✓ 色情文档识别

Task 7: Question Answering



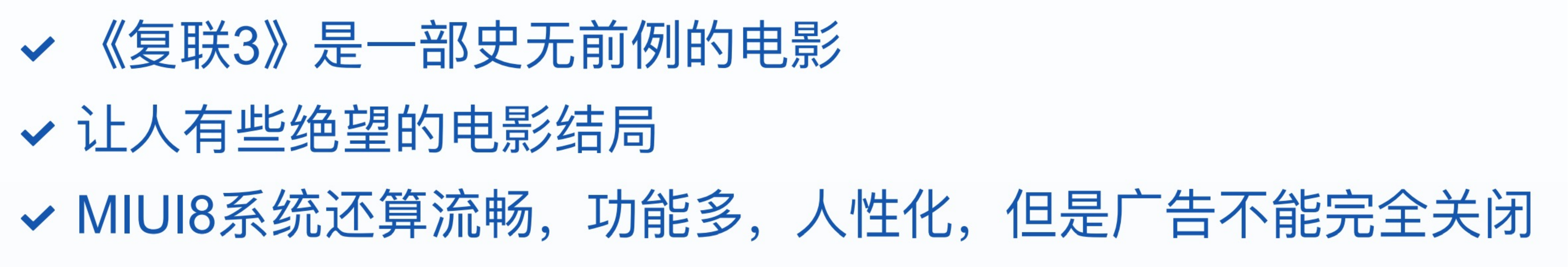
Big Oak Tree State Park is a state - owned nature preserve … in the Mississippi Alluvial Plain portion of the **Gulf Coastal Plain**.

The **Gulf Coastal Plain** extends around the Gulf of Mexico in the **Southern United States**…

The **Southern United States**, commonly referred to as the American South, Dixie, or simply the South, is a region of the **United States of America**.

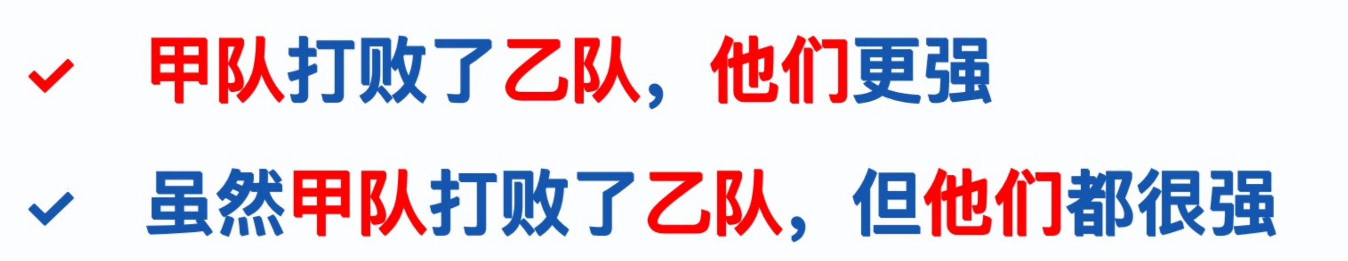
Q: (**Big Oak Tree State Park**, located in, ?)
A: United States of America

Task 8: Sentiment Analysis



- ✓ 《复联3》是一部史无前例的电影
- ✓ 让人有些绝望的电影结局
- ✓ MIUI8系统还算流畅，功能多，人性化，但是广告不能完全关闭

Task 9: Coreference Resolution



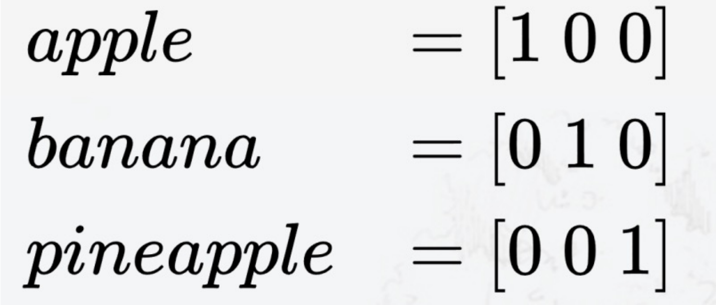
- ✓ 甲队打败了乙队，他们更强
- ✓ 虽然甲队打败了乙队，但他们都很强

Contents

- Introduction to NLP
- Word Embedding and RNNs
- Attention and Transformers
- Self-supervised Learning
- Pre-trained Language Models

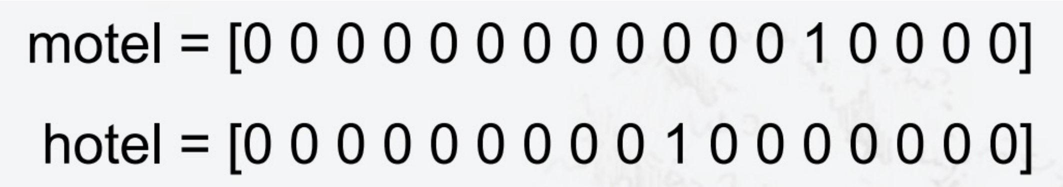
Why word embedding?

One-hot Encoding



$$\begin{aligned} apple &= [1\ 0\ 0] \\ banana &= [0\ 1\ 0] \\ pineapple &= [0\ 0\ 1] \end{aligned}$$

- Problems
  - Embedding size
  - transductive
  - No meaning in it

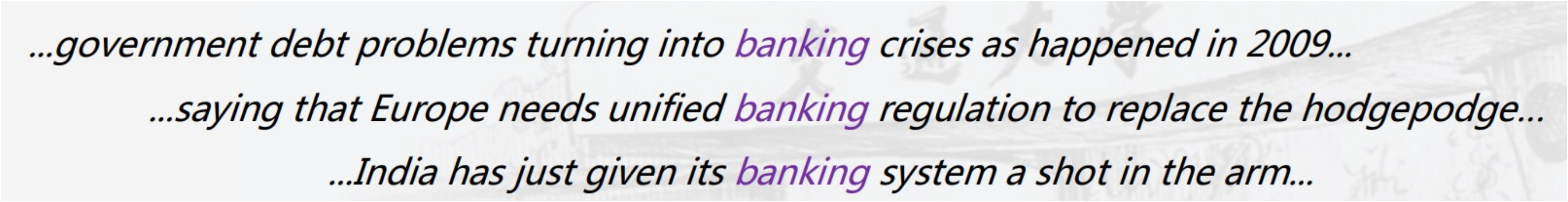


motel = [0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

hotel = [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Idea

- “You shall know a word by the company it keeps.”



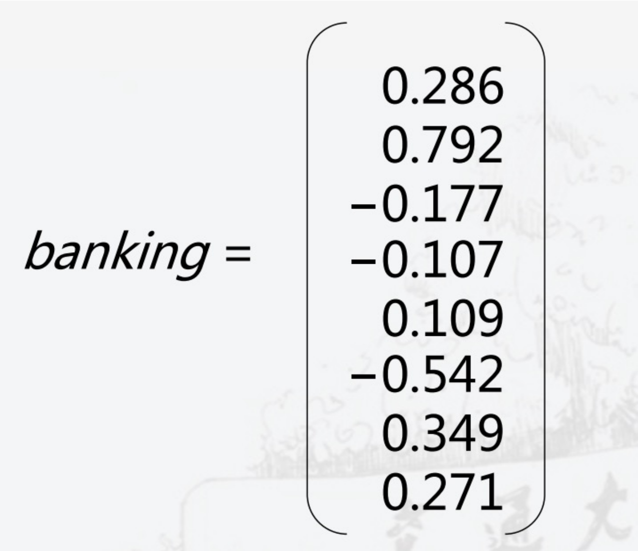
...government debt problems turning into banking crises as happened in 2009...

...saying that Europe needs unified banking regulation to replace the hodgepodge...

...India has just given its banking system a shot in the arm...

Word Embedding

- Distributed embedding
- Dense vector
- Word vector
- Word representation
- Distributed representation



$$\textit{banking} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

- Limited dimension / word meaning included

Word2Vec

- “You shall know a word by the company it keeps”
  - CBOW
  - Skip-gram

CBOW

- Continuous bag of words

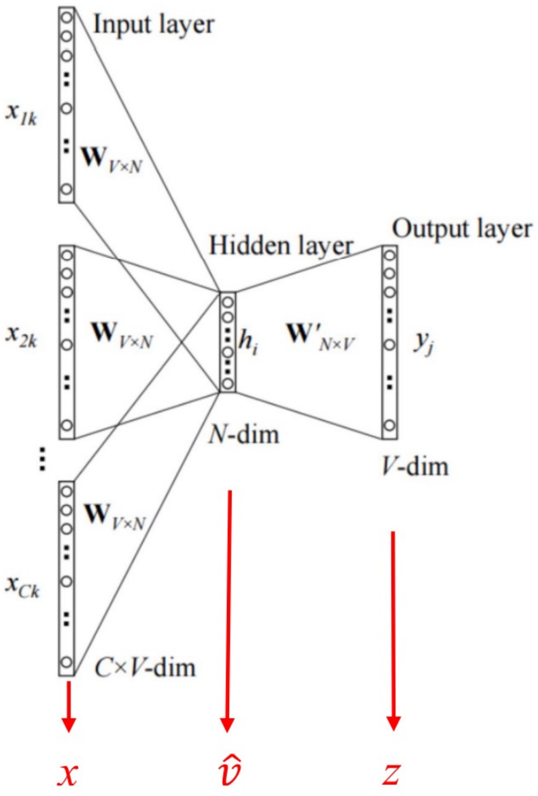
1. Generate one-hot word vectors for the input context of size m: $(x^{c-m}, \dots, x^{c-1}, x^{c+1}, \dots, x^{c+m} \in R^{|V|})$.

2. Get embedded word vectors for the context.
   $v_{c-m} = x^{c-m} \cdot W, \ldots, v_{c+m} = x^{c+m} \cdot W \in R^N$

3. Average context word vectors get $\hat{v}$.
   $\hat{v} = \frac{v_{c-m+\cdots+}v_{c+m}}{2m} \in R^N$

4. Generate a score vector z.
   $z = \hat{v} \cdot W' \in R^{|V|}$

5. Turn the score vector into probabilities $\hat{y}$.
   $\hat{y} = softmax(z)$



Skip-gram

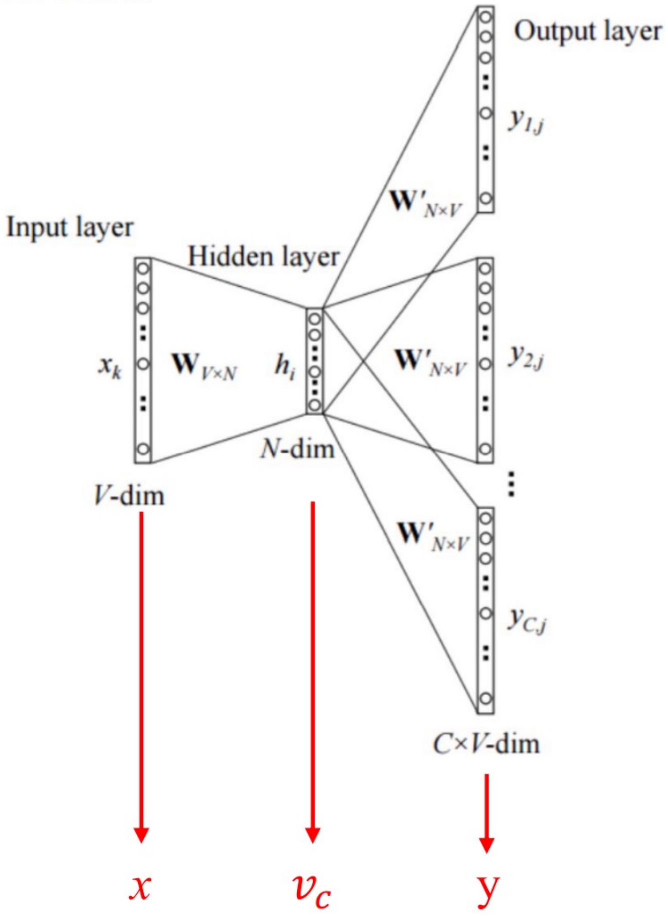
1. Generate one-hot input vector $x \in R^{|V|}$ of the center word.

2. Get embedded word vectors for the center word.
   $v_c = x \cdot W \in R^N$

3. Generate a score vector z.
   $z = v_c \cdot W'$

4. Turn the score vector into probabilities $\hat{y}$.
   $\hat{y} = softmax(z)$

5. Note that $\hat{y}_{c-m}, \ldots, \hat{y}_{c-1}, \hat{y}_{c+1}, \ldots, \hat{y}_{c+m}$ are the probabilities of observing context word.



Summary of Word2Vec

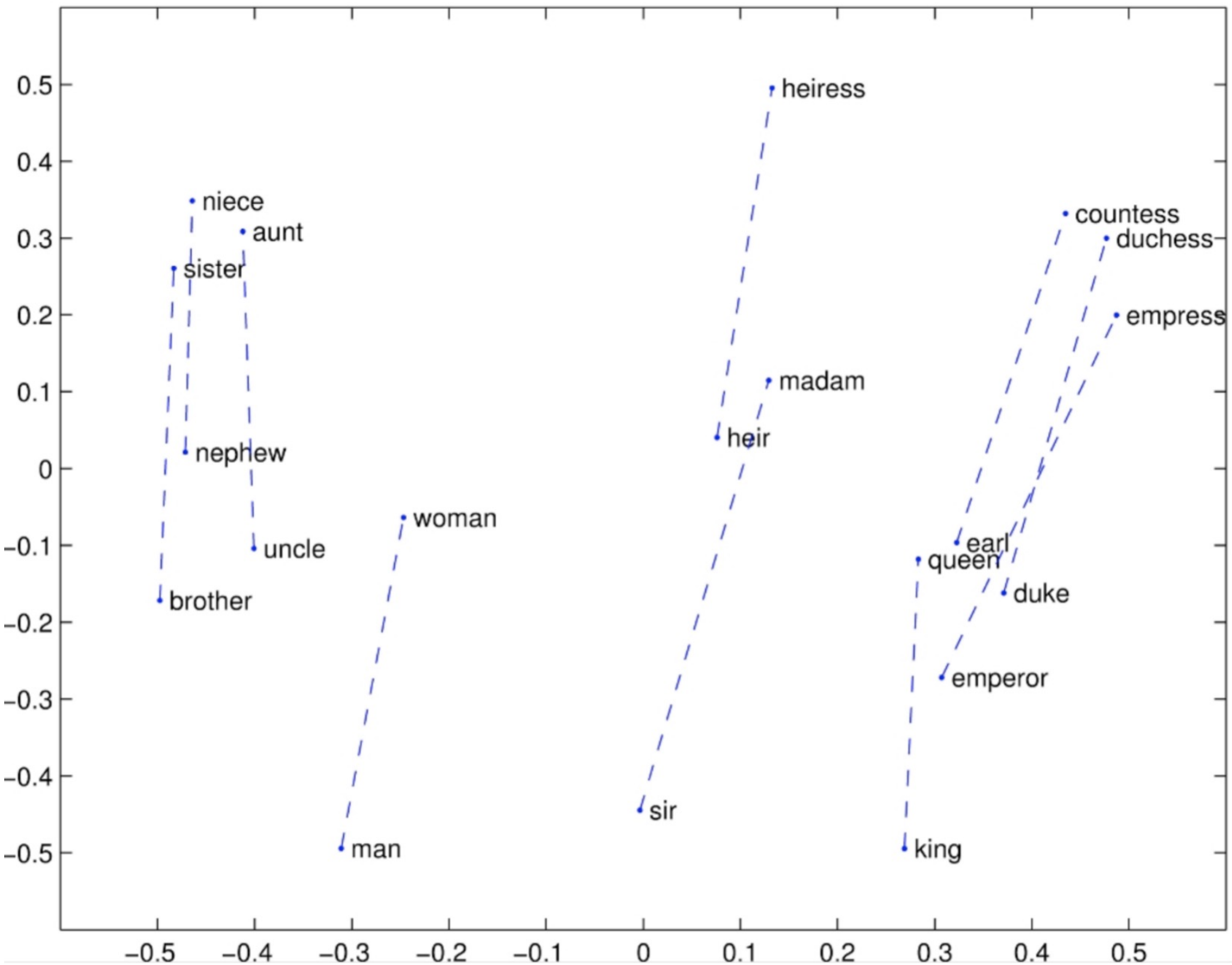
- What’s the idea?
- What’s the difference between CBOW and Skip-gram?

- Problem of vocabulary size
    - Negative sampling & hierarchical softmax

How do we evaluate word embeddings?

- Intrinsic evaluation
- Extrinsic evaluation

Intrinsic Evaluation



Extrinsic Evaluation

- Use word vectors on downstream tasks

Glove

- Word embeddings by Christopher Manning @ Stanford

- Global Vectors for Word Representations
- https://nlp.stanford.edu/projects/glove/

- Download
  - Code
  - Trained word embeddings

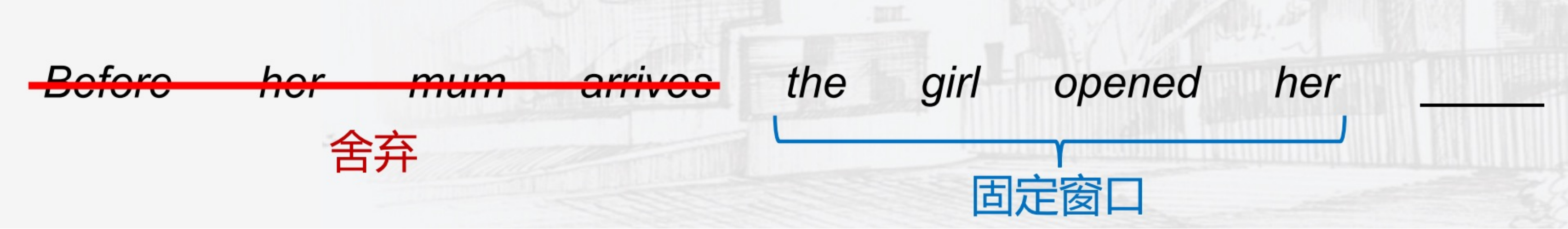
Language Model (LM)

- The girl opened her ______
  - Laptop
  - Books
  - …

- A language model tries to **predict the next word (token)** given the previous token sequence.

$$P(x^{(t+1)}|x^{(t)}, \ldots, \ x^{(1)})$$

Context window



~~Before her mum arrives~~ the girl opened her ______

舍弃

固定窗口

Linear Layer LM

- Problem
  - Window size
  - W parameter size
  - Never enough W

输出层

$$\hat{y} = softmax(Uh + b_2) \in \mathbb{R}^{|V|}$$

隐藏层

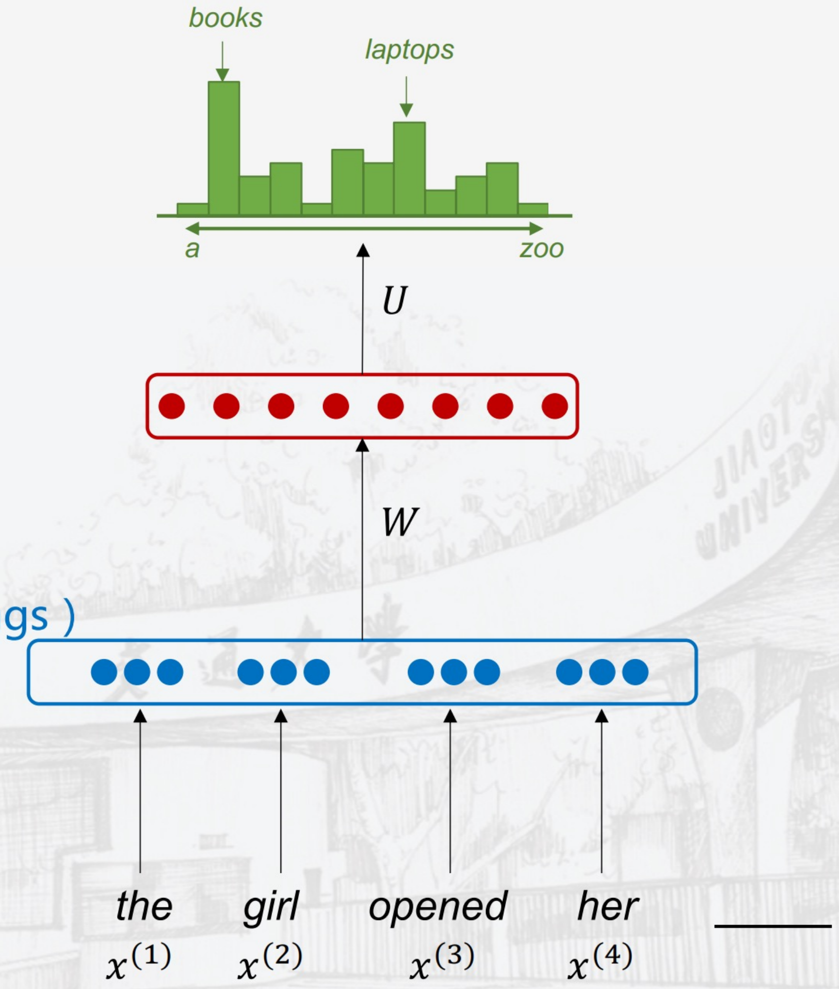
$$h = f(We + b_1)$$

连接嵌入式词向量（ word embeddings ）

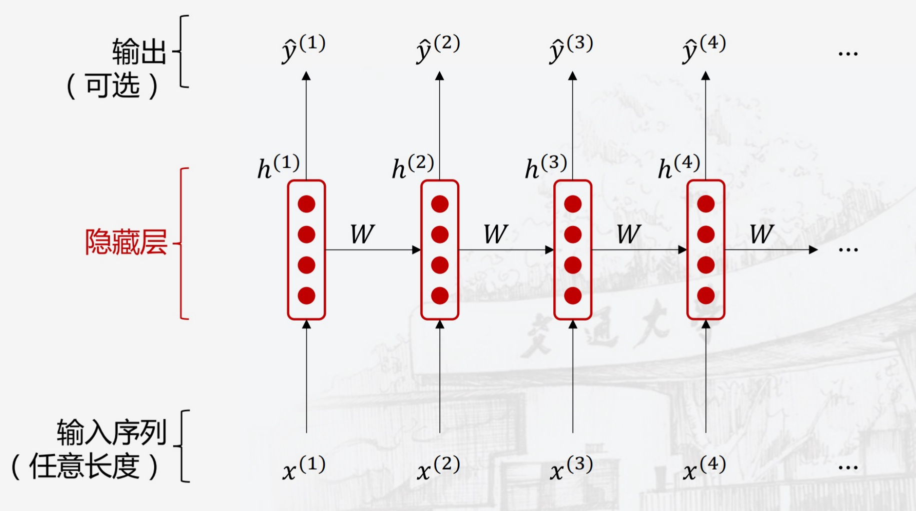
$$e = [e^{(1)};\ e^{(2)}; e^{(3)};\ e^{(4)}]$$

词向量（ one-hot、分布式表示 ...... ）

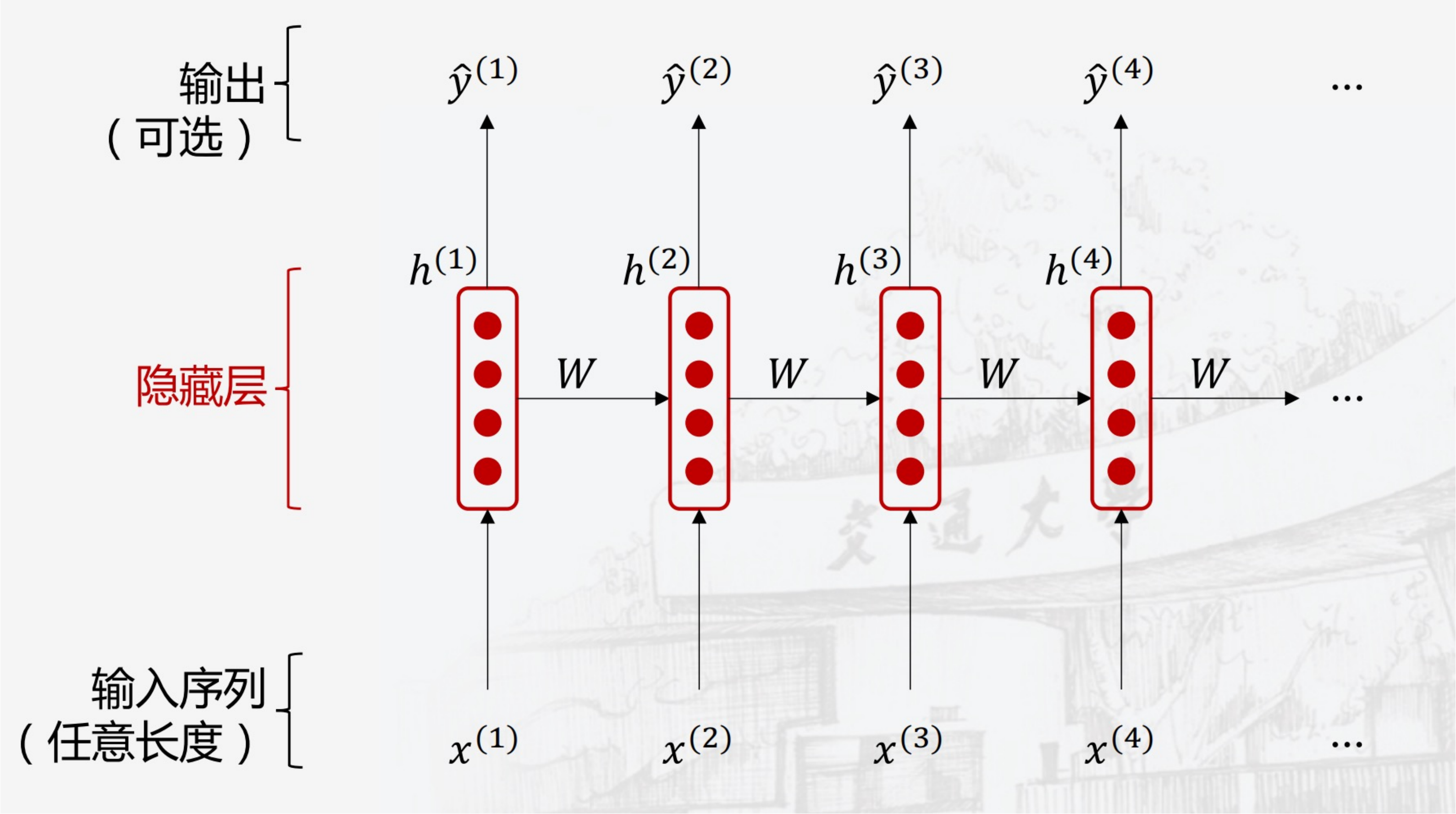
$$x^{(1)}, x^{(2)}, x^{(3)},\ x^{(4)}$$



Recurrent Neural Networks (RNNs)



Recurrent Neural Networks (RNNs)



RNN LM

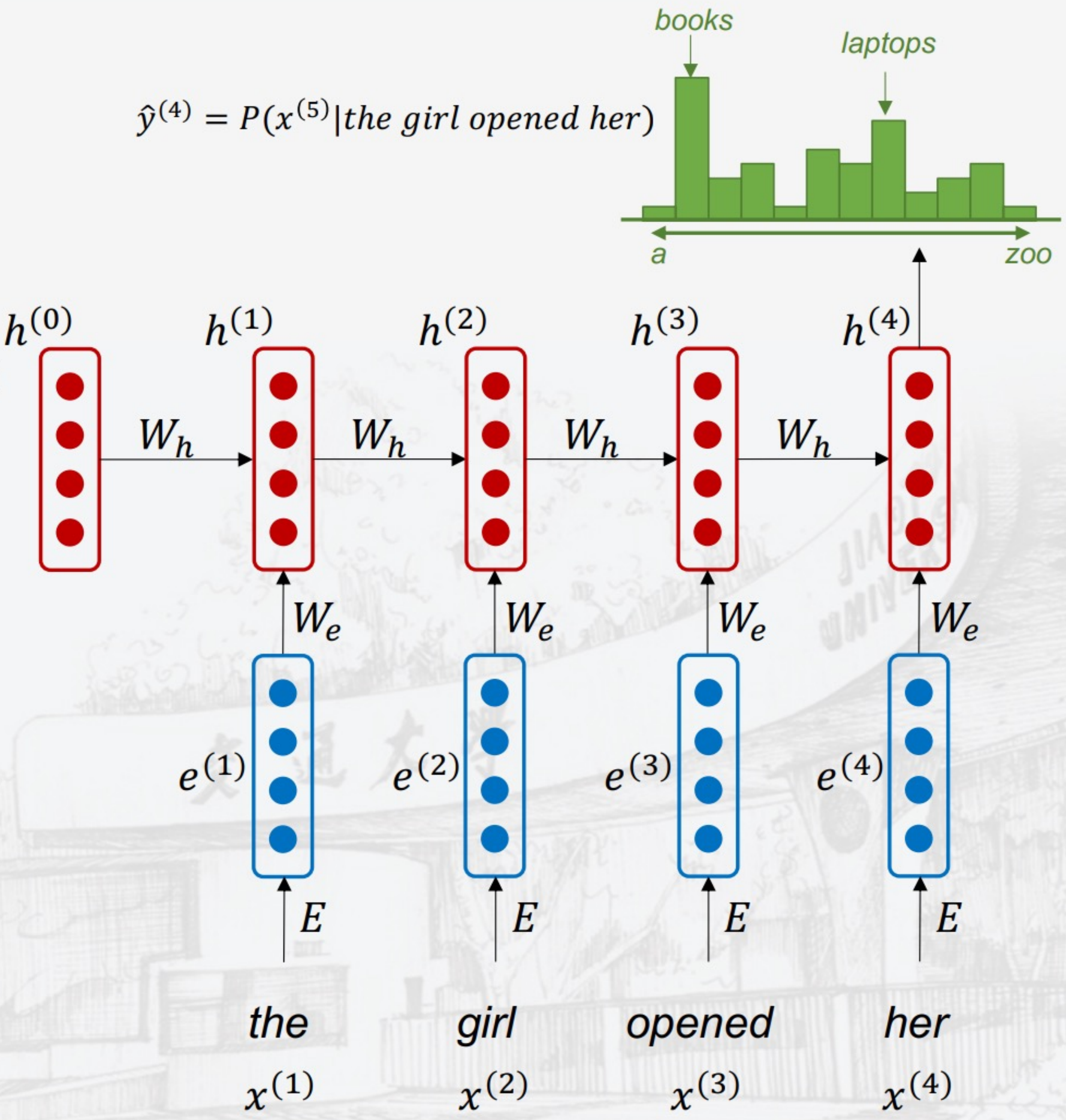
基于递归神经网络的LM

优点

- 可处理任意长度句子；
- 第 t 步的计算（理论上）使用了前面多步的信息；
- 模型体量不随着输入变长而增加；
- 每一步使用同一个 W ，降低计算量。

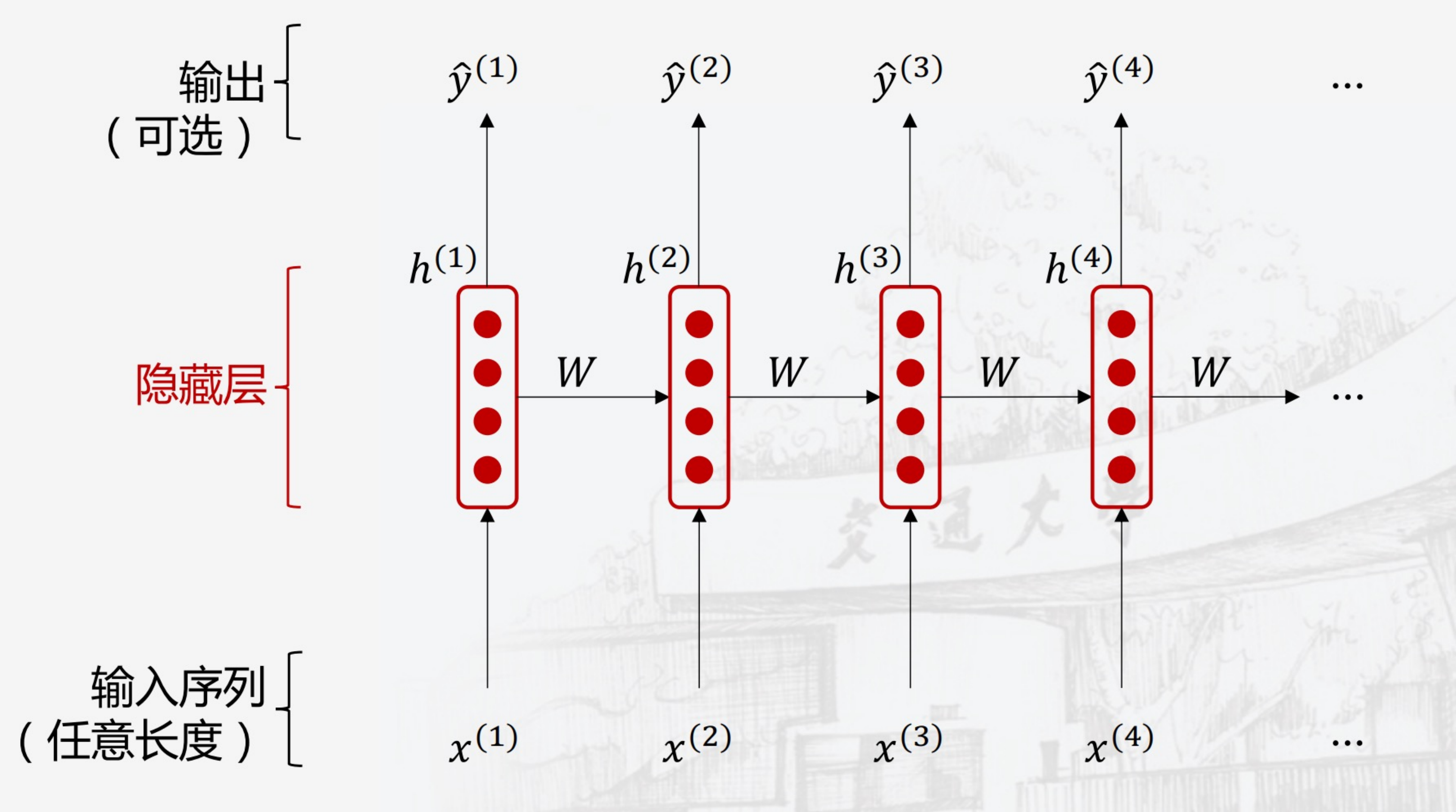
缺点

- 递归计算缓慢；
- 实际上，将前面很多步的信息完整传递是困难的。

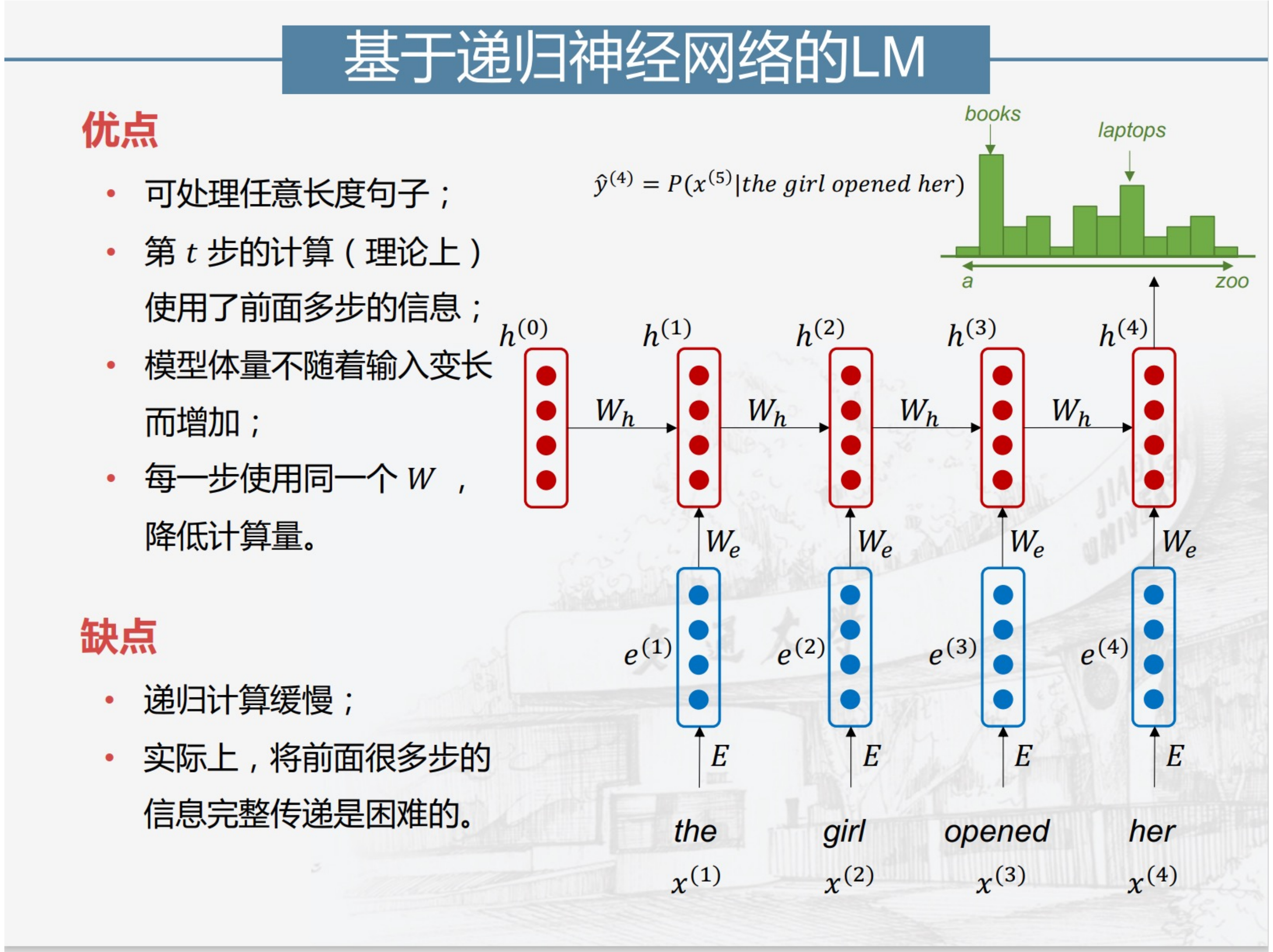


Important Issues

- When do we initialize W?
- When do we initialize h(0)?



Text Classification with RNN



基于递归神经网络的LM

优点

- 可处理任意长度句子；
- 第 t 步的计算（理论上）使用了前面多步的信息；
- 模型体量不随着输入变长而增加；
- 每一步使用同一个 W ，降低计算量。

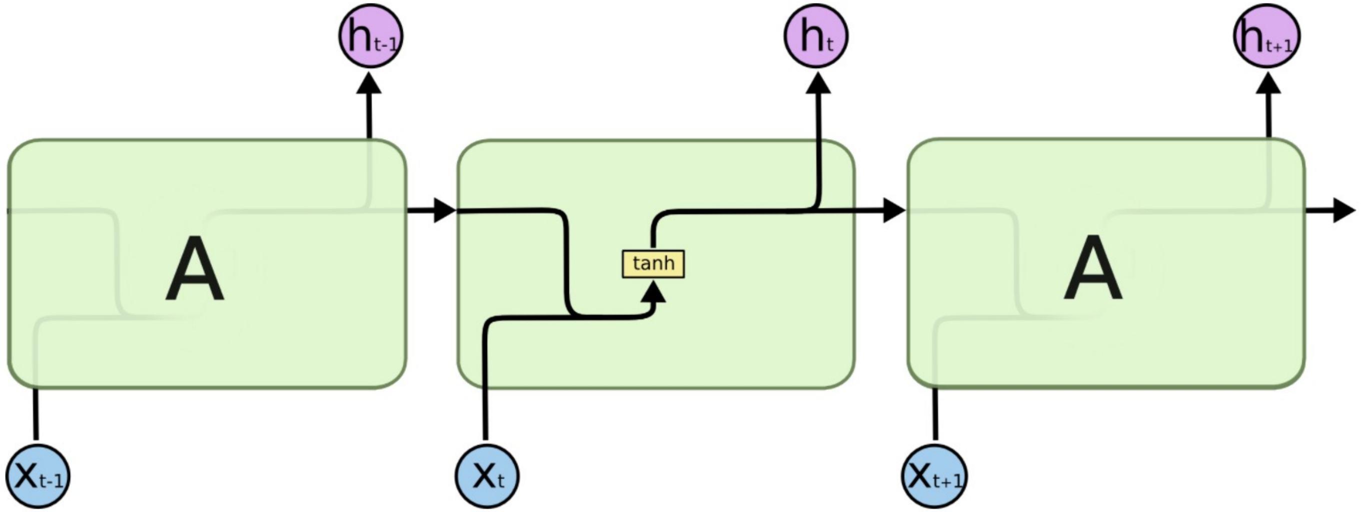
缺点

- 递归计算缓慢；
- 实际上，将前面很多步的信息完整传递是困难的。

RNN Problems

- Long-term dependencies

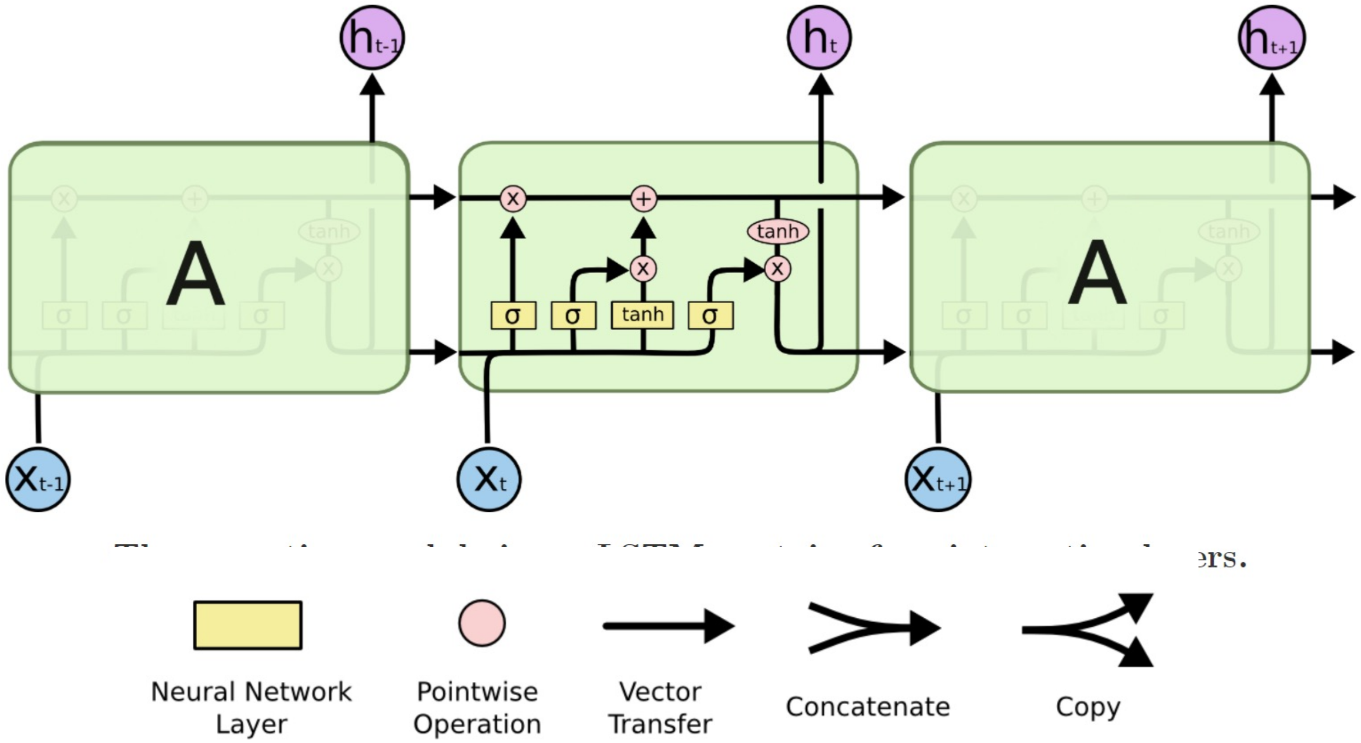
- Vanishing gradient problem



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

LSTM comes to the rescue

- Long short-term memory



Summary

- NLP tasks
- “You shall know a word by the company it keeps”
- Word2Vec
  - CBOW
  - Skip-gram
- Language Model
- RNN and LSTM

Project Tip

- Given textual input and word embeddings,
    - Approach A: average word embeddings as language representation
    - Approach B: RNNs + word embeddings

- A or B?
- Issues
    - OOV (out-of-vocabulary)
    - Happy, happier, happiest, …
    - .,?/~!@
    - …

Thx for Attention

Shangbin Feng

LUD Lab, Xi’an Jiaotong University

wind_binteng@stu.xjtu.edu.cn

February 6, 2022