组会（2019秋）

Word2vec简介

张义策

2019/09/07

0、welcome

欢迎各位来到ITNLP实验室。

1、符号与信号

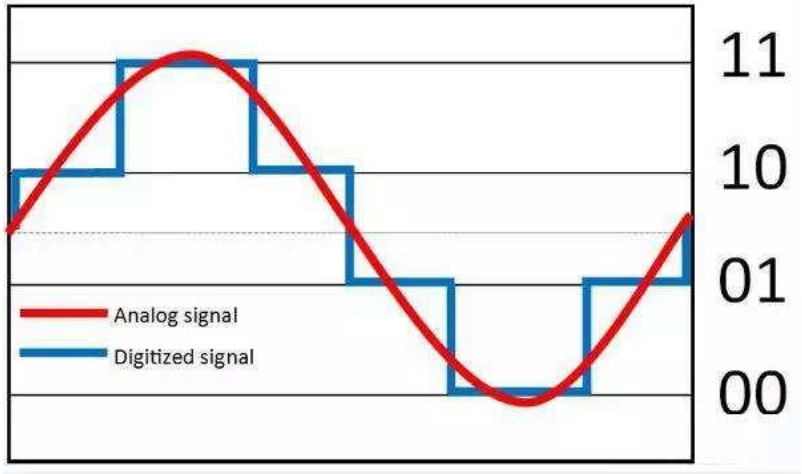
1、符号与信号

符号 符号是一种象征物，用来指称和代表其他事物。

α, β,新,新技术楼

信号 信号是一种表示消息的物理量。

数字信号
模拟信号



1、符号与信号

符号还是信号?

(1) yes的时间波形

(2) 一句话　　I'm fine, thank you. And you?

(3) 词表中的序号　　221 0 100 653 359 523 654 1 523 655

序号	单词
0	am
1	and
100	fine
221	i
359	thank
523	you
653	,
654	.
655	?

1、符号与信号

- 符号和信号都是信息的载体。
- 图像、语音是信号，而自然语言是符号。

- 信号可以自然地输入到神经网络中，而符号则不行。
- 我们需要对自然语言进行进一步的表示，进而输入到神经网络中。

- 将词表示成向量?

i you	含义比较相似	(1,1,0,0) (1,0,0,0)
, . ?	含义比较相似	(0,0,1,0) (0,0,1,1) (0,0,1,1.5)

2、词嵌入

2、词嵌入 – 分布假设

分布假设

1954年，Harris提出分布假说：

上下文相似的词，其语义也相似。

这是词的分布表示的理论基础。

统计语言模型

给定句子 $s = w_1 w_2 \cdots w_T$ ， 那么有

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)\cdots p(w_T|w_1 \cdots w_{T-1})$$

引入马尔科夫假设

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_2)\cdots p(w_T|w_{T-1})$$

二阶马尔可夫假设

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)\cdots p(w_T|w_{T-2} w_{T-1})$$

统计语言模型

二阶马尔可夫假设

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\cdots p(w_T|w_{T-2}w_{T-1})$$

$$p(w_3|w_1w_2) = \frac{p(w_1w_2w_3)}{p(w_1w_2)}$$

$$p(w_1w_2w_3) \approx f(w_1w_2w_3) = \frac{\#w_1w_2w_3}{N_3}$$

$$p(w_1w_2) \approx f(w_1w_2) = \frac{\#w_1w_2}{N_2}$$

存在稀疏性问题!

2、词嵌入 – NNLM

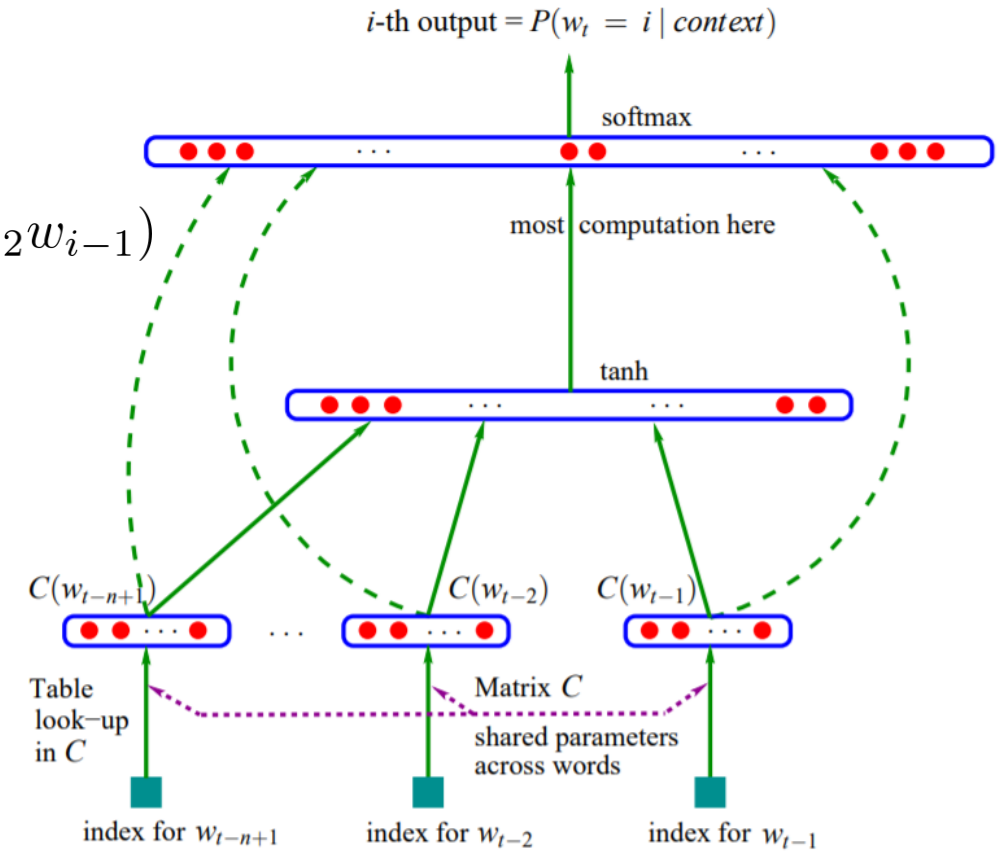
神经网络语言模型 NNLM[1]

使用神经网络计算

$$p(w_i|\text{context}_i) = p(w_i|w_{i-n+1}\cdots w_{i-2}w_{i-1})$$

目标是最大化 $p(w_i|\text{context}_i)$

词向量是其副产品。



[1] A neural probabilistic language model(Bengio Y等, JMLR2003).

2、词嵌入 – NNLM

word2vec [2,3]

2013, Mikolov等人在NNLM的基础上提出了CBOW模型和skip-gram模型，着重在计算效率上进行了改进。

在这两个模型的基础上，同年google开源了一款训练词向量的高效工具，名为word2vec。

[2] Efficient estimation of word representations in vector space(Mikolov等, arXiv2013).
[3] Distributed representations of words and phrases and their compositionality (Mikolov等, nips2013).

3、word2vec

3、word2vec – CBOW

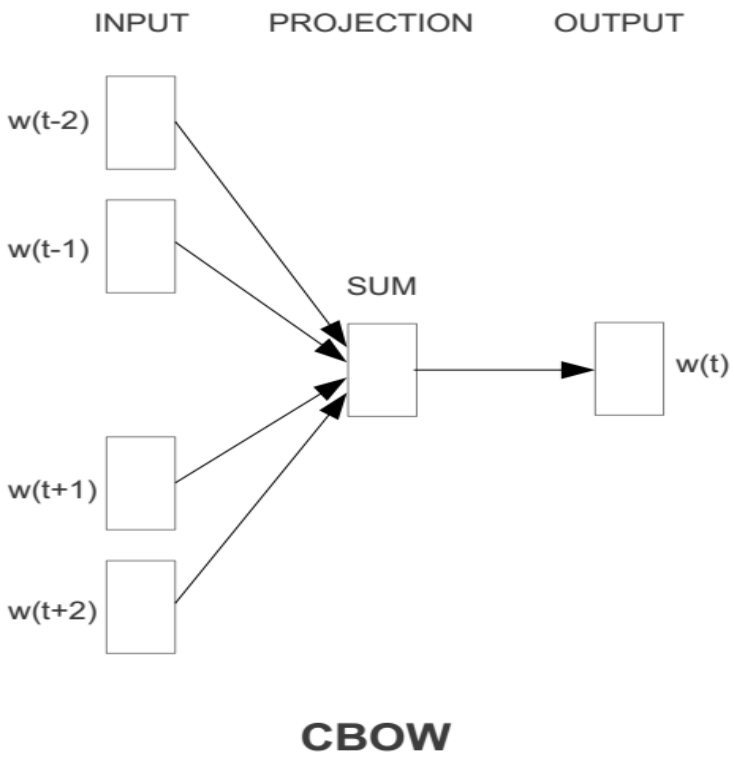
CBOW(Continuous Bag-of-Words Model)

不仅用到了上文，还用到了下文。

$$\mathbf{x}_{w(t)} = \sum_{i=1}^{c} \mathbf{v}_{w(t-i)} + \sum_{i=1}^{c} \mathbf{v}_{w(t+i)}$$

$$p(w|\text{context}(w)) = \text{softmax}(\mathbf{x}_w^\top \boldsymbol{\theta}_w)$$

训练目标，最大化 $p(w|\text{context}(w))$ 。



3、word2vec – skip-gram

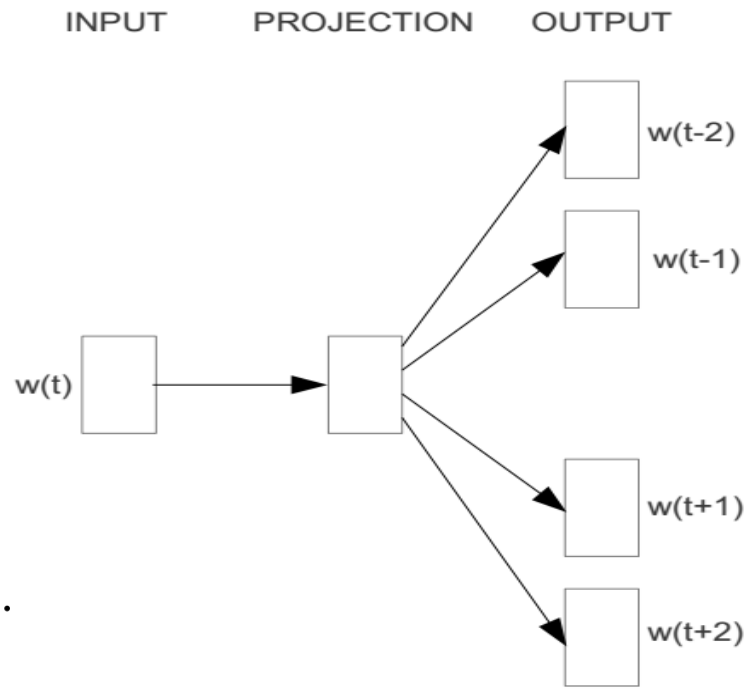
skip-gram

用单词预测上下文

$$p(w_i|w_j) = \text{softmax}(\mathbf{v}_{w_j}^{\top}\boldsymbol{\theta}_{w_i})$$

训练目标，最大化

$$p(\text{context}(w)|w) = \prod_{w^c \in \text{context}(w)} p(w^c|w).$$



Skip-gram

3、word2vec – softmax计算量太大

问题：softmax计算量太大

以skip-gram为例

$$p(w_i|w_j) = \text{softmax}(\mathbf{v}_{w_j}^\top \boldsymbol{\theta}_{w_i}) = \frac{e^{\mathbf{v}_{w_j}^\top \boldsymbol{\theta}_{w_i}}}{\sum_{k=1}^{|V|} e^{\mathbf{v}_{w_j}^\top \boldsymbol{\theta}_{w_k}}}$$

归一化需要太多计算量

两种解决方案：层次化softmax和负采样。

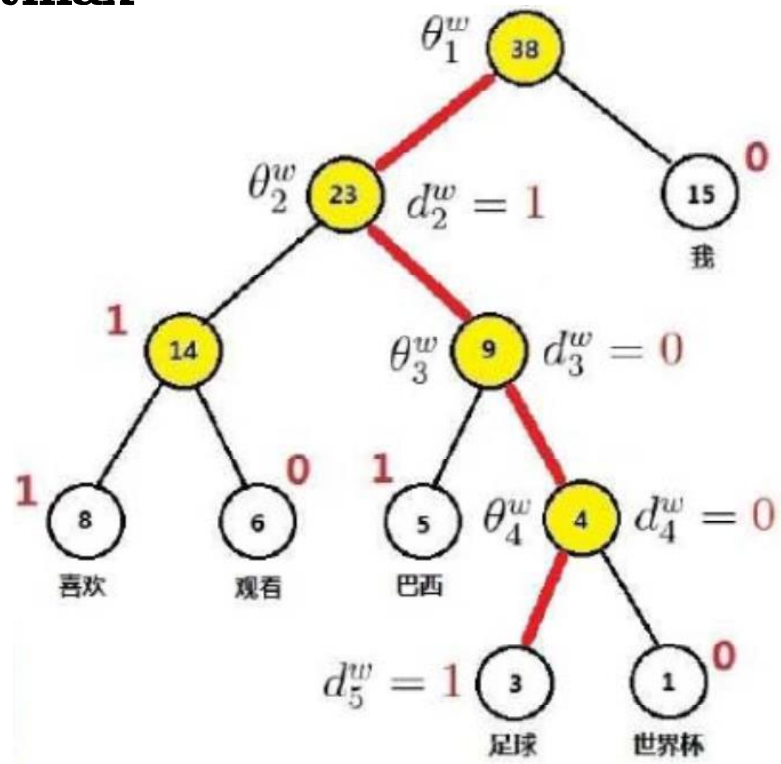
3、word2vec – 层次化softmax

问题：softmax计算量太大 – 层次化softmax

计算 $p(w_i|w_j)$ 成了若干个2分类问题。

$$p(d_k = 0|h_k, w_j) = \sigma(\mathbf{v}_{w_j}^\top \boldsymbol{\theta}_k)$$

$$p(d_k = 1|h_k, w_j) = 1 - \sigma(\mathbf{v}_{w_j}^\top \boldsymbol{\theta}_k)$$



$$p(\text{足球}|w_j) = p(d_2^w|h_{38}, w_j)p(d_3^w|h_{23}, w_j)p(d_4^w|h_9, w_j)p(d_5^w|h_4, w_j)$$

3、word2vec – 层次化softmax

问题：softmax计算量太大 – 层次化softmax

普通的softmax计算复杂度　$O(|V|)$

层次化softmax计算复杂度　$O(\log|V|)$

为什么现在都不用层次softmax了？

并行性不够友好。当显存足够大的时候，普通的softmax的时间复杂度为O(1)，而层次化softmax的时间复杂度还是 $O(\log|V|)$ 。

3、word2vec – 负采样

问题：softmax计算量太大 – 负采样

从希望最大化

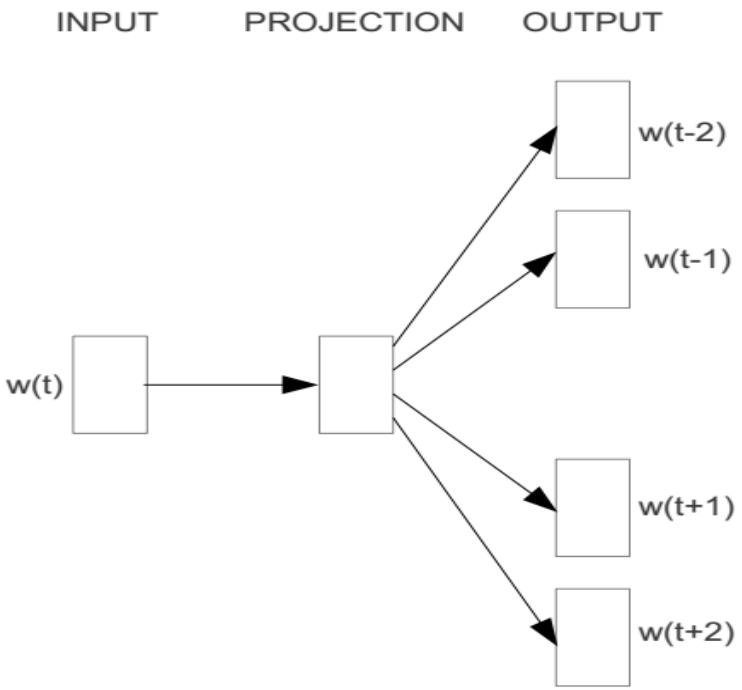
$$p(w_{t-2}|w_t) = \text{softmax}(\mathbf{v}_{w_t}^{\top}\boldsymbol{\theta}_{w_{t-2}})$$

变为希望

$$\sigma(\mathbf{v}_{w_t}^{T}\boldsymbol{\theta}_{w_{t-2}}) = 1$$

$$\sigma(\mathbf{v}_{w^{neg}}^{T}\boldsymbol{\theta}_{w_{t-2}}) = 0$$

其中，w^{neg} 为负样本，负样本是按照词频的0.75次方进行采样的。



Skip-gram

3、word2vec – 负采样

问题：softmax计算量太大 – 负采样

负样本是按照词频的0.75次方进行采样的

即一个词被采样到的概率

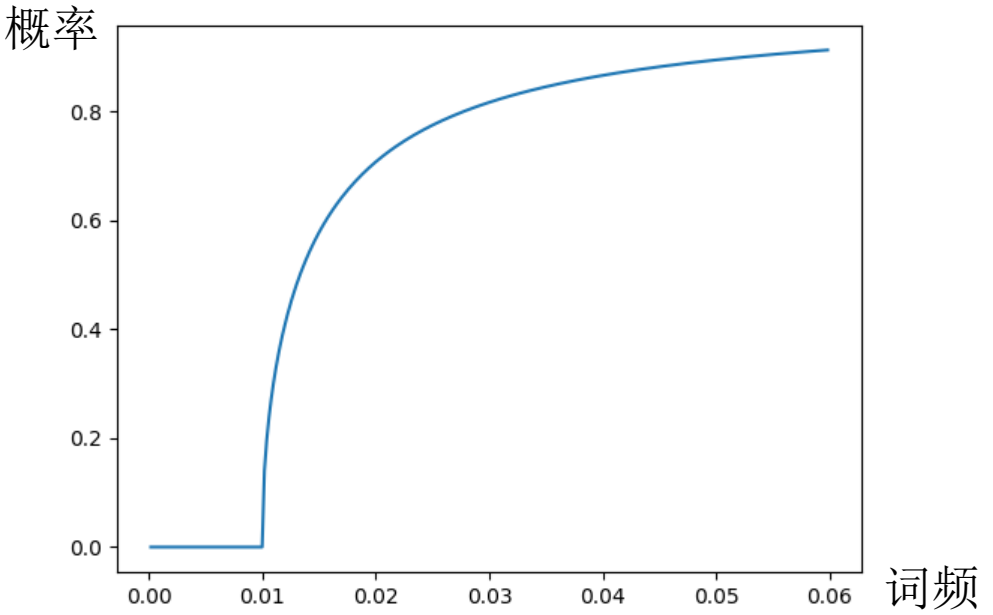
$$p_{\text{sample}}(w) = \frac{f(w)^{0.75}}{Z}.$$

3、word2vec – 高频词的处理

高频词如“的”、“是”所包含的有用信息较少，对这些高频词进行下采样可以提高训练速度

令 t 为词频阈值，一个词将以如下概率被丢弃

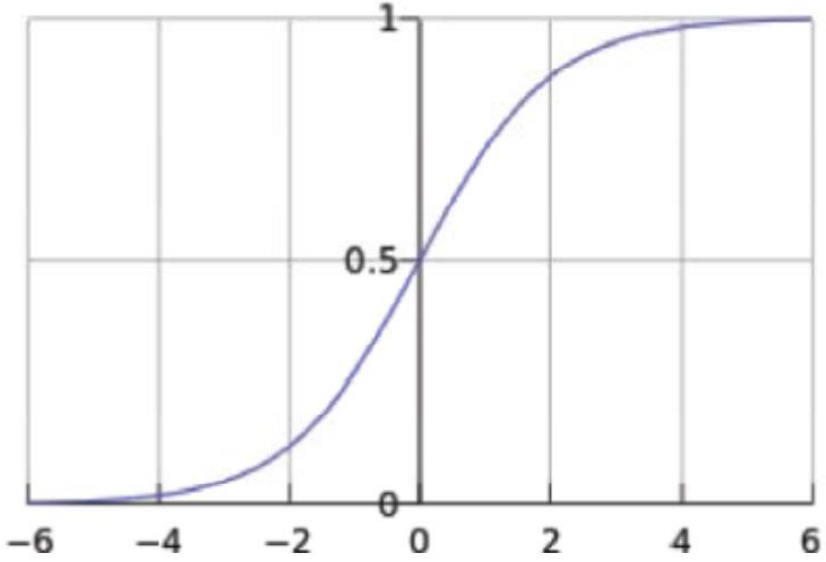
$$p_{\text{drop}}(w) = 1 - \sqrt{\frac{t}{f(w)}}$$

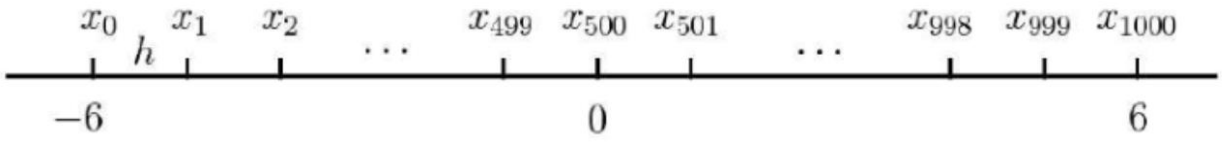


3、word2vec – sigmoid的近似

模型主要的计算量在sigmoid函数

$$\sigma(x) \approx \begin{cases} 0, & x \le -6 \\ \sigma(x_k), & x \in (-6, 6) \\ 1, & x \ge 6 \end{cases}$$





3、word2vec – 效果

4、word2vec 后续

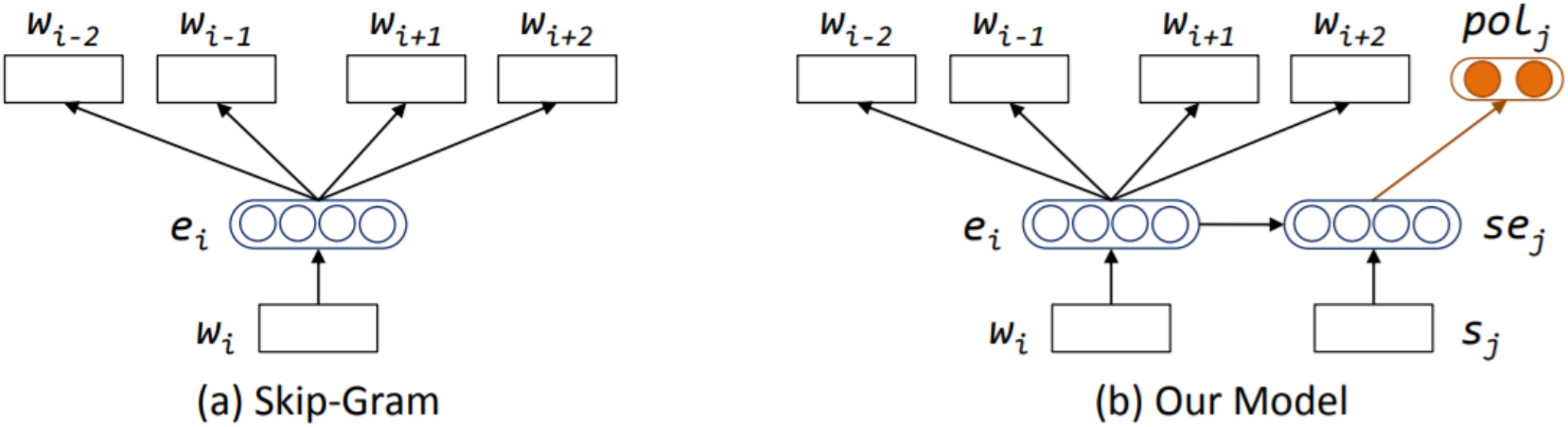
4、word2vec 后续 – glove

GloVe: Global Vectors

[4] Glove: Global vectors for word representation(Pennington等, EMNLP2014).

4、word2vec 后续 – SSPE

训练词嵌入时，预测情感倾向



[5] Learning sentiment-specific word embedding for twitter sentiment classification(Tang等, NAACL2014).
[6] Building large-scale twitter-specific sentiment lexicon: A representation learning approach (Tang等, coling2014).

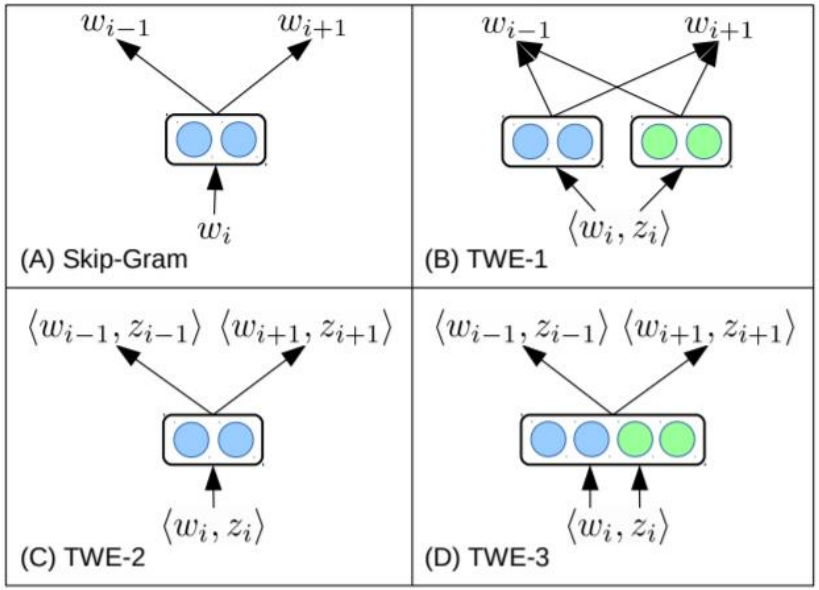
4、word2vec 后续 – TWE

考虑词的主题

歧义问题

晚饭前去打点酱油。

这次面试，我就是个打酱油的。

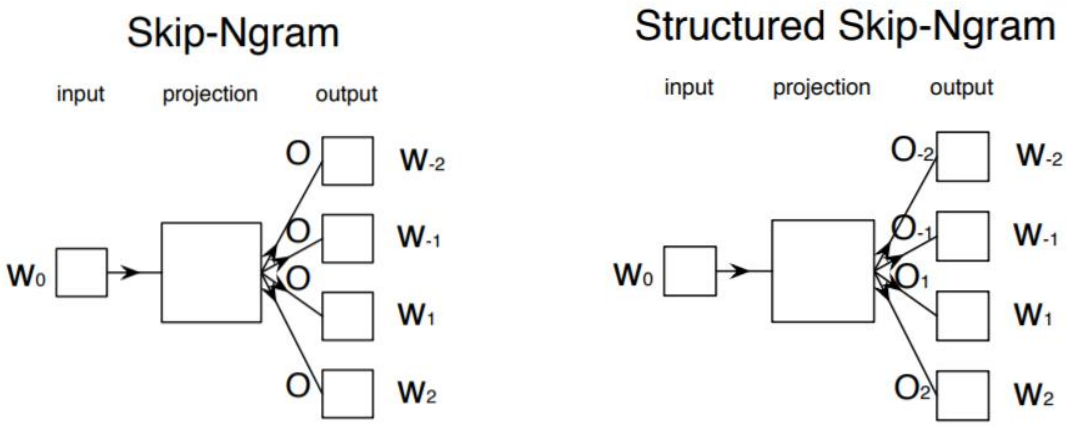


[7] Topical word embeddings(Liu等, AAAI2015)

4、word2vec 后续 – structured Skip-gram

考虑上下文的相对位置

用在词性标注中。

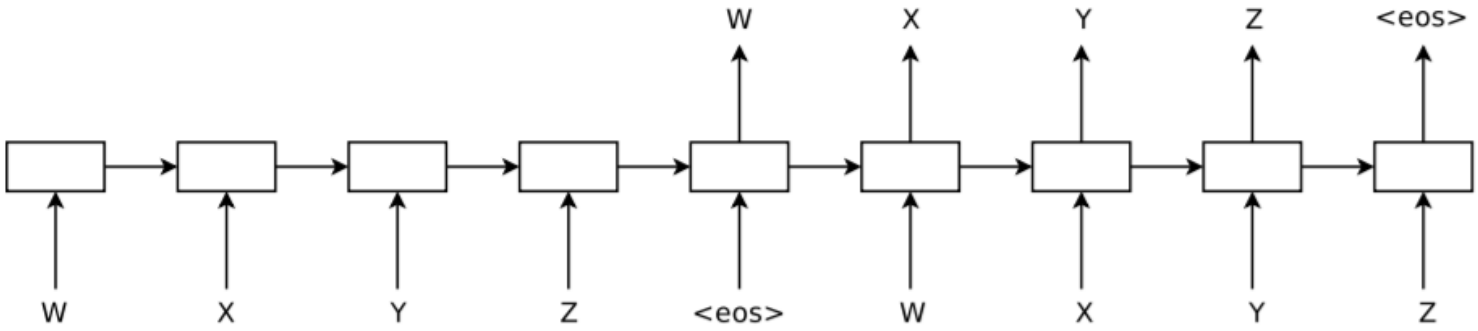


[8] Two/too simple adaptations of word2vec for syntax problems(Ling等, NAACL2015).

5、预训练语言模型

5、预训练语言模型

Dai等[9]提出使用预训练的语言模型/自编码器初始化分类器

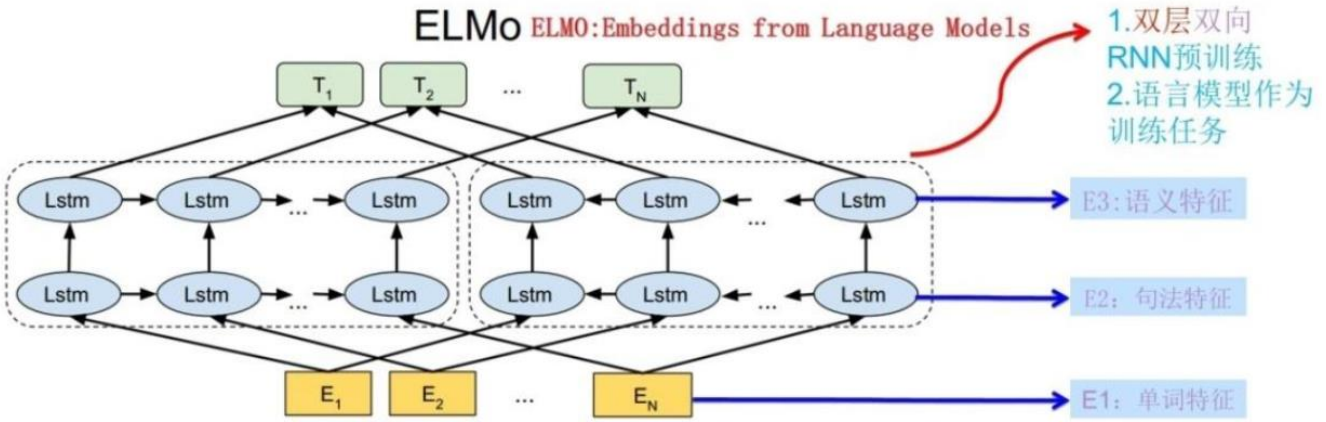


[9] Semi-supervised sequence learning(Dai等, NIPS2015).

5、预训练语言模型

ELMo: Embeddings from Language Models

双层、双向



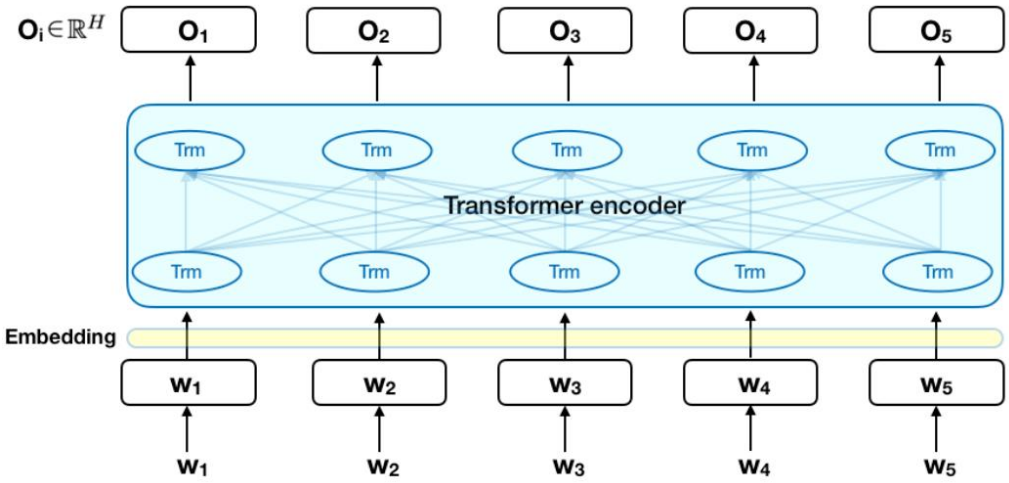
[10] Deep contextualized word representations(Peters等, NAACL2018).

5、预训练语言模型

BERT: Bidirectional Encoder Representations from Transformers

序列一长，LSTM就会很慢。

使用Transformer[12]替代LSTM，层数变成了12/24层。

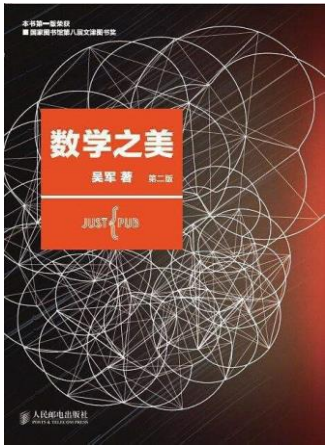


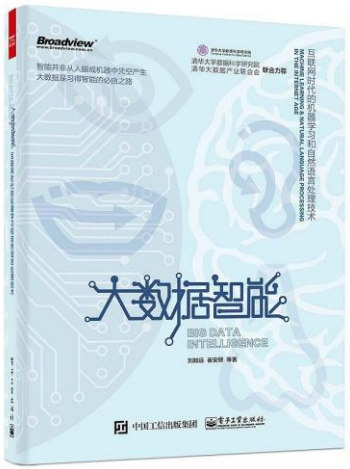
[11] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(Devlin等, NAACL2018).
[12] Attention is all you need(Vaswani等, NIPS2017).

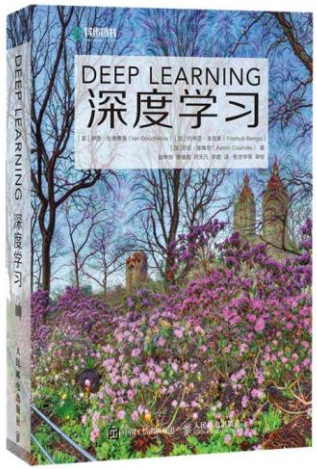
6、一些建议

6、一些建议

- Github
- 多看看文章 ACL/EMNLP/NAACL AAAI/NIPS
- 图书推荐
  - 数学之美(吴军)
  - 大数据智能(刘知远等)
  - 深度学习(Ian Goodfellow等)







6、一些建议

- 推荐的组会主题
  - LSTM/textCNN/HAN/Transformer
  - 门控机制/注意力机制
  - 自编码器/变分自编码器
  - 图神经网络
  - 对抗生成网络
  - 阅读理解/序列标注
  - HMM/CRF/xgboost
  - 优化器/损失函数/网络初始化方法/Dropout/BN
  - BERT/xlnet

谢谢

2019/09/07　张义策

参考文献

1. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
2. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
3. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
4. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.