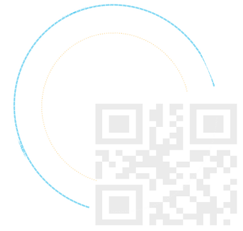
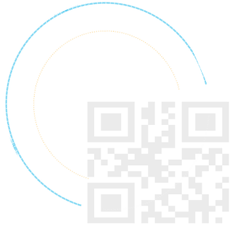
文本检索

玖强



Outline

- Introduction to text retrieval
- Basics of indexing and retrieval
- Deep learning based text retrieval



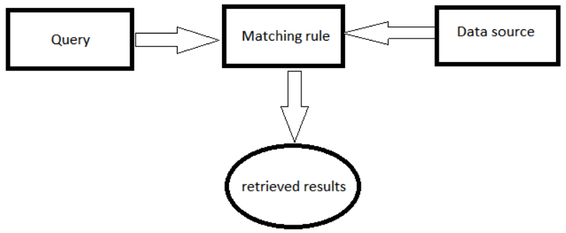
Introduction

Text matching history

- 文本匹配是自然语言理解中的一个核心问题

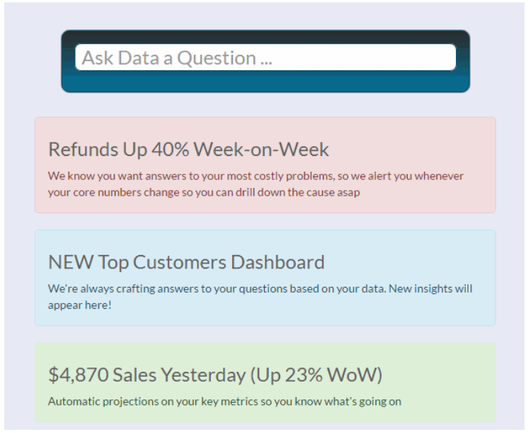
文本匹配用在哪里？

信息检索



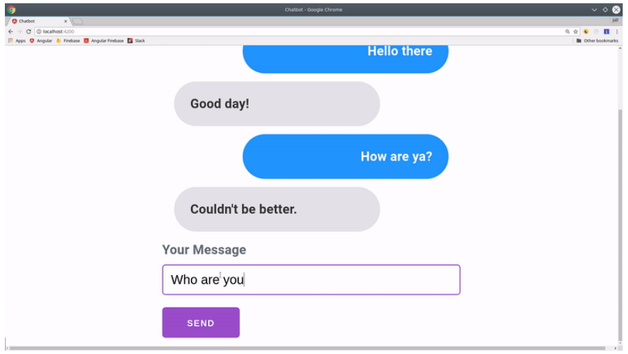
basic model of an information retrieval system

自动问答



机器翻译

机器对话



文本匹配面临的挑战

- 词语匹配的多元性

“荷花”、“莲花”、“水芙蓉”、“芙蕖”



“Apple” or “Apple”


文本匹配面临的挑战

- 短语匹配的结构

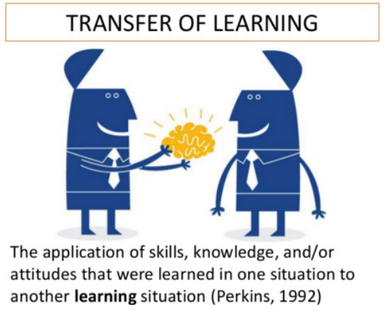
“机器学习” vs. “机器学习”

“机器学习” vs. “学习机器”

成为考试机器，丢人，
成为学习机器，光荣。

“迁移学习” vs. “迁移学习”

“迁移学习” vs. “学习迁移”



文本匹配面临的挑战

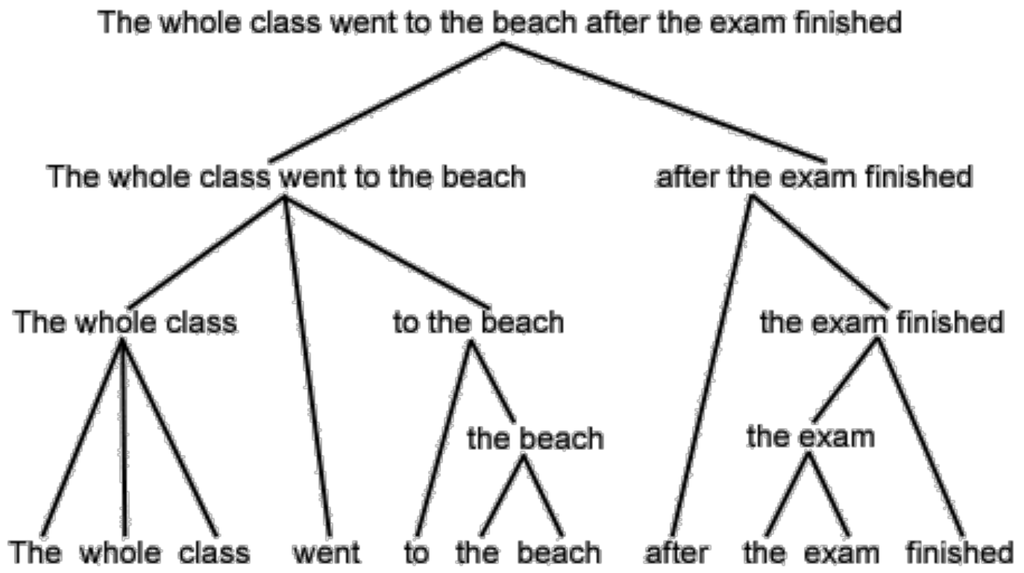
- 文本匹配的层次

  词语组成短语;

  短语组成句子;

  句子组成段落;

  段落组成篇章



需要考虑不同层次的匹配信息

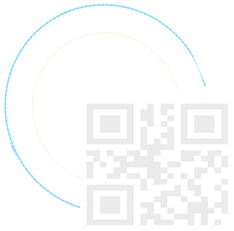
文本匹配的趋势：从传统方法到深度学习

- 传统文本匹配模型 ==> 深度文本匹配模型
- 传统方法的问题

1. 大量的人工定义和抽取的特征：特征总是根据特定的任务（信息检索，或者自动问答）人工设计的，很大程度上限制了模型的泛化能；
2. 迁移成本高：传统模型在一个任务上表现很好的特征很难用到其他文本匹配任务

- **Deep learning方法的优势**

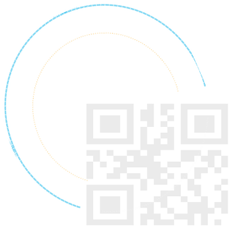
1. 自动从原始数据中抽取特征，免去了大量人工设计特征的开销；
2. 迁移成本低：特征的抽取过程是模型的一部分，根据训练数据的不同，可以方便适配到各种文本匹配的任务当中；
3. 深度文本匹配模型结合**Word2Vec**技术，更好的解决了词语匹配的多元性问题（词语匹配的多元性. 不同的词语可能表示的是同一个语义，比如同义词，“荷花”、“莲花”、“水芙蓉”）；
4. 更好的满足短语匹配的结构性和文本匹配的层次性的特征



基于深度学习的文本匹配

- 一般来说，可以将**deep learning**的文本匹配分为3大类：

1. 单语义文档(**Single Semantic Document**)表达的深度学习模型.
   - **Single Doc => Dense Vector (Distributed Representation)**
   - **Calculate the similarity or distance between two vectors**
   - **E.g., *Deep Semantic Similarity Model (DSSM)/ Deep Structured Similarity Model; Convolutional Latent Semantic Model (CLSM);LSTM-RNN Model***
2. 多语义文档(**Multi Semantic Document**)表达的深度学习模型
   - **Single Granularity Dense Vector (单一粒度的稠密向量) is not good enough ;**
   - **Multi-Granularity (多粒度) and Multi-Semantic(多语义) Vectors are better!**
   - ***E.g., MultiGranCNN, uRAE***
3. 直接建模匹配模式的深度学习模型
   - 更**Fine-Grained (精细)**的建模匹配的模式
   - 更早地让两段文本进行交互，然后挖掘文本交互后的模式特征，综合得到文本间的匹配
   - ***E.g., DeepMatch, Match-SRNN***



定义

文本匹配问题的定义

- **Given：Training data**
  - 定标注训练数据集 $S_{\text{train}} = \{(s_1^{(i)}, s_2^{(i)}, r^{(i)})\}_{i=1}^{N}$
  - $s_1^{(i)} \in S_1, s_2^{(i)} \in S_2$是两段文本（e.g., 查询项 vs. 答案/问题 vs. 答案）；
  - $r^{(i)}$表示对象$s_1^{(i)}$和$s_2^{(i)}$的匹配程度(Similarity, e.g., 问题和答案的相关程度)
- **Target：Mapping function**
  - $f = S_1 * S_2 \rightarrow R$
  - 对于测试数据集 $S_{\text{test}} = \{(s_1^{(i)}, s_2^{(i)})\}_{i=1}^{M}$上任意输入$(s_1^{(i)} \in S_1, s_2^{(j)} \in S_2)$。能够预测出$s_1^{(i)}$和$s_2^{(j)}$的匹配程度(Similarity)$r^{(i,j)}$；
  - 然后通过匹配度排序(Ranking)得到最后结果
- **Example**
  - $s_1^{(i)}$：从古至今，面条和饺子是中国人喜欢的食物；
  - $s_2^{(i)}$：从古至今，饺子和面条在中国都是人见人爱

排序问题

文本匹配问题的评价

- **Precision at k (P@k) and Recall at k (R@k)**
  - 定义真实排序前k个文本中，匹配文本的数量为G_k，而在预测排序中前k个文本中，匹配文本的数量为Y_k.
  - 评价指标P@k和R@k定义为：
    - P@k=$\frac{Y_k}{k}$: 所有"正确被检索的$item$"占所有"应该检索到的"的比例
    - R@k=$\frac{Y_k}{G_k}$: 所有"正确被检索的item"占所有"实际被检索到的"的比例.
  - 形象直观的理解就是Recall要求的是全，宁可错杀一千，不能放过一人，这样Recall就会很高，但是precision就会最低
- **MAP (Mean Average Precision/平均准确率)**
  - 假设预测排序中的真实匹配的文本的排序位置分别为k_1，k_2，...，k_r，其中r为整个列表中所有匹配文本的数量.
  - MAP的定义为：MAP=$\frac{\sum_{i=1}^{r} P@k_i}{r}$
  - MAP是为解决P，R，F-measure的单点值局限性的，同时考虑了检索效果的排名情况。
- **Mean Reciprocal Rank (MRR). Multiple levels of relevance. Normalized Discounted Cumulative Gain (NDCG)**

传统文本匹配学习模型

传统文本匹配

- 焦点：如何设置合适的文本匹配学习算法来学习到最优的匹配模型？
  - 例子： 互联网中搜索时，**query和web page时两个异质空间中的对象，多种匹配学习模型被提出来去计算query与web page的相关度**
  - ***Berger, Adam, and John Lafferty. "Information retrieval as statistical translation." ACM SIGIR Forum. Vol. 51. No. 2. ACM, 2017. 他们提出用统计机器翻译模型计算网页词和查询词间的"翻译"概率，从而实现了同义或者近义词之间的匹配映射***

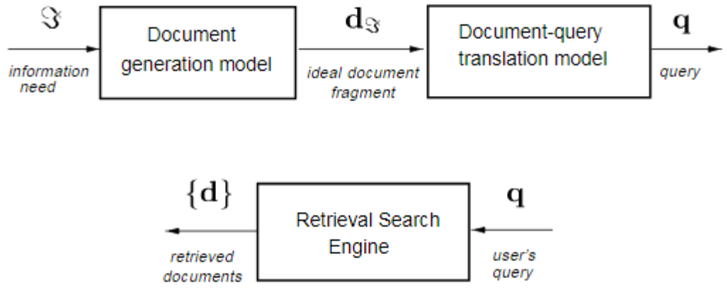


Figure 1. *Model of query generation and retrieval*

  - ***Gao, Jianfeng, Xiaodong He, and Jian-Yun Nie. "Clickthrough-based translation models for web search: from word models to phrase models." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010. 他们在词组一级训练统计机器翻译模型并利用用户点击数据进行模型训练。***
  - ***Wu, Wei, Zhengdong Lu, and Hang Li. "Learning bilinear model for matching queries and documents." The Journal of Machine Learning Research 14.1 (2013): 2519-2548. 他们提出正则化隐空间映射（Regularized Mapping to Latent Space, RMLS)把查询项和网页映射到同一隐空间中，并在模型训练中引入了正则化因子以避免奇异解。***

传统文本匹配

- 缺点
  - 人工代价大： 大量的人力物力才能提取到较少的比较有效的特征；有经验的工程师设计特征并选择。
  - 不精确：**Topic model (**主题模型)的隐空间模型还比较粗糙，不能精确建模文本匹配中的语义相近程度；
  - 性能低：传统模型很难发掘一些隐含在大量数据中，含义不明显的特征，而往往有些特殊情况需要这样的特征才能提高性能.

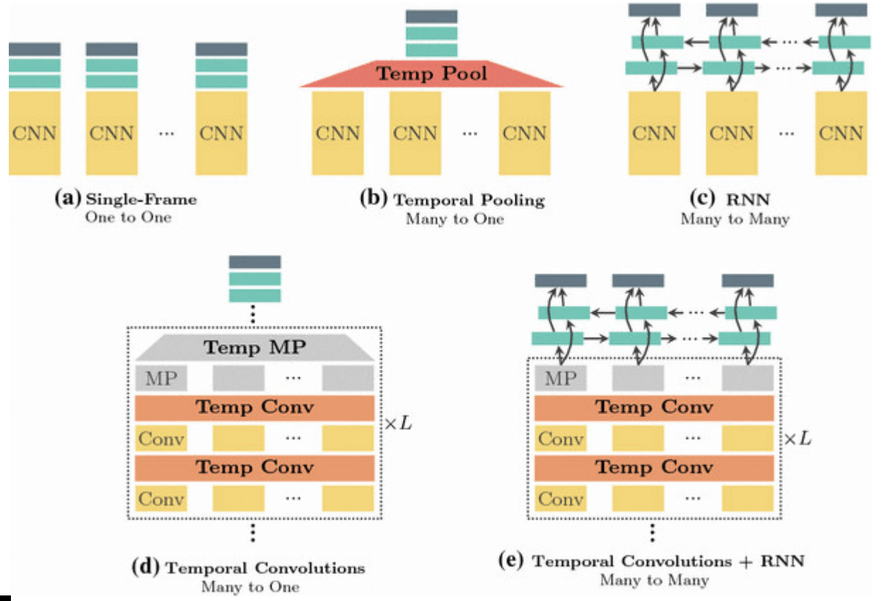
深度学习文本匹配

❑ 优势

1. **Deep learning model**可以将单词表示为语义空间中的向量，利用向量之间的距离运算更准确地描述两个单词之间的语义关系；

2. **Deep learning model**自身的结构是层次化和序列化的，能够比较自然地描述自然语言中的层次结构、序列结构和组合操作；

3. 深度学习模型很好地利用大规模数据的优势和日益发展的高性能计算的能力



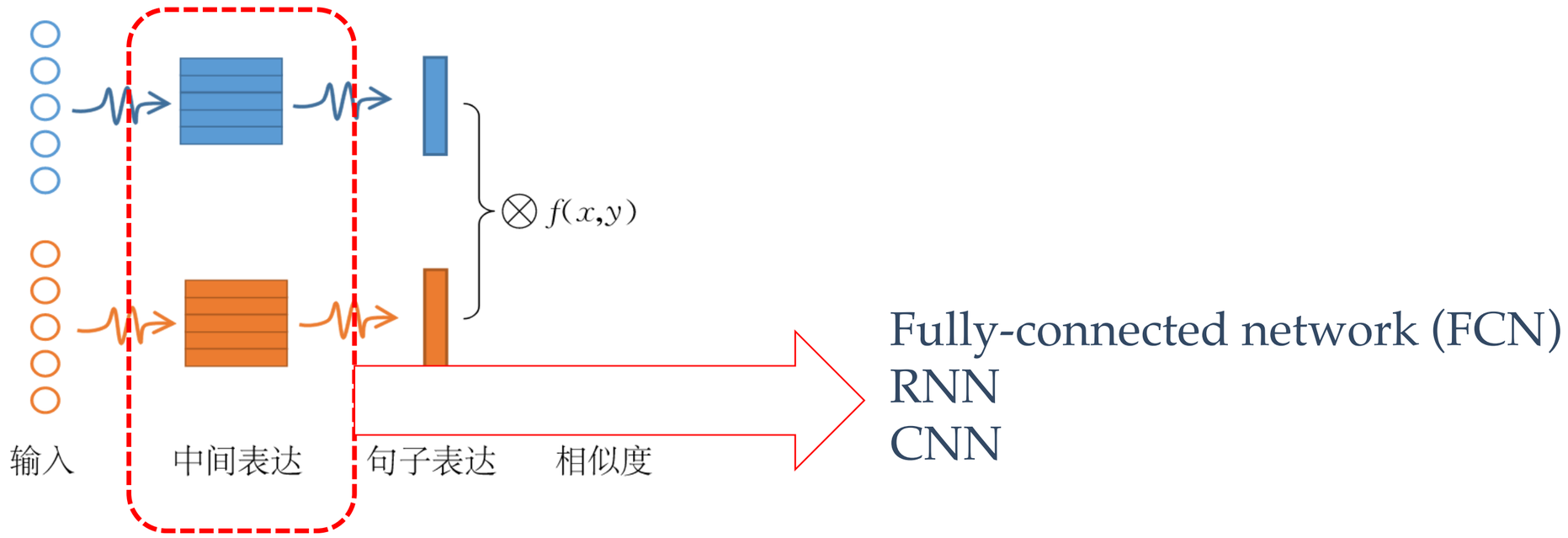
❑ 进展

- ❑ 词性标注、语法分析、情感分析、关系分类 等
- ❑ **CNN, RNN**

基于单语义文档表达的深度学习模型

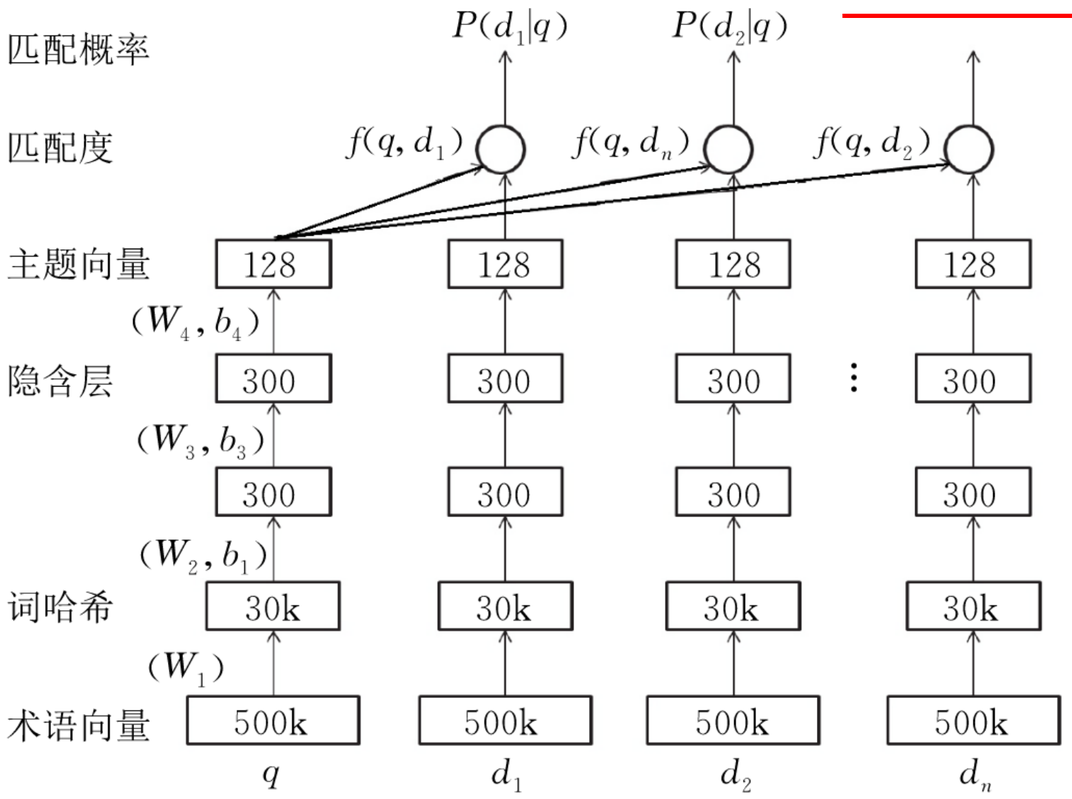
- 文档的表达：将文档表达成一个向量。

1. 利用深度学习的方法生成一个文档的高维度稠密向量
2. 得到两个文档的表达之后，通过计算这两个向量之间的相似度便可输出两者的匹配



基于单语义文档表达的深度学习模型

- **基于全连接神经网络(FCN)**
  - **Deep Semantic Structured Model, DSSM**



最大化所有正例的匹配概率的似然函数：

$$P(d \mid q) = \frac{\exp(\gamma f(q,d))}{\sum_{d' \in \boldsymbol{D}} \exp(\gamma f(q,d'))},$$

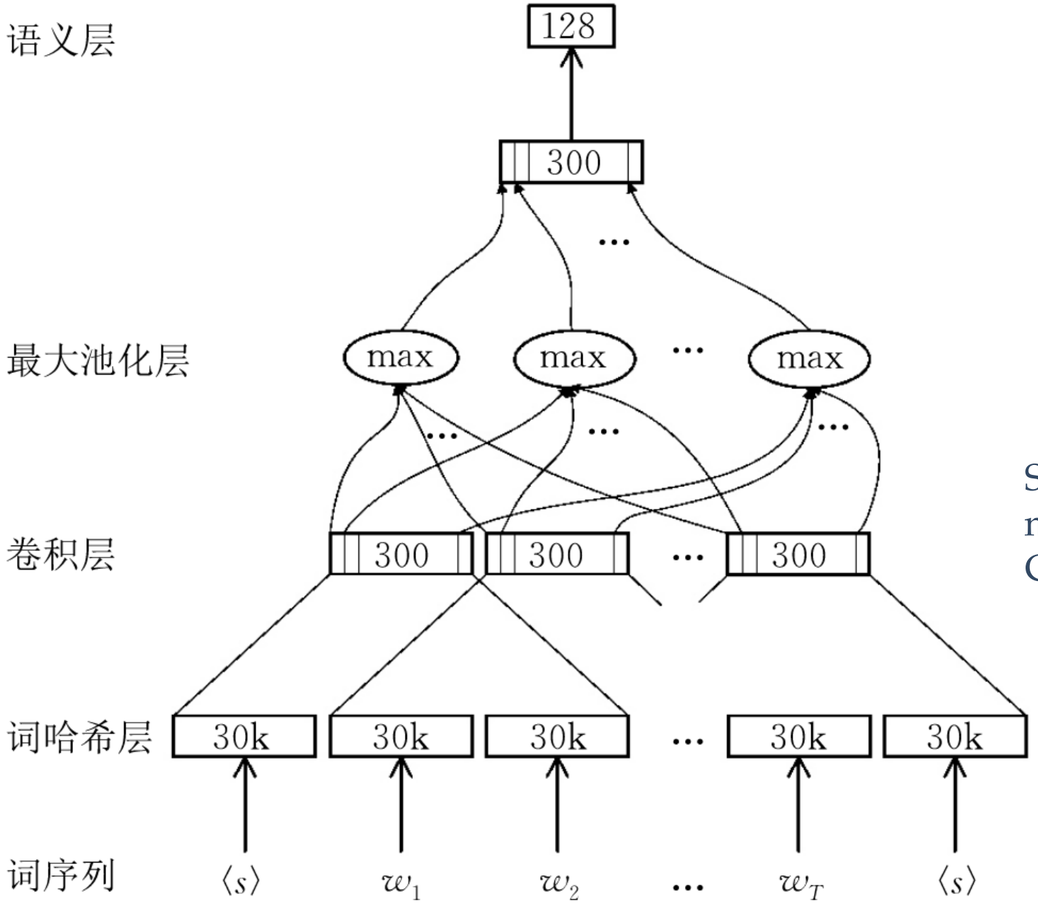
$$L(\Lambda) = -\log \prod_{(q,d^+)} P(d^+ \mid q)$$

其中，γ是Softmax函数的平滑参数，f(q,d)表示一个查询项和文档d之间的匹配度. D表示所有文档的集合。
在实际应用中我们一般采样若干正例d^+以及采样若干负例d^-，来取代整个集合D

Huang, Po-Sen, et al. "Learning deep structured semantic models for web search using clickthrough data." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

基于单语义文档表达的深度学习模型

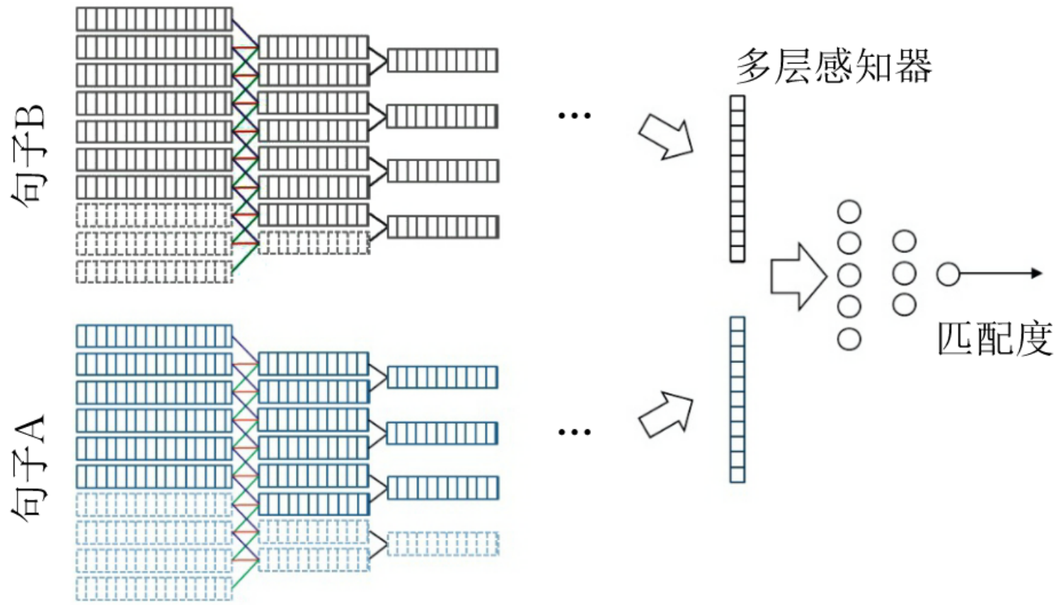
- **基于卷积神经网络(CNN)**
  - **Convolutional Deep Semantic, CSDSSM**



Shen, Yelong, et al. "Learning semantic representations using convolutional neural networks for web search." Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014.

基于单语义文档表达的深度学习模型

- **基于卷积神经网络(CNN)**
  - **Architecture-I (ARC-I) model,**



$$L(\Lambda) = \max(0, 1 + f(q, d^-) - f(q, d^+))$$

基于排序的损失函数，旨在拉大正负样本之间的匹配度数值的差距，而并不在意匹配度的绝对值的大小，这个损失函数更接近排序的应用场景。

Hu, Baotian, et al. "Convolutional neural network architectures for matching natural language sentences." Advances in neural information processing systems. 2014.

基于单语义文档表达的深度学习模型

- 基于循环神经网络(RNN)
  - Palangi, Hamid, et al. "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval."

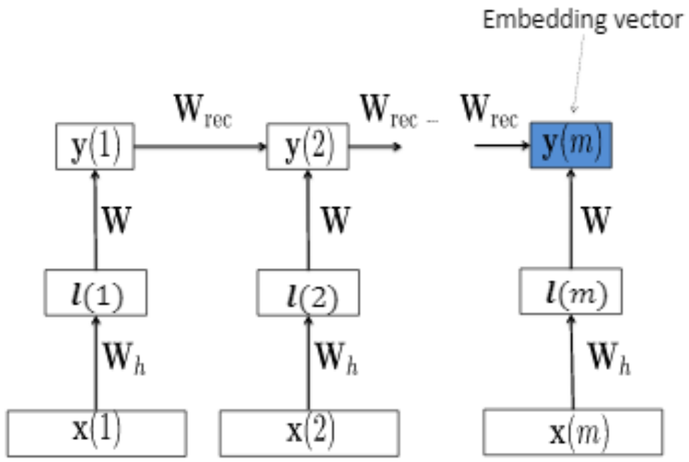


Fig. 1. The basic architecture of the RNN for sentence embedding, where temporal recurrence is used to model the contextual information across words in the text string. The hidden activation vector corresponding to the last word is the sentence embedding vector (blue).

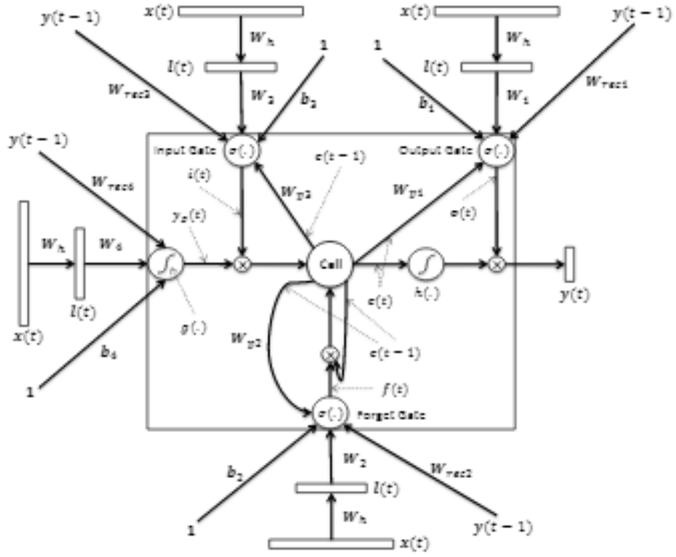


Fig. 2. The basic LSTM architecture used for sentence embedding

$$
\begin{aligned}
\mathbf{y}_g(t) &= g(\mathbf{W}_4\mathbf{l}(t) + \mathbf{W}_{\text{rec4}}\mathbf{y}(t-1) + \mathbf{b}_4) \\
\mathbf{i}(t) &= \sigma(\mathbf{W}_3\mathbf{l}(t) + \mathbf{W}_{\text{rec3}}\mathbf{y}(t-1) + \mathbf{W}_{p3}\mathbf{c}(t-1) + \mathbf{b}_3) \\
\mathbf{f}(t) &= \sigma(\mathbf{W}_2\mathbf{l}(t) + \mathbf{W}_{\text{rec2}}\mathbf{y}(t-1) + \mathbf{W}_{p2}\mathbf{c}(t-1) + \mathbf{b}_2) \\
\mathbf{c}(t) &= \mathbf{f}(t) \circ \mathbf{c}(t-1) + \mathbf{i}(t) \circ \mathbf{y}_g(t) \\
\mathbf{o}(t) &= \sigma(\mathbf{W}_1\mathbf{l}(t) + \mathbf{W}_{\text{rec1}}\mathbf{y}(t-1) + \mathbf{W}_{p1}\mathbf{c}(t) + \mathbf{b}_1) \\
\mathbf{y}(t) &= \mathbf{o}(t) \circ h(\mathbf{c}(t)) \qquad (2)
\end{aligned}
$$

基于单语义文档表达的深度学习模型—总结

- 基于单语义文档表达的深度学习算法的重心在于得到一个适合的文档表

- 优点：

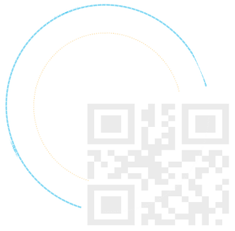
1. 文本映射为一个简洁的representation;
2. 匹配的计算速度
3. 大量无监督的数据进行预训练

- 缺点：

1. 很多匹配问题不具备传递性，因此不适合用一个参数相同的神经网络来描述
2. 文本的表示学习本身是非常困难的问题，只有效捕捉与描述对匹配有用的局部化（细节）信息

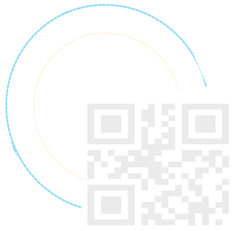
基于多语义文档表达的深度学习模型

- 一些新的深度匹配模型被提出来去综合考虑文本的局部性表达（词，短语等）和全局性表达（句子）
  - 例如：可伸展递归自动编编码器
- 多粒度的匹配可以很好地补充基于单语义文档表达的**Deep model**在压缩整个句子过程中的信息损失，而达到更好的效果
  - 例如：**Multi-grain CNN**，**Multi-view RNN**



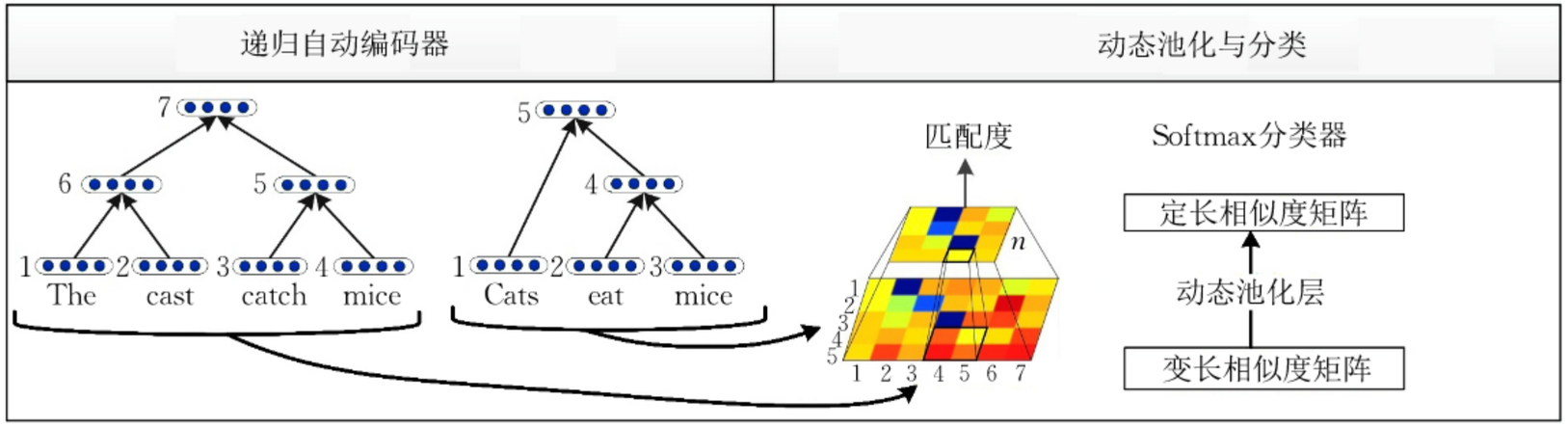
基于多语义文档表达的深度学习模型

- 一些新的深度匹配模型被提出来去综合考虑文本的局部性表达（词，短语等）和全局性表达（句子）
  - 例如：可伸展递归自动编编码器
- 多粒度的匹配可以很好地补充基于单语义文档表达的**Deep model**在压缩整个句子过程中的信息损失，而达到更好的效果
  - 例如：**Multi-grain CNN**，**Multi-view RNN**



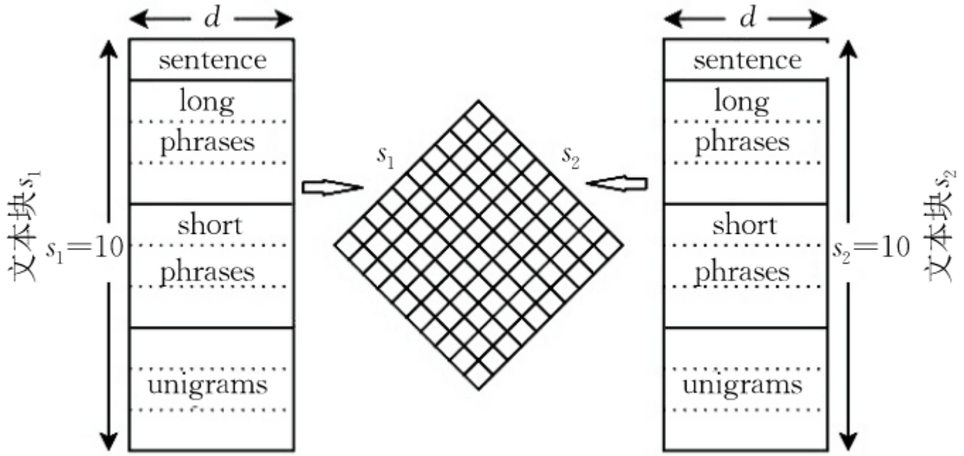
基于多语义文档表达的深度学习模型

- 可伸展递归自动编码（**unfolding Recursive Auto-Encoder, uRAE**）
  - **Wan, Shengxian, et al. "A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations." AAAI. Vol. 16. 2016.**



基于多语义文档表达的深度学习模型

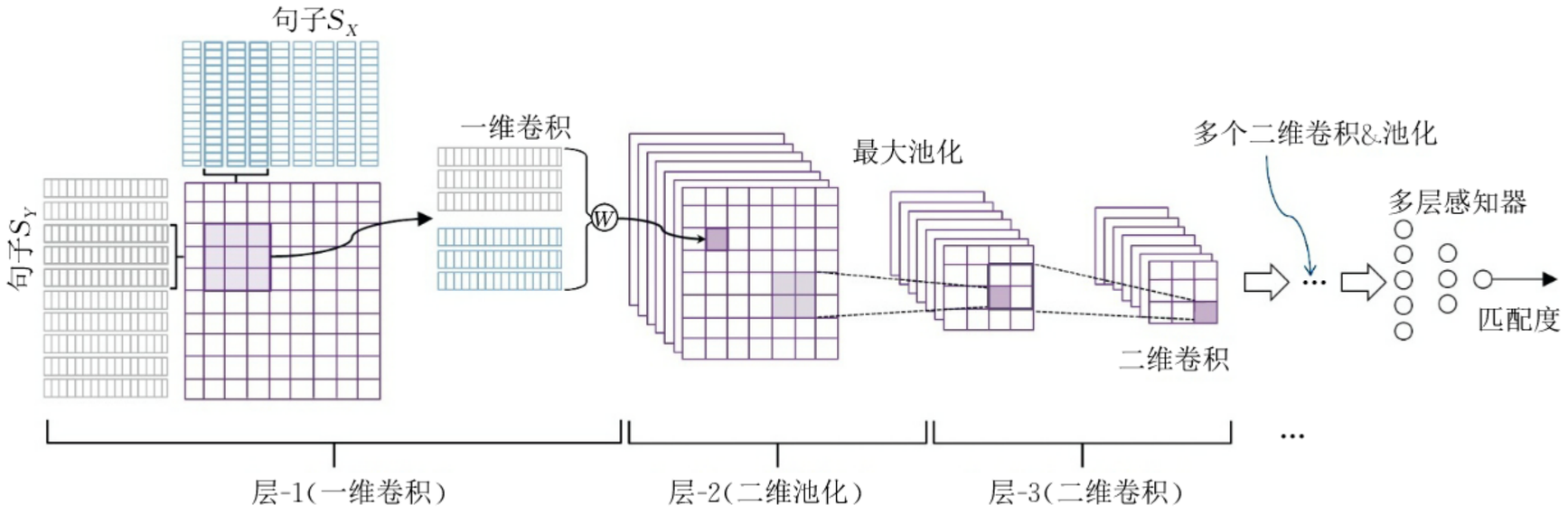
- 多粒度卷积神经网络
  - **Yin, Wenpeng, and Hinrich Schütze. "Convolutional neural network for paraphrase identification." Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015.**



直接建模匹配模式的深度学习模型

❑ CNN深度匹配模型

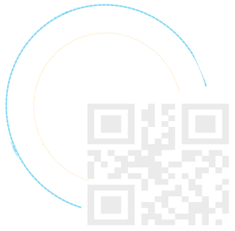
- ❑ **CNN-based model (例如，ARC-II)首先将橘子表达成句中单词的向量序列**
- ❑ **然后用滑动窗来选择词向量作为基本单元进行卷积操作，得到一个3d的张量，作为2个句子互作用的一个初步表示**
- ❑ **随后的卷积以这个三维向量为基础进行conv+pooling若干次，最后得到描述两个句子整体关联的向量**
- ❑ **最终由一个多层神经网络来综合这个向量的每个维度从而得到匹配值**



直接建模匹配模式的深度学习模型

❑ 深度模型的缺点

- ❑ 大量有监督的文本匹配的数据训练，所以无监督学习还是需要探索的；
- ❑ 计算复杂，每一对文档都得完全通过一遍网络。



END