机器阅读理解

讲师：杨博

目录
CONTENTS

01 MRC基础理论

机器阅读理解简介

机器阅读理解：机器阅读理解，又称**阅读理解问答**，要求机器阅读并理解**人类自然语言文本**，在此基础上，解答跟文本信息相关的问题。该任务通常被用来衡量机器自然语言理解能力，可以帮助人类从大量文本中快速聚焦相关信息，**降低人工信息获取成本**，在文本问答、信息抽取、对话系统等领域具有极强的应用价值。近年来，机器阅读理解受到工业界和学术界越来越广泛的关注，是自然语言处理领域的研究热点之一。

基础任务：机器阅读理解**基础任务**是根据问题，从**非结构化文档中寻找合适的答案**，因此，研究人员通常将机器阅读理解形式化为一个关于**（文档，问题，答案）**三元组的监督学习问题。

任务类型：根据答案的形式，机器阅读理解任务被细分为**完形填空式**、**多项选择式**、**片段抽取式**和**自由作答式**四类，这四类任务从易到难，见证了机器阅读理解技术的发展。

机器阅读理解任务详解

MRC四个任务	对应的数据集
完形填空（Ioze Test）	CNN & Daily Mail，CBT (The Children's Book Test)，LAM-BADA dataset (LAnguage Modeling Boardened to Account for Discourse Aspects)，Who-did-What，CLOTH，CliCR
多项选择（Multiple Choice）	MCTest，RACE
片段抽取（Span Extraction）	SQuAD，NewsQA，TriviaQA，DuoRC
自由作答（Free Answering）	bAbI，MS MARCO，SearchQA，NarrativeQA，DuReader

机器阅读理解任务详解

完形填空：完形填空式任务通常将文档中的某个实体用占位符替换，机器阅读残缺段落后寻找正确的词进行补充，使原文完整。这类任务的代表数据集有 CNN & Daily Mail，The Children’s Book Test，CMRC2017等，其中，CNN & Daily Mail 是2015年谷歌发布的首个大型阅读理解数据集，收集了9.3万篇 CNN 和22万篇 Daily Mail 新闻稿，通过实体替换等技术构造百万级别的三元组语料库。

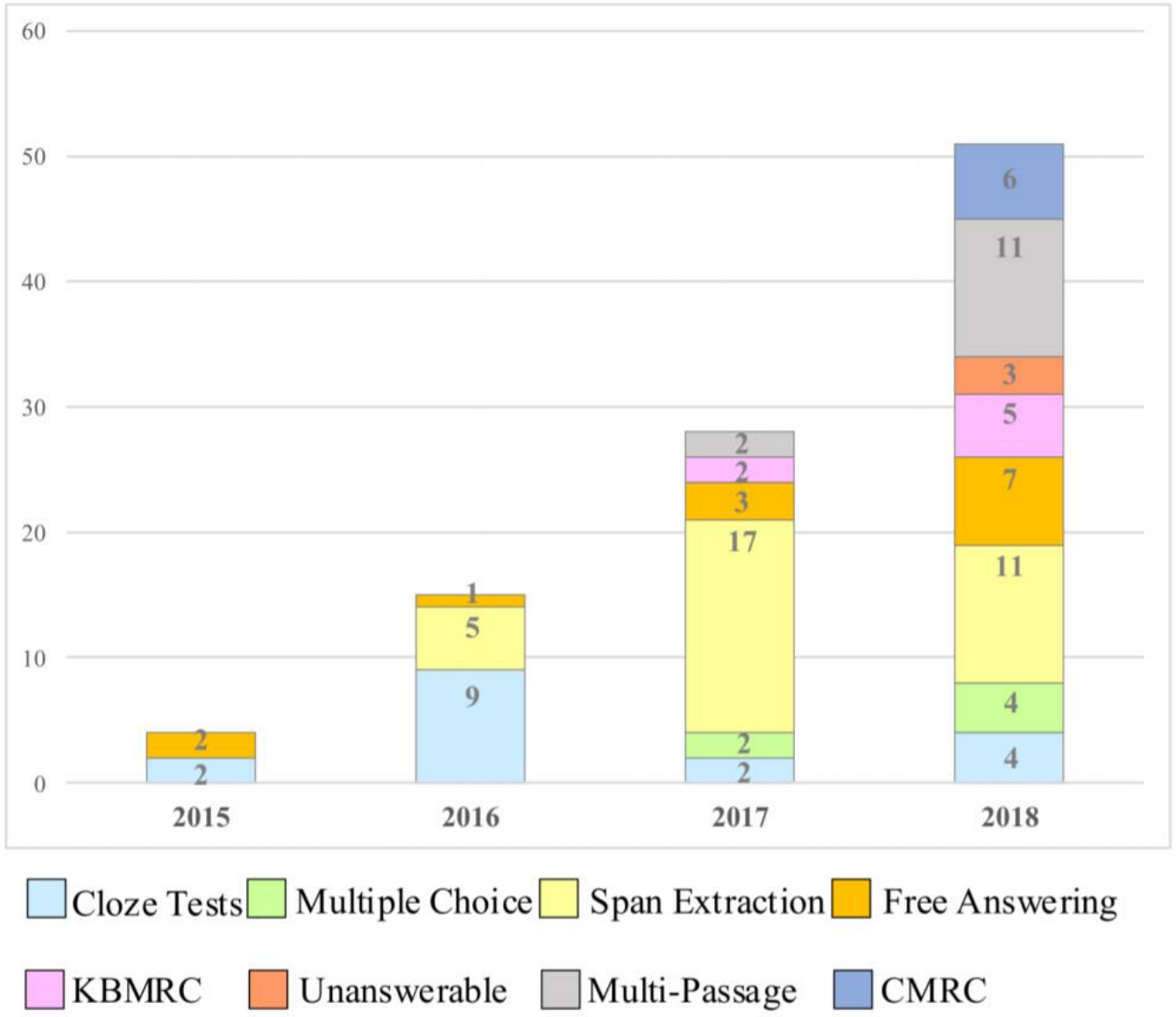
多项选择：多项选择式数据集为（文档，问题，候选答案集，答案）四元组形式，机器阅读文档和问题后，从候选答案集合中挑选正确的答案，如 MCTest 和 RACE。RACE 数据集源自初高中英语考试试题，包含约2.8万篇文章和10万个专家问题，用于测试机器的理解和推理能力。

机器阅读理解任务详解

片段抽取：片段抽取式任务要求从原文中抽取一段连续的句子或短语作为问题的答案，相比于完形填空任务填充单一实体，该任务面临更大的搜索空间，因此更具挑战性。2016年，斯坦福大学发布了SQuAD数据集，该数据集包括500多篇WiKi百科文章以及人工构建的10万多问题。2018年，Rajpurkar等人对SQuAD数据集进一步扩增，在原有基础上新增5万多无法回答的对抗性问题，这些问题在原文中都存在似是而非的迷惑性答案，模型不仅仅需要准确应答受原文支持的可回答问题，还需要避免对不可回答问题作出回应。因此，该任务难度更高，更能检验机器的阅读理解能力。

自由作答：与片段抽取式任务不同，自由作答式阅读理解的答案形式更加灵活，正确答案可能需要从原文进行推理或归纳总结，不限制于是否来自原文句子片段，与现实人类作答习惯最为贴近。代表数据集有 CoQA、MS-MARCO、DuReader 等，通常涉及到多轮问答、多跳推理等技术。

机器阅读理解任务热度



片段抽取任务详解

片段抽取：学者C. Snow在2002年的一篇论文中定义阅读理解是“**通过交互从书面文字中提取与构造文章语义的过程**”。而机器阅读理解的目标是利用人工智能技术，使计算机具有和人类一样理解文章的能力。下图给出了一个机器阅读理解的样例。其中，模型需要用文章中的一段原文回答问题。

段落

工商协进会报告，12 月消费者信心上升到 78.1，明显高于 11 月的 72。另据《华尔街日报》报道，2013 年是 1995 年以来美国股市表现最好的一年。这一年里，投资美国股市的明智做法是追着“傻钱”跑。所谓的“傻钱”策略，其实就是买入并持有美国股票这样的普通组合。这个策略要比对冲基金和其它专业投资者使用的更为复杂的投资方法效果好得多。

问题 1：什么是傻钱策略？

答案：买入并持有美国股票这样的普通组合

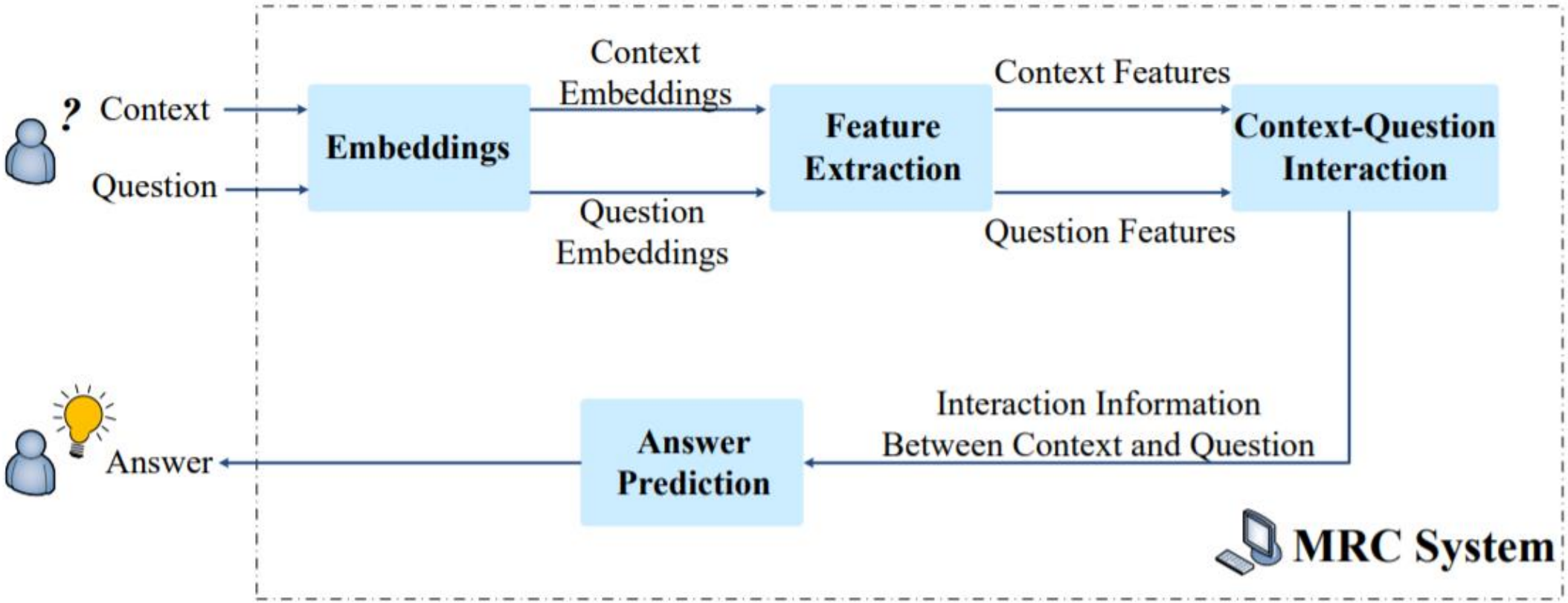
问题 2：12 月的消费者信心指数是多少？

答案：78.1

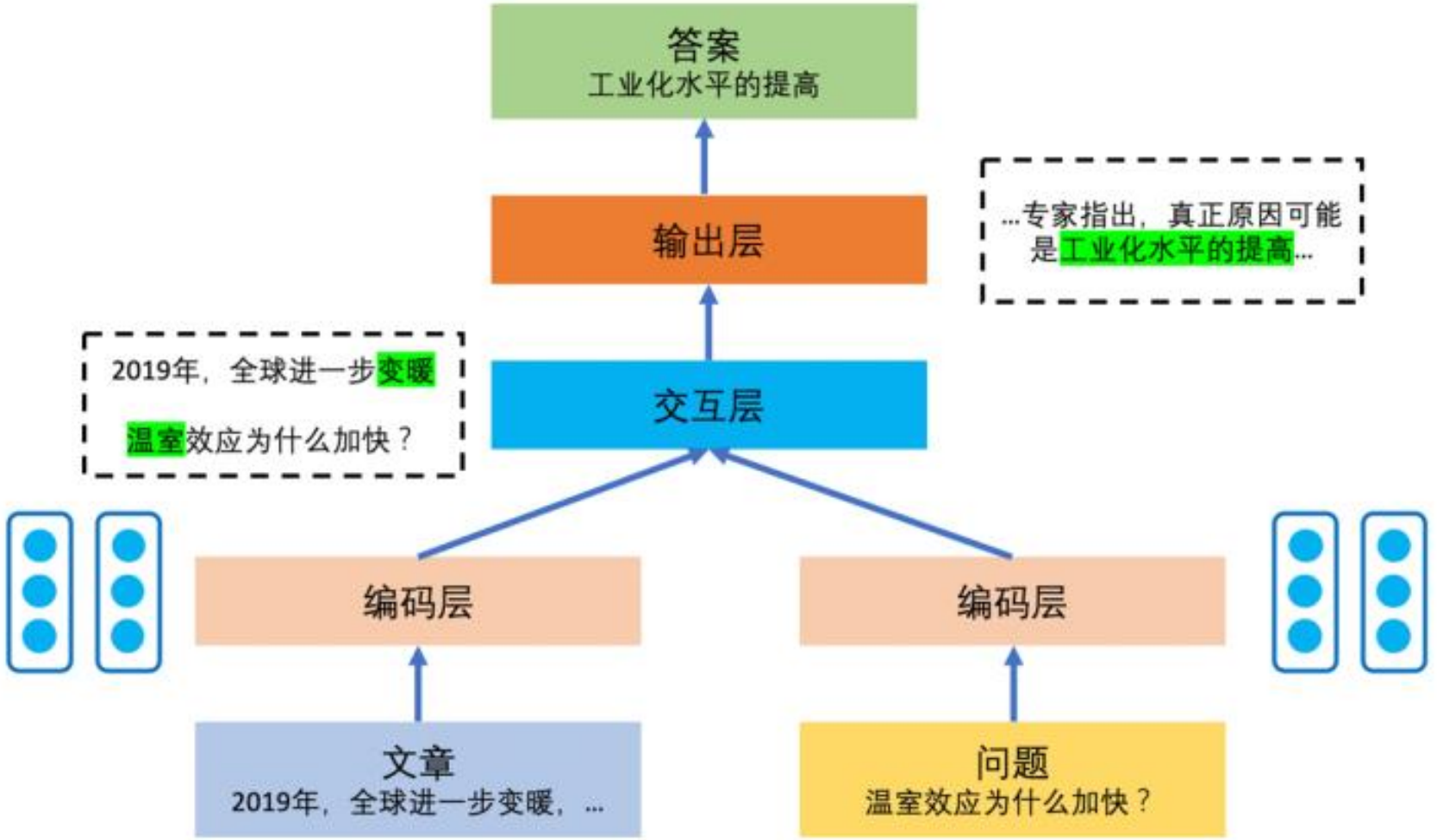
问题 3：消费者信心指数由什么机构发布？

答案：工商协进会

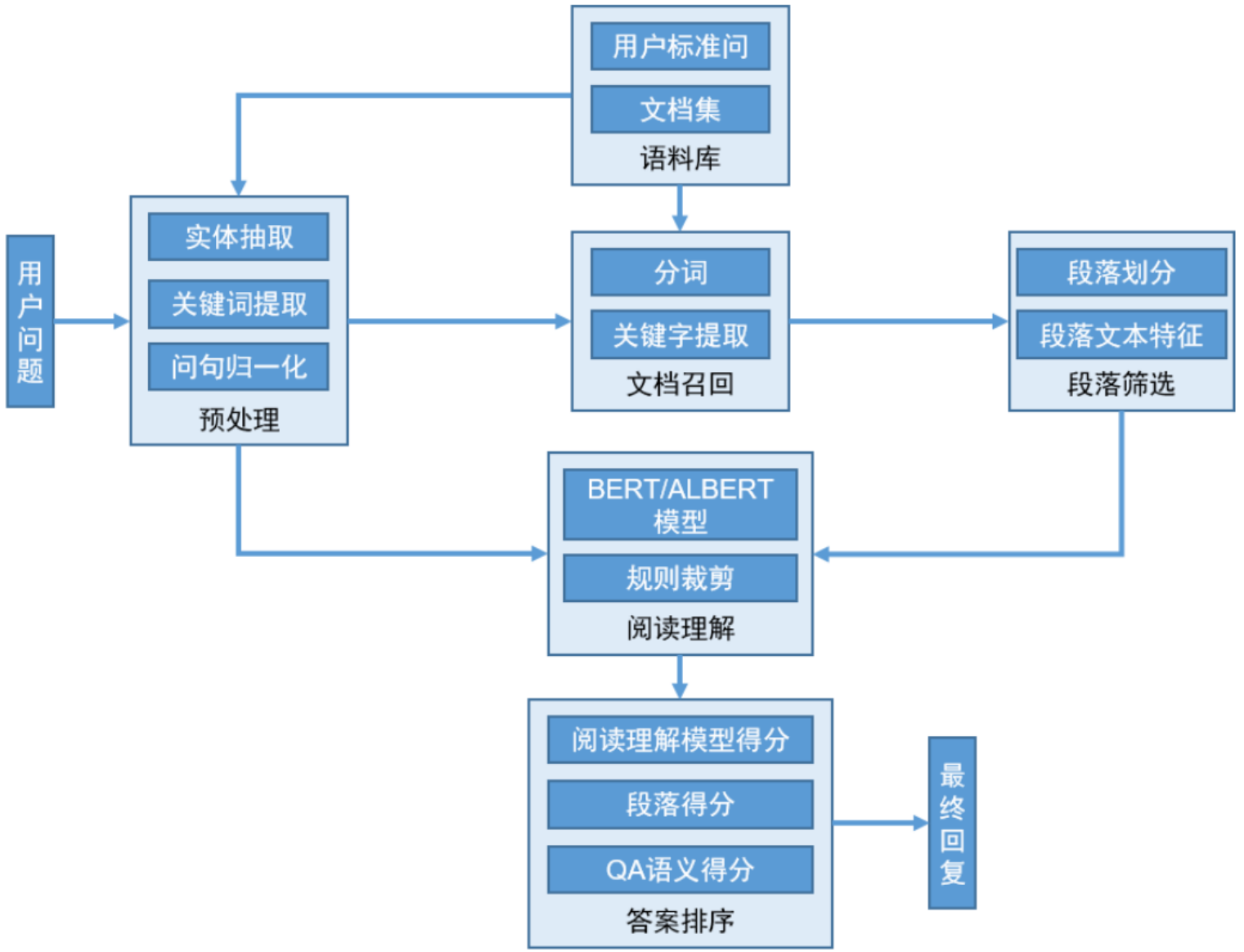
机器阅读理解通用架构



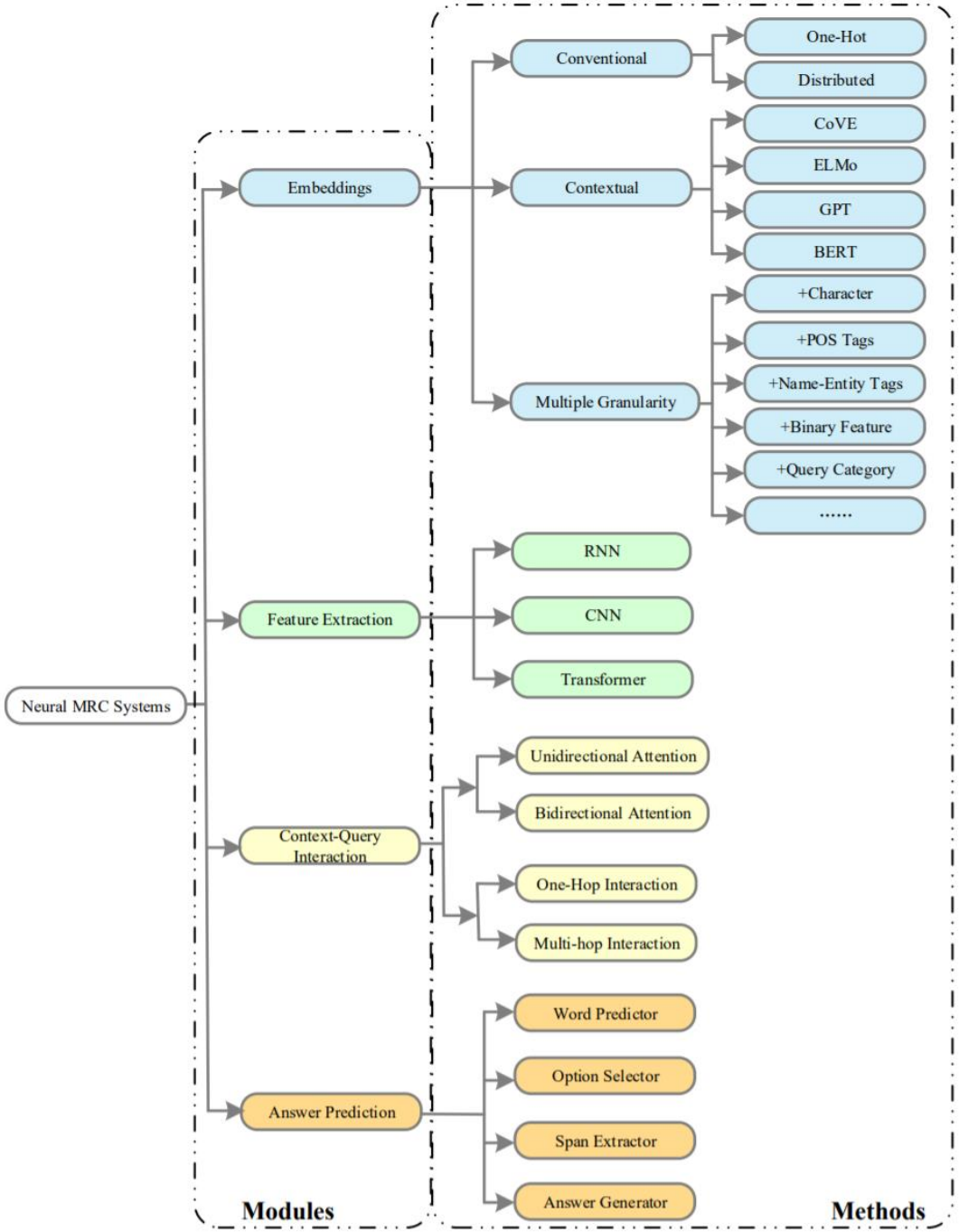
机器阅读理解通用架构



机器阅读理解工业架构



机器阅读理解典型技术



02 深入理解BiDAF

深入理解BiDAF

BiDAF：Bidirectional Attention Flow for Machine Comprehension，出自2017年ICLR的一篇论文。

核心内容：这篇论文主要对 attention 机制做了改进，为此作者总结了 MC 任务上过去常用的三类 attention

- Attention Reader：通过**动态** attention 机制从文本中提取相关信息（context vector），再依据该信息给出预测结果；(代表论文：Bahdanau et al. 2015, Hermann et al. 2015, Chen et al. 2016, Wang & Jiang 2016)
- Attention-Sum Reader：只**计算一次** attention weights，然后直接喂给输出层做最后的预测，也就是利用 attention 机制直接获取文本中各位置作为答案的概率，和 pointer network 类似的思想，效果很**依赖**对 query 的表示； (代表论文：Kadlec et al. 2016, Cui et al. 2016)
- Multi-hop Attention。计算**多次** attention； (代表论文：Memory Network(Weston et al., 2015)，Sordoni et al., 2016; Dhingra et al., 2016., Shen et al. 2016.)

深入理解BiDAF

主要贡献：在此上面基础上，作者对注意力机制做出了改进，具体 BiDAF attention 的特点如下

- 并没有把 context 编码进**固定大小**的 vector，而是让 vector **可以流动**，减少早期加权和的信息损失；
- Memory-less，在每一个时刻，**仅仅对 query 和当前时刻的 context paragraph 进行计算**，并不直接依赖上一时刻的 attention，这使得后面的 attention 计算不会受到之前错误的 attention 信息的影响；
- 计算了 query-to-context（**Q2C**） 和 context-to-query（**C2Q**）两个方向的 attention 信息，认为 C2Q 和 Q2C 实际上能够相互补充。实验发现模型在开发集上去掉 C2Q 与 去掉 Q2C 相比，分别下降了 12 和 10 个百分点，显然 C2Q 这个方向上的 attention 更为重要；

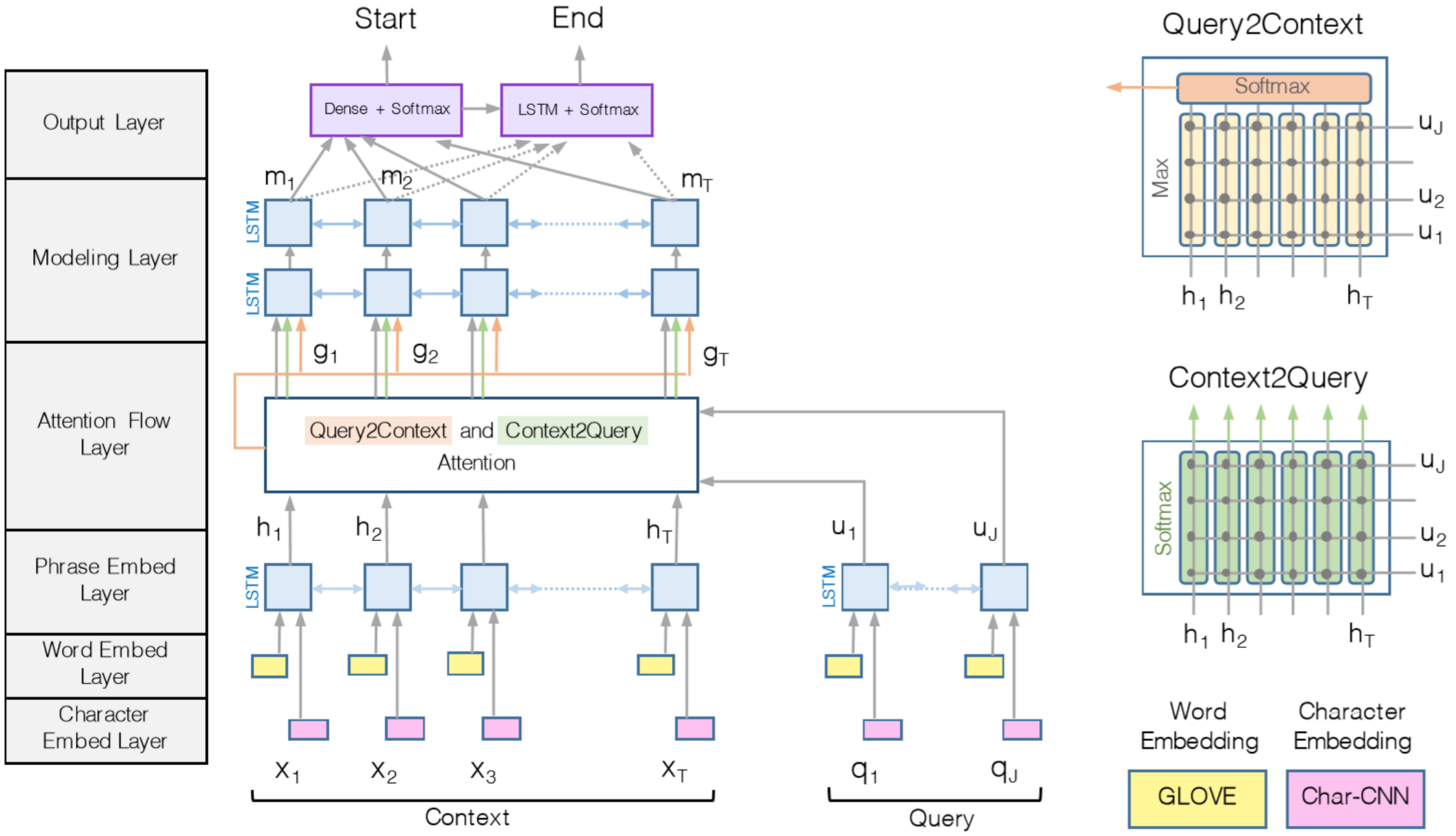


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

深入理解BiDAF

具体任务

Context:

Singapore is a small country located in Southeast Asia.

Query:

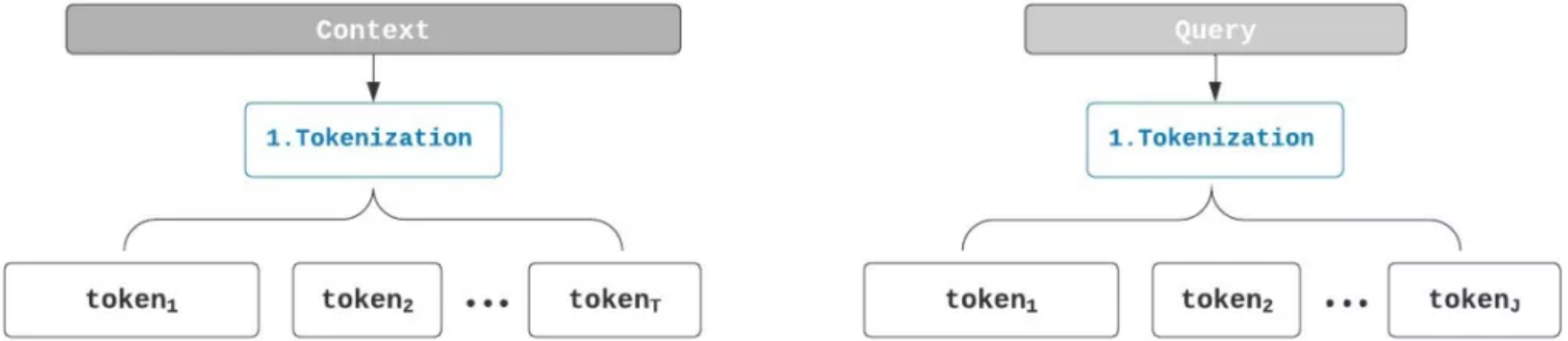
Where is Singapore situated?

BiDAF's Answer:

Southeast Asia.

深入理解BiDAF

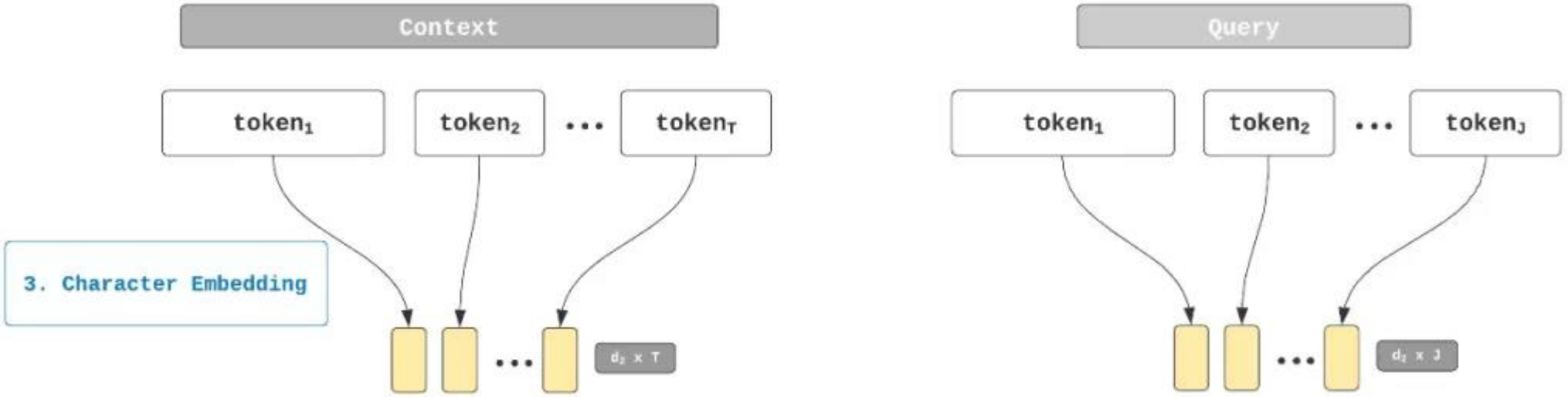
Word embedding layer（词嵌入层）：



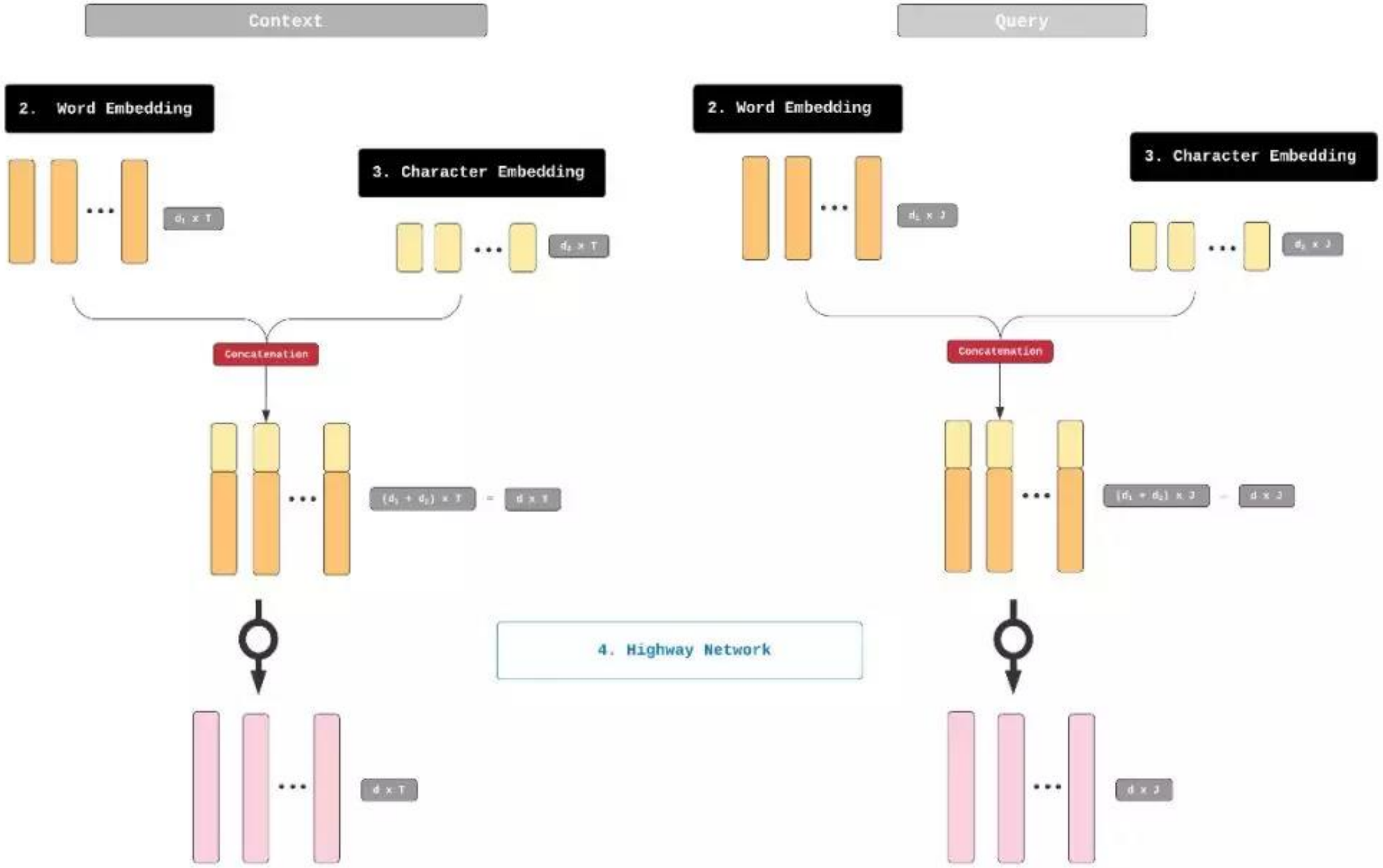
论文中使用的是Glove

深入理解BiDAF

Character embedding layer（字符嵌入层）：



论文中使用的是1D-CNN，主要是解决OOV问题


Context
Query
2. Word Embedding
3. Character Embedding
2. Word Embedding
3. Character Embedding
Concatenation
Concatenation
4. Highway Network

深入理解BiDAF

Highway Network:

$$\mathbf{z} = g(\mathbf{W}\mathbf{y} + \mathbf{b})$$

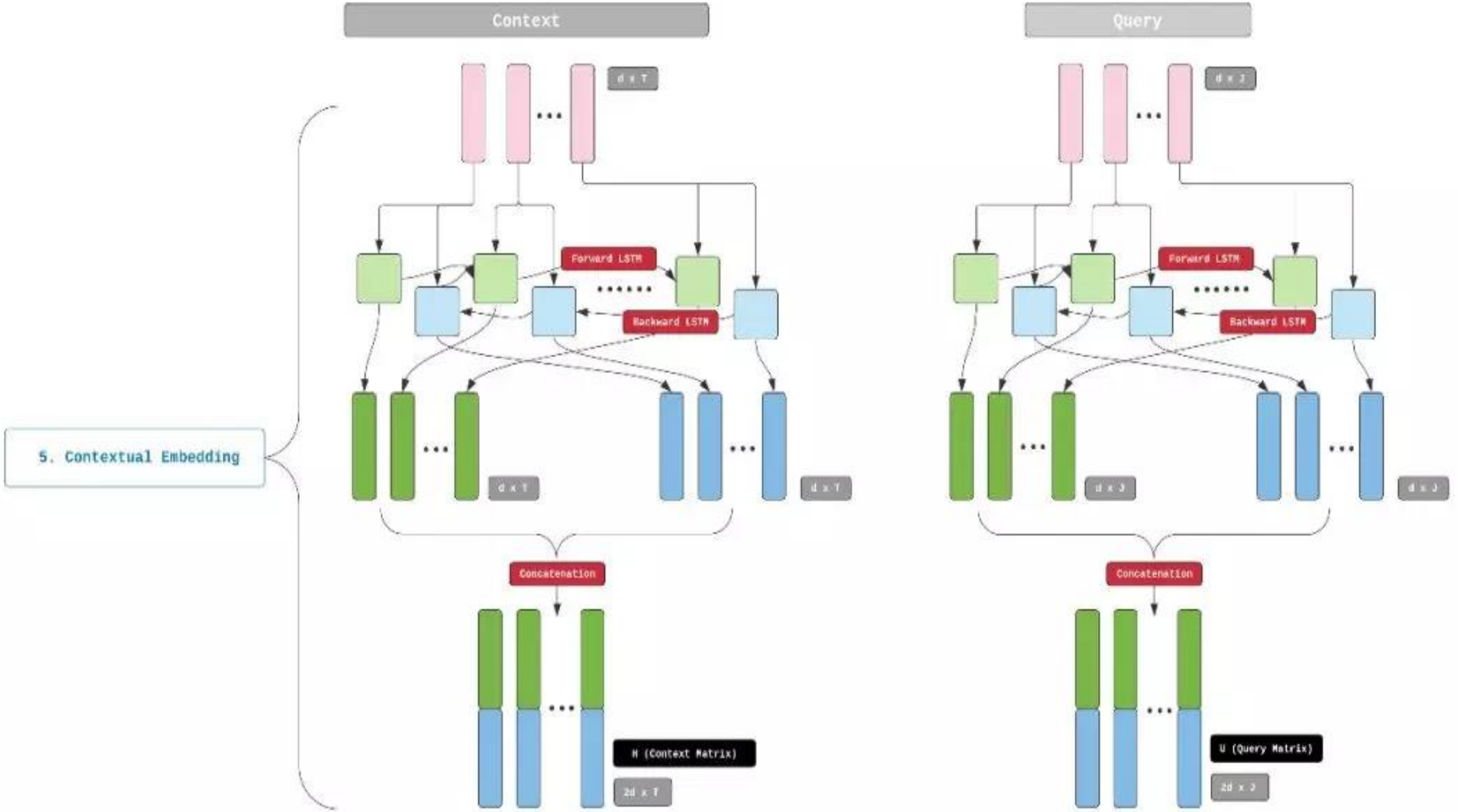
$$\mathbf{z} = \mathbf{t} \odot g(\mathbf{W}_H\mathbf{y} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{y}$$

$$\mathbf{W}_H, \mathbf{b}_H : \text{Affine transformation}$$

$$\mathbf{t} = \sigma(\mathbf{W}_T\mathbf{y} + \mathbf{b}_T) : \textit{transform}\ \text{gate}$$

$$\mathbf{1} - \mathbf{t} : \textit{carry}\ \text{gate}$$

高速神经网络的作用是**调整单词嵌入和字符嵌入步骤的相对贡献配比**，逻辑是，如果我们处理的是一个像“misunderestimate”这样的OOV词，会希望增加该词1D-CNN表示的相对重要性，因为我们知道它的GloVe表示可能是一些随机的胡言乱语。另一方面，当我们处理一个常见而且含义明确的单词时，如“table”时，我们可能希望GloVe和1D-CNN之间的贡献配比更为平等。


5. Contextual Embedding
Context
d x T
Forward LSTM
Backward LSTM
d x T
d x T
Concatenation
H (Context Matrix)
2d x T
Query
d x J
Forward LSTM
Backward LSTM
d x J
d x J
Concatenation
U (Query Matrix)
2d x J

深入理解BiDAF

模型架构：分为六层，以此是

（1）Word embedding layer（词嵌入层）

（2）Character embedding layer（字符嵌入层）

（3）Contextual embedding layer（上下文嵌入层）

（4）Attention flow layer（注意流层）

（5）Modeling layer（模型层）

（6）Output layer（输出层）

03 BiDAF论文讲解

联系方式

群名称:Python自然语言处理与知识...
群　号:1053360392

扫一扫上面的二维码图案，加我微信