腾讯知文团队负责人钟黎：从 0 到1 打造下一代智能对话引擎 | CCF-GAIR 2018

本文作者：汪思颖 2018-07-12 10:26 专题：2018 CCF-GAIR 全球人工智能与机器人峰会

导语：知文团队在智能对话系统上的一些探索经验~干货满满，值得关注

雷锋网(公众号：雷锋网)按：2018 全球人工智能与机器人峰会（CCF-GAIR）在深圳召开，峰会由中国计算机学会（CCF）主办，由雷锋网、香港中文大学（深圳）承办，得到了深圳市宝安区政府的大力指导，是国内人工智能和机器人学术界、工业界及投资界三大领域的顶级交流盛会，旨在打造国内人工智能领域最具实力的跨界交流合作平台。

在 7 月 1 日上午的 NLP 专场上，腾讯知文算法负责人钟黎为大家带来了题为《从 0 到 1 打造下一代智能对话引擎》的主题演讲。

作为腾讯知文算法负责人，钟黎与大家分享了他们在智能对话系统上的研究经验。

他表示，在业界打造通用智能问答平台通常需要解决如下三种问答类型：一是任务驱动型，二是信息获取型，三是通用闲聊型。

他重点描述了第二种类型，即如何让问答系统解决用户的信息获取类问题。围绕这一类型，他讲解了智能问答系统的基础架构，以及非监督学习和监督学习在这里所起的作用。

此后，他阐述了目前业界较为通用的快速召回方案：第一种，基于词汇计数（Lexical term counting）的方法；第二种，基于语言模型的方法；第三种，基于向量化的方法。

在演讲的最后，他谈到知文团队在建设业界问答系统的一些心得体会。

- 第一，要重视 Baseline。
- 第二，尽早建立起整个流程的 pipeline。
- 第三，没有免费午餐定理，不存在万能算法。
- 第四，领域相关的数据准备、数据清洗非常重要。

以下为钟黎的演讲内容，雷锋网做了不改变原意的编辑整理。

大家好，我是腾讯知文的负责人钟黎，今天很荣幸站在这里跟大家分享我们团队在过去一年里打造智能问答的一些心得。前面几位老师从学术角度讲了自然语言处理技术最新的发展，我会更多地讲到如何打造一个业界可用的智能问答平台。

专题
2018 CCF-GAIR 全球人工智能与机器人峰会
本专题其他文章
中科院赵军：开放域事件抽取
AI安全大牛都来了！智能安全的
国内最高规格AI盛会来袭！CC
今天，人工智能在 CCF-GAIR
CCF-GAIR 袭：100位
哈工大秦兵：机器智能中的文本
more


汪思颖
编辑
关注 AI 学术动态以及各类数据挖掘比赛，加好友请备注人信息~谢谢。微信：awanglala
发私信

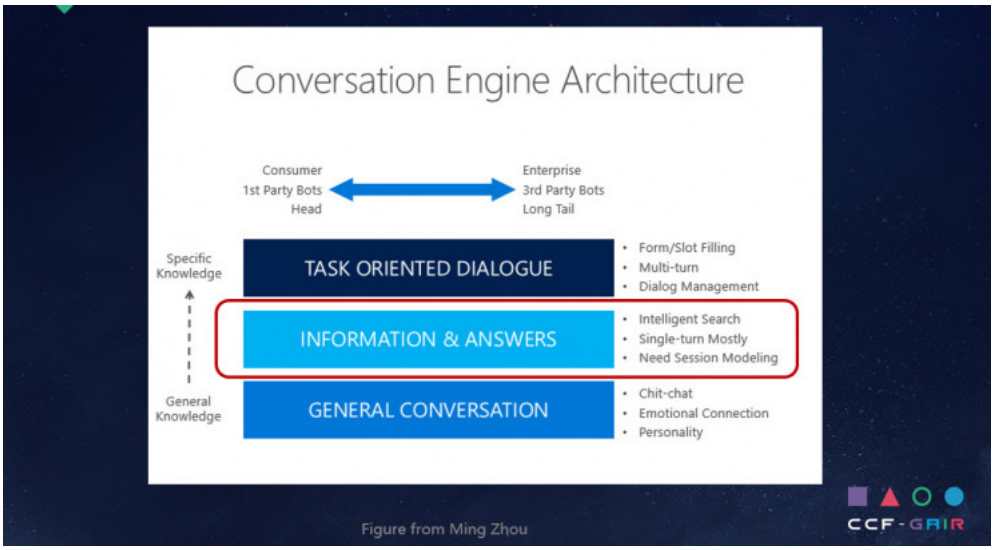
首先对我们团队做一个简短介绍，我们团队成立时间不长，到现在不足一年，成员来自五湖四海。我们研究的重点是自然语言的智能交互，围绕着这一研究重点的内涵和外延，我们在问答、对话、搜索这些领域都做了一些探索和尝试，也在 AAAI、IJCAI、SIGIR、EMNLP 等学术会议上发表了多篇论文。我们和腾讯金融云的同事一起打造了金融行业的智能客服解决方案，和腾讯视频云的同事一起打造了通用行业的小微智能客服解决方案，另外我们研发和支持了腾讯云内容理解产品。

下面这张图是微软周明老师的一张图，也是我非常赞同的一种分类。我们在业界打造通用智能问答平台的时候要解决如下几种问答类型：

第一种类型，任务驱动型。这种类型通常是用户希望去完成一些任务，比如查天气、查汇率等。

第二种类型，解决用户信息获取类的问题。这种类型也是我们这次分享的重点，我们将主要在这点展开。这也是目前业界落地最多的一种问答系统类型。

第三种类型，通用闲聊型。比如微软的小冰、苹果的 Siri 都支持通用闲聊，通用闲聊的加入会使对话系统更富于人性化，也可以加入个性化信息、用户画像信息，包括前面教授们提到的情感信息。



今天我将重点分享第二种问答类型，即如何让问答系统解决用户的信息获取类问题。这可以看作是一种问答，在问答领域可以将数据分为三种类型：

第一种，基于标准的、结构化的知识，比如说 FAQ 和 KG。FAQ 是常见问题解答，KG 是组织好的知识图谱，这两种都是比较结构化的数据类型。

第二种，数据以非结构化的形式存在，比如说表格、文档。

第三种，多模态、跨媒体问答，比如说 VQA，或可能存在视频、音频问答的语料库。

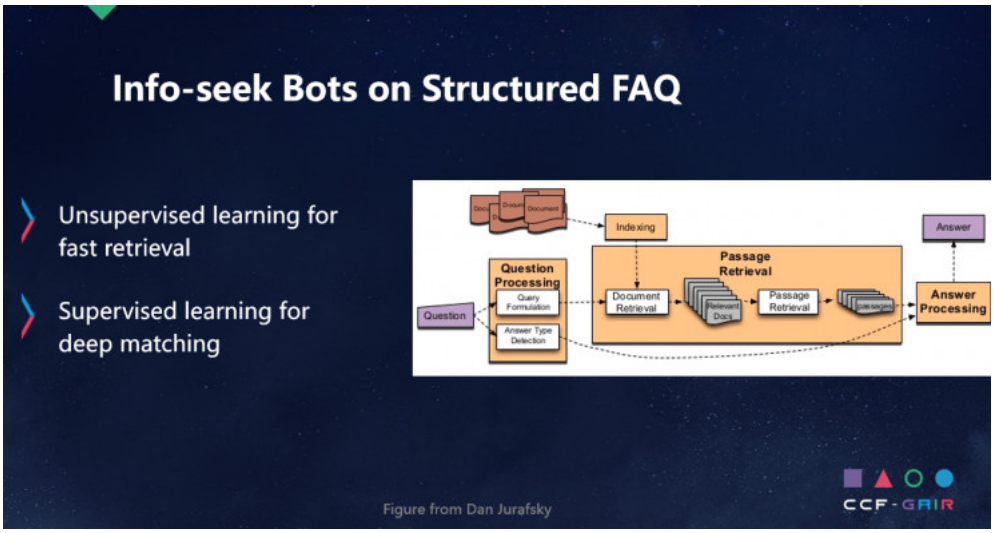
接下来讲我们在结构化的 FAQ 上怎么打造智能问答系统。

下图右边所示是一个非常通用的框架，这个框架跟搜索引擎的框架非常类似，主要包括如下模块：

首先是问题处理模块，这一模块的工作包括查询、问询改写，错词纠正，同义词替换。第二步是召回，即在 FAQ 里召回文档，最主要的目标是召回要快，召回率要很高，准确性可以比较低，可以召回不那么相关的信息。之后，我们会做一个匹配。

这跟搜索有什么区别？搜索会得到搜索结果列表，有很多的评价方式，比如说基于列表的评价，然后再用一些指标来评价搜索结果的好坏。问答的要求更高，有时候是没有列表显示出来，只有一句话或者只有一个答案，我们要追求 top1 的准确率，对匹配的要求会更高一些。

这里提到两种方式，非监督学习和监督学习，大家可以用非监督学习快速召回，但监督信号的加入可以较大提升匹配的准确性。



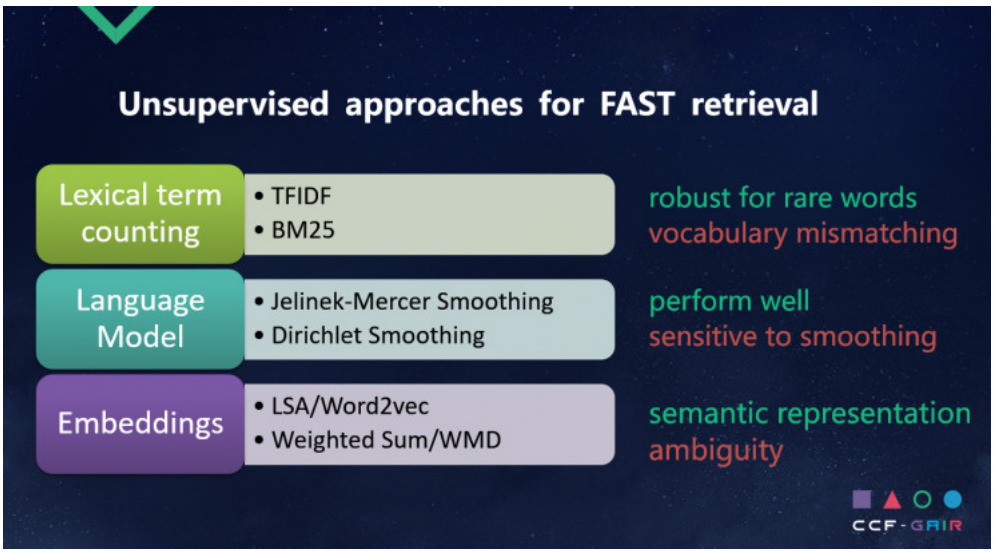
讲一讲比较流行或者在业界我们用得比较多的快速召回的方案。

第一种，基于词汇计数（Lexical term counting）的方法。大家都很熟悉这类方法，它基于字面匹配，好处在于很简单，对长尾的词有很好的鲁棒性，只要在标准问里有出现过，做匹配的时候一定可以召回。但是它的缺点很明显，它基于符号，没有语义层面的理解，所以很难处理字面不同语义相近的表述。

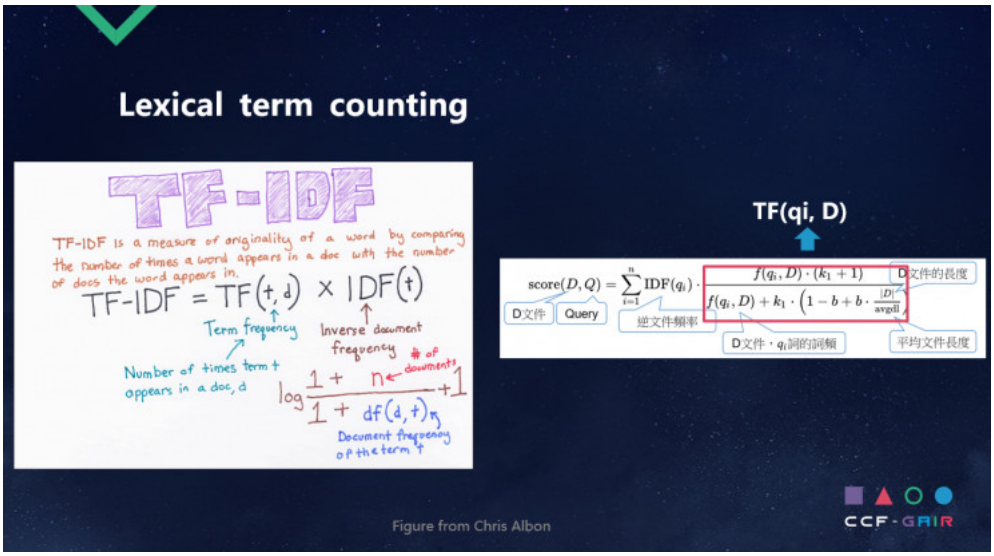
第二种，基于语言模型，主要思想是用概率的方法来判断知识库里面的 FAQ 和用户问询在哪一种在概率上更为接近。它的实战表现更好一些，但是它对语言模型参数的优化非常敏感，所以要做很多平滑实验。

第三种，基于向量化的方法。我把用户的问询投射到这样的向量空间里去，把知识库的 FAQ 也投射到这样的向量空间里去，在向量空间里用距离的方法去做度量。目前存在很多种投射方案，比如基于矩阵的分解，可以把向量拿出来，还可以基于一些其他方法做向量化，向量空间算距离的时候也有很多种方法，比如用平均求和来算这两个点之间的距离。

WMD 是 2015 年的工作，它用了一些更加新的方法来算这种距离，这样的方法比简单的平均化求距离要更好一些。但存在一个问题，这种方法对多义性的解决不太好。



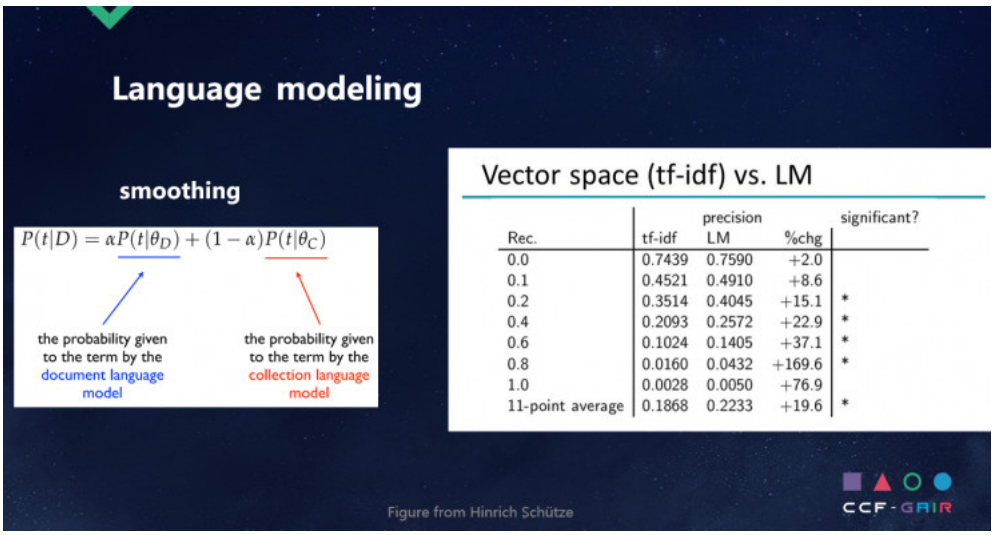
后面先讲 TF-IDF，这个想法非常直观。TF 表示这个词在当前文档的频繁程度，IDF 表示这个词的性质。如果 IDF 非常高，说明这个词是一个比较独特的词，如果比较低，说明在很多文档中普遍出现，是一个比较泛的词。我们可以对它进行求乘积，得到 TF-IDF 的分数。



语言模型的基本思想是用概率分布的方式去描述句子。语言模型在很多地方都有广泛应用，比如说在机器翻译、拼写纠错中，它可以判断哪种可能性更高。在我们自己的召回里，我们基于文档或者标准 FAQ 来生成当前用户 Query 的概率大小，进而判断得分，这是语言模型用在 IR 上的基本思想。

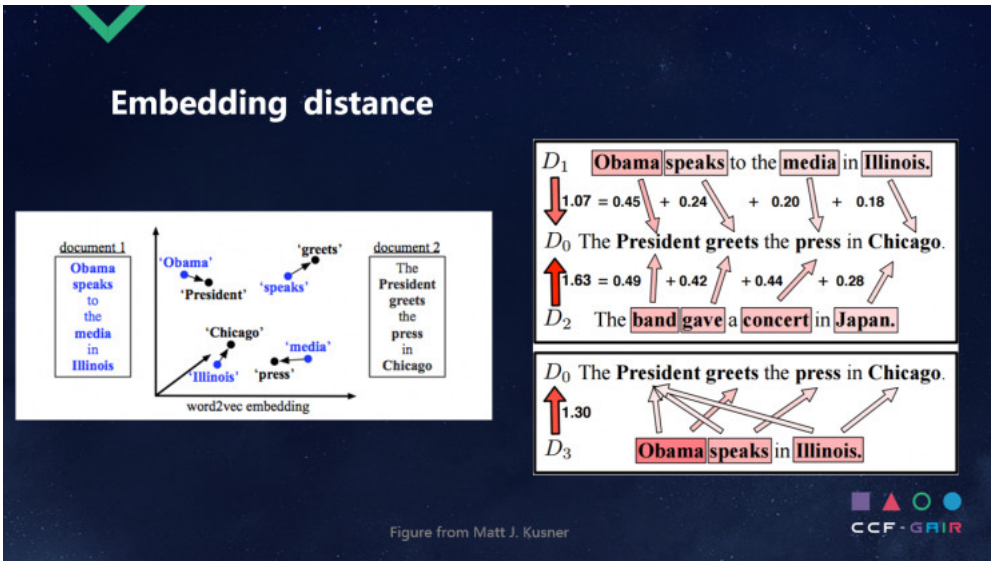
要解决的问题和遇到的困难是，很可能用户 Query 中的词并没有在 FAQ 里出现，所以我们要做如下的平滑——如果该词出现了要怎样，如果没有出现要怎样。

不同平滑的方法对应不同的语言模型。从实践角度来看，TF-IDF 和语言模型比较，语言模型有比较大的提升，可以看到图右几项比较里有五项取得显著提升。

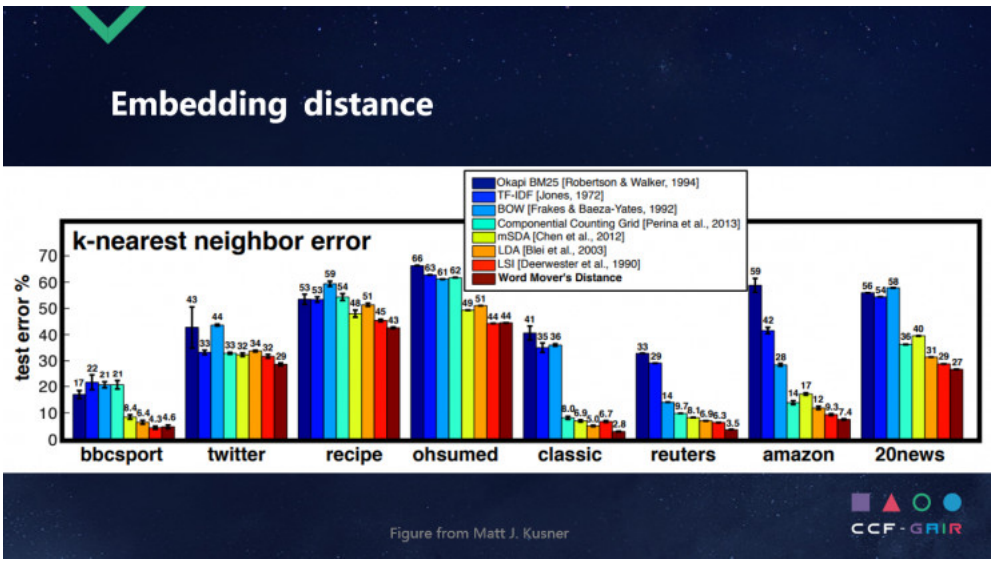


刚才提到词移距离的方法，这个方法就是 WMD，基于加权平均的方法比较简单，这里主要讲一下 WMD。我们投射的每个词都要算距离，需要找到与这个词最像的那个词，而不是简单地把这个词和所有词加权平均以后才会扩散。

看这个例子，对于「奥巴马」这个词来说，跟奥巴马最相近的词是总统，把这个词算出来，求的应该是最小的距离。这有点像旅行的问题，我有一些货品要从这边移到那边，总是要找出每次移动的最小距离，把这些距离加起来。

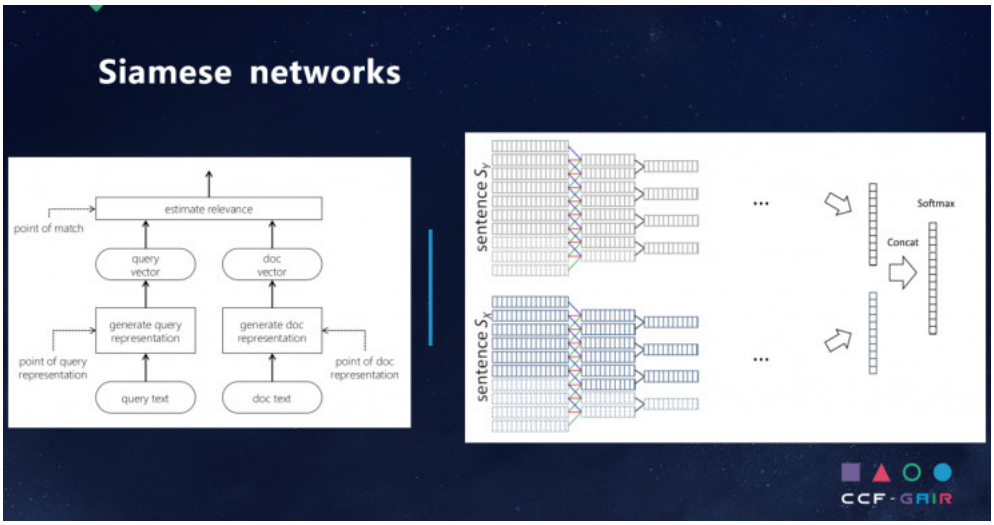


从下图可以看到 WMD 的效果，在几个评测里，它的错误率相对来说比较低，比其他方法低了将近十几、二十个百分点。它的实际效果确实不错，但是算法复杂度比较高，因为需要全部做两两比较来计算，所以耗时会长一点。我们一个很大的要求是快，对 WMD 有一些扩展研究，有兴趣的同学可以继续关注后面的一些工作。



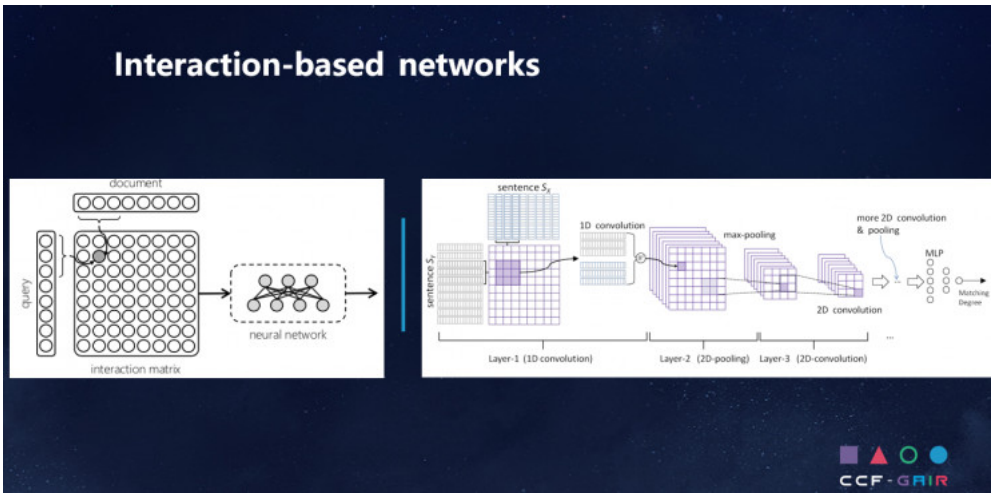
刚才讲的是快速召回，接下来一个很关键的点是做深度匹配。现在有很多深度匹配的方式，最多的是监督匹配，在这当中有两类比较多的方法，一类是 Siamese 网络，一类是基于交互矩阵的网络。

像 CNN 的 ARC-1 就是 Siamese 网络典型的例子，Siamese 网络比较直观，它的想法很简单，把两个输入用同样的编码器做一个表达，把表达做出来以后，可以用一个模块来做相似度的计算，它的特点是共享网络结构和参数。右边是具体的实现，可以用 CNN 实现 Encoder。



基于交互矩阵的网络的不同之处在于，除了最终表达相关性度量时，中间某些词可能会有更强的交互，特别是在文档很长的时候。这一类型的网络和 Siamese 网络相比，在两个问句很短的时候打成平手，但是如果问题很长，包含了很多的内容，里面有一些关键信息，这一网络就会有更好的表现，当我们做好表达以后，会看这个表达里面每一个小的词组之间交互的情况。

下图是用的比较好的一个网络，左边是刚才提到的结构，右边加入了交互。左边非常简单，Question 和 Answer 进入以后得到了表达的矩阵，然后再得到向量，最后求出得分，这是非常直观的流程。在 Attentive Pooling 网络里，会把交互放在求向量之前，想要在交互矩阵中得到行的取值和列的取值，就要得到它们重新的表达，再用最后的表达求扩散的分数。对于长文档，特别是如果 FAQ 很长，基于交互矩阵的网络会带来更多信息。

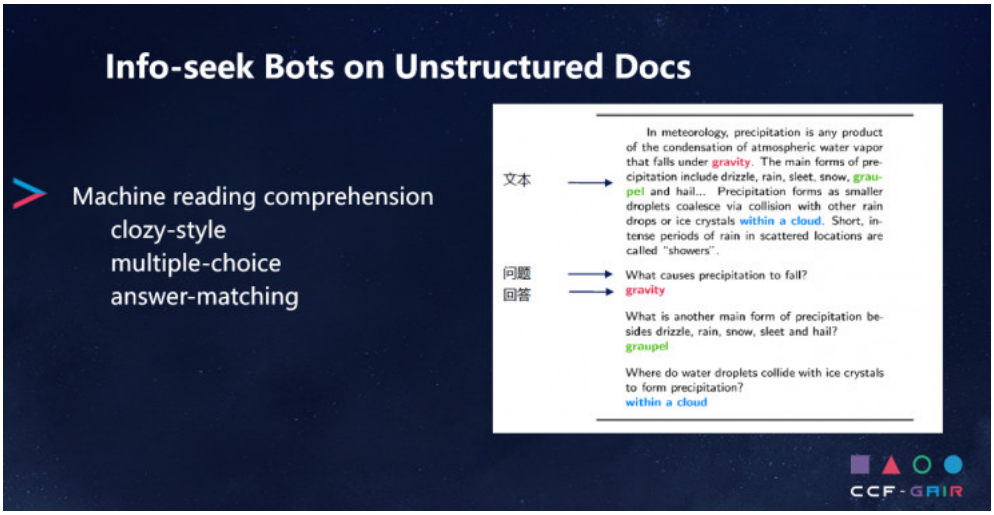


刚才讲的是结构化文档构建的情况。在实际场景下，结构化的数据很少，因为结构化意味着人力的投入，意味着很多人要去做数据标注、做知识库的构建。目前非结构化的数据更多一些，这也是我们团队研究的重点，也是我们认为非常有前景的方向，即如何在非结构化文档里寻找信息和答案。

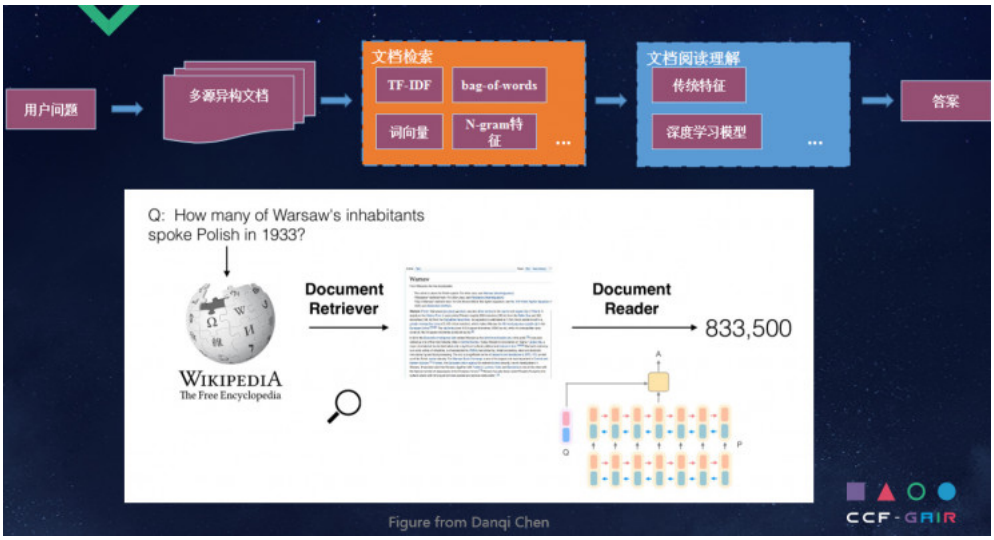
非常相似的一个领域是机器阅读理解，它有如下几个类型：

> - 完形填空。在文章里挖几个实体词，用模型算法把实体词填上。
> - 多项选择。读了这篇文章以后我会有问题，接下来会有几个答案，我会从里面选择所对应的答案。
> - 答案匹配。通常给定一个问题，这个问题在原文中有出现，找到原文中哪些内容可以回答这个问题。

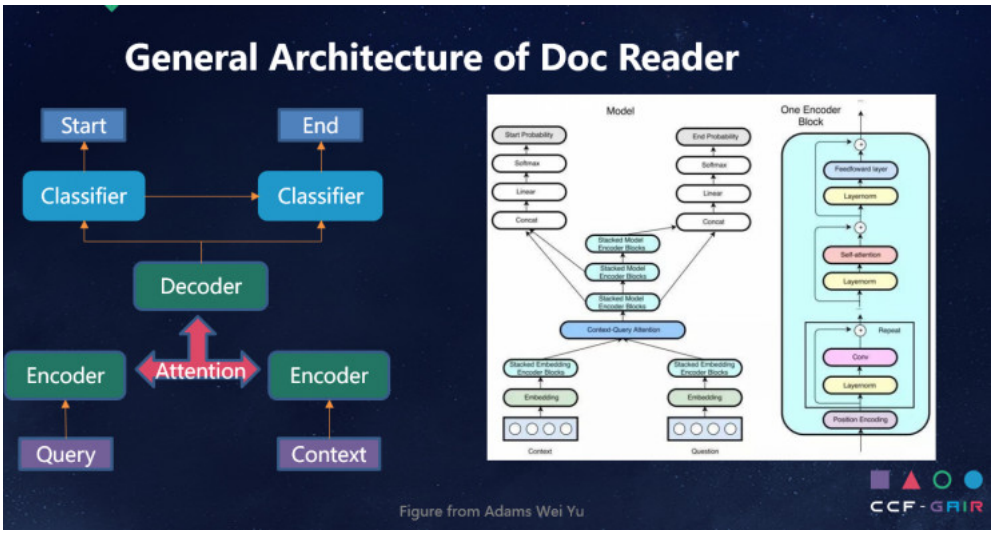
下图右边是比较典型的答案匹配的例子，这是斯坦福比较著名的 SQuAD 比赛。SQuAD 现在已经出了 2.0，在 1.0 的时候，所有的问题答案都在原文中出现过，所以很多学者觉得这并不是特别符合实际，所以现在升级到 2.0 版本。我们现在研究的场景还是基于答案是原文中有的，类似政府的一些文书、材料、文件，可以在里面找到答案。



实际上我们离真正做阅读理解还存在差异，我们在业界做阅读理解时，首先要做召回，因为不知道哪篇文章里包含了问题的答案。这里先要快速去做检索，做完检索才到下一个部分——文档阅读理解模块。文档阅读理解并不是这两年才发展起来的，很多年前就有相关工作了，以前是基于传统特征，比如说基于三元组的关系抽取的方法，现在更多想用一些深度模型的方法来做阅读理解。



下面是一般的 Doc Reader 框架。这几年阅读理解框架非常火，也有很多相关工作。下图右所示是谷歌最近得分最高的一个单模型 QANet，可以看到，首先会有一些 Encoder，把用户的问题都读取进来，它的核心在 Attention，可以设置很多 Attention 机制，比如基于字、基于短语，或基于时刻、步长的 Attention。Attention 是一个很大的舞台，可以尝试很多方法。当拿到问题和文档交互的内容信息，送到 Decoder 部分，来生成文档在文章中的位置。所以这是一个分类问题，即这个词是不是这个档案的开头，另外是找结尾，看这个词是不是档案的结尾。这是一个比较通用的框架。



最后谈谈我们在业界的一些心得。

首先，要重视 Baseline，这一点非常重要。不要把 Baseline 搞得太复杂，因为要通过 Baseline 理解数据和问题。

第二，尽快地构建 pipeline。我们的 pipeline 是一整套系统，包括数据处理、模型训练、模型加载、模型预测、模型评价，特别要注意评价指标和整个流程的打通，只有建立 pipeline 才有迭代的基础，如果没有 pipeline，就没办法迭代，没办法评价模型，也没办法更新框架。

第三，没有免费的午餐，没有倚天屠龙刀，不存在一种可以解决所有问题的算法，算法一定有其适用的数据和场景。有了基准和评价标准，我们才可以尝试更多模型，才能知道模型在哪种条件下更加合适，做到扬长避短。

最后，要有领域相关的数据。领域相关的数据不只是指训练数据，也包括该领域的专家经验和知识，与该领域相关的框架和模型。就我们的经验来讲，领域数据的优化，比如清洗领域数据，或者构建领域词典、词表，这些方法带来的提升比较显著，甚至比模型带来的提升更加显著，所以要重视领域数据的工作。

我今天的分享就到这里。感谢大家！

相关文章

gair 自然语言处理 对话系统


当区块链、生物识别遇上金融，将碰撞出怎样的火花？


哈工大秦兵：机器智能中的文本情感计算 | CCF-GAIR


犹他大学计算机系终身教授承恒达：人工智能中不确定

为什么自动驾驶芯片是人工智能芯片中的珠穆朗玛峰？

提交

热门关键字

热门标签 人工智能 机器人 机器学习 深度学习 金融科技 未来医疗 智能驾驶 自动驾驶 计算机视觉 激光雷达 图像识别 智能音箱 区块链 智能投顾 医学影像 物联网 IoT 微信小程序平台 微信小程序在哪 CES 2017 CES 2016年最值得购买的智能硬件 2016 互联网 小程序 微信朋友圈 抢票软件 智能手机 智能家居 智能手环 智能机器人 智能电视 360智能硬件 智能摄像机 智能硬件产品 智能硬件发展 智能硬件创业 黑客 白帽子 大数据 云计算 新能源汽车 无人驾驶 无人机 大疆 小米无人机 特斯拉 VR游戏 VR电影 VR视频 VR眼镜 VR购物 AR 直播 扫地机器人 医疗机器人 工业机器人 类人机器人 聊天机器人 微信机器人 微信小程序 移动支付 支付宝 P2P 区块链 比特币 风控 高盛 人脸识别 指纹识别 黑科技 谷歌地图 谷歌 IBM 微软 乐视 百度 三星s8 腾讯 三星Note8 小米MIX 小米Note 华为 小米 阿里巴巴 苹果 MacBook Pro iPhone Face GAIR IROS 双创周 云栖大会 先打 智能硬件公司 智能硬件 QQ红包 支付宝红包 敬业福 vuforia 体验设计 网络 家用摄像头 google 新操作系统 360度视频 第一台计算机 freeme os amazon echo 仓库机器人 阿里云人工智能 小米概念机mix怎么样 5g信号穿墙 htc vive 体验 阿里buy+ 更多

0