



雷锋网

读懂智能&未来

AI研习社

AI投研邦

活动

专题

爱搞机

业界

人工智能

智能驾驶

AI+

金融科技

未来医疗

网络安全

智慧城市

机器人

行业云

人工智能

正文

专访腾讯钟黎：知文团队在智能问答系统方面的探索

本文作者：汪思颖

2018-07-29 12:18

导语：理解语言不仅涉及感知层面，还涉及认知、推理、知识、演绎，语言背景还和文化、历史息息相关。从这一点看，语言理解技术还在早期，还需要我们有更多的坚持、耐心。

腾讯知文

知文团队

+8

AI影响因子

活动

企业：腾讯

操作：

事项：专访

对于刚刚成立的腾讯知文团队来说，过去是收获颇丰的一年。虽然成立尚不足一年，但这一团队已在AAAI、IJCAI、SIGIR、EMNLP、COLING等顶会上发表数十篇论文。

知文团队隶属于腾讯SNG数据中心，服务和支撑着包括腾讯云、QQ、QQ空间及相关产品和业务。这一团队目前的研究重心在自然语言智能交互，围绕智能交互的内涵和外延，团队的学术研究涉猎广泛，包括但不限于问答、对话系统、文本摘要、知识图谱、机器阅读理解等。

据知文团队负责人钟黎介绍，知文核心算法团队在15人左右（hiring!），此外还有若干产品和工程职位。钟黎目前主要负责知文智能产品（包括智能对话、智能问答、智能搜索等）的研发和落地，此前他曾在软银AI Lab、微软亚洲研究院、IBM研究院（新加坡）从事人工智能与机器学习的相关研究工作。

最新文章

日活超1.6亿，揭秘快手背后的AI技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启元，美的决胜智能，专访一家有愿景的中国业「启元世界」

面向数据科学和AI的开发者推荐：Python、R各7个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- 汪思颖
- IoT
- 自行车
- 谷歌地图
- 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

当月热门文章

- 重读 YouTube 学习推荐系统论文，字字珠玑，大神文

YouTube 深度学习推荐系统的工程问题

12月19日 顶级论文报告会暨 CAAI 0 发成果奖报告将于哈工大（深圳）开幕

CCS 2018论文解读：使用少量样破解文本验证码

张潼离职腾讯，或赴港科大接班海微众的杨强



围绕知文团队目前的发展以及在 NLP 领域中的心得体会，雷锋网对这位腾讯高级研究员进行了独家专访。（凭借此次专访，腾讯知文团队在雷锋网学术频道 AI 科技评论旗下数据库项目「AI 影响因子」中增加 8 分。）

此前，他在由雷锋网(公众号：雷锋网)承办的 CCF-GAIR 大会上谈到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、数据分析非常重要。

这些建议给大家带来了全新思路，而在专访中，钟黎进一步围绕 NLP 的发展与难点，全方位展示了智能问答平台的打造之路。

钟黎表示，在打造智能问答平台中，一个主要的难题是领域迁移。提到解决方案，目前他们希望先打造通用、多领域多任务、自动学习的智能问答平台，让新的业务、新的场景、新的领域能够以较低的成本和门槛迁入进来，然后再提供一系列方法和工具，帮助做领域的定制。

对于 NLP 技术在腾讯的重要性，他表示，目前 NLP 已经为业务带来了许多价值。但是，自然语言处理和理解本身是非常困难、非常复杂的领域，理解语言不仅涉及感知层面，还涉及认知、推理、知识、演绎，语言背景还和文化、历史息息相关。从这一点看，语言理解技术还在早期，还需要我们以更多的坚持、耐心，一点一滴去突破。

以下为雷锋网对钟黎的采访内容，做了不改变原意的编辑整理。

问：你的 NLP 研究经历如何？你是从大学就开始这一研究吗？

答：其实不是，我最开始的经历和视觉相关，后来又转到推荐，再之后又研究机器人、语音。语音其实也是 NLP 的一部分，但它跟 NLP 又不太一样。我是来到腾讯才开始做纯 NLP 研究。

基于过去较为多样的研究和实践经历，现在在做 NLP 任务时，我更多地会思考这些任务之间的共性，会尝试一些其他学科的研究工具、经验和知识是否有助于解决当前的问题，相对于完全 NLP 背景出身的研究人员来说，我觉得这也是自己的一个特点。

最新文章

日活超1.6亿，揭秘快手背后的 AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国业「启元世界」

面向数据科学和 AI 的开发者推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图

对于我们这一团队来讲，更多地是要考虑到更加复杂的生产环境、业务场景以及实际问题，就像前面提到的学科交叉，如果具有多学科交叉背景，其实会更有优势。

问：结合你此前的工作经验，NLP 技术在落地上经历了怎样的发展历程？

答：NLP 技术其实在企业界的应用非常早，例如在 PC 时代，搜索是最早把 NLP 技术作为核心技术的应用之一。我此前做过推荐、人机对话、语音识别，这些研究都涉及到 NLP 的一些应用。

它在工业界的落地上，会有这样几个历程：首先还是问题和数据驱动，先出来了一些实际问题，比较适合用NLP相关的技术去解决，然后在这些问题里，我们再去提炼比较具有共性的一些方案，把它从业务逻辑里抽离出来。

问：你提到此前的三个方向，一是推荐，二是人机对话，三是语音识别，这三个方向各自的难点在哪里？

答：这三个方向都是很大的领域，有很多的难点与挑战，这里仅仅简单提及一下。

在推荐里，基本需要做到内容（物品）理解，用户理解，场景理解，并且还要做好三者的关联。这里就仅提下内容理解的难点，内容有两点特征：

第一，多模态内容的融合，可能不光有文本内容，还有视觉内容，实时地处理这些多模态的数据会有些挑战。

第二，内容更新很快，时效性很短，所以必须要挖掘出最新的热点、事件，但想利用 NLP 技术完全自动地去发现热点话题、事理图谱也比较难。

人机对话是我们现在做的核心，难点也非常多。这里最大的问题在于我们并没有实现真正的智能。机器本质上还是数据驱动的算法模型，需要大量的数据来进行训练。数据和模型其实是受限于领域的，一旦做领域的迁移，可能就做不好了。所以在对话里，如何做好领域的迁移，我觉得这也比较难。

谈到语音识别，其实近场语音做得很好，但如果是远场，比如说收集器在比较远的地方，然后又是多人对话，语音有重叠，这就会涉及到很多问题，比如语音分离，识别谁是说话人，另外，多人会话比较嘈杂，信噪比比较低。远场语音可能还涉及回声消除。现在远场语音还比较难，也是目前重点的研究方向。

问：这些学科之间有哪些共性？

答：举个例子，语言模型这一工具在语音里其实用得比较多，因为我们研究语音的时候，是用语言模型来判断转移概率，当然语言模型在NLP里也用得非常多。另外，谈到推荐的架构，推荐有好几层，有召回、重排、精排等，这一架构其实跟 NLP 里搜索的架构有些类似。所以这些技术会有一些共通点。

问：深度学习的发展会促进 NLP 的发展吗？

深度学习更多是一个工具，它在 NLP 里也取得了很多进步，但相对来说还没有到一统天下的程度。在今天（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）讲到的模型既有深度学习模型，也有非深度学习模型，还涉及到很多比较简单的模型。

我之前也提到了，没有完美的模型，在实际运用的时候，一定要找到适合数据和场景的模型。深度学习并不是万能的，最重要的是理解问题和数据。

问：从技术的曝光度和产品的落地情况来看，NLP 离计算机视觉还存在一定距离，这其中的主要因素是什么？

答：这个问题和之前有些相似。语言本身有它的复杂性，最简单的例子，图片的基础存储都是像素矩阵，美国的图片、中国的图片、日本的图片的基本表示都是一样的像素点，动物的图片和飞机的图片基本表示都是一样的像素点；然而在语言里的情况要复杂很多，没有 universal 的表示、没有 atom unit、本身具有先验结构、需要外部语境与知识。这些困难使得 NLP 在应用上和视觉有些差距，但目前应该是有越来越多的落地了。

问：知文团队近年来的发展路径如何？挑选研究方向的主导因素是什么？

最新文章

日活超1.6亿，揭秘快手背后的AI技术
TensorFlow 2.0开发者测试版啦，正式版推出指日可待
深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」
面向数据科学和 AI 的开发库推荐：Python、R 各 7 个
AAAI 牵头示范如何正确地给小学生教人工智能
深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊智能手表
- 摩托罗拉Magic Leap
- 自然语言处理HoloLens
- iPhone 5S丰田
- IoT自行车
- 谷歌地图



答：虽然知文团队成立尚不足一年，但此前在 NLP 方向有诸多探索和应用。对于业界的 NLP 团队来说，我们的研究方向更多是问题驱动型，首先是解决业务技术难题和挑战，在众多的业务需求中，提炼共性，抽象成更为广泛的课题，再进行深入的研究、探索，最后将成果反哺给业务，产生更大的实际价值。

问：知文团队在多个学术顶会上都有论文被录取，目前这一团队与哪些高校有着合作？与高校的合作主要集中在哪些研究领域？

答：我们和 MIT、ETH、哥伦比亚大学、中科院、北京大学等高校均有一些合作，目前合作研究领域也是围绕知文团队的研究重心而展开，以自然语言智能交互为核心，包括问答、对话、机器阅读理解等。我们同时也希望进一步加深同海内外高校的合作和交流，我们有丰富应用场景、来自业务一线的难题和工业级大数据，希望能携手学术科研机构，一起促进产学研的联动和落地。

问：目前在打造智能问答平台的过程中，从技术和落地上来看，面临的主要难点有哪些？

答：难点还是不少，这里列举一些：

- 第一，领域迁移问题。从业界的经验来讲，有效融合领域相关的知识、经验和数据，往往比算法优化能够带来更显著的提升。但领域迁移涉及到数据、模型和架构的迁移，特别对于定制化要求较高的领域，目前还是很难做到规模化。我们现在希望先打造通用、多领域多任务、自动学习的智能问答平台，让新的业务、新的场景、新的领域能够以较低的成本和门槛迁入进来，有一个较好的 warm-start，然后再提供一系列方法和工具，帮助做领域的定制。
- 第二，问答分为多种类型，比如说任务型、寻求信息型、聊天型，但想要判断用户问题的类型并不是那么容易，问题的分界没那么清晰，需要对真实的意图、背后的语境有较好的把握。
- 第三，如何构建垂直领域的知识图谱。目前数据很小，用自动化的方法去构建垂直领域的知识图谱会出现很多错误，可能需要人工构建。
- 第四，问答、对话其核心都是语言的交互。既然是交互，多轮会话是难以绕开的问题。在单次会话时期中，让机器学会适当的反问，获取更多的信息来完成对话，也是目前的热点研究方向。

问：互联网上的数据很多都是视觉与文本信号共同出现，譬如腾讯视频，不仅有视频信息，还有音频信息以及相应的字幕、评论和弹幕等，你们团队是否有将计算机视觉与 NLP 相结合的交叉性研究？你如何看待这些学科的交叉发展？

答：image captioning、video question answering、multi-modal/cross-media search 这些都是跟我们非常相关的领域，事实上知文团队在 CIKM2017 上就有篇基于对偶学习做 image captioning 的工作。

在实际的业务场景中，学科交叉是非常普遍的现象，现在大量的业界数据都是多模态数据，能在业界实用的工作，一定是充分融合了各个学科的技术，充分挖掘出多模态的信息。

此外，学科间的交叉融合也能带来新的创新和突破，例如 CNN 是视觉领域的基础网络，但目前 NLP 领域基于 CNN 的工作也非常普遍。知文就在 IJCAI 2018 上发表了一篇基于 CNN seq2seq 做文本摘要的工作。

问：接下来的研究重心会放在哪里？

答：还是会放在自然语言的交互，包括之前提到的一些难点，我们会不断探索更好的解决方案。

例如迁移是其中一个需要解决的重点，还有非结构化问题也是我们要处理的挑战之一。另外，怎么让回复更加自然、人性化，让机器学会问问题，这些都是需要探索的问题。

要解决的问题很多，未来的路仍然很长。

问：NLP 在腾讯有哪些典型落地场景？你对哪一个场景最为看好？

最新文章

日活超1.6亿，揭秘快手背后的 AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和 AI 的开发者推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图

答：NLP 在腾讯有着丰富而广泛的落地场景。例如，在社交方面，腾讯使用 NLP 技术进行热点和话题的挖掘和发现；在通讯方面，腾讯使用 NLP 技术进行各类服务的唤醒，使用户的体验更便捷；在内容方面，腾讯使用 NLP 技术进行千人千面的内容推荐；在企业服务和云方面，腾讯知文使用 NLP 技术打造智能问答平台，帮助企业降低人力成本，提升客服、营销的效率。

我非常看好知文团队一直深耕的自然语言智能交互场景。搜索是 PC 时代的流量和服务入口，它是中心化的、寡头的。自然语言的智能交互将会是 AI 时代的最重要入口，它是分布式的、去中心化的，每个设备、每个企业、每个用户都会有自己的自然语言智能交互入口。这是从铁器时代走向蒸汽时代，不仅是互联网和商业模式，它将会给全社会、全人类带来影响深远、深刻的变革。

问：从战略层面上看，目前 NLP 技术在腾讯处于什么地位？

答：NLP 技术在腾讯的应用已经很广泛、很久远，为业务带来了许多价值，是非常重要的基础技术地位。但是，正如前文中提到，语言处理和理解本身是非常困难、非常复杂的领域，需要联系实际业务场景、业务数据，才能发挥更大的价值。近年来可以看到，NLP 技术在腾讯各个业务上多点开花，百花齐放，知文也非常幸运身在其中，贡献自己的力量。

问：学术界和工业界需要如何做，才能促进 NLP 的稳步进展？

答：知文团队非常希望能够进一步加深和海内外高校和学术结构的合作和交流，我们有来自业务一线有趣、有挑战、有意义的课题、有丰富的应用场景、有工业级的大数据，希望能够携手学术界的教授专家们，一起来促进 NLP 产学研的联动和落地。

（完）

雷锋网原创文章，未经授权禁止转载。详情见[转载须知](#)。

2人收藏

分享：



相关文章

自然语言处理

NLP



复旦大学陈俊坤：自然语言处理中的多任务学习 | AI ...



AMiner 新功能：技术趋势分析—挖掘技术源头、近期



手把手教你NumPy来实现 Word2vec



从语言学角度看词嵌入模型

文章点评：

我有话要说.....

☐ 同步到新浪微博

提交

最新文章

日活超1.6亿，揭秘快手背后的 AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和 AI 的开发库推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二分）

热门搜索

亚马逊

智能手表

摩托罗拉

Magic Leap

自然语言处理

HoloLens

iPhone 5S

丰田

IoT

自行车

谷歌地图

2018 机器阅读理解技术竞赛颁奖仪式：竞赛冠军 Naturali 分享问答系统新思路

本文作者：汪思颖2018-07-28 23:35

导语：2018 机器阅读理解技术竞赛提供了一个基于真实场景的大规模中文阅读理解百度数据集，本文是冠军团队对比赛方案解读。

雷锋网 AI 科技评论按：7 月 28 日，由中国中文信息学会和中国计算机学会联合举办的第三届语言与智能高峰论坛于北京语言大学举办，Naturali 奇点机智团队作为 2018 机器阅读理解技术竞赛冠军团队，受邀参加本次活动的「机器阅读理解评测论坛及颁奖仪式」，Naturali 联合创始人兼 CTO、国际计算语言学协会会员（ACL Fellow）林德康作为代表进行了 2018 阅读理解技术竞赛系统报告。

获奖名单

1	Naturali 北京奇点机智科技有限公司		
2	ZWYC 北京大学Olib研究组	NEUKG-NReader 东北大学	
3	iiioio 上海理工大学	Delta Text Analysis 台達電子-台達研究院	iDeepWise 深思考人工智能机器人科技(北京)有限公司

机器阅读理解任务一直是自然语言处理领域的重要问题。2018 机器阅读理解技术竞赛提供了一个基于真实场景的大规模中文阅读理解百度数据集，共包含来自百度搜索的 30 万个来自真实用户的问题，对中文阅读理解系统提出了很大的挑战。Naturali 的系统在正式测试集上 ROUGE-L 和 BLEU-4 分别达到了 63.38 和 59.23，在 800 多支报名队伍、105 支提交最终结果的队伍中取得了第一名。

以下为 Naturali 奇点机智联合创始人兼 CTO 林德康在现场分享的竞赛报告：



一、DuReader 阅读理解数据集的特点

日活超1.6亿，揭秘快手背后的 AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」



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

当月热门文章

重读 Youtube 深度学习推荐系统文，字字珠玑，惊为神文

YouTube 深度学习推荐系统的工程问题

12月19日，人工智能顶级论文报告会暨 CAAI 青年科技成果奖报告将于哈工大（深圳）开幕

CCS 2018论文解读：使用少量样本破解文本验证码

张潼离职腾讯，或赴港科大接班海微众的杨强



最新文章

北大开源中文分词工具包 pkuseg

YouTube 深度学习推荐系统的大工程问题

四万高手过招，这份阿里全球学竞赛试题你真的不要看吗

司法鉴定牵手深度学习：Kag 相机型号识别大赛深度分析

CIKM AnalytiCup 2018冠军案出炉，看他们构造模型的诀窍

干货 | BERT fine-tune 终极教程

热门搜索

现在有很多阅读理解问答数据集，百度数据集算是中文阅读理解数据集里最好的。百度数据集相比Stanford 经典数据集 SQuAD 更真实、更具挑战性。SQuAD 是从维基百科的信息来源集成问题，答案必定出现在文本里面，句法比较标准。而此次比赛的百度数据集都是用户的真实提问，不仅包括事实性的问题，还包括意见性的问题，部分问题在百度搜索里没有标准答案，并且问题的表达方式不一定很直接。

比如说意见性问题「iPhone X 好不好用」，或现象描述「响一声就说正在通话中」，这类问题回答起来难度更高。

二、竞赛题目实例

问题：昆特牌什么时候公测

人工标注答案：[‘时间为 6 月 6 日，暂定为期两周，即 6 月 6 日-6 月 19 日。’]

Naturali 答案：[‘巫师之昆特牌国服山丘试炼开启时间为 6 月 6 日，暂定为期两周，即 6 月 6 日-6 月 19 日。’]

参考文档

[‘文章 阅读’, ‘巫师之昆特牌山丘试炼马上开启了，帅编来告诉大家开启时间。’, ‘巫师之昆特牌国服山丘试炼开启时间为 6 月 6 日，暂定为期两周，即 6 月 6 日-6 月 19 日。’, ‘参与过 “青草试炼” 的玩家将直接获得本次测试的资格，无需激活码。’, ‘国服公测时间暂未公布。’, ‘声明：本文由入驻搜狐公众平台的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。’, ‘一款专为游戏动漫爱好者打造的 app 全面的资讯福利，热门资讯图鉴攻略应有尽有。国内外热门手游推荐，精彩不容错过。’, ‘itmo 爱萌游戏-二次元游戏第一门户 itmo 爱萌游戏是国内第一二次元游戏门户网站，致力于打造全新型的手机游戏网站。’]

[‘南方公园游戏在 U2 上放出了新的宣传片昆特牌公测日期发布南方公园游戏这都从去年 4 月延到今年 10 月不过动画 21 季今年 9 月开始播刚好可以衔接到游戏发售日期’, ‘反正昆特牌打了一下午电脑一盘没赢我就放弃了’, ‘我巫师 3 二周日开始玩昆特牌，毕竟一周目没钱，二周目也不继承。现在走到哪打到哪。’, ‘下周就公测？好突然，这么快’, ‘昆特盘看测试录像，氪金也是厉害啊。。。这南方公园竟然跳票到这个时候。。。’, ‘昆特牌国际服已经激活就是看不懂挺期待南方公园的’, ‘期待 spOktoberfest! 另外希望昆特正式服早日上线’, ‘南方公园一听就是垃圾游戏，大家千万别买’, ‘应用吧活动，去领取’, ‘活动截止：2100-01-01’, ‘要不是川普赢了，也不至于这么跳票’, ‘昆特牌还没公测啊，我都以为大家已经玩了很久了’, ‘打牌才是正事 打牌打到十一月玩高清二战 美滋滋’]

解析说明：参考文档是从搜索引擎得到的排名靠前结果的网页全文，一个问题会对应多篇长文档；标注答案 是人工根据文档总结撰写而成，一个问题可能对应多个答案，特别是对意见性的问题来说，有多个答案 是很常见的。从以上案例可见，Naturali 阅读理解系统给出的答案比人工答案甚至还要全面。

三、数据预处理

百度提供五个篇文章作为参考文档。由于文章没有长度限制，我们根据关键词密度，句子位置等信息将超过 500 词的文章压缩到 500 词以内。

以下是我们数据预处理的具体方法：

- 1. 如果标题和各段内容中间插入特殊分割符号连接在一起，没有超过预设最大长度，则将得到结果作为预处理的结果；
- 2. 否则，我们计算各段落和问题的 BLEU-4 分数，以衡量段落和问题的相关性；
- 3. 在分数排名前 k 的段落中，选择最早出现的段落；
- 4. 选取标题，这个段落以及下一个段落；
- 5. 对于第 3 到第 10 个段落，选取每个段落的第一句话；



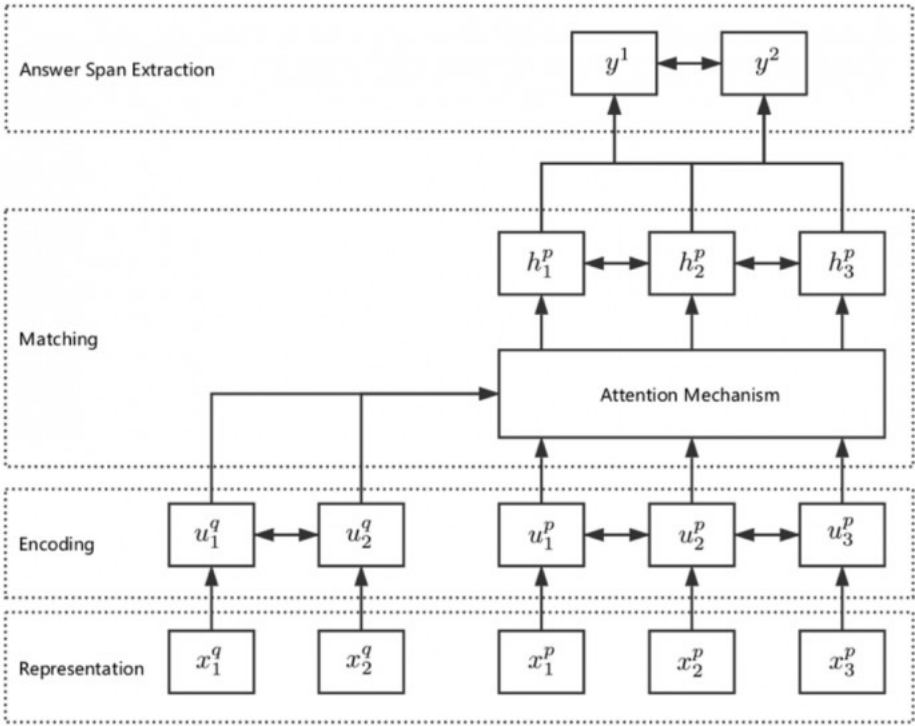
热门搜索



6. 将所有选取的内容以特殊分隔符连接在一起，截取最前面不超过预设最大长度的内容，将得到的结果作为预处理的结果。

四、模型整体结构

我们使用的模型整体结构，是经典的端对端阅读理解模型结构，分为四层：



- 第一层：特征表示层（Representation）
- 第二层：编码层（Encoding）
- 第三层：匹配层（Matching）
- 第四层：答案片段抽取层（Answer Span Extraction）

下面我们对每一层进行简单介绍。

第一层：特征表示层

首先，给定一个问题的词序列和篇章的词序列，我们要对它进行特征抽取，将它变成一个特征向量序列。

我们是在搜狗互联网语料库上进行的预训练。这个数据集比百度数据集还要大好几个量级，所有中文网页都在里，每一词用什么向量表达就是在这里面训练的。

第二层：编码层

得到问题和篇章的向量特征表示序列后，我们分别进行编码。

第三层：匹配层

匹配层是模型比较核心的部分，我们利用注意力机制融合问题和篇章信息。

经过了几种模型的测试，最后我们的系统里用到了 Match-LSTM、BiDAF、DCA 这三种集成模型，相比其他模型，这几种模型效果接近，训练速度较快。在单一模型中我们运用 BiDAF，在集成模型中则会运用到不同的匹配层得到的结果进行集成。

第四层：答案片段抽取层

最新文章

日活超1.6亿，揭秘快手背后的AI技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和AI的开发库推荐：Python、R各7个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图

最终，我们利用指针网络进行答案抽取。

根据百度数据集特点，回答里面可能会包含多个答案，所以我们采用的是第二种利用多个参考答案的公式进行计算，即在多个答案上损失的平均数作为损失函数。

- 常用损失函数

$$L = -\frac{1}{N} \sum_{i=1}^N (\log y_{a_1}^1 + \log y_{a_2}^2)$$

- 利用多个参考答案

$$L = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|A_i|} \sum_{k=1}^{|A_i|} (\log y_{a_1^k}^1 + \log y_{a_2^k}^2)$$

五、最小风险训练

通常的 RC 系统是以提高标准答案概率作为训练的目标，但实际评测的标准是 ROUGE。最小风险训练是拿评测的标准作为训练的目标，需要对每个片段都计算损失函数，所以优化的时间比较长。我们的系统首先用最大似然估计训练得到初始模型，然后直接优化 ROUGE 函数，让我们的 ROUGE 值达到最高。这里 $\text{delta}(y_i, y_i^*)$ 是候选答案 y_i 与标准答案 y_i^* 在 ROUGE 函数上的差。

- 最小风险训练

$$J_{MRT}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{y_i|x_i;\theta} [\Delta(y_i, y_i^*)]$$

- 以最大似然估计训练得到的模型初始化继续训练

$$J = J_{MLE} + \lambda_{MRT} J_{MRT}$$

六、单一模型实验结果

我们的 ROUGE 分数最终能够远远超出基线系统分数，是通过篇章预处理、预训练词向量、其他特征、多个答案、联合训练、最小风险训练等方法综合累计得来的。

最新文章

日活超1.6亿，揭秘快手背后的AI技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和 AI 的开发库推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图

模型	模型各部分的影响			
	Search		Zhidao	
	ROUGE-L	Δ	ROUGE-L	Δ
基线系统	30.77	-	45.90	-
+ 篇章预处理	42.07	+ 11.30	50.52	+ 4.62
+ 预训练词向量	46.98	+ 4.51	55.02	+ 4.50
+ 其它特征	47.15	+ 0.17	55.67	+ 0.65
+ 多个答案	48.75	+ 1.60	56.97	+ 1.30
+ 联合训练	48.76	+ 0.01	57.20	+ 0.23
+ 最小风险训练	49.61	+ 0.85	58.11	+ 0.91

七、集成模型

我们提交的数据是通过集成模型计算出来的，最终根据不同种模型（BiDAF, MatchLSTM, DCA）和不同参数（Dropout：0.1, 0.15, 0.2，联合学习比率：4.0, 5.0）做成了一个集成模型，比单一模型的 ROUGE 分数又高出了 1.5 个点。

在线测试集结果				
	ROUGE-L	Δ	BLEU-4	Δ
单一模型	62.04	-	57.23	-
集成模型	63.62	+1.58	59.51	+2.28

八、总结及展望

我们本次竞赛用的是神经网络端到端的系统，而我曾经在谷歌做搜索问答用的是模块化的系统。模块化系统会把问题分成几部分，首先识别答案类型，再根据类型和问题、文本的匹配度去计算分数。而神经网络系统把所有的步骤放到一个网络里面，虽然没有专门为不同的答案类型建模，但是训练完成后仍然能够覆盖到不同问题类型，比如问「什么时候.....」，找的答案里就有日期。

相比模块化系统，神经网络端到端的系统代码简单很多，并且每次改动、优化都是全局的优化。然而模块化系统的优化，是优化某一个模块，模块之间已经适应了各自的缺陷，其中一个模块变好，其他模块不一定随之改善，使整个系统的优化会变得比较困难。

目前有很多问答系统已经上线了，但是背后的实现还是模块化的机制。很有可能神经网络系统现在的表现暂时还不及模块化系统，但到现阶段它的准确度已经可以有一些应用。比如我们做语音助手的时候，经常将搜索作为一个「兜底」的功能。比如在智能客服应用中，拿客服文档做关键词匹配，但用自动阅读理解就可以把阅读理解系统当做一个「兜底」，找到文档以后便可以把更精简、准确的答案找出来。

最后，如果有对我们的团队感兴趣的同学，欢迎投简历到 jobs@naturali.io，期待你与我们一起，在 AI 语音交互这条路上共同前行。

谢谢大家。

附 Naturali 奇点机智简介：

Naturali 创立于 2014 年 11 月，目前已经将业务聚焦在两个方面：一个是 NI 开放平台——「零编码、五分钟，创造属于你的语音技能」，为各类硬件、APP 赋予 AI 语音交互能力。另一个是一款带有学习功能的第三方语音助手 APP「布点语音」，已经在各大安卓商店上线，目前已经可以覆盖 300+APP，支持 12000+ 语音技能。

最新文章

日活超1.6亿，揭秘快手背后的 AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和 AI 的开发库推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图



雷锋网雷锋网(公众号：雷锋网)雷锋网

雷锋网原创文章，未经授权禁止转载。详情见[转载须知](#)。

4人收藏

分享：

相关文章

自然语言处理

机器阅读理解



复旦大学陈俊坤：自然语言处理中的多任务学习 | AI ...



AMiner 新功能：技术趋势分析—挖掘技术源头、近期



手把手教你NumPy来实现 Word2vec



从语言学角度看词嵌入模型



百度两论文被 ACL 2018 录用，阅读理解新进展一览 |



百度NLP团队在微软 MARCO阅读理解数据集上



360企业安全集团宣布完成 9亿元B轮融资



腾讯知文团队负责人钟黎：从 0 到1 打造下一代智能对

文章点评：

我有话要说.....

☐ 同步到新浪微博

提交

最新文章

- 日活超1.6亿，揭秘快手背后的 AI 技术
- TensorFlow 2.0开发者测试版啦，正式版推出指日可待
- 深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」
- 面向数据科学和 AI 的开发库推荐：Python、R 各 7 个
- AAAI 牵头示范如何正确地给小学生教人工智能
- 深度强化学习从入门到大师：过Q学习进行强化学习（第二分）

热门搜索

- 亚马逊智能手表
- 摩托罗拉Magic Leap
- 自然语言处理HoloLens
- iPhone 5S丰田
- IIoT自行车
- 谷歌地图

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虚拟现实 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 xplay5旗舰版 kdd 毫米波 高交会产品 盲人 p85 开源无人机 华为 人工智能 rokid机器人 真空压缩袋 无限流量上网手机卡 微信 小程序 vr元年 kindle管理软件 大容量硬盘 更多



最新文章

日活超1.6亿，揭秘快手背后的AI 技术

TensorFlow 2.0开发者测试版啦，正式版推出指日可待

深度强化学习+启发人类的决策智能，专访一家有愿景的中国企业「启元世界」

面向数据科学和 AI 的开发库推荐：Python、R 各 7 个

AAAI 牵头示范如何正确地给小学生教人工智能

深度强化学习从入门到大师：过Q学习进行强化学习（第二部分）

热门搜索

- 亚马逊
- 智能手表
- 摩托罗拉
- Magic Leap
- 自然语言处理
- HoloLens
- iPhone 5S
- 丰田
- IoT
- 自行车
- 谷歌地图

