文章编号：1003-0077(2018)10-0118-12

2018 机器阅读理解技术竞赛总体报告

刘 凯，刘 璐，刘 璟，吕雅娟，佘俏俏，张 倩，时迎超

（百度 自然语言处理部，北京 100190）

摘 要：机器阅读理解是自然语言处理和人工智能领域的前沿课题，“2018 机器阅读理解技术竞赛”旨在推动相关技术研究和应用的发展。竞赛发布了最大规模的中文阅读理解数据集，提供了先进的开源基线系统，采用改进的自动评价指标，吸引了国内外千余支队伍参与，参赛系统效果提升显著。该文详细介绍技术竞赛的总体情况、竞赛设置、组织流程、评价结果，并对参赛系统结果进行了分析。

关键词：机器阅读理解；自动问答；数据集；技术评测；

中图分类号：TP391 **文献标识码**：A

Overview of 2018 NLP Challenge on Machine Reading Comprehension

LIU Kai, LIU Lu, LIU Jing, LV Yajuan, SHE Qiaoqiao, ZHANG Qian, SHI Yingchao

(BaiduInc., Beijing 100190, China)

Abstract: Machine Reading Comprehension (MRC) is a challenging task in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI). 2018 NLP Challenge on Machine Reading Comprehension (MRC2018) aims to advance MRC technologies and applications. The challenge releases the largest scale, open-domain, application-oriented Chinese MRC dataset, provides an open sourced baseline systems and adopts improved evaluation metrics. Over one thousand teams registered for this challenge and the overall performance of the participant systems have been greatly promoted. This paper presents an overall introduction to MRC2018, and gives a detailed description of the evaluation task settings, evaluation organization, evaluation results and corresponding result analysis.

Keywords: machine reading comprehension; question answering; dataset; technology evaluation

0 引言

机器阅读理解（Machine Reading Comprehension）是指让机器阅读文本，然后回答和阅读内容相关的问题。这项技术可以使计算机具备从文本数据中获取知识并回答问题的能力，是构建通用人工智能的关键技术之一。作为自然语言处理和人工智能领域的前沿课题，机器阅读理解研究近年来受到广泛关注。

“2018 机器阅读理解技术竞赛”由中国中文信息学会、中国计算机学会主办，百度公司承办，旨在为研究者提供开放的学术交流平台，提升机器阅读理解的水平，推动语言理解和人工智能领域技术研究和应用的发展。

竞赛数据集采用了百度公司发布的当前最大规模的中文阅读理解数据集 DuReader[1]。该数据集中的问题和文档均来自搜索引擎的真实场景，符合用户实际需求。在传统阅读理解自动评价指标基础上，此次竞赛针对特定类型问题的评价进行了适当的调整，使其与人工评价标准更为一致。除此之外，竞赛还提供了先进的阅读理解基线系统①，为参赛者快速实验和提升阅读理解技术提供了便利。竞赛吸引了来自国内外的千余支队伍报名参与，参赛阅读理解系统的整体水平得到了显著提升。

本报告详细介绍了此次阅读理解竞赛的整体情况、评测方法、评测结果以及相应的结果分析等。希

① https://github.com/baidu/DuReader

收稿日期：2018-06-27 **定稿日期**：2018-07-26

望能够为国内外学者和单位提供有益的信息，对阅读理解技术发展起到积极的推动作用。

1　竞赛设置

1.1　竞赛任务

本次竞赛任务设置为：对于给定问题 q 及其候选文档集合 $D=d1, d2, \cdots, dn$，要求阅读理解系统输出能够回答问题的文本答案 a。目标是 a 能够正确、完整、简洁地回答问题 q。其中对于是非类型问题 q，我们期望参赛者能够进一步给出相应答案的是非判断信息（Yes/No/Depends）。

1.2　数据简介

竞赛采用的 DuReader[1] 阅读理解数据集是当前规模最大的中文阅读理解数据集。数据集的构建基于真实的应用需求，所有问题都是百度搜索中用户提出的真实问题。文档来自全网采集的网页（Search）和百度知道（Zhidao）文档，答案是基于问题与文档人工撰写生成的。数据集中标注了问题类型、实体答案和观点答案等丰富信息。其中问题分为描述类、实体类和是非类三种类型，而实体类问题和是非类问题中分别包含了进一步的实体答案和观点答案。关于 DuReader 数据集的构建和详细的数据分布信息请参见参考文献[1]。本次竞赛的数据集的分布如表 1 所示，划分为 Search 和 Zhidao 两个不同数据来源的集合，并在测试集中随机添加了 10 万的混淆数据，以避免参赛系统针对性调节参数，保证竞赛的公平公正。

表 1　DuReader 数据分布

	训练集	开发集	测试集	混淆集
Search 部分	135k	5k	10k	50k
Zhidao 部分	135k	5k	10k	50k
全集	270k	10k	20k	100k

1.3　基线系统

本次竞赛为参赛者提供了数据集相应的基线系统源代码。参赛队伍可以有针对性地对基线系统进行改进升级，构造自己的参赛系统。基线系统实现了 BiDAF[2] 和 MatchLSTM[3] 两个阅读理解神经网络模型，二者均为当前主流的阅读理解模型，很多阅读理解模型是以这两个模型为基础进行创新的。本文中将采用基于 BiDAF 模型的系统作为基线系统。

1.4　评价方法

竞赛结果采用自动和人工两种评价方法进行评价。其中自动评价指标将作为直接的评价指标对提交的全部系统结果进行效果评价，用于系统排名和最终成绩认定。而人工评价指标将作为对前 10 名（TOP10）系统进行效果评价和问题分析的主要依据。

1.4.1　自动评价

在自动评测中采用 ROUGE-L[4] 和 BLEU-4[5] 两个指标，其中 ROUGE-L 将作为主要参考指标用于排名。对于数据集中的是非类型问题和实体类型问题，答案中包含观点判断或实体答案枚举的片段对于答案应当有着更大的影响。因此本次竞赛采用了改进的 ROUGE-L 和 BLEU-4 指标[6] 进行效果评价，对于是非类型问题，希望参赛者能够对自己找到的答案做进一步的观点判断，如果判断正确，评估时将会得到一定的奖励；而对于实体类型问题，将直接在评价时对答案中包含的正确实体在评价中进行一定的奖励。关于改进的评价指标及改进效果详见参考文献[6]。在本次竞赛的自动评价计算中，取 $\gamma=1.2$，而是非问题和实体问题类型的激励权重则分别设置为 $\alpha=1.0$，$\beta=1.0$。

1.4.2　人工评价

为了更好地评价系统结果并进行系统问题分析，本次竞赛对自动评价排名靠前的系统进行了人工采样打分评价。评分的主要依据为该答案是否正确、完整并简洁地回答了对应问题。人工评分原则上依据表 2 中的标准，为每个系统的答案给出 0-3 分的打分。对于每一条待评分答案安排五个标注者进行评分标注，最终评分结果采用五人的均值。

表 2　人工评分标准

3 分（完全正确）	答案能够满足问题的需求，答案正确完整，不存在问题无关内容
2 分（基本正确）	答案总体能够满足问题的需求，存在可接受的内容遗漏或无关冗余
1 分（部分回答）	答案仅能满足部分问题的需求，存在难以接受的内容遗漏或过多的无关内容
0 分（回答错误）	答案不能回答对应问题，答案与问题完全不相关

对于不同的待评估系统，评测组织方随机采样相同的 1 000 条问题进行评分，且对不同类型的问题（描述类/是非类/实体类）均依据总体一致的原则

进行打分评估,不同类型问题的具体打分标准略有不同,人工评分样例详见附表1。对于有瑕疵或者错误的答案,我们进一步地考察了候选答案存在的具体问题,以便进行问题分析。

2 组织流程

本次阅读理解技术竞赛为期两个月,具体竞赛组织流程如表3所示。竞赛测试集分两次发放,首次发放一部分测试集供参赛者在线自助评估并查看排名。在线自动评估阶段每个参赛系统每天最多可以提交两次结果。完整的测试集于竞赛结束前一周发放,作为最终排名依据。

表3 竞赛组织流程

3月1日	启动报名,发放部分训练及验证数据
3月31日	报名截止,发放全部训练数据、验证数据及测试集第一部分,开放在线评测
4月23日	发放完整测试集
4月30日	结果提交截止,进行最终离线评测
5月15日	公布竞赛结果,接收系统报告和论文

此次竞赛总注册报名的队伍达1062支,覆盖众多高校、科研机构及企业,其中包含了128支来自美、英、日等14个国家的国际队伍。最终共有153支队伍累计提交了1 489份系统结果。竞赛期间,参赛系统整体水平提升显著,ROUGE-L评价指标上由最初的35.96提升至终赛的63.62,超过半数系统的效果都优于官方提供的基线系统。

3 评价结果

在本报告中对参赛系统依据自动评价的ROUGE-L评分排序进行顺序编号,将系统编号替代系统名称指代各个系统。本报告中将重点就TOP10系统进行评价和分析,完整系统结果详见竞赛官网①。

3.1 自动评价结果

排名前10系统整体的自动评价效果如表4所示,排名前10系统在不同问题类型下的自动评价效果如表5所示。各系统在不同数据来源及问题类型下的对比如图1所示。

表4 TOP10系统自动评价结果

系统编号	Search部分		Zhidao部分		全集	
	ROUGE-L	BLEU4	ROUGE-L	BLEU4	ROUGE-L	BLEU4
人工结果	**68.85**	**71.1**	**68.30**	**68.22**	**68.58**	**69.60**
S1	57.04	54.96	69.93	63.81	63.38	59.23
S2	54.97	49.83	67.20	63.67	60.99	55.93
S3	52.96	47.51	63.36	58.27	58.08	52.49
S4	50.71	43.95	64.62	59.14	57.55	50.87
S5	49.52	41.36	63.85	56.93	56.57	48.03
S6	48.68	43.64	61.15	53.07	54.81	47.98
S7	47.81	43.21	61.35	58.07	54.47	49.58
S8	45.5	40.26	63.61	55.43	54.41	47.77
S9	48.78	44.15	59.81	57.41	54.2	49.14
S10	43.51	39.21	64.01	55.21	53.59	47.21
基线系统	**33.47**	**26.51**	**55.73**	**53.26**	**44.24**	**39.03**

① http://mrc2018.cipsc.org.cn/

表 5　TOP10 系统在不同问题类型下的自动评价结果

系统编号	描述类问题		实体类问题		是非类问题	
	ROUGE-L	BLEU4	ROUGE-L	BLEU4	ROUGE-L	BLEU4
人工结果	**70.76**	**73.12**	**64.09**	**56.49**	**67.42**	**69.00**
S1	65.00	60.60	61.92	54.24	56.32	53.47
S2	61.82	57.79	60.77	50.51	55.61	48.52
S3	58.89	55.60	57.47	44.47	54.14	45.59
S4	58.11	52.67	57.79	45.91	52.67	43.05
S5	56.13	49.58	58.99	44.36	51.74	38.44
S6	57.52	52.63	51.54	36.17	45.82	29.65
S7	54.78	52.17	55.97	43.71	47.26	36.26
S8	56.26	49.55	51.07	40.84	51.77	44.56
S9	54.03	51.56	55.59	43.10	50.93	39.42
S10	53.41	48.57	54.77	43.47	51.00	41.39
基线系统	**44.18**	**39.85**	**45.73**	**38.27**	**41.82**	**31.1**

从数据集来源上看，如图 1 所示，Zhidao 来源的结果普遍优于同系统的 Search 部分结果。相比之下，如表 4 所示，人类阅读理解的效果在不同来源的数据上未显示出明显效果差距。在不同问题类型方面，如图 1 所示，各系统在描述类型和实体类型问题上的答案的自动评价效果相对较好，而在是非类问题上效果相对较差。而如表 5 所示，人工的效果则在实体类型的问题上表现相对一般，在其他两类问题上效果相对较好。

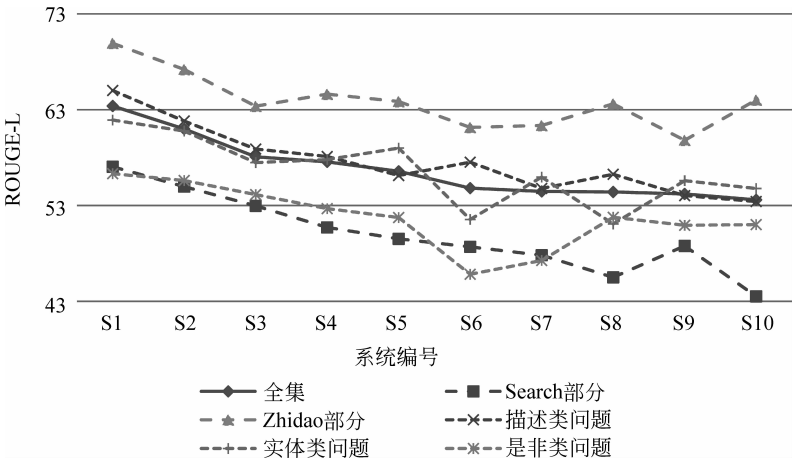

图 1　各系统在不同数据来源及问题类型下的效果对比

3.2　人工评价结果

自动排名前 10 系统的人工评价评分均值效果如表 6 所示。对于所有系统和问题，五人评分的多数一致率达 94.7%，评分质量相对可靠。系统间的人工评价结果显著性检验见附表 2。

如表 6 所示，参赛系统整体最高分为 2.20，距人工评价的 3 分满分评价仍有一定差距。在不同类型问题方面，描述类/实体类/是非类问题的最高人工评分分别为 2.25/2.07/2.33，其中是非类型答案在人工评价标准下为效果最好部分，与自动评价中是非类型答案效果最差的结论不一致。在不同数据来源方面，各系统的 Zhidao 部分结果的人工评价均高于 Search 部分的结果，该结论与自动评价结论一致。

表6 TOP10系统人工评价结果

系统编号	人工评分					
	描述类问题	实体类问题	是非类问题	搜索部分	知道部分	全集
S1	2.25	2.07	2.26	2.16	2.24	2.20
S2	2.18	2.07	2.28	2.13	2.18	2.16
S3	2.16	2.01	2.19	2.09	2.15	2.12
S4	2.16	2.02	2.33	2.10	2.17	2.14
S5	2.11	2.03	2.22	2.08	2.11	2.09
S6	2.13	1.86	2.09	2.00	2.09	2.05
S7	2.07	1.92	2.18	2.03	2.04	2.04
S8	2.08	1.70	2.04	1.79	2.13	1.96
S9	2.06	1.94	2.05	1.98	2.06	2.02
S10	2.08	1.9	2.18	1.94	2.13	2.04

人工评估结果与自动评估结果在不同情况下的排序相关性如表7所示，其中在测试集全集上的自动/人工排序相关性达0.92，整体排序基本一致。在不同类型问题方面，描述类和实体类问题排序基本与自动排序结论一致，其中实体类型自动/人工排序相关性最高，而是非类型问题上当前自动/人工评价相关度较低。在不同数据来源方面，自动/人工评价相关度均较高，相对而言Search部分来源排序相关性较Zhidao部分略高。因此自动评价指标在效果在整体上效果良好，但对于是非类型的评估有待进一步改进。

表7 人工评估与自动评估的系统排序相关性

	描述类问题	实体类问题	是非类问题	Search部分	Zhidao部分	全集
相关度	0.93	0.98	0.65	0.93	0.87	0.92

TOP10系统总体和TOP1参赛系统人工评分分值分布如表8所示。其中可以看出TOP10系统平均可以基本解决(答案评分达2～3分)75%以上的阅读理解问题，而TOP1系统可以基本解决82%的问题。完全回答错误的部分占比均小于10%。

表8 TOP10总体/TOP1参赛系统人工评分分布

评分	TOP10系统/%	TOP1系统/%
3分	39	46
2分	38	36
1分	14	10
0分	9	8

4 结果分析

4.1 主要错误分析

为了更好地进行错误分析，人工评价时对主要错误类型进行了标注。主要错误类型如表9所示。不同的错误类型可能同时出现在一个答案中，在标注时仅标注该答案的一个最主要错误类型。

表9 答案主要错误类型

错误类型	错误描述	占比/%
无答案	空答案、答案无意义或完全错误的答案；	14
不通顺	答案语言不通顺或缺乏合理标点断句	4
不完整	答案中缺少应有的必要信息	48
有冗余	答案中存在冗余无关内容	19
部分相关	答案为与问题相关内容但不是正确答案	15
逻辑不自洽	答案中的内容逻辑矛盾，结论前后相反	<1
是非有误	答案内容中传达的是非结论信息有误(仅针对是非问题)	<1

表9中给出了所有参赛系统的错误类型分布。

其中所有错误中的“不完整”和“有冗余”类型错误的占比最大，占错误总量的 67%。这两类错误的直接原因可以归结为，参赛阅读理解系统有能力找到相关答案，但答案边界定位不够准确。因此，当前阅读理解系统主流的答案边界预测框架的改进空间仍然很大，这类问题也是当前阅读理解技术所需重点解决的问题之一。相比之下，由于相关性问题导致的“无答案”的错误占错误总量 14%，说明当前系统在答案相关性匹配上获得的效果较好，但仍然有改进空间。而错误中涉及到逻辑类型的“部分相关”和“逻辑不自洽”错误也占有相当部分，该类型错误的主要原因可能为系统未能深入理解答案内容逻辑，给出了相关但错误的答案。因此当前阅读理解技术在答案内容上如何进行进一步的逻辑建模仍然有待深入研究。

4.2　不同数据来源错误分析

所有参赛系统在不同数据来源下的错误类型分布如图 2 所示。其中 Zhidao 来源上的错误相对集中，有超过 56%来自于“不完整”错误，而其他问题相对 Search 来源数据错误较少。其可能的主要原因为 Zhidao 来源数据为已经人工处理的问题相关数据，因此文档数据中天然存在的内容冗余和不相关问题较少，所以答案边界定位的问题易集中体现在“不完整”的错误上。

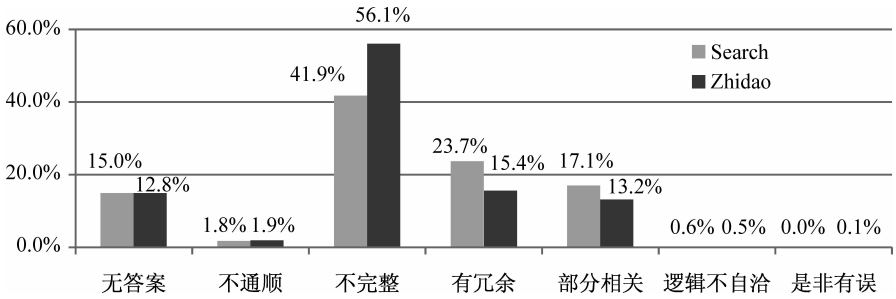

图 2　不同数据来源条件下的错误类型分布

4.3　不同问题类型错误分析

不同问题类型下参赛系统的错误类型分布如图 3 所示。在描述类问题中最突出的错误为“不完整”，实体类问题中分布突出的错误为“无答案”及“有冗余”错误，是非类问题相对突出的错误为涉及答案逻辑的“部分相关”、“逻辑不自洽”以及特有的“是非有误”错误。由此我们可以看出，不同问题类型上的错误分布不同、特点明显，所需解决的难点均不相同，因此针对不同问题类型进行差异性建模对于提升已有阅读理解系统效果具有积极意义。

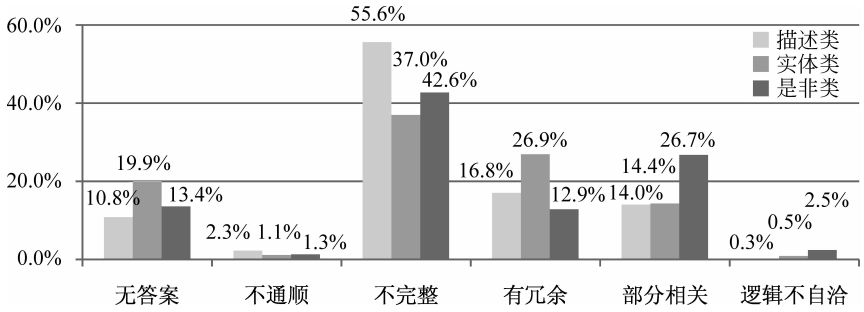

图 3　不同问题类型条件下答案的错误类型分布

4.4　系统技术应用统计

我们采用调查问卷的方式对参赛系统所采用的技术进行统计分析，梳理当前阅读理解技术方面流行或有效的技术模块。其中发放 110 份问卷，返回有效数据 39 份，其中 TOP10 系统均提交了有效问卷，具体 TOP10 应用技术统计点如表 10 所示。大部分参赛系统均采用了基线系统进行改进，少量参赛系统采用了自研或其他开源系统。在建模方法方面，多数参赛系统选择的是流行的多层次注意力建模方法，并采用了是非判断和文档排序的算法模块，仅有少量的系统采用了语言生成改写及强化学习方法。TOP10 各系统的详细系统描述参见附表 3。

表 10 参赛阅读理解系统采用的技术统计

	系统/技术/工具/数据	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
阅读理解系统	基线阅读理解系统	●	●	●	●	●	●	●		●	●
	自研阅读理解系统	●		●				●	●	●	
	开源阅读理解系统①			●						●	●
建模方法和算法模块	是非判断	●		●	●	●		●		●	●
	实体定位				●			●			
	问题分析						●		●		●
	文本结构化						●				
	多层注意力建模	●		●		●		●	●	●	
	文档/段落排序	●		●			●	●	●	●	●
	多文档/答案校验	●		●			●	●	●		●
	答案重排序						●		●		●
	自然语言生成/改写						●				
	系统融合	●	●			●		●		●	
	强化学习训练	●							●		
前后处理和外部数据	改进预处理	●	●	●	●	●	●	●			●
	其他自然语言处理工具②				●	●	●			●	●
	评测外部数据③	●	●							●	

实心圆点代表该系统采用了相关技术。

5 总结

2018 机器阅读理解技术竞赛得到学术界和工业界学者的广泛关注和参与。参赛系统效果提升显著,对推动阅读理解技术发展起到了积极的作用。在人工评价标准下对参赛系统的分析发现,当前优秀的参赛系统已能基本正确回答 75%以上的问题,但与人类阅读理解能力相比仍然存在一定差距。其中,阅读理解系统的错误主要集中在答案边界定位、答案冗余等方面,现有专注答案边界定位的阅读理解技术和模型仍然有很大的改进空间。对于不同的问题类型,参赛系统所表现出来的错误分布有显著不同,针对不同问题类型进行差异性建模是可行的改进方向。在评价标准方面,当前的阅读理解自动评价指标整体上与人工评价具有较好的相关性,但对于是非类型问题答案的自动评价仍然需要进一步的研究和探索。

参考文献

[1] Wei He, Kai Liu, Jing Liu, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications [C]//Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne, Australia: Association for Computational Linguistics, 2018: 37-46.

[2] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, et al. Bidirectional attention flow for machine comprehension[C]//Proceedings of ICLR, 2017.

[3] Shuohang Wang, Jing Jiang. Machine comprehension using match-lstm and answer pointer[C]//Proceedings of ICLR, 2017.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries[C]// Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Bar-

① 参考实现:https://github.com/HKUST-KnowComp/R-Net , https://github.com/NLPLearn/R-net

② 主要为 FastText 和 Jieba 工具包:https://github.com/facebookresearch/fastText,https://github.com/fxsjy/jieba

③ 主要为 Glove[7]预训练向量

celona，Spain：Association for Computational Linguistics，2004：74-81.

[5] Kishore Papineni，Salim Roukos，Todd Ward，et al. Bleu：A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics，2002：311-318.

[6] An Yang，Kai Liu，Jing Liu，et al. Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task[C]// Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne，Australia：Association for Computational Linguistics，2018：98-104

[7] Pennington Jeffrey，Richard Socher，Christopher Manning. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing（EMNLP）. Association for Computational Linguistics，2014：1532-1543.

刘凯（1987－），博士，工程师，主要研究领域为机器阅读理解、机器翻译。
E-mail：liukai20@baidu.com

刘璐（1992－），硕士，主要研究领域为应用语言学、语料库语言学、跨语言语法化。
E-mail：liulu27@baidu.com

刘璟（1984－），博士，工程师，主要研究领域为问答、信息抽取、社会计算。
E-mail：liujing46@baidu.com

附录：

附表 1　人工评分样例

打分	问题	答案	备注
3 分	宁静致远是什么意思	宁静致远：平稳静谧心态，不为杂念所左右，静思反省，才能树立（实现）远大的目标。	准确地给出了定义
	公积金缴款比例	公积金缴存比例区间为 5%－12%。	准确给出了答案实体，没有无关信息
	于谦是八旗子弟吗	谦哥满族人都不是，跟八旗子弟更是八竿子打不着边了。	给出了正确的判断和适当的补充
2 分	银行卡交易明细删除	银行卡交易记录是不可以删除的；记录是长期保存的，没得删除，销户也能查询到，这个是银行系统的对账凭证。	给出了正确答案，但含有无关冗余内容
	喝酒吐完吃什么好	1、喝加有蜂蜜的柠檬汁或橘子汁。2、饮温牛奶，牛奶与酒精混合后，可以减轻人的醉酒程度。	给出了答案实体，但含有无关冗余内容
	喝柠檬水后可以晒太阳吗	喝柠檬水后不可以晒太阳。	给出了结论信息，但未给出适当的解释
1 分	什么是情商	情商往往是决定命运的情商是一种能力，情商是一种创造，情商又是一种技巧。	内容相关但未给出明确定义答案
	血型有哪些	血型有 A 型	含有部分答案实体，但实体召回严重不全
	GTX1050 显卡可以玩绝地求生吗	GTX1050 是最近一款性能比较高的主流显卡	回答有助判断，但未提供有效结论
0 分	青岛极地恐龙游乐园怎么样	青岛极地海洋世界服务不怎样，不值那个钱。	问答对象不一致，答非所问
	查阅的近义词	cháyuè。	无正确答案实体
	磨毛是棉吗	磨毛面料是一种比较具有功能性的产品，在冬天的时候使用磨毛面料的床品，舒适度柔软度及保暖性是比较好的。	未提供有助判断的任何信息

附表 2 TOP10 系统人工评价显著性检验①

	S1	S2	S4	S3	S5	S6	S10	S7	S9	S8
S1	—	●	●	●	●	●	●	●	●	●
S2	—	—	○	●	●	●	●	●	●	●
S4	—	—	—	○	●	●	●	●	●	●
S3	—	—	—	—	○	●	●	●	●	●
S5	—	—	—	—	—	○	●	●	●	●
S6	—	—	—	—	—	—	○	○	○	●
S10	—	—	—	—	—	—	—	○	○	●
S7	—	—	—	—	—	—	—	—	○	●
S9	—	—	—	—	—	—	—	—	—	●
S8	—	—	—	—	—	—	—	—	—	—

附表 3 TOP10 系统描述

系统编号	系统描述
S1	**软硬件环境：** 操作系统：Ubuntu 16.04，64 位； 硬件配置：CPU：Inter(R) Xeon(R) CPU E5-2630 2.20GHz(4 处理器)；内存：128GB；Titan X Pascal GPU，显存 12GB **测试集运行时间：**100min **技术概要：** 采用启发式文档抽取模块； 采用包括预训练词向量、词性标注信息在内的丰富特征； 使用辅助任务联合训练； 考虑多个答案的信息； 利用最小风险训练； 使用集成模型； **参数概要：** Embedding 为 256 维，所有隐层表示为 150 维，使用最大似然估计训练 10 轮，取最好结果再用最小风险训练 1 轮 **训练数据处理说明：** 对于每个文档，使用启发式文档抽取技术抽取不超过 500 词。 **外部数据说明：** 在部分 SougoT 数据上预训练词向量
S2	**软硬件环境：** 操作系统：Ubuntu 16.04.3 LTS (GNU/Linux 4.13.0-39-generic x86_64) 硬件配置：Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz(6 处理器)；内存：132GB；TITAN 12G 显存 **测试集运行时间：**3h，(主要耗费在 search boundary) **技术概要：** 共有八个特征：词向量、文档排序、问题类别、词性特征、精确匹配、上下文匹配、是否由数字和是否由字母组成。 采用基线系统作为基础系统改进； **参数概要：** Embedding 为 64 维，整体训练 5 轮，所有隐层为 156 维，最大答案长度为 300，最大文档长度为 1000 **训练数据处理说明：** 从文档级出发，基于每个真实标注答案对文档进行匹配。每个文档选择匹配分数最大的作为伪答案片段。这样真实的标注答案信息会出现在多个候选文档中，其分别匹配不同的真实答案。本文基于 F1 的词匹配指标，对提供的训练集进行重标注，同时过滤掉匹配分数小于 0.65 的答案片段。为了加速数据重构的速度，使用多进程并行处理方式。

① 按系统人工评分排序，$p<0.05$，实心原点表明两个系统间效果具有显著差异，空心原点则代表没有差异。

续表

系统编号	系统描述
S3	**软硬件环境：** **操作系统：** CentOS Linux release 7.3.1611，64 位； 硬件配置：Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz (4 处理器)； 内存：64GB；GeForce GTX TITAN X 4 GPU，显存 12GB **测试集运行时间：** 12 万测试集，运行 3h **技术概要：** 采用相关段落排序，把最相关的段落提取出来，补全文档到固定长度 500； 采用基线系统作为基础系统改进，基线系统 bidaf 得到的表示进行双层的 self-Attention 来获取最终文档的表示。 **参数概要：** 在实验中，输入的文档长度设置为 500，问题长度 n 设置为 60，答案的最长长度设置为 200，词 embedding 的维度设置为 300，GRU 的隐藏层的维度 d 设置为 150，batch 的大小设置为 24。训练中采用了 Adam 算法进行优化，学习率设置为 0.001。整体训练了两轮，基本上最好的结果在第一轮的时候出现。 **训练数据处理说明：** 对于训练数据没有进行额外的预处理，因为我们发现经过处理之后的线上效果反而不好。 **外部数据说明：** 采用了 Glove 利用百度的语料训练了一份 embedding，作为初始化词嵌入变量
S4	**软硬件环境：** 操作系统：Ubuntu 16.04，64 位 软件环境：Python 3.6，Pytorch 0.4 硬件配置：CPU：Intel® I7-7700k @ 4.2G * 8 内存：32GB DDR4. 显卡：NVIDIA GTX 1080，显存 8GB. **测试集运行时间：** 32min **技术概要：** 采用基线系统作为基础系统改进，仅仅端到端监督训练答案的起始和终止位置。 运用了 BiLSTM，Conv1d，Highway 等多种网络层来提取特征 **参数概要：** 词向量为 200 维，训练时固定不变 循环网络隐藏层统一为 100 维 卷积核大小为 4，步长为 2 dropout 率统一为 0.2 batch 大小为 24 训练 7 轮 学习率为 0.001，并每轮减半 **训练数据处理说明：** 在训练集上自主对答案片段进行定位，得到答案开始与结束位置的训练标签(使用了提供的 preprocessed 数据，但并未使用其标注的 answer doc 与 answer span 字段) 长段落截断为 500 问题截断为 50 测试时答案截断为 300 **外部数据说明：** 无外部数据，其中 GloVe 词向量仅在该数据集上训练，参数为默认值(50 次迭代，学习率 0.05)
S5	**软硬件环境：** 操作系统：Ubuntu 16.04 LTS，64 位 硬件配置：CPU：Intel Core E5-2698；GPU：NVIDIA DGX-1 搭载 Tesla V100；显存 128GB； **测试集运行时间：** 单一模型预测 Test1＋Test2 约 2h：40m，集成模型为六个单一模型组合在一起，集成计算 10min (multi-thread，max_workers＝60)，Yes_No Classifier：8min **技术概要：** 采用基线系统作为基础系统改进； 以完整段落来预测答案； 採用 fastText 訓練詞向量；

续表

系统编号	系统描述
S5	集成 6 个单一 BiDAF 模型； 集成注意力(Attention)機制與相似度(Similarity)機制兩個模型，進行是非類型的答案分类； 标点符号正规化； **参数概要：** FastText：dim＝300，epoch＝5，window＝4，max_ngram＝2，char_ngram＝1～4，lost function＝hs BiDAF (train)：batch_size＝64，dropout_keep_prob＝1，embed_size＝300，epochs＝2，hidden_size＝150，learning_rate＝0.001，max_a_len＝250，max_p_len＝500，max_p_num＝5，max_q_len＝60，optim＝'adam'， BiDAF (predict)：batch_size＝32，max_a_len＝250，max_p_len＝1000 **训练数据处理说明：** 对每一个参考答案皆进行预处理，故原先的一笔数据会变成数笔训练数据（但因时间与设备限制未完成，最终是 347k 笔） 过滤掉相似度低于 0.7 的训练数据 **外部数据说明：** 采用了 fasttext 作为初始化词嵌入变量(以 DuReader 数据训练)
S6	**软硬件环境：** 操作系统：Linux ubuntu 14.04.5 ,64 位； 硬件配置：CPU：Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz (8 处理器)；内存：128GB；Titan 4 GPU，显存 48GB **测试集运行时间：**51min **技术概要：** 该模型在 BiDAF 模型四层网络框架的基础上添加了段落 Ranking 层，针对段落 Ranking，本文提出了多特征融合的段落 Ranking 算法，主要包括段落过滤、段落重组、语义匹配、特征加权、最大覆盖度以及多文档投票。此外本文还在答案预测层，利用候选答案交叉验证以及段落位置信息对答案加权从而对答案进行综合预测； **参数概要：** Learning_rate：0.001，optim：adam，embed_size：150， pretrained_embedding：true，hidden_size：150，batch_size：48， epochs：3，max_p_num：10，max_p_len：600，max_q_len：60，max_a_len：300， splice_L：400； **训练数据处理说明：** 为了数据方便清洗，本文选择直接对 raw 数据进行清洗，由于 raw 数据是没有标签的原始数据，所以清洗完成后还需要生成含有标签的训练数据和验证数据。此外本文对生成算法进行了改进，使生成算法在时间效率上提升了两个数量级。改进的基本原则是首先将可以重复使用的值提到尽量提到循环外层计算，避免重复计算，如 answer 的长度，paragraph 的长度，Counter 计算等，其次是根据预判对一些不用计算的循环选择跳过，从而可以减少大量的低层计算。 **外部数据说明：** 采用了 Glove 作为初始化词嵌入变量
S7	**系统描述：** 软硬件环境：操作系统：Ubuntu 16.04，64 位； 硬件配置：CPU：Interl(R) Xeon(R)CPU 2.00GHz(4 处理器)；内存：32GB；NVIDIA 1080 GPU，显存 8GB **测试集运行时间：**(单 GPU) 50min **技术概要：** 采用段落排序模块； 采用 Gated Recurrent Network，self-attention； **参数概要：** Embedding 为 100 维，整体训练 10 轮，所有隐层为 150 维 **训练数据处理说明：** 采用竞赛提供的分词，链接了(concatenate)每个 document 里的所有段落，并针对其中的长段落进行截断处理。 **外部数据说明：** 采用了 word2vec 作为初始化词嵌入变量

续表

系统编号	系统描述
S8	**软硬件环境：** 操作系统：Ubuntu，64 位； 硬件配置：Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz(48 处理器)；内存：128GB；Titan XP 8GPU，显存 12GB **测试集运行时间：**6h **技术概要：** 采用段落排序模块； 采用基线系统作为基础系统改进； 采用强化学习方法直接优化评价指标 **参数概要：** Embedding 为 300 维，整体训练 10 轮，所有隐层为 150 维 **训练数据处理说明：** 针对其中的长段落进行截断处理。未对标点符号进行归一化处理。 **外部数据说明：** 无
S9	**软硬件环境：** 操作系统：Windows10/Ubuntu16.04，64bit 硬件配置：CPU：Intel Core i7-4790，3.5GHz? 内存：32GB，TitanXP GPU * 1，显存 12GB **测试集运行时间：**约 2h(Ensemble 模型) **技术概要：** 使用基线系统为基础改进； 为验证集和测试集中段落长度不足的样本采用段落选择填充策略； 自注意力机制； 基于词频特征的 Yes/No 态度分类； 在 DuReader 及中文维基语料预训练的 Glove 词向量。 **参数概要：** Embedding 为 300 维，所有隐层为 150 维，整体训练 10 轮(一般第 6 轮收敛到最佳效果) **训练数据处理说明：** 在训练时，将 HTML 特殊符号以及重复的标点符号去除，更新相应的 span_answer，对数字和部分标点进行归一化处理。 **外部数据说明：** 使用了中文维基百科语料，用来预训练词向量。
S10	**软硬件环境：** 操作系统：Debian 硬件配置：CPU：Intel(R) Core(TM) i7-5930K 3.50GHz；内存：128G；GPU：GeForce TITAN XP，显存 12G **测试集运行时间：**30min **技术概要：** 采用字向量，词向量，词性向量，exact match 共 4 个特征 采用基线系统作为基础系统改进 建立知识库进行段落筛选 使用 passage ranking 多任务学习 使用多答案投票机制 **参数概要：** 字向量，词向量维度 300，词性向量维度 10，字向量通过的 GRU 隐层为 100 维，其他所有隐层 150 维。整体训练 5 轮。 **训练数据处理说明：** 采用 ltp 进行分词和词性标注，对长段落截断，短段落进行填充。后处理所有标点转为中文标点。 **外部数据说明：** 使用 Wikipedia Dump 作为字向量，词向量训练语料。