文章编号：1003-0077(2011)06-0021-05

从 IBM 深度问答系统战胜顶尖人类选手所想到的

黄昌宁

（微软亚洲研究院，北京 100080）

摘 要：2011 年 2 月 14-16 日，IBM 的深度问答系统在美国 Jeopardy 电视竞答节目中一举打败该节目的两位前冠军，凸显了计算机在自然语言处理（NLP）技术上超越人类的智能行为，这是人工智能研究历史上意义非凡的里程碑。该文通过这一事件来回顾国内外自然语言处理和自动问答技术研究的某些得失，借此纪念中国中文信息学会成立 30 周年华诞。

关键词：人工智能；自然语言处理；自动问答系统；deepQA；IBM watson

中图分类号：TP391　　**文献标识码**：A

Thinking about DeepQA beating Human Champions

HUANG Changning

(Microsoft Research Asia, Beijing 100080, China)

Abstract: The DeepQA question answering system of IBM beat two human champions on U. S. Jeopardy Show in 14th-16th February, 2011. It obviously shows that Watson, IBM's super computer, outperforms human intelligent behaviors. It is a great milestone in the history of Artificial Intelligence. This paper reviews some gain and loss of natural language processing and automatic question answering technologies. The paper is written for the thirty-year ceremony of the Chinese Information Processing Society of China.

Key words: artificial intelligence; natural language processing; automatic question answering; deepQA; IBM watson

Jeopardy（危险边缘）是美国著名的电视竞答节目，有长达 25 年的历史。它有两位最负盛名的冠军：一位是在该节目中连赢 74 场的 Ken Jennings，另一位是获得奖金总额最高的选手 Brad Rulter，他在这档节目中累计赢下 235 万美元。在今年 2 月 14-16 日的 Jeoparty 电视节目中，装载着深度问答（DeepQA）系统的 IBM 超级计算机沃森（Watson）以 77 147 分的绝对优势战胜两位前冠军，Ken 和 Brad 只分别赢得 24 000 分和 21 600 分。

这是继 1997 年 IBM 超级计算机深蓝（Deep Blue）在一场国际象棋比赛中击败世界冠军卡斯帕洛夫之后，人工智能研究历史上的又一个里程碑式的辉煌胜利，意义非凡。本文通过深度问答系统的成功来回顾国内外自然语言处理（NLP）和自动问答技术研究的某些得失，借此纪念中国中文信息学会成立 30 周年诞辰。

1 DeepQA 系统获胜的主要原因

David Ferrucci 是 IBM DeepQA 项目[①]的负责人。他在论文[1]中称，为了面对 Jeopardy 的挑战，DeepQA 采用了海量并行和基于证据的概率型架构，整个系统体现了高级自然语言处理、信息检索、知识表示、自动推理、机器学习等开放式问答技术。Ferrucci 认为，比分支技术重要得多的是怎样把它们适当地整合到问答系统中来，从而使得这些原本互相交叉重叠的技术都能够把各自的优势贡献给整个系统的三个关键指标——正确率、自信度和速度。Ferrucci 在文中介绍了 DeepQA 的以下三个主要技术：

① http://www.research.ibm.com/deepqa/deepqa.shtml

收稿日期：2011-10-17　定稿日期：2011-10-20

作者简介：黄昌宁（1937—）男，清华大学教授、博导（退休），微软亚洲研究院资深顾问，主要研究方向为计算语言学和中文信息处理。

(1)海量并行主义(massive parallelism):在对回答的多重解释和假设的考察中,充分贯彻并行主义。(2)无处不在的自信度计算(pervasive confidence estimation):系统中没有一个部件可以独立地对一个回答的正误负责,每个部件都在各自产生的结果上伴有一个自信度值,后者随着问题和内容解释的不同而相应变化。底层的自信度处理机制将自动学习如何累加这些自信度分值。(3)整合浅层和深层知识、权衡严格语义学和浅层语义学的使用、利用多种随处可见的本体知识(ontology)。

由于Jeopardy竞答节目要求计算机像人一样能够在3~5秒钟内迅速对一个选定问题做出响应,因此响应速度成为计算机参赛的必要条件。由90台IBM Power-750组成的服务器机群总共拥有2 880颗服务器核心和16TB(16万亿字节)内存。这台海量并行的超级计算机Watson是DeepQA系统取胜的硬件保障。

电视竞答节目允许每个参赛选手对一个选定的问题进行抢答,这不仅要求计算机寻找答案的速度足够快,而且只有当计算机对答案的正确性具有足够高的自信度时,才应当按下蜂鸣器实施抢答。调查显示,Jeoparty的参赛选手至少在40%~50%的问题上达到85%~95%的正确率,才有可能在竞答中胜出。Ken Jenning平均回答62%的问题,回答的正确率达92%。所以要面对Jeoparty的挑战,计算机不仅要像人类选手那样在60%左右的选定问题上达到85%以上的回答正确率,而且要求对每一个回答的自信度做出准确的估值。无处不在的自信度评估(上述(2))强调了自信度计算的重要地位,指出了系统每个部件都将对回答的自信度做出相应的贡献(本文在下一节中还要讨论这个问题)。上述(3)则强调了万维网各种结构化与非结构化知识以及各种NLP技术对回答正确率的贡献。尽管Watson在比赛中没有链接互联网,它的4TB磁盘上存有超过200万页结构化与非结构化的文档供其查询,内容包括百科全书、词典、新闻、文学等领域,如维基百科全文数据库DBpedia①、WordNet②和Yago③等。

2 自信度计算的重要性

传统的问答系统一般包括三个主要部件:问题分析、信息检索和回答抽取。为了面对Jeoparty的挑战,IBM团队在DeepQA的设计中特别强调了对每个候选回答的自信度计算,而且把这项计算分布到上述每个部件的输出结果上,称之为无所不在的自信度计算。

其实,对自信度计算的重视可以一直追溯到1999年第八届国际文本检索大会(TREC-8)举办的首次大规模问答系统评测活动。

TREC-8公布了一个规模约为100万篇文档、近30亿字节(3GB)的TREC新闻文本语料库。评测用的200个问题主要是关于简短事实的问题(fact question),即答案是命名实体(如人名、地名、机构名等)和数字串(如日期、时间、款额、温度、重量、长度等)一类的问题(factoid question)。组织方规定,每个问题的回答必须限定在50字节(或250字节)之内,而且每个回答都必须出自上述TREC语料库中的某个文档,回答形式为:[文档ID编号,回答字符串],否则即使回答正确也不能得分。很显然,这样的评测规定反映了组织方推进QA技术的宗旨,即尽可能节省用户在信息检索过程中所消耗的时间和精力,这是下一代信息检索技术的目标。相比之下,传统搜索引擎根据查询命令回送一个冗长的相关文档列表,已满足不了广大互联网用户的需求。

TREC-QA评测活动虽然没有像Jeopardy竞答节目那样对回答的自信度提出如此苛刻的要求,但自信度计算从来都是QA评测中的一项重要指标。例如,TREC-8允许参评系统对每个评测问题报送5个按自信度高低序号(1~5)排列的回答,而每个问题的得分是按照正确答案所在序号的倒数来计算的,叫做"平均排序倒数"*MRR*(Mean Reciprocal Rank)[2],*MRR*的定义如下:

$$MRR = \frac{1}{N}\sum_{i=1}^{N} 1/\text{正确答案所在序号}$$

公式中,N是问题总数(对TREC-8评测来说,$N=200$)。请注意,参评系统对每个问题报送的5个回答,当它们被裁定为正确答案时由于自信度高低序号不同,其相应得分分别为1,0.5,0.33,0.25,0.2和0(没有找到正确答案)。组织方也承认*MRR*指标不甚合理,因为它对序号大于1的回答打分过低。比如,系统的整体性能完全可以改用平均一选

① http://wiki.dbpedia.org
② http://wordnet.princeton.edu
③ http://www.mpi-inf.mpg.de/yago-naga/yago

正确率、平均二选正确率、平均三选正确率等传统的评价指标。TREC-8 决定采用 *MRR* 指标的用意显然在于突出自信度计算在 QA 系统评价中的重要地位。到了 2002 年 TREC-11 的 QA 评测，测试问题总数增加到 500 个，但不保证每个问题都能在 TREC 语料库中找到答案。参评系统对每个问题只允许报送一个伴有自信度值的回答（对找不到答案的问题应回送 *NIL*），而且系统报送的全部 500 个回答都要按自信度值降序排列。组织方规定，参评系统的性能评价指标用“自信度权重分”*CWS*（Confidence Weighted Score）来计算[3]：

$$CWS = \frac{1}{N}\sum_{i=1}^{N} \text{前 } i \text{ 个回答中正确答案的次数 } / i$$

不难看出，*CWS* 同样强调了自信度值在问答系统整体评价中的权重。可是，依然有学者对此提出异议，认为 *CWS* 同样会把自信度稍低的回答排斥在问答系统的输出之外，不利于系统总体性能的最优化。

尽管 TREC 的 QA 评测活动对推动国际 QA 技术的进步产生过巨大影响，事实上大多数 QA 研究的文献都可以在历届 TREC 的会议论文集中找到。但据 Ferrucci 的论文披露[1]，在 DeepQA 项目起步的 2008 年，TREC QA 评测中性能在前 3 至 5 名的问答系统，如 IBM 的 PIQUANT 问答系统[4]曾先后参加过 6 届 TREC 的 QA 评测，在 Jeoparty 节目 2 万题题库中随机选出的 500 个问题上进行测试，平均正确率只达到 33%。可见，DeepQA 项目面临的挑战是十分严峻的。IBM 团队不仅需要大幅度提高问答系统的平均正确率，而且需要从根本上改进自信度计算的机制。PIQUANT 问答系统在 2003 年就已经形成多资源和多智能代理（multi-agent）的串行—并行混合式体系结构[5-6]。DeepQA 系统沿用了这种架构，在问题分析和回答抽取之间插入了假设生成、假设与证据的寻找、假设的合并与排序等部件。使得寻找回答的过程从一开始就形成并行的多个假设，并使用概率信息在问题、检索和回答等三个层面上合并假设。这种方法使得后续步骤有机会改正先前步骤产生的错误，而不像传统的串行体系那样造成错误的逐级累加。这里每个部件的输出结果都伴有自信度计算，并为此引入计算开销很高的推理和证明机制。实践证明这些考虑提高了回答的正确率和自信度精确率。这也是 DeepQA 留给我们的一条重要经验。

3　DeepQA 给我们的启示

3.1　DeepQA 的创新

电视节目 Jeopardy 的竞答对手是冠军级的人类选手，即使对于一流水平的 TREC 问答系统来说，也是莫大的挑战。敢于选择 Jeopardy 竞答作为自动问答研究的新目标就是一个极高的创意，而要在速度、平均正确率和自信度精确率等指标上完善 DeepQA 的各项设计，最终战胜 Jeopardy 的前冠军们，需要更多的创新。所以，笔者认为 DeepQA 项目的创新是它留给全世界同行最珍贵的启示。我们要在 NLP 和信息技术上赶超世界顶尖水平，一定要有这样的创新意识，敢于挑战过去，在自己的研究工作中不断树立新的标杆。苹果公司的奇才 Steve Jobs（1955-2011）把创新视为生命，他是全世界科技工作者的楷模。

3.2　OAQA 专题研讨会促成了 DeepQA 的立项和起步

2008 年初 IBM 举办了一个名为 OAQA（Open Advancement Question Answering）的专题研讨会[7]，出席会议的有 IBM 的 14 位研究人员和来自卡内基·梅隆大学、麻省理工大学、南加州大学、马萨诸塞大学、德克萨斯大学等院校的 7 位代表。会议讨论了自动问答系统评价指标的 8 个维度（dimensions）：平均正确率、自信度精确率、领域的宽泛度、问题的难度、查询语言的复杂度、内容语言的复杂度、速度（响应时间）、用户互动/可用性。代表们通过 TREC-QA，Jeopardy，LbR（Learning by Reading）等五种 QA 有关的研究应用领域来考察它们各自对 QA 性能所提出的要求。如果以 TRECQA 的性能要求为基准，那么 Jeoparty 竞答节目对平均正确率、自信度精确率、领域的宽泛度和速度等维度提出了更高的标准。这次研讨会为 DeepQA 的立项和起步铺平了道路，并在整个项目的研究过程中加强了 IBM 同高等院校之间的合作。此外，美国计算语言学会（ACL）、计算机学会（ACM）和人工智能学会（AAAI）在每年举办的国际会议上，通常都会同时组织一些类似的专题研讨会（workshops），以便同行对彼此感兴趣的问题开展深入的交流与切磋。这三个学会前几年在 QA 领域曾先后组织过 QA 研究的新方向、开放领域 QA、多语种 QA、文本

类型 QA 的推理技术等专题研讨会，对推动 QA 技术的进步产生过很大的影响。

三十而立，今年是中国中文信息学会成立 30 周年华诞，笔者期待着她能带领我国语言信息技术研究的同行大步走上创新、发展的高速路。然而，与 ACL，ACM 和 AAAI 相比，差距是明显的。比如，国内期刊和全国性学术会议的论文评审制度就有很多需要改进的地方；又如我们两年一届的计算语言学大会，似乎很少组织专题研讨会，今后也可以改进。在国内情况未能改观之前，国内各学科的研究团队应积极参加国际上相关的专题研讨会和附带评测项目的系列会议，以便通过交流更真切地了解国际学术前沿。

3.3 政府主管部门的规划和组织是 DeepQA 成功的保障

众所周知，自然语言问答系统集成了 30 年来 NLP 研究与应用的各种成果，包括词法分析、词性标注、浅层或深层句法分析、命名实体识别、指代消解、词义消解、文本检索、信息抽取（包括关系抽取）、机器学习、本体知识获取、知识挖掘、知识表示、逻辑推理等等。DeepQA 系统集成的上百种技术正是 30 年来 NLP 研究与应用的各种成果。上面提到的大多数关键技术在项目定义和评测活动方面都可以追溯到 NIST（National Institute of Standards and Technology）、DARPA（Defense Advanced Research Project Agency）和 ARDA（the US Department of Defense Advanced Research and Development Activity）等美国政府部门的直接领导。比如，TREC-QA 的评测始终是在 NIST 的指导和资助下开展的。MUC（Message Understanding Conference）定义了命名实体识别（NER）、指代消解、事件识别，ACE（Automatic Content Extraction）定义了自动文摘和信息抽取（包括关系抽取），它们背后也都有政府部门的直接领导。此外，ACL 下属的 SIGHAN（汉语处理专业委员会）自 2003 年起举办的 Bakeoff 系列评测活动，内容包括中文自动分词、中文词性标注、中文命名实体识别等；SIGNLL（自然语言学习专业委员会）自 1999 年起逐年举办 NLP 关键技术的评测活动，如词性标注、语块识别、子句句法分析、词义消解、依存句法分析、语义角色标注等。应当指出，以上许多专项评测是多语种的，如 2009 和 2010 年的依存句法分析和语义角色标注评测项目就包括英、德、中、日等七种不同语言。我国的参评选手车万翔和刘挺、赵海和揭春雨都曾分别获得七语种平均总分第一名。这在一定程度上说明，汉语和西方语言一样可以用几乎相同的机器学习方法和统计模型来实现句法和语义的分析。十年来 NLP 关键技术的长足进步和主管部门的直接领导以及上述种种评测活动是分不开的。

可是回顾国内的情况，国家科技部、国家自然科学基金委和国家 863 委员会等技术主管部门似乎只关注项目审批和经费分配，很少考虑关键技术的规划和评测。为此笔者建议，在这种情况下中国中文信息学会应率领下属的专业委员会积极向政府主管部门反映相关情况，争取他们采取行动尽快改变目前这种领导和一线科研团队脱节的状况。比如，可以说服自然科学基金委和 863 委员会每年拿出一部分经费来资助关键技术的项目定义和规划，包括资源建设和评测方案的设计。这类课题完成后，其成果和资源应在学术界的研究工作中共享。相信这样可以有效地减少重复性的资源建设和应用研究，有利于缩小我们同发达国家之间在科研项目规划管理上的差距。

4 结束语

今年是中国中文信息学会成立 30 周年，30 年来我们在汉字进入计算机、汉字激光照排系统等领域取得了里程碑式的辉煌成果，值得庆贺。但也应当清醒地看到，同印欧语言相比我们在汉语的句法-语义信息处理的研究应用方面，还存在一定的差距，值得引起同行的关注和讨论。上一节说了 IBM 深度问答系统的成功给我们的启示，当然是个人的粗浅认识，意在期望中国中文信息学会在新的十年中做出更大的作为。

董振东先生在今年 8 月我国计算语言学大会（洛阳）的邀请报告中也承认，我国的语言信息处理研究，同印欧语言相比还有相当的差距。不过他认为，这是因为汉语的句法不像印欧语言的句法那样有严格的形式（或形态）约束，汉语是一种意合语言等等。他建议，要改变目前的落后状况必须另辟蹊径：跳过句法，直接在句子的语义和理解层面上下功夫，以便反超西方现有的理论与应用成果，后来居上。董先生的建议很大胆，不过他在报告中没有进一步说明这一建议的实施细节，因此恕我不便评论。这里我愿意引用石定栩的一篇论文[8]，看看作者站在当代语言学理论的角度上是怎样来评论汉语意合

论的，供有兴趣的同行们参考。

其实关于寻找汉语特点的主张，早在上世纪 30 年代就出现了，上世纪 80 年代国内的部分语言学家更把这样的诉求推向极致，目标据说是要创立有中国特色的汉语语法理论。回过头来看，这类研究同当代语言学理论对语言共性和语言类型学研究的宗旨和方法论是背道而驰的，其结果当然也可想而知了。现在看来，我们在汉语动词的语义分类、句子中词语之间的依存关系、汉语的谓词—论元结构、中心语—附加语结构以及并列结构等基本语法关系上的研究与描写都落后于印欧语言，这在一定程度上与国内语言学和计算语言学研究的取向是关联的，因为恰恰是这些基本语法关系反映了不同语言之间的共性。

参考文献

[1] David Ferrucci, et al. Building Watson: an overview of the DeepQA project[J]. Artificial Intelligence Magazine2010, 31(3): 39-79.

[2] Ellen M. Voorhees. The TREC-8 question answering track report[C]//Proceedings of the Eighth Text Retrieval Conference (TREC-1999), 1999.

[3] Ellen M. Voorhees. Overview of the TREC 2002 question answering track[C]//Proceedings of the Eleventh Text Retrieval Conference (TREC), 2003.

[4] John Prager, et al. IBM's PIQUANT in TREC 2003 [C]//Proceedings of the Twelfth Text Retrieval Conference (TREC-2003), 2004.

[5] Jennifer Chu-Carroll, et al. A multi-strategy and multi-source approach to question answering[C]//Proceedings of AAAI Workshop on New Directions in QA, 2004.

[6] Jennifer Chu-Carroll, et al. In question answering, two heads are better than one[C]//Proceedings of HLT/NAACL, 2003.

[7] David Ferrucci, et al. Towards the open advancement of question answering systems[R]. IBM Research Report, RC24789 (W0904-093), April 22, 2009.

[8] 石定栩. 汉语句法的灵活性和句法理论[J]. 当代语言学，2000，(1)：18-26.