

沃森机器人给我们的启示

潘 燕¹,程显毅²

(1 南通大学 电气工程学院, 江苏 南通 226019; 2 南通大学 计算机科学与技术学院, 江苏 南通 226019)

摘 要: 就人机大战超级电脑“沃森”(watson)的胜出,讨论了沃森对人工智能研究给我们的启示.和银河相比,沃森的计算能力还较弱.沃森可能进入不了超级计算机的计算能力前 100 名,但是沃森能把一件事情(智力问答)干到最好.沃森的强项也不在于深蓝系统的国际象棋,但日常知识类的问题的挑战不在国际象棋之下,甚至有过之而无不及.这正是人工智能一个新的里程碑.

关键词: watson; 人工智能; 自然语言理解; 问答系统; 机器人

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)10-0060-03

Inspiration of Watson Robot

PAN Yan¹, CHENG Xian-yi²

(1 College of Electrical Engineering, Nantong University, Nantong 226019, China;

2 College of Computer Science, Nantong University, Nantong 226019, China)

Abstract: In this paper, super computer Watson's win, in man-machine war, we discusses the nspiration of researching artificial intelligence. And compared to Galaxy computer, Watson's computing power is still relatively weak. Watson may not enter the super-computing power top 100, But one thing can Waston(quiz) to the best dry. Watson's strength nor the Deep Blue chess system, But the challenges of common sense is not worse under chess, or even better. This is a new milestone in artificial intelligence. .

Key words: Watson; artificial intelligence; natural language understanding; Q & A system; robot

1 引言

沃森是一个能够与人类回答问题的能力相匹敌的超级计算机系统,具有足够的速度、精确度和可信度.沃森的技术可以理解自然语言的提问,分析数以百万计的信息碎片,并且根据寻找到的证据提供最佳答案.

2 沃森战胜人类获冠军

2.1 沃森超级电脑概况

沃森是一款由国际商业机器公司(IBM)开发,以公司创始人托马斯·沃森命名的人工智能软件系统.沃森评估了大约 2 亿页的内容,使用人类自然语言编写的沃森采用 10 支架、运行 Linux 系统、IBM

POWER 750 服务器,使用 15 万亿字节的内存,2880 处理器核心,并且操作能力为每秒钟 80 万亿浮点^[1].

要准确地回答问题沃森的首要任务是理解问题.精通语言学研究人员通过程序算法让沃森能够从题目中找到其中的关键字,给出很多可能的答案,然后把答案按照可信度进行排序,决定是否参加抢答.这些复杂的过程,沃森运行起来只需要 3 s.它每秒可以进行 80 万亿次运算.当它认定答案可信时,这台身躯笨重的机器只需要 10 ms 就可以按下抢答器.

为了让沃森能够学会人类的知识,在沃森的大脑中放入的信息相当于一本超过 2 亿页的文档,其中包含《全球百科全书》这样的正式出版物,也包含

收稿日期: 2011-04-20; 修回日期: 2011-06-25

基金项目: 国家自然科学基金资助项目(60873069)

由网民共同维护更新的“维基百科”。

2.2 如何与人智力竞答

在比赛过程中,当主持人阅读线索时,它们会以电子内容的形式传输给沃森,然后这台超级电脑会对接收到的信息进行分析,并做出各种构想,接着搜索所有信息检验构想的正误,得出5个最佳答案,并对每个答案设定信心级别,最终决定选用哪个答案。

在Jeopardy!中,沃森在北京时间2月17日上午战胜了该节目有史以来最优秀的两位人类冠军Ken和Brad,圆满结束了这场历时三天的人机大战。三轮比赛中,沃森在第一、二轮几乎抢答近90%题目,第三轮抢答逾60%,答对率逾95%。

2.3 如何决策是否继续答题

“沃森”会估计《危险边缘》节目组剩下的奖金数额、自己比其他对手落后或领先多少、自己在特定类别的题目上的表现,及时调整自己的信息级别。如果“沃森”落后对手太多,它的信心级别会较低,例如40%,这时它仍会继续努力;不过在领先对手时它也会想:“我已经领先那么多了,为什么还要冒险呢?”因此,即使它的信心值是75%,它也有可能选择不再回答。

《危险边缘》是哥伦比亚广播公司一档长盛不衰的电视问答节目,沃森之所以能够应付这种狡猾的试题,主要依靠的是它对自然语言的理解和高速的计算。

2.4 回答问题的内容来源是什么

“沃森”会运用百科全书、字典、新闻报道、书籍、网络内容等资源。但在参加《危险边缘》节目时它不能与互联网连接。

3 沃森的自然语言理解

所谓自然语言,是人类借以思维和互相交际的一个音义结合的符号系统,是一个变动的音义结合的系统结构^[2]。

自然语言理解就是研究如何能让计算机理解并生成人们日常所使用的语言,目的在于建立起一种人与机器之间的密切而友好的关系,使之能进行高度的信息传递与认知活动。自然语言理解的研究从1950年代开始到1970年代,可以说基本上停留在实验和纯理论的探讨阶段。到了1980年代,由于计算机硬件技术飞速发展和自然语言理解的理论水平的提高,自然语言理解的应用研究广泛展开^[3]。

3.1 自然语言的特点

相对于规范语言,自然语言的理解具有更大的

难度。这是因为自然语言有如下特点^[4]。

首先,自然语言是极其复杂的符号系统。符号的形式和意义的结合是由社会“约定俗成”,人类用有限的语言符号构建自然语言这一结构,但又不失系统性。自然语言的系统性表现在它是一个分层装置,底层是一套音位,上层是音义结合的符号和符号的序列,通过结构关系中的组合和替换运转分层装置。因此,自然语言是在诸多符号构成的系统中有序地完成人类交际的过程。

其次,自然语言的各个层次都有巨大的不确定性。在语音和文字层次上有一音多字和一字多音的问题存在;在语义和语用层次上也有由种种原因造成的内涵、外延、指代、言外之意的不确定性,语言学上称其为“歧义”。歧义不能通过造成歧义的语言单位获得解决,而必须借助更大的语言单位乃至环境背景常识来解决。

再次,自然语言不是一成不变的语言。自然语言诞生于人类社会,在社会中不断发展,在不同语言和同一语言的不同变体的人们相互影响中变化。现今计算机技术日益发展,面对人类独有的交际工具——自然语言,计算机将如何理解并运用它更好地实现与人类的沟通。这必然涉及自然语言理解问题。

3.2 自然语言理解

自然语言理解的本质就是知识工程,即挖掘出人在理解语言时所运用的知识,然后“教会”计算机使用这些知识,也就是说,它包括两个方面:对人类知识的再认识和知识表达^[5]。

从计算机处理的角度,对自然语言的研究和处理,一般应经过如下三个方面的过程^[3]:(1)把需要研究的问题在语言学上加以形式化,使之能以一定的数学形式,严密而规整地表示出来;(2)把这种严密而规整的数学形式表示为算法,使之在计算上形式化;(3)根据算法编写计算机程序,使之在计算机上加以实现。

语言的分析和理解过程应当是一个层次化的过程,它包括:词汇分析和句法分析(语法层),语义分析(语义层),基于语境和世界知识的篇章分析和自然语言生成(语用层)。虽然这种层次之间并非是完全隔离的,但是这种层次划分的确有助于更好的体现语言本身的构成^[3]。自然语言理解的基本模型如图1所示。

目前,在人工智能界尚无统一的机器自然语言理解的定性准则和量化标准。一般认为,根据著名的

表 3 PKU、NCC 和 CTB 语料上的训练过程记录数据

模板集 序号	PKU 语料训练数据			NCC 语料训练数据			CTB 语料训练数据		
	特征数	训练时间/ s	模型大小/ MB	特征数	训练时间/ s	模型大小/ MB	特征数	训练时间/ s	模型大小/ MB
1	145 224 232	无法训练		74 761 652	无法训练		52 577 888	无法训练	
2	17 614 133	186 178.49	69.7	10 179 464	94 819.77	43.8	6 606 572	51 617.17	30.0
3	17 603 524	152 027.86	69.7	10 176 100	74 384.22	43.8	6 605 203	39 865.86	30.0
4	91 420 534	无法训练		47 666 314	无法训练		33 512 676	无法训练	
5	10 997 310	148 211.24	44.3	6 555 566	84 061.73	28.2	4 228 841	47 858.08	19.2
6	10 986 701	113 086.12	44.3	6 552 202	54 584.97	28.2	4 227 472	32 182.33	19.2

表 4 不同特征模板集的词性标注结果

模板集序号	特征模板集名称	PKU 语料上评测结果	NCC 语料上评测结果	CTB 语料上评测结果
		Accuracy/%	Accuracy/%	Accuracy/%
2	T10-Single+B	94.12	90.05	91.94
3	T10-Single	93.90	90.02	91.71
5	T6-Single+B	94.78	90.48	92.07
6	T6-Single	94.52	90.53	91.65

4 结束语

汉语词性标注中应用条件随机场对上下文建模时会扩展出数以亿计的上下文特征,如此多的特征使得模型训练的时间过长甚至无法进行.针对这一问题,本文通过深入分析上下文特征产生的内在机理,对采用 CRF++ 工具包实现的汉语词性标注所使用的特征模板集进行了优化,对比实验表明,优化后的特征模板集在模型训练时间、训练后模型大小、词性标注精度等指标上达到了整体最优.

参考文献:

[1] 姜维,王晓龙,关毅,等. 基于多知识源的中文词法分析系统[J]. 计算机学报, 2007,30(1):137-145.
[2] 洪铭材,张阔,唐杰,等. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学,2006,33(10):148-155.
[3] Pereira L J, Mccallum F A. Conditional random fields:

probabilistic models for segmenting and labeling sequence data [C]// Proc. of 18th ICML. San Francisco, USA: AAAI Press, 2001: 282-289.
[4] 于江德,睢丹,樊孝忠. 基于字的词位标注汉语分词[J]. 山东大学学报:工学版,2010, 40(5):117-122.
[5] 冯冲,陈肇雄,黄河燕,等. 基于条件随机域的复杂最长名词短语识别[J]. 小型微型计算机系统,2006,27(6): 1134-1139.
[6] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006,34(5): 804-809.
[7] 于丽丽,德鑫,曲维光,等. 基于条件随机场的古汉语词义消歧研究[J]. 微电子学与计算机,2009,26(10):45-48.

作者简介:

于江德 男,(1971-),博士,副教授.研究方向为计算语言学、中文信息处理、文本信息抽取等.

(上接第 62 页)

[2] 程显毅,朱倩,王进. 中文信息抽取原理与应用[M]. 北京:科学技术出版社,2010.
[3] 刘小东. 自然语言理解综述[J]. 统计与信息论坛,2007, 22(2):5-12.
[4] 任洁. 自然语言与自然语言理解及其应用[J]. 科教文汇,2006:69-70.
[5] 高文利,李德华. 对自然语言理解的思考[J]. 牡丹江大学学报,2007,16(5):75-76.

[6] 程显毅,朱倩. 文本挖掘原理[M]. 北京:科学技术出版社,2010.

作者简介:

潘燕 女,(1988-),硕士研究生.研究方向为自然语言处理.
程显毅 男,(1956-),博士,教授.研究方向为人工智能.