文章编号: 1003-0077(2007)03-0003-05

编者按: Internet 时代对中文信息处理提出了更多、更新的需求,同时,致力于中文信息处理研究的队伍也在不断地壮大。在这支队伍中,既有在这个领域里长期辛勤耕耘的老兵,也有初出茅庐的新人。为了使研究者们得以在更高的起点上开展研究,我们特向该领域(或相关领域)的资深专家和学者约稿,这些稿件或是多年研究成果的厚实积累以及发轫于斯的深刻思考,或是具有前瞻性的前沿课题探索,或是相关研究工作系统而深入的综述。我们设立了一个约稿专栏,陆续刊登此类稿件,以飨读者。本期刊登其中的2篇,分别是张钹院士的"自然语言处理的计算模型"、黄昌宁教授等的"中文分词十年回顾"。相信这些论文对读者全面、深刻地了解乃至理解相关学术问题,一定会大有裨益。

自然语言处理的计算模型

张 钹

(清华大学 计算机系,北京 100084)

摘 要: 本文讨论自然语言处理的计算模型。目前已经存在有各种类型的语言计算模型,如分析模型、概率统计模型、混合模型等,这些模型各具特色,并存在其自身的局限性。自然语言处理作为一个不适定问题,我们将讨论求解这类问题的本质困难,面临的挑战,以及解决这些困难的途径。

关键词: 人工智能;自然语言处理;计算模型;分析模型;概念统计模型;混合模型;不适定问题
中图分类号: TP391 文献标识码: A

The Computational Models of Natural Language Processing

ZHANG Bo
(Department of Computer Science & Technology
Tsinghua University, Beijing 100084, China)

Abstract: In this paper, we will discuss the computational models of natural language processing. There have been several kinds of computational models such as analytical model, statistical model, hybrid model, etc; each has its own characteristics and limitations. As an ill-posed problem, we 'll discuss what the essential hardness the natural language processing has, what challenge we will confront with, and what measures we 'll adopted to solve the difficulty.

Key words: artificial intelligence; natural language processing; computational model; analytical model; statistical model; hybrid model; ill-posed problem

1 引言

本文讨论的"自然语言处理"都是指利用电子计算机对自然语言的各级语言单位进行的自动处理,包括对字、词、句、篇章等进行转换、分析与理解等等[1]。与电子计算机的发展历史相比,自然语言处理算是一门很"老"的学科了。电子计算机刚刚问世,计算机科学家就对语言的机器处理备感兴趣,不久语言学、心理学、认知科学、人工智能等不同领域

收稿日期: 2007-03-01 定稿日期: 2007-03-01
基金项目: 国家自然科学基金资助项目(60621062);国家973资助项目(2003CB317007,2004CB318108)
作者简介: 张钹(1935—),男,中国科学院院士,主要研究方向为人工智能。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的学者也纷纷参入他们的研究队伍,一门新的研究领域——自然语言处理从此诞生。翻开它的历史,人们会发现,自然语言处理的发展道路并不平坦,研究工作跌宕起伏,时而乐观,时而悲观。人们对自然语言自动处理的困难通常估计不足,对它发展的前景往往过于乐观。可是,实践却一再表明事实并非如此,研究工作总是困难重重,进展缓慢,于是引来了悲观情绪。奇怪的是,这种乐观与悲观情绪的交替、循环在半个多世纪自然语言处理的发展历史上却不断地重演着。

早在二次世界大战期间,现代电子计算机还处于襁褓之中,利用计算机来处理自然语言的想法就已经出现了。当时人们从破译军事密码的工作中得到启示,以为不同的语言(中文,英文,还有其他语种)只不过是对"同一语义"的不同编码而已。于是想当然地认为,采用译码技术"破译"(理解)这些"码"(语言)应该不成问题。结果却大大出乎人们的意料,自然语言自动处理居然比破译密电码困难得多!

1956 年人工智能诞生之时,该领域的创始人就把计算机国际象棋(Computer Chess)和机器翻译(Machine Translation)作为两个标志性的任务提出来,认为只要计算机的象棋程序打败国际象棋世界冠军,机器翻译程序达到人类翻译的水平,就可以宣告人工智能的胜利。他们对此充满信心,以为凭借计算机的计算能力,将会在很短的时间里达到预定的目标。如认知心理学家 H. Simon 认为十年内这两项目标都可以实现。大家知道,实际上,直到 1997 年,即 40 年(不是 10 年)以后,IBM 的国际象棋程序——深蓝(Deep Blue)打败国际象棋世界冠军卡斯帕罗夫,才宣告第一项任务的胜利完成。而机器翻译呢?至今依然是一项十分困难的任务!这些过分乐观的估计至今一直成为人们质疑人工智能的一个口实和笑柄。人们一再低估自然语言处理的困难。

然而,跌宕起伏的历史也正是自然语言处理研究工作的魅力所在,它吸引着千千万万的研究者去研究自然语言的复杂性,探索其中的原因,寻求机器自动处理的方法。至今大多数研究者主要从语言本身的复杂性来探讨这些问题,找到了其中的许多原因[2~7],其中包括:存在于各级语言单位(字、词、句、篇章等)的局部歧义性(Local Ambiguity),上下文的影响(Contextual Dependency),语法与语义的相互依赖关系,语言环境,知识背景等等。毫无疑问,语言处理的复杂性来源于语言本身的复杂性,因此上述研究成果对于进一步理解自然语言的特点,以及改进机器处理的性能,都起过很好的作用。不过,在自然语言自动处理过程中,计算机处理的直接对象并不是实际的自然语言,而是它的计算模型,因此要真正理解自然语言自动处理的问题,并找出解决的办法,还需要从语言处理建模的角度来探讨这些问题,可惜目前这方面的探讨还不多,本文将着重讨论它。

2 不适定问题

现实的自然语言系统 N(Natural Language)十分复杂,不可能作为计算机的直接处理对象。为了使它成为可处理的对象,首先需要根据处理的要求,把它抽象为一个问题 P(Problem),比如 P 是自然语言 N 中的分词问题。然后根据给定的输入、输出集(I,O),以及问题 P,建立一个数学模型 M(Model),以及与其相关的有效算法 A(Algorithm)。M 与 A 组成了问题 P 的计算模型 F(Computational Model)。显然,同一个模型 M,可以采用不同的算法,因此计算模型 F 取决于采用的数学模型 M,M 是模型的本质,而算法只是实现的手段。有了计算模型 F,在给定的输入集(I)下,就可以计算出输出 O,因此 O 也可称为 F 的解。或者说,通过计算模型 F,我们对自然语言中的 P 问题进行处理(如图 1 所示)。因此研究自然语言自动处理的关键是研究计算模型 F。

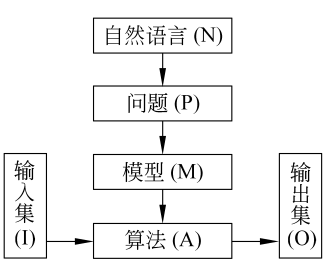

图 1 自然语言处理建模

给定计算模型 $F(I,O)$,其中 I 是输入集,即一组数据,O 是输出集,通常由语义空间的元素组成。以汉语分词为例,输入一个句子"南京市长江大桥",对于计算机来讲,这个句子只不过是由"0"和"1"组成的机器码,即一组数据。我们要求的输出是:按照语义切分出句子中的词。因此模型 F 的作用就是按语义对数据 I 进行分类,分类的结果就是输出 O。可以说,F 是数据空间 I 到语义空间 O 的映射

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(Mapping),即映射 $F: I \rightarrow O$。一切自然语言的自动处理问题 P,都可以抽象为这样一个映射问题。于是我们把所有的自然语言处理(分词、词性标注、词法分析、语言理解等等)归结为一个普适的科学问题 - 映射问题 F,从而便于进行理论研究。

不幸的是,由自然语言处理抽象出来的计算模型-映射 $F: I \rightarrow O$,由于自然语言的复杂性,往往是一个不适定(Ill-posed)的逆问题,这就使得自然语言处理问题的求解十分困难。本来,对于一个问题解的存在性,唯一性,以及稳定性中任何一条不满足,就算是难解的不适定问题,而自然语言处理的计算模型往往这三个条件都不满足,因此是一个强不适定问题(Strongly Ill-posed Problems)。仍以汉语"南京市长江大桥"的分词为例,首先它的解不唯一,至少有两种可能的分词结果:"南京市| 长江| 大桥","南京| 市长| 江大桥"。解的存在性和不稳定性也十分明显。如果改动上面句中的某个字,比如,将"京"字改为"景"字,根据"分词"的定义:"依语义(词义)对数据进行切分",因为无论是"南景"还是"南景市"这两个词都不存在,因此无法从词义上对该句子进行切分,问题也就变成无解的了。目前已有许多关于不适定问题求解理论与方法的研究成果[8,9],自然语言处理完全可以借鉴这些理论来探索新的解决方案。不适定问题的求解方法[8],简单地讲,就是加入适当的约束(Constraint)条件,使问题的一部分变成适定的(Well-Posed)。约束条件可以加到输入集、输出集、模型本身等。比如,著名的求解不适定问题的正则化方法(Regularization),就是对输出集(解集)进行约束,把它限制在具有稳定解的范围之内,从而使问题在这个范围内变成适定的。本文将从这个角度研究自然语言处理的相关计算模型,探索一条新的研究途径。

3　分析模型

语言学家 N. Chomsky 认为人类生成合乎文法的语句的能力是生来具有的,为此他提出一种称为生成句法(Generative Grammar)的理论[10],这个理论对人类语句的生成做了如下的解释,即人们通过一组有限的规则作用于一个有限的词汇上,从而本能地生成无限的可接受的、合乎文法的句子(Acceptable Grammatical Sentences)。这个理论的提出马上得到语言学界的广泛兴趣,并对自然语言自动处理产生深刻的影响。这个理论表明在自然语言的各级语言单位中都存在一定的内在规律性,因此依据这种规律性,就可以为语言处理建立一种计算模型,比如基于规则(Rule-Based)的模型。由此可见,一切理性分析的语言计算模型(Analytical Model)都是建立在这种理论假设之上。

如果对输入集加以适当的限制,比如假定有限的输入集,理性分析模型一般可以满足适定性的条件,因此这种模型对于解决较小规模的自然语言处理问题具有一定的效果。可是,由于语言的输入集(I)是无限的,这种通过有限规则集,特别是少量规则集的建模方法,显然很难满足自然语言处理的全部需要。因此语言的理性分析模型面对大规模的真实文本时,都难以通过"可扩展性"(Scalability)的考验。因为当问题的规模扩大之后,理性分析模型在大型的输入集上,难以使问题的全部解达到适定性的要求。这也就是理性分析模型的局限性所在。

4　概率统计模型

Chomsky 关于语言获取(Language Acquisition)的理论也受到一部分学者的质疑,他们认为人类自然语言与人造的形式语言不同,并不遵循严格的规律,因此语言理性主义的分析方法难以克服语言复杂性带来的困难。与 Chomsky 理论相反,行为心理学家 B. F. Skinner[11] 提出另一种语言理论。这个理论认为人类语言能力的获得来自于学习,语言是通过不断地实践而"约定俗成"的结果。这就是自然语言形成的经验主义解释。概率统计模型(Statistical Model)[12]属于经验主义的语言计算模型。概率统计建模采用从数据中学习(Learning From Data)的方法,至今取得很大的成功,目前已成为自然语言处理中占统治地位的建模方法。概率模型的成功应该归功于网络时代信息的数字化和网络化,正因为这些变化,为我们带来了取之不尽、用之不竭的数据。"数据驱动"(Data Driven)法应运而生,正是这种新的研究方法促成了当今以概率建模为代表的经验主义方法的繁荣与发展。比如,目前流行的基于语料库(Corpus-Based)的语言处理方法就是一种典型的数据驱动方法。

但概率统计建模也不是无懈可击的,面对大规模的真实文本,它面临着许多挑战。首先,语言的计算模型 $F: I \rightarrow O$ 是不连续映射,根据统计学习理论[13],不难知道,通过学习与训练获取不连续映射的困难很大,通常存在学习不收敛、学习误差大、推

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

广能力弱等诸多问题。因此基于概率模型的大规模文本处理的结果通常准确度受到一定的限制。其次,从建模的角度看,由于自然语言的层次结构,在各个层次的语言单位之间存在着大量的依存关系,特别是远距离的依存关系(Long-Distance Dependency),如上下文关系等。如果建模时,把这些可能的关系都考虑进去,模型将会变得极其复杂而无法处理。但是语言计算模型的解通常是不稳定的,任何一个未加考虑的微弱因素(例如,长距依存关系,以及其他小概率事件等)都可能引起解的巨大变化,从而带来严重的错误,因此许多场合下,不能忽略微弱参数的影响,这就使概率建模方法陷入两难的境地。最后,虽然网络上的文本数据(生语料)几乎是无限的,但带有正确层次结构标注的数据依然匮乏,统计模型仍然面临严重的"数据稀疏"问题。因此单纯的概率模型也不能完全解决自然语言处理的自动化问题。

5 混合模型

以上讨论使我们认识到,无论理性的分析模型,还是经验的概率模型都不能解决语言自动处理的全部,特别是大规模的真实文本。其原因还需要从自然语言本身的特点去寻找,人类不仅利用自然语言表意,同时也用它来言情,一段语言中往往既有理性的思考,又有感情的流露,意中有情,情中有意,情景交融。因此自然语言处理既需要理性分析,也需要感性经验,二者互相补充。就是说,需要走理性主义与经验主义结合的道路,即混合模型(Hybrid Model)的道路。目前已有许多研究工作试探混合模型的方法,已经取得一些成果[14~17]。但困难依然存在,比如,感性经验的表达与运用就是其中关键之一,也就是说,如何考虑语感、语境和知识背景等问题。

在机器翻译研究的初期(上个世纪60年代),美国人经常举以下的例子来说明机器翻译任务的艰巨性。

英文的原句是:

(1) The spirit is willing but the flesh is weak.

(心有余而力不足)

经机器翻译成俄文之后(在文法分析、双语词典等支持下),再把它翻译回英文,得到的结果如下:

(2) The Voltka is strong but the meat is rotten.

(伏特加酒是浓的,但肉却腐烂了)

这也许只是一则笑话,可是它充分说明自然语言处理的困难所在。显然,机器将句子(1)的意思翻译错了。但不幸的是,我们从中竟然找不出错在何处。因为(1)与(2)两个句字的语法完全一样,可见机器并没有犯任何语法错误。从语义层面看,"spirit"(精神,烈性酒)译成"Voltka"(伏特加酒)并无错误,同样,"flesh"(肉体,肉)译成"meat"(肉)等等也并没有犯语义上的错误。如果错在何处不容易找到,能否找出错误来自何处?的确,词的多义性是错误的始作俑者。可是问题并没有因此解决,进一步的问题是,如何消解这些歧义,找到正确的答案?对此我们似乎无计可施。因为任何的理性分析都难以纠正上述错误,唯一有效的解决办法,似乎只能直接"告诉"机器,它就是"心有余而力不足",换句话说,这是约定俗成,没有什么理由可讲。说明这里需要的是感性体验,而非理性分析。其实,当我们把一个文件输进计算机,文件里描绘的如果是一幅乡间的景色,讲述的是一段男女的情感故事,机器如何"看懂"它,如何对它进行处理?显然,要解决这类问题,机器除具备理性的分析能力之外,更重要还要有丰富的感性经验与知识。

机器是否可以具有感性经验,又如何得到这种体验?这是人工智能研究的重要课题,至今已经取得一些成果。以计算机国际象棋程序为例,其实,从理性分析的角度看,计算机分析棋局的能力早已超过人类,但是长期以来计算机象棋程序一直无法打败人类象棋大师。其中主要的原因是,人类具有"棋感"和下棋的经验,而计算机没有。IBM的象棋程序所以能够最后战胜人类高手,是因为同时在以上两个方面下了功夫,采取了相关的措施。一方面,通过各种渠道,提高机器的计算速度,使它在下棋过程中,可以往前预测10—15步,而象棋大师一般只能预测3—5步,机器的分析能力远超过人类棋手。另一方面,为了弥补机器在"棋感"与下棋经验方面的不足,在IBM机器中存储了大量的下棋经验与知识,包括60多万种的棋谱(以往的下棋经验),棋局的评价标准(启发式的决策经验)等。

换句话讲,需要依靠理性分析与感性经验的密切结合,但此项研究工作才刚刚开始,至今依然远未解决。

6 结论

一台电子计算机不管性能多么的高,本质上,都

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

只是会计算“0”和“1”的机器。从计算的角度看,自然语言处理是一个强不适定问题,因此简单的建模方法,无论是确定性的,还是不确定性的都无法解决其全部。根据不适定问题的求解原理,只有通过提供大量的“约束”(包括知识,经验等),才能使之成为适定性的、可解的问题。因此出路是,通过计算机科学、语言学、心理学、认知科学和人工智能等多学科的通力合作,将人类认知的威力与计算机的计算能力结合起来,才可能提供丰富的“约束”,从而解决自然语言处理的难题。

参考文献:

[1] 王晓龙,关毅,等. 计算机自然语言处理[M]. 北京: 清华大学出版社,2005.

[2] Gibson, E., Linguistic complexity: Locality of syntactic dependencies [J]. Cognition, 1998, 68: 1-76.

[3] Daniel Grodner, Edward Gibson and Duane Watson. The influence of contextual contrast on syntactic processing: evidence for strong-interaction in sentence comprehension [J]. Cognition 2005, 95: 275-296.

[4] Silvia Gennari and David Poeppel. Processing correlates of lexical semantic complexity [J]. Cognition 2003, 89: B27-B41.

[5] Tessa Warren and Edward Gibson. The influence of referential processing on sentence complexity [J]. Cognition 2002, 85: 79-112.

[6] Gerry Altmann, Mark Steedman. Interaction with context during human sentence processing [J]. Cognition 1988, 30: 191-238.

[7] Douglas Roland, Jeffrey L. Elman and Victor S. Ferreira, Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences [J]. Cognition 2006, 98: 245-272.

[8] Tikhonv, A. N., Arsenin, V. Y.. Solution of III—posed problems [M]. New York: Winston/Wiley 1977.

[9] Bakushinsky, A., Goncharsky, A.. Ill-posed problems: Theory and Applications [M]. Dordrecht/Boston/London: Kluwer Academic Publishers, 1994.

[10] Chomsky, N.. Syntactic structures [M]. The Hague: Mouton, 1957.

[11] Skinner, B. F., Verbal Learning [M]. New York: Appleton-Century-Crofts, 1957.

[12] Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing [M]. Cambridge, Massachusetts: The MIT Press 1999.

[13] Vladamir N. Vapnik, Statistical Learning Theory [M]. New York: John Wiley & Sons, Inc., 1998.

[14] Aue, Anthony, Arul Menezes, Robert Moore, et al. Statistical Machine Translation Using Labeled Semantic Dependency Graphs [A]. In: Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation [C]. Baltimore, 2004.

[15] Pinkham, J, and M. Corston-Oliver, Adding Domain Specificity to an MT System [A]. In: Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics [C]. Toulouse, France, 2001, 103-110.

[16] Richardson, S., W. Dolan, A. Menezes et al. Achieving Commercial-quality translation with example-based methods [A]. In: Proceedings of MT Summit VIII [C]. Santiago De Compostela, Spain, 2001, 293-298.

[17] Wu Andi. Statistically-Enhanced New Word Identification in a Rule-Based Chinese System. In: Proceedings of the Second Chinese Language Processing Workshop [C]. Hong Kong, China, 2000, 46-51.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net