

面向自然语言处理的人工智能框架

蔡艳婧¹, 程显毅², 潘 燕²

(1 南通商贸高等职业学校, 江苏 南通 226000; 2 南通大学 计算机科学与技术学院, 江苏 南通 226019)

摘 要: 在自然语言的处理研究中, 一般对语义的处理只是在某范围内选定若干低级的原语, 用这些原语的结构组织来解决高层次词条的语义, 并且这些所需要原语的语义必须清楚, 这样可以解决一定范围的语义问题. 实际仅仅用一定范围的语法结构代替语义描述, 并没有真正解决语义的表达问题. 随着人工智能的研究取得一定进展, 计算机已从单纯的数值计算进入到知识处理阶段. Agent 理论有助于从语言的本质使用的角度来处理自然语言处理中的人工智能框架, 为语义分析提供哲学的视域和分析方法.

关键词: 人工智能; 自然语言处理; Agent; 语义

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)10-0173-04

A Framework of Artificial Intelligence Oriented Natural Language Processing

CAI Yan-jing¹, CHENG Xian-yi², PAN Yan²

(1 Nantong Business Vocational School, Nantong 226000, China;

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

Abstract: In the study of natural language processing, in general, semantic processing is within a certain number of low-level primitives selected, with the structural organization of these primitives to solve the high-level semantics of term, and these semantic primitives need to be clear, this would solve a range of semantic issues. Actually, it is only a certain range instead of the grammatical structure of semantic, does not really solve the problem of semantic expression. With some progress made in the research of artificial intelligence, computer from a simple numerical calculation into knowledge processing stage. Agent theory helps to use from the perspective of the nature of language to deal with natural language processing in artificial intelligence framework, for the semantic analysis provides the horizon of philosophy and methods of analysis.

Key words: artificial intelligence; natural language processing; agent; semantic

1 引言

人工智能框架问题研究一直存在三种理论假设: 一种是物理符号假设(认知主义, 规则推理); 一种是仿生学假设(联结主义, 神经网络); 一种是控制进化假设(行为主义, 环境交互, Agent)^[1]. 人工智能框架在自然语言处理中分别对应知识工程方法、机器学习方法和智能体方法.

前两种方法在实际应用中, 显示了不同方面的局限性. 知识工程方法(也称基于规则语言模型), 它

几乎不可能表达自然语言处理所需的各种知识, 这是因为各种知识不仅在数量上浩瀚无际, 而且还存在着知识获取代价高等问题. 为了弥补这方面的局限性, 机器学习方法(也称基于统计语言模型)被提出, 试图从大规模真实文本的语料库中获取语言知识, 以求得对自然语言的准确认识, 但处理大规模真实文本的目标与处理受限语言的目标大相径庭, 目标的不同必然导致意向观念的不一致而产生歧义. 另一方面, 从语料库中采集、整理、表示和应用知识的统计模型并不十分完善, 研究工作仍然比较困

难. 因此, 尽管基于大规模语料库的方法为自然语言处理领域带来了不少成果, 但这些方法都是统计学中的方法和一些其他的简单方法与技巧. 在实际应用中, 研究者似乎已将它们的潜能发挥到了极致. 那么, 自然语言处理所面临的问题是: 要取得新的重大进展是提出新的方法, 还是在已有方法的基础上进行改良优化或综合. 人工智能的物理符号假设框架在哲学意义上具有脱离经验的根本局限性, 联结主义研究框架实质上与此相同^[2]. 面向自然语言处理的人工智能的框架需要基于自然机制为内核、算法机制为辅的哲学纲领. 现代自然语言处理的句法和语义研究是脱离使用者与语境知识而处于指称层面, 只有增加基于语用层面的处理才是完整的. 本文基于 Agent 理论, 提出了面向自然语言处理的人工智能框架.

2 计算机处理自然语言的困难及对策

在处理语言的时候, 需要一种关于常识推理的理论, 即人工智能框架. 大脑与计算机最大的不同是: 计算机不能进行常识推理或有限的常识推理, 带来的问题是对接收到的语句不能进行信息的“补充”和“校正”.

(1) 自然语言的陈述中, 常常有很多词语. 这些词语在形式上不出现, 然而, 实际上是存在的. 人们能在生活常识的基础上通过联想非常自然地补充上去.

(2) 自然语言的陈述中, 也常常有很多不合语法、不合常理的地方, 听者在解读这些陈述时会在生活常识的基础上自然地加以校正.

例 1: 那天在商店里, 看见一盆花, 漂亮极了, 但是价格很高, 买回去恐怕要挨骂.

任何人听了都会在生活常识的基础上迅速地通过“联想、推理(判断)与选择”理解为:

那天(我)在商店里, (我)看见一盆花, (这盆花)漂亮极了, 但是(这盆花的)价格很高, (如果我把它)买回去恐怕(我)要挨(家里人的)骂^[3].

在实际应用中需要计算机能作出这样合乎情理的补充.

例 2: “我认为这种食品是很环保的!”把名词性的“环境保护”放在“很”后面是很不“规范”的, 但是我们会正确地把它理解为“我认为这种食品很符合环境保护的要求”

人们把这种“缺胳膊少腿”, “转弯抹角”称为“刁钻”句. 如何避免“刁钻”有两种策略:

(1) 提高计算机的智能框架水平(把大脑的联想、推理(判断)与选择过程)也纳入语言处理模型中去.

计算机解决问题的前提是“建立模型”(或“形式化”), 目的是对所研究的事物进行抽象, 去掉“无关紧要”的细节, 保留对研究目的来说有重要意义的方面. 如果所建立的模型“简繁得当”, 它的行为就会和原来的研究对象基本一样. 换句话说, 虽然对客观事物的考察或者描述无论在广度或者深度上都是可以无穷尽地拓展下去的, 但是在“模型”中我们可以只收入对我们的研究目标有意义的那些东西.

“自然语言”这一研究对象和天文学的天体、电子学中的器件和网络、生物学中的动植物……这些对象有些不同, 因为后者是完全独立于我们主观世界以外的客体, 而“自然语言”则和我们的主观世界有密切关系, 最主要的差异有两点:

①自然语言中的各种成分: 字词、短语、语块、句子、句群、段落、篇章……只是大脑使用的一些“符号”, 并不是像“地球、氧气、昆虫……”那样, 是具有稳定不变的属性、独立于我们主观世界以外的物质实体. 因此, 把自然科学研究中的思想方法一成不变地用来研究自然语言中的各种成分就会出现困扰.

②自然语言中语句的生产和理解都离不开大脑的加工. 所以在“建模”时, 本来应该把人类大脑中的处理过程包括在内. 然而, 以 Chomsky 为代表的经典的语言理论就完全没有涉及这一环节. 因此这些理论虽然对“仅供计算机处理的”计算机编程语言作出了很大贡献, 但是在自然语言面前就显得无能为力.

如果让计算机具有大脑的联想、推理(判断)与选择功能, 就要把背景知识全部装进计算机, 并且设法使计算机具有“运用”这些知识的能力. 可是, 自然语言和人类千百年来所积累的知识是密切关联的, 因而知识不仅是无涯的, 而且还是在爆炸性地增长的, 这就使自然语言处理系统的开发者面临着一个无底的黑洞, 不知何时可以到头.

(2) 形式化文本和非形式化文本处理两头逼近.

首先, 要“写下来”, 因为口里说的话常常会受到周围语境、包括说话人面部表情的干扰. 当说话人说: “你可真行啊!”的时候, 脸上带着诚挚还是冷笑的表情, 意思就完全不一样. 而写成书面语言以后, 脱离了说话现场, 这种干扰就可以大大减少.

其次, 要“说清楚”. 怎样才算说清楚了呢? 这倒

是一个不容易说清楚的问题. 同样一种表述, 对人是“说清楚”了, 对计算机就往往“没有说清楚”; 对大人算是“说清楚”了, 对小孩就往往“没有说清楚”; 对专家算是“说清楚”了, 对一般人就往往“没有说清楚”. 一段正确的C语言程序对于装备有‘C编译器’的计算机算是“说清楚”了, 对没有装备有‘C编译器’的计算机就“没有说清楚”. 所以, “说清楚”的第一个必要条件就是“给出的信息所表达的内容能够和对方已有的知识相结合”, 能够为对方所“理解”. “说清楚”的判断标准是: 听话的一方“是否只需通过常规的脑力劳动(照章办事型信息处理)即可完成理解过程”.

第三, “形式化程度”要适宜. 人脑的“短期记忆”是有限的. 高度形式化以后产生的过长的句子使人无法把握. 反过来, 计算机在记忆方面没有问题. 这就是为什么. html文件同时提供给人看的“浏览”格式和冗长得不得了的给计算机看的“代码”格式的原因. 换句话说, “形式化文本”是为当代计算机这种“木头头脑但是手脚飞快的脑力劳动奴隶”准备的; 而“非形式化文本”是为以人为代表的“智慧生物”准备的. 当然, 这里面有一个“分寸”问题, 即使对于人, 话也不是讲得越简略越好. 话讲的过于简略就会引起误解.

“形式化”程度高的文本可以通过信息的缺失、扭曲而退化, 使得其中“形式化”成分的比重下降. 反过来, “非形式化文本”可以通过向其中添加、补充信息来提高“形式化”成分的比重. 按照XML格式来改写(补充)用自然语言书写的文章, 从而便于计算机处理, 就是一种典型的做法.

因此, 按照两头逼近的方式来解决NLP的问题: 一头是丰富计算机的知识库、提高计算机的知识运用能力; 另一头是通过人工标注、或者借助工具直接书写出便于计算机处理的形式化程度比较高的文本.

3 自然语言处理的智能特性分析

人工智能是用计算机模拟人类智能的一个重要方面. 人们通过程序使计算机能进行一些思维推理, 使其具备一定的环境适应、自动学习、自动决策、自然语言理解等人类高级智能. 所以, 自然语言处理是人工智能中研究最早并取得试验性成果的方向之一. 实际上, 在人工智能这一概念出现之前, 使用计算机处理自然语言的设想就已经产生了.

自然语言处理就是研究如何能让计算机理解并

生成人们日常所使用的(如汉语、英语等)语言, 使得计算机懂得自然语言的含义, 并对人给计算机提出的问题, 通过对话的方式, 用自然语言进行回答. 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.

自然语言是人类区别于其他动物的重要标志之一. 具有以下特征:

(1) 语言是人赖以从事复杂思维的工具, 思想是语言的内容(语义), 两者相辅相成. 但语言既不同等于说出的话, 也不能等同于写下的句子. 语句所表达的意义分为句义和话语义两部分.

定义1 句义是字义和词义根据一定规则组合的产物^[4].

我们可以相对孤立地考察字义和词义, 比如在查阅字典时, 我们可以机械地研究组字成词, 组词成句后的句义, 而不必考虑句子使用时涉及的语境因素. 句义是一般的意义、稳定的意义, 浅层语义. 包括逻辑语义(真值条件义)和字面语义.

句义是客观的、不包括人的主观因素的那部分内容, 被许多哲学家称为命题. 也就是说, 客观的句义以命题的形式出现, 一个命题或者真实地反映了外部世界的某个现象, 或者对某现象做出了不正确的、虚假的描述.

定义2 话语义是言者在特定语境中所表达的意义^[4].

话语义可以等同于句义, 但也可能超越句义, 有额外附加的意义, 甚至与句义完全不同. 话语义可被看成是由命题组成的集合, 然而, 只有说出的命题才构成话语义, 此外却还有存储在记忆中未经表述的命题.

虽然话语义并不等同于承载该话语的句义, 但听者总能从某个句义出发, 借助知识, 经推理而得到话语所表达的命题. 同时, 他可以用与言者原来使用的句子不一定相同的一个或几个句子, 把有关命题表达出来.

话语义是个别的意义、临时的意义, 深层语义.

(2) 人类对语言的表达处理能力的优势集中地体现在对语义、语用、语境和常识的应用及处理上, 而不是语法, 这正是为什么很多人并不懂得多少语法, 甚至一无所知, 但他们对语言的运用能力丝毫不逊色. 而在计算机自然语言处理的研究中, 人们很早就创立了一系列偏于形式化结构的语法理论, 对于自然语言的语义处理至今尚没有一种十分有效的处理方法, 这种先语法后语义的研究顺序恰好与人

脑对处理语言方法的掌握顺序相反。

(3)在人工智能界或语言信息处理领域中,人们普遍愿意接收著名的图灵(Turing)试验来作为判断计算机是否“理解”了某种语言的标准。具体的判断准则至少有如下四条:

①问答(Question-answering):机器能正确地回答输入文本中的有关问题;

②文摘生成(Summarizing):机器有能力产生输入文本的摘要;

③释义(Paraphrase):机器能用不同的词汇和句型来复述其输入文本;

④翻译(Translation):机器具有把一种语言(源语言)翻译成另一种语言(目标语言)的能力。

这个定义实际上表明了自然语言将会与人工智能碰到一样的难题——语用层次的世界语境知识的表达与识别。自然语言是人类智能的独特表现,实现了自然语言的理解也就意味着实现了一定意义的智能。

4 面向自然语言处理的人工智能框架

对智能的理解,一般有两个层次:第一个层次是在自我意识基础上的具有较高复杂性处理的理性能力,如数理逻辑推理能力、抽象思考等;第二个层次是在无意识基础上关于日常世界的经验能力,如行走、感知对象、条件反射及日常活动等。可以看出,一般的动物都具有第二个层次智能,而第一个层次智能只有人类才具有。对智能的定义,应包括对这两类心智能力的理解才是完整的。因此,可以认为人工智能的目的是提供一个系统的理论,既可以解释意向性的一般范畴,也可以解释以此为基础的各种基于自我意识与无意识的心智能力。

语言的表现是智能的一个突出特征,对自然语言理解的深入研究非常有助于智能研究的突破。当前,自然语言理解的研究主要限于语言的精细计算模型的探索,亦即计算语言学的主要研究范围。从更大的范围来讲,自然语言的理解需要哲学家、语言学家、心理语言学家、计算机科学家的共同协作。自然语言理解所涉及的知识是多层次的,包括语音和音韵知识、词语形态学知识、句法知识、语义知识、语用知识以及世界知识等。

从语用学的维度来看,自然语言的理解是主体(Agent)以知觉的形式对世界的表现,它具体地反映了主体对世界的感受程度及其富有个性化的呈现方式。仅从指称维度或统计概率维度来构建自然语言

处理系统显然是艰巨而永无止境的,因此,自然语言处理研究要想取得重大进展,显然需要在人工智能框架上实现重大突破。

也就说,完整地表达一个句子的语义信息,不但依赖于组成句子词汇的统计概率(仿生学假设),而且还包含指称信息(物理符号假设)及 Agent 信息(上下文、背景、常识、主体思维状态等,控制进化假设)。所以概率、指称、Agent 三位一体即是面向自然语言处理的人工智能框架,它类似于全息照片,可以使句子包含的信息更准确,更全面地表示出来(如图1所示)。

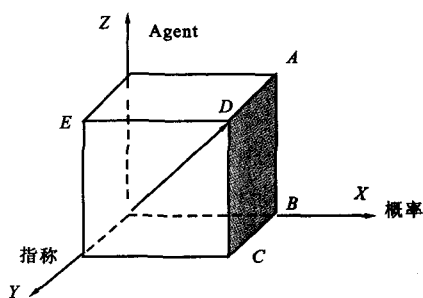


图1 面向自然语言处理的人工智能框架

图1中,X轴表示词序列,是外延的(显性结构),称为语言空间,句子的表层结构;Y轴表示词的指称信息,是内涵的(隐性结构),称为概念空间,句子的深层结构;Z轴表示句子的背景信息,包括主体的思维状态和上下文,称为知识空间。X,Y是客观的,静态的,语言的内部信息;Z是主观的,动态的,语言的外部信息。

在言语交际过程中,交际双方能直接凭感官得到句子的表层结构,指称是一个中介层面。它们能帮助我们得到语言的层次结构,以协助词义遵循语言空间提供的结构有规律地组合起来,从而得到句义。Agent是语境的简称,Agent的行为包括四个过程(感知、规划、落实和行为)与三个状态(信念、期望和意图),Agent利用感知得到的信息来动态地更新信念,利用信念进行推理和规划,然后基于信念和期望来着手落实某种意图,最后通过行为来实现这些意图^[5],通过Agent的学习、推理、协作达到理解语言的目的。

5 结束语

对语言的把握仅从指称维度是远远不够的,而应将语言视为一种行为,一种受语用力量和语境等诸多条件制约的行为,只有将这些诸多制约条件及

(下转第180页)



图 8 概率霍夫变换效果

0.016 s,这对于实时应用有很大的意义。

6 结束语

本文讨论了在智能车辆中车道线检测的方法,先进行图像的预处理,再利用概率霍夫变换的方法有效的提取出了车道线,并比较了概率霍夫变换和霍夫变换的执行速度,实验证明了概率霍夫变换的实时应用要明显优于霍夫变换。

参考文献:

[1] 鲁曼,蔡自兴,李仪.道路区域分割的车道线检测方法

(上接第 176 页)

指称的命题内容完整地形式化才能实现自然语言理解的完整意义上的研究.从这一意义来看,无论是基于句法—语义的规则语言模型框架,还是基于统计概率理论的语料库建模方法,自然语言理解都只是脱离使用者与语境知识而处于指称层次的片面理解,而只有基于语用学层次的理解才是完整的.因此,自然语言理解的关键是通过对话用力要素及语境的分析,识别言语的意向性,才能实现公共语言框架下的主体间性解释,进而实现智能的解剖。

参考文献:

[1] Stuart Russell. 人工智能——一种现代方法[M]. (2 版). 姜哲,金奕江,张敏,等译.北京:人民邮电出版社,2010.

[J]. 智能系统学报, 2010, 5(6):007—009.

- [2] Hsiao P Y, Yeh C W, Huang S S, et. A portable vision based real-time lane departure warning system: day and night [J]. IEEE Transactions on Vehicular Technology, 2009, 58(4): 2089—2094.
- [3] 陈光,任志良,孙海柱. 最小二乘曲线拟合及 Matlab 实现[J]. 软件技术, 2005, 24(3): 107—108.
- [4] 杨喜宁,段建民,高德芝,等. 基于改进 Hough 变换的车道线检测技术[J]. 计算机测量与控制, 2010:292—298.
- [5] Galamhos C, Matas J, Kittler J. Progressive Probabilistic hough transform for line detection[C]// Proceedings of Computer Vision and Pattern Recognition Fort Collins, co, USA: IEEE, 1999:23—35.
- [6] 张书真,向晓燕,赵英男,宋海龙. 基于二维直方图的曲线分割法[J]. 微电子学与计算机, 2010, 27(9): 21—23.

作者简介:

胡彬 男, (1985—), 博士生. 研究方向为模式识别与图像处理.

赵春霞 女, (1964—), 教授. 研究方向为智能机器人与智能检测系统、图形图像技术与模式识别.

[2] James Allen. 自然语言理解[M]. 北京: 电子工业出版社, 2005.

- [3] 晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用, 2004(16): 405—409.
- [4] 朱倩,程显毅,韩飞. 汉语单句语义三维表示模型的研究[J]. 智能系统学报, 2009, 4(2): 122—130.
- [5] Jiming Liu. 多 Agent 原理与技术[M]. 北京: 清华大学出版社, 2003.

作者简介:

蔡艳婧 女, (1979—), 硕士研究生, 讲师. 研究方向为人工智能.

程显毅 男, (1956—) 博士, 教授. 研究方向为人工智能.

潘燕 女, (1988—) 硕士生. 研究方向为自然语言处理.