·专辑:人工智能与情报学·

智能信息处理和智能信息分析前瞻

叶 鹰

(1.南京大学信息管理学院 江苏南京 210023)

摘 要:在智能信息系统整体架构下,智能信息处理和智能信息分析的应用前景包括智能分析、机器翻译和自动简报。DIKW概念链可以提供智能信息处理和智能信息分析的理论基础,自动简报可作为智能信息处理和智能信息分析的标志性应用,自然语言理解是智能信息处理和智能信息分析的关键技术。

关键词:智能信息处理;智能信息分析;自然语言理解;DIKW概念链

中图分类号:TP18;G250.252 **文献标识码:**A DOI:10.11968/tsyqb.1003-6938.2017116

A Prospect on Intelligent Information Processing and Intelligent Information Analysis

Abstract Under the framework of intelligent information system, the prospect applications of intelligent information processing (IIP) and intelligent information analysis (IIA) include intelligent analysis, machine translation and automatic summary report. It is pointed out that DIKW chain provided a theoretical foundation of IIP and IIA, and it is proposed that automatic summary report can be significant application of IIP and IIA. Natural language understanding (NLU) as key technology is strengthened.

Key words intelligent information processing; intelligent information analysis; natural language understanding; DIKW chain

在部署智能制造等国家重点研发计划和实施“互联网+”行动方案基础上,国务院于2017年7月发布了《新一代人工智能发展规划》[1],把发展人工智能提升到了国策高度。这一发展规划以“科技引领、系统布局、市场主导、开源开放”为基本原则,计划分三步实现战略目标:

第一步,到2020年人工智能总体技术和应用与世界先进水平同步,人工智能产业成为新的重要经济增长点,实现人工智能核心产业规模超过1500亿元,带动相关产业规模超过1万亿元。

第二步,到2025年人工智能基础理论实现重大突破,部分技术与应用达到世界领先水平,实现人工智能核心产业规模超过4000亿元,带动相关产业规模超过5万亿元。

第三步,到2030年人工智能理论、技术与应用总体达到世界领先水平,成为世界主要人工智能创新中心,实现人工智能核心产业规模超过1万亿元,带动相关产业规模超过10万亿元。

在这一发展规划中,与信息科技和情报学密切相关的既有大数据智能理论、类脑智能计算理论等新一代人工智能基础理论,也有自然语言处理技术、跨媒体分析推理技术等新一代人工智能关键共性技术,以及知识服务体系。本文沿袭作者对智能信息处理(Intelligent Information Processing,IIP)和智能信息分析(Intelligent Information Analysis,IIA)的前期探讨[2-3],概略前瞻融入当今人工智能的信息处理和信息分析,以期为情报界参与智能前沿领域和智能综合应用的创新提供微薄参考。

1 智能信息处理和智能信息分析的理论架构

人工智能研究无疑有计算机学界一马当先,纯粹技术不是情报学界所长,而信息处理与信息分析才体现情报学优势,因而人工智能与情报学的最佳结合非智能信息处理和智能信息分析莫属。

智能信息处理既包括海量多媒体信息检索与处理、大数据挖掘与集成、机器翻译、乃至生物信

收稿日期:2017-10-26;责任编辑:魏志鹏

息处理与量子计算等,也包括电子政务、电子商务、电子金融等领域中的智能化数据处理,总之以处理复杂信息和海量信息为己任。智能信息分析则以从处理过的信息中发现情报和知识为目标。尽管现有智能信息处理迷失在大数据里或淹没在各种算法中[4-5],新一代人工智能的曙光正让智能信息处理和智能信息分析在理论与技术的黎明中复苏。

一个完整的智能信息系统架构是一个有机体。其中智能信息处理作为前端,智能信息分析作为后端,以智能机把两者耦合为一体(见图1)。

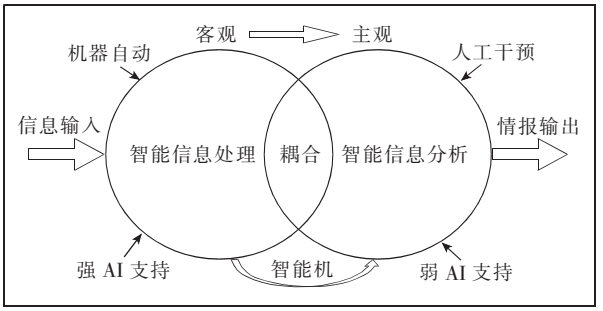

图1 智能信息系统架构

这样,信息由智能信息系统前端输入,经智能信息处理并提交智能信息分析后,从系统后端输出情报。智能信息处理多为客观成分,适用强人工智能技术支撑;智能信息分析则需主观介入,适用弱人工智能技术支持。

依照DIKW概念链量化模型[6],客观数据D经输入传递系统成为物理信息i;物理信息i经社会传递,转化为可接收的客观信息I;可接收的客观信息I经主体吸收,转化为带有主体价值判断的主观信息J即情报;情报J经结构化体系化而成为知识K;矩阵化个性化知识则构成智慧W。

从客观信息I到情报J间的转化是一关键环节。根据对数透视原理[6],从客观到主观需经对数转换,同时,为描述主体价值判断,引进价值系数 $v \in [0,1]$(匹配Rescher模型),可得如下关系式:

$$J=\log I^{v}=v \log I \quad (1)$$

式(1)确定了信息I和情报J的关系,即情报是信息的对数与价值系数的乘积。

在情报J进一步转化为知识K的过程中,采用分析信息学的合理假说[7-8]:有价值的信息才会使知识增加,单位信息增量产生的单位知识增量应与有价值信息量(情报量)成正比,即:

$$\frac{dK}{dI}=kJ=k\ ln\ I^{v} \quad (2)$$

其中k是信息的知识转化系数。于是,知识K是情报J对信息I的积分:

$$K=k\int JdI=k\int vlnIdI=kvI(lnI-1)+K_0=K_0+\triangle K \quad (3)$$

其中K_0是积分常数,代表原有的知识;而ΔK代表了新增加的知识。这正是著名的布鲁克斯基本方程,该推导过程的优势是给出了机理解释[9]。

以上内容可作为智能信息处理和智能信息分析的理论基础。

2 智能信息处理和智能信息分析的应用

作为智能信息处理的先驱,Luhn和Salton等已对智能分类、智能标引、智能文摘等进行过开拓性研究[10-13],智能检索也在计算机科技的推动下走向成熟,这些领域的智能化技术皆渐趋完善。未来的发展预期将是智能分析、机器翻译和自动简报。

2.1 智能分析

智能分析面临的很多问题需要自然语言理解支撑,尤其是中文信息的智能分析至少涉及:(1)词切分和词性标注;(2)概念标注与分析;(3)语义知识表示;(4)词典与知识库;(5)句法及语义分析等。因此,智能分析的前景是在自然语言理解基础上,融合已有的智能分类、智能标引等技术,发展出结合算法分析与计算智能的综合应用。

2.2 机器翻译

机器翻译的基本方法可分为基于规则(Rule-based)的方法和基于语料库(Corpus-based)的方法两大类。基于规则的机器翻译又可以分为基于转换的方法(Transform-based)和基于中间语言(Interlingua-based)的方法;而基于语料库的方法又可以分为基于统计(Statistic-based)和基于实例(Example-based)的方法。从实用效果看,混合(Hybrid)方法是最有前途的方法。当前,Google翻译器已显现出强大的人工智能特性,尤其是能实现多语种之间自由组合的智能化句级翻译和段落翻译,为今后的多语种机器翻译提供了现实前景。

2.3 自动简报

自动简报是自动文摘的升级,当年由Luhn首先

提出[11]、后来由 Salton[12-13]等不断推进改良的智能摘要已趋完善,如今一般通过原文文本分析、全文-文摘转换、重组生成文摘即可实现自动文摘。采用的方法既有基于符号、规则的方法,也有基于词频等文本表层特征的统计学方法。以后的自动简报将期望对文本、多媒体信息等进行智能化分析后提供类似摘要性质并加以特征分析的报告,报告长短可调控,真正实现输入信息后自动生成简报输出。

以上应用中自动简报可作为标志性应用。由于这些应用均涉及自然语言理解,因而自然语言理解技术作为关键技术若有突破就能带动智能信息处理和智能信息分析快速进步。

3 自然语言理解是智能信息处理和智能信息分析的关键

要进行完善的智能信息处理和智能信息分析,关键技术在于自然语言理解(Natural Language Understanding,NLU)[14]。由于人类智能在很大程度上需要通过自然语言表达,因此对自然语言的理解是智能信息处理和智能信息分析的关键。计算机能否实现智能信息处理和智能信息分析,关键就在于能否理解自然语言。因此,《新一代人工智能发展规划》把自然语言理解列入核心技术非常合理,自然语言理解的确是智能信息处理和智能信息分析的关键技术。

就目前国内外较有代表性影响也较大的自然语言理解理论而言,有主要作用于英语理解的 Chomsky 转换生成语法[15-16]、Schank 概念依存理论[17-18]和主要作用于汉语理解的鲁川句模理论[19-20]、黄曾阳概念层次网络(Hierarchical Network of Concept: HNC)[21-22]以及具有可比性的 WordNet[23]和 HowNet[24]等。如今真正能用于支撑技术研发的是 WordNet 和 HowNet。

3.1 WordNet

WordNet 最初由普林斯顿大学认知科学实验室的心理学教授 George A. Miller 创建于 1985 年,后由 Christiane Fellbaum 领导建设。该项目得到美国自然科学基金等的资助,其成就让创始人 George A. Miller 和 Christiane Fellbaum 于 2006 年获得 Antonio Zampolli 奖。

WordNet 的发展受益于语义网络和概念依存思想的综合,作为一个在线的英语词汇数据库(语义关系系统),WordNet 的一个重要理论基础是"可分离性假设"(Separability Hypothesis),即认为语言的词汇成分可以被离析出来并有专门针对性地加以研究。

在设计原理与方法上,WordNet 以同义词集合作为基本构建单位进行语义组织的,其基本设计原理是用"词汇矩阵模型",而一个词汇矩阵从理论上可以用单词及其同义词集合之间的映射来表示。当某个词有多个同义词时,通常同义词集合足以满足差异性的要求。虽然同义词只是词形之间的一种词汇关系,但由于这种关系在 WordNet 中被赋予了中心角色,因此同义关系的词被放在{ }中,与其他被放进[ ]中的词汇关系的词区别开来。

这样,用同义词集 Synsets(在一定语境中可以互换的同义词的列表)来表示词义,词汇关系存在于词形间,语义关系存在于词义间。WordNet 2.0 就把包括 152059 个词(words)、115424 同义词集(synsets)、203145 个词义对(word-sense pairs)等联系成为一个包括了上下位、同义、反义、部分、整体等词汇的语义关系网。至 2012 年 11 月发布 WordNet3.1 时,该联机数据库已包含 155287 个词、117659 个同义词集、206941 个词义对,可压缩成约 12 MB 数据集。

WordNet 中只对自然语言理解分析过程中较为重要的名词、动词、形容词、副词四类词进行处理,尤其注重名词和动词。WordNet 采用层次体系结构来表示名词,所有三种语义关系(下位义、部分义和反义)均被包含在内,结果组成一个互相连通的名词概念网络。WordNet 原初目标是要建立一个词典浏览器,如今已发展成自足的词汇数据库和语义机读词典。

3.2 HowNet

董振东、董强父子在 WordNet 启发下从 1988 年开始建立 HowNet(知网),这是一个结合中英文语料、以汉语和英语的词语所代表的概念为描述对象、以揭示概念与概念之间以及概念所具有的语义关系和语义网络为基本内容的语义知识库。

HowNet 与 WordNet 的最重要差异在于其哲学思想,即认为世界上一切事物(物质的和精神的)都在特定的时间和空间内不停地运动和变化,一个事物可

以被视为是整体,也可以被认为是部件;每一事物都包含有多种属性;事物之间的异同是由属性决定的。

在设计理论与方法上,HowNet采用与WordNet类似的自上而下的建设方法。其基本设计原理是把概念与概念之间的关系以及概念的属性与属性之间的关系组成一个网状知识系统,采用自上而下的归纳方法,通过对全部基本义原进行观察分析并形成义原标注集,然后再用更多的概念对标注集进行核实并据此建立完善的标注集。因此,提取义原作为基本构建单位进行语义组织是HowNet的关键。

在语义关系的描述上,HowNet中的上下位关系由概念的主要特征体现,也具有继承关系,而WordNet只是词义之间的上下位关系;HowNet对于同义的定义与WordNet相似,但WordNet的同义关系是显性的,而HowNet的同义关系是隐性的;HowNet中的反义关系则比WordNet定义的要宽泛些。

至2007年,HowNet形成了围绕800多个事件义原构成的标注集及其标注出的事件概念为网络的知识库。而HowNet的目标是要建立一个面向计算机的多重语义关系及知识网络,为建立自然语言处理系统提供所需知识库。

总的来看,WordNet拥有丰富的词语概念,由于许多国家都在WordNet基础上建立了词汇数据库,所以WordNet已有多国语言处理的词汇转换接口,且一直在持续发展更新中,这是其显著优势。HowNet则在语义知识构建和推理设计方面有优势,只可惜2007年后似已停滞。

从智能信息处理和智能信息分析的理论需要看,自然语言理解及其技术可以提供指导思想和操作技术,因此具有作为基础理论和关键技术的潜质。但仅仅依靠自然语言理解在技术上也是不够的,智能信息处理和智能信息分析不仅需要NLU,也需要计算智能与算法技术的集成,并与语义网(Semantic web)、关联数据(Linked data)等研究[25-26]整合发展。

4 结语

展望未来,智能信息处理和智能信息分析的基础理论可望形成,以自动简报为前瞻标志的智能信息处理和智能信息分析应用可望实现,而作为智能信息处理和智能信息分析关键技术的自然语言理解问题依旧。借力国家新一代人工智能发展规划,自然语言理解理论与技术可能持续进步,进而推动智能信息处理和智能信息分析获得突破。

参考文献:

[1] 国务院.国务院关于印发新一代人工智能发展规划的通知[EB/OL].[2017-09-10].http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[2] 叶鹰.智能信息处理的基础理论探讨[J].情报科学,2008(9):1281-1285,1291.

[3] 叶鹰.智能信息分析的理论基础与技术模型[J].情报学报,2005,24(2):233-236.

[4] 王耀南.智能信息处理技术[M].北京:高等教育出版社,2005.

[5] 郑家恒.智能信息处理[M].北京:科学出版社,2010.

[6] 叶鹰,马费成.数据科学兴起及其与信息科学的关联[J].情报学报,2015,34(6):575-580.

[7] 叶鹰.信息科技基础理论的分析建构[J].情报学报,1999,18(2):160-166.

[8] 叶鹰.分析信息学的理论基础[J].情报学报,2000,19(4):380-384.

[9] Ye F Y.Measuring Knowledge:A Quantitative Approach to Knowledge Theory[J].International Journal of Data Science and Analysis,2016,2(2):32-35.

[10] Luhn H P.A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J].IBM Journal of Research and Development,1957,1(4):309-317.

[11] Luhn H P.The Automatic Creation of Literature Abstract[J].IBM Journal of Research and Development,1958,2(2):159-165.

[12] Salton G.Automatic Text Processing:The Transformation,Analysis,and Retrieval of Information by Computer[M].Reading,MA:Addison—Wesley,1989.

(下转第95页)