文章编号：1003-0077(2014)01-0001-08

语言计算的重要国际前沿

孙茂松[1]，刘 挺[2]，姬东鸿[3]，穗志方[4]，赵 军[5]，张 钹[1]，吾守尔·斯拉木[6]，俞士汶[4]，
朱 军[1]，李建民[1]，刘 洋[1]，王厚峰[4]，吐尔根·依布拉音[6]，刘 群[7]，刘知远[1]

(1. 清华大学 计算机系，北京 100084；2. 哈尔滨工业大学 计算机学院，黑龙江 哈尔滨 150001；
3. 武汉大学 计算机学院，湖北 武汉 430072；4. 北京大学 信息学院，北京 100871；
5. 中国科学院自动化研究所，北京 100190；6. 新疆大学 信息学院，新疆 乌鲁木齐 830046；
7. 中国科学院计算技术研究所，北京 100190)

摘 要：该文在互联网规模语言信息处理的语境下，从语言计算基础模型、语言分析、语言资源建设、机器翻译、文本内容理解与问答等多个方面，对国内外相关重要动态进行了评述，讨论了语言计算的若干前沿问题及其对中文信息处理近期研究工作所提出的要求。

关键词：语言计算；研究前沿；评述；中文信息处理

中图分类号：TP391 **文献标识码**：A

Frontiers of Language Computing

SUN Maosong[1], LIU Ting[2], JI Donghong[3], SUI Zhifang[4], ZHAO Jun[5], ZHANG Bo[1], WUSHOUER Silamu[6], YU Shiwen[4], ZHU Jun[1], LI Jianmin[1], LIU Yang[1], WANG Houfeng[4], TURGUN Ibrahim[6], LIU Qun[7], LIU Zhiyuan[1]

(1. Department of Computer Science, Tsinghua University, Beijing 100084, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China;
3. School of Computer Science, Wuhan University, Wuhan, Hubei 430072, China;
4. School of Information Science and Technology, Peking University, Beijing 100871, China;
5. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
6. School of Information Science and Technology, Xinjiang University, Urumqi, Xinjiang 830046, China;
7. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper surveys research frontiers of language computing in the context of Web-scale text information processing, covering the perspectives of fundamental computational model, language analysis algorithm, linguistic resource construction, machine translation, content understanding as well as question and answering. Several related key issues are discussed, and their significance to Chinese information processing in the near future is also addressed.

Key words: language computing, research frontiers, survey, Chinese information processing

1 引言

赋予机器以人类的语言能力，一直是科学家们的梦想，其研究几乎与计算机的问世同步，几个里程碑式的发展阶段，体现了人类对语言计算本质的认识不断深化的过程。1947 年，美国著名科学家 Weaver 在给控制论之父 Wiener 的一封信中首次提到了利用计算机进行自然语言翻译的可能性。1949 年，他发表了《翻译备忘录》，正式提出机器翻译的思想(同年他还与信息论之父 Shannon 合著出版了影响深远的《通信的数学理论》)。受信息论的影响和鼓舞，这个阶段的研究把句子看作是串行的字符流，把机器翻译看作是一种机械地解读密码的过程，乐观地认为借助计算机的能力，通过词与词的对应即可实现机器翻译。由于忽视了语言的本质——具有结构性，这种尝试当然碰得头破血流。1966 年，美国科学院语言自动处理咨询委员会公布了耗时两年

收稿日期：2013-10-16 **定稿日期**：2013-12-16

调查完成的、语言信息处理史上著名的ALPAC报告《语言与机器》,指出机器翻译研究遇到了机器难以逾越的"语义屏障"(Semantic Barrier),全面否定了机译的可行性。

此后,学者们日益认识到语言结构分析的重要性,沿着两条主线进行了系统深入的探索。一条主线以句法为主,始自20世纪50年代中期贯穿至80年代末期,经典工作包括现代语言学之父Chomsky的短语结构语法和转换生成语法,以及一批著名学者对短语结构语法的扩展,例如,词汇功能语法、中心语驱动短语结构语法、广义短语结构语法、扩充转移网络等,引入了复杂特征和词汇化信息(主要在句法层面,但也在模型中为语义留出了位置)。另一条主线以语义为主,主要集中在20世纪60年代末期至70年代中期,经典工作包括著名语言学家Fillmore的格语法(后演变为框架语义学),著名数理逻辑学家Montague的蒙太古语法,著名计算机科学家Schank的概念依存理论,著名人工智能学家Simmons的语义网络理论以及图灵奖获得者、人工智能之父Minsky的框架表示法等。这两条主线上的研究工作在理论深刻程度上无与伦比,闪烁着人类智慧的熠熠光辉(一般被归入理性主义的范畴),但也存在严重的不足。主要问题是,根据语言学家的思辨和语感人工编制句法规则集,难以保证对复杂语言现象的覆盖能力;而由于受到语义资源、计算能力等各方面的限制,语义分析仅限于在受限领域研制一些"玩具"系统,距离真实应用遥不可及。

有鉴于此,1990年在芬兰赫尔辛基召开的第13届国际计算语言学大会适时地提出了处理大规模真实文本的战略任务,开启了语言计算的一个新的历史阶段——基于大规模语料库的统计自然语言处理(属于经验主义范畴),并在语音识别、文字识别、机器翻译、信息检索等领域中取得了巨大进展,因此迅速壮大成为引领自然语言处理研究领域至今的主流方法。

耐人寻味的是,统计自然语言处理的基石是Hartley和Shannon的信息论以及建于其上的"语义无关"假设。信息论主要从统计的角度研究由串行字符流组成的消息的编码与解码问题,与语言具有丰富的结构(语义)这一根本性质并不契合,却能够取得如此骄人的成绩,确乎有些令人惊讶。非常重要的因素是:互联网的蓬勃兴起为这种模型的充分训练提供了优越的语料库条件。现今的统计模型虽然能够进行部分的语言结构分析,但它是在马尔可夫化的假设之下,是对语言结构分析的一个简化,所以只能部分缓解但不可能完全克服"语义屏障"问题。我们正处于下一轮螺旋式上升周期的开端:带统计的理性主义,或者具深度的经验主义,两大方法范畴应殊途同归,汇流合进。这个新的历史进程在宏观上至少呈现出以下三个重要态势。

(1) 从句法分析深入到语义分析。英文语言分析从深度上已超越句法整体上推进到了语义层面,从广度上则扩张到了互联网规模。IBM的DeepQA在强大的硬件平台和知识资源的支持下,融合了多种语言分析技术,包括浅层分析、命名实体识别及关联、深层分析、语义角色标注、逻辑表达式演算甚至共指消解等。美国华盛顿大学Etzioni教授在美国国防部先进项目研究局(DARPA)支持下于2009年启动的"Machine Reading(机器阅读)"重大项目,试图利用深度语言分析技术自动阅读整个互联网的文本,得到表示句子语义的逻辑表达式,从而构造互联网规模的知识库。

(2) 经验主义和理性主义的深度融合。语言计算往往是一个欠约束的不适定问题(Ill-posed Problem)[1],其求解迫切需要新的计算模型与理论。值得庆幸的是,近年来,机器学习理论取得了重大进展,为互联网条件下的语言结构学习及分析打下了理论和方法上的坚实基础。2011年图灵奖获得者Pearl教授的"基于图结构的概率推理"正在对自然语言处理、语音处理方向产生重要影响;2011年*Science*上发表了题为"心智何来?统计、结构与抽象"的文章,阐发了更"类似人"的机器学习系统能在柔性结构表示的层级体系上进行概率推理,抽象知识可引导从稀疏数据中进行学习和推理等[2]。这启示我们,基于统计的语言结构学习模型和主要以规则形式存在的语言知识的融合将是语言计算很有前途的解决方略。

(3) 互联网海量弱标注数据的利用。互联网上海量、繁杂又包含大量噪声的数据给语言计算带来了严重困难,但同时也为解决这些困难创造了新的可能性,为关键技术的突破带来了契机。如互联网上广泛存在的弱标注数据资源(所谓弱标注是指观测数据不能完全表示模型中隐含变量的取值,或标注数据带有噪声,或与直接任务间接相关的标注数据以及无标注数据等)为语言结构学习算法提供了丰富的语言资源,互联网上信息的高度冗余性使准确抽取知识更加可行。

以下从语言计算基础模型、语言分析、语言资源建设、相关关键技术(包括机器翻译、文本内容理解与问答)等方面,对国内外重要动态进行评述。

2　语言计算的若干国际前沿：问题与进展

2.1　语言计算基础模型

不同于线性的信号序列，语言是一种具有复杂结构的对象，语言的自动分析与理解，需要借助于结构化学习的理论与方法。

从复杂数据中学习具有结构的统计模型是过去20年统计机器学习领域的核心问题之一。图和一阶谓词逻辑是表示结构信息的两个有效框架，代表性的工作分别是条件随机场和马尔可夫逻辑网络。对于同一类模型，从参数学习的角度又可分为最大似然估计、最大间隔学习以及综合两者优点的最大熵判别式学习等。

在给定充足的完全标注样本的情况下，一般可以学到鲁棒的模型对未知样本进行结构预测。但是，在训练样本有限的情况下，学习具有复杂结构的统计模型是一个不适定的问题，即能够充分描述给定数据样本的模型可能有很多个（有可能是指数多个）。理论研究表明[2]，为了学习一个稳定的统计模型必须借助“额外的信息”或“额外约束”。这里的额外信息可以分为以下两个方面：①先验假设或先验知识。主观地对可行的模型空间及其分布进行先验假设或者尽量引入客观的先验知识（语言计算中如句法、语义、情境等知识）作为约束，代表性的工作包括基于稀疏正则化的概率图模型学习[3]、具有树状结构的回归分析[4]以及基于贝叶斯推理的拉普拉斯最大间隔马尔可夫网络[5]、后验正则化方法[6]等；②未标注数据。为了弥补完全标注数据不足的问题，针对具有结构的统计模型的半监督以及无监督学习方法得到了广泛的研究，代表工作包括半监督的最大间隔马尔可夫网络[7]、无监督的马尔可夫逻辑网络[8]以及无监督的语法学习[9]等。

面对越来越复杂的现实数据，人们更希望发现其中隐含的深层结构，而不仅仅停留在表面的一两层。深度学习（Deep Learning）致力于从数据中自动学习更一般的从底层特征到高层概念的多层抽象表示，逐渐成为近年来的研究热点。2006 年以后，以 Hinton 关于深层信念网络 DBN[10] 的革命性工作为代表，出现了 DBN、Autoencoder 等学习深层结构的算法[10-12]。Hinton 等人[13]在 *Science* 杂志上提出，可以利用多层受限波尔兹曼机 RBM 的 Pre-training 方法学习到很好的低维表示。

深度学习已经在分类、回归、维数约简等学习问题中取得了成功，并被有效应用到图像分析、语音识别和自然语言处理等众多具体领域中。例如，在语音识别中，Seide 等人[14]将深层神经网络 DNN 与传统的 HMM 相结合，在大规模语料上的转写任务中错误率较现有方法下降了 30% 左右。又如，Collobert 等人提出了一个基于深度学习的自然语言处理框架[15]，可以进行 POS、Chunking、NER 和 SRL 等多种自然语言处理典型任务。

互联网环境给基于结构的统计学习既带来了机遇也带来了挑战。如何有效利用弱标注资源，同时避免噪声的负面影响是目前机器学习[16]及不同应用领域关注的热点，这方面的研究刚刚起步。也有一些工作研究如何从极少数种子样本（有标注的示例）进行迭代的、滚雪球似的增量学习，例如，用于自动抽取互联网上实体关系的 StatSnowball 系统[17]和美国工程院院士、卡内基梅隆大学 Mitchell 教授带领的 NELL（Never Ending Language Learning，“永不停止的语言学习”）项目。虽然上述工作已经取得初步成功，如何自动或者半自动地从互联网上获取有用信息仍然是一个难题。

语义和内容的结构属于深层结构，语言计算的实质是深层结构的分析问题。由于自然语言的模糊性、歧义性和复杂性，人工编制规则的理性主义方法难以满足互联网环境下语言深层结构分析的需求。另一方面，由于深层结构包含密集的语义关联，在保证计算深度的条件下，还需要大量实例以保证统计模型的学习性能，传统经验主义的统计模型遇到了深刻的困难。而深度学习可望把一个复杂任务的学习过程分解为多层抽象表示的非线性推导过程，从而保证统计上的可行性和计算上的可操作性。深度学习在图像分析和语音识别等领域已经显示出卓尔不群的优越性，在自然语言处理的若干具体任务中也取得了初步进展（虽然其成效并不很显著）。我们认为，深度学习的理论与方法对构建语言计算的基础模型具有重要的启发性和参考价值。

2.2　语言分析

自然语言分析按处理对象由低向高分为：词汇分析、句子分析与篇章分析。其中，句子分析占据核心地位，一直以来都是自然语言处理研究的重点和难点。

句子级的语言分析主要包括句法分析和语义分析，句法分析目前相对成熟，按照所使用文法的不同主要分为短语结构文法和依存文法。由于依存结构相对于短语结构来说，具有形式简洁、易于标注、便于应用、时间复杂度低等优点，因此逐渐受到更多的重视。语义分析目前主要采用语义角色标注的形

式,它标注句子中主要动词的语义角色。CoNLL 2009年组织了一次7国语言句法分析和语义角色标注的联合任务评测,句法分析采用依存文法,国内外20多家单位参与了这次评测,最终的结果表明:英文句法分析准确率最高93.5%,而中文只有83.3%;英文语义角色标注准确率最高86.2%,而中文为78.6%,中文句法分析和语义角色标注比英文低8%~10%。这些工作都需要建立在大规模的句法语义语料库的基础之上,而语料库的建立需要大量的专家标注,因此一些利用弱标注知识的方法也纷纷涌现出来,例如,利用生文本对词语进行聚类自动产生词类标签[18],使用双语语料产生可信度较高的依存词对结构[19],以及从海量网络资源中挖掘对句法有帮助的知识[20]。

国际上对句子级深度语义分析研究的关注程度在逐年增加。主要的研究方法包括:采用同步上下文无关语法将句子映射成逻辑表示式[21];组合范畴语法(CCG)和lambda逻辑演算相结合[22];采用依存组合语义树(DCS)表示句子语义[23];基于无监督的方法进行语义分析[24]。上述英语语义分析方法的共同特点是依赖于句法分析的结果。此外,值得注意的是,近年来国际上还进一步提出了Parsing the Web("分析互联网")的理念和任务。

与词语、句子等更小的语言单位相比,篇章能够从宏观上反映信息的整体结构和主题内涵,对于内容理解和语言交流具有更直接的作用。因此,在句子分析基础上进一步研究篇章分析,是实现深度计算的必要途径。

共指消解是篇章分析中传统的研究方向,ACL、COLING、EMNLP、EACL、NAACL等重要的国际会议都召开过共指消解的专题会议,*Computational Linguistics*也出版了专辑,先后出现了MUC、ACE等与共指消解相关的国际评测。初期的共指消解研究以语言学方法为主,随后引入机器学习方法,多采用二元分类模型。目前,共指消解逐渐向多资源、跨文档、海量数据统计的方向发展,典型的工作如使用世界知识的共指消解方法[25]。

句间关系识别是篇章语义分析的重要组成部分,以美国国家科学基金会NSF资助的PDTB(Penn Discourse Tree Bank)项目为代表。该项目的目标是通过为句间关系建模来分析篇章结构、挖掘语义信息。早期的句间关系识别以关联词语为中心[26]。目前,越来越多的研究者提出不依赖关联词语的新方法[27],例如,核函数方法被用于引入结构化信息帮助识别句间关系,同时使用事件时序信息帮助句间关系识别。挖掘语义信息来帮助识别句间关系,并用它来支持其他任务[28],是该方向未来的发展趋势。

2.3 语言资源建设

语言知识资源主要包括句法资源和语义资源。20世纪50年代以来,句法分析占据主流地位,相应的句法资源的发展与建设相对成熟,例如,在英文语言信息处理领域影响较大的美国宾夕法尼亚大学开发的英语句法树库Upenn Treebank,北京大学开发的现代汉语语法信息词典和大规模词性标注语料库,基本满足了浅层语言分析的需求。然而,对语言进行深层分析需要语义知识资源的支撑。近年来,许多语言学家、心理语言学家和计算语言学家从不同研究角度出发,组织研制了众多的语义知识库。

认知层面的概念、框架、情境等语义信息,在语言层面主要通过词汇、句子、语篇等语言单位来承载和实现。

在概念语义方面,以词汇为单位组织语义信息的典型工作包括WordNet、VerbNet、HowNet、MindNet等。其中,WordNet从认知语言学的角度描述概念。描述信息包括同义词集合(Synset)及其概念层级关系,是一种外延式的知识描述方式。HowNet描述的是概念及概念属性之间的关系,是一种内涵式的知识描述方式。VerbNet在对英语动词进行分类的基础上描述了动词的论旨角色、角色的语义选择限制以及简单的事件框架信息。美国微软公司开发的MindNet是利用句法分析器自动分析词典释义文本,通过自动构建的方式而建立。

在框架语义方面,近年来一个重要进展是从谓词—论元(Predicate-Argument)关系入手把句法关系和语义角色描述联系起来,形成句法语义链接知识库。宾州大学在宾州树库基础上,进一步发展了语义角色标注的命题库(PropBank)[29]和NomBank[30],在句法关系链上添加相应的特定谓词(包括名词化谓词)的论元结构。加州大学伯克利分校的FrameNet计划[31]以Fillmore框架语义学理论为基础,试图用语义框架对语义(包括词义、句义和情境义)进行系统的描述和解释。

在情境语义方面,在词义、句义描写的基础上,语义资源建设又向更高层次语义的描写发展,出现了篇章级标注语料库,包括RST-DT[33]、宾州语篇树库PDTB等。其中,RST-DT在系统功能理论框

架下创建，在宾州语料的基础上，描述了语篇单位之间的修辞结构关系。PDTB 是目前规模最大的篇章级标注语料库，其标注语料也来源于宾州树库，将语篇连接词看作二元的语篇关系的谓词，目标是标注语篇连接词以及语篇连接词所支配的论元。

在多类型、多层面语言资源共存的现状下，多源异构语言知识资源的融合成为迫切需要解决的问题。OntoNotes[34]在句法结构上，实现了词义知识、指代关系等语义知识的标注。但目前只是把现存的比较典型的语言资源简单地连接在一起，包括：词汇、句法、篇章级语言资源，还没有对语言知识实现真正的融合。

尽管上述语义资源在描述规模和深度上都达到了一定水平，但是对于面向互联网深度计算的目标，仍存在问题和不足。

2.4　机器翻译

得益于互联网文本的持续快速增长，数据驱动的统计方法近年来逐渐成为机器翻译领域的研究热点，其发展趋势可以归纳为以下两个方面。

第一、语言层次持续加深。统计机器翻译近 20 年的发展是一个沿着机器翻译先驱 Vauquois 提出的著名的“机器翻译金字塔”(Machine Translation Pyramid)从底层不断向顶层攀爬的过程：在保持从大规模真实文本中自动获取翻译知识的同时不断加深语言分析的层次。早期的统计机器翻译方法以词作为翻译的基本单元，属于位于机器翻译金字塔最底端的直接翻译方法。本世纪初，基于短语的方法由于能够有效地对局部的择词和调序进行建模，开始成为统计机器翻译的主流，并在 Language Weaver、Google、Microsoft、百度和有道等商用机器翻译系统中得到广泛使用。2005 年后，基于句法的方法利用同步语法对语言的层次结构进行建模，实现了机器翻译金字塔中句法层次的转换。

尽管统计机器翻译取得了长足的发展，但是目前仍未达到语义层次。保证源语言文本和目标语言文本的语义相同是机器翻译的首要目标，只有实现了对语义的分析、转换和生成的统计建模，并在大规模真实数据上自动获取语义翻译知识，统计机器翻译才有可能逼近这一目标。虽然美国卡内基梅隆大学、美国罗切斯特大学、新加坡信息通讯研究院和香港科技大学的一些学者尝试将语义引入统计机器翻译，但是所采用的语义角色标注和潜在语义分析层次较浅，无法真正利用深层次的语义知识来指导翻译过程[35-37]。更重要的是，这些工作并未建立真正意义上的语义翻译模型，只是对基于短语的系统的输出结果做后处理，或者将语义信息作为基于句法的系统中的特征函数。

第二、语言种类不断拓广。统计机器翻译的研究对象开始从英语、汉语和阿拉伯语等少数几种资源丰富的语言向更多的资源匮乏的语言拓广。2002 年，美国国家标准技术研究院(NIST)开始组织一系列国际机器翻译评测，对机器翻译的发展起到了巨大的推动作用。出于政治因素的考虑，NIST 评测将汉语—英语和阿拉伯语—英语设为固定评测任务，引导学术界将英语、汉语和阿拉伯语作为机器翻译的主要研究对象。欧洲的 EuroMatrix 项目(2006～2009)和 EuroMatrixPlus 项目(2009～2012)更是试图将统计机器翻译技术扩展到欧洲所有的语言对(如捷克语、丹麦语、荷兰语、芬兰语等)，形成一个巨大的欧洲语言机器翻译矩阵。

2.5　文本内容理解与问答

自动问答是自然语言处理、人工智能和信息检索领域的热点研究方向之一。它接受用户用自然语言提出的问题，并返回该问题的答案。华盛顿大学 Etzioni 教授 2011 年在 *Nature* 上指出问答系统是互联网搜索引擎发展的方向[39]。

问答系统的发展经历了几个阶段。20 世纪 60 到 80 年代随着人工智能技术的发展，基于知识推理的问答系统在有限领域获得成功，例如，MIT 开发的数学符号运算系统 MACSYMA；20 世纪 90 年代到本世纪初期，随着大规模语料库的建立和互联网的发展，自然语言处理、信息检索、信息抽取、人工智能、机器学习等多种技术相互融合，形成了一种新的问答技术—问答式检索技术，并在 TREC、TAC、CLEF 等评测计划的推动下得到迅速发展，例如，MIT 开发的 Start、Umass 开发的 QuASM 以及 Microsoft 开发的 Encarta 等。但是，由于受限于自然语言处理和人工智能技术的水平，问答式检索系统只能较好地回答一些相对简单的事实性、列表性和定义性提问，离用户更广泛的真实信息需求存在巨大的差距，这极大地限制了自动问答系统的实用性。近年来，随着多层次自然语言处理技术的不断融入，问答系统向深层次发展。2008 年微软以 1 亿美元收购了语义搜索引擎 Powerset，其核心是基于自然语言处理技术的问答系统。2011 年，IBM 基于深层问答技术 DeepQA “沃森”系统再一次在具有历史

意义的“人机大战”中战胜人类;之后,苹果公司在Wolfram Alpha知识计算引擎之上推出了智能生活助手Siri系统。以上事件成为问答系统发展的重要里程碑。

实现网络环境下的深度问答这一目标,需要文本内容理解技术的支撑。文本内容理解最理想的途径是对文本中每个句子所包含的语义内容自动地进行形式化描述(例如,表示为谓词逻辑表达式),然后融合这些语义内容并在大规模知识系统中进行推演得到新的知识或事实,从而实现对文本内容全面、深入的理解。要达到这个“理想”境界,无疑极具难度,还有一段较为漫长的路要走。

为了降低文本内容理解的难度,一个替代的方法是文本内容抽取。文本内容抽取的任务是:从自然语言文本中抽取指定类型的实体、关系、事件等事实信息,并形成结构化数据输出。从20世纪80年代开始,在MUC、ACE、TAC等评测会议的大力推动下,文本内容抽取技术的研究得到蓬勃发展。但是,传统内容抽取评测任务是面向限定领域文本的、限定类别实体、关系和事件的抽取。近年来,为了适应互联网实际应用的需求,人们开始以较大的热情关注开放域内容抽取技术[40],其特点在于:①文本领域开放:处理的文本是不限定领域的网络文本;②内容单元类型开放:所抽取的内容单元不限定类型,而是自动地从网络中挖掘内容单元的类型,例如,实体类型、事件类型和关系类型等。

目前,文本内容抽取大多只能抽取文本中显式表示的内容,对于文本中隐含的内容基本无能为力,学者们于是开始研究文本内容推演问题。Schoenmackers在把文本内容表示成一阶谓词逻辑的基础上,利用自动习得的推理规则在已有知识库上进行推演,得到新的事实以满足用户的知识需求。实验显示,受限于文本内容抽取性能的影响,逻辑推理效果一般;同时由于推理规则学习方法的局限,当面对深层推理时性能尚不能满足实际需求[41]。这方面的研究还比较初步。

文本内容抽取和内容推理技术日益受到工业界和学术界的高度关注。例如,Google自2010年收购了FreeBase后一直致力于构建相互关联的实体及其属性的规模巨大的“知识图谱”。目前这个知识图谱所包含的实体已数以亿计。CMU在DARPA、NSF、Google、Yahoo! 共同资助下正在开展的研究Read the Web(“阅读互联网”),致力于研发一个不停学习的计算机系统—NELL,不间断地从互联网上抽取和挖掘知识,以构建一个可以支持多种智能信息处理应用需求的海量规模网络知识库[42]。从2010年系统开始运行以来,NELL已经收集了超过1 500万候选事实,其中具有很高可信度的事实有将近90万,关系和类别有810种。

互联网环境的深度问答需要开放域文本内容理解技术,分析文本所蕴含的实体、事件及其关联演化关系等内容信息。这涉及到开放域内容抽取技术和内容推演技术。开放域内容抽取研究目前大多以实体为中心,停留在实体及其关系抽取的层面上,对事件抽取、事件关系抽取和事件关系推演方面的研究才刚刚起步。

3 结语

如上所述,近年来语言计算的国际前沿正经历着深刻的变化和拓展,各种重要动态如“山阴道上行,山川自相映发,使人应接不暇”。“分析互联网”、“阅读互联网”、“永不停止的语言学习”、“知识图谱”,这些以前对自然语言处理而言难以想象的困难任务,目前都已经驶入研究的轨道上了。在互联网规模语言信息处理这个基本需求的“压迫”之下,语言计算研究终于被彻底地“倒逼”出了“象牙塔”而置身于互联网这个复杂巨系统中,带着兴奋,也无可避免地带着几分忐忑和迷惘。显然,无论是挑战还是机遇都是空前的,我们的学术研究能力和学术组织能力目前都很不适应,亟需鼎新求变。

《国家中长期科学和技术发展纲要》(2006~2020)中将以自然语言理解为基础的“以人为中心”信息技术列为前沿技术。这是国家重大科技需求的体现。在中文信息处理领域,“分析中文互联网”、“阅读中文互联网”、“永不停止的中文语言学习”、“中文知识图谱”等与英文平行的大规模深入研究,几乎都还没有开展起来。中文的特点所导致的中文信息处理与生俱来的困难性,使得这些任务更加艰巨。但这种状况也提示着我们,中文信息处理很可能正处于一个重大的创新窗口期。我们必须认清并瞄准国际重要前沿,迎难而上,攻坚克难,谋求中文信息处理研究产生实质性突破,进而占据中文信息处理技术的战略制高点。

致谢

本研究受到教育部哲学社会科学研究重大课题

攻关项目(10JZD0043)和国家自然科学基金项目(61170196)的支持。本文是以共同作者为主要成员的国家重点基础研究发展计划 2013 年度重要支持方向“互联网环境中文信息处理与深度计算的基本理论与方法”申请团队在项目申请时期集体思考、研讨的结晶。

参考文献

[1] 张钹,自然语言处理的计算模型[J]. 中文信息学报,2007,21(3):3-7.

[2] Tenenbaum J, Kemp C, Griffiths T, et al. How to Grow a Mind: Statistics, Structure, and Abstraction [J]. Science, 2011,(331): 1279-1285.

[3] Zhu J, Lao N, Xing E. Grafting-Light: Fast, Incremental Feature Selection and Structure Learning of Markov Networks[C]//Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010.

[4] Kim S, Xing E. Tree-guided Group Lasso for Multi-task Regression with Structured Sparsity [C]//Proceedings of International Conference on Machine Learning (ICML), 2010.

[5] Zhu J, Xing E, Zhang B. Laplace Maximum Margin Markov Networks [C]//Proceedings of International Conference on Machine Learning (ICML):1256-1263, 2008.

[6] Ganchev K, Gra a J, Gillenwater J, et al. Posterior Regularization for Structured Latent Variable Models [J]. Journal of Machine Learning Research. 2010 (11):2001-2049.

[7] Altun Y, Tsochantaridis I, Hofmann T. Hidden Markov Support Vector Machines[C]//Proceedings of International Conference on Machine Learning (ICML), 2003.

[8] Poon H, Domingos P. Unsupervised Ontology Induction from Text[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2010.

[9] Cohen S, Smith N. Covariance in Unsupervised Learning of Probabilistic Grammars[J]. Journal of Machine Learning Research, 2010(11):3017-3051.

[10] Hinton G, Osindero S, Teh Y. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006(18): 1527-1554.

[11] Bengio Y, Lamblin P, Popovici D, et al. Greedy Layer-Wise Training of Deep Networks[C]//Proceedings of Advances in Neural Information Processing Systems 19 (NIPS 2006): 153-160, MIT Press, 2006.

[12] Ranzato M A, Poultney C, Chopra S, et al. Efficient Learning of Sparse Representations with an Energy-Based Model[C]//Proceedings of Advances in Neural Information Processing Systems (NIPS 2006), MIT Press, 2007.

[13] Hinton G E, Salakhutdinov R. Reducing the dimensionality of data with neural networks[J]. Science, 2006(313): 504-507.

[14] Seide F, Li G, Yu D. Conversational Speech Transcription Using Context-Dependent Deep Neural Networks[C]//Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH), 2011:437-440.

[15] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011(12): 2493-2537.

[16] Raykar V C, Yu S, Zhao L H, et al. Learning from Crowds[J]. Journal of Machine Learning Research, 2010(4):1297-1322.

[17] Zhu J, Nie Z, Liu X, et al. StatSnowball: a Statistical Approach to Extracting Entity Relationships[C]//Proceedings of International Conference on World Wide Web (WWW), 2009: 101-110.

[18] Koo T, Carreras X, Collins M. Simple Semi-supervised Dependency Parsing [C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Columbus, Ohio, June, 2008, 595-603.

[19] Chen W, Kazama J. Bitext Dependency Parsing with Bilingual Subtree Constraints[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala, Sweden, 2010, 21-29.

[20] Bansal M, Klein D. Web-Scale Features for Full-Scale Parsing[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL), Portland, Oregon, USA, 2011, 693-702.

[21] Wong Y, Mooney R. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus [C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), Prague, Czech Republic, 2007(6): 960-967.

[22] Kwiatkowski T, Zettlemoyer L S, Goldwater S, et al. Inducing Probabilistic CCG Grammars from Logical Form with Higher-Order Unification[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, October, 2010: 1223-1233.

[23] Liang P, Jordan M I, Klein D. Learning Dependency-Based Compositional Semantics [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), Portland, Oregon, USA,

2011, 590-599.

[24] Poon H, Domingos P. Unsupervised Semantic Parsing[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 2009,8: 1-10.

[25] Rahman, V. Ng. Coreference Resolution with World Knowledge [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), Human Language Technologies, 2011: 814-824.

[26] Lin Z, Kan M, Ng H T. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 2009,8: 343-351.

[27] Wang W, Su J, Tan C. Kernel-based Discourse Relation Recognition with Temporal Ordering Information [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala, Sweden, 2010, 710-719.

[28] Lin Z, Kan M, Ng H T. Automatically Evaluating Text Coherence Using Discourse Relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL), USA, 2011, 997-1006.

[29] Palmer M, Kingsbury P, Gildea D. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics,2005, 31(1): 71-106.

[30] Meyers A. Annotation Guidelines for Nombank—Noun Argument Structure for Propbank. Technical report, New York University. 2007

[31] Baker F, Fillmore J, Lowe B. The Berkeley FrameNet Project[C]//Proceedings of the the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL). 1998.

[32] Xue N, Palmer M. Annotating Propositions in the Penn Chinese Treebank[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03. Sapporo, Japan, 2003.

[33] Mann C, Thompson A. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization [J]. Text, 1998,8(3):243-281.

[34] Pradhan S, Xue N, OntoNotes: the 90% Solution [C]//Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Tutorial, 2009.

[35] Wu D, Fung P. Semantic Roles for SMT: A Hybrid Two-Pass Model[C]//Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2009.

[36] Liu D, Gildea D. Semantic Role Features for Machine Translation[C]//Proceedings of the conference of the International Committee on Computational Linguistics (COLING), 2010.

[37] Gao Q, Vogel S. Corpus Expansion for Statistical Machine Translation with Semantic Role Label Substitution Rules[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 2011.

[38] Oflazer K. Statistical Machine Translation into a Morphological Complex Language [C]//Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing), 2008.

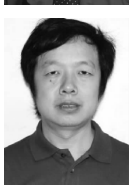
[39] Etzioni O. Search needs a shake-up[J]. Nature, 2011 (476): 25-26.

[40] Etzioni O. Anthony Fader, Janara Christensen. Open Information Extraction: the Second Generation[C]// Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2011.

[41] Schoenmackers S. Inference over the Web[D], Ph. D thesis, Washington University. 2011.

[42] Carlson A, et al. Toward an Architecture for Never-Ending Language Learning[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI), 2010: 1306-1313.

孙茂松(1962—),博士,教授,主要研究领域为中文信息处理、计算语言学、Web智能、社会计算。
E-mail: sms@tsinghua.edu.cn



姬东鸿(1967—),博士,教授,主要研究领域为中文信息处理、信息检索。
E-mail: dhji@whu.educ.cn

刘挺(1972—),博士、教授,主要研究领域为中文信息处理、社会计算、信息检索。
E-mail: tliu@ir.hit.edu.cn