深度学习在汉语语义分析的应用与发展趋势

王睿怡,罗森林,吴舟婷,潘丽敏
(北京理工大学 信息系统及安全对抗实验中心,北京 100081)

摘 要:人工智能分为感知和认知两个研究阶段。近年来,随着大数据技术和以深度学习为代表的机器学习技术的迅猛发展,人工智能在感知阶段进展飞速。然而,在认知阶段,尤其是在自然语言理解方面的发展仍较为有限。与人类丰富的语言经验、语言知识储备相比,仅仅依靠基于数据驱动的深度学习很难产生真正的智能。为了打破深度学习的性能瓶颈,必须将语义分析的理论和技术与深度学习模型相结合。因此,汉语语义分析理论和技术具有重要研究价值。汉语语义分析可以从海量的中文文本信息中挖掘语义信息,并提供智能的知识服务。文中主要描述了目前主流的汉语语义体系及其语义知识库的构建情况,介绍了汉语语义自动分析方法的研究进展和将汉语语义信息融入深度学习模型中的应用,最后对汉语语义分析的发展与态势进行了展望。

关键词:自然语言处理;深度学习;语义知识库;汉语语义分析;发展趋势
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2019)09-0110-07
doi:10.3969/j.issn.1673-629X.2019.09.022

Application and Development Trend of Deep Learning in Chinese Semantic Analysis

WANG Rui-yi, LUO Sen-lin, WU Zhou-ting, PAN Li-min
(Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Artificial intelligence is divided into two research directions: perception and cognition. Recently, with the rapid development of big data technology and machine learning technology represented by deep learning, artificial intelligence has progressed rapidly in the perception. However, it is still limited to develop cognitive intelligence, especially in natural language understanding. Compared to rich experience of language and language knowledge reserves, it is difficult to generate true intelligence by relying solely on data-driven deep learning. In order to have a breakthrough of deep learning, the combination of theory and technology of semantic analysis and the deep learning model must be promoted. Therefore, it is rewarding to do a research on Chinese semantic analysis theory and technology. Semantic information can be selected from a large amount of information in Chinese semantic analysis, which provides intelligent knowledge services. Based on the current mainstream Chinese semantic system and the construction of its semantic knowledge base, we mainly focus on the introduction of the research progress in Chinese semantic automatic analysis method and the application of Chinese semantic information in the deep learning model, which will provide an outlook of the developments and situations in future Chinese semantic analysis.

Key words: natural language processing; deep learning; semantic knowledge base; Chinese semantic analysis; development trend

0 引 言

人工智能的发展可分为感知智能和认知智能两个阶段。近年来,随着大数据技术和以深度学习为代表的机器学习技术的迅猛发展,人工智能在感知智能阶段进展飞速,在图像识别、语音识别等任务中均可达到人类专家的水平。然而,在认知智能阶段,尤其是在自然语言理解方面的发展仍较为有限。与人类丰富的语言经验、语言知识储备相比,仅仅依靠基于数据驱动的深度学习很难产生真正的智能。为了打破深度学习的性能瓶颈,尝试进行语义分析与深度学习模型的结合,

收稿日期:2018-10-31 **修回日期**:2019-02-26 **网络出版时间**:2019-04-24
基金项目:国家 242 信息安全计划(2017A149)
作者简介:王睿怡(1992-),女,硕士研究生,研究方向为文本安全、自然语言处理等;罗森林,博士,教授,博导,研究方向为网络安全、文本安全、媒体安全、数据挖掘;潘丽敏,硕士,高级实验师,通信作者,研究方向为网络安全、文本安全、媒体安全、数据挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1051.050.html

将成为人工智能在认知功能方面的下一个突破口。

为了将语言知识运用到机器学习的算法当中,首先需要将现有的语言知识量化为可直接与计算机应用相结合的量化模型,即开展语义体系、语义知识库构建等工作。研究者借鉴国外的经典语义理论,结合汉语自身的语义学基础,研究出适合中文的汉语语义体系。汉语语义知识库是通过利用汉语语义体系对原始语料库的加工、以形式化结构来描述汉语语言的一种语义资源库。例如,董振东的知网(HowNet)、袁毓林的论元系统、起源于格语法的谓词-论元结构、汉语语义依存分析和汉语句义结构分析等。

汉语语义分析是从海量的中文文本信息中挖掘语义信息,以此提供智能的知识服务。研究者选取特定的汉语语料、结合语义体系的标注规则来完成相应汉语语义知识库的构建工作,并结合统计知识进行汉语语义自动分析。早期,汉语语义分析遵循传统机器学习的步骤,即进行特征构建、特征抽取、特征选择和传统机器学习模型的训练。随着训练数据量的增大以及计算机计算能力的提高,研究者发现深度学习模型可以从大量原始数据自动提取构建特征,而不需要进行特征工程,并在特定领域任务中有很好的效果。因此,研究者开始尝试将深度学习模型应用到汉语语义自动分析的研究上,利用深度学习模型来自动提取有效的特征,从而完成汉语语义自动分析任务。

虽然目前深度学习模型在自然语言处理的多个任务中取得了不错的效果,但是深度学习模型的不可解释性以及缺乏标签数据的问题也一直无法得到解决。在深度学习模型中融合语义分析的基础研究,能够为任务提供更深层的语义先验信息,增强深度学习模型的可解释性和泛化性,让机器更好地理解人的语言,为人类提供更智能的服务。因此,研究者对在深度学习模型中融合先验语义信息、提高深度学习模型可解释性做了很多新的尝试,将融合多元知识库应用在深度学习模型中,为解决分析系统的可扩展性进行很多新的探索。

文中将按照汉语语义分析发展的主线,概要介绍汉语语义分析中的语义体系及其对应的语义知识库,重点阐述汉语语义分析的自动分析方法的研究情况,并介绍融合先验语义信息的深度学习模型的应用研究,最后对汉语语义分析存在的问题和发展进行分析和展望。

1　汉语语义知识库

研究者对汉语语义结构进行研究,得到各具特点的汉语语义体系,并希望通过这些语义体系制定的规则,将汉语的语义转换成计算机可处理的结构化信息。计算机想要通过这些结构化的语义信息学习到语义体系的规则,就需要通过统计学习的方法、利用大量的语义知识库来实现。因此,汉语语义体系的研究和语义知识库的构建至关重要。不少研究者一直致力于这两方面的研究,并获得了可喜的成果。例如,董振东开发的知网、袁毓林构建的中文网库、山西大学创建的汉语框架语义网库(Chinese FrameNet,CFN)、美国宾州夕法尼亚大学建立的中文命题库(Chinese proposition bank,CPB)、哈尔滨工业大学的语义依存树库和北京理工大学的汉语句义结构标注语料库(Beijing forest studio-chinese tagged corpus,BFS-CTC)。下面将对这些基于相应语义体系建立的汉语语义知识库进行介绍,其中汉语语义知识库对比分析如表 1 所示。

表 1　汉语语义知识库对比分析

汉语语义知识库	创建时间	创建者	标注对象	标注集	标签特点
知网	1999	董振东、董强等	义原和义项之间的语义角色	10 万种义项、2 000 个义原、90 个语义角色	利用义原和丰富细致的语义角色来描述概念,更加灵活、精细地对句中词语进行更好的解释和标注
中文网库	2007	袁毓林等	动词的论元角色	23 种论元角色	覆盖面广和精炼、每个论元的区别特征清晰
汉语框架语义网	2004	山西大学	语义框架中对应的框架元素	361 个语义框架、4 547 个词元	表达的语义内容丰富、深入,自然语言的语义适当且实用性强
中文命题库	2003	薛念文、Palmer 等	谓词的语义角色	6 种核心语义角色、10 余种非核心语义角色	标注简洁使得计算机进行自动分析更加容易
汉语语义依存树库	2011	哈尔滨工业大学	修饰词与核心词对间的语义关系	123 种语义关系、外加 20 种句法关系	标注清晰、全面,包含有非主要谓词包含的语义信息,如数量、属性和频率等
汉语句义结构标注语料库	2009	北京理工大学	句义成分和句义成分之间的关系	4 种句义类型、7 种基本格类型、12 种一般格类型、4 种谓词类型、3 种时态类型、4 种时态特征词位置	标注丰富、完整,能反映出细致的细节信息;不仅描述了谓词相关项的语义功能,还描述了项与项之间的关系

1.1 知 网

知网是董振东和董强组织建立的常识知识库。采用《分类体系》、《事件角色与典型演员》、《对义表》和《公理关系与角色转换》等多种理论作为它的理论基础。它的基本思想是以汉语和英语的词语所代表的概念为描述对象,并且揭示概念与概念之间以及概念所具有的属性之间的关系。知网利用义原和丰富细致的语义角色来描述概念,可以更加灵活精细地对句子中的词语进行解释和标注。然而,知网仅依赖单个词语的语义知识,没有考虑词语相互之间的关系。同时,标注过程不是在句法分析的基础上进行的,因而标注结果缺少句法关系的信息。

1.2 中文网库

中文网库是北京大学袁毓林教授在北大汉语句法分析树库的基础上,对新闻语义真实文本进行论元角色标注的语料库。采用国内外的论元结构理论、生成语法、格语法和配价语法作为构建该语料库的理论基础。袁毓林总共定义了 23 种论元角色,并根据这些论元角色提出对应的层级关系,如图 1 所示。与知网相比,中文网库的标注过程是在句法分析的基础上进行的,给标注结果增加了一定的句法信息;但由于中文网库只标注句法成分上标定动词的论元角色,因此对句子的语义分析结果缺少一定的完整性。

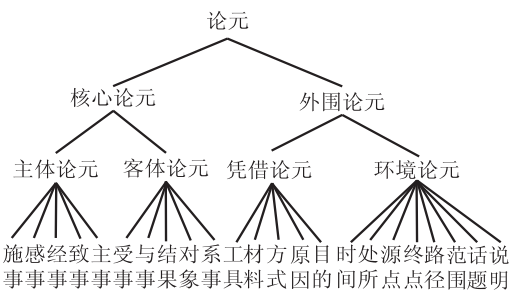

图 1 汉语动词论元角色的层级关系

1.3 汉语框架语义网

汉语框架语义网是由山西大学在 2004 年开始建立的、架构参照了英文框架网(FrameNet)的汉语词汇语义数据库。采用 Fillmore 的框架语义理论为其数据库构建的理论基础。该数据库用框架来描述词义、句子意义和文本含义,其中框架中的框架元素类似于语义角色。汉语框架样例如表 2 所示。汉语框架语义网的优点在于将框架与框架之间的关系展示出来,使语义的表达层次更加丰富,同时,框架元素表达的语义内容深入和实用性强。但是,汉语框架语义网对语义的刻画过于细致,给计算机完成框架元素的自动分析增加了难度。

1.4 中文命题库

中文命题库是薛念文和 Palmer 等基于“谓词-项”论元结构、参照英文命题库(proposion bank,PB)在宾州中文树库(Penn Chinese treebank,PCT)的句法分析树的基础上进行语义角色标注的语料库。中文命题库中一个句子的标注实例如图 2 所示。中文命题库的优点在于简洁的标注使得计算机进行自动分析更加容易。同时,它考虑了名词也可以作为谓词的情况,在一定程度上克服了论元结构仅以动词作为考察对象的缺点。但是,中文命题库只使用数个标记来表示语义角色,标记没有清晰的语义信息,使得语义角色不够丰富和统一,并且在标记时容易造成混淆。

表 2 汉语框架样例

框架名	量变
定义	该框架表示实体在某个维度上(即某属性)的相对位置发生变化,其属性值从初值变至终值
核心框架元素	实体(Ent),属性(Att),初值(Vall),终值(Val2),初状态(Inis),终状态(Finis),变幅(Diff),值区间(Val_ran)
非核心框架元素	环境条件(cir),倚变因素(Cor),动作时间量(Dur),倚变起点(Corl),倚变终点(Cor2),修饰(Manr),路径(Path),空间(Place),速度(Speed),时间(Time)
框架关系	父框架:无 总框架:无 后续过程:无 子框架:[增值] 分框架:无 结果状态:[数量] 参照:无
词元	波动 v,增加 v,增长 v,提高 v,减少 v,降低 v,上升 v,攀升 v,升 v,增 v,下降 v,降 v

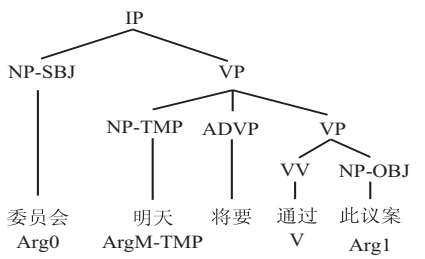

图 2 中文命题库中一个句子的标注实例

1.5 汉语语义依存树库

汉语语义依存树库是由哈尔滨工业大学的研究者们采用依存语义分析构建的能够完整地对句子语义进行分析的语义知识库。2011 年,哈工大社会计算与信息检索研究中心与北京语言大学合作推出了一套依存语义体系——HIT 语义依存。该体系是以依存分析为基础,将知网的语义框架与袁毓林、鲁川的语义体系相结合。汉语语义依存树库就是利用这套体系完成句子的标注,对句子进行深层的语义分析,从而更好地表达句子的结构信息和语义信息。

1.6 汉语句义结构标注语料库

汉语句义结构标注语料库是北京理工大学信息安

全与对抗技术实验室根据句义结构模型(Chinese sentential semantic model,CSM)构建的语料库。汉语句义结构模型以中文语言学家贾彦德提出的《汉语语义学》为理论基础、研究句子句义成分及各成分之间关系的句义结构表示模型。该模型分别由句型层、描述层、对象层和细节层组成,其中每一层所包含的句义成分如图 3 所示,句义成分之间的关系包含了谓词间关系、基本项和谓词之间的关系以及一般格与各句义成分之间的关系。句义结构模型不仅能够提供更为丰富的汉语语义特征,而且是一个能够完整地反映出句义成分以及成分组合关系的模型。

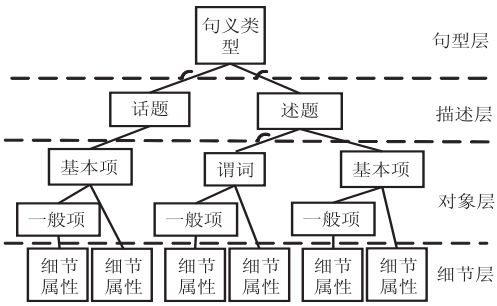

图 3　句义结构模型的基本形式

2　汉语语义深度分析

早期,汉语语义自动分析是运用传统机器学习方法自动分析汉语句子的语义结构。其中,在特征构建时,人工总结规律构建特征的过程必不可少。然而,随着深度学习的发展,研究者发现可以利用深度学习模型自动提取特征,从而取代传统机器学习中人工构建特征的步骤。同时,深度学习模型学到的上下文特征更加完备和有效,可以包含句子中更深层的含义。近年来,汉语语义自动分析的研究开始从人工构建特征进行传统机器学习的语义分析转向利用深度学习模型完成端到端的语义分析。

根据对语义分析程度的深浅不同,可以将汉语语义分析分为浅层语义分析和深层语义分析两种。浅层语义分析只要求标注与句子中的谓词相关的语义成分。深层语义分析不再以谓词为中心,而是将整个句子转化为某种形式化表示。两种语义分析方法的对比分析如表 3 所示。

其中,浅层语义分析方法主要有语义角色自动标注方法和框架元素自动标注方法,方法特点和研究进展如表 4 所示。

表 3　汉语语义自动分析方法对比

汉语语义自动分析方法	标注对象	常见方法	优点	缺点
浅层语义分析	与谓词相关的语义成分	语义角色自动标注、框架元素自动标注	标注简洁、易于自动分析和实现,可以让计算机更高效、直接地理解汉语的句子	容易丢失一些语义信息
深层语义分析	整个句子	语义依存分析方法、句义结构分析方法	能更好地表达句子的结构和隐含信息、挖掘句子更加丰富的语义信息	分析方法相对复杂,分析结果的准确率也相对难以提高

表 4　浅层语义分析的研究方法

研究方法	方法定义	研究者	具体方法
语义角色自动标注	给定动词,自动识别出对应的语义角色,并对其进行分类	Sun 等[1]	参照英文的方法,使用手工标记的短语结构树、利用支持向量机(support vector machine,SVM)进行训练和分类实现对中文语料的语义角色自动标注
		Xue 和 Palmer[2]	对中文命题库的最初版本,使用最大熵分类器进行分类,完成语义角色自动标注
		丁伟伟和常宝宝[3]	利用 CRF 在英文语料上能够利用论元之间的相互关系、提高标注准确率的特点,将其运用到中文命题库,使用 CRF 对中文语义组块分类也取得好的效果
框架元素自动标注	给定目标词以及目标词所属框架,自动识别出对应的框架元素,并对其进行分类	王蔚林[4]	采用最大熵模型把汉语框架语义角色标注问题转化为序列标注问题
		杨杏丽[5]	使用支持向量机作为分类器来标注汉语框架的语义角色
		刘开瑛等[6]	最早尝试用层叠条件随机场模型对汉语框架元素进行自动标注
		王智强等[7]	尝试基于树结构的条件随机场(tree structured conditional random field,TCRF)模型,通过在词与词性特征的基础上加入依存句法特征对核心框架元素进行自动标注

深层语义分析方法主要有汉语语义依存分析和汉语句义结构自动分析,方法特点和研究进展如表 5 所示。

表 5 深层语义分析的研究方法

研究方法	方法定义	研究者	具体方法
汉语语义依存分析	第一步,结合语义信息在依存句法结构的构建规则上稍作改动得到依存结构;第二步,利用 HIT 语义依存体系在弧上标注语义关系	李明琴等[8]	研究汉语语义依存,调整了知网的语义角色关系,同时提出一种统计语义分析器,完成了语义依存关系识别的任务
		王丽杰[9]	提出了基于图的自动汉语语义依存分析方法,利用哈工大构建的汉语语义依存树库完成了依存弧和语义关系的分析
		丁宇[10]	针对基于 HIT 语义依存构建的依存树受到树形结构的限制、丢失部分语义信息的问题,提出依存图的结构,并利用多分类方法识别出弧上的语义关系
汉语句义结构自动分析	对句义成分及句义成分的关系进行自动识别	罗森林等[11]	利用 C4.5 和 SVM 对句义成分中的 4 种句义类型进行自动识别
		王倩等[12]	基于谓词和句义类型块,利用 C4.5 对句子的句义类型进行识别
		韩磊等[13]	利用规则和 C4.5 完成句义成分中的谓词识别任务
		韩磊[14]	为提高谓词和句义类型的准确性,同时增加词关系识别功能,提出利用条件随机场完成句义结构中每一个环节的自动识别,同时提出基于转移的依存分析方法完成词关系识别。该套句义结构分析方法能够一次性完整地识别句义结构模型的句义成分和句义成分间关系,进一步推进汉语句义结构自动分析的研究

然而,传统的方法在系统性能上严重依赖于领域知识,并且需要人工选择特征来完成特征工程,同时,人工选择特征的有效性和完备性无法保证。随着训练数据量增大、计算能力提高,深度神经网络在多个自然语言处理任务上都取得了非常好的效果,因此也受到语义分析研究者的关注。研究者尝试将深度神经网络应用到浅层汉语语义分析的研究上,利用深度神经网络能自动提取有效的特征,解决浅层汉语语义分析在特征选择上的限制问题。例如,党帅兵[15]利用深层神经网络进行基本块识别,并将隐层向量作为基本块的分布表征,让其与角色识别任务的神经网络模型的中间层做级联,提高了汉语框架语义角色识别模型的标注性能;赵红燕等[16]利用深度神经网络自动学习目标词上下文特征,来提高语义角色标注中框架识别的准确率;王宇轩[17]对丁宇利用规则和 SVM 构建依存图的方法进行改进,提出一种基于转移的分析器,使用 list-based arc-eager 算法的变体对依存图进行分析,同时提出了两种有效的神经网络模块,分别用于获得转移系统中缓存和子图更好的表示。该系统在中英数据集上都取得了很好的效果,并且还能通过简单的模型融合方法进一步提高性能。尽管复杂的特征被设计,但是在句子中长距离的依赖关系很难被构建,因此有研究者尝试利用 BRNN(bidirectional recurrent neural networks,双向循环神经网络)解决语义标注中两个方向的依赖关系无法捕获的问题。Wang 等[18]利用了基于 LSTM 的 BRNN 完成中文语义角色标注任务,解决语义分析中长距离依赖关系难以构建的问题。

3 深度学习与汉语语义的结合

如今,深度学习在自然语言处理的多个任务中取得了不错的效果,深度神经网络的不可解释性以及缺乏标签数据的问题也随之暴露。因此,将先验语义信息加入到深度学习模型中,可以增强深度神经网络的可解释性和泛化性,让机器更好地理解人的语言,为人类提供更智能的服务。

下面主要介绍在深度学习模型中融合先验语义信息来提高深度学习模型可解释性的应用成果,以及融入多元知识库后,解决了单一特定标注集运用在深度学习模型中可扩展性受限的问题。

研究者对在深度学习模型中融合先验语义信息、提高深度学习模型可解释性做了很多新的尝试和探索。2017 年,牛艺霖等[19]在 word2vec 中的 Skip-Gram 模型的基础上提出 SAT(sememe attention over target model)模型。与 Skip-Gram 模型相比,SAT 模型不仅考虑了上下文信息,还考虑了单词的义原信息,借助义原信息使模型更好地“理解”单词,从而验证了分布式表示学习与义原知识库之间的互补关系。同年,谢若冰等[20]综合利用矩阵分解和协同过滤两种手段,利用词汇表示学习模型,对新词进行义原推荐,辅助知识库标注工作。2018 年,曾祥楷等[21]尝试利用词语表示学习与知网知识库进行词典扩展。通过实验表明,引入义原信息能够使层次分类效果得到提升。

同时,研究者还发现引入语义角色标签和标注模式不同、但表达潜在语义相同的异构数据(heterogeneous data)可以解决单一语义知识库规则不完备导致分析系统扩展性受限的问题。例如,2015 年,Wang 等[18]引入异构数据—中文网库来预训练词向量。他们基于中文网库学习 LSTM-RNN 模型,利用从中文网库中获得的预训练的词向量来初始化一个新模型,最后用中文命题库来训练。实验结果表明,该方法引入异构数据解决了单一标注集扩展性受限的问题。2016 年,Li 等[22]利用 RNN 模型做汉语语义角色标注。他使用英文主题库去提高汉语语义角色标注的性能。实验结果表明相对于先进的方法有显著提升,F1 值能达到 78.39%。2017 年,Xia 等[23]提出一种渐进式的神经网络模型(progressive neural network,PNN),并发布了一个新的中文语义角色标注数据集——Chinese SemBank 作为异构数据。PNN 模型能够充分地容纳和利用异构数据更好地完成语义角色标注任务。

4　发展趋势

汉语语义分析发展日趋成熟,但对它的研究还有很多值得深入探索的问题。在该部分,根据目前的研究现状指出汉语语义分析存在的问题,并对其改进方案和发展趋势作简要的介绍。

(1)目前,汉语语义知识库已经有足够大的规模,但随着信息时代的日新月异,汉语语义知识库需要相应的改变和适当的扩展。然而,知识库在不断更新的过程中,容易出现标注不一致的现象。因此需要探索以深度学习为代表的数据驱动和以知识库为代表的专家驱动相结合的技术,让计算机能够辅助人类专家更及时高效地完成标注知识库的工作。并且,在不断优化和扩充语义体系的同时,也能提高人类专家标注知识库的一致性。

(2)在语义分析模型训练的过程中,数据收集昂贵,并且只用一份标注规则相同的语料库训练模型是对语料库的浪费。因此,在后续工作中,考虑将主动学习应用于深度汉语语义分析任务中,从而大幅减少达到最先进结果所需的数据量。同时,也可以考虑将多种语义知识库进行融合,训练得到语义信息更加丰富的模型。这种通过融合不同知识库的语义信息来提高汉语语义自动分析系统性能的研究将成为语义分析的下一个研究热点。

(3)目前,通过结合深度学习模型,汉语语义分析效果有明显提升,利用深度学习模型自动提取特征取代了传统机器学习中需要人工构建特征的过程,提升了特征选择的有效性和完备性。同时,随着注意力机制在自然语言处理任务中的广泛应用,尝试利用注意力机制学习更多标签潜在的依赖信息,从而提升语义分析的效果。这也将成为今后研究的热点。因此,在标注语料达到一定规模的情况下,使用深度学习模型自动提取特征进行语义分析将成为汉语语义深度分析的研究趋势。

(4)分布式表示(distributed representation)在可解释性方面能力较弱,另一方面,利用端到端(end-to-end)框架训练得到分布式表示的效率较低且需要极大的训练语料。因此,在利用深度学习框架完成语义分析任务时,仍然需要加入语义知识库来为系统提供更多的先验知识,从而提高系统的分析效率和结果的可解释性。因此,如何在分布式表示中引入语义知识库作为先验知识是未来的重要挑战性问题。同时,如何利用先验知识实现无监督学习,使得较少标注数据通过先验知识的加入也可以训练出很好的模型,也将成为汉语语义分析中新的发展趋势。

5　结束语

文中在充分调研和深入分析的基础上对汉语语义分析的研究进展进行了总结。对目前常用的汉语语义知识库,如知网、中文网库、汉语框架语义网、中文命题库、汉语语义依存树库以及汉语句义结构标注语料库进行了说明;在对汉语语义自动分析方法的研究中,依据对句义分析的深浅程度的不同,将分析方法分为浅层语义分析和深层语义分析两种方法。对这两种方法的特点和研究进展进行列举,指出存在的问题,并对运用深度学习模型自动提取特征完成语义分析的方法进行介绍。在汉语语义分析的应用中,主要介绍了在深度学习模型中融合先验语义知识提高深度学习模型可解释性的应用成果,以及融入多元知识库后,解决了单一特定标注集运用在深度学习模型中的可扩展性受限的问题。最后,指出目前汉语语义分析存在的问题,对每个问题提出可行的解决办法,并对深度学习与汉语语义分析结合的应用进行了展望,希望对该领域的其他研究者有所启发。

参考文献:

[1] SUN Weiwei. Improving Chinese semantic role labeling with rich syntactic features[C]//Proceedings of ACL. Uppsala, Sweden:ACL,2010:168-172.

[2] XUE Nianwen, PALMER M. Automatic semantic role labeling for Chinese verbs[C]//Proceedings of international joint conference on artificial intelligence. Edinburgh, Scotland:Morgan Kaufmann Publishers Inc.,2005:1160-1165.

[3] 丁伟伟,常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报,2009,23(5):53-61.

[4] 王蔚林. 基于最大熵模型的汉语框架语义角色自动标注[D]. 太原:山西大学,2010.
[5] 杨杏丽. 基于支持向量机的汉语框架语义角色自动标注[D]. 太原:山西大学,2010.
[6] 刘开瑛,陈雪艳,李济洪. 汉语框架元素自动标注实验报告[C]//全国信息检索与内容安全学术会议. 北京:清华大学出版社,2008:48-55.
[7] 王智强,刘海静,李双红,等. 基于 TCRF 的核心框架元素标注[C]//第五届全国青年计算语言学研讨会. 武汉:中国中文信息学会,2010:812-820.
[8] LI Mingqin, LI Juanzi, WANG Zuoying, et al. A statistical model for parsing semantic dependency relations in a Chinese sentence[J]. Chinese Journal of Computers, 2004, 27(12): 1679-1687.
[9] 王丽杰. 汉语语义依存分析研究[D]. 哈尔滨:哈尔滨工业大学,2010.
[10] 丁 宇. 基于依存图的中文语义分析[D]. 哈尔滨:哈尔滨工业大学,2014.
[11] 罗森林,王 倩,刘莉莉,等. 融合 C4.5 与 SVM 算法的汉语句义类型识别方法[J]. 北京理工大学学报,2012,32(10):1036-1041.
[12] 王 倩,罗森林,韩 磊,等. 基于谓词及句义类型块的汉语句义类型识别[J]. 中文信息学报,2014,28(2):8-16.
[13] 韩 磊,罗森林,潘丽敏,等. 融合词法和句法特征的汉语谓词高精度识别方法[J]. 浙江大学学报:工学版,2014,48(12):2107-2114.
[14] 韩 磊. 汉语句义结构模型分析及其文本表示方法研究[D]. 北京:北京理工大学,2016.
[15] 党帅兵. 基于词分布表征的汉语框架语义角色识别研究[D]. 太原:山西大学,2015.
[16] 赵红燕,李 茹,张 晟,等. 基于 DNN 的汉语框架识别研究[J]. 中文信息学报,2016,30(6):75-83.
[17] WANG Yuxuan, CHE Wanxiang, GUO Jiang, et al. A neural transition-based approach for semantic dependency graph parsing[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. New Orleans, Louisiana, USA: AAAI, 2018:5561-5568.
[18] WANG Zhen, JIANG Tingsong, CHANG Baobao, et al. Chinese semantic role labeling with bidirectional recurrent neural networks[C]//Conference on empirical methods in natural language processing. Lisbon, Portugal: ACL, 2015: 1626-1631.
[19] NIU Yilin, XIE Ruobing, LIU Zhiyuan, et al. Improved word representation learning with sememes[C]//Proceedings of ACL. Vancouver, Canada: ACL, 2017:2049-2058.
[20] XIE Ruobing, YUAN Xingchi, LIU Zhiyuan, et al. Lexical sememe prediction via word embeddings and matrix factorization[C]//Proceedings of the twenty-sixth international joint conference on artificial intelligence. Melbourne, Australia: AAAI Press, 2017:4200-4206.
[21] ZENG Xiangkai, YANG Cheng, TU Cunchao, et al. Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. New Orleans, Louisiana, USA: AAAI, 2018:5650-5657.
[22] LI Tianshi, LI Qi, CHANG Baobao. Improving Chinese semantic role labeling with English proposition bank[C]// China national conference on chinese computational linguistics. Yantai, China: Springer International Publishing, 2016: 3-11.
[23] XIA Qiaolin, CHANG Baobao, SUI Zhifang. A progressive learning approach to Chinese SRL using heterogeneous data[C]//Proceedings of f the 55th annual meeting of the association for computational linguistics. Vancouver, Canada: ACL, 2017:2069-2077.

(上接第 54 页)

[5] 陈振洲,李 磊,姚正安. 基于 SVM 的特征加权 KNN 算法[J]. 中山大学学报:自然科学版,2005,44(1):17-20.
[6] 周 靖,刘晋胜. 一种采用类相关度优化距离的 KNN 算法[J]. 微计算机应用,2010,31(11):7-12.
[7] 杨 立,左 春,王裕国. 基于语义距离的 K-最近邻分类方法[J]. 软件学报,2005,16(12):2054-2062.
[8] 余小鹏,周德翼. 一种自适应 k-最近邻算法的研究[J]. 计算机应用研究,2006,23(2):70-72.
[9] 江 昆,白旭英,车金星. 基于随机森林的 K 最近邻算法[J]. 南昌工程学院学报,2016,35(6):99-102.
[10] CHEN Yewang, ZHOU Lida, TANG Yi, et al. Fast neighbor search by using revised K-D tree[J]. Information Sciences, 2018,472:145-162.
[11] 杨金福,宋 敏,李明爱. 一种新的基于距离加权的模板约简 K 近邻算法[J]. 电子与信息学报,2011,33(10):2378-2383.
[12] 肖辉辉,段艳明. 基于属性值相关距离的 KNN 算法的改进研究[J]. 计算机科学,2013,40(S2):157-159.
[13] 孙 新,欧阳童,严西敏,等. 基于训练集裁剪的加权 K 近邻文本分类算法[J]. 情报工程,2016,2(6):8-16.
[14] 张 宇. K-近邻算法的改进及实现[J]. 电脑开发与应用,2008,21(2):18-20.
[15] KAHRAMAN H T. A novel and powerful hybrid classifier method:development and testing of heuristic k-nn algorithm with fuzzy distance metric[J]. Data and Knowledge Engineering, 2016, 103:44-59.
[16] HUPP A M, PERRON J, ROQUES N, et al. Analysis of biodiesel-diesel blends using ultrafast gas chromatography (UFGC) and chemometric methods: extending ASTM D7798 to biodiesel[J]. Fuel, 2018, 231:264-270.
[17] CHEN Jindong, TANG Xijin. The distributed representation for societal risk classification toward BBS posts[J]. Journal of Systems Science and Complexity, 2017, 30(3):627-644.