基于本体语义网络的语言理解模型

王 飞[1] 易绵竹[1] 谭 新[2]
(信息工程大学洛阳校区 河南 洛阳 471003)[1] (91709 部队 吉林 珲春 133300)[2]

摘 要 传统的知识表示存在涵盖知识面不够和语义形式化描述不够全面的问题,导致计算机理解自然语言不够准确。受大脑神经元工作原理的启发,从语义剖析的角度出发,基于本体语义,在概念和词汇两个层次构建了本体语义网,使其具有神经网络的特性,既能准确理解文本语义,刻画词在不同领域内的不同含义,又涵盖了文本生成过程中的语义组合特点。为使模型进一步形式化,采用矩阵的方式表示,并用奇异值分解来降低矩阵规模复杂度,以便于描述词汇与概念之间的关系。

关键词 概念,词汇,神经网络,本体语义,矩阵
中图法分类号 TP391 **文献标识码** A

Language Understanding Model Based on Ontological Semantics Network

WANG Fei[1] YI Mian-zhu[1] TAN Xin[2]
(Information Engineering University Luoyang Campus, Luoyang, Henan 471003, China)[1]
(91709 Troops, Hunchun, Jilin 133300, China)[2]

Abstract The traditional knowledge representation has limited scope of knowledge and incomprehensive formal semantics description, thereby causing the computer's accurate portrayal of natural language. This paper proposed ontological semantics network on the levels of concept and lexicalfor semantic analysis. Its brain-like neural language network drawing from the inspiration of human brain neural cell's work principle, could both accurately portray different meanings of a word in different domains, understand the text meaning and cover the elements and characteristics in the process of words's generating into sentences. Matrix is employed to further formalize the model, with singular value decomposition to reduce the scale complexity, which makes it more convenient to describe the relationship between lexical semantics.

Keywords Concept, Lexical, Neural network, Ontological semantics, Matrix

1 引言

客观世界中的事物是离散的,相互之间是独立的,人的认知却是连续的、结构化的,未形成结构化的表示会逐渐被遗忘。在认知体系中,人将客观世界映射到连续的思维空间中表示成概念,概念并非离散孤立地存在于人类的思维中,而是相互联系的。这些联系体现为多种语义关系,语义关系是二元的,从一个语义节点到另一个语义节点,这样就将连续的意义表示转化为离散的节点之间的关系。概念和词汇通过语义联系起来,形成知识,并作为"记忆"储存起来。概念属性与概念间的关系是语义中最关键的特征,它们以适当的概念组织形式将离散的客观世界连接起来,从而建立一个连续的低维语义表示模型,成为了人类认知的基础。

人类使用自然语言描述世界,在脑中构建世界模型。当人象征性地表示对象和事件时,使用的是心理模型,也就是概念;但自然语言表述的复杂性与不确定性,使得计算机难以理解。计算机要想对现实世界进行建模,就必须用形式化的语言对现实世界进行抽象表示。客观世界中的某个对象在人脑中形成概念,概念不会因不同的人或不同的描述语言而具有不同的内涵。这种思维共性形成了基本的概念系统,也就是语义系统,它决定了不同语言文化群体对同一对象的认知也是基本相同的,这说明不同的语言具有语义共性。概念(本体)是将自然语言表示(词典)与客观存在(实例)联系在一起的中介[1]。人工智能领域的本体应当是"存在"与外部世界的联系构成的,不仅符合人类认识世界的基本规律,也使计算机能够理解语义[2]。基于此,本体语义学提供了一种恰当的概念表示形式,将客观世界抽象成一个概念层次体系,建立语义表示框架,通过语义属性准确地描述事物特征,能够对整个世界或者某一领域的知识进行描述[3]。这种形式化的描述能够让计算机理解,从而实现计算机智能。基于本体概念结构和词汇表示的关系,本文提出了类似于神经网络的语言理解模型,使得计算机能够对语义进行分析和理解。

2 相关研究

本体语义所建立的心理智能模型与互联网模型、CPU 类似,都是受大脑(神经网络)启发而建立的。大脑实质上是一

本文受国防科技创新特区项目:面向开放数据的大规模知识图谱构建及其应用(17-H863-01-ZT-005-008-03),国家自然科学基金项目:多语言言语数据获取、标注与分析研究(11590771)资助。

王 飞(1983-),男,博士生,助理工程师,主要研究领域为自然语言处理、数据挖掘,E-mail:89738764@qq.com(通信作者);**易绵竹**(1964-),男,博士,教授,主要研究领域为自然语言处理,E-mail:mianzhuyi@gmail.com;**谭 新**(1994-),男,助理工程师,主要研究领域为自然语言处理,E-mail:1154338328@qq.com。

个神经元网络,由大约 1000 亿个神经元构成[4]。它通过大量神经元激活和抑制单元,形成不同层次的认知。由于一个神经细胞通过多个树突与其他多个神经细胞相连,从这个神经细胞传出的信号实际上是与其连接的多个神经细胞的信息融合后发出的,传入这个神经细胞的信号也会传遍与其相连的每个神经细胞[5]。复杂的信息通过辐射、聚合、链锁和环状等多种神经元联系方式,在巨大的神经网络系统中得以迅速处理[6]。基于神经网络这种对离散世界的连续表示机制,大脑具备了高度的学习能力和智能水平。在对世界的认知过程中,大脑是无法存储所有的现象的,而是通过分析少量数据进行推理,进而泛化,才产生了大量的描述。

近年来,针对如何模拟大脑对客观世界的认知过程以及对语言的理解和描述,用计算机能够理解的方式表示词汇语义的问题,研究者们进行了积极探索。一方面要求全面涵盖各个领域,避免歧义的产生;另一方面需要清晰且准确的形式化表征。受大脑和互联网的启发,高效的计算和智能化发展需要具备两个关键要素:分布式存储和层次化[7]。从人类认知的角度出发,依照大脑和互联网的结构和特点,基于本体语义理论构建了具有两个层次的静态知识源,形成本体语义网,模拟神经网络的结构,包括语言无关的概念层(本体和事实数据库)和语言相关的词汇层(词典和专名库),层次间通过语义关系相互映射,层次内通过语义属性相互关联,描述更加明确、清晰,具有很好的知识表示逻辑性和技术实现可操作性。人的记忆是按列表模式存储的,本体按照结构化模式存储,并且模拟人的认知进行学习和保存记忆,经过对特征的多重选择最终学习到最佳组合,完成识别和预测任务。

3 层次化理解过程

人对事物的认知过程其实是一种模式识别。人具有很强的模式识别能力,这种模式识别是分层级的,并形成语言表示的基础。本体语义学对文本的理解模型就是对词汇语境捕获、结构化、形式化和认知计算的过程。理解模型的主体就是具有两个层次的静态知识源,在两个层次内部,概念与概念通过属性相互连接形成具有分布式特点的本体网络,词汇也通过句法模式的组合形成分布式词汇网络,两个层次之间通过词汇到概念的映射表示文本语义。概念和词汇内部还会再分层次,每个层次都包含一种模式。低层模式识别都向高层模式识别输出,每个输入/输出节点都有“权重”来表示重要性。模式中的因素越重要,考虑是否触发该模式被识别的权重越大。层次知识结构对概念和语言的正确识别很重要,人的语言就具有复杂的层次结构。

3.1 概念层

概念层的本体是语言无关的。本体正是从人类认知角度出发构建的心理模型,它通过一组概念和属性互相连接来描述整个世界模型,要求建模者必须寻找基本概念,使用尽可能少的基本概念来以组合方式构建复杂对象和过程的描述。需要注意的是,基本概念集规模太小,模型简单,但会影响对描述的完整性和明确性;而基本概念集太大,意义表示的完整性和明确性能得到保证,但是模型也会包含太多的冗余信息。因此,根据任务的需要,在明确性和完整性不受影响的情况下,基本概念的选择应当尽可能地简单。概念的属性和关系能够体现基本概念集的语义关系,在基本概念组合成复杂对象时可以通过继承推演得到新对象的属性。本体具有相对较小的原始概念集,但同时具有强大的继承特性和推理能力,使得本体在演化的过程中内容会越来越丰富[8]。如果用神经元节点表示一个概念,则每个神经元包含多个树突和一个轴突,神经元接收来自多个树突的输入,向一个轴突输出。树突对应为每个概念与多个概念相关联,说明一个概念可以作为其他概念的属性存在,轴突对应于概念之间存在语义关系或者与词汇的映射关系,整个本体就构成了一个有向无环的网络。

3.2 词汇层

词汇层是语种相关的,专名库包含的专有名词可以看作是词典的一部分。词汇的特征比较突出,同属于一个语义类的词语,其内部属性及其搭配要求却存在很大差别。语义类能够满足解释上的充分性,但不能满足描述上的充分性。因此,为了更好地为自然语言处理系统提供高性能和高准确度的保障,在构造词汇层时,应当在语义类的基础上依靠属性描述来刻画每个动词和与之相联系的名词之间所建立的句法模式,即要建立面向自然语言处理的语义词典。由于自然语言文本类型多变、样式复杂,在本体语义理论中,所有的意义都被归结为了形式化的概念,数量也是有限的,本体语义文本处理可以将多个同义的文本表示减少成唯一的形式意义表示。同时,有些功能上多余的措辞也必须从输入文本的意义中去除,以使意义表示与存储的记忆相匹配。与概念层类似,每个神经元节点表示一个词汇,不同的是词汇节点不区分树突与轴突,每个词汇通过树突与多个词汇相连,这部分词汇与当前词形成句法模式,表示词与词之间具有组合和聚合关系,表现为词汇的句法语义属性;轴突与概念相连,表示词汇的含义通过与之相连的概念传递,构成了组词成句的基础。最终输出的轴突代表着模式,轴突激活后,相应的模式就被识别了。由于词汇语义特征具有多样性,不同概念激活的词的内涵也不同,从而形成了一词多义现象。同时,其也描述了含义越丰富的词所连接的词汇节点也越多,使用越频繁的词所连接的词汇节点也越多。词汇网络的分布式表示从理论上很好地解释了深度神经网络训练得到的分布式词向量[9]。

3.3 本体语义网

文本或者句子表示在概念层面就是 TMR,计算机理解或者生成语义的过程都离不开 TMR。本体以概念为中心,概念之间通过属性互相连接,为语言理解和生成提供了语义框架,这在本体语义系统中起着基本作用,它能够通过各种复杂的语义组合对事件和对象进行表示,并且其语法被特别设计以便于表达复杂的词汇含义,从而使得本体能够使用尽可能少的基本概念来以复合方式构建复杂对象和过程的描述。语义框架通过语言的实例化表示就形成了 TMR,本质上是语义框架构成的网络,每个框架具有中心和多个属性值对,一个框架通过属性值链接到另一个框架,从而形成语义网络。本体语义通过这种概念和词汇的分布式表示既能够准确描述自然语言的语义,又符合人类的认知模式。

要理解、建模和模拟大脑的工作模式,就须建立分层结构。大脑的记忆是分层且连续的。人在理解一句话的含义时,首先利用的是已掌握的知识,然后才是经验。本体语义模拟人的理解,理解时首先通过本体语义分析技术去理解输入,按照本体语义分析流程和推理算法(包括共指消解的推理和话语行为的解释)识别句子中心,激活相应的概念节点,根据

句法模式将句子分解成词汇单元,通过语义框架的选择限制为词汇单元分配语义角色,得到最佳的意义组合[10],并将其形式化表示后作为实例存储在事实数据库中,不断为整个本体语义系统创造新的记忆,在语义生成时使用。当有新的句子输入时,就需要与系统自身存储的"记忆"匹配。若语义分析无法解释,则说明知识不足,这时需要根据经验(以数据驱动的方式)以一定概率进行分析,得出结论,该过程可被看作是模式识别。概念层和词汇层共同构成了本体语义学的静态知识源,这个知识源就相当于人的"记忆"模型,或者说已掌握的知识。

句子生成的过程与理解相反,大脑首先产生意识(概念),或者称为想法,这种想法就是一种模式。想法产生后才会通过语言进行表述。简单模式由单个概念形成,复杂模式由多个概念组合而成。概念层接收到输入信号,然后会刺激到一系列复杂的递归链接算法,多个层次的概念节点被激活,组合成一个复杂的语义框架,由映射关系输出至词汇层,相应的词汇节点被激活,词汇代替概念填充到语义框架中,这就是模拟人的认知过程。相应地,语义框架还激活了中心节点句法模式,更多的词汇按照句法模式的链接被概念激活,作为中心节点的语境。不同的模式被识别出来后就会通过轴突发送信号,轴突会与下一个更高层级的模式建立连接,输出变成下一个模式识别的输入,当多个概念输出到词汇层后,激活词汇的过程是并发执行的,同时被激活的词汇通过句法模式和语义关系相连,最终形成了句子,句子组成段落,段落组成篇章。图 1 形式化地描述了该模型。

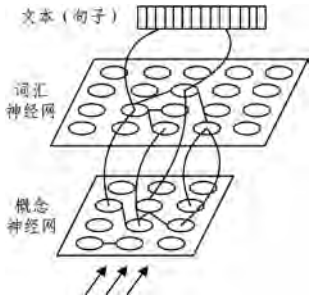

图 1　本体语义网络模型描述示意

整个模型由两层节点组成,这些节点在理解和生成语义时被"触发",其工作原理类似于大脑或者 CPU 的工作模式。本体语义网络的概念神经网络包含了概念和实例的语义框架,好比是人的记忆或者 CPU 的存储器,利用存储的相关知识来识别和分析问题的性质和范围,分析的过程类似于人的思维活动或运算器工作。词汇、句法代表了意义的表层形态,既有助于分析输入,又完成了输出表示。三者之间的类比关系如图 2 所示。

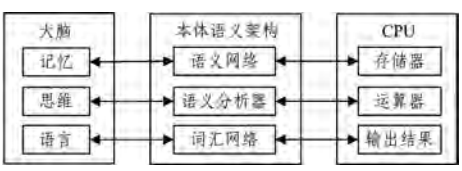

图 2　语义处理结构的类比关系

4　本体语义的网络特性

作为对大脑计算机制的一种模拟,本体语义网具有与大脑、互联网相似的特性,通过这些特性描述,能够更好地构建和理解本体语义网,并做出更好的开发和应用。

4.1　聚类性

概念与词汇构成了大脑中的两级网络,由内在的语义关系将节点与节点相连,甚至是跨层级的节点相连。在大脑中,神经元通过自我开关来工作,语义网络通过外界输入的信息激活相关节点来工作。语义相关度高的词汇连接在一起形成模式(子网),或者称框架。例如,词汇"网球""网球拍"和"网球场"在语义上具有相关性,相互之间通过边联系,构成了一个小的模式。类似的例子还有"厨房""餐具""烹调"和"食材"等词汇。宏观上看,整个网络又可以看成是模式与模式相连的网络。整个网络就像一个大的图,每个特征或者词汇就是其中的一个节点,节点之间通过边相连,节点之间会受到语言内部层级的相互影响。这种连接具有特定的语义模式,在工程实践中,一般使用矩阵对图结构进行表征。

4.2　演化性

概念层与词汇层不是固定不变的,而是动态发展的。纵向来看,人在婴幼儿时期所掌握的词汇量很小,因此脑中的词汇网络的规模很小;随着后天知识学习,人脑词汇网络的规模会越来越大,并最终到达一个成熟、稳定的状态。横向来看,成熟、稳定状态的词汇网络依然是发展变化的,但是处于一种动态平衡。大致来说,受外界知识和新事物的影响,有的词汇会加入进来,形成了新的节点,与周围的词汇建立新的连接;有的词汇因为各种原因而用法受限,或者由于遗忘从词汇网络中消失。大致来说,这是一个创造到消亡的过程,大脑创造了一些新的思想,重要的思想被保存了下来,创造了一些新的词汇节点,无关重要的思想被过滤掉,同时这些思想所表现出的词汇节点也就消失了。思维的不断更新产生了创造性思想,使得整个网络都发生了变化。整体来看,网络结构处于一个稳定状态,但细分到一个个子网中时都是在不断变化的。例如,单词"dog"曾经有猎犬的含义,随着时间的推移,含义渐渐被放大,今天就只指称普遍意义的"狗"。随着社交网络的发展,许多网络新词又会诞生出来,例如"给力"等。

4.3　等级性

词汇网络中的每个词由于使用频率和含义丰富程度的不同,表现为节点的连接数也不同,高频词和多义词的节点较多,连接数较大。其中,词与其他词结合形成更大语言单位的能力表现在它的连接数上,这一点由配价理论决定。这种能力在语句中受句法、语义和语用等因素的约束。图 3 是词连接的示意图。

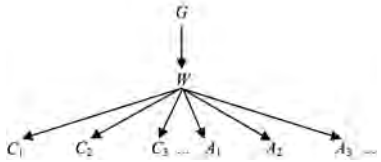

图 3　词节点连接示意图

图 3 中的 W 表示一个词,C_1-C_3 是为了完善或明确 W 的意义所需要的补足语,A_1-A_3 是可对 W 做进一步说明和限定的说明语,G 为 W 潜在的支配词[11]。一个词的结合能力,可以分为支配词(输出,激活)和被支配力(输入)两种:支配力是它支配其他词的能力,被支配力表示该词受别的词支配的能力。一旦 W 出现在真实的文本中,那么它就打开了一些需要填补的空位,对此提出了所需补足语的数量和类型。

同时，W 在进入具体文本时也显现了它是否能满足别的词支配的需要。至于真正的词汇组合是否合理，要看句法、语义等方面的结合能否满足要求，这样句法、语义特征限制也就成为了整个组词成句的一部分。在词汇组合成句的过程中，应该按照句法模式、语义角色进行填充，同时需要语义选择限制的约束，才能构成合理的、有具体含义的句子。受 PageRank 算法[12]思想的启发，可以认为一个词的重要程度与它连接的节点数以及连接的节点词汇的关联度有关，如果与 A 节点相连的 B 节点也同样是一个高频词，那么 A 节点词汇也是一个重要性较高的词。如果一个词被很多其他词连接，则说明它受到普遍的认可，用途广泛，它的地位也就较高。例如，动词“take”“play”和“get”它们的组合能力比较强，表现在网络中就是该词节点连接了更多的词节点，它们在网络中就处于一种“区域中心”的地位。对来自不同词汇节点的链接要区别对待，那些地位高的词汇节点更重要，于是需要给这些链接赋予较高的权重。宏观上看，词汇节点的重要程度来自与它连接的子网的大小。每一个节点可以看作是计算机网络中的路由器，路由器又连接着另一个网络或子网。所有的节点通过网络连接起来，变化出无限的信息，并按照人的想法灵活地组织起来。

4.4 可计算性

文本理解与生成的过程中，概念的触发、词汇的激活都是并行的，同时根据语义框架和句法模式等内在关系对概念属性和词汇间的复杂计算，最终形成合理的句子，计算过程就是大脑思维的过程。可计算意味着可被形式化表征，自然语言在大脑中存在着计算性，这一点可以从人能识别两个相似的句子得出。当看到两个句法结构不同而语义相似的句子时，大脑先将两个句子抽象成概念和含义，再通过复杂的计算判断其是否相似，得出结论。本体语义网通过语义的形式化表征也具有了可计算性，语义计算可以看作是矩阵的乘法。大多数(开放类)词汇的语义被分配给一个或多个本体概念，概念的输入激活相应的属性和词汇，一组相关的属性组成某种语义的向量，这构成了一种语义的分布式表示，每个词可以通过刻画它的各种属性来高效表示，属性又与多个概念相关联。每个词既包含从上位词继承来的公有属性，又包含自身的私有属性。理解与生成的过程就是基本属性的激活和矩阵相乘的运算。本体语义网将词汇映射到语义空间中，每个词汇用向量表示，属性就是构成词汇的各个维度的值。词汇之间的相似度可以通过计算词汇向量之间夹角的余弦值来刻画，例如，“岛屿”与“岛礁”的距离相对较近，而“岛屿”与“机场”的距离相对较远。这样，词与词之间的关系都可以在语义空间中通过计算表示。

4.5 传递性

一个词被触发后，与之相连的句法模式中的词都被语义框架激活，共同构成一个上下文环境，或者称之为语境。对于具有多种含义的词汇而言，不同的含义被不同的概念触发，所激活的上下文也是不同的，能够准确描述词汇意义的概念和句法模式中的其他词汇被“高亮”显示，并通过关系和属性连接起来，从而消除了歧义。例如，当词汇“ask”被激活时，它包括“提问”“请求”等多个含义，而激活它的概念是“REQUEST-ACTIVITY”，根据概念映射就能选择出合适的词义。“说话”被激活时，要求施事具有“生物”的语义属性，如果施事具有多个选择，就可以根据语义选择限制得到正确的意义。在分布式假设中，拥有相似上下文语境的词，它们的词义也相似。从搭配的角度来看，能够相互搭配的词在语义上也是相容的，并且句法会对远距离的搭配有一定的限制。因此，每一个词节点与其周围一定范围内的词节点除了相似性，还具有一定的关联性，共现的频率比较高。这种关联性随着激活的词汇而向外扩散传递出去，又有与新激活的词节点关联的词被激活，最后构成句子。

4.6 预测性

大脑真正能体现智能的地方就是预测能力，它不会凭空猜测，很大程度上依靠的是经验，即通过已具备的知识来推测产生某一结果的概率。语言神经网中，事实的积累记录了已经发生的事情，从而在一定程度上也能够预测未发生的事，类似于人的“经验”。为了模拟这一决策过程，词汇网中需要设定一个阈值(T)，当大脑要表达一个预测结果时，根据经验进行预测表达，根据结果调整节点之间的连接，如果表达正确，脑中的算法就会“鼓励”节点之间的连接，所用到的词节点也会保存记忆，该节点和到该节点的连接就会加固；相反，如果预测表意错误，算法就会放弃惩罚部分节点之间的连接。如果出错次数达到阈值，则连接到表意错误节点之间的连接就会被撤销，下次进行预测时就不会选择这条路径，也就不会用该词表达。这种机制不断完善和修正着人在使用语言时尽可能地选择正确的词，也就是强化学习的过程。例如，句子“卧室的地板上趴着一只__。”根据经验，语言神经网就会预测空位该填入的词可能是“猫”或者“狗”，而不太可能是“老虎”或者“狮子”，显然“猫”或者“狗”出现在这个句子中的概率要大于“老虎”或者“狮子”，类似的现象在其他词类中也有很多。拥有足够的知识进行推理预测，才能使计算机更具有人的“智慧”。

5 语言神经网的矩阵表示

5.1 模型的形式化表示

本体语义网连接关系具有明显的聚合性，所有的神经元聚合成多个簇。簇内实现高密度连接，簇间的连接密度较稀疏。在不影响计算的条件下，神经网络模型可以转换成矩阵表示，描绘出每一类词的属性状况，矩阵间的转换通过激活函数作用完成。这种网络结构所形成的是一种非线性的计算过程，这也恰好符合计算机指数级的计算增长的过程。

矩阵 M 描述了上百万个词汇与数千个概念的关联性，它的行表示词汇，一行表示一个词向量，因为概念取不同值组合在一起可以生成更多的词汇。列表示概念，描述同一语义类，概念组合起来构成领域，整个矩阵涵盖所有领域。M 中根据通用的分类标准进行描述，为了做到不重、不漏、不杂，存在一个矩阵的秩。矩阵 M 表示如下：

$$M=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

理论上，矩阵的规模 m 和 n 是一定的，且 $m>n$。实际中受个体认知的影响，矩阵规模会在小范围内有变化。一个行向量 X 表示一个词，例如向量 $X_1=(a_{11},\cdots,a_{1n})$ 具有 n 个元素，X_1 表示第 1 个向量，每个元素的值即为概念。

输入本体语义网络的是一个抽象的想法，可以称之为意

识。意识会激活多个概念,组成每个词的概念向量。激活函数用 f 表示,在一个具体的词汇语义组合关系框架内,该函数根据输入意识将概念矩阵的部分概念激活,得到用于反映某一确定含义的词的一个概念向量。与意识相关的概念被激活,而与意识无关的概念则没有被激活。不同的概念使激活函数作用于词汇矩阵,得到具体的词向量 X_i。i 表示向量的索引号。

$$f_i(M)=X_i, i\in[1,m] \tag{1}$$

句子 S 表示为被激活的各个词向量的组合,在不考虑依存关系的情况下简单表示如下:

$$S=\sum_i X_i \tag{2}$$

其中,S 就是最终的句子表示。

5.2　矩阵的分解

矩阵 M 中包含了词汇、概念、语义类,并按照不同的范围划分子领域,比如社会、政治、科技和艺术等各个领域,因此这个矩阵非常庞大,即使是某个具体的领域构成的子矩阵,规模也会很大。通过矩阵奇异值分解近似方法,用多个小规模的矩阵相乘来表示原来的大规模矩阵,相应的存储量和计算量都会小很多。根据奇异值分解的思想,取一个较小值 t,得到式(3):

$$A_{m\times n}=X_{m\times t}\cdot B_{t\times t}\cdot Y_{t\times n} \tag{3}$$

应用到词汇概念矩阵 M,就得到了矩阵的形式化分解式:

$$M_{\text{词}\times\text{概念}}=X_{\text{词}\times\text{语义类}}\cdot S_{\text{语义类}\times\text{领域}}\cdot Y_{\text{领域}\times\text{概念}} \tag{4}$$

分解后的每个小规模矩阵都有清晰的含义。第一个矩阵 X 是对词分类的结果,它的行数为 m,每一行表示一个词,列数为 t,每一列表示一个语义相近的词类,简称为语义类。于是,矩阵中的每一行就构成一个 t 维词向量,每个词的语义就分布在 t 个语义类中。矩阵 X 中的每个元素表示当前词与该语义类的相关性,维度值大小表示相关的程度。下面以一个 4×2 的矩阵为例来进行说明。

$$X=\begin{pmatrix}0.8 & 0.13\\ 0.23 & 0.5\\ 0 & 0.95\\ 0.35 & 0.02\end{pmatrix}$$

这个小矩阵里有 4 个词和两个语义类。第一个词与第一个语义类比较相关(相关度为 0.8),与第二个语义类不太相关(相关度为 0.13);第二个词与第一个词的情况相反。第三个词只与第二个语义类相关,与第一个完全无关;第四个词和每一类都不太相关,因为它对应的两个值都不大。

第三个矩阵 Y 的行数为 t,每一行对应一个领域,列数为 n,每一列对应一组概念。每个领域的概念就分布在各个列所对应的元素中,元素的值表示概念与该领域的相关性。相关性强的数值就大,相关性弱的就小,从而构成概念权重。

中间的矩阵 S 是个方阵,行数和列数都为 t,行表示语义类,列表示领域。矩阵中的元素表示该元素所在行列所代表的语义类与领域的关系。每一行是一个语义类,该语义类出现的频率分布于相关的各个领域中。矩阵元素值代表了相关性,如果该语义类构成某领域的专业术语,那么它们的相关性就强,数值就大;相反,如果该语义类与某领域无关,则该维度值就可能很小甚至为零。

将矩阵 M 分解之后,不仅降低了矩阵的维数,还得到了词、语义类、领域和概念相互之间的关系。这里的 M 是一个形式化的矩阵,在具体的构建中,适合应用于某一类文本中,如百科、新闻、军事类等。然后,再根据奇异值分解的思想对其进行分解,以降低计算复杂度。

结束语　受神经网络的启发,本文基于本体语义理论提出了概念网络和词汇网络的分布式模型,重点讨论了与大脑、CPU 运算机制相类似的一些特点,并给出了词汇的语义概念矩阵表示,从理论上对分布式词向量进行了解释。由于矩阵的规模非常大,因此基于奇异值分解对矩阵进行了理论分解,建立了数学逻辑推导模型。构建的目的就是面向语言信息处理,重视系统的可操作性,不至于太过抽象或繁琐。同时,模型的提出也结合认知将描述性知识通过学习转化成程序性知识,进而泛化,识别更多的模式。本体语义网络将人类的认知模式和目前得到较快发展的计算能力和精确度相结合,有望获得语言分析更丰富的成果,并且推动该领域继续发展。

本体概念集的构建较为复杂,仅靠词汇模板的方式获取概念的难度很大,借助深度学习方法,可以避免枚举所有的概念,通过词汇的语义聚类逐层构建,可以自动学习到所有的概念。在大的范畴内,根据语义分类划分词汇,在小的领域内,根据概念属性划分词汇,将词表示成分布式词向量,以获得更多的语义信息。

参 考 文 献

[1] MCSHANE M,NIRENBURG S,BEALE S. Two kinds of paraphrase in modeling embodied cognitive agents[C]// Proceedings of the Naturally-Inspired Artificial Intelligence AAAI Fall Symposium. 2008.

[2] 崔晓菊,易绵竹. 面向文本语义自动分析的本体语义学述要[J]. 解放军外国语学院学报,2013,36(2):39-43.

[3] NIRENBURG S,RASKIN V. Ontological Semantics (Language, Speech, and Communication)[M]. Cambridge: The MIT Press,2004.

[4] 曾毅,刘成林,谭铁牛. 类脑智能研究的回顾与展望[J]. 计算机学报,2016,39(1):212-222.

[5] 顾宗华,潘纲. 神经拟态的类脑计算研究[J]. 中国计算机学会通讯,2015(10):10-18.

[6] 唐华锦,胡隽. 神经拟态认知计算[J]. 中国计算机学会通讯,2015(10):27-31.

[7] 斯蒂伯. 我们改变了互联网,还是互联网改变了我们?[M]. 北京:中信出版社,2010.

[8] 王向前,张宝隆,李慧宗. 本体研究综述[J]. 情报杂志,2016,35(6):163-170.

[9] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estimation of Word Representations in Vector Space[OL]. http://www.surdeanu.info/mihai/teaching/ista555-spring15/readigns/mikolov2013.pdf.

[10] BEALE S,LAVOIE B,MCSHANE M,et al. Question answering using ontological semantics[C]// The Workshop on Text Meaning and Interpretation. Association for Computational Linguistics,2004:41-48.

[11] 刘海涛. 依存语法的理论与实践[M]. 北京:科学出版社,2009.

[12] HAVELIWALA T H. Topic-Sensitive PageRank:A Context-Sensitive Ranking Algorithm for Web Search[J]. IEEE Transactions on Knowledge & Data Engineering,2003,15(4):784-796.