

《智能信息处理》课程考试

基于本体的中文浅层语义神经网络 的构造

赵心田

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 19 日

基于本体的中文浅层语义神经网络 的构造

赵心田

¹⁾ (大连海事大学 信息科学技术学院, 大连 116026)

摘 要 在自然语言理解方面, 符号主义与联接主义目前都相继遇到了巨大的障碍。在这种情况下, 一种非常自然的考虑就是能否将两者结合在一起, 来摆脱目前自然语言研究的困境。而语义神经网络就是这样一种思想方法。它打破传统的基于规则的线性的理解模式, 把自然语言理解过程看成是语言(文字)信号在人脑中激活相应的神经元并建立/激活语义神经网络的过程。我们在文中设计了语义神经网络神经元的模型, 给出了神经元的结构、方法的定义和描述, 并对汉语表层语义关系进行了初步的分析, 同时创建了概念本体库, 给出了概念本体的定义以及描述, 建立了它与神经元的映射关系。

关键词 自然语言; 语义神经网络; 神经元; 概念本体

中图法分类号 TP36 DOI 号

Construction of Chinese Shallow Semantic Neural Network Based on Ontology

Zhao Xintian¹⁾

School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract In natural language understanding, symbolism and connectionism have encountered great obstacles one after another. In this case, a very natural consideration is whether the two can be combined to get rid of the current dilemma of natural language research. Semantic Neural Network is such an idea. It breaks the traditional rule-based linear understanding model and regards the natural language understanding process as a process in which language (text) signals activate corresponding neurons and establish / activate semantic neural networks in the human brain. In this paper, we design the neuron model of semantic neural network, give the definition and description of neuron structure and method, preliminarily analyze the surface semantic relationship of Chinese, create the concept ontology library, give the definition and description of concept ontology, and establish its mapping relationship with neurons.

Key words Natural language; Semantic Neural Network; Neuron; Concept Ontology

1 引言

人类历史上用语言文字形式记载和流传的知识占到知识总量的 80% 以上。因此, 如果计算机能够较好地理解自然语言, 还有助于揭示人类智能的奥秘, 深化我们对语言能力和思维本质的认识。而

要让计算机理解语言则离不开语言学, 将计算机与语言学有机结合起来, 彼此互助, 彼此互补, 利用计算机来处理自然语言, 产生了人工智能中最有挑战性的研究领域——自然语言理解。从计算机科学特别是从人工智能的观点看, 自然语言理解就是建立一种计算机模型, 这种计算机模型能够像人那样理解自然语言(即人们日常使用的汉语、英语等), 并

据此做出人们期待的各种正确响应。说的全面一点，就是如何去获取各种不同的知识，并以一种计算机可以使用和处理的方法表达知识。为了达到理解语言的目的，需要进行三步工作：理解所出现的每个词；从词义构造表示语句意义的结构；从句子的语义结构表示言语的结构。这三个过程中，句子分析上接篇章分析，下联词汇分析，起着承上启下的作用，因此自然语言理解的核心技术是语句分析技术。它可分为基于规则和基于统计两大类。

2 语义网的概念及形式

2.1 语义神经网络

自然语言中的句子是由词语通过组合而成的，但是词语之间的组合方式很多而且关系复杂，故直接给出词语与词语之间的关系不可能。而从认识史的角度来讲，人们对客观世界认识首先是具有了一个相应的语义概念以后再组织成自然语言的方式表达出来。也就是说词语只是概念或概念类的语言标记。因此我们认为词语的相关实为概念相关，更进一步可用概念类相关来初步刻画概念相关。在概念知识的表达方式中，语义网络一直以来都是一种能有效刻画概念之间语义结构的语义分析和表达方式。它以概念为结点，按一定的语义关系将不同概念连接起来，构成一种语义表达形式，很符合人脑对于知识的存储方式，但传统的语义网络的结点较单一，而且网络的生成采用的均为传统的产生式规则方法，对于丰富的语言现象其描述力就显得有点力不从心。这也是基于规则的符号主义的主要缺点。另一方面，从生物学的角度来看，人的大脑是一个由许许多多简单的神经元所构成的庞大而又复杂的神经网络。大脑的神经网络又被分成了许多的子网络，并且彼此有着不同的分工。既然人们对客观世界的认识被抽象为大脑中的概念，那么概念的存储必然是由这些神经元来共同承担的。把语义网络中的每个结点看成是一个有着自发行为的神经元，而 SNN 就是这些神经元按照语义网络的结构，也即按照人们所认识的概念间的相互关系连接而成的。在我们的 SNN 中，神经元可以被外部信息、其它神经元送来的信息所激发，也可以自主地激发。它可以和其它的神经元相连接，并依据其内部嵌套的知识规则或函数向其它的神经元发送信息。

2.2 SNN 形式

一个神经元用来表示一个概念类，并承担了该概念类的识别、存储、和知识处理任务。这些神经元按照语义网络的结构形式连接构成 SNN，其形式定义为：

一个神经语义网络 u 是一个五元组 (T, V, L, E, n_0) ，其中：

1、 T 是一个非空有限集，其中元素称之为概念符号；

2、 V 是一个有限集，其中元素称为符合概念；

3、 L 是一个有限集，其中元素称为语义角色；

4、 $E \subset V \times L \times (V \cup T)$ 是一个弧集；

5、 $n_0 \in V$ 称为 u 的根，它是 u 的所有结点的祖先；

其中，所有的结点、概念符号和复合概念都是神经元。它是由这些神经元依据它们所表达的概念的继承、关联等关系而联结在一起所构成的一个有向图。

通过概念之间的合理搭配来取得正确的语义理解，因而具有很好的适应性和更高的灵活性，不拘泥于固定的句式规则，更适合表示丰富的自然语言。同时用概念类描述语义关系有小而多的歧义性并且关系表达非常直观；其次这种描述支持一种类似人类学习机制的语义自学心里模拟。另外 SNN 具有并行处理的特点，运行速度快，因而一方面可存储大量的知识，另一方面可保持运行速度。它克服了传统自然语言理解模式的不足，为自然语言理解的研究提供了一种新的方法。而我们要做的工作就是构造 SNN 来实现汉语表层语义理解进行研究。

3 语义神经元设计

3.1 神经元概述

SNN 是由代表概念类的神经元按照人们所认识的概念间的相互关系连接构成的。它可以被外部信息、其它神经元送来的信息所激发，也可以自主地激发。同时还可以和其它的神经元相连接，并依据其内部嵌套的知识规则或函数向其它的神经元发送信息。根据 SNN 的以上特点，我们认为 SNN 中的神经元具有下述部分或全部特征：

(1)自治性。神经元的运行不需要人或其他神经元的直接干预，神经元能够根据内部的语义知识控制自己的行为 and 内部状态。

(2)协作性。它往往不是单独存在的，它要通过

与其他神经元进行交互,以共同来生成语义网络。

(3)自发性。它可以被动地被环境的变化所激发,做出反应,也能够自主的激发,表现出一种有目的的行为。

显然 SNN 神经元继承了人工神经元的自发性和分布式知识表示等特点。但是单个人工神经元的内部结构过于简单和机械,不适合直接用于表述自然语言概念知识以及它们之间的复杂关系。另外概念类之间的协作和自治性也单靠人工神经元是无法实现。因此我们将它设计成为具有人工神经元和概念类双重特性的结点。

3.2 神经元的模型设计

跟人工神经元类似, SNN 中的每个神经元都具有自己的一组输入、输出、以及内部处理。输入输出根据神经元类型以及内部描述的不同按需要设定。神经元接受其他神经元所发出的信号作为它的输入内容,然后将这些信号根据自身所具有的知识规则来进行相应的处理,再以信号形式输出到其他神经元,其输出信号就是相关神经元的输入信号。

下面是我们模拟人工神经元[23]模型并结合概念类的特征所给出的一个 SNN 神经元的模型:

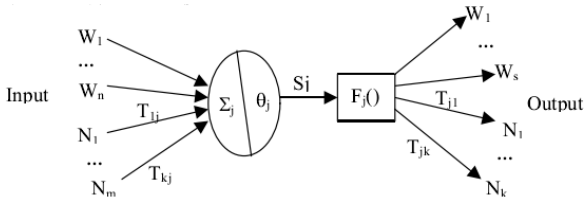


图 1 SNN 的神经元模型

图中: 输入信号分为两类: 1、 $\{W_1, \dots, W_n\}$ 为其他神经元, 它们发送过来信息信号。2、 $\{N_1, \dots, N_m\}$ 为相关神经元, 它们发送过来作用(刺激)信号。前一类信号主要用于语义神经网络的深层语义的计算, 而在本文中涉及到的主要是后一类信号。

$\{T_{1j}, \dots, T_{kj}\}$ 和 $\{T_{j1}, \dots, T_{jk}\}$ 为信号的权重, 表示信号的强弱或连接的强度。

Σ_j 表示神经元 j 的综合输入权值数, 记为:

$$\Sigma_j = \frac{\text{激活信号的权值和}}{\text{激活信号的个数}} + \frac{\text{一般刺激信号信号的权值和}}{\text{一般刺激信号的个数}} +$$

$\frac{\text{抑制信号的权值和}}{\text{抑制信号的个数}}$

我们根据权重的大小来判断信号的类别。当 $T_{ij} > 1$, 则认为是激活信号, $0 < T_{ij} \leq 1$ 认为是一般刺激信号, 而 $T_{ij} = -1$ 则是抑制信号。

θ_j 为阈值, 我们均设为 1。 S_j 表示神经元的

激活水平, 也称为状态函数。 $S_j = \sum j - \theta_j$ 。我们将其分为两个等级, 活跃状态 $S_j > 1$ 和抑制状态 $S_j \leq 1$ 。 $F_j()$ 表示神经元的输出函数。它对接受到的信号结合内部知识进行操作和处理, 以此来决定输出的信号。

我们把表层语义层面上神经元之间的信号分为两类: 通信信号和作用信号。对神经元的状态和地位都不产生任何影响的信号我们称之为通信信号, 它用于神经元之间的信息传递。而把对神经元的状态或所处地位产生一定影响的信号称为作用信号。作用信号又分为语义作用信号和语义组合信号两大类。

语义作用信号对神经元的状态产生影响。它包括(1)刺激信号 $\text{Signal1}(N_i, N_j)$, 其中 $T_{ij} \in \{0.1, 1\}$ 。刺激信号有强弱之分, 只有达到一定的阈值才能激活该神经元。(2)抑制信号 $\text{Signal2}(N_i, N_j)$, 权值 $T_{ij} = -1$; (3)唤醒信号 $\text{Signal3}(N_i, N_j)$, 权值 $T_{ij} = 1$ 。语义组合信号是指对神经元在语义网络中所处的地位产生一定影响的信号。表示为 $\text{Signal4}(N_i, N_j) = \{<R_k, D_k> \mid \text{其中 } R_k \text{ 为 } N_i \text{ 的某个语义角色名, } D_k \text{ 为语义关系类型代码, 令 } D_k \in [2, 5]\}$ 。

3.3 神经元的结构设计

神经元的构造与概念类的构造是一致的。神经元在拥有多个接口与外界通讯的同时, 内部还具备对应概念类的语义特征知识, 以及实现内部通讯和神经元作用的方法。从自然语言理解出发, 我们认为神经元也对应由表层语义和深层语义两个层次构成。表层语义反映的是概念之间固定的语义搭配关系, 以及一些语用、逻辑、语序等信息。深层语义则是结合具体的环境, 在表层语义基础上的推理及运算。鉴于框架良好的继承性和结构化[28], 我们在内部把神经元的表层语义知识设计为一个框架, 抽象为一个输入函数集、处理函数集、和输出函数集。输入函数集负责对接受到的信号进行处理, 包括信号的分类、权值求和等等。处理函数集则根据带有的概念类知识规则和信号里的信息进行相应地处理。处理函数集抽象为一个属性集和一个内部函数集。属性集描述该概念的语义属性, 它由若干语义关系(Semantic Relation)构成。一个关系包含一个语义角色名和一个与之相关联的概念类的集合。内部函数负责神经元内部操作、以及神经元的内部通讯。输出函数集负责与其它神经元的各种通讯, 它由作用算子(Influences)和外部通讯函数组成。作用算子是神经元对其它神经元的某种作用。

由此，我们描述神经元如下：

<神经元>

{

 输入函数集；

 处理函数集；

 输出函数集；

};

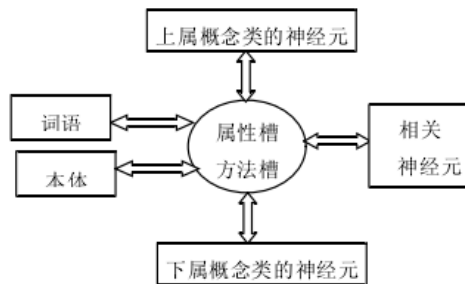


图2 SNN 神经元抽象模型

在具体实现上，神经元由两种类型的槽组成：属性槽和方法槽。SNN 神经元对应的抽象模型如图2所示。属性槽（简称为属性），表示为一个属性列表，用来刻画概念类的属性集。其中公有属性槽刻画神经元的一般语义关系。私有属性槽刻画该神经元的语义搭配关系。方法槽（简称方法）描述神经元的作用算子和内部函数，其值为程序。由此，神经元的框架表示为：

<神经元描述> ::= def class <概念类名>[<概念类型名>]

{

 {<神经元的属性槽（公共属性槽、私有属性槽）>}

 {<神经元的方法槽>}

}

<公共属性槽> ::= <属性名><指针>

[<注释：包含该神经元可被其子类继承的公共属性和关系>]

<特殊属性槽> ::= <属性名><指针>

[<注释：包含该神经元私有的语义属性和语义关系>]

<方法槽> ::= <算子槽><算子>

<算子> ::= <作用算子槽>:<算子名><方法值>

.....

[<注释：方法值为对应的算子函数，

 同一算子可被多个神经元同时引用>]

<内部函数槽>:<函数名><方法值>

.....

[<注释：方法值为对应的内部函数>]

<属性名> ::= <字符串>

<注释> ::= <字符串>

<算子> ::= <子框架>

<算子名> ::= <字符串>

<函数名> ::= <字符串>

<方法值> ::= <函数>

神经元所包含的属性集，也即神经元所具有的语义知识我们将在继的章节里将给予了较详细的刻画和描述，这里我们主要对神经元的方法集中所包含的一些常用函数来给以简要介绍和描述。

4 本体库

4.1 本体的引入

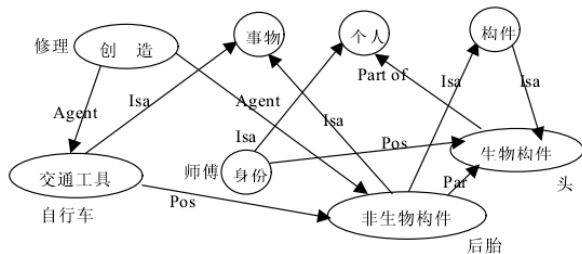
Ontology 的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇间相互关系的明确定义。具体地说,某个领域的本体就是关于该领域的公认的概念集,该概念集包含确定的语义和概念之间的关联,概念是以本体表征出来的。因此 Ontology 的本质就是概念模型,表达的是概念及概念之间的关系。我们将要刻画的表层语义体现的就是概念之间的固定语义关系,这些关系是共同认可的知识,独立于任何语言环境。

基于以上这些,我们在处理概念知识的时候,引入本体论的观点,创建了概念本体。我们把概念本体作为神经元语义知识的描述手段,以一个概念或概念类为基本知识元素,通过对概念本体的描述来建立概念(类)之间的语义关系,如:格关系、槽关系、对立关系、上下位关系、整体部分关系等,为神经元提供完备的表层语义知识。这里引进本体论是想从学科领域的概念和关系、属性集出发,建立一个便于理解和分析的汉语概念知识结构,来更好的描述概念本身及概念之间的错综复杂的约束搭配关系。并支持满足一致性的汉语知识库的开发,同时便于实现通用领域上的知识共享。

4.2 概念本体库的构造

在建立汉语概念本体之前,必须弄清楚什么是一个概念,概念是对客观事物及其特有属性的概括的反映。一个概念表示为一个语言符号串集合。在我们的库中,概念(类)就是一个同义词的集合。在概

念本体库中,我们把每个概念(类)看作一个本体。汉语本体体系是一个由概念按照一定的语义关系连接在一起的语义网络结构(图 3),网络节点由概念名标识。SNN 的连接形式也是语义网络的结构形式,故两者就很好的耦合在一起了。



4.3 基于本体的 SNN 的构造

SNN 是模拟人脑神经系统的基本组织特性构成的新型信息处理系统。该网络一般由多个神经元组成,每个神经元有多个输入和多个输出,它跟其它的神经元以语义网络的结构形式互连在一起,构成自适应、非线性的动态系统。SNN 对于自然语言的理解是由本体知识库结构、神经元的输入输出特性、神经元的方法集、以及神经元间的连接状况等决定。SNN 构造流程如图 4 所示。

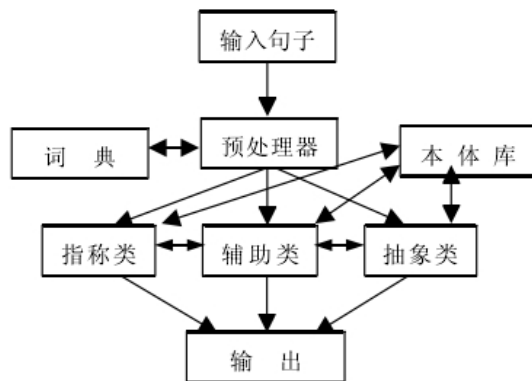


图 4 基于本体的 SNN 构造流程

其中预处理器根据词典取得每个词所对应的概念类;同时删除一些无意义的助词如:的、得等词,并根据需要添加一些被省略的词。由输入带输入的已经分词的句子经预处理以后就对应为概念串。同时激活对应的神经元。三类神经元分别从本体库获得所需的语义知识,并根据各自的方法集自主地进行协商,做出合适的动作,建立/激活 SNN。

5 总结

本文首次把语义神经网络应用于汉语自然语言句子的分析,用神经元的激活和自主连接的方法

来实现汉语自然语言句子的表层语义分析。具体来说,本文的主要工作有如下几个方面:

(1) 本文将人工神经网络神经元的结构特点与自然语言概念类的特性有机结合,构造了语义神经网络神经元的模型结构,并给出了具体的属性和方法描述。

(2) 本文提出将本体与概念相结合,建立了概念本体库,用来描述和存储神经元知识,作为神经元描述方法的一个有效补充手段。同时针对汉语表层语义的特点,总结了一套表层语义关系体系,具有一定的通用性。

(3) 本文通过协商来实现语义神经网络神经元之间正确、有序地自主联接。为自然语言的智能化理解进行了初步的探索。

参考文献

- [1]边慧珍,哈斯.知识图谱概念获取研究进展[J].广西科学院学报,2018,(1).
- [2]文亮,李娟,刘智颖,等.基于概念层次网络的知识表示与本体建模[J].中文信息学报,2018,(4).doi:10.3969/j.issn.1003-0077.2018.04.008.
- [3]Lin Tsung-Yi,Goyal Priya,Girshick Ross,et al.Focal Loss for Dense Object Detection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(2).doi:10.1109/TPAMI.2018.2858826.
- [4]黎雪微,应时,& 洪伟. (2021). 基于本体和信息量融合的个性化推荐方法研究. (2019-9), 90-95.
- [5]张玉峰,& 何超. (2021). 基于本体的竞争情报语义整合与分析研究. (2011-7), 64-66.
- [6]梁昌勇, 杨大寨, 张什永, & 梁祥君. (2021). 基于本体的政务信息资源个性化应用研究. (2012-9), 66-70.
- [7]彭文博, 李炳军. 基于本体论的大学本体的探究[J]. 2021(2016-9):7-12.
- [8]张德政, 谢永红, 李曼,等. 基于本体的中医知识图谱构建[J]. 2021(2017-1):35-42.