WordNet

1327405063 李佳鹏

WordNet是由Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典。它不是光把单词以字母顺序排列，而且按照单词的意义组成一个“单词的网络”。

是一部在线词典数据库系统，采用了与传统词典不同的方式，即按照词义而不是词形来组织词汇信息。WordNet将成为一种国际标准。目前许多国家在筹划和建立于英文WordNet兼容的本国语言WordNet系统。

**1）WordNet的词汇组织结构大致包含:**

WordNet 将英语的名词、动词、形容词、和副词组织为Synsets，每一个Synset表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系（synonymy）、反义关系（antonym） 、上下位关系（hypernymy & hyponymy） 、部分关系（meronymy）等多种语义关系。 主要的词汇关系

1. 同义关系（构成Synsets）

WordNet 用同义词集表示概念，最重要的关系就是同义关系。如果两种表达方式在语言文本中相互替代而不改变其意义，则着两种表达就是同义的。

例如：

{heavy, weighty, ponderous}

{light, weightless, airy}

1. 反义关系 （指针 !）

反义关系不是WordNet的基本组织关系。反义关系是一种词形关系，不是概念之间的语义关系。

例如：

男人 女

上升/下降 升高/下落

heavy/light weighty/weightless

（形容词的反义关系

* + - * 描述性形容词的语义组织方式与名词完全不同，采用N维超空间结构，而不 是树形的层次结构。
      * 描述性形容词的基本语义关系是反义关系。
      * 形容词的同义词集用形容词簇表示
      * 形容此词簇通过反义关系联系在一起。

例如：

* + - * + ponderous 的语义近似于（similar to ）heavy
        + 直接反义词对(light/heavy)
        + 间接反义词对(light / ponderous)

3〉 上下位关系（hypermymy & hyponymy）

上下位关系是词汇之间的语义关系。 例如：{枫树}是{树}的下位词，{树}是{植物}的下位词

上下位关系也称为从属/上属关系，子集/超集关系，或ISA关系。

上下位关系不是一种对称关系，通常一个同义词集只有唯一的上位，而下位却很多。

名词的上下位关系

* + - * 名词的组织依照上下位关系构造名词语义树
      * 名词的继承系统，构成一种层次关系(hierarchies)

例如，橡树@--->树@--->植物@--->生物

* + - * 下位词汇的词条继承了他们上位词汇的所有属性(attributes)
      * 名词的语义元素
        + 25个起始概念：一般性概念，作为独立的层次结构
        + 各领域的名次数量不等，互不排斥
        + 大体覆盖了明确的概念和词汇范围

4〉部分关系 （名词、指针%m/%s/%p）

“部分-整体”关系

例如：

if “A Y has an X ” or “A n X is a part of Y”

then{x1, x2, …, xn} are meronym

{y1, y2, …, ym} are holonym

部分关系是一种不对称的关系

部分关系是名词概念之间的一种语义关系。

动词蕴涵关系(entailment) 、动词的上位关系(hypernymmy) 、方式关系(troponymy)、因果关系(cause relation)

动词之间的组织关系成为动词的蕴涵

例：

打鼾逻辑上蕴涵睡觉，因为句子“他正在打鼾”蕴涵了句子“他正在睡觉”；如果第一个句子成立，则第二个句子必定成立。

蕴涵是一种单向关系

方式关系是一种特殊的蕴涵。

例如：

一个较一般化的动词V2，它的每一个方式词V1也蕴涵V2V1是V2的一种方式。

2）WordNet 的算法

wordnet是一个免费的软件包，可以用来计算两个概念（或者word sense）之间的语义相似度，它提供了六种计算相似度 和三种计算概念之间的关联度方法，所有的这些方法都是基于wordnet这个词汇数据库。这些计算方法已经用Perl实现，算法模型的输入是两个概念，返回结果就是用来表示相似或者关联度的一个数量值。

概述：

计算两个概念相似的程度，主要是基于包含在层次结构“is－a”结构中的信息。比如：如果automobile和boat拥有同一个祖先vehicle，那么automobile 会被认为 与boat 的相似程度大于和tree的相似度。

自从wordnet把名词和动词组织成is－a这种层次关系，wordnet 就非常适合进行形似度的计算。在2.0版，有九个名词层次结构 包含80000个概念，还有554个动词层次结构包含13500个概念。

Wordnet中的is－a关系并不是跨词性的，所以，基于wordnet的相似度计算只是局限在名词之间（比如：cat 和dog）和动词对之间（比如：run和walk）。而wordnet也包含形容词和副词，这些词不能组织成is－a关系，所以这些词也就不能进行相似度的计算。

然而，概念之间除了通过相似度关联外，还可以通过很多种其他方法。比如：wheel 是car 的一部分，night是day的反面，snow是由water变成的，knife是用来切bread等等。Wordent也提供额外的联系 比如：has－part，is－made－of，is－an－attribute－of 等等。除此之外，每一个概念（或者word sense）都通过一段简短的定义或者注释加以描述。

计算关联的方法是基于这些额外的信息，并且这些方法可以运用的更广泛的概念对中。比如:可以跨越词性,然后可以达到使动词murder和名词gun相关联的程度。也可以计算不是is－a关系的词汇比如形容词 violent和harmful。

相似度计算

三种相似度的计算是基于概念之间路径长度的：lch（Leacock&Chodorow 1998）,wup(Wu&Palmer 1994),和path，其中lch方法是找到两个概念之间的最短路径，然后在is－a层次中找到最大路径长度。Wup方法是找到两个概念从最近的公共祖先节点到root 节点的路径长度。Path方法是两个概念之间最短路径长度的倒数。

另外，概念的信息内容来自SemCor的sense－tagged 集合。然而，也有现成的程序可以对没有标注的集合中的信息进行计算，比如Brown Corpus，the Penn Treebank, British Nation Corpus ,或者任意形式的普通文本。

WordNet::Similarity支持两种假定的根节点，这两种根节点可以存在也可以不存在，当存在时，一个根节点包含所有的名词节点，另一个包含所有的动词节点。这样便允许把相似度计算应用到任意的名词和动词对之间。如果两个根节点不存在，那么只有在同一个is－a层次中的词才能计算相似度。

计算关联度

有三种wordnet支持的计算关联度的方法分别是：hso（Hirst&St-Onge1998）,lesk(Banerjee&Pedersen 2003),和vector(Patwardhan 2003).hso方法是基于路径并且根据方向把关系分类。比如is－a关系是纵向的，has－part关系是横向的。最好能在两个概念之间找到一条路径，这条路径不能太长并且方向也不能改变的太频繁。

Wordnet中的每一个概念（或者word sense）都是通过一个短注释进行定义的。Lesk方法和Vector方法是把注释作为概念的唯一的标识方法。Lesk方法通过寻找和计算两个概念的注释的交叉部分，和在wordnet中同时与一个概念相连接的方法来计算相似度。