中外语言文化比较研究丛书
丛书主编　许明武　谭　渊

搭配与谓词语义计算

唐旭日　著

Collocation and Computation of Predicate Semantics

WUHAN UNIVERSITY PRESS
武汉大学出版社

中外语言文化比较研究丛书
丛书主编　许明武　谭　渊

本书受华中科技大学人文社科自主创新重大交叉项目（2018WKZDJC003）资助

搭配与谓词语义计算

唐旭日　著

Collocation and Computation of Predicate Semantics

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

搭配与谓词语义计算/唐旭日著. —武汉：武汉大学出版社，2018.11
中外语言文化比较研究丛书/许明武，谭渊主编
ISBN 978-7-307-20637-3

Ⅰ.搭… Ⅱ.唐… Ⅲ.汉语—主谓—研究 Ⅳ.H146.3

中国版本图书馆 CIP 数据核字(2018)第 264014 号

责任编辑：罗晓华　　责任校对：汪欣怡　　版式设计：韩闻锦

出版发行：**武汉大学出版社**　(430072　武昌　珞珈山)
(电子邮件：cbs22@ whu.edu.cn　网址：www.wdp. com.cn)
印刷：北京虎彩文化传播有限公司
开本：720×1000　1/16　印张：14.25　字数：194 千字　插页：1
版次：2018 年 11 月第 1 版　2018 年 11 月第 1 次印刷
ISBN 978-7-307-20637-3　定价：38. 00 元

中外语言文化比较研究丛书

自　序

语义理解是自然语言处理研究的终极目标，而谓词是句法分析和语义理解的核心成分。谓词语义计算不仅能够推进词汇语义研究的发展，而且对于理解句法词汇交互，进而确定句法层面的语义表达具有重要意义。本书以谓词语义为研究对象，采用计算语言学方法，以词语搭配为主要语言知识来源，讨论了谓词的词义消歧、义项区分和词义发现三个方面的问题。

本书首先从语言信息处理的角度对词语搭配进行了深入分析，并提出了新的词语搭配模型。该模型认为，搭配作为一种重要的语言知识，其基元是“词语—语义类”，搭配基元之间的关系是句法结构框架下的语义关系，搭配强度是一个连续统，搭配强度计算存在客观概率分布和主观概念层次两种方法。与现有词语搭配模型相比，“词语—语义类”搭配模型能够更准确地反映词语搭配的本质和界定词语搭配知识的颗粒度。

为确定搭配中谓词的语义，本书提出基于冗余原则的词义消歧算法。这种消歧算法利用语义的组合冗余和平行冗余，建立了基于关联的义项组合过滤机制和基于最小描写长度的义项组合过滤机制，采用半监督方法确定了主谓搭配和动宾搭配中歧义性较大的谓词语义，并取得了较高的精确度，运用这种算法，我们在较少人工参与条件下构造了大规模的基于“词语—语义类”模型的主谓和动宾搭配知识库。

在义项区分方面，本书讨论了知识本体在提高义项确定的可操作性

和一致性方面的作用，并建造了基于 DBSCAN 聚类的义项区分辅助平台，进行了谓词语义区分实验，实验结果充分表明自动聚类在解决义项区分的可操作性、义项颗粒度大小以及义项发现等问题中的作用。

谓词语义的自动发现是本书中最主要的部分。现有计算词汇语义学研究主要集中在词义消歧，而缺少对语言动态变化导致的词典义项缺失问题的研究。本书提出了基于概念隐喻的词义发现模型，从隐喻理解角度对谓词词义的自动发现进行了探索性研究：

首先，本书在分析计算语言学领域广泛应用的词语选择限制隐喻识别模型的局限性基础上，提出了基于语义关系模式的隐喻识别模型。该模型利用多种语义关系来建立语义关系模式，并应用支持向量机进行分类识别。实验表明，该模型隐喻识别 F1 值达 89.15%，比词语选择限制模型高出 37%，显示出较强的泛化能力，摆脱了对隐喻词表的依赖。

其次，本书提出了基于经验的概念隐喻知识库构建方法。依据概念隐喻知识库的结构与知识本体中的概念结构的同构性特点，构建了概念隐喻知识库整体框架。该框架通过多义词分析自动构造了包含 440 个属性概念隐喻，1097 个属性值概念隐喻的基础概念隐喻知识库，并通过词语概念关联对部分形容词进行了概念隐喻知识库扩展实验。

最后，本书分析了词义的动态性，并以“主语+形容词谓语”主谓结构为例，构建了谓词词义发现模型。该模型利用概念隐喻知识库和实体(事件)属性知识库，通过属性显著度计算发现谓语形容词的语义。

唐旭日

2018 年 6 月 3 日

目　录

第一章 绪 论

冯志伟(2010)在总结自然语言处理发展的几个特点时认为："自然语言处理中越来越重视词汇的作用，出现了强烈的'词汇主义'的倾向。"这种倾向性表现在词汇知识库建设(如 WordNet①、HowNet②(Dong, 2006))、词义消歧(如 SENSEVAL③)、语义相似度、语义相关性以及语义框架等相关研究的增多，也表现在国内、国际层面各种词汇语义专题讨论会议的增多。例如，针对知识本体和词汇语义的会议 Ontolex④ 自 2000 年至 2010 年几乎每年都会召开一次；汉语词汇语义学会议至 2010 年已是第 11 次会议；此外，还有其他各种类型的专题研讨会，如 2009 年的 UMSLLS⑤(Unsupervised or Minimally Supervised Learning of Lexical Semantics)等。

---

① WordNet 是一个英语词汇知识库，由普林斯顿大学 George A. Miller 主持开发，可通过 http://wordnet.princeton.edu/获取。

② HowNet 又称知网，是我国学者董振东和董强研制的语言知识库。HowNet 以汉语和英语作为概念的描述手段，来描写汉语和英语中的词语所表达的概念。HowNet 是一个常识知识库，给出了概念与概念之间，以及概念与其所具有的属性之间的关系。该知识库的详细信息可以通过网址 http://www.keenage.com/获取。本书采用 HowNet 义原表示词语的意义解释，并采用"[English | 中文]"的表示形式，如"[big | 大]"。

③ http://www.senseval.org/.

④ http://www.ontotext.com/OntoLex/.

⑤ http://aclweb.org/anthology-new/W/W09/#1700.

词汇语义学的兴起并非偶然。自然语言理解的终极目的是语义理解(宗成庆，2008：190；陈小荷，1998)，语义理解是一个“人工智能完备性(AI-Complete)”问题(Ide & Veronis，1998)。而词是最基本的语言单位，也是语义的基本单位，为达到语义理解的终极目标，需要首先从词汇语义的研究开始。Ide 等(1998)在讨论词义消歧研究的意义时指出，词义的研究对于一些基于深层语义理解的应用，如消息理解(Message Understanding)、人机对话等是不可或缺的一部分。此外，词义对于机器翻译、信息抽取、超文本导航(Hypertext Navigation)、内容和主题分析、句法分析、语音处理以及文本处理等自然语言处理任务，至少是有益的，在某些情况下则是必须的。

计算语言学中的词汇语义学主要关心两个问题：词义知识的形式化表征和词义计算的模型与算法。词义知识的形式化表征可以包括词义的聚合语义知识的表征和组合语义知识的表征(俞士汶，2003)，其中聚合语义知识可以包括近义关系、反义关系、上下位关系、整体—部分关系、成员关系等。组合语义知识则包括词语选择限制、词语搭配、语义角色等。词汇语义的计算在 Jurafsky 和 Martin (2008)中称为计算词汇语义学(Computational Lexical Semantics)，主要任务是词义消歧(Word Sense Disambiguation)、词语相似度计算(Similarity Computation)以及语义角色自动标识(Semantic Role Labeling)、隐喻(Metaphor)的识别与理解等。

本书讨论的内容归属词汇语义学范畴，以搭配研究为基础，探讨了词义消歧、义项区分和词义发现问题。本章详细介绍本书的研究对象、假设、思路及意义。

1.1 研究对象

本书以谓词语义为研究对象。谓词语义是指句子中充当谓语中心词的词义。在动宾结构中，谓词语义是结构中动词的语义；在主谓结构

中，谓词语义是结构中谓语中心词的语义，谓语中心词[①]可以是动词，也可以是形容词。

当试图解释例 1-1、例 1-2、例 1-3 和例 1-4 中词语“大”的语义时，可以发现几种不同的现象：

例 1-1 a. 车子大　　b. 面包大　　c. 电脑大

例 1-2 a. 雾大　　b. 声音大　　c. 问题大

例 1-3 a. 心胸大　　b. 魄力大　　c. 耐性大

例 1-4 a. 话大　　b. (这个问题令人)头大

在例 1-1 中，利用词语的语义选择限制知识，可以确定“大”的语义解释为“[big | 大][②]”，该义项能够在 HowNet 中找到，且一般认为该义项是词语“大”的基本意义(或称为字面意义)。利用语义选择限制，也能够确定例 1-2 中“大”的语义解释分别为“[strong | 强]”、“[loud | 高声]”和“[serious | 严重]”。这些义项也能在 HowNet 中找到。例 1-1 与例 1-2 的不同之处在于例 1-2 中词语“大”的义项可以看成隐喻词汇化的结果。

在解释例 1-3 中主谓结构的语义时发现，HowNet 中给出的词语“大”的义项都不能作为合适的语义解释，但是 HowNet 的义素集合提供了适合的义原。如例 1-3a 可解释为“[broadminded | 心胸开阔]”，例 1-3b 为“[brave | 勇]”，例 1-3c 为“[patient | 有耐性]”。与例 1-3 相比，例 1-4 中结构的语义解释更加困难。HowNet 中不仅没有定义谓词“大”的语义解释，寻找词语“话”“头”的合适语义解释也是困难的。

对于例 1-1、例 1-2 而言，谓词语义的确定问题也就是词义消歧(Word Sense Disambiguation)问题。所谓词义消歧，就是依据上下文语

① 本书没有考虑谓词是名词的情况。

② 参阅第 1 页脚注②。

境确定多义词的意义。具体而言，在例 1-1、例 1-2 的主谓结构中，是指利用主语中心词以及主谓结构特征信息确定“大”的意义。

词义消歧的研究始于 20 世纪 50 年代，在最初的机器翻译研究中就提出了确定词义、依据语境确定多义词词义的任务。在人工智能领域，词义消歧被认为是最难解决的问题之一。Jurafsky 和 Martin(2008)定义“词义消歧”如下：在给定目标词语出现的上下文语境以及一个封闭词语义项集合的条件下，依据上下文语境信息从目标词义项集合中确定一个合适的词语义项。可以看出，上述定义是在一系列假设条件下定义的。其中一个重要的假设是，所给出的封闭的词语义项集合能够满足目标词语意义解释的需要。换言之，目标词在任何语境下的合适语义解释都能够在这一封闭义项集合中找到。这一假设可称为义项的完备性(吴云芳、俞士汶，2006)，“操作者(计算机或者人)可以顺利地对语料中的每一个目标词标注出义项”。词义消歧的另一个假设是，在具体上下文语境中，通过选择一个词语义项即能够达到词义歧义消解的目标，可称为“歧义消解的确定性”假设。

然而，上述假设对于例 1-3、1-4 中“大”的义项获取并不适用。在例 1-1、例 1-2 中，词语“大”的义项能够在 HowNet 中找到。而例 1-3、例 1-4 中词语“大”的义项并没有在 HowNet 中给出，因而义项完备性假设不能成立。但是这并不能说明 HowNet 词语义项分析不充分。对于词语“大”而言，HowNet 区分了 12 种不同义项，而《现代汉语词典》仅区分了 9 种义项。例 1-3、例 1-4 中所揭示的词典义项不完备性问题不是义项区分不充分造成的，其根本原因在于义项的表征机制。词典表述义项的一般表现形式是静态的“枚举”，通过列表形式给出词语的可能义项集合。然而语义的动态变化是语言的本质特征，是否能够通过枚举形式达到义项的完备性是一个需要进一步求证的问题。

词义消歧定义的第二个假设是语义解释的确定性。一般而言，词语在具体的语境中其意义是确定的(俞士汶，2003)。这种意义的确定性是通过不确切语义表征方式(Underspecification)，即选择能够涵盖所有可

能语义解释的概念(Lenci, 2001)而实现的。如例 1-1 中“大”表示体积上的尺寸比较大。但是在不同的结构中，大的具体尺寸大小并不一样。例 1-1a“车子”的体积要远远大于例 1-1b 中“面包”的体积，但都可以使用“大”进行描述。不确切语义表征方式也表现为使用一个抽象的概念指代具体的概念，如使用“[food | 食品]”表示面包、饼干、米饭等不同类型食品。

然而，并不是在所有情况下词语意义都是确定的。一些比喻性语言(如隐喻表达式)，给定的上下文语境可以允许词义采用多个语义解释(Nirenburg & Raskin, 2004：15)。例如在主谓结构“纰漏大”中，“大”可能表示错误的严重性，或者表示错误数量比较多，甚至可以蕴涵造成错误的人比较粗心的意义。例 1-5 列出了另外一些典型例句。从这些实例看来，语义确定性假设并不是在所有情况下都成立：

例 1-5

a. My lawyer is a shark. (Glucksberg, 2008)

b. Zizek is another Derrida. (Sperber & Wilson, 2008)

c. Locomotive is in the bed. (MacCormac, 1985)

g. 丑女是恐龙。 (摘自互联网)

h. 丑男是青蛙。 (摘自互联网)

本书研究就是围绕上述问题展开的。具体而言，我们采用例 1-1、例 1-2、例 1-3 以及例 1-4 形式的大规模词语搭配库为研究对象(其中包括主谓搭配和动宾搭配)，探索了两种谓词语义获取方法。一种方法是词义消歧方法，在认同义项完备性假设和歧义消解确定性假设的基础上，构建面向词语搭配的词义消歧模型，确定主谓搭配和动宾搭配中谓词的语义。这一类型的词义获取方法是本体层次上的语义处理(俞士汶，2010)。另一种方法则是基于概念隐喻机制的词义发现方法。这一方法在义项完备性和歧义消解不确定性两个假设不成立情况下，通过概

念隐喻机制发现词语在该语境中的合理词义解释。该方法在构建大规模概念隐喻知识库的基础上，通过知识推理机制给出特定语境中谓词的可能义项解释。这种方法被称为认知层次上的语义处理(俞士汶，2010)。

1.2 研究假设

语义的研究具有悠久的历史。古代的训诂学即语义研究的一部分。近代语义学研究由法国语言学家布雷阿尔在1893年提出，距今已有一百多年历史。在此期间，语义学的理论层出不穷，如指称论、意念论、真值条件论、用法论、境况论等(徐烈炯，1990)。不同的语义学理论研究出发点不尽相同，研究对象也不尽相同，用以解释语义现象的理论框架也存在较大差别。因此，语义研究往往需要首先建立语义概念的一些假设性认识，并在这些假设基础之上开展研究工作，从而使得研究具有可操作性。

语义研究首先面临的是语义与思维和客观现实的关系。对语义与思维及客观现实之间关系问题的不同认识，形成了两种不同的语义观(Saeed，2000：24)。一种语义观强调语言的指称作用，语义是对客观现实和客观情景的指称，称为语义的指称理论。另一种语义观则认为语言是概念结构的一种外在体现，语言并不能直接指称客观世界，而需要依赖于人脑中的内在客观世界模型。本项研究采用的是第二种语义观，即认为语义是概念结构的语言外化形式，并认同认知语言学关于语言与思维的一些认识原则：

(1)语言是人类思维和认知的一种形式，语言的运用体现了人类认知的一般原则。反之，人类认知的一般原则也可应用于语言结构和语言运用规律的解释(Saeed，2000：300)。

(2)语义以习俗化的概念结构为基础，是对人们在与客观世界的交互过程中形成的概念范畴的反应(Saeed，2000：301)。