第一图书网, tushu007.com

<<现代汉语词语级歧义自动消解研究>>

图书基本信息

书名：<<现代汉语词语级歧义自动消解研究>>

13位ISBN编号：9787030236463

10位ISBN编号：7030236467

出版时间：2008-12

出版时间：科学出版社

作者：曲维光

页数：255

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

前言

欣闻曲维光博上的专著《现代汉语词语级歧义自动消解研究》即将出版，我由衷地感到高兴。
曲维光博士要我写个序言，实在是盛情难却。
为他人的著作作序，在我的学术生涯中还是第一次。
我以为，写“序言”是一件极其困难的任务，不仅要领会全书的精要，还要了解相关学科的全局以及该书对学科发展的贡献。
就能力和精力而言，我确实难以胜任。
然而，曲维光博士2006年初进北京大学汁算机科学技术博士后工作站，两年期『日J与我密切合作。
他不仅刻苦努力，勤于思索，出色完成了博士后研究任务，为我承担的973课题“文本内容理解的数据基础”贡献了力量；而且富有协作精神，与北京大学汁算语言学研究所师生结下了深厚的友谊。
同时，我知道曲维光博士的导师陈小荷教授已经为本书写了序言，相信“序言”的任务已经完成。
我自觉压力不那么大了，只不过是再加上自己的读后感而已。
当前自然语言处理研究的主攻方向，是让机器能够自动地识别和消解自然语言的歧义。
曲维光博士的研究重点是词语级的各种类型的歧义消解，这是自然语言处理研究的基本课题，已经研究很多年了，但没有彻底解决，甚至离彻底解决尚有很长的路要走。
这种情况一方面i兑明，这里有创新的机会和发展的空间，另一方面也i兑明，创新和发展的难度很大。
可以说，曲维光博士是在打攻坚战。
任何一个语言单位脱离其语境（不妨狭义地理解为该语言单位的上下文）都有可能产生歧义，消解歧义的所有方法都要利用其语境信息。
不同的问题、不同的方法所利用的语境的范围各不相同。
就词语级歧义而言，语境通常约束为研究对象在语句中左右相邻的若干个词语。
曲维光博士提出的浯境计算模型RFR_SUM利用了研究对象在整个语料库中的相关信息，取得了很好的消歧效果。
这是本书最重要的创新成果，值得向读者推荐。
在这里试做一个浅显的解说。

内容概要

《现代汉语词语级歧义自动消解研究》提出基于词语搭配强度计算的语境计算模型RFRSUM（SUMofRelativeFrequencyRatio），用于处理各类词语级的歧义消解问题。
各章节的顺序大致勾勒出该模型形成和发展的轨迹。
提出广义组配理论框架，并据此建立语境计算模型RFR_SUM，用以处理语言中广泛存在的词语级歧义现象。
将RFR—SUM模型应用于中文信息处理中的组合型切分歧义和交集型切分歧义的消解、兼类词的消解、多音词的消解以及词义消歧、语料库精加工、隐喻识别等多项任务中，均取得满意的结果，验证了该理论的普适性。
《现代汉语词语级歧义自动消解研究》可以作为从事自然语言处理和计算语言学相关研究人员的参考书。

书籍目录

序一序二绪论1 自然语言处理的根本问题2 词语搭配问题的研究3 本书的主要研究内容第1章 词语组配的研究现状1.1 汉语词语组配及其性质1.2 国外词语搭配研究现状1.3 国内词语搭配研究现状第2章 词语搭配的自动抽取研究2.1 词语搭配的抽取方法2.2 搭配抽取框架的建立2.3 实验及其结果第3章 广义组配理论3.1 广义组配理论的提出3.2 语境的可计算性第4章 语境计算模型RFR_SUM4.1 相对词频比RFR4.2 基本RFR_SUM模型第5章 RFR_SUM模型在分词消歧中的应用5.1 RFR_SUM模型应用于组合型消歧5.2 RFR_SUM模型应用于交集型消歧第6章 兼类词与多音词的消歧6.1 RFR_SUM模型在兼类词消解中的应用6.2 基于RFR_SUM模型的多音词的消歧第7章 词义消歧研究7.1 RFR_SUM模型在词义消歧中的应用7.2 无需词性标注语料的词义消歧实验第8章 词义消歧的二元模型及集成研究8.1 81_RFR_SUM模型8.1.1 二元搭配强度和二元相对词频比（B1_RFR）8.1.2 81_RFR_SUM模型8.1.3 实验及结果8.2 UNl_RFR_SUM与Bl_RFR_SUM的集成8.3 多分类问题研究第9章 超大规模语料精加工技术研究9.1 问题的提出9.2 现有标注软件的性能指标的计量研究9.2.1 ICTCl.AS系统标注结果分析9.2.2 系统改进探讨9.3 语料精加工的方法9.3.1 词表校对法9.3.2 基于简单词语组合特性的方法9.3.3 基于多元组比对的方法9.3.4 基于RFR_SUM模型的方法9.4 初步实验结果第10章 隐喻识别研究10.1 隐喻研究现状10.2 隐喻研究的意义l0.3 隐喻研究的内容和方案10.4 初步的研究成果结语1 本研究完成的主要工作2 进一步研究计划主要参考文献附录l 北京大学汉语文本词性标注集附录2 组合型切分歧义强弱势比例附录3 “从小／学”圳练用例句附录4 “应／用于”训练用例句附录5 “应用于”测试集．_附录6 “从小学”测试集附录7 “科学”词性标注开放测试中标注错误句子附录8 “黄色”词义消歧中错误句子附录9 “黄金”讧J义消歧中错误句子附录10 经改进后，“黄金”词义消歧中错误句子附录11 经改进后，“黄色”词义消歧中错误句子附录12 “黄色”词义开话测试错误句子附录13 “黄金”词义开放测试错误句子附录14 “分子”分类错误的句子附录15 “材料”分类错误的句子-_附录16 “着／u”和“着／v”校对出错洪的句子附录17 “本书／r”和“本／q书／n”校对对错误的句子后记

章节摘录

第2章 词语搭配的自动抽取研究　　对于词语搭配的自动抽取，国外较早开展了相关领域的研究。
Smadja的Xtract系统是迄今为止关于搭配定量分析最为成功的工作。
在Xtract系统中，Smadja提出了度量词语对之间搭配强度的计算公式，引入了位置信息以及相关统计数据分布的离散度计算公式，集成了语料库语言学中词性自动标注技术，在一个规模为一千万词语的股票市场闻报告语料库上运行Xtract得到的结果显示，搭配提取的准确率达到80%。
　　我国学者也在汉语词语搭配研究领域做了大量辛勤的工作，出版了多部词语搭配词典。
但这些词典的编纂，主要还是使用手工抽词的方式，其搭配词语的客观性、覆盖度，以及对中文信息处理的贡献都有待进一步检验。
对中文词语搭配的自动抽取研究相对比较少，其中以孙松的工作最为完整和深入，但其算法自动发现搭配的准确率只有33。
94%。
这对于建立大规模词语搭配知识库来说，无疑会加重人工校对的负担，而且使搭配获取的客观性受到影响。
　　对现有中文词语的搭配抽取方法进行研究，发现需要在以下几个方面加以改进：　　（1）实验所用的语料，大多只经过分词处理，没有经过词性标注，这使得语料中缺少了搭配所需的重要信息。
　　（2）抽取搭配词汇的同时，没有抽取出搭配的结构信息。
　　（3）搭配抽取方案中没有充分利用语言学知识。
　　……

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com