模糊限制信息检测研究

The Research of Hedge Detection

周惠巍 黄德根 著

大连理工大学出版社

2013 国家自然科学基金项目

模糊限制信息检测研究

The Research of Hedge Detection

周惠巍 黄德根 著

大连理工大学出版社

图书在版编目（CIP）数据

模糊限制信息检测研究 / 周惠巍，黄德根著. — 大连：大连理工大学出版社，2017.3

ISBN 978-7-5685-0677-9

Ⅰ. ①模… Ⅱ. ①周… ②黄… Ⅲ. ①模糊语言学－研究 Ⅳ. ① H087

中国版本图书馆 CIP 数据核字（2017）第 008746 号

大连理工大学出版社出版

地址：大连市软件园路 80 号　邮政编码：116023

发行：0411-84708842　邮购：0411-84708943　传真：0411-84701466

E-mail：dutp@dutp.cn　URL：http：//www.dutp.cn

大连金华光彩色印刷有限公司印刷　　大连理工大学出版社发行

幅面尺寸：147mm × 210mm　印张：6　字数：150 千字

2017 年 3 月第 1 版　2017 年 3 月第 1 次印刷

责任编辑：王晓历　责任校对：康 佳

封面设计：张 莹

ISBN：978-7-5685-0677-9　定 价：32.00 元

前言

模糊性是人类语言的一种属性。由于各种局限性，在语言交流以及科技论文写作中，常常借助模糊限制语来表达不确定性和可能性的含义。模糊限制语最早是由 G. Lakoff 提出的，用来指“把一些事情弄得模模糊糊的词语”。由模糊限制语所引导的信息为模糊限制性信息。为进行事实信息的挖掘，应将模糊限制信息与事实信息区分开来。作为信息抽取的一个重要环节，模糊限制信息检测旨在区分文本中的模糊限制信息与事实信息，避免将模糊限制信息作为事实信息用于信息抽取。

随着网络信息量的迅猛增长以及信息抽取技术的快速发展，作为信息抽取源的网络信息的真实性和可靠性日益受到关注。模糊限制信息检测包含模糊限制语识别和模糊限制信息范围检测两个子任务。2010 年，国际计算语言学协会将模糊限制语识别和模糊限制信息范围检测定为 CoNLL（Conference on Computational Natural Language Learning）的共享任务。

近年来，随着大规模英文模糊限制语语料库的构建，英文模糊限制信息检测研究已经取得了一定的进展，但是模糊限制信息范围检测性能尚未达到实用化要求。这是由于模糊限制信息范围检测任务比较复杂，具有依赖语义和句法结构的特点，单纯基于一个统计模型难以满足它的处理需求。融合方法可以将自然语言处理任务中的多类特征、多种方法、多个模型有效地结合起来，避免单一模型的片

面性，实现自然语言处理的准确性。

中文模糊限制语被广泛地应用于中文各个领域，但是中文模糊限制信息检测的研究却非常少。如果不借鉴英文模糊限制信息的相关知识，从零开始中文模糊限制信息检测研究，将难以促进其研究发展。能否针对中文模糊限制信息的特点，利用已有的英文模糊限制信息检测研究成果，辅助中文模糊限制信息检测研究，实现跨语言的模糊限制信息检测成为摆在我们面前的一个亟待解决的课题。

模糊限制语被广泛地应用于医学文献、百科全书和新闻报道等各个领域。然而，不同领域使用的模糊限制语及其使用频率均不相同。在每个领域标注大量的训练语料耗时耗力，同时也给模糊限制信息检测的应用和推广带来一定困难。应用已有领域的语料实现跨领域模糊限制信息检测，以及应用有限的训练语料提高模糊限制信息检测性能，对于模糊限制信息检测的应用和推广具有重要意义。

基于以上问题，本书主要阐述了英文模糊限制语识别及模糊限制信息范围检测研究、中文模糊限制信息语料库的构建研究、跨语言模糊限制语识别研究、跨领域中文模糊限制语识别研究以及中文模糊限制信息范围检测研究。

本书共分为 8 章，具体研究内容如下：

第 1 章 本章主要介绍研究背景和意义、研究现状以及常用的统计机器学习方法和半监督学习方法。

第 2 章 本章研究基于复合核函数融合结构特征与平面特征的模糊限制信息范围检测。复合核函数包含多项式核函数和卷积树核函数两个独立的核函数。利用多项式核函数挖掘平面特征信息，构建模糊限制信息范围检测模型；利用卷积树核函数捕捉句法结构信息，建立模糊限制信息范围检测模型。重点探索了模糊限制范围的结构化信息表达形式，降低了传统的将结构信息平面化时导致的信息丢失的可能。采用句法结构特征的卷积树核函数与采用平面特征的多项式核函数融合，得到的复合核函数能显著提高

模糊限制信息检测性能。

第 3 章 本章研究基于规则和统计相结合的模糊限制信息范围检测。结合系统以统计为主，以规则为辅，包含支持向量机（Support Vector Machine, SVM）统计子系统、规则子系统以及条件随机域（Conditional Random Field, CRF）融合系统三部分。其中，SVM 统计子系统利用短语结构检测模糊限制信息范围；规则子系统利用依存结构检测模糊限制信息范围；CRF 融合系统将两个子系统的检测结果作为两个独立的特征对规则和统计进行融合，这种将规则和统计相结合的模糊限制信息范围检测有效利用了依存结构和短语结构，实现了 CRF 和 SVM 两种机器学习方法的融合。

第 4 章 本章研究多分类器融合的模糊限制信息检测，提出一种基于投票策略的模糊限制信息范围检测方法。首先分别基于 SVM、CRF、M3N 以及规则和统计相结合的方法，以前向和后向两个解析方向构建 8 个基本分类器，再分别采用多数投票、分类器加权投票和词性加权投票 3 种投票策略融合 8 个基本分类器的分类结果。

第 5 章 本章研究中文模糊限制语的分类，并在生物医学和维基百科两个领域，设计构建中文模糊限制语及其限制信息范围语料库。语料库分别标注了模糊限制语及每个模糊限制语的限制范围。这些资源对于中文模糊限制信息检测研究以及中文事实信息的抽取具有重要意义。同时，为语言学家从语义和语用等方面进行模糊限制语的研究提供了强大的知识库支持。

第 6 章 本章研究基于双语语义信息和模糊限制信息联合表示学习的跨语言模糊限制语识别。在词语义表示基础上，引入上下文模糊限制信息表示，联合表示学习在特定语境下的模糊限制信息。将双语语义学习和模糊限制信息学习合为一个学习阶段，提高跨语言模糊限制语的识别性能。

第 7 章 本章研究跨领域中文模糊限制语识别。首先，结合迁移学习和半监督学习，利用大量源领域标注数据、极少量目标领域标注数据以及大量的目标领域未标注数据训练获取跨领域模糊

限制语分类模型。然后，基于词向量和迁移学习相结合进行跨领域中文模糊限制语识别。

第8章 本章研究中文模糊限制信息范围检测。首先，分别基于复合核模型挖掘词法和句法信息，基于长短时记忆递归网络（Long-Short Term Memory，LSTM）模型挖掘语义信息。然后，将复合核模型和LSTM模型整合到一个统一的框架中，基于混合系统检测模糊限制信息范围。

本书全面细致地阐述了英文和中文模糊限制信息检测研究。本研究为模糊限制信息检测研究以及自然语言处理领域的数据挖掘研究提供了有益的借鉴。开展模糊限制信息检测研究，有助于提高抽取信息的真实性和可靠性，对事实信息抽取意义重大。

本书得到了国家自然科学基金项目“基于翻译学习和核方法的中文模糊限制信息检测研究”（61272375）的资助。大连理工大学学生张严、杨欢、陈龙、邓会杰、徐俊利、杨云龙、刘壮和宁时贤等参与了研究中的创新实验、文稿撰写等工作。借此书稿付梓之际，仅致谢忱！

在编写本书的过程中，我们参考、借鉴了许多专家、学者的相关著作，对于引用的段落、文字尽可能一一列出，谨向各位专家、学者一并表示感谢。

限于水平，书中仍有疏漏和不妥之处，敬请专家和读者批评指正，以使教材日臻完善。

著 者

2016年12月

目 录

第1章 绪 论

模糊性是人类语言的一种属性，由于各种局限性，在语言交流、科技论文写作特别是生物医学论文写作中，常常借助模糊限制语来表达不确定性和可能性的意义。由模糊限制语所引导的信息为模糊限制性信息（Hedge Information）。为进行事实信息（Factual Information）的挖掘，应将模糊限制信息与事实信息区分开来。生物医学领域的模糊限制信息检测成为生物医学信息抽取的首要步骤。

本章详细叙述了模糊限制信息检测的研究背景和意义，介绍了模糊限制信息的相关概念、研究现状和发展趋势，说明了本书的主要工作以及贡献。

1.1 研究背景和意义

自然语言处理（Natural Language Processing, NLP）："利用计算机作为工具对人类特有的书面形式和口头形式的语言进行各种类型处理和加工的技术"[1]。自然语言处理又称自然语言理解（Natural Language Understanding, NLU）、计算语言学（Computational Linguistics, CL）等。

自然语言处理存在两种不同的研究方法：理性主义方法（Rationalist Approach）和经验主义方法（Empiricist Approach）[2]。理性主义方法相信在人类头脑中重要的知识不是由感官得到的，而是提前固定在头脑中，由遗传基因决定的。理性主义方法希望建立一个智能系统，在这个智能系统中通过手工编码大量的先验知识和推理机制，得以复制人类大脑中的语言能力。由于用于自然语言处理的符号系统通常表现为规则的方式，

因此理性主义方法在自然语言处理中又常常被称为基于规则的方法（Rule-based Method）。其基本思想是建立语言规则库，根据语言规则分析语言。然而，由于自然语言具有认识性、复杂性和不确定性，自然语言的现象并不能全部用确定性的规则来刻画，而且规则的使用同样具有不确定性。

经验主义方法同样假设大脑中存在某些认知的能力，该方法和理性主义方法的区别不是绝对的，只是在某种程度上有所区别。经验主义方法不同于理性主义方法之处在于，它认为人类的智能不是固定在头脑中的规则集，也不是针对各种各样语言结构和其他感知领域的程序集。经验主义方法认为可以使用概率的方法研究语言，将统计学和机器学习方法应用于大规模语料，学习获得语言的概率模型。在自然语言处理领域，经验主义方法被称为统计自然语言处理方法（Statistical Natural Language Processing），或简称为基于统计的方法（Statistic-based Method）[2]。

统计自然语言处理方法在处理大规模真实文本方面，表现出了规则方法无法比拟的优越性。从20世纪80年代后期开始至今，机器学习方法得到快速的发展，K-近邻分类器（K-Nearest Neighbor，KNN）[3-5]、决策树（Decision Tree，DT）[6,7]、最大熵（Maximum Entropy，ME）[8,9]、支持向量机（Support Vector Machines，SVM）[10,11]、条件随机域（Conditional Random Fields，CRF）[12]、最大间隔马尔可夫网络（Max-Margin Markov Networks，M3N）[13]等模型相继提出，并应用于自然语言处理任务。

随着计算机技术和生物技术的高速发展，生物医学领域的文献和数据正在以指数方式增长。例如，美国国家医学图书馆（The National Library of Medicine，NLM）提供的在线生物医学文献数据库MEDLINE，收录了自1966年以来70多个国家和地区出版的生物医学期刊文献近2000万篇，而且每天以2000篇以上的速度在增长。

大规模的生物医学文献和数据是信息和知识的来源，但不等于信息和知识。与正在以指数级增长的生物医学文献相比，生物医学知识的增长却十分缓慢。研究人员迫切渴望从大量的生物医学文献中挖掘新的知识，推动生物医学的发展，帮助人们改善生活环境，提高生活质量。这一需求推动了生物医学信息抽取（Information Extraction, IE）技术的产生与发展。

生物医学信息抽取技术的主要任务是从生物医学文本中抽取出特定的事实信息。而在生物医学领域，由于各种局限性，当信息的撰写者不可能提供完全准确、肯定的信息时，往往使用模糊限制语（hedges），使自己的陈述更客观，避免使用绝对的方式阐述自己的观点。模糊限制语，这个术语最早由 G. Lakoff 提出，用来指那些“把一些事情弄得模模糊糊的词语”，表示的是不确定性、临时性和可能性的意义[14]。

在生物医学论文中，模糊限制语使命题的表达更严谨，更周全；在临床诊断中，模糊限制语的使用可以起到保护诊断者，降低诊断者所承担的责任等作用。在生物医学领域，模糊限制语是一种常用的语言使用策略。

统计表明，在 MEDLINE 摘要中，11% 的句子包含模糊限制信息[15]；而在用于模糊限制信息检测研究的 BioScope 语料中，摘要中 17.69% 的句子、正文中 22.29% 的句子包含模糊限制信息[16]；Szarvas[17] 统计指出，在 Medlock 和 Briscoe[18] 标注的模糊限制信息测试语料中，37.57% 的基因名（1698 个基因名中的 638 个）出现在含有模糊限制信息的句子中。这表明，基因关系抽取系统产生的许多错误正例是由于未进行模糊限制信息检测导致的。

为进行事实信息抽取，模糊限制信息检测成为生物医学信息抽取的首要步骤，因此引起了国内外许多研究人员的广泛关注，并成为生物医学领域信息抽取的一个热点和关键研究问题。

国际计算语言学协会（Association for Computational Linguistics，ACL）将模糊限制性句子识别及模糊限制信息范围检测（Learning to detect hedges and their scope in natural language text）定为2010年CoNLL（Conference on Computational Natural Language Learning）共享任务[19]。CoNLL-2010共享任务包含两个子任务，任务一为识别含有模糊限制信息的句子，包含生物医学和维基百科两个领域；任务二为确定所识别的模糊限制语的模糊限制信息范围，仅包含生物医学领域。此次共享任务吸引了来自全世界二十多个研究机构的参与，其中包括国外的斯坦福大学、剑桥大学、哥伦比亚大学、东京大学等科研院校，国内的哈尔滨工业大学、大连理工大学、复旦大学、上海交通大学、国防科技大学等科研院校。研究人员进行了大量的研究，会议也取得了一些重要的研究成果，但仍然存在模糊限制语识别，尤其是模糊限制信息范围检测准确率和召回率偏低的问题。究其原因，主要是由于模糊限制信息存在模糊性和复杂性的特点。

模糊限制语识别的难点在于：

（1）同一词语根据语境的不同，在有些句子中是模糊限制语，而在有些句子中却不是模糊限制语。

（2）模糊限制语不但包含单个词语，还包含多词短语，这样导致模糊限制语的边界难以确定。

（3）有些模糊限制语不连续，如“either…or…”是一个模糊限制语，因为其在句子中的位置不连续，所以很难正确识别。

模糊限制信息范围检测的难点在于：

（1）模糊限制语识别是模糊限制信息范围检测的首要任务，模糊限制语识别错误必然导致模糊限制信息范围检测错误。

（2）虽然模糊限制信息范围的左右边界相互照应、相互联系，但是其左右边界往往距离较远，而传统的建模方法的特征窗口有限，因此标注其左（右）边界时，很难引入其右（左）边界信息。

（3）模糊限制信息范围检测有依赖于语义和句法结构的特点，这两点一直是自然语言处理公认的研究难点。

（4）有的句子中包含多个模糊限制语，每个模糊限制语对应一段模糊限制范围，一段模糊限制信息可以是另一段模糊限制信息的子串，模糊限制信息范围存在嵌套关系。

（5）模糊限制语在其模糊限制范围内的位置不确定，有时位于模糊限制信息范围的开始位置，有时位于结束位置，有时位于中间位置。

现阶段使用机器学习方法进行模糊限制语识别和模糊限制信息范围检测成为模糊限制信息检测研究的主流方法。然而由于模糊限制信息检测的复杂性和模糊性，基于单一分类器或基于规则的方法，均难以满足模糊限制信息检测的需求。而从近期的研究成果来看，多种特征和模型的融合，尤其是各种机器学习方法的融合，已经成为改进和提高自然语言处理系统性能的有效途径。因此，一种可能的改进检测系统性能的途径是采用融合的方法来建立模糊限制信息检测的综合性模型。

中文生物医学文献和维基百科等领域同样包含大量模糊限制信息。研究发现，基于维基百科抽取的部分错误的实体关系源于未进行模糊限制信息检测。如“许多人认为悉尼是澳大利亚的首都”为模糊限制性句子，因为未进行模糊限制信息检测，导致错误地将“悉尼－澳大利亚”抽取为首都－国家关系。未经模糊限制信息检测，直接进行信息抽取，将获得大量不可靠信息，严重影响中文信息抽取的准确性。开展中文生物医学领域及其他领域的模糊限制信息检测研究，有助于提高抽取信息的真实性和可靠性，对中文事实信息抽取意义重大。

本书将研究基于多种融合方法的英文模糊限制信息检测，重点探讨多种类型特征的融合、规则方法和统计方法的融合、多分类器的融合三种融合方法对模糊限制信息检测系统性能的影响，

从而构造高性能的模糊限制信息检测系统。近年来，融合方法得到了自然语言处理、模式识别等领域专家的广泛关注，它是自然语言处理发展到一定阶段的必然课题。模糊限制信息检测中的融合方法研究，不但能提高模糊限制信息检测性能，而且能为今后自然语言处理领域中融合方法的研究提供有益的借鉴。

英文模糊限制信息检测已经取得了一些有价值的研究成果，但是中文模糊限制信息检测研究尚处于起步阶段，存在一些困难和挑战。本书针对中文模糊限制信息的特点，利用已有的英文研究成果，辅助中文模糊限制信息检测研究，实现跨语言的模糊限制信息检测。同时，模糊限制语还广泛地用于中文的各个领域。模糊限制语的使用具有领域性。中文模糊限制信息检测是一个领域性较强的任务，在一个领域中训练出来的分类模型，通常不能适用于其他的领域。然而，针对每个领域分别标注大规模的训练数据是不现实的。本书研究利用中文已有领域的训练语料，检测其他领域的模糊限制信息，实现跨领域的模糊限制信息检测。最后，针对中文词法、句法和语义的特点，分别基于复合核模型挖掘词法和句法信息，基于长短时记忆递归网络（Long-Short Term Memory，LSTM）模型挖掘语义信息。并将复合核和 LSTM 模型整合到一个统一的框架中，基于混合系统检测中文模糊限制信息范围。

1.2 模糊限制信息检测研究综述

1.2.1 模糊限制语的分类

模糊限制语是美国语言学家 Lakoff 于 1972 年首次提出的，他认为模糊限制语的作用是增加或减少语言的模糊程度[14]，即其