文章编号：1003-0077(2015)06-0083-07

中文模糊限制语语料库的研究与构建

周惠巍[1]，杨 欢[1]，张 静[2]，亢世勇[2]，黄德根[1]

(1. 大连理工大学 计算机科学与技术学院，辽宁 大连 116024；2. 鲁东大学 文学院，山东 烟台 264025)

摘 要：模糊限制语常用来表示不确定性和可能性的含义，由模糊限制语所引导的信息为模糊限制信息。为进行中文事实信息的抽取，应将模糊限制信息与事实信息区分开来。然而中文模糊限制语语料资源却十分缺乏，影响了中文模糊限制语和模糊限制信息检测的研究。该文研究了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，设计构建了一个具有 2.4 万句规模的中文模糊限制语语料库。统计分析了语料标注的一致性，以及模糊限制语的类型和领域之间的关系。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。

关键词：中文模糊限制语；分类；语料库；一致性分析

中图分类号：TP391 **文献标识码**：A

The Research and Construction of Chinese Hedge Corpus

ZHOU Huiwei[1], YANG Huan[1], ZHANG Jing[2], KANG Shiyong[2], HUANG Degen[1]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China;
2. School of Liberal Arts, Ludong University, Yantai, Shandong 264025, China)

Abstract: Hedge is usually used to express uncertainty and possibility. When authors cannot back up their statements, they usually use hedge to express uncertain information. To avoid extracting uncertain statements as factual information, uncertain information should be distinguished from factual information. However, inadequate Chinese hedge corpus limited the research of Chinese hedge. This paper discusses the categorization of Chinese hedge, introduces the design and construction of a 24,000-sentence Chinese hedge corpus in the biomedical and Wikipedia domains. We calculate agreement rates for the corpus and reveal the domain and genre dependency of hedges. The construction of the corpus is of great significance in the research of Chinese hedge detection and Chinese information extraction. Meanwhile, the resource provides a great support for linguists to study the semantic hedge and the pragmatic hedge.

Key words: Chinese hedge; categorization; corpus; agreement analysis

1 引言

模糊限制语(Hedges)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”，表示的是不确定性和临时性的观点[1]。由模糊限制语所引导的信息为模糊限制信息(Hedge Information)。当信息的撰写者不能提供完全准确、肯定的信息时，往往使用模糊限制语，使自己的陈述更客观。模糊限制语广泛地用于自然科学、新闻等各个领域，为进行事实信息的抽取，应将模糊限制信息与事实信息区分开来，模糊限制信息检测成为信息抽取的首要步骤。由于网络信息量的迅猛增长，及信息抽取技术的高速发展，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。2010 年国际计算语言学会(Association for Computational Linguistics, ACL)将模糊限制语识别和模糊限制信息检测定为 CoNLL (Conference on Computational Natural Language Learning)共享任务[2]。其中模糊限制语识别包含生物医学和维基百科两个领域。生物医学

收稿日期：2014-09-05 定稿日期：2014-10-30

基金项目：国家自然科学基金(61272375,61173100)

领域源语料源自 BioScope[3] 语料库，该语料库包括生物医学论文摘要、全文及临床诊断报告三类文献；维基百科语料源自 WikiWeasel[2] 语料库。

英文模糊限制语语料库的研究与构建取得了长足的进展，除上述两个语料库外，公开发表的语料库还有 Medlock 和 Briscoe[4]，Kim 等[5]，Settles 等[6]，Shatkay 等[7]，Nawaz 等[8]，和 Uzuner 等[9]构建的生物医学领域语料库；Rubin 等[10]，Wilson[11]，Saurí 和 Pustejovsky[12]，和 Rubin[13]构建的新闻领域语料库。王舟[14]调查了中英文医学论文摘要各 80 篇，中文论文中模糊限制语累计出现 205 次，英文论文中出现 305 次。中文医学文献中同样包含大量模糊限制语[15-17]。除医学文献外，模糊限制语还广泛地用于中文的各个领域。维基百科作为一个以开放和用户协作编辑为特点的知识系统，其中蕴涵了丰富的信息，成为目前研究人员进行信息抽取的重要语料资源。但是当撰写者不能提供完全准确、肯定的信息时，往往使用模糊限制语，使自己的陈述更客观。

语言学界从语义、句法、词性等方面对模糊限制语进行了长期的研究，中英文研究人员分别对中英文模糊限制语的表现形式和分类进行了探讨，将模糊限制语从语义、词性、结构和句法功能等方面进行分类。英文模糊限制语的研究开始于 20 世纪 70 年代，Prince 等[18]从语用功能上将英语模糊限制语划分为变动型（approximators）和缓和型（shields）。前者改变原话题的真值条件，对话题进行某种程度或范围的限制，如“a little bit”、“almost”等。而后者不能改变原话题的真值条件，但可以反映说话人对话题所持有的态度，缓和了话题的肯定语气，如“think”、“perhaps”等。Szarvas 等[19]根据话题在真实世界的真假性，将英文模糊限制语分为假设型（hypothetical）与认知型（epistemic）两类。两者的主要区别是前者认为话题在真实世界里可能为真、假或者不确定三种情况，如“He believes that the Earth is flat.”而后者则是就目前所知无法判断话题在真实世界是正确的还是错误的，如“It may be raining.”

我国英语界于本世纪 80 年代对模糊限制语进行了初步探讨，何自然[20]在 Prince 等[18]的研究基础上，将变动型模糊限制语细分为程度变动型和范围变动型，将缓和型模糊限制语细分为直接缓和型和间接缓和型。认为变动型模糊限制语属于语义范畴，缓和型模糊限制语属于语用范畴。语言学家对中英文模糊限制语进行了翻译研究[21]，认为中英文模糊限制语存在一定的差异，很少存在等值译文。苏远连[22]对中文模糊限制语进行了对比研究，他赞同何自然的观点，认为中文模糊限制语也可以按照同样的方法进行分类，并将变动型模糊限制语细分为程度变动型，如“有点”、“相当”等；范围变动型，如“上下”、“左右”等；和频率变动型，如“经常”、“不时”等。缓和型模糊限制语仍然分为直接缓和型，如“我认为”、“看来”等；和间接缓和型，如“听说”、“据报道”等。

目前国内模糊限制语呈现出多理论、多角度和多方面的研究[23]。模糊限制语语料库的构建是模糊限制语研究与模糊限制信息检测的基础，然而中文模糊限制语语料资源缺乏，至今尚未发现公开发表的模糊限制语语料库。本文研究了中文模糊限制语的分类，设计并构建了一个具有 2.4 万句规模的中文模糊限制语语料库。语料选自生物医学和维基百科两个领域，生物医学文献包括摘要、实验结果、讨论、结论和全文五个部分，维基百科文献选取了包含国家介绍、历史人物介绍、事件介绍等 242 篇文章。实验分析了语料标注的一致性，并统计了不同领域各类模糊限制语的使用比例。本文构建的中文模糊限制语语料库，涵盖了丰富的中文模糊现象，为语言学家从语义、语法、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。语料库中的医学文献和维基百科文献分别包含 9 946 个和 1 958 个模糊限制语，在各自的领域足以训练出一个比较准确的模糊限制语识别模型，用于模糊限制信息检测研究。同时，还可以应用两个领域的语料库进行跨领域模糊限制语识别研究。

本文组织结构如下：第二节提出了本文对中文模糊限制语分类方法；第三节阐述了语料库的构建过程；第四节统计分析了语料标注的一致性，以及模糊限制语种类和语料领域之间的关系；第五节是结论与展望。

2 中文模糊限制语的分类

本文根据 Prince 等[18]和何自然[20]的分类方法，将模糊限制语分为变动型和缓和型两类。在此基础上，根据模糊限制语的语义和语用功能，将这两大类模糊限制语进行了更细致的划分，如图 1 所示，各类模糊限制语的定义如下。

（1）变动型模糊限制语

变动型模糊限制语是对话题本身进行某种程度的限制，它能修改话题原来的真值，当说话人不能准确说出某个话题的真值或有意模糊某个话题的真值

时用到变动型模糊限制语。根据变动话题的类型，此类模糊限制语可细分为数量变动、程度变动、范围变动和频率变动四个类型。

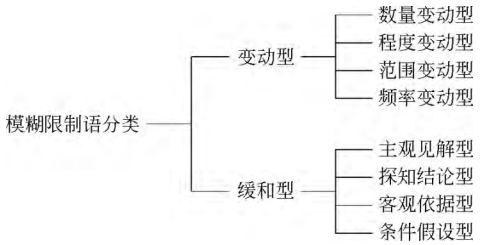

图 1　中文模糊限制语的分类

数量变动型　当说话人不能明确地说出具体的数字，但是能估计出一个大概的数量时，常会用到数量变动模糊限制语。如："少数"，"大部分"等。

程度变动型　把一些接近正确但不敢肯定完全正确的话题说得更得体些，与实际情况更接近些，避免过于武断，表明话题与真实情况的接近程度。如："有点"，"稍微"，"十分"等。

范围变动型　在话题中往往提供了具体数字，使用这类模糊限制语时，听话人不必考虑具体情况与所说的话题的接近程度如何，而只考虑范围大小，听话人可以在一定的范围内去理解话题意义。如："大约"，"在一定范围内"，"将近"等。

频率变动型　用于反映一个事件发生的频率。如："常常"，"偶尔"等。

（2）缓和型模糊限制语

当说话人提出某一个论断时，缓和型模糊限制语可以缓和说话人的语气，为说话人留有余地，减轻说话人为此论断所应付的责任，这类模糊限制语不改变话题原来的意思。根据缓和型模糊限制语的语用功能将其细分为主观见解型、探知结论型、客观依据型和条件假设型四类。

主观见解型　用来表示说话人对某事的直接推测及所持的态度，其所阐述的话题只是个人的主观见解。使用这类模糊限制语可以在一定程度上削弱说话人对话题所承担的责任。如："我认为"，"就我所知"等。

客观依据型　通过借助第三方或大家普遍认同的观点，间接地表达说话人对某事所持有的态度，说话人在一定程度上同意第三方的观点，只是他对此依据究竟有多大程度的赞同，在话语中看不出来，只能另作推断。例如，据说"，"有人说"等。

探知结论型　用来表示对某个结论的推测，根据存在的现象推知未来可能会发生的事情或待证明的结论。例如，"表明"，"可能"，"调查"，"仍不清楚"等。

条件假设型　通过给出假定的前提条件表明说话人的意愿，但现在事实是怎样的并不知晓。例如，"如果"，"假定"等。

3　中文模糊限制语语料库构建

3.1　语料的选取与预处理

本文构建的中文模糊限制语语料库覆盖了生物医学与维基百科两个领域。生物医学领域语料选自《中国生物医学工程学报》、《中国生物化学与分子生物学报》和《生物医学工程学杂志》等权威性中文生物医学类期刊的 2011～2013 年间的科研论文。分别摘取部分文献的摘要、实验结果、讨论、结论并选取部分文献的全文，分别标注以便统计分析模糊限制语在文献不同章节的使用频率。

维基百科的组成单元称为"概念"或"词条"，每个词条对应一篇文章，由不同用户一次次编辑形成。本文选取国家介绍、历史人物介绍、事件介绍等方面的 242 篇词条构建维基百科语料库。

从 CNKI（中国知网）上下载的文献需要转化为文本格式。人工修正文本转化产生的乱码，并将其中的所有英文及数字统一为半角格式，去掉多余的空格。由于存在中英文标点符号混合使用的情况，将所有标点符号统一为中文格式。

3.2　结构设计与一般标注规则

标注语料采用一种特定的 XML 格式，每一个句子显示为一行，如图 2 所示。首先，标注句子号，

```
<Document type="Wikipedia_text_weasel">
  <DocID> 38</DocID>
    <DocumentPart type="Title">……</DocumentPart>
      <sentence id="S38.1" certainty="certain">……</sentence>
      ……
      <sentence id="S38.9" certainty="uncertain">希波克拉底使用体内平衡理论来解释为何疾病和压力<ccue id="38.9.1" type="频率变动">总是</ccue><ccue id="38.9.2" type="探知结论">被定义</ccue>为一种对威胁一个系统的平衡或生理平衡的状况的反应。</sentence>
```

图 2　模糊限制语标注语料示例

如“S38.9”,其中“38”为文章号,“9”为该句在文章中的序号。然后,标注该句的模糊限制类型,其中“certain”为确定性句子,“uncertain”为模糊限制性句子,当一个句子包含有一个或一个以上的模糊限制语时,这个句子就是模糊限制性句子。对于模糊限制性句子,标注模糊限制语。分别使用标记“<ccue>”和“</ccue>”标注模糊限制语的起始和结尾,同时给出模糊限制语的标号,如“S38.9.1”,和细分类,如“频率变动”。文章号、句子序号和模糊限制语标号采用层次结构,并有且仅有一个标号。

模糊限制语的标注遵循最小原则:标注能表明模糊限制性的最小单元为模糊限制语,多个模糊限制语组合起来表示模糊限制性时,分别标注每个模糊限制语,如图2中,“总是”和“被定义”被分别标记为模糊限制语,而不是将“总是被定义”作为一个模糊限制语。

3.3 特殊词语标注规则

除了一些明确具有模糊限制含义的词语外,还有一些词语需要根据上下文语境判断其是否表模糊性,这是模糊限制语标注的一个难点。判断一个词语是否是模糊限制语,主要是看它对所陈述的命题是否产生不确定的影响。为减少标注错误,提高标注速度,增加标注语料的一致性,我们研究制定了一些特殊词语标注规则,这些规则随着标注过程动态更新。部分特殊词语标注规则如下。

1. 词语“根据”引用的是第三方的观点或理论,间接地表达说话人对某事所持有的态度时,认为是模糊限制语,如例句(1)中的“根据”是模糊限制语;当命题中未表达个人观点时,认为不是模糊限制语,如例句(2)。

例句1 <ccue>根据</ccue>染色体分离机理,Cdc20的表达是PBEⅠ所必需的。

例句2 大部分代表是直接民选产生,100人则是根据政党得票率按比例分配。

2. 词语“或者”是在每个领域都经常出现的词,通过研究我们认为当“或者”连接的同位词只有一个正确的时候,是模糊限制语,如例句(3),不确定是“第三或者第四大”,但对的只能选择其中的一个,所以是模糊限制语;而当“或者”连接的同位语无论选哪个都正确的时候,认为不是模糊限制语,如例句(4),选择“tartuffolo”或“小松露”,命题都正确,所以认为不是模糊限制语。

例句3 美国的国土面积是世界第三<ccue>或者</ccue>第四大。

例句4 在十五世纪时,马铃薯在意大利被叫作“tartuffolo”或者“小松露”。

3. 词语“表明”在生物医学文献中常用于推测某个结论,当根据某些现象或条件,推测出一个结论时,认为是模糊限制语,如例句(5),“表明”连接的是一个推测性的结论,所以认为是模糊限制语;当只是客观地描述了一个结果或现象时,认为不是模糊限制语,如例句(6),只是客观地陈述了一个实验的结果,所以认为不是模糊限制语。

例句5 在晚期动脉相因子图中,肿瘤完全增强,并且周围组织也增强,<ccue>表明</ccue>有肝动脉血流开始进入周围组织区域。

例句6 通过对30位受试者的对比实验,结果表明,本监护仪的测量验证的平均准确率达到92.2%。

4. 词语“证明”常出现在生物医学文献中,后面跟随一个命题。我们规定,当该命题需要加以证明时,“证明”是模糊限制语,如例句(7),“b和c两条带为Pil1磷酸化状态”这一命题在此例句中是有待证明的命题,所以认为是模糊限制语;而命题已得到证明了,则“证明”不是模糊限制语,如例句(8),“高糖可以通过线粒体凋亡途径诱导成骨细胞凋亡”这一命题已经通过实验得到了验证,所以认为不是模糊限制语。

例句7 为了<ccue>证明</ccue>这b和c两条带为Pil1磷酸化状态,Fig.2C表示将蛋白提取物加入磷酸酶处理后作免疫印记检测。

例句8 本研究证明,高糖可以通过线粒体凋亡途径诱导成骨细胞凋亡。

3.4 语料库的构建

基于已有的英文模糊限制语语料,和中文待标注语料,收集各类中文模糊限制语,整合成为一份完备的模糊限制语词典。为了减轻标注人员的负担,采用正向最大匹配算法,标注中文语料中的词典词,形成初始标注语料。

本文参照英文生物医学领域的BioScope[3]语料库的标注过程进行标注。首先,分别由两名语言学专家按照标注规则,判断初始标注语料中的词典词在句子中是否表示模糊性,人工修正初始标注语料的错误,形成两份标注结果。然后,规则的制定者对两份标注结果中不一致处进行统一,形成最终语料。具体标注过程如图3所示。

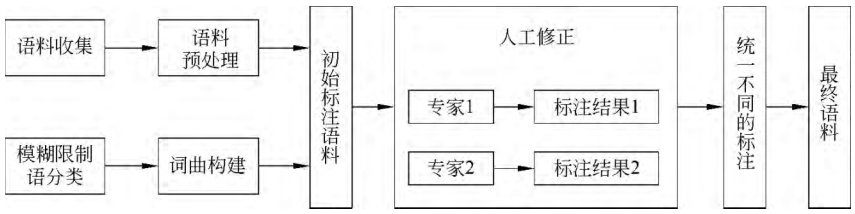

图 3　语料标注过程

4　语料库的统计数据与分析

4.1　语料库的统计数据

语料库的统计信息如表 1 所示，生物医学和维基百科两个领域共标注语料 24 414 句，约 100 万词。其中模糊限制性句子8 160 句，模糊限制语11 904 个。维基百科文章中，33.78％的句子包含模糊限制信息；生物医学文献中，摘要中 25.28％的句子，全文中 35.09％的句子包含模糊限制信息，可见中文模糊限制语广泛地存在于中文文献中。此规模的语料库足够用于模糊限制信息检测的研究。

表 1　语料库的统计信息

统计条目 \ 文献领域	生物医学						维基百科	总数
	摘要	实验结果	讨论	结论	全文	总数		
篇章数	1 316	388	349	141	16	2 210	242	2 452
句子数	8 089	4 277	6 117	655	989	20 127	4 287	24 414
模糊限制性句子比例/％	25.28	27.08	47.69	37.41	35.09	33.35	33.78	33.42
模糊限制语个数	2 759	1 622	4 674	353	538	9 946	1 958	11 904

4.2　不同领域模糊限制语的分布

模糊限制语的分布具有领域性[19]，为了探究不同类型的模糊限制语在生物医学和维基百科领域的分布，对语料库中的各类模糊限制语进行统计，结果如图 4 所示。

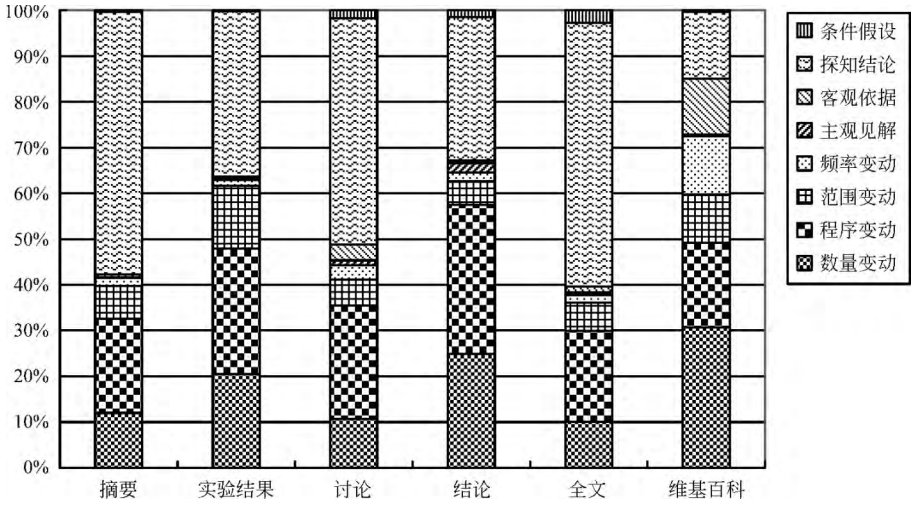

图 4　语料中模糊限制语的类型分布

由图 4 可见，生物医学文献中缓和型模糊限制语的使用频率较高，维基百科文章中变动型模糊限制语的使用频率较高。其中，在生物医学领域，探知结论型模糊限制语较多。因为在生物医学论文写作中，当作者根据实验现象推测结论时，常常使用探知结论型模糊限制语。在维基百科领域，客观依据型模糊限制语所占比例明显高于生物医学领域。其主要原因是本文选取了国家介绍、历史人物介绍、事件

介绍的文章，所以往往借用别人的观点来表述自己态度。一般而言，程度变动型和数量变动型模糊限制语在各个领域都比较常用，因此，这两类模糊限制语在生物医学和维基百科中都占有较大的比重。

4.3 一致性分析

标注完成后，对标注语料进行一致性分析。先比较两份独立标注的语料，将其中一份语料作为标准语料；再分别将两份独立标注的语料（标注结果1，标注结果2）与最终语料进行比较，最终语料作为标准语料。采用式(1)、式(2)和式(3)计算获得F值作为一致率。

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$F = \frac{2\mathrm{PR}}{P+R} \tag{3}$$

上式中，TP(True Positives)表示两份语料中相同的模糊限制语的个数。FP(False Positives)表示评测语料中被标注为模糊限制语，而标准语料中未被标注为模糊限制语的个数。FN(False Negatives)表示评测语料中未被标注为模糊限制语，而标准语料中被标注为模糊限制语的个数。一致性分析结果如表2所示。

表2 一致性分析结果

	模糊限制语级别F值/%	句子级别F值/%
摘要	78.55/91.13/86.6	88.11/96.17/91.66
实验结果	73.9/78.77/85.44	82.22/85.83/89.69
讨论	74.83/78.25/87.62	90.98/93.13/95.59
结论	77.32/82.16/93.23	92.75 94.94/97.6
全文	80.99/83.22/84.5	88.15 /91.88/92.92
维基百科	73.58/88.91/80.01	87.9/95.68/90.67

模糊限制语级别的F值采用精确匹配，即左、右边界完全匹配时认为识别正确，而句子级别的F值只需句子的模糊性判断正确即可。各列中的第一项表示两份独立标注的语料间的一致率，第二项和第三项表示两份独立标注语料与最终语料间的一致率。由表2可见，模糊限制语级别的一致性明显低于句子级别的一致性，说明模糊限制语的识别比模糊限制性句子识别更具难度。同时，由于模糊限制语没有明确的定义，有一些词语需要根据上下文语境判断其是否表模糊性，因此，模糊限制语的标注具有一定的主观性。但是，在语料的标注过程中，规则的制定者与两名语言学专家对前两份独立标注语料的不一致处进行了深入的探讨，反复修改了标注规则。两名语言学专家又根据新的规则分别修改了各自的标注语料。这也说明中文模糊限制语具有较大的歧义性，中文模糊限制信息检测存在较大的难度。两份独立标注的语料间的一致性低于它们分别与最终语料间的一致性，这是因为最终语料是规则的制定者对两份独立标注语料的不同之处再修改获得的，所以有可能和二者之一相同。当然，规则的制定者也对全部语料进行了审查，修改了部分独立标注语料的相同标记，最终语料具有较高的质量。

5 总结与展望

本文根据中文模糊限制语的语义和语用功能，对其类型进行了更细致的划分。在生物医学和维基百科两个领域，设计构建了中文模糊限制语语料库。在语料库构建过程中，从语料收集、标注规范制定和语料标注等多方面提高语料库的质量。目前已标注完成了一个具有2.4万句规模的中文模糊限制语语料库。统计表明，生物医学文献全文中35.09%的句子，维基百科中33.78%的句子包含模糊限制信息。两个领域中，由于词语的使用频率不同，所以模糊限制语的类型分布具有较大的差异。实验检测了语料标注的一致率，其中模糊限制语的一致率不高，表明中文模糊限制语具有歧义性，中文模糊限制语识别存在较大的难度。语料库的建设是一项长期而艰巨的任务，下一步我们将继续完善标注规范，改进标注质量，扩大语料规模。此外，本文仅标注了中文模糊限制语及其所属类别，标注模糊限制语的限制范围也将是本文下一步的研究工作。最后，我们希望尽快推出一个语料库的在线版本，为中文模糊限制语的研究提供共享资源。并基于中文模糊限制信息语料库，进行模糊限制信息检测研究。

参考文献

[1] Lakoff G. Hedges: a study in meaning criteria and the logic of fuzzy concepts [J]. Journal of Philosophical Logic, 1973, 2(4): 458-508.

[2] Farkas R, Vincze V, Móra G, et al. The CoNLL 2010 shared task: learning to detect hedges and their scope in natural language text [C]//Proceedings of the CoNLL, Uppsala, Sweden, 2010, 1-12.

[3] Szarvas G, Vincze V, Farkas R, et al. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes [J]. BMC Bioinformatics, 2008, 9(11): S9.

[4] Medlock B and Briscoe T. Weakly supervised learning for hedge classification in scientific literature [C]// Proceedings of the ACL, 2007: 992-999.

[5] Kim J D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature [J]. BMC Bioinformatics, 2008, 9(10): 1-25.

[6] Settles B, Craven M, Friedland L. Active learning with real annotation costs [C]//Proceedings of the NIPS Workshop on Cost-Sensitive Learning, Vancouver, Canada, 2008: 1-10.

[7] Shatkay H, Pan F, Rzhetsky A, et al. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users [J], Bioinformatics, 2008, 24(18): 2086-2093.

[8] Nawaz R, Thompson P, Ananiadou S. Evaluating a meta-knowledge annotation scheme for bioevents [C]//Proceedings of the Workshop on Negation and Speculation in Natural language Proceeding, Uppsala, 2010: 69-77.

[9] Uzuner O, Zhang X R, Sibanda T. Machine learning and rule-based approaches to assertion classification [J]. Journal of the American Medical Informatics Association, 2009, 16(1): 109-115.

[10] Rubin V L, Liddy E D, Kando N. Certainty identification in texts: Categorization model and manual tagging results [J]. Computing Attitude and Affect in Text: Theory and Applications, 2006, 20: 61-76.

[11] Wilson T A. Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states [D]. Ph. D. thesis, University of Pittsburgh, PA. 2008.

[12] Saurí R, Pustejovsky J. FactBank: A corpus annotated with event factuality [J]. Language Resources and Evaluation, 2009, 43(3): 227-268.

[13] Rubin V L. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts [J]. Information Processing and Management, 2010, 46(5): 533-540.

[14] **王舟. 英汉学术论文摘要中模糊限制语的对比研究——一项基于语料库的研究[J]. 华中科技大学学报：社会科学版**, 2008, 22(6): 59-63.

[15] **陈萍，蒋跃. 中英医学论文摘要中模糊限制语的对比研究[J]. 外语艺术教育研究**, 2009, 3(1): 15-20.

[16] **范晓晖，李晓，李莹. 中英作者医学论文英文摘要中模糊限制语的对比研究[J]. 西北医学教育**, 2010, 18(5): 1019-1021.

[17] **顾敏，周红. 英汉访谈节目中模糊限制语语用功能的对比研究[J]. 嘉兴学院学报**, 2013, 25(1): 87-91.

[18] Prince E F, Frader J, Bosk C. On hedging in physician-physician discourse [J]. Linguistics and the Professions, 1982: 83-97.

[19] Szarvas G, Vincze V, Farkas R, et al. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty [J]. Association for Computational Linguistics, 2012, 38(2): 335-367.

[20] **何自然. 模糊限制语与言语交际[J]. 外国语(上海外国语学院学报)**, 1985, (5): 27-31.

[21] **文旭. 语义模糊与翻译[J]. 中国翻译**, 1996, (2): 5-8.

[22] **苏远连. 英汉模糊限制语的分类和功能[J]. 广州大学学报：社会科学版**, 2002, 1(4): 29-32.

[23] **蒋平. 国内模糊语言研究:现状与目标[J]. 外国语(上海外国语大学学报)**, 2013, 36(5): 43-49.

周惠巍(1969—)，**博士，副教授，主要研究领域为句法分析、生物医学信息挖掘和自然语言处理。**
E-mail: zhouhuiwei@dlut.edu.cn

杨欢(1988—)，**硕士研究生，主要研究领域为生物医学信息挖掘、机器学习和自然语言处理。**
E-mail: yanghuan_dlut@mail.dlut.edu.cn

张静(1989—)，**硕士研究生，主要研究领域为中文信息处理。**
E-mail: zhangjing891010@126.com