

# 构建否定与模糊信息识别汉语语料库

陈站成<sup>1,2</sup>, 邹博伟<sup>1,2</sup>, 朱巧明<sup>1,2</sup>, 李培峰<sup>1,2</sup>

(1. 苏州大学计算机科学与技术学院, 江苏 苏州 215006;

2. 江苏省计算机信息处理技术重点实验室, 江苏 苏州 215006)

**摘要:** 否定与模糊信息识别是目前自然语言处理中的一个研究热点。本文介绍了汉语否定与模糊识别语料库构建的过程和方法, 讨论了语料库设计和建设过程中的几个重要问题, 包括语料的选取、标注体系、规则和一致性, 以及语料库的相关数据统计等。汉语否定与模糊识别语料库标注了表示否定或模糊信息的触发词及其作用范围, 体现了汉语中否定与模糊表达的语言现象, 为否定与模糊信息识别的相关研究提供了有力的资源支持。

**关键词:** 否定; 模糊; 汉语语料库

**中图分类号:** TP391.1

## The Construction of Chinese Negation and Uncertainty Identification Corpus

Chen Zhancheng<sup>1,2</sup>, Zou Bowei<sup>1,2</sup>, Zhu Qiaoming<sup>1,2</sup>, Li Peifeng<sup>1,2</sup>

(1. School of Computer Science and Technology, Soochow University, JiangSu SuZhou 215006;

2. Key Lab of Computer Information Processing Technology of Jiangsu Province, JiangSu SuZhou 215006)

**Abstract:** Negative and uncertain information identification is becoming focus of researches in NLP recently. In this paper, we introduce the procedure and approach how to build Chinese negative and uncertain information identification corpus, and discuss several important problems in the process of design and building corpus including selection of raw text, framework of annotation, guidelines, data consistency and statistic data obtaining from corpus. In this corpus, we annotate negative and uncertain information cues and their scope, and results embody the negative and uncertain information expression phenomenon in Chinese, and supply resources for future relevance researches about negative and uncertain information identification.

**Keywords:** negation; uncertainty; Chinese corpus

## 0 引言

否定信息是指否认一个命题的成立、存在或真实性<sup>[1]</sup>; 模糊信息是指一个命题具有不确定性和推测的含义<sup>[2]</sup>。否定和模糊识别包括识别含有否定信息或模糊信息的句子, 判断表示否定信息或模糊信息的触发词, 识别触发词对应的作用范围。CoNLL'2010 评测<sup>[3]</sup>定义了这两个子任务: 1) 识别句子中是否含有否定信息或模糊信息并确定触发词; 2) 识别触发词作用的文本范围, 即覆盖域。

由于篇幅限制, [本文不再重复详述]<sub>scope1</sub>。

[这些路径可能同时被维护在优先队列中]<sub>scope2</sub>。

例 1.1 含有否定信息, 触发词是“不再”, 其作用范围为 scope1。例 1.2 中, “可能”表示“这些路径同时被维护在优先队列中”是一个不确定命题, 具有模糊信息, 因此触发词为“可能”, 其覆盖域为 scope2。

Vincze 等<sup>[4]</sup> (2008) 指出从否定词覆盖域或不确定词覆盖域中抽取出来的信息应该与确定的信息分开处理。这在科学研究领域尤其重要, 因为科学研究领域更倾向使用多样化的语

基金项目: 博士点基金 (20093201110006)

作者简介: 陈站成, (1989-), 男, 硕士研究生, 主要研究方向: 自语言处理。

通信联系人: 朱巧明, (1963-), 男, 教授, 博士生导师, 主要研究方向: 自然语言处理、web 信息处理、嵌入式系统。E-mail: qmzhu@suda.edu.cn

言方式来表达实验结果，并在实验结果论述中出现一些假设性的推论。Vincze 等<sup>[4]</sup>根据 BioScope 生物医学语料库，统计出文献全文中分别有 12.70% 和 19.44% 的句子包含否定和模糊信息，对这些信息的识别对于科技文献的信息抽取研究具有重要意义。

基于语料库的研究通常依赖于相关语料库的建立<sup>[5]</sup>，已公布的否定与模糊信息识别语料库仅有英语的 BioScope 语料库<sup>[4]</sup>和 CoNLL'2010 评测语料库<sup>[3]</sup>。BioScope 语料库<sup>[4]</sup>数据来源于生物医学论文，标注了否定触发词（Negation Cue）和模糊触发词（Speculative Cue）及其覆盖域。CoNLL'2010 评测语料库使用了 BioScope 语料库的部分语料以及带标注的 Wikipedia 文本。目前，汉语否定与模糊识别语料库的探索和研究还很少有人涉足。

本文以科技论文集《计算机学报》为基础，构建汉语否定与模糊识别语料库。选取该语料构建汉语否定与模糊识别语料库是由于：1）《计算机学报》作为中国计算机领域的权威学术刊物，在语言表达上相对严谨，具有较小的歧义性；2）否定与模糊识别在科学文献的信息抽取相关研究中具有重要地位，例如，科技文献中的模糊表达往往是尚未或较难被证明的命题，表达了作者的态度和观点<sup>[6]</sup>；3）文本数量充足，适合语料库规模的扩充，以及为将来进行半自动甚至自动标注提供了可能。

本文后续内容组织如下：第二章介绍语料标注过程，包括语料预处理、标注工具的介绍及语料的存储；第三章结合具体语言现象说明标注规则；第四章对已构建好的语料库进行数据统计和分析；最后对我们的工作进行了总结并提出了下一步的工作目标。

## 1 语料标注

否定与模糊语料库标注步骤：1）将原始语料进行预处理和分句，获得生语料；2）由标注者利用标注工具标注生语料，形成初步语料库；3）对所标注语料格式进行规范化，构建 XML 格式语料库。

### 1.1 语料的预处理

语料选自《计算机学报》2012 年第 11 期的 19 篇论文<sup>1</sup>，为了确保语料格式统一，我们将汉语标点符号全部统一为全角，去除了多余的空格。由于语料中具体的图表和公式对否定与模糊识别研究没有作用，我们将其从语料中直接删除。对嵌入到文本中的公式，我们用占位符“[公式]”将其替换掉。语料以句子作为标注单元，如果一个句子中包含多个触发词，则处理为两个独立的标注实例。

### 1.2 语料的存储

文本文件和 XML 文件是常用的语料存储格式。由于文本文件简单易用，很多语料库都采用这种格式进行存储，然而，文本文件不适合表示结构化标注元素的信息，例如在否定与模糊识别语料库中，多个触发词与其覆盖域在句子中可能会出现嵌套现象。因此，我们采用 XML 文件格式存储语料，可以清楚的显示不同触发词的覆盖域，如图 1 所示。

<sup>1</sup> <http://cjc.ict.ac.cn/qwjs/2012-11.asp>

80

<sentence id="207">首先从 EED 随机抽取规模 M 的数据集,再从中随机选取 N 个关键词作为规模 N 的词典, 并<scope id="s207.1"><cue ref="s207.1"> 去除</cue><cue ref="s207.2"><scope id="s207.2"> 不</scope></cue> 包含于词典中的数据关键词</scope>. </sentence>

85     ( “<sentence id="207">” 标签表示第 207 个句子; “<cue ref="s207.1">” 表示触发词, 其覆盖域为 “<scope id="s207.1">”; “s207.1” 和 “s207.2” 分别表示了句子 207 中的两个实例。)

图 1 汉语否定与模糊识别语料库标注格式

1.3 标注工具

90     为了减轻标注人员的负担, 提高标注的效率和精确度, 减少标注的错误率, 我们开发了一款标注工具 (图 2) 来辅助标注人员的开发工作。标注者选取句子中所要标注的触发词和覆盖域, 当有多个触发词时, 点击 “AddInstance” 添加。当选取完成后, 点击 “Generate” 生成最后的标注结果。



图 2 语料标注工具界面

95

2 标注规则

否定和模糊信息属于语言的表达方式, 通常反映了语言使用者的观点或意愿, 因此, 需要指定合理的标注规则, 使得能够既忠实于语言材料本身, 又标注出这些特殊的语言现象。本章详细描述了汉语否定与模糊识别语料库的标注规则。

100   2.1 否定实例标注规则

1) 否定信息触发词通常为对行为或性状进行否定的副词。常见的包括: “不”、“不能”、“不再”、“无”、“无法”、“并未”、“并非”、“不易”、“难以”等。而其覆盖域通常是触发词修饰的结构: 当副词触发词修饰动词或动词性词组时, 覆盖域通常是动词所在的子句; 当副词触发词修饰形容词时, 往往是形容词所在的名词性短语; 当副词修饰其它副词时, 就要看被修饰的副词所修饰的部分是属于前面两种情况中的哪种。

105

例 3.1   当存在多数据拥有者查询授权时, [用户查询过程难以抵御恶意数据拥有者与 CSP

的合谋攻击 ] scope1。

例 3.1 中, 否定触发词“难以”修饰动词性词组“抵御恶意数据拥有者与 CSP 的合谋攻击”, 因此, 其覆盖域为子句 scope1, 该子句中, 覆盖的句子结构是“主语+否定副词+谓语+宾语”。

110 由于汉语中具有省略现象, 因此覆盖域所在的子句并不一定能完整包含“主语+谓语+宾语”结构, 可能缺少主语或是缺少宾语, 该情况下, 覆盖域仍为缺少成分的短语或子句。

例 3.2 如下所示: [公式]. 式 (4) 计算 IDF 的最大特点是去除 [用户无法访问的无关数据] scope1 对其查询排序的影响。

115 例 3.2 中, 触发词“无法”修饰的中心词是“无关数据”, 而前面的“最大特点”是“去除”的主语, scope1 是“去除”的宾语, “无法”与谓语“去除”没有语义上的关系, 所以触发词“无法”的覆盖域是 scope1。

2) 否定触发词还可以是其它副词, 如程度副词“基本上”、“全面”等。副词之间的组合形式可以是“程度副词+否定副词”或“否定副词+程度副词”等。

例 3.3 因此, [目前解决 TDSP 问题的方法均不能解决本文面对的问题] scope1。

120 在例 3.3 中, 否定触发词“不能”是副词, 修饰动词“解决”, 而其覆盖域句型结构为“主语+程度副词+否定副词+谓语+宾语”。

3) 否定触发词也可能为动词, 如“没”、“没有”、“缺乏”、“避免”、“去除”、“排除”、“减少”等, 其中“没”、“没有”是副词和动词的兼类。这些动词在具体语境中具有否定信息, 在没有省略句子成分时, 动词所在的子句就是触发词对应的覆盖域; 当该子句缺少主语时, 覆盖域为省略主语的子句。

例 3.4 基本算法 BSL 遍历出的图中所有 q-clique 均 [缺乏良好的可伸缩性] scope1。

例 3.4 中, 触发词“缺乏”的覆盖域的句型结构为缺少主语的“动词+宾语”, 宾语是偏正短语“良好的可伸缩性”。

除了副词和动词, 还有少数否定触发词是形容词、名词或介词。

130 4) 形容词也可以作为否定触发词, 如“非”、“不可行的”、“不同”、“错误的”等, 作定语时, 形容词性否定触发词与它所修饰的名词一起构成覆盖域; 而作表语时, 覆盖域为具有否定信息的子句。

例 3.5 显然, [该结果是错误的] scope1。

例 3.5 中, 否定触发词“错误的”作表语, 否定了“该结果”。覆盖域为句子 scope1。

135 5) 当否定触发词为介词时, 如“除了”、“不同于”等。由于介词通常引导一个短语作状语, 所以大多情况下覆盖域为介词所引导的成分。

例 3.6 [不同于传统的 Trie 树] scope1, RegionTrie 中同一个前缀可能对应着多个结点。

在例 3.6 中, 将“传统的 Trie 树”与“RegionTrie”形成对比, 说明两者的不同, “不同于”作触发词, 引导状语 scope1, 因此覆盖域为子句 scope1。

## 140 2.2 模糊实例标注规则

1) 在模糊实例中, 作触发词的副词通常修饰动词、形容词或名词性成分<sup>[7]</sup>, 而覆盖域通常是一个包含主谓宾、或者省略主语、或者省略宾语、或者以逗号分开的子句。

例 3.7 现有方法的查询时间与 M 值同比增长, 使 [查询时间基本不受 M 值的影响] scope1。

145 在例 3.7 中, 表模糊信息的副词“基本”修饰作谓语的动词“不受”, 该句属于主谓宾结构, 所以该句的覆盖域为子句 scope1。

2) “基本”是程度副词, 表示模糊信息, 与此类似的还有“一般”、“往往”、“较大地”、“较大幅度地”等。也有表示模糊信息的动词触发词, 如“估计”、“试图”、“推理”、“认



为”、“可见”、“假设”等，这类动词触发词出现时，该句通常有完整的句子结构，覆盖域通常是该完整的句子。

150 例 3.8 [文献**试图**通过扰乱排序实现查询隐私保护] <sub>scope1</sub>，但其内部用户完全可以根据自己所掌握的密钥与 CSP 合谋[推理出其他用户的查询请求隐私] <sub>scope2</sub>。

在例 3.8 中，有两个模糊信息实例：1) <sub>scope1</sub> 中的“**试图**”表示一种尝试性的做法，表示一种不确定性事件，有模糊信息。覆盖域为以逗号分开的子句 <sub>scope1</sub>，是一个完整的句子；2) <sub>scope2</sub> 中，“**推理出**”本就强调由前者可以推理出后者，只是推理而已，故是一种模糊信息。在  
155 整句中“完全可以根据自己所掌握的密钥与 CSP 合谋”不算在覆盖域里面，因为推理出在这里作谓语，重点是主语和后面的宾语，作状语的“完全可以根据自己所掌握的密钥与 CSP 合谋”就省去，不影响覆盖域的识别。

3) 短语“成为……的问题”、“在很多情况下”、“当……时”等有时也具有模糊信息。“成为……的问题”，如果是说某件事成为了一个问题，是问题显然就具有不确定性，就具有模  
160 糊信息。“在很多情况下”表示情况不同，结果也不同，具有不可预测性，也有模糊信息。“当……时”具有假设的意思，是假设就具有模糊信息。

4) 除上述情况之外，类似于“任意”、“假定的”等形容词，具有不确定性，表示模糊信息。

165 例 3.9 然而，[这些工作**假定的**时间模型是离散的] <sub>scope1</sub>。

在例 3.9 中，该句表达的意思是这些工作的时间模型不一定是离散的，而在此处假定它是离散的，也就是说时间模型是不是离散是一件不可确定的事，有模糊信息。“**假定的**”作触发词，原句 <sub>scope1</sub> 为覆盖域。

5) 类似于“或”、“或者”、“若”、“如果”、“一旦”等连词，具有很强的不确定性。

170 例 3.10 其中，[顶点代表道路的交叉口**或者**道路的端点] <sub>scope1</sub>，边代表道路片段。

例 3.10 中，“**或者**”是模糊信息触发词，表示顶点代表什么是不确定的，子句 <sub>scope1</sub> 就是覆盖域。与“或者”有相同意思的“或”也是模糊信息触发词。

## 2.3 特殊标注规则

本文 3.1、3.2 节分别介绍了判断否定和模糊信息的规则，然而，有些触发词及其覆盖域的识别往往需要依赖其在上下文中的具体含义，本节对这些特殊情况进行了总结。

### 175 2.3.1 否定特殊标注规则

1) “不同”在某些情况下可以做否定触发词。

例 3.11 时间子序列匹配，根据不同查询标准，可分为范围查询和  $k$  近邻查询两类。

180 例 3.11 中，“不同”表示“查询标准”的差异，而不是强调对查询标准的否定，因此认为“不同”在此句中不作否定触发词。在标注语料的过程中，标注者根据上下文语境和作者表达的重点来判别。

2) “除了……之外”、“除了……”等，要对比上下文内容是否表示相同来识别。

例 3.12 而隐私保护则要求除了保护数据隐私，还需确保密文索引和查询过程无隐私泄露。

例 3.12 中，“除了……”表达了递进的含义不具有否定信息。

185 例 3.13 [除了叶节点对所有序列保存所必须的开销**以外**]，索引所需要的额外开销非常的小。

例 3.13 中，“除了……以外”标注为否定触发词，其否定的是“叶节点对所有序列保存所必须的开销”。

### 2.3.2 模糊特殊标注规则

1) “考虑”是否标注为模糊触发词, 需要根据其在上下文中的含义。

例 3.14 Zhu 等人对同时考虑结构信息和属性信息的近似图匹配问题进行了研究。

“考虑”在例 3.14 中表示“涉及”的含义, 不具有模糊信息。

2) “如何”修饰事实时不标注为模糊触发词。

例 3.15 在本节中, 我们介绍如何计算起点到终点的最小费用代价。

例 3.15 中, “如何”修饰“计算起点到终点的最小费用代价”这一事件, 没有推测或不确定的含义。

## 3 语料库数据统计与分析

在本章, 针对标注完成的语料进行数据统计工作, 进而分析结果出现的原因。具体包括说明语料选取问题, 统计语料中句子个数、触发词个数, 计算人工标注语料准确率、召回率和  $F_{\beta=1}$  结果, 计算一致性, 分析原因。

英语科技文献已被证明在否定与模糊识别方面具有很好的研究价值<sup>[6]</sup>, 因此, 本文选取《计算机学报》作为构建汉语否定与模糊识别语料库的原始语料。语料库句子数为 4842 句, 标注工作由两位标注者参与完成。

表 1 人工标注语料准确率、召回率和  $F_{\beta=1}$  结果

类别	P (%)	R (%)	$F_{\beta=1}$
否定信息	94.24	95.94	95.09
模糊信息	93.26	94.73	94.00

表 1 中我们对人工标注的否定信息和模糊信息中的覆盖域计算准确率、召回率和  $F_{\beta=1}$ 。根据准确率和召回率给出的综合评价指标  $F_{\beta=1}$  分别能取到 95.09%、94.00%, 在否定信息的标注过程中  $F_{\beta=1}$  更高, 因为否定相比模糊信息更容易识别。

表 2 针对汉语否定与模糊识别语料库的相关数据进行了统计。否定信息句子数占比和模糊信息句子数占比分别为 15.78% 和 13.88%, 与 Vincze 等<sup>[4]</sup>在 BioScope 生物医学语料库上统计出的数据 12.70% 和 19.44% 接近, 表明了否定和模糊信息在汉语科技文献中普遍存在。否定信息覆盖域平均字数和模糊信息覆盖域平均字数分别为 14.95 和 18.41, 在英语中 Morante 等(2009) 统计的覆盖域平均长度为 8.81<sup>[8]</sup> 和 14.37<sup>[9]</sup>, 在汉语中否定覆盖域长度和模糊覆盖域长度之差没有英语差别大, 表明英语和汉语中否定覆盖域长度和模糊覆盖域长度是有区别的。

表 2 汉语否定与模糊识别语料库相关数据统计

否定信息	触发词数	941
	句子数占比	15.78%
	覆盖域平均字数	14.95
模糊信息	触发词数	812
	句子数占比	13.88%
	覆盖域平均字数	18.41

(“句子占比”指含否定信息句子数与语料库句子数之比。)

我们采用了 Cohen's kappa<sup>[10]</sup> 值作为衡量语料标注一致性的指标, 抽取 1300 个标注句子计算一致性。计算 kappa 值时采取完全匹配的方式, 即只有当两位标注者 kappa 值的计算结果分别为 84.55%、83.04%, 表明两位标注者对语料中的实例都能较准确的识别, 也表明在该语料上进行标注工作具有很好的针对性, 该语料库能够为否定与模糊识别的相关研究提供

有力的资源支持。

## 4 总结及展望

225       本文介绍了否定与模糊识别语料库的构建方法。首先，将原始语料进行预处理和分句，  
获得生语料；其次，由标注者利用标注工具对生语料标注，形成初步语料库；最后对所标注  
语料格式进行规范化，构建 XML 格式语料库。标注过程中，我们发现与英语标注不同，汉  
语中的省略现象较多，缺少句子成分，判断覆盖域时不能仅依靠句子的句法结构；同时，由  
于标点符号使用的不同，英语句子通常只包含一个主谓结构，而汉语句子中可以包含多个主  
230       谓结构，因此，可以充分利用句子内部的上下文信息识别触发词及其覆盖域。

语料库建设是一项长期工作，未来我们将进一步扩大否定与模糊识别语料库的规模；此  
外，我们还将尝试发现其它类型的语言材料（如微博文本、新闻文本等）中是否具有这种语  
言现象，并有针对性地进行标注。

## 235 [参考文献]

- [1] Eduardo Blanco, Dan Moldovan. Semantic Representation of Negation Using Focus Detection[J]. In  
Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, 581-589.  
[2] Ben Medlock, Ted Briscoe. Weakly Supervised Learning for Hedge Classification in Scientific Literature[J]. In  
Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, 992-999.  
240 [3] Rich'ard Farkas, Veronika Vincze, Gy'orgy M'ora, J'anos Csirik, Gy'orgy Szarvas. The CoNLL-2010 Shared  
Task: Learning to Detect Hedges and their Scope in Natural Language Text[J]. Proceedings of the Fourteenth  
conference on Computational Natural Language Learning: Shared Task, 2010, 1-12.  
[4] Vincze V, Szarvas G, Farkas R, M'ora G. and Csirik J.. The BioScope corpus: biomedical texts annotated for  
uncertainty, negation and their scopes[J]. BioNLP 2008: Current Trends in Biomedical Natural Language  
245 Processing, 2008, 38-45.  
[5] John Sinclair. Corpus Concordance Collocation[M]. Oxford: Oxford University Press, 1991.  
[6] KEN HYLAND. Writing without conviction? Hedging in scientific research articles[J]. Applied Linguistics,  
1996, 17(4): 433-454.  
[7] 许利英. 试论现代汉语否定句[J]. 安庆师范学院学报, 1986, (3):108-118.  
250 [8] Roser Morante, Walter Daelemans. A metalearning approach to processing the scope of negation[J]. In  
Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), 2009:21-29.  
[9] Roser Morante, Walter Daelemans. Learning the scope of hedge cues in biomedical texts[J]. In Proceedings of  
the workshop on BioNLP, 2009:28-36.  
[10] Cohen. A coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement,  
255 1960:37 - 46.