农业病虫害词汇获取方法初探*

郑家恒 杜永萍 宋礼鹏
山西大学计算机科学系（030006）
E-mail:jhzheng@sxu.edu.cn

摘 要： 本文采取统计的方法获取农业病虫害词汇的词性搭配规则、语义类分布规则，并进一步利用这些规则在大规模语料中采用并列同现、模式匹配、特征词匹配等策略获取病虫害词汇，建立特定专业领域（主要为农业病虫害领域）词汇词典。

关键词： 病虫害词汇，专业词汇获取，中文信息处理

The Research on Lexical Acquisition of Agricultural Plant Diseases and Insect Pests

Zheng JiaHeng Du YongPing Song LiPeng
The Department of Computer Science of shanxi University(030006)
E-mail:jhzheng@sxu.edu.cn

Abstract: In this paper we use the statistic method to acquire the rules to combine the segmented characters which should be one word and the rules of semantic distribution. Furthermore, We adopt the strategy of co-occurrence, pattern matching, central matching to build the special lexicon(the field of agricultural plant diseases and insect pests).

Keywords: Diseases and Insect Pests lexicon , Domain Lexical Acquisition , Chinese Information Processing

1 前言

目前随着中文信息处理应用领域的扩展，已经提出了对专业词汇（术语）词典的需求，它的构建将为文档分类以及信息抽取任务提供有力依据。为了获取在词典中无法得到的专业词汇，我们必须转向其它资源，如大规模语料。

Ellen Riloff 等人曾就基于语料库方法，建造语义词典做了研究[2]。利用初始选定的种子词，在大规模语料中搜寻专业词汇，采取多重循环获取更多词汇。Brian Roark 等人

* 本文承国家 863 项目（2001AA114031）的资助

在 Ellen Riloff 的研究基础上，对初始种子词的选取、词汇的分值计算以及复合词的识别做了更为细化的探讨[1]。Marti A. Hearst 等人还对从大规模语料中自动获取词汇的上下位信息做了报道[3]，但模式自动获取的实现还有一定困难，并且有很多模式并不适合中文。

本文以 Inter 网（中国北方农业信息网等网站）上的最新信息为语料资源，采用并列同现、模式匹配、特征词匹配等策略，在语料中抽取农业病虫害领域词汇，并利用词与词的语义相似度对词汇噪音做进一步剔除，提高了词典质量。

2 研究难点

我们获取的对象是专业词汇，属于未登录词语的范畴，且针对性较强。现有的分词系统属于通用型，对于这样的词汇在分词过程中，很多词会被切分为散串。如："透|A|翅|N|蛾|N|，稻|N|象|N|甲|N|" 等等。我们将其在获取过程中遇到的难点分析如下：

① 构词无规律。专业术语的构成不象人名、地名有一定的构词规律。它的构成方式多样，有些是以复合词的方式出现，有些是由单字词或语素字组成；用字比较分散，有些是普通字，有些是生僻字；专业术语的长度没有一定的限制。

② 缺乏启发信息。人名、地名的识别有一定资源《中国人名用字库》、《中国地名用字库》可以借鉴。并且同中国人名相比，缺乏象姓氏一类的启发信息。

③ 专业术语指示词出现情况多样化。在真实文本中，一些介词，动词之类的指示词（防治、危害）经常同专业术语一起出现，对专业词汇识别能起标志作用，但这类词在文本中并不总是与专业词汇同时出现。如："危害广大人民群众的利益"、"防治策略有很多"。

④ 专业特征词出现情况复杂。专业术语经常伴随着一些专业特征词出现，如：病、虫、蛾等等。但是文本中出现的专业特征词，并不都表示真正的专业术语。如："重病"。

由上可见，种种现象使专业术语的识别变得复杂。

3 当用资源库的建立

3．1 词库的建立

1．专业词汇特征字（词）库

该库中的字或词通常都作为专业词汇用字（词），它们可以出现在专业词汇的首部、中部或尾部。结构： CWord Position （分别代表特征词和它在专业词汇中出现的位置）

Position={ B（Begin）、M（Middle）、E（End）}

2．指示词库（按《同义词词林》语义类分类）

该库的建立是为了确定专业词汇的出现位置，这类词通常同专业词汇一同出现。结构：CWord SemType （分别代表指示词和它在《词林》中的语义类别）。例：防治 Hg200000

3．关系词汇词库

（1）指示并列关系字词库

该词库由表并列关系的连词和顿号组成。结构： CWord （代表指示并列关系的字词）。

（2）指示上下位关系词库

该词库中的词指示词汇间的上下位关系，以获取具有上下位语义关系的词汇。结构：CWord　（代表指示上下位关系的字词），例：归于（归属）。

（3）指示同位语关系词库。结构：CWord　（代表指示同位关系的字词），例：俗称。

4．单字词库（不和其它字组成词）

该词库中的词不能和其它字组词，只能单独使用，共有 7415 个词，它们大部分为生僻字。结构：CWord　Tag　（分别代表单字词及其词性），例：懈　N。

3．2 模式库的建立

无论是专业词汇的获取，还是上下位、同位关系词汇的获取，都会用到模式匹配方法。这些模式只是一些简单的专业词汇及其相关知识的触发抽取模式，如：TriggerB[控制]<>的 TriggerE[危害]，TriggerB、TriggerE 分别代表前触发词和后触发词，<>的内容即为所要抽取的词汇，这些模式是由人工干预进行半自动获取。这些模式分为如下三类：

1．专业词汇获取模式库

举例：A：[防治]<>

B：<>[主要危害]

C：TriggerB[兼治]<>等 TriggerE[病害]

A、B、C 分别代表前触发型、后触发型、前后触发型。

2．上下位关系词汇获取模式库

举例：Noun（复、下）Trig[其它] Noun（单、上）

“复”代表可以是许多并列出现的词汇，“单”代表是单个词汇，“上”代表上位词，“下”代表下位词。

3．同位关系词汇获取模式库

举例：<>Trig[又名]<>

3．3 规则库的建立

在专业词汇识别的过程中需要用到两类识别规则库：词性搭配规则库、语义类别分布规则库。在词汇获取过程中，我们采取了三种策略：并列同现、模式匹配、特征词匹配，相应地建立了三类获取规则库—并列获取规则库、模式匹配规则库、特征词匹配规则库。举例如下：

1．获取专业词汇模式匹配规则

前触发型：　if　POS(TrigWord，Sentence)<>0　then　BackwardExtrWord

后触发型：　if　POS(TrigWord，Sentence)<>0　then　ForwardExtrWord

前后触发型：　if　((POS(TrigWordB，Sentence)<>0) and (POS(TrigWordE，Sentence)<>0)) then　MiddleExtrWord

TrigWordB、TrigWordE 分别代表前、后触发词，Sentence 代表真实文本中的某一句子。

2．获取上下位关系词汇模式匹配规则

if　POS(HypoTrigWord，Sentence)<>0 then HypoExtr(LefBord，RigBord)

HypoTrigWord 代表上下位关系的指示词，LefBord，RigBord 分别代表该指示词的左右界。

3．获取同位关系模式匹配规则

if POS(ParaTrigWord，Sentence)<>0 then ParalExtr(LefBord，RigBord)

ParaTrigWord 代表指示同位关系的指示词，LefBord，RigBord 分别代表该指示词的左右界。

4 专业词汇获取设计

我们获取的专业词汇，根据它们在语料中出现的规律，采用 4 种策略进行抽取。

4．1 并列同现获取

这种途径是基于这样一种认识：相同语义类名词经常共现，它们通常由连词、标点符号相隔构成一个序列或以同位语的形式出现。如：“主要病害包括稻恶苗病、稻蓟马、稻料黑粉病和稻瘟病。”该方法的实现不需要经过加工的语料，即不必事先分词、POS、语义类标注，这样避开了散串合并的问题。它首先定位出现专业词汇的句子，并进一步在该句中定位指示并列关系的词汇和标记，而后利用并列获取规则来抽取词汇。

4．2 特征词匹配获取

特征词我们特指专业词汇用词（字），它们可以出现在专业词汇的首部、中部或尾部。如果特征词是尾词，则该种匹配称为中心匹配（本文主要讨论中心匹配）：

词串 $X=X_1X_2\cdots X_n$　　$Y=Y_1Y_2\cdots Y_m$

$X_r=X_i\cdots X_n$（X 的右子串）　$Y_r=Y_j\cdots Y_m$（Y 的右子串）（$1<i<n$；$1<j<m$）

If $X_r=Y_r$ Then CenterMatching（X，Y）

例：X=小麦吸浆虫　Y=大豆食心虫　$X_r=Y_r$=虫

以该种方式获取的词汇我们认为属于同一种语义类，若 X 是专业词汇，则 Y 可直接加入到同类词汇词典，并且通过这种方式获得专业词汇的模式分值较高。

4．3 模式匹配获取

对在语料中出现的专业核心词学习其核心模式（核心词出现的上下文），按照模式的触发位置分为前触发型、后触发型、前后触发型，一旦在真实语料中发现模式触发词则根据其类型利用相应规则识别词汇，这些核心模式可进一步识别更多的专业词汇。

在语料库中识别出一些简单的出现频率较高的词汇模式来指明词汇之间的上下位关系，下面列出了一些模式，并伴有例句和从中推出的谓词关系：

① Noun（单、上）Trigger[包括] Noun（复、下）

例：造成严重损失的病害包括白发病、黑穗病、谷瘟病、谷锈病等。

上下位（“白发病”，“病害”）　上下位（“黑穗病”，“病害”）

上下位（“谷瘟病”，“病害”）　上下位（“谷锈病”，“病害”）

② Noun（单、上）Trigger[包含] Noun（复、下）

例：金针虫包含细胸金针虫、沟金针虫等。

上下位（“细胸金针虫”，“金针虫”） 上下位（“沟金针虫”，“金针虫”）

这是一种在语料中发现两个或多个词之间上下位词汇关系的方法。

4．4 词汇噪音的剔除

为了能够测量出词-词之间的语义相似度，需要考虑一个词在它所出现的文本中的分布情况。在本文中，我们建立一个矩阵 A 来完成这步工作，每一行由所要衡量相关度的词来标识，每一列由词所在的句子序列（包括单句和段落）来标识，矩阵中的元素值 a_{ij} 为第 i 行的词在第 j 列的文字片段中出现的次数。

按照如下公式计算任意两个词之间的相似度：

$a_i=(a_{i1}, a_{i2}, \cdots, a_{in})$ $a_j=(a_{j1}, a_{j2}, \cdots, a_{jn})$

$$\cos(a_i, a_j)=\frac{a_{i1}a_{j1}+a_{i2}a_{j2}+\ldots+a_{in}a_{jn}}{\sqrt{a^2_{i1}+\ldots+a^2_{in}}\sqrt{a^2_{j1}+\ldots+a^2_{jn}}}$$

a_i，a_j 分别代表两个行向量。

我们所计算的两个词当中，有一个是已经确认的专业术语，另外一个是新获取到的词汇 NewWord，通过计算二者之间的相似度来判断 NewWord 与该专业的相关度，如果值小于 0.8%，则认为它是噪音词汇，将其剔除。

5 实验

5．1 词汇可信度度量

算法直接抽取出的词汇未必都是专业词汇，为了从中选取可信度高的词汇加入到专业词汇词典中，我们对其中的每个词按如下公式进行分值计算。

$$SCORE\ (w, D)=\frac{Count_1}{Count_2}$$

$Count_1$ 为词 w 在专业领域 D 的上下文窗口（核心词汇及其周围的词组成）中的出现次数。$Count_2$ 为词 w 在语料中的出现次数

设定一阈值 0.02，我们认为分值小于该阈值的词为低可信度词汇，将其剔除，但有些真正的专业词汇分布较分散，在专业领域 D 的上下文窗口中出现次数远远小于它在语料中的总出现次数，成为低分值词汇，被剔除了。

5．2 实验分析

为测试算法的有效性，我们进行了实验，语料来源于中国北方农业信息网、中国农科

院植保所、通州农业信息网、山西农业信息网、中昊网等网站。对选取的农业病虫害60万语料进行了测试，经统计，获取了220余农业病害词汇，正确率 72.3%，180余虫害词汇，正确率 76.4%。实验结果错误原因分析：

1．并列抽取错例分析

例：玉米粗缩病涉及到病害和传毒介体灰飞虱。　误抽词：到病害。

核心词“粗缩病”出现在该句中，而且有两个并列触发词“及”、“和”，故而抽取了噪音词汇“到病害”。

2．模式匹配抽取错例分析

下表列出了在病虫害语料中出现的高分值模式及其5次循环获取专业词汇的实验结果。

高分值模式	“防治< >”
获取词汇	白粉病 干尖线虫病 策略 蝗虫
高分值模式	“< >主要危害”
获取词汇	条纹病 稻纵卷叶螟 病源 小菜蛾

表中这些抽取模式对于病虫害词汇获取具有指导作用，但也产生了一些非病虫害词汇（如：病源、策略）。

3．特征词匹配抽取错例分析

例：同一品种，低洼湿度大的田块较湿度小的田块发病重。误抽词：发病。

“病”是病害的尾特征词，向前匹配抽取出“发病”这一噪音词汇。

该系统还有其局限性，需要做进一步的改进工作：

（1）由于语料库的规模比较小，农业病虫害领域专业词汇用字在真实文本中的覆盖率不完全，使得用字的使用程度信息的计算结果不够客观。今后要进一步扩大语料库的规模，提高专业词汇用字的覆盖率。

（2）现有的规则库、模式库规模较小，且形式单一，需要进一步获取其它形式的规则。如：句法规则，以扩大规则集的规模，提高系统的性能。

参考文献

[1] Brian Roark, Eugene Charniak . Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. http://www.researchindex.com

[2] Ellen.Riloff, J.Shephed.1997. A corpus-based approach for building semantic lexicon. In proceedings of the Second Conference on Empirical Methods in Natual Language Processing, pages 127-132.

[3] Marti A.Hearst . Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes France, July 1992

[4] 梅家驹等，《同义词词林》，上海辞书出版社，1983