**《未登录中文领域术语识别研究》陈巧**

1.**内容摘要**：

领域术语的自动识别需要同时考虑术语的单元性和术语性，较一般领域未登录词识别更为困难。本文以系统生物学领域为例，设计了一个统计和规则相结合的未登录中文术语识别方法。通过分析语料库中术语的特征，本文挑选了6个体现术语统计学和语言学的候选特征，分别设计了以字特征和以词特征为基本特征的两组ＣＲＦｓ模型，最终得出词特征、词性特征、词长特征、相关度特征和信息熵特征的组合取得最好效果。针对ＣＲＦｓ模型的错误识别情况，本文总结了两条后处理规则。

2.**目的**：

完善的领域词典对于专业领域文献的语义理解十分重要。

3.**概念**：

术语：定义为“特定专业领域中一般概念的词语指称”。

4.**意义**：

①领域术语识别被广泛应用于本体构建、自动摘要、信息检索、领域知识库构建等研究领域。

②提高分词的速度和准确度。

③构建领域词典，提高信息检索准确度和完整度。

④通过自动抽取领域术语，促进对专业文献的自动化处理。

5.**本文创新点**：

①**构建分类领域词典**：本研究构建的是带有类别信息的领域词典，提供每个术语的上层分类信息，能够满足用户细粒度的信息需求。

②**识别低频术语**。

③**提出更合适的特征离散化方法**。

6.**研究内容**：

①**构建语料库**：对文本进行语义标注，人工标注句子中包含的术语，生成实验所需的语料库。

②**构建条件随机场模型**：分别建立以字和词作为基本特征的两组模型

③**模型评价**。

④**构建规则并对识别结果进行规则处理**：对模型识别的错误结果进行归纳总结，对有规律可循的、数量达到一定范围的、能够通过规则改正的错误制定人工规则集。对照规则集修正错误识别结果，通过比较系统的评价指标得出规则集的效果。

⑤**探讨系统对低频术语的识别效果**。

7.**基于统计的方法**：

**（一）基于统计特征的方法**：经常使用于术语识别的统计特征有词语频率、互信息、t-测试、t-测试差、信息熵等。

①**词语频率**

②**互信息**：反映相邻汉字间结合紧密程度的度量值。

公式： 表示字串ＡＢ结合越紧密，ＡＢ的互信息值越大。

③**t-测试**: 通过比较单个汉字与其前一个汉字和后一个汉字之间的连接趋势，判断该汉字与其前一个汉字还是后一个连接更紧密。

公式：

其中p(z|y)，p(y|x)分别代表z关于y，y关于x的条件概率。， 分别代表z关于y的方差，y关于x的方差。当：

（1）tx,z(y)>0时，ｙ与其后继字符Ｚ有相连的趋势，tx,z(y)值越大，代表二者联接趋势越强；

（2）tx,z(y)=0时，不反映任何趋势；

（3）tx,z(y)<0时，y与其前趋字符x有相连的趋势，tx,z(y)值越小，代表二者联接趋势越强。

④**t-测试差**

对有序中文字串abcd，汉字b、c之间的t-测试差的计算公式为：

（1）大于0时，b、c之间倾向于连接。

（2）等于0时，不反映任何倾向；

（3）小于0时，b、c之间倾向于断开。

⑤**信息熵**

词语的左、右信息熵分别衡量词语左侧、右侧的稳定程度。如果语料库中词语W左侧（或右侧）固定地出现词语W1，其左信息摘（或右信息摘）就越小，组合W1W（或WW1）就可能是一个词语。

**（二）基于机器学习的方法**：隐马尔科夫模型（HMM）、最大熵模型（MEM）、条件随机场模型（CRFs）和支持向量机（SVM）

①**隐马尔科夫模型（HMM）**

一个隐马尔可夫模型记作

:状态的有限集合

:观察值的有限集合

: 转移概率

: 输出概率

: 初始状态分布

在给定的模型下，要求从一定观察值序列的所有可能的状态中，选取概率最大的作为最终的状态序列。

②**最大熵模型（ME）**：保留最大的不确定性，使熵值到达最大，从而将风险降到最小。

熵的计算公式：

符合最大熵原理的概率分布公式:

③**条件随机场模型（CRFs）**：条件随机场(X,Y)就是一个以观测序列X为全局条件的无向图模型。可观测的输入序列，是待预测的有限标记序列集合，是对应的标签。线性CRFs定义标记序列Y的条件概率为：

Z(x)是归一化因子，F(y,x)是特征函数。是对应特征函数的权重，可通过最大似然估计法训练模型获得。模型在解码过程中采用Viterbi动态规划算法求取全局最优解：

在条件随机场中，特征函数的选取直接关系模型的性能。CRFs模型中特征函数的形式定义：。它是状态特征函数和转移特征函数的统一形式表示。特征函数通常是二值函数，取值要么为1要么为0。

④**支持向量机（SVM）**：其目标是找到分类边界与最近的训练资料点之间间隔最大的超平面作为分类边界。

8.**语料库的构建**：

步骤①：对文档进行语句层语义标注

步骤②：获取“茎”、“叶”描述文本

步骤③：分别对不同类型的术语进行标注，将目标术语标注在<terms>和</terms>之间

9.**术语的特征**：

①**领域性强**：一般不会出现在我们日常生活领域或是不相关的其他专业领域。

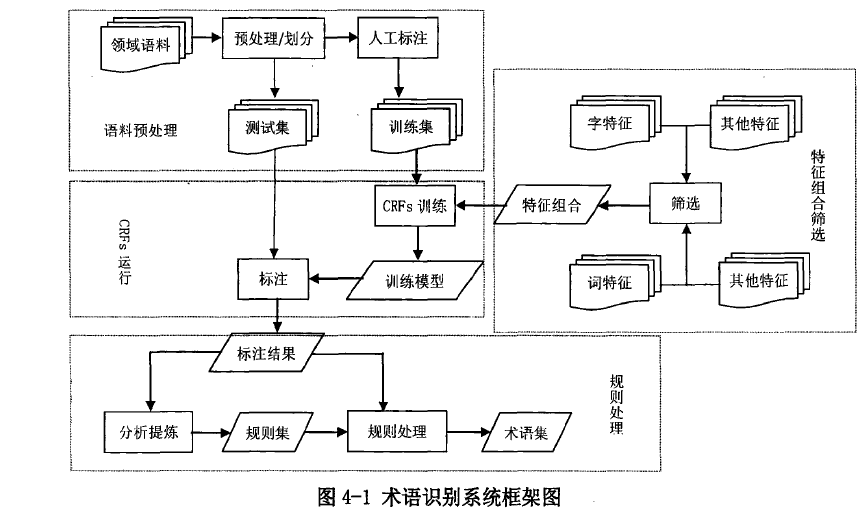
②**非名词性术语多**

③**中低频术语比重大**

10.**术语识别系统框架**：

本研究釆取统计和规则结合的方法，构建CRFs融合术语多特征，并构造规则对识别结果进行处理的术语识别系统。系统分为语料格式转换、特征计算组合、CRFs训练与测试模块和规则处理四个模块。

1. **语料格式转换模块**：首先随机选取一定比例的测试术语，根据每条语句是否包含测试术语对语料进行划分，包含测试术语的语句组成测试集，剩下的语句组成训练集。将训练集和测试集分别转换成CRF++工具包要求的格式：将文本转换成每行一个字或一个词的形式，为每个字或词进行SBIEO标记，将术语和非术语用符号区别开来。测试集和训练集都进行同样的标注，测试集的标注结果将作为正确结果与后续测试结果进行比较，计算模型评价指标。
2. **特征计算、组合模块**：在测试集和训练集中分别计算每个字或词的各个候选特征值，对于连续性特征值进行离散化，将这些候选特征添加到相应的字或词之后，以备后续实验进行特征组合选择。
3. **CRFs训练与测试模块**：调用CRFs工具对训练集进行学习，根据学习得到的模型对测试集进行标注。根据实验结果选取最佳特征组合模型。
4. **规则后处理模块**：对错误识别结果进行归纳总结，人工制定规则集，再应用规则集对实验结果进行处理，得到最终的术语集。



11.**比较隐马尔可夫模型和条件随机场模型**：

最大熵模型解决了隐马尔可夫模型输出独立性假设的问题，可以任意的选择特征，但由于其在每一节点都要进行归一化，只能找到局部的最优值，同时也带来了标注偏置 (Label Bias)的问题，即凡是训练语料中未出现的情况全都忽略掉。而条件随机场模型不在每一个节点进行归一化，而是所有特征进行全局归一化，因此可以求得全局的最优值。

12.**条件随机场模型的优点**：

①条件随机场模型具有表达字串长距离依赖性和交叠性的能力，能够捕捉观测序列之间的任意依存关系。

②CRFs模型推理能力比较强，能够充分利用上下文，即使是只给出很简单的特征，它仍然能达到较高的性能。

③CRFs模型能很好的克服数据稀疏问题。

13.**候选特征**：

1. **字（词）本身**

**②POS词性标记**。

1. **词长**
2. **互信息**
3. **相关度**

一个字串W的相关度定义如下：

其中，字串W的首字为A，次字为B，n为语料库所有二元组的串频，。表示首字为A次字为B的串频，表首字为A次字非B的串频，表示首字非A次字为B的串频，表首字非A次字非B的串频。

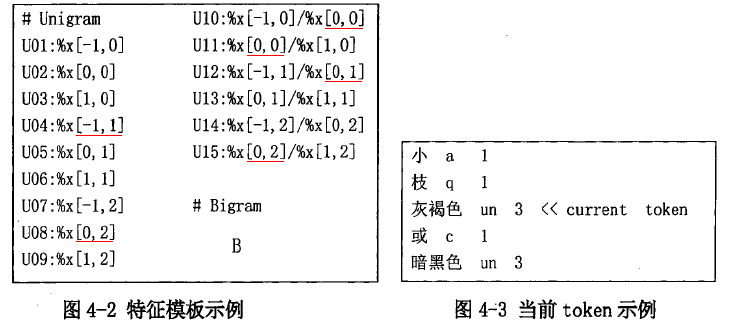
1. **信息熵**

14.**特征模板**：

①CRF++是著名的条件随机场开源工具

②CRF++提供两种特征类型，一种是Unigram特征，在特征文件中以字母”U”开头，代表一元特征；另一种是Bigram特征，在特征文件中以字母”B”开头代表二元特征。

1. 当选取包含一元、二元的词特征、词性特征和词长特征，窗口为１时的特征模板如下图：



在模板文件中，以U或B开头的每一条表示一个模板。**%x[row,col]**中row表示到当前token行的偏移行数，col指定第几列。

15.**标注集**：

1. 每个汉字在特定的词语中都占据着一个确定的构词位置，即词位。
2. 将术语识别问题转换为标注文本中每个字的词位，给术语和非术语标注不同的词位符号。
3. 经常使用的词位标记集有二字位标注集、四字位标注集、六字位标注集。

④本研究中术语字长是1到5，于是在四位标注集BIEO的基础上定义了SBIEO标注集。其中，S(Single)表示单字词，B(Begin)表示术语的第一个字，I(In)表示中间的字，E(End)表示最后一个字，O(Out)表示当前字不在术语中。每个S或连续的B(I)E构成一个术语。

16.**系统评价指标**：

17.**术语识别模型的构建思路**

**（一）设计两组识别模型**：

1. 一组以**字**作为基本特征，探讨它与互信息、相关度和信息熵特征的相互作用。
2. 另一种以**词**为基本特征，候选特征有词性、词长、互信息、相关度和信息熵特征。

分别选取最佳的一组特征组合进行比较，选择识别效果最佳的模型。

**（二）建立领域词典**：

使用常规分词工具的分词效果较差，由于ICTCLAS分词软件不具备植物学领域专业知识，在对实验文档切分时会出现两大类问题：①不能将语料中的词语划分在一起；②将几个词语的部分字符错分成另一个词。

本研究放弃使用ICTCLAS分词软件对语料进行自动切分，建立一个小型领域词典。

**（三）实验步骤**：

**（1）划分测试集和训练集**。为了达到测试各种长度、各种词频、各种结构、各种形式术语的目的，本研究采取通过选择测试术语来划分测试集和训练集的方法，将包含测试术语的句子组成测试集，剩下的句子组成训练集。

**（2）将测试集和训练集处理成字特征和词特征形式**。字特征形式即是将文本中的文字逐一拆分成行；词特征形式的处理首先将手工词典中的相应类别的词剔除，对文本分词后的每个词单独成行，将语料中标注在<terms>和</terms>之间的术语进行SBIEO标注。

**（3）为基本特征添加其他候选特征**。以字作为基本特征，候选特征有互信息、相关度和信息熵特征；以词为基本特征，候选特征有词性、词长、互信息、相关度和信息熵特征。

**①添加词性特征**：

本研宄釆用ICTCLAS提供的词性标注功能进行词性标注，将人工词典导入到ICTCLAS的用户词典，对文本进行分词和词性标注，ICTCLAS将用户词典包含的词的词性均标注为“un”。如果一个词没有在用户词典中，被ICTCLAS切分在一起，赋予了一个词性，则将这两个字拆分，赋予相同的词性标记。(例如，“主枝”不在用户词典中，被赋予了n（名词）词性，程序则将“主枝”拆分为“主”和“枝”，词性都标记为n。)

ICTCLAS提供了计算所一级标注集和计算所二级标注集，两种标注集的区别在于一级标注集是二级标注集的上位概念，两者的粒度不同，一级标注集的粒度较大，其类别少于二级标注集。实验表明计算所一级标注集效果优于计算所二级标注集。

**②添加词长特征**：

词长是指一个词包含的字数，对于字特征来说，词长都是一，不具备区别含义，因此只为词特征添加词长特征。

1. **添加互信息特征**：

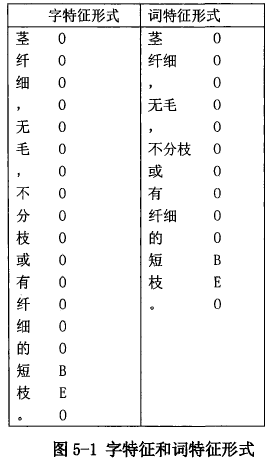
互信息值是一个连续值，不适于CRFs学习和预测，需要将其离散化。本研究将互信息等频率分为5个等级。

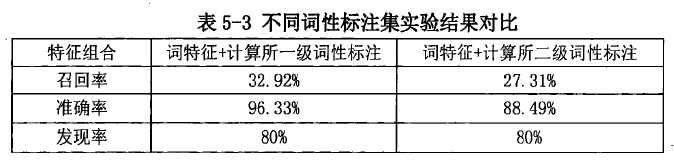
**④添加相关度特征**：

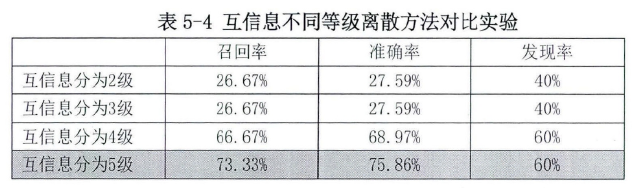
相关度仍需要进行离散处理，离散赋值方法与互信息的赋值方法一致。

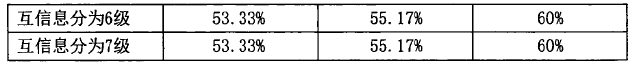
**⑤添加信息熵特征**

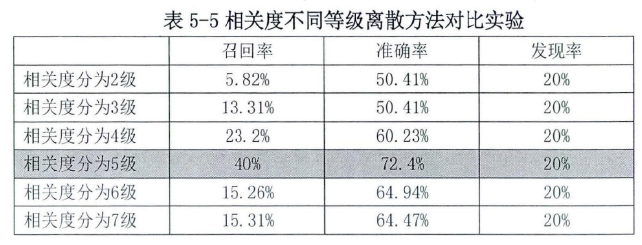
信息熵也是一个连续值，本研宄提出了另一种信息熵的离散方法：计算每个字（或词）的左右信息熵，比较大小，若左信息熵大于右信息熵，则将特征标记为“rgh”（right，表右），说明该字（或词）倾向于与右边的字（或词）链接，左边更可能是词语的边界，否则标记为”lft”（left,表左），忽略左右信息熵相等的情况。

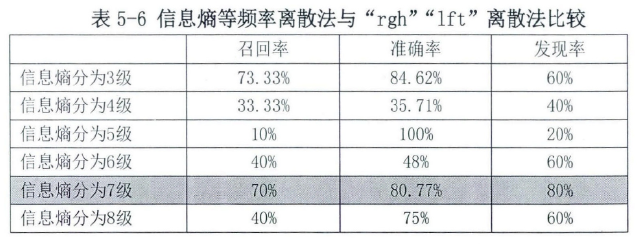






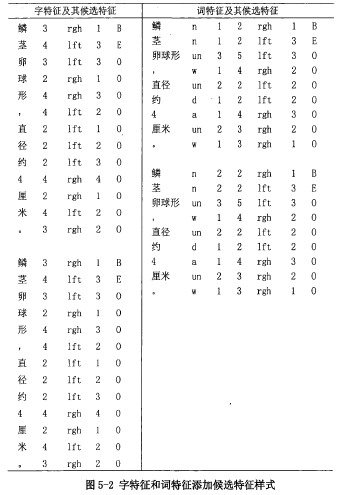








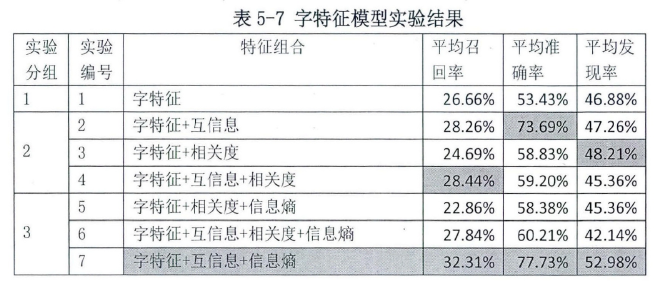
下图是字特征及其候选特征和词特征及其候选特征均添加之后的形式。字特征后面紧跟的是互信息，信息熵，相关度特征以及SBIEO特征后面依次是词性、词长、互信息、信息熵、相关度特征以及SBIEO标注。



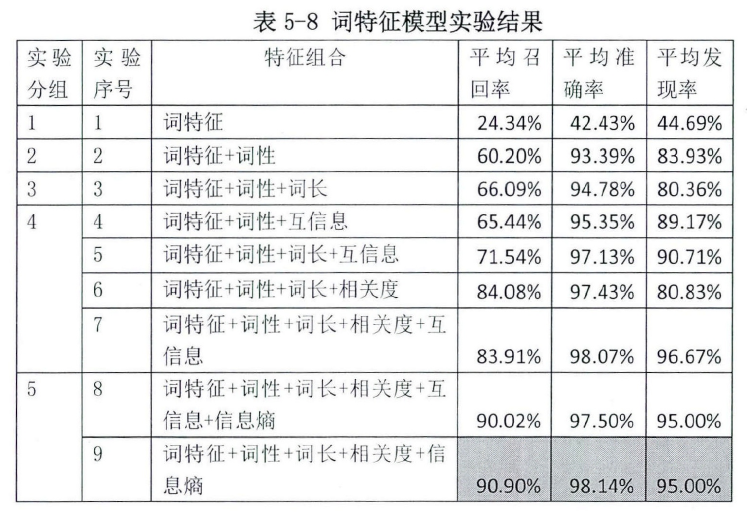
**（4）训练CRFs模型，选择识别效果最佳的模型**。

**①字特征模型最佳特征组合实验**：

语料中的术语长度最大为5，因此将字特征的窗口设置为2，其他特征的窗口设置为1。每个实验的结果中若三个指标都低于前一组实验或本组其他实验的三个指标，那么直接抛弃这种组合，如果有指标在前一组实验和本组实验中是最高值，则保留组合到下一组实验中讨论。



1. **词特征模型最佳特征组合实验**：



1. **选择最佳特征组合**。字特征模型最佳组合是字特征、互信息、信息熵特征，召回率、准确率和发现率分别是32.31%，77.73%，52.98%。词特征模型最佳组合是词特征、词性、词长、相关度、信息熵特征，召回率、准确率和发现率分别是90.90%，98.14%，95.00%。词特征模型明显优于字特征模型，于是选择词特征模型。

18.**术语识别模型评价**：

通过上一节的实验，确定了词特征模型及其最佳特征组合。本节将对该模型进行5轮检测，获取模型评价指标值。将以上5轮实验评价指标的均值作为该模型的最终评价指标。

19.**规则处理**：

**步骤（1）：错误分析**：主要存在以下几种错误情况：

1. 同一个术语在文中出现的部分位置被准确标注出来了，部分位置没有被标注出来。
2. 仅标注出多字术语中的一部分，即术语嵌套问题。
3. 术语完全不能识别。

**步骤（2）：规则总结**：

制定规则的前提：针对的错误是有规律可循的；针对的错误出现的频次占到一定的比例；针对的错误是可以通过规则修正的。

制定规则如下：①如果某字串（长度大于１）被识别为术语，则将其出现在文中的其他位置也标注为术语。②针对茎组成、茎类型、叶组成类名词术语的识别，若字串被识别为术语，如果其前一个字符为“大”、“小”、“短”、“长＂、“老”、“幼”、＂嫩”这些形容词，则将识别出来的字串与这些形容词组合成一个新的术语；针对茎形态、叶表面类包含动名词结构术语的识别，如果被识别为术语的字串前是动词“具”、“有”、＂无”、“带＂、＂近”、＂被”，则将识别出来的字串与这些动词组合成一个新的术语。

**步骤（3）：规则运用**：

这两条规则在运用中需要考虑先后顺序，规则一应先于规则二运用。

20.**模型对未登录低频术语的识别效果**

首先，需要对低频进行界定，文献将词频在1-5之间的词视为低频词。

按照以上定义，语料中的低频术语有144个，占总术语的41%。根据Zipf定律，每个词的词频与其在词频排列里的排名的乘积约是一个常量，因此低频词的词型数 理应较大。在5轮实验中，总的出现57个低频术语，约占总的低频术语的40%。使用规则前发现25个，占比43.86%，使用规则后发现率为49.12%。

**《基于卡方检验的汉语术语抽取》胡文敏**

原理：将长串分解为更短的子串, 通过卡方检验结合参数 F-MI衡量子串之间的关联程度来评估长串成为术语的可能性。

**一．语料收集**

1. **语料预处理**

建立 PatTree索引, 并在索引中统计出所有出现频率在两次以上的字串, 由此构成候选的术语集合表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 候选术语 | 频率 | C-value值 | F-MI |
| Id | term | tf | cvalue | fmi |
| int | string | Int | double | double |

**三．用基于卡方检验的质子串分解方法抽取术语**

**首先**：对候选术语集合进行C-value参数计算, 并填写表 1 中 C-value的值, 对于 C-value

小于给定阈值的候选术语将被从列表中删除;

**然后**：对表中的候选术语进行字符串分解, 并根据分解结果计算所有候选串的 F-MI参数值;

**最后**：根据给定的 F-MI阈值, 淘汰掉错误的候选术语, 并输出最终的术语列表

**1. 质子串分解**

**质词**：不可再分解为更小的词汇单元的词汇。

**合词**：由质词组合而成的词汇。

**合串**：对于串 S, 除了单字串和质串以外, 都是合串。单字既不是质串, 也不是合串。

对于合串 S, 如果 S可以串分解为, 其中 Si可以为质串或单字, 但必须至少有一个是质串, 则称是 S的一种**质子串分解**。

**2. 串分解的 F-MI（改进的互信息参数）**

对于串 S及 S的一种分解, 串分解的 F-MI的计算公式为：

其中：C-value(S)=F(S)-T(S)/C(S)

S：待计算的串

F(S)：S在文档集中出现的次数

T(S)：S所有父串在文档集中出现的次数

C(S)：S所有父串的个数

参数 的定义为：

其中：i表示表中的行变量；j表示列变量。

表示表单元（i,j）的观测值

表示期望值

对于2\*2的表：

其中，N是语料库中二元对的总数。

**3. 串的 F-MI**

对某一质串 (其中均为单字 ), 质串F-MI的计算公式为:

其中, 本文定义单字的 C-value(C)=F(C)。

对某一合串 S, 如果 S的所有质子串分解方式有n种，则分别计算每一种串分解的F-MI值，则合串 S的 F-MI的定义为: