基于自主学习规则的中文物种描述文本的语义标注研究*

段宇锋[1] 黑珍珍[1] 鞠 菲[1] 崔 红[2]
[1](华东师范大学商学院 上海 200241)
[2](美国亚利桑那大学图书馆学与信息资源学院 图森 85719)

【摘要】从《中国植物志》中随机采集 1 000 个文档作为数据集,采用自主学习规则与先导词相结合的算法实现中文物种描述文本的语义标注。实验数据表明,本研究设计的基于规则的算法整体标注效率(F 值)达到 0.930,大部分元素的 F 值在 0.724-0.964 之间,该算法优于朴素贝叶斯分类算法。同时证明,先导词对优化算法具有积极意义。

【关键词】规则 先导词 物种描述文本 语义标注
【分类号】G350

Study on Semantic Markup of Species Description Text in Chinese Based on Auto-learning Rules

Duan Yufeng[1] Hei Zhenzhen[1] Ju Fei[1] Cui Hong[2]
[1](Business School, East China Normal University, Shanghai 200241, China)
[2](School of Information Resource & Library Science, University of Arizona, Tucson 85719, USA)

【Abstract】This paper uses the algorithm of auto-learning rules combining with leading words to implement the semantic markup of species description text in Chinese with the data set of 1 000 documents collected from Flora of China randomly. Experimental results indicate that the whole markup efficiency (the values of F) of rule-based algorithm, which is designed by the study, generally reaches 0.930, and most elements are in the range of 0.724-0.964. Therefore, this algorithm is better than Naive Bayesian categorization algorithm, and it is also proved that leading words are positive for optimizing the algorithm.

【Keywords】Rules Leading words Species description text Semantic markup

日益严峻的环境问题对生态和生物研究提出了更高、更紧迫的要求。物种描述是生物学和生态学的起点,在过去的 200 多年间积累了海量文献。为了满足生物学和生态学领域对物种描述信息的自动语义处理和细粒度检索的需求,物种描述文本的语义标注成为需要解决的关键问题。本研究利用统计学习与先导词相结合的方法自动生成规则,进而依据获得的规则实现对物种描述文本的语义标注。

1 研究现状

物种描述文本的语义组织是生物学和生态学领域知识基础设施建设的重要组成部分。语义标注作为其中的

收稿日期:2012-03-26
收修改稿日期:2012-04-26
*本文系教育部人文社会科学青年项目"基于深度语义标注的网络中文学术信息抽取研究"(项目编号:10YJC870004)的研究成果之一。

关键技术,其研究获得美国、加拿大等国的大力支持。例如,2007 年加拿大自然科学和工程研究委员会(Natural Sciences and Engineering Research Council of Canada, NSERC)资助了项目"The Value of Automated Semantic Annotation for Biodiversity Informatics";2009 年美国国家科学基金会(National Science Foundation, NSF)资助了项目"Fine - Grained Semantic Markup of Descriptive Data for Knowledge Applications in Biodiversity Domains"。

根据实现原理,国内外的主要研究可以分为 4 类:

(1)基于自然语言处理中句法解析方法的研究。早期的研究通过人工分析原始资料的句法和词汇,自建解析器[1,2]。尽管针对特定文本集开发的句法解析器可能具有良好性能,但显然这并不是一种高效的方法,因为每个文本集的词汇和句法都有所不同。

(2)基于本体的研究。MultiFlora 依靠手工创建的领域本体与正则表达式结合,在有监督学习的基础上,实现植物描述特征抽取的准确率和召回率分别为 74% 和 66%[3]。本体也是目前国内相关研究采用最多的方法,例如, 罗贝等[4]利用已有生物本体构建植物本体,从 Web 中获取植物领域知识,建立植物学知识库; 沙丽华[5]依据领域文档特点,提出面向领域文档的语义标注方法 SAMDD,在人工构建玉米本体的基础上,实现玉米领域 Web 文档的语义标注;石静[6]手工构建植物本体,并结合基于正则表达式的句子分类规则,对《中国高等植物图鉴》中的植物描述信息进行抽取。但这些研究建立和使用的本体还不完善,达不到广泛应用的要求。

(3)基于规则的研究。GoldenGATE 是一个典型的基于规则的标注系统,用户使用正则表达式可以实现任意粒度的标注[7],但它需要许多人为干预,例如更正各种错误、调整正则表达式规则,正则表达式也可以通过有监督的机器学习获得。Tang 等[8,9]改造 Soderland 提出的有监督学习算法,将北美植物群落 1 600 种物种叶子的形状、大小、颜色、排列及果实的形状特征填充到预先定义的模板中,对不同特征标注的准确率介于 30% - 100%。

(4)基于统计学习方法的研究。郑家恒等[10]在聚类的基础上,利用主题分布的特点对农作物种子信息进行句子级标注;Cui 等[11]运用基于统计学习的数据挖掘方法对句子进行标注,准确率达到88% - 95%。但有监督学习需要较大规模的训练集,为降低学习成本,Cui 等[12]提出 Bootstrapping - based 无监督学习算法,测试结果显示,无监督学习算法达到了有监督学习算法的标注准确率。

本研究虽然是基于规则实现文本的语义标注,但不像 Soderland、Tang 等的研究那样首先定义模板,任何既定目标以外的领域概念都被忽略不计,而是利用机器学习算法最大限度地识别所有领域概念,提供标注目标,进而标注复杂文本。在本质上,笔者是借鉴 Cui 等的思路,构建适用于中文物种描述文本语义标注的过程和算法。

2 语义标注系统实现

2.1 实现原理

通过分析物种描述文本发现:

(1)物种描述使用一定的模式(习性、根、茎、叶、花、萼片、花瓣、雄蕊、心皮和果实等)和半技术性语言列举分类群的特征;

(2)"并列"和"包含"是模式构成元素之间最基本的关系,因此描述模式可以被表达成树形结构;

(3)文本的语义标注可以转化为文本语料的分类问题,即依据描述模式,逐层判断语句描述的对象(模式元素),并将相应信息插入文本,如图 1 所示:

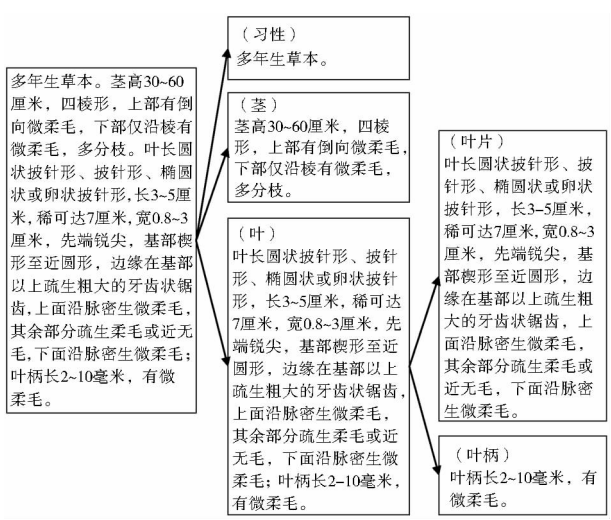

图 1 物种描述文本的语义标注原理

(4)将分析结果转化为 XML 文件,如下所示:

```
<? xml version = "1.0" encoding = "UTF - 8"? >
< description >
```

```
<plant – habit – and – life – style >
  <phls – general > 多年生草本。 </phls – general >
</plant – habit – and – life – style >
<stems >
  <stem – general > 茎高 30 – 60 厘米,四棱形,上部有倒向微柔毛,下部仅沿棱有微柔毛,多分枝。 </stem – general >
</stems >
<leaves >
  <leaf – blade > 叶长圆状披针形、披针形、椭圆状或卵状披针形,长 3 – 5 厘米,稀可达 7 厘米,宽 0.8 – 3 厘米,先端锐尖,基部楔形至近圆形,边缘在基部以上疏生粗大的牙齿状锯齿,上面沿脉密生微柔毛,其余部分疏生柔毛或近无毛,下面沿脉密生微柔毛;
  </leaf – blade >
  <petiole > 叶柄长 2 – 10 毫米,有微柔毛。 </petiole >
</leaves >
</description >
```

依据描述模式定义 XML 文件中的元素,本研究使用Cui 等标注英文植物物种描述文本所用的 Schema[13]。

2.2　学习与标注算法

(1)构建规则集

本研究采用基于规则的方法进行语义标注,规则的构建基于两方面的因素:

①词项的概率分布

文本的语义通过所使用的词项来表达。在表达特定语义时,出现频率高的词项对语义识别的参考价值大;表达任何语义都出现的词项,对语义识别的价值小。

②在句子中,越靠近句首的词项参考价值越高

在英文物种描述文本中,句子的先导词对语义标注具有重要参考价值。笔者发现,中文物种描述文本与英文极为类似,例如“轮伞花序腋生,球形,有总梗或无总梗;花梗纤细,长约 2.5 毫米;花萼管状钟形,长约 2.5 毫米,外有微柔毛及腺点,10 脉,萼齿 5,狭三角状钻形;花冠淡紫色,长约 4 毫米,外面略有微柔毛,内面在喉部以下有微柔毛,冠檐 4 裂,上裂片较大,先端 2 裂,其余 3 裂片等大,先端钝圆;雄蕊 4,前对稍长,稍伸出冠外,花药卵圆形,2 室;花柱顶端 2 裂,裂片近相等。小坚果卵球形,黄褐色,有窝点。”句首以波浪线标出的词明确反映出语句描述的对象,为标注提供了重要线索。因此,本研究设定表示先导词数量的参数 Fr(即语料块的前 Fr 个词),Fr 的取值范围为[0,10]。其中,0 代表采用所有词。

每个词项针对所有语义类都生成一条候选规则,候选规则的表达形式为:{词项,语义类,P,C}。其中:在每条语句中,前 Fr 个词才会生成候选规则,系统可以通过学习过程确定 Fr 的最佳取值。本研究为了验证先导词对标注结果的影响,以人工方式指定系统运行时截取的先导词数量(即 Fr);P = 词项在特定语义类中出现的频次/词项在文本集中出现的总频次;C = 词项在文本集中出现的总频次/文本集包含语句的数量,它用于消除罕用词的影响。

P、C 值大于阈值的候选规则入选规则集,P、C 的阈值系统通过学习过程获得。规则集示例如下所示:

腋芽,buds	疤痕,fruits	荚果,fruits
鳞片,leaves	花瓣,flowers	观赏,phenology
边缘,leaves	种子,seeds	山区,phenology
沿,leaves	泡,leaves	

(2)标　注

语句进行分词处理后,对先导词逐一在规则集中查询是否存在与之匹配的规则,并选择合适的规则进行标注。其标注流程如图 2 所示:

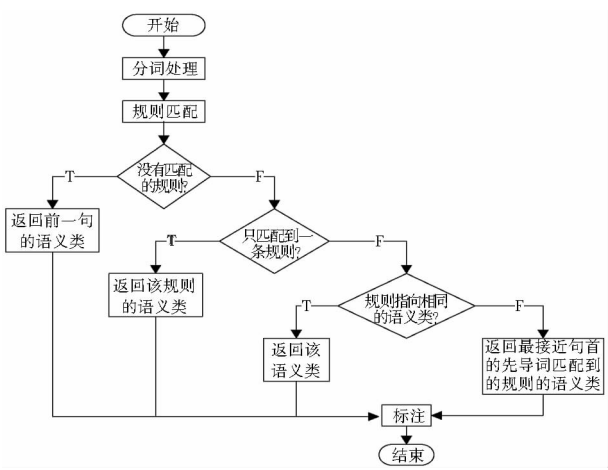

图 2　标注算法流程

(3)学　习

在基于规则的标注系统中,规则学习是语义标注的关键,而目标的标注则退居为次要过程。Fr 的取值是影响候选规则生成的关键因素,只有 P 和 C 的值大于阈值的候选规则{词项,语义类,P,C}才能入选规则集。因此,学习过程必须确定 P、C 的阈值及 Fr 的最佳值。

准确率(Precision)和召回率(Recall)是衡量系统效率最常用的指标。准确率和召回率反向相关,即提升准确率往往以损失召回率为代价,反之亦然。为了兼顾准确率和召回率,采用 F 值衡量系统效率,计算公式为:

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

学习过程如图3所示：

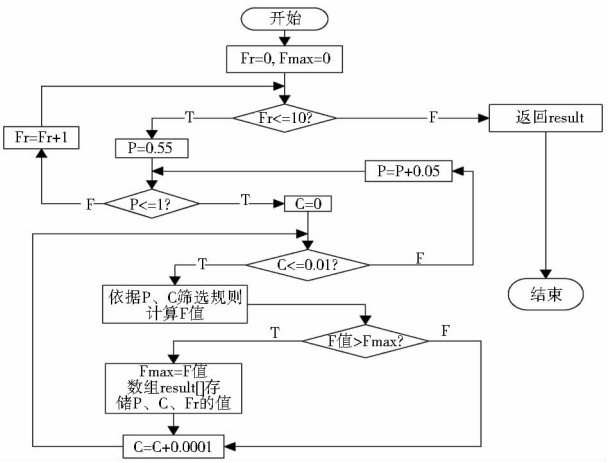

图3　学习算法流程

2.3　系统框架

在自动语义标注系统实现时，系统的学习和标注依据树形结构展开，如图4所示：

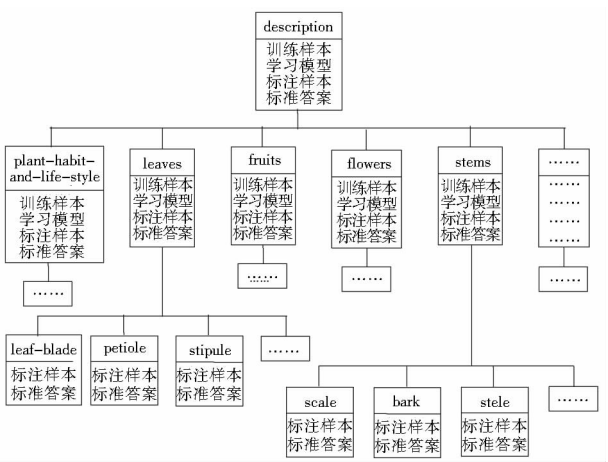

图4　自动语义标注系统的逻辑结构

具体过程为：

(1)系统从Schema中获取层级结构；

(2)学习模块基于层级结构从训练集中获取训练实例，并获得规则集。

以2.1节中的文本标注示例为例，当XML文本被读入至根节点description后，该节点根据XML文本中description包含的子元素plant－habit－and－life－style、stems、leaves、flowers、fruit和phenology，将子元素的文本内容分配至相应的子节点中。例如，子节点leaves获取“叶长圆状披针形、披针形、椭圆状或卵状披针形，长3－5厘米，……；叶柄长2－10毫米，有微柔毛。”leaves根据其内容中包含的子元素，继续将内容分配至其下的子节点，即节点leaves分别向子节点leaf－blade和petiole分配文本片段“叶长圆状披针形、披针形、椭圆状或卵状披针形，长3－5厘米，……；”和“叶柄长2－10毫米，有微柔毛。”这个过程将持续进行直至每个分支都到达终端节点(即不包含子节点的节点)。在此基础上，依据规则过滤条件获得规则集。

(3)根据学习到的规则，对新样例(测试文本)按照树形结构并依据深度优先的原则逐层标注。

当一个新的描述样例出现时，文本被读入根节点description中，并按“；”和“。”切分。对于切分出的每个语料块，系统在规则集中查询与之相匹配的规则，按照匹配结果将该语料块派遣至相应的子节点，如leaves；该节点(如leaves)按“，”继续切分语料，并依据获得的规则集继续查询匹配规则，按匹配到的结果将语料派遣至相应的节点，如把“叶柄长2－10毫米，有微柔毛。”派遣至节点petiole。petiole为终端节点，该节点只接受父节点分配的语料块。至此，该语料标注过程结束。

3　实验数据与结果

3.1　数据集

数据来源为中文版中国植物志(Flora of China，FOC)[14]。本研究采取随机抽样和分层抽样相结合的方式，从中国植物志中采集1 000个文档作为数据集，共涉及37个科，每科大约30个种。为消除偶然性所造成的影响，本研究运用十折交叉验证(10－fold Cross Validation)，每轮随机抽取数据集中的90%作为训练数据，剩余的10%作为测试数据。

3.2　实验结果

为了更直观地评价算法性能，本研究目前只进行了第一层级的标注，即将语料分配至plant－habit－and－life－style(phls)、roots、stems、buds、leaves、flowers(flow)、fruits、seeds、spore－related－structures(spore)、phenology(phe)和compound(comp)后不再进行更深层次的标注。“compound”用于标注描述两种或两种以上植物结构的语料，例如“苞片和小苞片线形”。

每个节点都在给定的Fr上进行自主规则学习分类。在实验中，Fr的值分别为0－10，系统在不同Fr值时得到的结果如表1所示，不同元素的标注效率如表2所示。

表 1　整体标注性能(F 值)

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	第 6 组	第 7 组	第 8 组	第 9 组	第 10 组	平均值
0	0.953	0.931	0.913	0.925	0.909	0.934	0.937	0.918	0.921	0.919	0.926
1	0.933	0.904	0.898	0.909	0.895	0.912	0.918	0.895	0.902	0.920	0.909
2	0.945	0.927	0.910	0.927	0.910	0.928	0.940	0.911	0.928	0.913	0.924
3	0.949	0.932	0.912	0.933	0.918	0.939	0.937	0.923	0.930	0.926	0.930
4	0.947	0.922	0.914	0.925	0.916	0.934	0.946	0.926	0.923	0.927	0.928
5	0.950	0.924	0.920	0.926	0.915	0.934	0.947	0.930	0.925	0.928	0.930
6	0.945	0.915	0.912	0.920	0.912	0.928	0.943	0.926	0.923	0.923	0.925
7	0.951	0.918	0.914	0.920	0.910	0.932	0.933	0.922	0.928	0.924	0.925
8	0.949	0.930	0.923	0.920	0.910	0.930	0.944	0.922	0.928	0.924	0.928
9	0.947	0.925	0.919	0.920	0.913	0.932	0.942	0.922	0.933	0.924	0.928
10	0.947	0.922	0.919	0.922	0.913	0.932	0.940	0.922	0.927	0.923	0.927

表 2　各元素的标注性能(F 值)

	0	1	2	3	4	5	6	7	8	9	10
buds	0.964	0.950	0.928	0.924	0.941	0.941	0.931	0.952	0.952	0.952	0.941
flow	0.943	0.963	0.956	0.956	0.952	0.953	0.948	0.945	0.947	0.944	0.944
fruits	0.848	0.888	0.862	0.866	0.865	0.868	0.862	0.845	0.859	0.855	0.857
leaves	0.938	0.949	0.957	0.958	0.955	0.955	0.949	0.955	0.953	0.954	0.953
phls	0.915	0.750	0.856	0.882	0.884	0.883	0.880	0.886	0.899	0.907	0.905
roots	0.266	0.000	0.040	0.147	0.173	0.340	0.192	0.226	0.266	0.266	0.266
seeds	0.945	0.926	0.903	0.922	0.934	0.934	0.934	0.934	0.927	0.927	0.924
spore	0.931	0.951	0.921	0.900	0.908	0.937	0.928	0.922	0.928	0.919	0.911
stems	0.872	0.724	0.847	0.874	0.886	0.889	0.878	0.881	0.875	0.876	0.874
comp	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
phe	0.956	0.945	0.955	0.956	0.942	0.944	0.943	0.947	0.952	0.952	0.950

3.3　分析与讨论

(1)与朴素贝叶斯分类算法的比较

本研究同时采用朴素贝叶斯(Naïve Bayes,NB)分类与先导词相结合的方法对相同的数据集进行标注,整体标注性能如表 3 所示:

表 3　NB 算法的整体标注性能(F 值)

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	第 6 组	第 7 组	第 8 组	第 9 组	第 10 组	平均值
0	0.801	0.793	0.817	0.803	0.770	0.812	0.798	0.777	0.801	0.781	0.795
1	0.924	0.891	0.868	0.867	0.891	0.901	0.900	0.892	0.885	0.888	0.891
2	0.916	0.901	0.886	0.898	0.891	0.913	0.922	0.885	0.911	0.899	0.902
3	0.909	0.891	0.877	0.898	0.882	0.916	0.901	0.878	0.898	0.889	0.894
4	0.893	0.884	0.870	0.876	0.861	0.911	0.887	0.868	0.879	0.872	0.880
5	0.887	0.865	0.869	0.881	0.858	0.896	0.879	0.866	0.875	0.871	0.875
6	0.884	0.848	0.857	0.870	0.846	0.882	0.873	0.853	0.868	0.868	0.865
7	0.880	0.857	0.852	0.862	0.843	0.878	0.864	0.857	0.855	0.860	0.861
8	0.868	0.857	0.836	0.849	0.834	0.861	0.853	0.854	0.842	0.844	0.850
9	0.861	0.847	0.836	0.849	0.825	0.861	0.844	0.840	0.834	0.837	0.843
10	0.870	0.840	0.834	0.848	0.833	0.866	0.850	0.839	0.829	0.836	0.844

运用 ANOVA 模型分析两种算法标注效率(以 F 值衡量)之间的差异。结果表明,本方法显著优于朴素贝叶斯分类算法($P < .0001$)。

(2)Fr 对标注效率(以 F 值衡量)的影响

根据表 1 和表 3 中的数据,对于 $Fr \in [0,10]$,采用 ANOVA 模型分析 F 值之间的差异,如表 4 所示。

表4　F 值的差异分析(基于朴素贝叶斯分类的算法)

Pr > |t|	Fr=0	Fr=1	Fr=2	Fr=3	Fr=4	Fr=5	Fr=6	Fr=7	Fr=8	Fr=9	Fr=10
Fr=0	—										
Fr=1	<.0001	—									
Fr=2	<.0001	0.0012	—								
Fr=3	<.0001	0.3341	0.0198	—							
Fr=4	<.0001	0.0032	<.0001	0.0001	—						
Fr=5	<.0001	<.0001	<.0001	<.0001	0.11	—					
Fr=6	<.0001	<.0001	<.0001	<.0001	<.0001	0.006	—				
Fr=7	<.0001	<.0001	<.0001	<.0001	<.0001	0.0001	0.2281	—			
Fr=8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.002	—		
Fr=9	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.063	—	
Fr=10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1191	0.7579	—

表4显示,基于朴素贝叶斯分类的标注算法在 Fr∈[1,10]时的 F 值与 Fr=0 时的 F 值具有显著差异(P<0.0001)。采用先导词后 F 值平均提高 0.048 – 0.107。

基于规则的算法对先导词的敏感性则相对较低。表5显示,当 Fr∈{1,3,5}时 F 值与 Fr=0 时的 F 值有显著差异(P< 0.05)。在表1中,当 Fr={1,2,6,7}时 F 值的平均值小于 Fr=0 时 F 值的平均值。因此,只有在 Fr∈{3,5}时,先导词对标注效率(F 值)的提升才具有统计意义。尽管如此,笔者认为先导词仍是设计标注算法时重点考虑的因素,其原因有二:Fr∈{3,5}时 F 值的提升具有统计学意义,这一点不可否认;采用先导词可以有效降低系统的计算成本。在小样本测试时可以忽视由此产生的优势,但对于实用系统在处理海量文本时则具有重大意义。

表5　F 值的差异分析(基于规则的算法)

Pr > |t|	Fr=0	Fr=1	Fr=2	Fr=3	Fr=4	Fr=5	Fr=6	Fr=7	Fr=8	Fr=9	Fr=10
Fr=0	—										
Fr=1	<.0001	—									
Fr=2	0.3070	<.0001	—								
Fr=3	0.0304	<.0001	0.017	—							
Fr=4	0.2375	<.0001	0.0292	0.3153	—						
Fr=5	0.0269	<.0001	0.0015	0.9591	0.2914	—					
Fr=6	0.5160	<.0001	0.7083	0.0054	0.0689	0.0047	—				
Fr=7	0.8285	<.0001	0.4200	0.0177	0.1631	0.0155	0.6647	—			
Fr=8	0.2318	<.0001	0.0300	0.3101	0.9913	0.2865	0.0705	0.1663	—		
Fr=9	0.3008	<.0001	0.0415	0.2499	0.8825	0.2296	0.0939	0.2116	0.8910	—	
Fr=10	0.6229	<.0001	0.1318	0.0916	0.4885	0.0823	0.2550	0.4791	0.4953	0.5855	—

(3)元素的标注效率

从表2中可以看到,大部分元素的标注效率较为理想。无论 Fr 取何值,F 值始终在 0.724 – 0.964 之间,如图5所示:

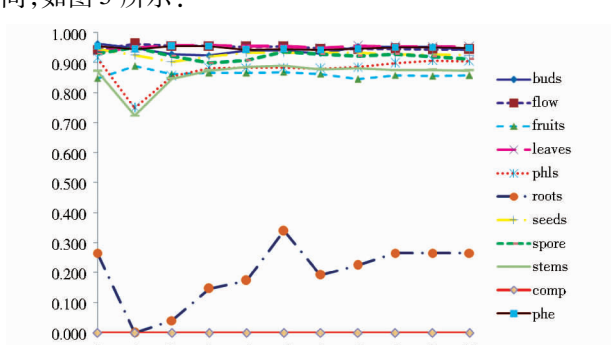

图5　Fr 值对各元素标注效率的影响(F 值)

标注效果最差的是 compound,其 F 值为0,这是因为任何描述内容都可能出现在该复合元素中,学习算法不能很好地处理这种情况;其次是 roots,F 值最高为0.340,最低为0,这主要是由于该元素的训练数据过于稀疏,且 roots 经常被嵌套到元素 stems 中。

4　结　语

实验结果表明,对于标注中文物种描述文本来说,本文的算法性能优于朴素贝叶斯分类算法,先导词对优化算法,提升标注效率具有一定价值。通过对标注结果的分析,可以发现专业领域语词切分错误是导致标注错误的重要原因。因此,后续研究应将专业领域的新词识别作为需要解决的重要问题,从而进一步提升标注性能。

参考文献：

[1] Taylor A. Extracting Knowledge from Biological Descriptions[C]. In: *Proceedings of the 2nd International Conference on Building and Sharing Very Large – Scale Knowledge Bases*. 1995:114 – 119.

[2] Vanel J M. Worldwide Botanical Knowledge Base [EB/OL]. [2011 – 10 – 11]. http://wwbota.free.fr/.

[3] Wood M M, Lydon S J, Tablan V, et al. Using Parallel Texts to Improve Recall in IE[C]. In: *Proceedings of International Conference on Recent Advances in Natural Language Processing* (*RANLP*). Amsterdam: John Benjamins, 2004:70 – 77.

[4] 罗贝,吴洁,曹存根,等. 从文本中获取植物知识方法的研究[J]. *计算机科学*,2005,32(10):6 – 13. (Luo Bei, Wu Jie, Cao Cungen, et al. Botanical Knowledge Acquisition from Text[J]. *Computer Science*, 2005,32(10):6 – 13.)

[5] 沙丽华. 面向领域文档的语义标注方法研究[D]. 长春:吉林大学,2009. (Sha Lihua. Research on Semantic Annotation for Domain Documents[D]. Changchun: Jilin University,2009.)

[6] 石静. 基于本体的植物信息抽取与分析研究[D]. 西安:西北农林科技大学,2010. (Shi Jing. Information Extraction and Analysis Based on Plant Ontology[D]. Xi'an: Northwest Agriculture and Foresty University, 2010.)

[7] Sautter G, Bohm K, Agosti D. A Combining Approach to Find all Taxon Names[J]. *Biodiversity Informatics*,2006(3):46 – 58.

[8] Tang X Y, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval[EB/OL]. [2011 – 08 – 10]. http://www.tdwg.org/proceedings/article/view/195/.

[9] Soderland S. Learning Information Extraction Rules for Semi – Structured and Free Text[J]. *Machine Learning*, 1999, 34 (1 – 3): 233 – 272.

[10] 郑家恒,营小艳. 农作物信息抽取系统的设计与实现[J]. *计算机工程*,2006,32(7):197 – 198,220. (Zheng Jiaheng, Jian Xiaoyan. Design and Realization of the System of Farm Crop Information Extraction[J]. *Computer Engineering*, 2006, 32(7):197 – 198,220.)

[11] Cui H, Heidorn P B. The Reusability of Induced Knowledge for Automatic Semantic Markup of Taxonomic Descriptions[J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(1): 133 – 149.

[12] Cui H, Boufford D, Selden P. Semantic Annotation of Biosystematics Literature Without Training Examples[J]. *Journal of the American Society of Information Science and Technology*, 2010, 61 (3): 522 – 542.

[13] Cui H. The XML Schema for MARTT[EB/OL]. [2012 – 08 – 08]. http://publish.uwo.ca/ ~ hcui7/research/xmlschema.xsd.

[14] 中国植物志编辑委员会. 中国植物志[M]. 北京:科学出版社,1959. (Flora of China Editorial Committee. Flora of China [M]. Beijing: Science Press, 1959.)

(作者 E – mail:damolvying@126.com)