



本体与条件随机场结合的涉农商品名称抽取与类别标注

黄念娥^{1,2}, 黄河^{1*}, 王儒敬¹

(1. 中国科学院 合肥智能机械研究所, 合肥 230031; 2. 中国科学技术大学 合肥物质研究院, 合肥 230027)

(*通信作者电子邮箱 hhuang@iim.ac.cn)

摘要:传统的基于条件随机场(CRF)的信息抽取方法在进行涉农商品名称抽取与类别标注时,需要大量的训练语料,标注工作量大,且抽取精度不高。为解决该问题,提出了一种基于农业本体与CRF相结合的涉农商品名称抽取与类别标注方法,将涉农商品名称的自动抽取与分类看作序列标注的任务。首先是原始数据的分词处理和词、词性、地理属性、本体概念特征选择;然后,采用改进的拟牛顿算法训练CRF模型参数,用维特比算法实现解码,共完成4组对比实验,识别出7种类别,并将CRF和隐马尔可夫模型(HMM)、最大熵马尔可夫模型(MEMM)通过实验进行比较;最后,将CRF应用于农产品供求趋势分析。结合合适的特征模板,本体概念的加入使CRF开放测试的总体准确率提高10.20%,召回率提高59.78%, F 值提高37.17%,证明了本体与CRF结合方法在涉农商品名称和类别抽取中的可行性和有效性,可以促进农产品供求对接。

关键词:条件随机场;农业本体;涉农商品名称;供求趋势;序列标注

中图分类号: TP391.1; TP18 **文献标志码:** A

Agriculture-related product name extraction and category labeling based on ontology and conditional random field

HUANG Nian'e^{1,2}, HUANG He^{1*}, WANG Rujing¹

(1. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei Anhui 230031, China;

2. Hefei Institute of Physical Science, University of Science and Technology of China, Hefei Anhui 230027, China)

Abstract: Traditional information extraction method based on Conditional Random Field (CRF) requires large-scale labeled corpus, it is expensive to label corpus manually and the extraction precision is low in processing agriculture-related product name extraction and category labeling. In order to solve this problem, a method of agriculture-related product name extraction and category labeling based on agricultural ontology and CRF was proposed, automatic extraction and classification of agriculture-related product names was regarded as sequence labeling. Firstly, original data was processed, word, part of speech, geographical attributes and ontology concept features were selected. Then, parameters of the CRF model were trained by the improved quasi-Newton algorithm and decoding was implemented by Viterbi algorithm. A total of four groups of comparative experiments were completed and seven categories were identified. CRF, Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM) were compared through experiments. Finally, the supply and demand trend analysis of agriculture produce was accomplished. The experimental results show that the overall precision, recall and F -score of the open test were increased by 10.20%, 59.78% and 37.17% respectively by adding ontology concepts with appropriate CRF features; it also proves the feasibility, effectiveness and practical significance of the method in promoting automatic supply and demand docking of agricultural products.

Key words: Conditional Random Field (CRF); agricultural ontology; agriculture-related product name; supply and demand trend; sequence labeling

0 引言

随着互联网的快速发展,目前已有超过30000家的涉农电商平台^[1],如阿里巴巴农业频道、中国惠农网、顺丰优选等,这些网站每天会发布大量种植业、林木花卉、农机、农具等各类涉农商品信息。通过对这些供求信息分析,有助于预测农产品市场趋势、及时发现买难卖难、促进供求自动对接。

然而,对这些涉农供求信息分析之前首先需要对涉农商品名称与类别进行抽取。如“厂家直销 两行玉米播种机 免剥皮玉米脱粒机”这条供求信息中,需要抽取“玉米播种机”和“玉米脱粒机”这两个涉农商品名称,同时类别标注为农业机械类。这样,就可以对一段时间内、不同地域的农业机械类的供求情况进行趋势分析。

涉农商品名称自动抽取与类别标注主要涉及农业领域术

收稿日期: 2016-08-02; 修回日期: 2016-09-19。

基金项目: 国家科技支撑计划项目(2013BAD15B03); 中国科学院重点部署项目(Y622A21291); 安徽省科技攻关项目(1401032010)。

作者简介: 黄念娥(1991—),女,安徽安庆人,硕士研究生,主要研究方向:信息抽取、垂直搜索引擎; 黄河(1980—),男,安徽合肥人,副研究员,博士,主要研究方向:农业大数据、农业智能系统; 王儒敬(1964—),男,安徽亳州人,研究员,博士,主要研究方向:知识表示与可视化、知识获取。



语自动抽取,包括基于规则与基于统计两种方法。基于规则方法依赖于语言和领域规则模板的建立^[2],需要人工编制大量规则和有经验的领域专家,系统可移植性差。基于统计的方法分为经典的统计方法和统计机器学习方法。经典的统计方法主要基于词频、互信息以及信息熵等。Guan 等^[3]利用关联规则、C-value 和词频-逆向文件频率 (Term Frequency-Inverse Document Frequency, TF-IDF) 混合算法提取中国戏曲领域的专业术语。该方法克服了基于规则的缺点,但低频术语通常不能被有效提取。统计机器学习方法中,最具代表性的为条件随机场 (Conditional Random Field, CRF)^[4-7],利用序列标注的思想,融合上下文多特征提取领域术语。由于其条件独立性,只需考虑当前已经出现的观测状态特性,对于整个序列内部的信息和外部观测信息均可有效利用,避免了标记偏置问题,被广泛应用。孟洪宇^[8]通过 CRF 融合字符本身、词性、词边界等多特征提取中医术语, F 值达到 75.56%。Zhan 等^[9]利用两层 CRF 提取简单和复杂的术语,并通过领域相关性和一致性提取最终领域术语, F 值为 82.01%。

传统 CRF 需要大规模的训练语料^[10-12]。针对涉农商品名称抽取与类别标注,由于涉农商品名称繁多,人工标注工作量大。如“玉米收割机”进行了标注,但当遇到“小麦收割机”时,如果样本没有标注,依然不能正确抽取,影响了抽取的精确率。而事实上,如果将“玉米”“小麦”的父类概念“粮油作物”作为 CRF 的一项特征,可实现由“玉米收割机”抽取出新词“小麦收割机”。因此为实现对属于同一概念的大量新词(指未在样本中标注的词)进行有效抽取,文中将农业本体与 CRF 相结合,引入词所对应的本体概念作为 CRF 的特征,赋予涉农商品名称以语义知识,同时结合词、词性、地理位置特征进行 CRF 训练,最终实现涉农商品名称的抽取与类别标注。通过学习样本,CRF 模型表现出一定的“推理”能力,如将概念为粮油作物和收获机械的相邻实例词作为一个涉农商品名称抽取,类别识别为农业机械类,概念为生鲜水果和农作物种子种苗的相邻实例词抽取为种植业类的涉农商品名称等;并将 CRF 与隐马尔可夫模型 (Hidden Markov Model, HMM)、最大熵马尔可夫模型 (Maximum Entropy Markov Model, MEMM) 进行比较,同时用于农产品供求趋势的分析。表明农业本体与 CRF 相结合进行涉农商品名称抽取与类别标注方法的有效性。

1 农业本体与 CRF

1.1 农业本体

本体是关于概念体系的明确的、形式化的规范说明^[13],农业本体是专业性的本体,表示的知识都是针对农业学科领域,提供了关于该领域中概念的词表以及概念之间的关系^[14-15]。

概念层次是本体的骨架,主要反映概念之间的父子类关系。文中使用阿里巴巴农业 (<https://www.1688.com/>) 概念层次体系,结构如图 1 所示,该分类体系有 4 个层次,包括 218 个叶子节点,目前已有超过 170 万个农业供求信息映射到该分类体系中,因此基本可以涵盖各种农产品供求类型,具有很强的覆盖性。利用本体中的父子类概念知识表示词所对

应的概念,赋予词以语义。生鲜水果作为苹果、草莓的父类概念,可用生鲜水果描述苹果、草莓;种植业作为生鲜水果、农作物种子种苗的父类概念,使用种植业来描述生鲜水果、农作物种子种苗,也可使用种植业来描述苹果、草莓、蔬菜种子种苗等,进一步增强知识泛化能力。

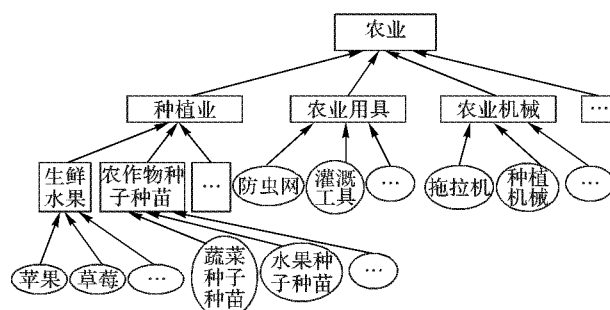


图 1 农业本体概念层次树

Fig. 1 Concept hierarchies of agricultural ontology

1.2 条件随机场

CRF 是用来标注和划分序列结构数据的概率化的无向图模型^[4],具有表达元素长距离依赖性和交叠性特征的能力,在模型中可包含众多领域知识^[16]。

1.2.1 CRF 模型

对于给定的输出标记序列 $\mathbf{y} = (y_1, y_2, \dots, y_n)$ 和输入观察序列 $\mathbf{x} = (x_1, x_2, \dots, x_n)$, CRF 通过定义条件概念 $p(\mathbf{y} | \mathbf{x}, \lambda)$ 来描述模型。图 2 表示 CRF 链式结构。

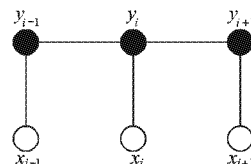


图 2 CRF 链式结构

Fig. 2 CRF chain structure

CRF 定义的条件概率公式为:

$$p(\mathbf{y} | \mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \exp \left(\sum_{i=1}^n \sum_j \lambda_j \cdot f_j(y_{i-1}, y_i, \mathbf{x}, i) \right) \quad (1)$$

其中: \mathbf{x} 为观察序列; \mathbf{y} 为标记序列; $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ 为权重向量; λ_j 为特征函数的权重; $f_j(y_{i-1}, y_i, \mathbf{x}, i)$ 为对应整个观察序列 \mathbf{x} , 标记位于 i 和 $i-1$ 的特征函数; 分母 $Z(\mathbf{x})$ 为归一化因子 (保证所有可能的状态序列概率之和为 1), 公式如下:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp \left(\sum_{i=1}^n \sum_j \lambda_j \cdot f_j(y_{i-1}, y_i, \mathbf{x}, i) \right) \quad (2)$$

1.2.2 参数训练

CRF 的参数训练过程是在训练数据集上基于对数似然函数的最大化进行^[17-18], 设一个标注过的数据序列集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, $\bar{p}(\mathbf{x}, \mathbf{y})$ 为训练样本中 (\mathbf{x}, \mathbf{y}) 的经验概率, $\bar{p}(\mathbf{x})$ 为随机变量 (观察序列) \mathbf{x} 在训练样本中的经验分布, $\ln Z(\mathbf{x})$ 表示 $Z(\mathbf{x})$ 的以 e 为底的自然对数函数, 则 CRF 模型中的极大似然函数为:

$$L(\lambda) = \sum_{\mathbf{x}, \mathbf{y}} \bar{p}(\mathbf{x}, \mathbf{y}) \sum_{i=1}^n \left(\sum_j \lambda_j \cdot f_j(y_{i-1}, y_i, \mathbf{x}, i) \right) - \sum_{\mathbf{x}} \bar{p}(\mathbf{x}) \ln Z(\mathbf{x}) \quad (3)$$



对 λ_j 求导:

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = \sum_{x,y} \hat{p}(x,y) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) - \sum_{x,y} \hat{p}(x) p(y|x, \lambda) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) = E_{\hat{p}(x,y)} [f_j(x,y)] - \sum_k E_{p(y|x_k, \lambda)} [f_j(x_k, y)] \quad (4)$$

$E_{\hat{p}(x,y)} [f_j(x,y)]$ 、 $E_{p(y|x_k, \lambda)} [f_j(x_k, y)]$ 分别表示经验分布和模型分布中特征的期望值;令式(4)等于0,求 λ 。

由于改进的拟牛顿算法(Limited Broyden-Fletcher-Goldfarb-Shanno, L-BFGS)只保存并利用近几次迭代(迭代次数由使用者控制)的曲率信息来构造海森矩阵的近似矩阵,每次迭代的开销小,执行速度快,能保证近似矩阵的正定,算法的鲁棒性强^[19]。本文选取 L-BFGS 算法估计似然参数 λ 。

1.2.3 解码问题

对于 x 来说,CRF 要做的就是搜索概率最大的 y^* ,即求解式(5):

$$y^* = \arg \max_y p(y|x, \lambda) \quad (5)$$

该式可通过维特比动态规划算法^[4]进行计算,对状态序列作出最优估计。

模型的具体实现中,使用了 Taku 开发的 CRF++-0.58 工具包^[20],该工具包支持自定义特征集,可输出所有候选的边际概率值,含训练参数时的 L-BFGS 算法以及解码时的维特比算法,可被应用到各种各样的自然语言处理任务中。实验在 64 位 Windows 7 下,装有 Java、C++ 编译环境进行,其他配置为 Intel Pentium P6200,2.13 GHz,2.00 GB RAM。

2 数据集和特征选择

2.1 数据集

数据集选自构建农业本体时使用的阿里巴巴网,从中抽取标题数据,包括七大类:种植业、园艺业、养殖业、化肥、农业用具、农业机械及鲜活水产品加工制品,覆盖了该网站中近 90% 的农产品信息,每类 500 条。

在转换原始语料格式,构造标准的数据集时,利用基于开源 HanLP 自然语言处理包^[21]的 CRF 分词。分词得到词和词性,并去除停用词,如“阿里巴巴”“淘宝”“顺丰”“包邮”等。如“大量供应优质红小麦”CRF 分词后为“大量/m, 供应/vn, 优质/b, 红小麦/nz”,首先利用 Java 程序经过“,”分隔,得到每个词的词和词性组合,再经由“/”分隔,即可转换为符合 CRF++-0.58 工具包的输入格式。因涉农商品名称很多由三个及以上词组成,选取 5 词位标注法,以词为单位进行序列标注,标注符号集为(B,M,E,S,O),为实现类别标注,添加符号集(Z,L,YZ,H,Y,J,X)作为序列标注符号的后缀,各个符号含义如表 1 所示。如涉农商品名称为“玉米小麦播种机”农业机械类中,标注为玉米(B-J)小麦(M-J)播种机(E-J)。

2.2 特征选择

CRF 标注算法中,特征选择以及特征函数的定义至关重要,直接关系到模型的性能。CRF 模型的特征一般分为三类^[22]:原子特征、复合特征以及全局变量特征,针对不同语料,选取的特征不同。选取词 Word、词性(Part-Of-Speech,

POS)、地理属性和农业本体概念作为特征。构建特征模板时,使用了对应的原子特征和复合特征,上下文特征窗口为 5。

表 1 序列标注符号含义

Tab. 1 Symbolic meanings of sequence labeling

符号	含义	符号	含义
B	涉农商品名称开始词	L	园艺业类
M	涉农商品名称中间词	YZ	养殖业类
E	涉农商品名称结尾词	H	化肥类
S	单个涉农商品名称	Y	农业用具类
O	非涉农商品名称	J	农业机械类
Z	种植业类	X	鲜活水产品加工制品类

2.2.1 词

由于涉农商品名称具有领域性,有些词只在本领域流通,故词本身包含了最有效的信息,可作为特征。如“拖拉机”“玉米渣”“叶面肥”可作为农业领域的商品名称。

2.2.2 词性

词性特征指当前字符的词性,是涉农商品名称的一个重要特征,一般而言涉农商品名称为名词,复合名词,还包括部分动词。如“麦麸/n”“狼/n 青犬/nz”“麦秆/n 捡拾/v 打捆机/n”可作为涉农商品名称。

2.2.3 地理属性

涉农商品名称中有些涉及到地理属性,如“山东开沟机”“河南特产玉米”“黑龙江大豆”。对于这类数据,应将其地理属性抽取出来,分词后词性标注为“ns”的表示地名,因此可很方便地将地理属性作为特征加入到 CRF 中。

2.2.4 农业本体概念

选取词在农业本体中所对应的概念作为 CRF 的一项特征,将词进行泛化,利用概念知识表示实例词,使词具有语义。共使用 2 种本体概念,一种是实例词在农业本体概念层次树中对应的叶子节点概念,特征表示为 F0;另一种是实例词在本体中对应的上层概念,在此指去除叶子节点和根节点后所对应的概念,特征用 F1 表示。文中使用的农业本体概念如表 2 所示。如“菠萝莓”对应的叶子节点概念为“草莓”,对应的上层概念为“生鲜水果”和“种植业”。

表 2 词所对应的本体概念关系

Tab. 2 Ontology concept relations of word correspondence

词在本体中对应的叶子节点概念	词在本体中对应的上层概念
山楂、甘蔗、草莓、西瓜	生鲜水果、种植业
玉米、薯类、麦类、高粱	粮油作物、种植业
水果种子种苗	农作物种子种苗、种植业
粮食作物种子	农作物种子种苗、种植业
花盆容器、庭院灯、亭子	园林资材、园艺业
园林植物	园艺业
猪、马、牛	家畜、养殖业
鸡蛋、鸭蛋、鹅蛋	禽蛋、养殖业
复合肥、氮肥、叶面肥	化肥
防虫网、遮阳网、农用工具、渔业用具	农业用具
拖拉机、收获机械、种植机械	农业机械
干制水产品、腌制水产品	鲜活水产品加工制品

词所对应的农业本体概念通过维护领域词典实现。而中



国搜农网供求搜索栏目(<http://www.sounong.net/>)共搜集全国 1 万多个农业网站,拥有超过 3 万条农产品信息,实现了农产品到类别的映射,将该知识与阿里巴巴分类体系建立联系,实现涉农商品名称到概念的映射,降低人工维护领域词典的代价,提高自动化程度。图 3 表示词所对应的本体概念标注实现流程。

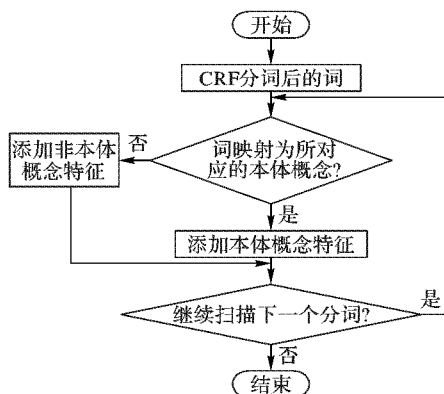


图 3 词所对应的本体概念标注流程

Fig. 3 Labeling for ontology concepts of word correspondence

3 实验及结果分析

3.1 实验评价指标

涉农商品名称抽取与类别标注的结果评价使用 3 个指标:准确率 P 、召回率 R 和 F -值^[23],公式表示如下:

$$\text{准确率}(P) = \frac{\text{正确的名称数}}{\text{抽取的名称总数}} \times 100\% \quad (6)$$

$$\text{召回率}(R) = \frac{\text{正确的名称数}}{\text{标准结果中名称总数}} \times 100\% \quad (7)$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (8)$$

3.2 基于 CRF 涉农商品名称抽取与类别标注

选取数据集中 70% 为训练数据,30% 为测试数据,实现开放测试。实验分为 4 组,每组包括 7 大类,即种植业、园艺业、养殖业、化肥、农业用具、农业机械和鲜活水产品加工制品。第 1 组选取词 Word、词性 POS、地理属性作为特征;第 2 组在前组的基础上,加入词在农业本体概念层次树中对应的叶子节点概念特征 F_0 ;第 3 组基于第一组实验的特征,直接加入词在农业本体中对应的上层概念特征 F_1 ;第 4 组在第 3 组实验特征基础上,添加特征 F_0 。实验总体流程如图 4 所示。

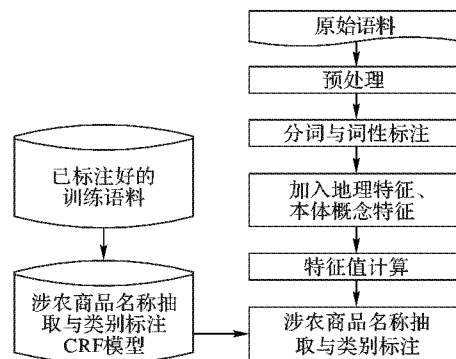


图 4 实验总体流程

Fig. 4 Experimental flow chart

实验结果如表 3 所示,在第 1 组特征基础上,加入本体中对应的叶子节点特征 F_0 ,总的准确率 P 和召回率 R 上升;加入本体中上层概念特征 F_1 ,总体召回率大幅度地上升;同时使用 F_0 和 F_1 特征,准确率 P 高的同时也保证了召回率 R 高,总体 F 值达到 92.32%,其中类别标记为化肥类的 F 值最高 96.00%,园艺业类的 F 值最低 87.50%,表明基于本体与 CRF 相结合进行涉农商品名称的抽取与类别标注的方法是有效的。

表 3 基于 CRF 实验结果
Tab. 3 Experimental results based on CRF

类别	第 1 组(词 + 词性 + 地理特征)			第 2 组(词 + 词性 + 地理特征 + F_0)			第 3 组(词 + 词性 + 地理特征 + F_1)			第 4 组(词 + 词性 + 地理特征 + F_0 + F_1)		
	P	R	F	P	R	F	P	R	F	P	R	F
种植业类	78.44	71.85	75.00	88.32	69.36	77.70	91.05	88.65	89.83	94.27	92.07	93.16
园艺业类	92.36	36.52	52.34	92.59	44.08	59.73	88.92	90.93	89.91	86.85	88.16	87.50
养殖业类	94.74	34.50	50.58	91.89	43.45	59.00	90.94	89.78	90.36	89.87	87.22	88.53
化肥类	96.12	52.66	68.04	92.59	53.19	67.57	93.58	93.09	93.33	96.26	95.74	96.00
农业用具类	83.68	70.37	76.45	94.67	69.72	80.30	90.91	89.32	90.11	93.13	91.50	92.31
农业机械类	83.37	68.65	75.30	88.92	71.63	79.34	91.25	91.07	91.16	94.40	93.65	94.02
鲜活水产品加工制品类	86.96	46.51	60.61	89.67	56.51	69.33	91.80	91.16	91.48	95.01	93.02	94.00
总体值	84.40	57.36	68.30	90.64	60.70	72.71	91.03	90.25	90.64	93.01	91.65	92.32

第 1 组实验错误主要有:名词组合“广西/ns 产地亚/nz 热带/n”“天山/ns 牌/n”“上海/ns 强力/n”“荷兰/ns 十五/nz”等提取为术语;“花卉/n”“磷肥/n”“滴灌管/n”“鲍鱼汁/nz”等未被正确识别;“玉米/nf./nz 小麦/n”“现货/n 鸵鸟蛋/nf”“爆款/nz 低价/n 香蕉/nf”“高产量/nz 玉米/nf 收割机/n”等作为一个整体抽取出来;养殖业、农业机械类的涉农商品名称如“比利时野兔”“山东开沟机”等错误抽取为种植业类。在大量新的涉农商品名称未被有效抽取与分类的前提下,保证准确率高,但召回率低,总体 F 值为 68.30%。

第 2 组实验中,加入词在农业本体概念层次树中对应的叶子节点概念特征 F_0 ,减少了错误分类的概率,可将第 1 组实验中错误分类的部分名称正确抽取分类;同时削弱词 Word、词性 POS 特征的权重,降低了将非涉农商品名称的名词组合错误识别为涉农商品名称的比率,但泛化能力较弱,对于新的涉农商品名称抽取与分类能力很差,准确率和召回率得到提升,总体 F 值为 72.71%。

第 3 组直接使用农业本体中的上层概念特征 F_1 ,赋予词以概念知识,大大增强泛化程度,抽取“菠萝”“浇花喷壶”



“芝麻香油机”“鱿鱼干”等新词。通过学习样本,CRF 模型表现出一定的“推理”能力,如将概念为生鲜水果的单独实例词抽取为种植业类的涉农商品名称,概念为粮油作物和种植机械的相邻实例词抽取为农业机械类的涉农商品名称等。最终召回率大幅度提升,总体 F 值达到 90.64%。

第 4 组综合第 2,3 组实验的特征,使用更详细的特征和特征模板,准确率和召回率有所提升,总体 F 值为 92.32%。其中园艺业、养殖业类的 F 值与其他 5 类相比较低,主要是由于分词错误影响较大以及地理属性未被有效抽取,如将“樟子松木”分词为“樟子/n 松木/n”,“河北小猪”抽取为“小猪”。表 4 列出了抽取的部分涉农商品名称以及标注的类别。

3.3 CRF 与 HMM、MEMM 算法的比较

利用相同的数据集,选取上述第 1 组实验中词、词性、地理属性作为特征,分别利用 CRF 和 HMM、MEMM 完成开放测试,其中后两种算法采用机器学习语言工具包 (Machine Learning for Language Toolkit, MALLET)^[24] 实现, MALLET 是用于文本分类、主题建模和序列标注等的 Java 工具包,实验结果如表 5。

表 4 抽取的部分涉农商品名称及类别标注

Tab. 4 The extracted agriculture-related product name and category label

序号	涉农商品名称	标注类别
1	海南香蕉	种植业类
2	塑料花盆	园艺业类
3	鸽子蛋	养殖业类
4	葡萄叶面肥	化肥类
5	行军铁锹	农业用具类
6	土豆地瓜花生收获机	农业机械类
7	海蜇丝	鲜活水产品加工制品类

表 5 CRF 与 HMM、MEMM (词 + 词性 + 地理特征) 的比较 %

Tab. 5 Comparative results of CRF, HMM and MEMM based on word, part of speech and geographical attributes %

类别	HMM			MEMM			CRF		
	P	R	F	P	R	F	P	R	F
种植业类	48.53	49.82	49.17	70.32	64.95	67.53	78.44	71.85	75.00
园艺业类	56.65	25.30	34.98	87.53	32.87	47.79	92.36	36.52	52.34
养殖业类	71.41	20.79	32.20	88.22	30.84	45.70	94.74	34.50	50.58
化肥类	78.56	43.42	56.07	92.70	47.21	62.56	96.12	52.66	68.04
农业用具类	64.58	53.12	58.29	76.11	60.58	67.46	83.68	70.37	76.45
农业机械类	44.05	31.60	36.80	69.45	42.79	52.95	83.37	68.65	75.30
鲜活水产品加工制品类	72.36	31.23	43.63	81.57	37.94	51.79	86.96	46.51	60.61
总体值	62.30	32.59	42.79	79.41	43.01	55.80	84.40	57.36	68.30

实验显示,CRF 的性能优于 HMM、MEMM。主要由于 HMM 为产生式模型,具有严格的输出独立性假设,不能充分利用上下文多特征信息,对于由 3 个及以下的词组成的涉农商品名称抽取效率差,如将“玉米小麦收割机”抽取为两个涉农商品名称“玉米”“小麦收割机”,容易出现类别识别错误;MEMM 克服了 HMM 的缺点,但使用每一个状态的指数模型来计算给定前一个状态下当前状态的条件概率,容易陷入局部最优,存在标注偏置的问题;而 CRF 在所有特征上进行全局归一化,能得到全局最优解,避免了 MEMM 缺点。因此文中选取 CRF 抽取涉农商品名称与类别标注是有效的。

3.4 基于本体与 CRF 的农产品供求趋势分析

涉农商品名称及类别标注的有效抽取,不仅有助于促进农业供求交易的智能对接,而且可用于农业供求趋势分析,了解市场动态。利用中国搜农网供求搜索栏目抓取的网站数据作为原始数据,通过第 4 组实验的方法,抽取涉农商品名称及分类,图 5(a)~5(d)表示 2016 年 5 月 3 日到 6 月 6 日连续 5 周内的供应求购趋势。由图 5 可知,四川省种植业类的商品求购量高于湖北省,两省在第 5 周都有大幅度的上升;河北省农业机械类的农产品周供应量较为平稳,而山东省在第 5 周时上升幅度大,达到 591;山东省养殖业类的供应量远高于江苏省,而园艺业的供应量则低于江苏省,反映出各地区农产品供应的差异性。根据这些供应求购趋势信息,买卖双方可根据地理位置,来选择适合的产品,更好地促成实时交易,如山东省的客户想购买玉米剥壳机,通过供应趋势图,则可就近选择较好的相关产品,给购买者提供方便。

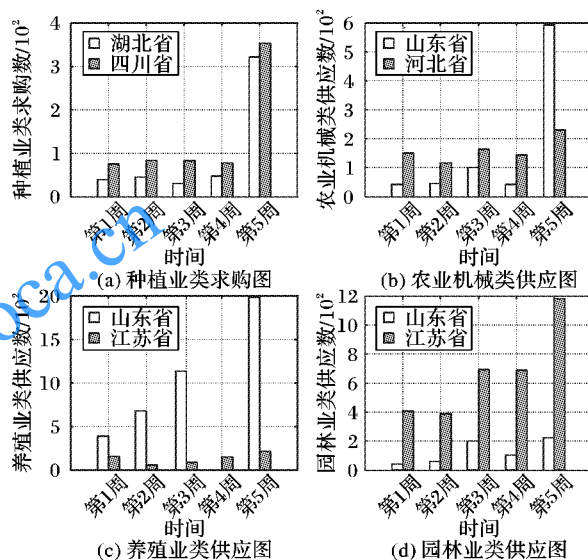


图 5 各类供应求购趋势

Fig. 5 Supply and demand trends

4 结语

本文基于农业本体与条件随机场 CRF 相结合抽取涉农商品名称实现类别标记,在词、词性和地理属性特征基础上,自动添加词所对应的农业本体概念特征,对实例名称进行不同程度的泛化,赋予词以语义和概念知识。通过实验,在一定范围内,泛化程度越高,CRF 模型表现出的“推理”能力越强,可有效地抽取测试语料中首次出现的涉农商品名称并分类,在准确率高的前提下,也保证了召回率,大量减少训练语料,降低人工工作量,与 HMM、MEMM 比较,体现出 CRF 的性能更优,并将此方法用于农产品供求趋势分析,可了解市场动态。原始语料以及分词工具的选取直接关系到 CRF 模型的性能,在今后的研究工作中,一方面将进行分词方法改进,选取不同的训练语料,进行 CRF 涉农商品名称抽取研究,进一步提升准确率和召回率,另一方面尝试从降低算法的复杂度入手,提高效率。

参考文献 (References)

[1] 于连军. 基于互联网+的农业电子商务发展模式的研究[J]. 农



- 业网络信息, 2015(11): 19 - 21. (YU L J. Research on the development model of agricultural E-commerce based on Internet + [J]. Agriculture Network Information, 2015(11): 19 - 21.)
- [2] LI L S, DAND Y Z, ZHANG J, et al. Domain term extraction based on conditional random fields combined with active learning strategy [J]. Journal of Information & Computational Science, 2012, 9(7): 1931 - 1940.
- [3] GUAN A Q, WANG Y B, YANG L F. Automatic term extraction for Chinese opera domain ontology [C]// Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE, 2015: 1372 - 1376.
- [4] 宗成庆. 统计自然语言处理[M]. 2 版. 北京: 清华大学出版社, 2013: 110 - 128. (ZONG C Q. Statistical Natural Language Processing [M]. 2nd ed. Beijing: Tsinghua University Press, 2013: 110 - 128.)
- [5] WALLACH H M. Conditional random fields: an introduction, technical report MS-CIS-04-21 [R]. Philadelphia, PA: University of Pennsylvania, 2004: 262 - 272.
- [6] FU W J, LI L. A method and application of automatic term extraction using conditional random fields [C]// Proceedings of the 2009 International Conference on Natural Language Processing and Knowledge Engineering. Piscataway, NJ: IEEE, 2009: 1 - 5.
- [7] ZHANG C Z, WANG H L, LIU Y, et al. Automatic keyword extraction from documents using conditional random fields [J]. Journal of Computational Information System, 2008, 4(3): 1169 - 1180.
- [8] 孟洪宇. 基于条件随机场的《伤寒论》中术语自动识别[D]. 北京: 北京中医药大学, 2014: 41 - 48. (MENG H Y. Automatic identification of TCM terminology in Shanghan Lun based on conditional random field [D]. Beijing: Beijing University of Chinese Medicine, 2014: 41 - 48.)
- [9] ZHAN Q, WANG C H. A Hybrid strategy for Chinese domain-specific terminology extraction [C]// Proceedings of the 11th International Conference on Semantics, Knowledge and Grids. Piscataway, NJ: IEEE, 2015: 217 - 221.
- [10] 王春雨. 基于 CRF 的农业命名实体识别研究[D]. 保定: 河北农业大学, 2014: 19 - 23. (WANG C Y. Study on recognition of Chinese agricultural named entity with CRF [D]. Baoding: Agricultural University of Hebei, 2014: 19 - 23.)
- [11] CAO Y S, WANG J, LI L. Word-level information extraction from science and technology announcements corpus based on CRF [C]// Proceedings of the 2nd IEEE International Conference on Cloud Computing and Intelligence Systems. Piscataway, NJ: IEEE, 2012: 1529 - 1533.
- [12] IZUMI M, MIURA T, SHIOYA I. Estimating the date of blog authors by CRF [C]// Proceedings of the 2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. Piscataway, NJ: IEEE, 2007: 249 - 252.
- [13] GRUBER T R. A translation approach to portable ontology specifications [J]. Knowledge Acquisition, 1993, 5(2): 199 - 220.
- [14] 李传席. 基于本体的自适应 Web 信息抽取方法研究[D]. 合肥: 中国科学技术大学, 2012: 15 - 17. (LI C X. Adaptive Web information extraction method research based on ontology [D]. Hefei: University of Science and Technology of China, 2012: 15 - 17.)
- [15] LIU X G, DUAN X H, ZHANG H Y. Application of ontology in classification of agricultural information [C]// Proceedings of the 2012 IEEE Symposium on Robotics and Applications. Piscataway, NJ: IEEE, 2012: 451 - 454.
- [16] 周晶, 吴军华, 陈佳, 等. 基于条件随机场 CRF 模型的文本信息抽取[J]. 计算机工程与设计, 2008, 29(23): 6094 - 6097. (ZHOU J, WU J H, CHEN J, et al. Using conditional random fields model for text information extraction [J]. Computer Engineering and Design, 2008, 29(23): 6094 - 6097.)
- [17] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 282 - 289.
- [18] Sunfox66. 条件随机场详解[EB/OL]. (2015-10-25) [2016-01-17]. <http://wenku.baidu.com/view/bbd57f82fc4ffe473268ab59.html>. (Sunfox66. Conditional random field introduction [EB/OL]. (2015-10-25) [2016-01-17]. <http://wenku.baidu.com/view/bbd57f82fc4ffe473268ab59.html>.)
- [19] LIU D, NOCEDAL J. On the limited memory BFGS method for large scale optimization [J]. Mathematical Programming, 1989, 45(45): 503 - 528.
- [20] TAKE K. CRF++ toolkit [EB/OL]. (2014-10-15) [2016-01-15]. <http://download.csdn.net/detail/linson3344/8039087>.
- [21] HANKCS. Han language processing [EB/OL]. (2015-03-27) [2016-01-28]. <http://www.hankcs.com/nlp/hanlp.html>.
- [22] 施水才, 王锴, 韩艳铎, 等. 基于条件随机场的领域术语识别研究[J]. 计算机工程与应用, 2013, 49(10): 147 - 149. (SHI S C, WANG K, HAN Y H, et al. Terminology recognition based on conditional random fields [J]. Computer Engineering and Applications, 2013, 49(10): 147 - 149.)
- [23] 贾美英, 杨炳儒, 郑德权, 等. 采用 CRF 技术的军事情报术语自动抽取研究[J]. 计算机工程与应用, 2009, 45(32): 126 - 129. (JIA M Y, YANG B R, ZHENG D Q, et al. Research on automatic military intelligence term extraction using CRF model [J]. Computer Engineering and Applications, 2009, 45(32): 126 - 129.)
- [24] MCCALLUM A K. MALLET: a machine learning for language toolkit [EB/OL]. (2002-02-28) [2016-02-25]. <http://mallet.cs.umass.edu>.

This work is partially supported by the National Science and Technology Support Program (2013BAD15B03), Chinese Academy of Sciences Key Deployment Project (Y622A21291), the Scientific and Technological Project of Anhui Province (1401032010).

HUANG Nian'e, born in 1991, M. S. candidate. Her research interests include information extraction, vertical search engine.

HUANG He, born in 1980, Ph. D., associate professor. His research interests include agriculture big data, agricultural intelligent system.

WANG Rujing, born in 1964, Ph. D., professor. His research interests include knowledge representation and visualization, knowledge acquisition.