doi:10.6041/j.issn.1000-1298.2017.S0.029

基于条件随机场的农作物病虫害及农药命名实体识别

李　想[1]　魏小红[1]　贾　璐[1]　陈　昕[1]　刘　磊[2]　张彦娥[1]

(1.中国农业大学信息与电气工程学院,北京 100083;2.山东老刀网络科技有限公司,潍坊 261000)

摘要:互联网农技问答平台现仅依靠人工提供答题服务,响应速度慢,回答质量难以保证。实现智能农技问题解答,构建农技知识库,需要从现有问答数据提取“农作物-病虫害-农药”命名实体三元组。现有对农业中文命名实体识别的研究较少,且准确率较低。根据农作物、病虫害及农药命名实体的特点,针对农技问答数据,提出基于条件随机场的农作物、病虫害及农药命名实体的识别方法。对数据集进行格式整理及自动分词,并对分词后的语料,针对是否包含特定界定词、是否含特定偏旁部首、是否是数量词、是否是特定左右指界词及词性等特征进行自动标注。利用标注后的数据训练 CRF 模型,可以对语料进行分类,包括判断语料是否属于农作物、病虫害、农药 3 类命名实体并识别该语料在复合命名实体中的位置,从而实现了对 3 类命名实体的识别,由此可自动构建关联三元组。通过试验选择特征组合和调整上下文窗口大小,提高了本方法的识别准确度,降低了模型训练时间,对农作物、病虫害、农药命名实体识别的准确度分别达 97.72%、87.63%、98.05%,比现有方法有显著提高。

关键词:病虫害;农药;知识库;命名实体识别;条件随机场

中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1000-1298(2017)S0-0178-08

Recognition of Crops, Diseases and Pesticides Named Entities in Chinese Based on Conditional Random Fields

LI Xiang[1]　WEI Xiaohong[1]　JIA Lu[1]　CHEN Xin[1]　LIU Lei[2]　ZHANG Yan'e[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Shandong Laodao Network Technology Co.*, *Ltd.*, *Weifang* 261000, *China*)

Abstract: On internet agricultural technology platform, thousands of new questions are waiting to be answered by experts every day. It is generally doubted because of slowly response time and uncertain quality of the manual services. An intelligent response system based on agricultural technology knowledge base can help to answer some questions automatically. To build the knowledge base, it is necessary to recognize triples of “crop-disease-pesticide” named entities from mass of existing questions and answers data. However, fewer studies are reported on recognition methods for named entities of diseases and pesticides in Chinese, and accuracies of those for named entities of crops are low. Thus, a recognition method based on conditional random fields (CRF) was proposed, which recognized crops, diseases, and pesticides named entities from agricultural technology questions and answers data. In the method, question and answer texts was formatted and split to pieces of corpus. Each corpus piece was automatically annotated with several features, including whether it contained characteristic Chinese characters and characteristic radicals, whether it was numeral, whether it was the left or right bound of a compound word, and part of speech. A CRF model was trained with these annotated texts to classify pieces of corpus, including judging whether they were parts of crop, disease, or pesticide named entities and recognizing positions in named entities. With the trained model, three types of named entities could be accurately recognized and triples could be associated automatically. Recognition accuracies and time cost of model training were optimized by choosing input feature combinations and adjusting sizes of context windows in experiments. Accuracies of recognizing crops, diseases, and pesticides of this method were 97.72%, 87.63% and 98.05% respectively, which were significantly higher than existing methods.

Key words: disease; pesticide; knowledge base; named entities recognition; conditional random fields

收稿日期:2017-07-15　修回日期:2017-11-21

基金项目:国家自然科学基金项目(61502500)、北京市自然科学基金项目(4164090)和中央高校基本科研业务费专项资金项目(2017QC077)

作者简介:李想(1983—),男,讲师,博士,主要从事农业大数据挖掘和实时复杂事件处理研究,E-mail:cqlixiang@cau.edu.cn

通信作者:陈昕(1974—),男,副教授,主要从事农业大数据挖掘算法研究,E-mail:chxin@cau.edu.cn

引言

随着“互联网 +”的快速发展，全国性的农资电商行业呈爆发型增长趋势，已有超过 30 000 家的涉农电商平台[1]，农技咨询服务成为大多数农资电商平台的标准配置服务。农户可通过手机 APP 描述农作物生长情况和病情，及时获得平台专家的解答。面对呈几何级数增长的问题数据，现有平台还依赖于领域专家人工在线回答，时效性和专业性成为农技咨询服务水平的短板。如何构建“农作物-病虫害-农药”知识库，从而支撑实现农技问题解答的智能化，成为一个重要问题。

自动问答系统知识库构建的基础是命名实体识别（Named entity recognition，NER），即识别文本中出现的专有名词，并建立它们的联系[2]。随着农技咨询服务数据的累积，从问题和回答数据中智能识别农作物、病虫害和农药等命名实体，成为构建知识库的一种简便易行的方式。

NER 已在很多领域进行了广泛应用，例如：Web 领域[3]、产品识别[4]、微博文本[5]、医学病历[6]、军事文本[7]等。在农业领域，方莹[8]提出基于层叠条件随机场（Conditional random field，CRF）模型的农业领域命名识别方法，针对互联网网页数据对农作物命名实体进行识别；王春雨等[9]对网站和刊物中获取的信息，采用 CRF 模型识别农作物、家禽等命名实体；黄念娥等[10]抓取了阿里巴巴网的标题数据，采用本体与 CRF 结合的方法对园林业、养殖业、化肥、农业用具、农业机械等命名实体进行了识别。但据研究发现，在农作物命名实体识别方面的研究准确度不高，在病虫害及农药的命名实体识别方面的研究成果还相对缺乏。

现有的 NER 方法主要分为基于规则的方法和基于统计的方法，基于规则的 NER 方法过度依赖词典和规则库，对于歧义词和未登录词的识别能力较低[11]；而基于统计的方法快速高效，具有较好的移植性，常见的方法主要包括隐马尔克夫模型（HMM）[12]、最大熵隐马模型（MEMM）[13]和条件随机场模型（CRF）[14]。经过比较，CRF 在易用性、稳定性和准确性等综合方面的表现最好[15]。CRF 既具有判别式模型的优点，又有产生式模型考虑到上下文标价间的转移概率[16]，以序列化形式进行全局参数优化和解码的特点，克服了 HMM 输出独立性假设的问题，解决了 MEMM 难以避免的标记偏见问题[17]。

本文基于 CRF 模型，研究农作物、病虫害及农药命名实体识别方法。利用农管家 APP 抓取的 7 万余条问答数据，通过分析农作物、病虫害及农药名称的词性、偏傍部首、左右指界词、附近数量词等特征，训练 CRF 模型，建立上述特征与命名实体类别和词位间的关联关系，从而识别命名实体。

1　数据集

本文分析数据集来自于山东老刀网络科技有限公司的农管家 APP 产品中 2016 年每月 1 日和 15 日的农户及专家的问答数据，包括：7 870 条农户描述农作物病害的问题数据、66 559 条平台专家的解答数据。对比农药目录等数据集，农业问答数据具有时效性的优势，通过挖掘“农作物-病虫害-农药”信息并关联地址等信息，可以准确地为农民提供近期某地某种作物病虫害常用药建议。

每条问题数据的数据项包括：问题 ID、问题内容、省份、回答数等，涵盖了蔬菜、粮食作物、果树等农作物。例如，问题 ID 为 6114 ~ 6116 的数据如表 1 所示。

每条回答数据的数据项包括：回复 ID、问题 ID、回复内容等。回答与问题存在对应关系，通过问题 ID 可以将回答与问题关联起来。问题 ID 为 6114 对应的全部回答数据如表 2 所示。

表 1　问题数据示例

Tab. 1　Examples of questions

问题 ID	问题内容	省份	回复数
6114	请教专家这是什么病？黄瓜叶子一摸就掉	辽宁省	5
6115	羊毛脂膏能用于果树发枝吗？	山西省	2
6116	核桃树能施氯化钾吗？	山西省	5

表 2　回答数据示例

Tab. 2　Examples of answers

回复 ID	问题 ID	回复内容
25265	6114	时间紧，先回答药，腈菌唑 + 叶面肥，喷透
25272	6114	不是药害是细菌性叶斑病简单的用农莲加喹啉酮连打两次降低湿度
25283	6114	看不清，好像有点霜霉病。
25305	6114	你用丁子香酚加白粉病的在配合霜霉病的，只要你的药物好一次搞定
26627	6114	是细菌性叶枯病。

2　基于 CRF 的命名实体识别方法

2.1　CRF 模型

本文提出一种基于 CRF 的命名实体识别方法，用于从农技问答数据中，识别农作物、病虫害及农药命名实体，并建立“农作物-病虫害-农药”三元知识集。

CRF 模型[18]是用来标注和划分序列结构数据的概率化的无向图模型,被广泛用于自然语言处理方面的机器学习。如图 1 所示,对于给定的输入文本序列 x,通过分词及特征标注得到(x_1,x_2,…,x_{v-1},x_v)语料序列,通过训练得到的模型参数,预测需要的语料标注组合 y 的条件概率。

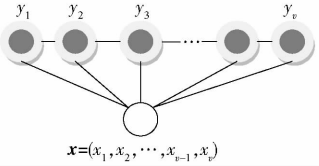

图 1 CRF 模型结构

Fig. 1 CRF model structure

描述 CRF 模型的条件概率 $P(y|x)$ 计算公式为

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_{j}w_jf_j(y_{i-1},y_i,x,i)\right) \tag{1}$$

其中 $Z(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}\sum_{j}w_jf_j(y_{i-1},y_i,x,i)\right)$

式中 x——观测语料序列

y——需要的语料标注序列

w_j——特征函数的权重

$f_j(y_{i-1},y_i,x,i)$——对应 x 标记位于 i 和 $i-1$ 的特征函数

训练过程是已知一个对输入和输出均进行过标注的语料集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 和训练样本中(x,y)的经验概率分布 $\tilde{p}(x,y)$,求模型参数。测试过程是给定 $P(y|x)$ 和观察序列 x,求条件概率最大的标注序列 y。

本文提出的方法,首先对农技问答数据自动进行格式预处理并分词形成语料序列。

对分词后的语料观测集 x,选择是否包含特定界定词、是否含特定偏旁部首、是否是数量词、是否是特定左右指界词及词性等 6 个特征进行自动标注。因为语言呈现链式序列的特点,因此,对 x 的标注不仅考虑当前语料特征,还考虑了语料的上下文特征,以提高识别准确率。

输出特征 y 包括两类:类别,即一个语料是否属于某个农作物、病虫害或农药命名实体;词位,即该语料在复合词中的位置。

通过训练完成的 CRF 模型经过测试,可用于准确识别问答文本中的农作物、病虫害、农药 3 类命名实体。具体过程如图 2 所示。

2.2 数据格式整理及分词

问答数据在数据格式上有不规范情况,不能直接用于处理,需要对数据格式进行整理:

(1)在训练工具中,“[ ]”是标注复合词专用符号,因此,将语料中原有的“[ ]”非复合词的改为“( )”,如:“[访问验证码是:929 307 请妥善保管]”改为“(访问验证码是:929 307 请妥善保管)”。

(2)“/”是标注词性专用符号,因此语料中原有“/”改为它代表的实际意义(“或”、“每”)。如:“混合复配浓度 1 ~ 2 毫克/升的细胞分裂素即可”,将句中的“/升”改为“每升”。

(3)去掉无效空格以及特殊符号。如:“不是药害是细菌性叶斑病简单的用农莲加喹啉酮连打两次降低湿度”改为“不是药害,是细菌性叶斑病,简单的用农莲加喹啉酮连打两次,降低湿度”。

对句子进行精确分词,将句子切割为语料,如“不是药害,是细菌性叶斑病,简单的用农莲加喹啉酮连打两次,降低湿度。”经过调查发现,现有分词工具较为成熟,这里不再赘述。

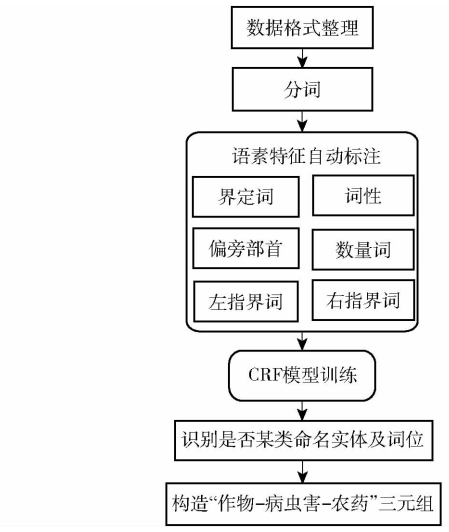

图 2 农作物、病虫害、农药命名实体识别方法流程

Fig. 2 Flow chart of the recognition method for named entities of crops, diseases and pesticides in Chinese

2.3 观测语料序列的特征自动标注

本文选取界定词(Word)、词性(Part-of-speech, POS)、左右指界词(Left bound + Right bound, LB + RB)、偏旁部首(Radical, Rad)、数量词(Numeral, Num)作为特征对观测语料序列中每个语料及其上下文进行标注。这些特征,具有较好的区分度,且容易通过程序实现自动标注。

首先,界定词(Word):农作物常包含“瓜、豆、菜”等特征语素,病虫害常包含“病、虫”语素,农药常包含“酚、脂、酯、胺”等语素。这些特征语素成为识别作物、病虫害、农药命名实体的重要依据。但由于会出现“什么病”、“啥毛病”,所以限制从第 2 个词符位置开始查找出现特征词语,且排除“毛病”、“生病”等类似这种无效的词语。

第二,词性 POS:农作物为名词,而病虫害及农

药除此以外，还有复合名词，还包括部分动词、形容词等。因此，需要对词性进行标注。如“不/d 是/vshi 药/n 害/ng，/wd 是/vshi 细菌/n 性/ng 叶斑病/n ，/wd 简单/a 的/ude1 用/v 农/ng 莲/ng 加/v 喹啉/n 酮/w 连/d 打/v 两/m 次/qv，/wd 降低/v 湿度/n”并去除停用词，如“谢谢”、“吗”等。词性含义见 ICTCLAS 汉语词性标注集。

第三，指界词 LB、RB：农作物、病虫害、农药等命名实体常与特定的谓词、动词和副词一起出现，某些词出现在命名实体左边，称为左指界词，出现在右边的称为右指界词。常见的指界词如表 3 所示。句首位置也是一种常用的左指界，标点符号也可作为右指界。

表 3　农作物、病虫害、农药的指界词

Tab.3　Left bound or right bound of crops, diseases and pesticides

类型	左指界词	右指界词
农作物	种植、看看、句首位置	名词
病虫害	是、得了、预防、注意	句末位置
农药	喷、施、用、打、加	打、加、试试、防治

第四，偏旁部首 Rad：农作物、病虫害、农药的名称常包含有特定偏傍部首的字，可以作为识别命名实体的特征。

农作物方面，农业问答数据集，主要集中于包括瓜菜、根菜、叶菜等蔬菜，苹果、香蕉等水果及土豆、大豆等大田作物。可用“木”、“艹”、“豆”偏旁对农作物识别。

病虫害是病害和虫害的并称，病害可以通过后缀特征词“病”进行识别；常见虫害有：稻飞虱、玉米螟、棉铃虫、棉蚜、麦蚜、麦红蜘蛛、蝗虫等，可用偏旁部首“虫”作为虫害识别的特征。

农药方面，目前使用最多的一类农药是有机合成农药，主要包括有机磷酸酯类、氨基甲酸酯类和拟除虫菊酯类 3 大类[19]，这类农药名称的偏旁部首主要是“气”、“石”、“酉”、“月”等。

第五，数量词 Num：数量词即数量形容词，是表示数、量或程度的形容词。通过观察语料发现数量词和农药同现的概率很大，可通过判断附近词是否为数量词对农药识别。

为了对 CRF 模型进行训练，训练语料集除了对观测语料序列标注外，还需要标注输出特征项。这里选择的输出特征为语料的类别及其在命名实体中的词位。这两个特征组合起来，能够识别出复合命名实体。

病虫害及农药多由 3 个及以上语料组成。词位是某个语料在复合词中的位置。选取五词位标注法，标注符号集为（B，I，E，S，O），其含义如表 4 所示。类别是指“农作物”、“病虫害”、“农药”3 类。符号集为（CRO ，DIS，PES），其含义如表 5 所示。如名称为“娃娃菜”农作物，标注为“娃娃（B-CRO）菜（E-CRO）”；名称为“细菌性叶斑病”病虫害，标注为“细菌（B-DIS）性（I-DIS）叶斑病（E-DIS）”；名称为“丁子香酚”农药，标注为“丁子（B-PES）香（I-PES）酚（E-PES）”。

表 4　命名实体内部词位标注

Tab.4　Annotation of positions in named entities

符号	含义
B	当前词为命名实体的首部
I	当前词为命名实体的内部
E	当前词为命名实体的尾部
S	当前词是命名实体
O	当前词不是需要的命名实体

表 5　农作物、病虫害、农药的命名实体类别标注

Tab.5　Annotation of named entity classes of crops, diseases and pesticides

符号	含义
CRO	农作物
DIS	病虫害
PES	农药

2.4　语料序列生成

通过分词及特征自动标注，能够生成语料观察序列和输出特征序列。表 6 是一个示例，ID 为 25 272 的回答数据为“不是药害是细菌性叶斑病简单的用农莲加喹啉酮连打两次降低湿度”。经过数据格式整理、分词及标注后，可表示一张标注表。这里用“T”表示该语料符合标注特征，“F”表示该语料不符合标注特征。用后缀“CRO”、“DIS”、“PES”分别表示农作物、病虫害、农药。如“叶斑病”语料的偏旁部首标注项标注为“T-DIS”，表示符合“疒、虫”偏旁部首特征，并且是病虫害类语料。测试语料也需要自动标注，标注方法与训练语料类似，区别在于最后一列词位及类别标注结果由训练完成的 CRF 模型预测得到。

考虑到农作物、病虫害、农药命名实体较为复杂，往往被分为多个语料，因此，在识别这 3 类命名实体时，可以将其前数个语料和后数个语料纳入识别范围，用于支持对该语料的识别。前后纳入识别范围的语料数称为上下文窗口。

使用语料特征序列和输出特征序列对 CRF 模型进行训练，优化模型参数。训练完成的模型即可用于农作物、病虫害、农药命名实体识别。

表 6 语料的分词及标注示例

Tab. 6 Example of corpus segmentation and annotation

语料	语料观测序列标注						出序列标注
	界定词 Word	词性 POS	数量词 Num	偏旁部首 Rad	左指界词 LB	右指界词 RB	(词位及类别)
不	F	d	F	F	T-CRO	F	O
是	F	vshi	F	F	T-DIS	F	O
药	F	n	F	T-CRO	F	T-CRO	O
害	F	ng	F	F	F	F	O
是	F	vshi	F	F	T-DIS	F	O
细菌	F	n	F	T-CRO	F	F	B-DIS
性	F	ng	F	F	F	T-CRO	I-DIS
叶斑病	T	n	F	T-DIS	F	T-CRO	E-DIS
简单	F	a	F	F	F	F	O
的	F	ude1	F	F	F	F	O
用	F	v	F	F	T-PES	F	O
农	F	ng	F	F	F	F	B-PES
莲	F	ng	F	T-CRO	F	F	E-PES
加	F	v	F	F	T-PES	T-PES	O
喹啉	F	n	F	F	F	T-CRO	B-PES
酮	F	w	F	T-PES	F	F	E-PES
连	F	d	F	F	F	F	O
打	F	v	F	F	T-PES	T-PES	O
两	F	m	T	F	F	F	O
次	F	qv	F	F	F	F	O
降低	F	v	F	F	F	F	O
湿度	F	n	F	F	F	T-DIS	O

2.5 “农作物-病虫害-农药”三元组构建

通过对命名实体的准确识别,可以在问答数据中构建关联命名实体集合,为进一步与图片分析进行关联,以及构建“农作物-病虫害-农药”知识库提供了支撑。如表 1 中 ID 为 6114 的问题和表 2 中 ID 为 25272 的答案可以构成关联的命名实体集合“黄瓜-病虫害-农莲、喹啉酮”。

3 试验及结果分析

选用 ICTCLAS 作为分词工具。CRF + + 0.58 是一个可用于分词/连续数据标注的条件随机场工具,对农作物、病虫害、农药的命名实体识别也是需要生成相应的词位标注及类别标注,因此,本文采用 CRF + + 0.58 工具包进行训练与测试。

设计 3 组试验,分别考察用单个特征、不考虑上下文的特征组合、考虑上下文的特征组合进行标注时 NER 方法的准确率。并集中探讨了训练样本数量、标注特征数量和上下文窗口大小对本文 NER 方法模型训练时间的影响。

在现有研究中,常以正确率 P、召回率 R 和 F 值作为评测中文 NER 系统识别准确率的指标[20],公式为

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{2}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{3}$$

$$F = \frac{(1 + \beta^2) PR}{(\beta^2 P) + R} \times 100\% \tag{4}$$

式中 T_P——真正例:预测结果和实际结果都为正例的样例数

F_P——假正例:预测结果为正例、实际结果为反例的样例数

F_N——假反例:预测结果为反例、实际结果为正例的样例数

β——调节正确率和召回率的比重,通常 $\beta = 1$

3.1 单个特征标注

本试验考察仅采用 Word、POS、Num、Rad、LB、RB 6 类特征中的一项分别进行标注时,本文 NER 方法的准确率。数据集采用 7 870 条问题数据和 66 559 条回答数据的全集。其中随机划分 75% 用作模型训练,25% 用作测试模型。

试验结果如表 7 所示。仅采用单个特征时,除了 Word 特征对 3 类识别保持较高准确率外,采用其他特征对命名实体识别准确率均较低,尤其特征

选择为 LB、RB 时，由于文本内容语法本身结构不完整，识别效果很差。因此，需要多种特征联合使用，提高识别准确率。

表 7　仅用单个特征标注时的农作物、病虫害、农药命名实体识别准确率比较

Tab.7　Accuracies of NER for crops, diseases and pesticides under different individual features　%

特征选择	类型	*P*	*R*	*F*
Word	农作物	64.84	63.37	64.09
	病虫害	58.38	95.92	72.58
	农药	82.69	84.84	83.75
POS	农作物	10.23	33.33	15.66
	病虫害	0	0	0
	农药	68.45	15.76	25.62
Num	农药	2.913	24.90	5.210
Rad	农作物	19.23	31.13	17.83
	病虫害	24.52	15.17	18.74
	农药	24.98	15.41	19.06
LB	农作物	16.67	33.33	22.23
	病虫害	0	0	0
	农药	38.71	20.67	26.95
RB	农作物	0	0	0
	病虫害	0	0	0
	农药	9.71	25.02	14.01

3.2　不考虑上下文时特征组合标注

采用不同的特征组合进行标注，考察农作物、病虫害和农药命名实体识别的准确度。这里标注仅对当前语料，不考虑上下文。数据集采用问题和回答数据全集。

试验结果如表 8 所示，标注并非越多越好。对于农作物，效果最好的特征选择为“Word + LB + Rad + POS”，*F* 值达 97.72%。而增加 RB 时，*F* 值并未上升，因此为了提高识别效率，可以不考虑 RB。

表 8　在不同特征组合时的农作物、病虫害、农药命名实体识别准确率比较

Tab.8　Accuracies of NER for crops, diseases and pesticides under different feature combinations　%

类型	特征选择	*P*	*R*	*F*
农作物	Word + LB	99.95	91.94	95.78
	Word + LB + Rad	98.73	96.02	97.31
	Word + LB + Rad + POS	99.45	96.16	97.72
	Word + LB + Rad + POS + RB	99.45	96.16	97.72
病虫害	Word + LB	96.44	68.76	80.28
	Word + LB + POS	96.08	68.99	80.31
	Word + LB + POS + Rad	96.33	69.50	80.74
	Word + LB + POS + Rad + RB	96.34	69.43	80.70
农药	Word + LB	95.94	87.83	91.71
	Word + LB + POS	96.00	88.28	91.98
	Word + LB + POS + Rad	96.03	88.60	92.16
	Word + LB + POS + Rad + RB	96.04	88.56	92.15
	Word + LB + POS + Rad + Num	96.05	88.56	92.15

对于病虫害，特征选择为“Word + LB + POS + Rad”时，效果最好，*F* 值达 80.74%，高于特征最多的组合“Word + LB + POS + Rad + RB”。

对于农药，特征选择为“Word + LB + POS + Rad”时，*F* 值达 92.16%，而特征选择最多的组合“Word + LB + POS + Rad + Num”*F* 值为 92.15%。

3.3　考虑上下文时特征组合标注

不考虑上下文时，即便选择多特征组合，对病虫害和农药的识别准确度还是较低，分别只有 80% 和 90% 左右。因此，本试验加入对上下文联合标注的考虑。数据集采用问题和回答数据全集。其中随机划分 75% 用作模型训练，25% 用作测试模型。对比如下 4 种上下文窗口：窗口 A，仅当前词；窗口 B，当前词 + 上文一个词；窗口 C，当前词 + 下文一个词；窗口 D，当前词 + 上下文各一个词。

试验结果如表 9 所示。可以看到，窗口 C 即考虑当前词和下文一个词时，表征识别准确率的 *F* 值最高，对病虫害命名实体识别的 *F* 值提升至 87.63%，对农药提升至 98.05%。

表 9　不同上下文窗口时病虫害和农药命名实体识别准确率

Tab.9　Accuracies of NER for diseases and pesticides under different sizes of context windows　%

窗口	类型	*P*	*R*	*F*
A	病虫害	96.33	69.50	80.74
	农药	96.05	88.56	92.15
B	病虫害	89.76	71.11	79.35
	农药	99.00	95.29	97.11
C	病虫害	90.89	84.59	87.63
	农药	99.01	97.10	98.05
D	病虫害	78.62	82.55	80.54
	农药	98.42	96.02	97.21

3.4　模型训练效率

本节考虑特征数量、上下文窗口大小对模型训练性能的影响。测试机环境配置为处理器 Intel(R) Core(TM) i5 - 3470 3.20 GHz 3.20 GHz、运行内存 8 GB DDR3。

首先，考察特征组合中的特征数量对模型训练时间的影响。这里，使用数据集采用 7 870 条问题数据和 66 559 条回答数据的全集，数据量为 27 245 KB，随机采样 75% 样本进行训练。试验结果如图 3 所示，模型训练时间随特征数量的增加而上升，但增速较缓。对比仅采用 1 个特征标注，采用 5 个特征标注时，对病虫害和农药命名实体识别的时间分别增加了 147.11% 和 71.07%。

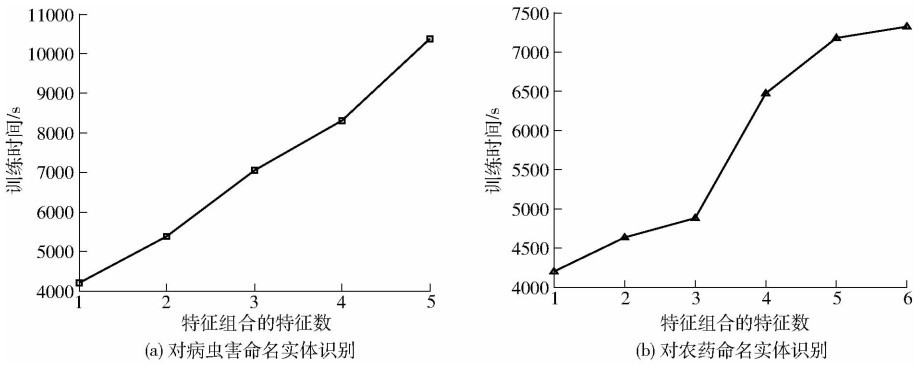

图 3 特征组合中特征数量对识别病虫害、农药命名实体性能的影响

Fig. 3 Time cost of model training of NER for diseases and pesticidesunder different numbers of features in feature combinations

其次,考察上下文窗口大小对模型训练效率的影响。上下文窗口大小表示窗口中上下文语料的数量。由于在窗口大小大于 5 时,训练时间过长,因此仅使用了 7 870 条问题数据进行训练,观察变化规律,数据量为 2 663 KB。

试验结果如图 4 所示。识别准确度均随上下文窗口扩大而增加。窗口大小为 1 时,表示仅考虑当前词,未考虑上下文,识别效果较差。窗口大小小于 7 时,F 值呈上升趋势,当窗口处于[1,3]时,F 值上升较为明显,大于 3 时上升趋势变缓。当上下文窗口大于 7 时,由于特征模板复杂导致训练时过度拟合,F 值下降。当上下文窗口等于 7 时,对于病虫害及农药的识别均取得最高的识别准确度。

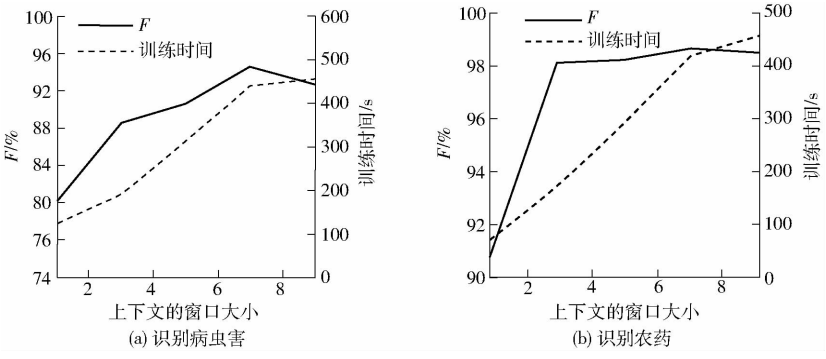

图 4 上下文窗口大小对识别病虫害、农药命名实体性能的影响

Fig. 4 Accuracies and time cost of model training of NER for diseases and pesticides under different sizes of context windows

识别准确度均随上下文窗口扩大而急剧增加。对比窗口大小为 1 时,在窗口大小为 3 和 7 时,对病虫害的识别模型训练时间分别增长了约 57.4% 和 265.03%,对农药的识别模型训练时间分别增长了约 141.26% 和 497.2%。如果考虑 7870 条问题数据和 66 559 条回答数据全集时,当上下文窗口为 9 时,训练时间已长达 48 h 以上,不具有实用性。

因此,平衡识别的准确度和性能,推荐上下文窗口大小采用 3,适当增加特征标注类型以提高识别准确度。而在性能较优的机器上,上下文窗口大小推荐为 7。

3.4 与现有方法对比

通过查阅文献,方莹[8]提出的基于层叠条件随机场模型的农作物命名实体方法,F 值仅达 92.70%;王春雨等[9]提出的基于条件随机场对农作物命名实体的识别,F 值在 93% 左右;而本文对农作物命名实体识别的 F 值高于两者,达到 97.72%。

4 结论

(1)提出了基于条件随机场的农作物、病虫害及农药命名实体的识别方法。方法考虑了界定词、词性、左右指界词、偏旁部首、数量词作为特征,以及语料间的上下文关系。

(2)试验表明,方法对农作物、病虫害、农药识别准确度的测度 F 值达 97.72%、87.63%、98.05%。均高于现有文献的识别方法。

(3)平衡运算性能和识别准确度,当机器性能较弱时,建议采用上下文窗口大小为 3,特征数量为 4。当机器性能较好时,建议采用上下文窗口大小为 7,特征数量为 4。

参 考 文 献

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 Lianjun. Research on the development model of agricultural e-commerce base internet + [J]. Agriculture Network Information, 2015(11):19 - 21. (in Chinese)

2 陈基. 命名实体识别综述[J]. 现代计算机, 2016(3):24 - 26.
CHEN Ji. Survey of named entity recognition[J]. Modern Computer, 2016(3):24 - 26. (in Chinese)

3 蔡爱杰. 基于 Web 的命名实体提取的研究方法[J]. 哈尔滨师范大学自然科学学报, 2010, 26(2):90 - 94.
CAI Aijie. Research of Web-based named entity extraction[J]. Natural Sciences Journal of Harbin Normal University, 2010, 26(2):90 - 94. (in Chinese)

4 谷川, 周宏宇, 于江德. 融合多特征的中文产品命名实体识别[J]. 科学技术与工程, 2013, 13(31):9417 - 9421.
GU Chuan, ZHOU Hongyu, YU Jiangde. Fusion of multiple features for Chinese product named entity recognition[J]. Science Technology and Engineering, 2013, 13(31):9417 - 9421. (in Chinese)

5 姜仁会, 王挺, 唐晋韬. 面向微博文本的命名实体识别[J]. 计算机与数字工程, 2014, 42(4):647 - 651.
JIANG Renhui, WANG Ting, TANG Jintao. Named entity recognition for micro-blog[J]. Computer & Digital Engineering, 2014, 42(4):647 - 651. (in Chinese)

6 李丽双, 何红磊, 刘珊珊,等. 基于词表示方法的生物医学命名实体识别[J]. 小型微型计算机系统, 2016, 37(2):302 - 307.
LI Lishuang, HE Honglei, LIU Shanshan, et al. Research of word representations on biomedical named entity recognition[J]. Journal of Chinese Computer Systems, 2016, 37(2):302 - 307. (in Chinese)

7 冯蕴天, 张宏军, 郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学, 2015, 42(7):15 - 18.
FENG Yuntian, ZHANG Hongjun, HAO Wenning. Named entity recognition for military text[J]. Computer Science, 2015, 42(7):15 - 18. (in Chinese)

8 方莹. C - CRF 模型在农作物名识别中的应用[J]. 广东农业科学, 2011, 38(6):197 - 199.
FANG Ying. Research on application of C - CRF model on named entity recognition of the agricultural product[J]. Guangdong Agricultural Sciences, 2011, 38(6):197 - 199. (in Chinese)

9 王春雨, 王芳. 基于条件随机场的农业命名实体识别研究[J]. 河北农业大学学报, 2014, 37(1):132 - 135.
WANG Chunyu, WANG Fang. Study on recognition of Chinese agricultural named entity with conditional random fields[J]. Journal of Agricultural University of Hebei, 2014, 37(1):132 - 135. (in Chinese)

10 黄念娥, 黄河, 王儒敬. 本体与条件随机场结合的涉农商品名称抽取与类别标注[J]. 计算机应用, 2017, 37(1):233 - 238.
HUANG Niane, HUANG He, WANG Rujing. Agriculture-related product name extraction and category labeling based on ontology and conditional random field[J]. Journal of Computer Applications, 2017, 37(1):233 - 238. (in Chinese)

11 ZHANG H P, LIU Q, HONG-KUI Y U, et al. Chinese named entity recognition using role model[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(2):29 - 59.

12 ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002:473 - 480.

13 PIETRA S D, PIETRA V D, MERCER R L, et al. Adaptive language modeling using minimum discriminant estimation[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Computer Society, 1992:633 - 636.

14 HOBERG T, ROTTENSTEINER F, FEITOSA R Q, et al. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery[J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53(2):659 - 673.

15 GUO H. Accelerated continuous conditional random fields for load forecasting[C]// IEEE, International Conference on Data Engineering. IEEE, 2016:1492 - 1493.

16 PAISITKRIANGKRAI S, SHERRAH J, JANNEY P, et al. Effective semantic pixel labelling with convolutional networks and conditional random fields[C]//Computer Vision and Pattern Recognition Workshops. IEEE, 2015:36 - 43.

17 朱艳辉, 刘璟, 徐叶强,等. 基于条件随机场的中文领域分词研究[J]. 计算机工程与应用, 2016, 52(15):97 - 100.
ZHU Yanhui, LIU Jing, XU Yeqiang, et al. Chinese word segmentation research based on conditional random field[J]. Computer Engineering and Applications, 2016, 52(15):97 - 100. (in Chinese)

18 LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 2001:282 - 289.

19 林海琳, 杨辉荣. 农药有机合成的进展与展望[J]. 化工科技, 1998(4):17 - 20.
LIN Hailin, YANG Huirong. The development and prospect for organic synthesis of pesticides[J]. Science & Technology in Chemical Industry, 1998(4):17 - 20. (in Chinese)

20 LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016:260 - 270.