Named Entity Recognition for the Horticultural Domain

Ziyu Liu\textsuperscript{1*}, Meng Luo\textsuperscript{2}, Hehe Yang\textsuperscript{2} and Xingwang Liu\textsuperscript{3}

\textsuperscript{1} Information Research Center, China Agricultural University, Beijing, China
\textsuperscript{2} Library & Information Science Program, China Agriculture University, Beijing, China
\textsuperscript{3} College of Horticulture, China Agricultural University, Beijing, China
Email: liuziyu@cau.edu.cn

Abstract. Named Entity Recognition (NER) is one of the fundamental tasks in natural language processing and knowledge engineering, as well as a prerequisite step of many downstream applications. Horticulture, a major branch of agricultural science, means the cultivation, processing, and sale of fruit, nuts, vegetables, and ornamental plants as well as numerous additional services. NER for the horticulture domain means to find key biological traits and state-of-art experimental methods for the horticulturists, new cultivation methods and useful tools for farmers, as well as other important information for planners and policy makers to trigger decision-making procedures. In this paper we designed an NER tagging-set of 7 fine-grained types, and since there is no publicly shared annotated corpus available in horticulture domain, we constructed training and testing corpora manually. Thus, we realized Horticulture NER in scientific literature abstracts with CRF method. Results showed that our system's accuracy and precision were satisfactory but still have room for improvement.

1. Introduction

1.1. Why NER Is Important
Name Entity Recognition (often referred as NER) is one of the fundamental tasks in information extraction. Its purpose is to identify parts of natural language text that represent entities and classify them into predefined categories. It also serves as the prerequisite step of many downstream applications of knowledge engineering and natural language processing, such as entity linking, machine translation, knowledge base completion etc. After thriving on almost a quarter-century, the field of NER had made a huge accomplishment in the diversity of languages and entity types. However, most of the work has concentrated on limited domains and textual genres. Only a few specific domains had to develop their own NER paradigm, such as Biomedical NER or chemical NER, and we still need more research on fine-grained NER in other domain-specific areas to support various knowledge engineering missions.

1.2. Necessity for NER in Horticulture Domain
Agriculture is the vital source of human livelihood. Agronomic information is essential to improve agricultural productivity, research and development efforts. FAO (Food and Agriculture Organization) emphasized that agronomic information is the crucial factor of global sustainable development [1]. Horticulture is a major branch of agricultural science and means the cultivation, processing, and sale of fruit, nuts, vegetables, and ornamental plants as well as many additional services [2], while other major branches include animal science, crop science, plant protection science etc. Horticulture
represents industries worth thousands of billion dollars, and has a profound impact on both of the global sustainable development and human well-being. However, despite the importance of agricultural science and horticultural science, there is always a lack of knowledge engineering research for these domains. Traditional knowledge organization studies carried by the library science community, such as AGROVOC by FAO, were limited in vocabulary and functions, since new phrase like CRISPR-Cas9 (an epoch-making gene editing technology) wasn't even included in this taxonomy.

1.3. Related Work

Since the ability to recognize previously unknown entities is an essential part of the NER system, related studies hinged upon recognition and classification rules triggered by distinctive features associated with positive and negative examples. While early studies mostly used handcrafted rules and lexicon (also called by dictionary, thesaurus, gazetteers, etc.) to achieve this goal [3, 4], most recent research used supervised, semi-supervised or unsupervised learning technologies to automatically induce rule-based systems or sequence labeling algorithms [5, 6].

We did an explorative survey in the core collection of the Web of Science database. With the search query “TI= ((Named entity recognition) OR NER OR (entity extraction) NOT (entity relationship extraction))” (retrieval on May 24, 2019), a total of 1 090 records were obtained. Publication dates are from 1991 to 2019, and the literature type is Article and Review. Clustering results by keywords co-occurrences are shown in figure 1. The applications of NER (in blue frames) mainly included entity extraction, relationship extraction, event extraction, automatic speech recognition, electronic health records, etc. The learning method of NER (in orange frames) mainly included Conditional Random Fields, Maximum Entropy, Hidden Markov Models, Supported Vector Machine, Association Rules, Recurrent Neural Network, and Bi-LSTM etc.

Systematic surveys that were separately conducted by Li et al. [7] covered related studies from 1996 to 2018. Meanwhile, systematic surveys also revealed that the statistical learning methodology could effectively reduce the human cost in NER with acceptable or even satisfying precision and recall rate. Hence, machine learning dominated recent study in both of the general NER and domain-specific NER. Vijayakrisha et al. [8] established Tamil NER for Tourism domain using CRF. Patil et al. [9] and Malarkodi et al. [10] successively worked on agricultural NER. Lu et al. [11, 12] used neural networks and achieved promising results in Biomedical NER. In previous agronomic NER research [8, 9], text genre was mainly Wikipedia articles that were edited and structured through crowdsourcing. The degree of difficulty was therefore relatively petty. Furthermore, previous study defined NER types without consultancy from agronomist, and this reduced the application value of their study.

2. Methodology and Data

2.1. Designing Tag-set for Horticulture Domain

Drawing up a fine-grained tag-set is prerequisite requirements for good NER. We used a classical Delphi Method to generate our tag-set. After consultancy and voting from 3 senior Horticulturalists affiliate to CAU, we designed a tag-set consisting of 7 types of Horticultural NER. Table 1 indicates the types, descriptions along with examples.

2.2. Conditional Random Field

Conditional Random Fields (CRF)[13] is a statistical model, a typical undirected graph, widely used in multi-linguistic NER mission. Lafferty J. et al. first proposed it in 2001. The main idea of CRF came from the Maximum Entropy model. Named entity recognition can be described as a sequence-labeling problem, aiming to judge whether the observation words belong to the predefined tag-set. There isn't any independent assumption of CRF. It can arbitrarily select and globally normalize multiple features to generate the global optimal solution. Hence, CRF not only retained the advantages of the Maximum Entropy, Hidden Markov model and other conditional probability frameworks, but also solved the problem of tagging-bias.
For a given word sequence \( x = \{x_i\} (i = 0, 1, \ldots, n) \) and the annotation sequence \( y = \{y_i\} (i = 0, 1, \ldots, n) \), a CRF model could be defined as following:

\[
p(y|x, \lambda) = \frac{1}{z(x)} \exp \left( \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_i) \right)
\]

\[
z(x) = \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x_i)
\]

where \( z(x) \) is the normalization factor, \( n \) represents the length of a given word sequence, \( f_j(y_{i-1}, y_i, x_i) \) is the feature function of an observation sequence and the labels \( i \) or \( j \) stands for word position in sequence, \( \lambda_j \) is the weight coefficient of the corresponding feature function, which is also the key parameter in machine learning.

### Table 1. NER tag-set with examples.

| NE Types       | Description                                                                 | Examples                           |
|----------------|----------------------------------------------------------------------------|------------------------------------|
| Organism       | Names of plants, fruits, nuts, flowers and other horticultural materials    | Cucumber, Rosa, Tomato, etc.       |
| Trait          | Names of all kinds of biological traits, such as fruit traits, quality traits, development traits, etc. | size, length, shape, sex differentiation, spine, tumor, fruit frost, aroma, sweetness, etc. |
| Method/Equipment | Names of research methods or laboratory equipment                           | facility culture, genetic population mapping, RNA-Seq, GWAS, gene cloning, etc. |
| Chemical       | Names of phytohormone, protein, fertilizers, pesticides, fungicides, etc.   | Nitrate, nitrogen, ethylene, ABA, auxin, gibberellin, etc. |
| Gene           | Names of structural genes and regulatory genes                             | NIP, PLA, NAC, MYB, etc.           |
| Environment    | Names of climate factors, natural disasters, and environmental stress       | low temperature, weak light, high temperature summer, winter, temperature, etc. |
| Miscellaneous  | Other entities that could not be classified into above categories           | Money, distance, numbers, date, etc. |

### 2.3. Corpus Collection and Manual Annotation

CRF was used only for supervised learning and its performance depended on training data. Hence, the corpus collection is a crucial factor affecting the final results. The dataset used in this paper was scientific literature abstracts downloaded from the web of science database. Suggested by the same horticulturists group in Delphi Interview, we collected article abstracts of 10 typical horticultural journals, published in 2018, a total of 2028 records (retrieval on May 24, 2019). The corpus covered major aspects of horticulture research from traditional plant breeding to molecular/genetics research, and from wine brewing to cut flower preservation. The corpus comprised of 600k tokens and 68923 named entities. Named entities are manually annotated with the help of nine post-graduate students who all had horticulture research background. The corpus is randomly divided into 80:20 for training and test data.

### 3. Experiments and Results

#### 3.1. Experimental Procedures

The raw text is tokenized and pre-processed to get part of speech and chunking information in the Pretreatment procedure, mainly using the NLTK library [14].
Though we used a machine-learning model, we also embraced with linguistic rules. Within the Feature Extraction procedure, we took advice from horticulture specialists and chose syntactic features as below and the whole experiment procedure was showed in figure 2:

1. Word: word itself as in the corpus;
2. Part of speech: mainly referring to proper noun, since noun and noun phrase play significant roles in NER and help identify the entity boundary.
3. Cue phrases: keywords occurring as part of an entity. The key-terms like blight, rot are occurring as part of plant disease, and terms like flavor, scent are commonly seen in fruit quality traits.
4. Parentheses: we had examined the parentheses following the proper noun or noun phrase. Our analysis revealed that in most of the cases characters in parentheses suggest abbreviation, which meant noun or noun phrase before parentheses would probably be some types of entity, such somatic embryogenesis (SE), grafted overhead-sweet potato cultivation (GOSC), etc.
5. Capital letters: when every word in a noun phrase begins with a capital letter, this noun phrase would highly probably be a named entity, such as EST-SSR, RAPD, CRISPR-Cas9, etc.
6. Numerical feature: As we classify pure digital patterns as miscellaneous, we also define the characters with connector and digit numbers would probably be entities like gene or protein. For instance, numerical values of length four usually tend to identify year names, and CgY1, CsUBL5, S8-RNase are genes and genetic markers.

The CRF model proposed in this paper was implemented in Python programming, and the main parameters of all the model training computing environments in this experiment were CPU 2*Intel Xeon, CPU E5-2620 v2@2.10 GHz, GPU 1*Nvidia Tesla K20, and 32 GB of memory.

3.2. Results and Evaluation
We randomly divided the horticulture corpus in the ratio of 80:20 for training and test data, and we measured the performance of our model in terms of precision, recall and F1 measure. In order to locate the best feature set, we have conducted several experiments using various combinations of features. Feature-wise performances are given in figure 3.

Initially, we simply applied word feature to determine the baseline performance of our model. We obtained F1 of 68.55 for the basic CRF model. With the increase of feature combination, the accuracy, recall rate and F1 value of model recognition of horticultural named entities were gradually improved. Results showed as in table 2 that the selected features had achieved good results in the recognition of entities in horticulture domain. With the increase of the feature combination, the recognition rate also enlarged. In this process, the increase of Cue phrases had the greatest impact on precision, reaching 4.15%, while Numeric feature’s contribution is the secondly, reaching 2.15%. However, Parentheses, POS and Numeric feature contributed most to the revision of recall rates, reaching 4.44%, 3.61% and 3.24% respectively.

![Figure 1. Keywords co-occurrences clustering of NER literature using VOSviewer.](image1)

![Figure 2. Experiment procedure.](image2)
Table 2. Feature-wise results.

| Feature Combination                      | Precision | Recall | F1    |
|-----------------------------------------|-----------|--------|-------|
| 1 Word                                  | 73.54     | 64.20  | 68.55 |
| 2 Word+POS                              | 74.62     | 67.81  | 71.05 |
| 3 Word+POS+Cue                          | 78.77     | 70.19  | 74.23 |
| 4 Word+POS+Cue+Parenthese               | 79.32     | 74.63  | 76.90 |
| 5 Word+POS+Cue+Parenthese+Capital       | 80.41     | 76.34  | 78.32 |
| 6 Word+POS+Cue+Parenthese+Capital+Numeric | 82.56     | 79.58  | 81.04 |

Figure 3. Tag-wise Results from CRF.

We also conducted Tag-wise comparison among precision, recall rate and F1 parameter. Results showed that our model performed very well as for recognizing organism, gene and miscellaneous, and acceptably with recognizing traits and method/equipment. However, the CRF model performed mediocrely for chemicals and environment factors. The main reasons for the satisfactory recognition results of organisms, genes and miscellaneous entities included: (1) the names of organisms were usually written regularly, limited in numbers and grammar; (2) the recognition efficiency of gene names was greatly improved by increasing capital letters and numerical features; (3) miscellaneous entities are recognized because of the large amount of them. The main reasons for the poor recognition results of other entities included: (1) the biological traits were very complex, and the expression forms are diverse, involving a large number of adjectives and adverbs besides nouns. (2) Different author expressed the research methods or experimental equipment quite differently. (3) Too many abbreviations of chemical substances lead to high rate of false recognition. (4) Environmental factors are the most complex named entity in horticultural domain, such as temperature, light, biological stress and etc., and it could be written in single word, phrases, numbers, or even a whole sentence in some abstracts, so it is very difficult to identify them.

4. Conclusion
In this paper, we have presented a NER system for the horticulture domain. To the best of our knowledge, this work is the first attempt in generating a NE annotated corpus and NE system with an expertise-based tag-set for this domain. We had collected corpora data from scientific literature and established our corpus through manually annotation. We designed the NER tag-set of 7 fine-grained tags with the help of horticulture specialists, so that it could cover prominent entities in this specific field. Our system exploits both machine learning techniques and linguistic rules. With our model, we achieved a basic recognition accuracy of 68.55% with minimal feature of word. Based on a detailed corpus analysis, more features were incorporated and our results were enhanced therefore by more than 11%.
However, due to the nature of entity types, our model worked unsatisfyingly for recognizing biological trait, method or equipment, chemical and environment factor in horticulture domain. In future we plan to extend this work in three way: (1) developing more effective syntactic rules or using more efficient statistical learning model to increase the performance of horticulture NER; (2) embracing neural networks methodology to generate semi-supervised or unsupervised system to reduce hand-crafted workload; (3) developing an robust relation extraction system and knowledge base for the horticulture domain.

Acknowledgments
This work was supported by CALIS Open Funding (funding number 2019kt110) to Ziyu Liu and National Key Research and Development Program of China (2018YFD1000800) to Xingwang Liu. It is the result of the collaboration between Information Research Center and Beijing Key Laboratory of Growth and Development Regulation for Protected Vegetable Crops, both affiliated to China Agricultural University.

References
[1] Vidanapathirana N P 2012 Agricultural information systems and their applications for development of agriculture and rural community a review study Proceedings of 35th Information Systems Research Seminar in Scandinavia pp 1-14.
[2] Wikipedia contributors 2019 Horticulture Wikipedia the Free Encyclopedia.
[3] Mikheev A, Moens M and Grover C 1999 Named entity recognition without gazetteers Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics).
[4] Hasegawa T, Sekine S and Grishman R 2004 Discovering relations among named entities from large corpora Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics pp 415-422.
[5] Petkova D and Croft W B 2007 Proximity-based document representation for named entity retrieval Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management pp 731-740.
[6] Qu L, Ferraro G, Zhou L, Hou W and Baldwin T 2016 Named entity recognition for novel types by transfer learning arXiv preprint 1610.09914.
[7] Li J, Sun A, Han J and Li C 2020 A survey on deep learning for named entity recognition IEEE Transactions on Knowledge and Data Engineering.
[8] Vijayakrishna R and Sobha L 2008 Domain focused named entity recognizer for Tamil using conditional random fields Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.
[9] Patil S, Pawar S and Palshikar G 2013 Named entity extraction using information distance Proceedings of Sixth International Joint Conference on Natural Language Processing pp 1264-1270.
[10] Malarkodi C S, Lex E and Devi S L 2016 Named entity recognition for the agricultural domain Research in Computing Science 117 121-132.
[11] Leaman R, Wei C H and Lu Z Y 2015 tmChem: A high performance approach for chemical named entity recognition and normalization Journal of Cheminformatics 7 (S1) S3.
[12] Leaman R and Lu Z Y 2016 TaggerOne: Joint named entity recognition and normalization with semi-Markov models Bioinformatics 32 (18) 2839-2846.
[13] Lafferty J, McCallum A and Pereira F C 2001 Conditional Random Fields for segmenting and labelling sequence data Proceedings of ICML-01 pp 282-289.
[14] Bird S, Klein E and Loper E 2009 Natural Language Processing with Python (O’Reilly Media Inc.).