IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 12, 2021, accepted November 8, 2021, date of publication November 15, 2021,
date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3128178

Domain Specific Entity Recognition With Semantic-Based Deep Learning Approach

QUOC HUNG NGO, TAHAR KECHADI, AND NHIEN-AN LE-KHAC, (Member, IEEE)
University College Dublin, Dublin 4, D04 V1W8 Ireland

Corresponding author: Quoc Hung Ngo (hung.ngo@ucdconnect.ie)

This work was supported in part by the SFI Strategic Partnerships Programme under Grant 16/SPP/3296, and in part by Origin Enterprises Plc.

ABSTRACT In digital agriculture, agronomists are required to make timely, profitable and more actionable precise decisions based on knowledge and experience. The input can be cultivated and related agricultural data, and one of them is text data, including news articles, business news, policy documents, or farming notes. To process this kind of data, identifying agricultural entities in the text is necessary to update news with agricultural orientation. This task is called Agriculture Entity Recognition (AGER - a kind of Named Entity Recognition task, NER, in the agriculture domain). However, there are very few approaches on AGER because of a lack of the consistent tagset and resources. In this study, we developed a new tagset for AGER to cover popular concepts in agriculture and we also propose a process for this task that consists of two stages: in the first stage, we use semantic-based approaches for detecting agricultural entities and semi-automatically build an annotated corpus of agricultural entities, while in the second stage, we identify the agricultural entities from the plain text using a deep learning approach, train on the annotated corpus. For the evaluation and validation, we build an annotated agriculture corpus and demonstrated the efficiency and robustness of our approach.

INDEX TERMS Agriculture entity recognition, WordNet, semantic class, named entity recognition, deep learning.

I. INTRODUCTION

Timely and cost-effective data updating plays a significant role in the development of digital agriculture. In this context, data is well-known as digital data, such as remote sensing data, soil nutrients, yield maps, weather data, etc. However, data can also be text data, as in agriculture news articles, farming business news, farming policy documents, mass media, and rural issues. Moreover, text data sources can be considered as social sensors, which can be useful sources of information. Analysing text data effectively can not only help people keep up-to-date with recent practices and trends, but also support human experts when making decisions in agriculture task management. Therefore, entity tagging is very important in a news processing system [5].

In a news processing system or natural language processing (NLP) domain, Named Entity Recognition (NER) aims at classifying words of a document into predefined target entity classes. It is now considered to be fundamental for many NLP and information retrieval (IR) tasks. In addition, NER is also fundamental in the ontology construction process or building knowledge graphs. As shown in Fig. 1, NER is a basic stage of the instance extraction and fact extraction modules in an (ontological) information extraction process as well as in building knowledge maps or knowledge banks.

Unstructured data, such as digitized news articles or reports, is usually stored as raw text articles. Valuable information in such text is difficult to find and extract. Hence, NER is useful to identify meaningful entities and make them a valuable input information for data analytic applications. For instance, named entities, such as location, person, organisation names, are firstly recognised and extracted from documents, [4], [24]. Then, the number of tags in NER tagsets is increased to give more specific meanings, [12] [19], [30] e.g., the organisation tag is divided into Company, Institute, Military tags in the extended Name Entity (NE) hierarchy of the approach mentioned in [30]. The current research direction is to extract meaningful entities in many domains,

The associate editor coordinating the review of this manuscript and approving it for publication was Gianmaria Silvello.

TABLE 1. Abbreviation terms.

Term	Description
NLP	Natural Language Processing
NE	Named Entity
NER	Named Entity Recognition
AGER	Agriculture Entity Recognition
IR	Information Retrieval
ML	Machine Learning
DNN	Deep Neural Network
CNN	Convolution Neural Networks
CRF	Conditional Random Fields
LSTM	Long-short Term Memory
BiLSTM	Bi-direction Long-short Term Memory
MUC	Message Understanding Conference
GMB	Groningen Meaning Bank
LDOCE	Longman Dictionary of Contemporary English
LLOCE	Longman Lexicon of Contemporary English

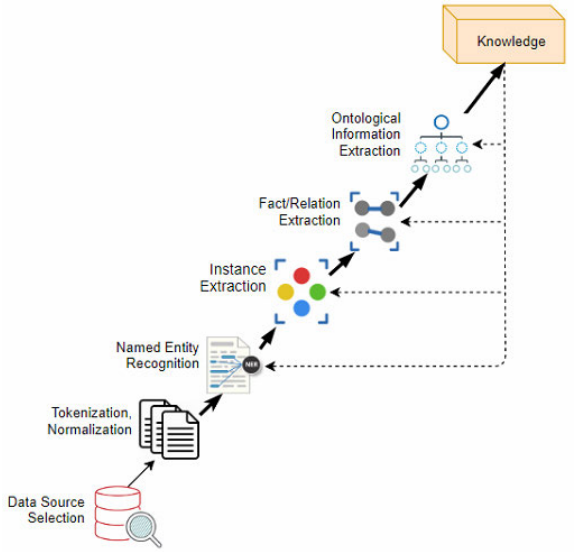

FIGURE 1. Role of NER in information extraction process, [11].

such as bioinformatics, [15] [16], and agriculture domains, [2] [6]–[8].

Usually, named entities are a kind of proper names while others are general entities within the agriculture domain. Therefore, NER can be carried out based on orthographic features, word-level features, and gazetteers. Data analytic techniques are used to classify named entity tags and the disambiguation among named entities based on their features, [10]. However, these features have very little contribution to the AGER task because general agricultural entities are not proper names and they do not have orthographic features like named entities in the NER task. This is a challenge for any entity recognition system.

In this paper, we propose a NER process in a specific domain which agriculture. The key highlights of the paper are: (1) design a new tagset for agricultural entities; (2) develop a semantic-based Agricultural Entity Recognition (AGER) tasks and a new efficient approach for AGER; (3) use deep learning (DNN) to train and identify agricultural entities from plain text. The semantic-based AGER approach combines a dictionary-based approach and semantic classes. This approach can recognise entities without training on the annotated corpora. We also describe a post-processing step to trim agricultural entities and get their root entities that are useful in text analytics and queries. Finally, we built an annotated agriculture corpus with more than 21, 000 news articles and 112, 500 agricultural entities for evaluation and validation purposes.

The next section presents an overview of general agricultural entity recognition. Section III reviews tagsets and corpus issues for agricultural entities. Then, in Section IV we present an efficient semantic approach, while Section V details experimental results and analysis of the AgNews corpus. We conclude and give some future research directions in Section VI. Moreover, Table 1 provides a list of abbreviation terms, which are used in this study.

II. RELATED WORK

The named entity extraction in agriculture domain has several applications. These include detecting and tracking the occurrence of some entities and patterns locality over time and summarising agriculture documents and reports. Another very popular application example is the detection and monitoring of disease outbreaks based on mass media. One of the mandatory modules for such application is NER. NER was introduced for the first time at the Sixth Message Understanding Conference (MUC-6) in 1995, [24]. At that time, MUC-6 and MUC-7 focused on Information retrieval, which can extract some patterns and models from unstructured text data. In information extraction, NER is used to recognise within text documents named entities, such as person, organisation, location names, numeric expressions including time, date, money, and regular expressions, [9]. Since 1995 (MUC-6), many conferences have been focusing on this research area. One of the well-established conference in this area is CoNLL-2003.

NER is a kind of sequence labeling tasks and uses token-based classification techniques. The tokens in input sentences can be classified as one of NER tags or NONE tag (for non-entity words). Types of features include orthographic patterns, part-of-speech tags, chunk tags, affixes, and gazetteers, [10]. In the last decade, many studies have been conducted on the application of machine learning to named entity recognition problem with published (i.e., CoNLL-2003[1] corpus for Shared Task; OntoNotes[2] corpus, and Groningen Meaning Bank[3] (GMB) dataset), [12]. For instance, Deep Learning (DL) approaches were used on NER with Bidirectional Long-Short Term Memory (BiLSTM), Convolutional Neural Networks (CNN) and Conditional Random Fields (CRF), [17], [18], [31] (as shown in Table 2). Specifically, these DL approaches are based on

[1]https://www.clips.uantwerpen.be/conll2003/ner/
[2]https://catalog.ldc.upenn.edu/ldc2013t19
[3]http://gmb.let.rug.nl/

word representation, such as GloVe[4], ELMo,[5] and BERT[6], are state-of-the-art approaches for the NER task.

TABLE 2. Previous NER's studies based on deep learning.

Approach	Experiment and Results	Authors
LSTM-CNN	Achieved an accuracy of 91.62% based on GloVe and CoNLL-2003 dataset	[17]
LSTM-CNN-CRF	Achieved an accuracy of 91.21% based on GloVe and CoNLL-2003 dataset	[31]
LSTM-CRF+ELMo	Achieved an accuracy of 92.22% based on ELMo and CoNLL-2003 dataset	[27]
BERT-base+LSTM	Achieved an accuracy of 92.4% based on the self-attention mechanics	[13]
Hierarchical BERT	Achieved an accuracy of 93.37% based on the self-attention mechanics	[20]
LUCE	Achieved an accuracy of 94.3% based on entity-aware self-attention mechanics	[32]

Named entity types are extended to other application domains. For example, many NER studies have been conducted in medical domain (with medication, disease names, drugs), [21], bioinformatics with GENIA corpus, [15], which triggered many other studies dedicated to specific types such as ''protein'', ''DNA'', ''RNA'', ''cell line'' and ''cell type'', [16]. NER is the main module in the analysis and monitoring systems architectures, like BioCaster, which is an ontology-based text mining system for detecting and tracking the spread of infectious disease outbreaks from global news articles, [25]. NER is also extended to more named entity types. The most well-known tagset is presented by Satoshi Sekine et al., with a named entity hierarchy containing about 150 NER types, [30]. This tagset is used in the GMB dataset, [12].

TABLE 3. Previous AGER's studies.

Approach	Experiment and Results	Authors
Semantic vectors and cosine distances	Recognise cereals and crops names, obtained an F-measure of about 45.1%	[2]
WordNet & gazetteer	Extract basic entities, obtained an accuracy of 72.28% on a corpus of 500 documents containing 171, 735 words	[7]
Expectation Maximization	Apply on 3 types of entities, achieved an accuracy of 76%	[3]
CRF	Apply on 19 fine grained tags, obtained an accuracy of 83% on a corpus of 100, 000 words	[8]

Although NER has been well applied in many domains, its deployment in digital agriculture is very limited. According to [2], the crop specific named entities can be recognised using semantic vectors and cosine distance functions. In their study, they recognised cereals and crops names in both Agro Products corpus and text data documents. Their experiments were performed on a corpus of 2, 206 sentences, 326 cereals and crops names. They obtained an F-measure of about 45% on 50 manually selected seeds. Their system, as presented in 2016, obtained low accuracy, while, all of these cereal and crop names were recognised as a Crop type in other systems. Therefore, it will be better if these entities are recognised as a Crop type and clustered based on their features.

Other studies on AGER were presented in [6], [7]. They recognised named entities in text documents using a potential word list (a type of gazetteer) and WordNet as a knowledge-base. However, the focus was only on extracting basic entities, such as rice, wheat, apple, pear, hazelnut, tomato, etc. They also built a dataset of 500 documents containing 171, 735 words. Their approach has an accuracy of 72.28% and recall of 53.69%. Another study presented in [3] also proposed an unsupervised learning method to extract crop names, diseases, and cure entities in text data. The authors used Expectation Maximisation (EM) algorithm and achieved an accuracy of 76% for three types of entities [3].

Finally, one of the most general AGER models was proposed in [8]. They created a tagset consisting of 19 fine grained tags (including named entity tags, numeric tags, and agricultural tags) and used CRF to identify and classify these tags. With syntactic and lexicon-syntactic features, they trained and evaluated the AGER model on a corpus of 100, 000 words (with over 11, 000 entities) and obtained an F-measure of 83%. Although the accuracy of the whole system is 83%, the accuracy of entity tags fluctuates widely from about 60% (with *Disease* and *Climate* tag) to over 80% (with Crop and Temperature tag) and the study did present an evaluation for only agricultural entity tags.

To summarise, there are several studies of AGER during the last few years, however, their results are still limited in the number of agricultural entity types and their efficiency. They do not have a unique tagset and available corpora for training and evaluation (as shown in Table 3). Furthermore, meaningful entities are not proper name entities, which leads to the difficulty of identifying them from unstructured text. Therefore, an efficient approach to identify general entities in a specific domain is required.

III. AGER TAGSET AND DATASET

A. AGER TAGSET

A seminal benchmark of NER was the CoNLL-2003 shared dataset. Since all NER studies were using CoNLL-2003 to train, analyse and test their corresponding NER systems to assess its performance and compare it to other different NER systems. In this shared dataset, NER task starts with six general named entity tags, including location (LOC), organisations (ORG), person (PER), number (NUM), money (MON) and time (TIM) [10]. The general NER problem has been studied for a long time, with several shared datasets; including CoNLL-2003 Shared dataset, GermEval 2014 NER Shared dataset.[7] However, each domain should have a specific tagset with more tags. For instance, BioNER and GenneNER [16] are more popular studies in many biomedical research.

[4]https://nlp.stanford.edu/projects/glove/
[5]https://allennlp.org/elmo
[6]https://github.com/google-research/bert
[7]https://sites.google.com/site/germeval2014ner/home

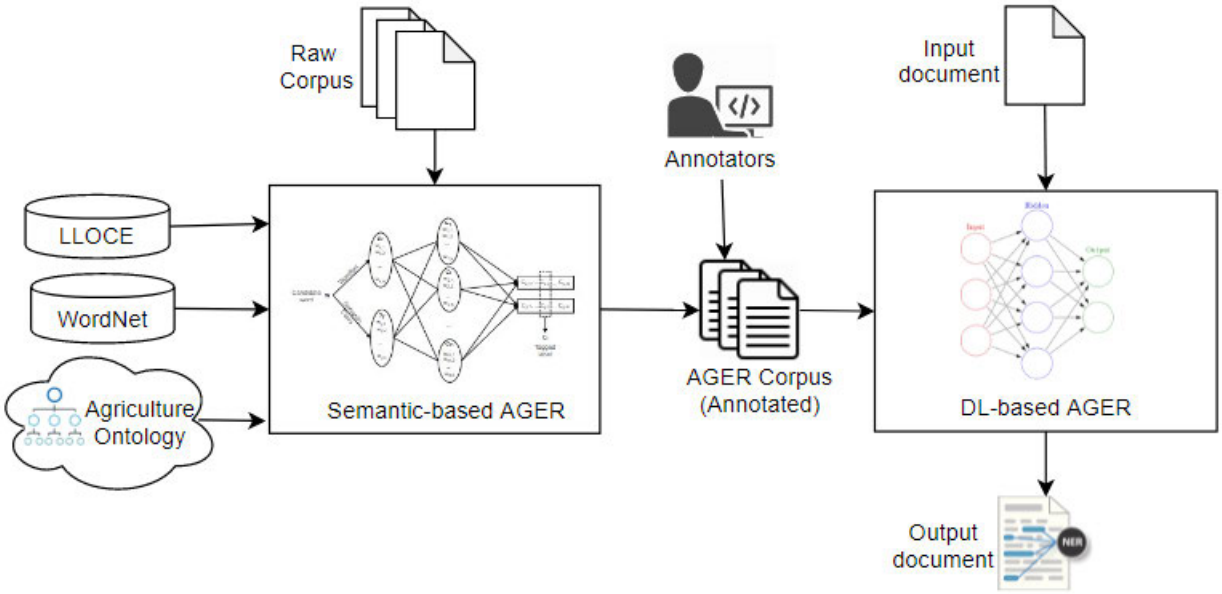

FIGURE 2. Architecture of AGER system.

As mentioned in the previous section, AGER's tagset was defined with 19 fine grained tags, including Person, Location, Organisation, Chemicals, Crop, Organism, Policy, Climate, Food, Disease, Natural Disaster, Events, Nutrients, Count, Distance, Quantity, Money, Temperature, and Year_Month_Date, [8]. This tagset includes both named entity and agricultural entity tags. In our study, we define an AGER tagset with 12 tags, as shown in Table 4. Named entity tags and numeric tags (i.e., Person, Location, Organisation, Count, Distance, Quantity, Money, and Year_Month_Date) are removed when compared with the Malarkor's tagset because these tags can be carried out with available NER toolkit. Moreover, this study focuses on only agricultural entities to improve the AGER model.

TABLE 4. Agricultural entity tagset.

Tag	Name	Description
CRP	Crop	Fruits, vegetables, cereals, grains
ANI	Animal	Name of animals
FOD	Food	Plant/animal products
FAM	Farm	Area of land, used for growing crops
MIO	Microorganism	Name of micro-organism
NUT	Nutrients	Fats, minerals, vitamins
TEM	Temperature	Numerical measure of climate
CLI	Climate	Denotes the climatic conditions
FER	Fertiliser	Bio fertiliser, chemical fertiliser
CHE	Chemical	An agrochemical or chemical
DIS	Disease	Diseases affecting crop/livestock
DST	Disaster	Disasters affecting crop production

B. AGER DATASET

For a general NER model, there are several common corpora that were built for developing and evaluating NER systems, while there is not any available corpus for the AGER model. Therefore, our study is to build an agricultural dataset (named AgNews), which is crawled from available news channels in 5 topics (including Crops, Livestock, Markets, Weather, and Machine) with details shown in Table 5.

TABLE 5. Details of AgNews corpus for agriculture domain.

Topic	Articles	Sentences	Named Entities	Agri. Entities
Crops	7,358	32,221	67,494	45,501
Livestock	5,523	27,410	53,145	36,072
Markets	5,719	45,985	127,896	63,175
Weather	1,682	11,891	27,900	16,689
Machine	981	2,516	6,185	3,285
Total	21,263	120,023	282,620	164,722

AgNews is a set of agricultural news articles collected over a period of 10 years (from 09/2007 to 06/2018). Although AgNews was mainly built to evaluate the AGER system, this corpus was also tagged general NER with SpaCy toolkit for future text analytics. **SpaCy**[8] is a free open-source library for NLP written to improve performance and models. This corpus contains 21,263 articles and 120,000 sentences with 282,600 named entities and 164,700 agricultural entities. Specifically, the corpus has more than 400,000 sentences, however, it only contains 120,000 sentences, which contain at least one agriculture entity.

IV. AGER ARCHITECTURE

In this study, we focus on tagging noun phrases that are recognised as one of agricultural classes, such as [a big rabbit]$_{ANI}$, [winter wheat]$_{CRP}$, [the snow storm]$_{DST}$, or [cold weather]$_{CLI}$. However, there are ambiguousness to recognise

[8]https://spacy.io/

entities from input sentences. For instance, in the sentence "The wheat area is increasing quickly in 2018.", [The wheat area]$_{NP}$ is tagged as a noun phrase and there are three cases to recognise entities:

1) [The wheat area]$_{CRP}$ - whole noun phrase is recognised as a CROP entity.
2) The [wheat]$_{CRP}$ area - only wheat is recognised as a CROP entity; and
3) The wheat area - no entity is recognised in this text.

In Case 1, the whole noun phrase "The wheat area" is recognised as a CROP entity, but it is a type of area and it mentions information of area rather than wheat. In Case 2, "wheat" is recognised as a CROP entity, however, in this sentence, wheat is a modifier of area. Finally, in Case 3, no entity is recognised in this text because "area" is not tagged. In our approach, we choose Case 3 for this situation. Eventually, tagged entities will have an integral meaning of objects, which can be classified into one of agricultural tags.

According to some studies, [9], [22] [28], the NER model approaches mainly use orthographic, word-level, and gazetteer features for machine learning. However, these features' named entities are represented in orthographic, while agricultural entities are general entities in a specific domain. Therefore, orthographic features are not effective while semantic approaches are applied for specific-domain entity recognition like AGER, [34].

The proposed AGER system has two stages (as shown in Figure 2). In the first stage, we propose a semantic-based AGER approach to annotate agricultural entities and build the AGER corpus. It is based on three semantic resources (WordNet, semantic classes, and an agriculture ontology) to identify entities and find matching entities from the raw corpus. Then, this corpus is reviewed and corrected by annotators to build a final AGER corpus. In the second stage, we use the BiLSTM combined with CNN or CRF approaches to classify input tokens into one of 12 agricultural entity labels or NONE. In general, these techniques have been widely used for the NER task (as mentioned in Table 2), however, they are rarely used for AGER due to the lack of annotated corpora. In the proposed process, the first stage is used as a baseline system to annotate agriculture entities from raw text for semi-automatically building an annotated corpus, while the DL model in the second stage, is used to recognize agriculture entities with higher accuracy.

A. SEMANTIC RESOURCES

WordNet[9] is a large English lexical database. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (called *synsets*). The English version of WordNet contains 117, 000 synsets. Based on these synsets, WordNet is used to construct the NER model, [4] to measure meaning distances between candidate words and words in the gazetteer of each class. In our experiments, we use WordNet 3.0 in the NLTK toolkit[10] to look up synsets of words.

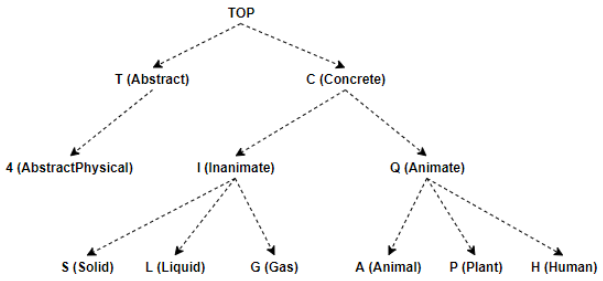

FIGURE 3. Hierarchy of LDOCE classes.

The other semantic resources used are **LDOCE** (Longman Dictionary of Contemporary English) and **LLOCE** (Longman Lexicon of Contemporary English) dictionaries. They are both widely used for daily life and NLP research. LDOCE is organised as a dictionary of hierarchical semantic classes. It has 32 major semantic codes in total; including 19 basic major codes and 13 linking major codes (Figure 3) [26]. LLOCE is a smaller dictionary, which is extracted from LDOCE and it is organised around semantic. Specifically, there are 16, 000 words organised in 2, 441 sets. The hierarchical structure of semantic classes of LLOCE is organised in three levels with increasing details between any two consecutive levels.

For example:

```
<MAJOR: A> Life and living things
|
<GROUP: A50-61> Animals/Mammals
|
<SET: A53> The cat and similar animals:
cat, leopard, lion, tiger,\ldots
```

With the hierarchical structure, sets of words in LDOCE and LLOCE are based on semantic, synonym or related meaning. Therefore, these sets are useful for looking up synonyms or related meaning of words of any words. In this study, we use LLOCE to look up synonyms of given words, and then compare them with words in the gazetteer of each class.

Finally, **AgriOnt** - an agricultural ontology [29] is applied to extract an AGER gazetteer. AgriOnt has 447 classes and over 700 agricultural instances and thousands diseases of plants and animals. This ontology has the main agricultural classes covering all types of the AGER tagset (such as *Crop*, *Animal*, *Farm*, *Food*, *Microorganism*, *Nutrient*, *Temperature*, *Climate*, *Fertiliser*, *Disease*, *Disaster*, *Soil*, etc) [29]. Therefore, instances or entities of this ontology are put into the AGER gazetteer. In addition, this gazetteer can be built manually from AGROVOC.[11]

[9]https://wordnet.princeton.edu/

[10]http://www.nltk.org/howto/wordnet.html

[11]http://aims.fao.org/vest-registry/vocabularies/agrovoc

TABLE 6. Detail of AGER gazetteer.

Tag	Name	Examples	Count
CRP	Crop	Apple, carrot, wheat, rice, tomato	371
ANI	Animal	Sheep, cow, chicken, pig	127
FOD	Food	Cheese, milk, bread	189
FAM	Farm	Farm, field, land, grange, farmhouse	10
MIO	Microorganism	Escherichia coli, fungi, bacteria, virus	32
NUT	Nutrients	Vitamin A, magnesium	91
TEM	Temperature	Celsius, Fahrenheit, °C, °F	14
CLI	Climate	Marine climate, desert climate	49
CHE	Chemical	Nitrogen, nitrate	76
FER	Fertiliser	Bio fertiliser, organic manure	12
DIS	Disease	Late blight, brown rot	566
DST	Disaster	Flood, storm, blizzard, tornado	37

B. SEMANTIC-BASED AGER MODEL

The Semantic-based AGER model consists of two semantic systems, WordNet and Semantic classes, with the view to identify agricultural entities from input text (Figure 4). First, we apply a shallow parser or partitioning to extract all noun phrases. These noun phrases are considered as seed words for classifying them into one of the AGER tags or none. However, seed words are only noun words, which are the most meaningful in the noun phrases. Next, candidate sets are created by looking up synsets of seed words in WordNet. Similarly, another candidate set is created by looking up seed words in semantic classes. Moreover, unit sets of AGER classes are create from AGER Gazetteer. Finally, the candidate sets of each model are compared with unit sets to try to match the result of the two models.

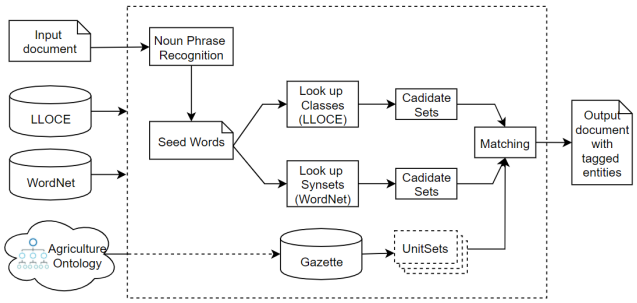

FIGURE 4. Architecture of semantic-base agriculture entity recognition.

The algorithm of the agricultural entity recognition is described as follows:

1) **Step 1**: Apply noun phrase recognition to extract noun phrases.
2) **Step 2**: Extract seed words from noun phrases. Only nouns in noun phrases are filtered.
3) **Step 3**: Look up synsets of longest seed words to build candidate sets by WordNet and by Semantic Classes.
4) **Step 4**: Compare candidate sets of seed words with unit sets of each agricultural class.
5) **Step 5**: Final label of each seed word (also of the noun phrase) is a class that has the highest score and matched of models.

According to the example shown in Figure 5, the candidate word w (a noun phrase of the previous noun phrase recognition module, which is the noun chunk module in SpaCy[12] toolkit in our experiments) can be used to generate a synset S_x by using WordNet. Synset S_x will be measured with *Unitsets* by the overlapping coefficient measure [23]. The resulting scores of WordNet model are $C_{x,i}, C_{x,2}, \cdots, C_{x,m}$. Similarly, the candidate word w is also carried on a similar process by using a Semantic Class dictionary and the resulting scores of Semantic Class model are $C_{y,1}, C_{y,2}, \cdots, C_{y,m}$. Finally, the candidate word w can be tagged label C_i if both $C_{x,i}$ and $C_{y,i}$ reach the highest score.

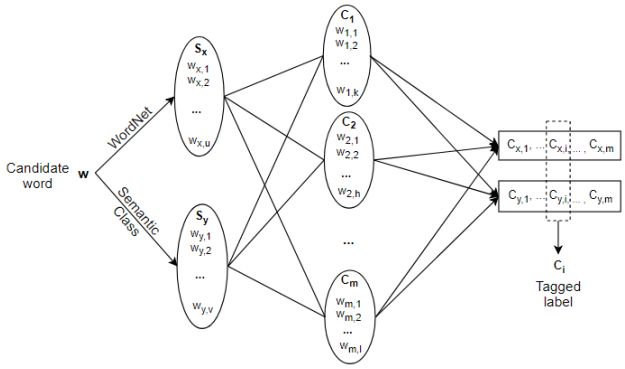

FIGURE 5. Example of agriculture entity recognition.

C. DL-BASE AGER MODEL

There are two main tasks in applying DL approaches into NLP (like the entity extraction task) including the word representation and neural networks for learning. For word representation, there are several approaches to transform text into vectors, such as Word2Vec,[13] FastText,[14] GloVe,[15] ELMo,[16] and BERT.[17] Basically, GloVe and BERT are currently state-of-the-art approaches for word representation in the last five years. GloVe [14] is an essentially log-bilinear model with a weighted least-squares objective and it provides 4 pre-trained models based on 6 billion tokens, 42 billion tokens, 840 billion tokens, and 2 billion tweets at https://nlp.stanford.edu/projects/glove/. BERT [13] is an attention-based model based on the attention mechanics [1] and it provides 24 pre-trained models for multilingual or English-only at https://github.com/google-research/bert. This model provides different vectors for a same word but in a different context.

The architecture of entity recognition based on DL approaches has two main steps: (i) converting the input sentence into feature vectors and (ii) BIO tagging vectors by neural networks. Specifically, the vector from an input

[12]https://spacy.io/
[13]https://code.google.com/archive/p/word2vec/
[14]https://github.com/facebookresearch/fastText
[15]https://nlp.stanford.edu/projects/glove/
[16]https://allennlp.org/elmo
[17]https://github.com/google-research/bert

sentence is converting it to a list of vectors, in which, each vector presents a token in the sentence. Each vector contains two parts: character representation and word embedding. The neural networks include three layers: Forward LSTM, Backward LSTM, and CRF layer (Figure 6).

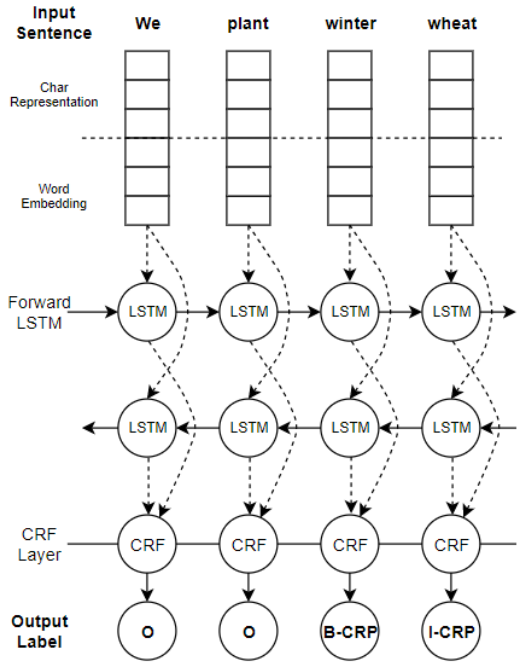

FIGURE 6. Architecture of agriculture entity recognition based on LSTM and CRF.

Our DL-based AGER models use the publicly available word embedding, namely Standford's GloVe[18] embedding, trained on 840 billion tokens from Common Crawl data [14]. In addition, we also use BERT with 12 layers (L) hidden size of 768 (H), and 12 self-attention heads (A) for another option of word embedding.

D. ENTITY STEMMING

According to the top 20 disaster entities from the dataset (see Table 7), they can be stemmed into five entities, which are drought, storm, thunderstorm, flood, and tornado. By clustering agricultural entities, the number of unique root agricultural entities will decrease and it is useful in monitoring and querying as it is more straightforward and effective.

The Entity Stemming mainly focuses on classifying tagged entities into similar groups of entities. This stage has four steps, which are:

1) **Step 1**: Remove stop words from tagged entities.
2) **Step 2**: Remove adjective and adverb from tagged entities.
3) **Step 3**: Remove proper names from tagged entities.
4) **Step 4**: Use stemming algorithms to get final root entities.

[18]https://nlp.stanford.edu/projects/glove/

TABLE 7. Top 20 DST entities.

DST Entities	Count	DST Entities	Count
drought	598	The drought	58
the drought	232	droughts	57
the storm	223	droughts	52
thunderstorms	131	severe thunderstorms	51
the worst drought	120	Storm	50
a drought	118	Drought	48
floods	113	Tornado	41
storms	108	the storms	35
tornadoes	104	the worst U.S. drought	34
a thunderstorm	60	a severe drought	31

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. RESULT OF AGER MODELS

In NLP, the data format for sequence tagging tasks (like noun phrase chunking, entity recognition) is based on the combination of the BIO format (B: Begin, I: Inside, and O:Out) and the tagset (AGER tagset, as described in Section III). Basically, each token in the input sentence can be labeled as B-CRP, I-CRP, B-CHE, I-CHE, etc, and O label. However, the number of O tags in tagging sequences is much larger than other tags. Therefore, we evaluate the accuracy of our AGER approaches based on predicted entities. Their accuracy is computed based on recognised entities as follows:

$$Precision = \frac{|\{ActualEntities\} \cap \{PredictEntities\}|}{|\{PredictEntities\}|} \times 100\%$$

$$Recall = \frac{|\{ActualEntities\} \cap \{PredictEntities\}|}{|\{ActualEntities\}|} \times 100\%$$

and

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

in which, $\{ActualEntities\}$ are manually stated to describe agricultural entities in the corpus, $\{PredictEntities\}$ are predicted entities with each model, and $\{ActualEntities\} \cap \{PredictEntities\}$ are correct entities.

Due to the imbalance in the number of agricultural entities in the corpus (as shown in the first graph of Figure 7), we chose three sizes of the datasets to evaluate the approaches. AgNews1 contains a total of 120,000 sentences with 164,700 agriculture entities, while AgNews2 has only 50,000 sentences with 67,910 agricultural entities. Finaly, AgNews3 is a half of AgNews2. AgNews2 was obtained by removing about 70,000 sentences, which contain most duplicates entities, from the corpus (AgNews1). As a result, AgNews2 and AgNews3 are much more balanced than AgNews1 (See the second and third graphs of Figure 7). The numbers of agriculture entities for each label are presented in Table 8.

In the first experience, we apply our semantic-based approaches to recognise agricultural entities for the corpus with each model, WordNet and Semantic Class, and the combined model (WN&SC). Each model is used to label the AgNews2 corpus and then compare it with the combined approach and the numbers of predicted entities are shown in

TABLE 8. Number of agriculture entities for each tag.

Tag	Name	Number of Agri. Entities AgNews1	AgNews2	AgNews3
CRP	Crop	80,861	13,614	5,826
ANI	Animal	27,345	10,959	5,044
FOD	Food	19,349	10,638	5,206
FAM	Farm	11,868	9,905	3,995
MIO	Microorganism	2,784	2,688	1,867
NUT	Nutrients	886	856	818
TEM	Temperature	341	334	326
CLI	Climate	5,300	3,350	1,116
CHE	Chemical	6,561	6,495	4,177
FER	Fertiliser	1,897	1,867	1,676
DIS	Disease	3,286	3,116	2,706
DST	Disaster	4,244	3,915	2,965
Total	-	164,722	67,737	35,722

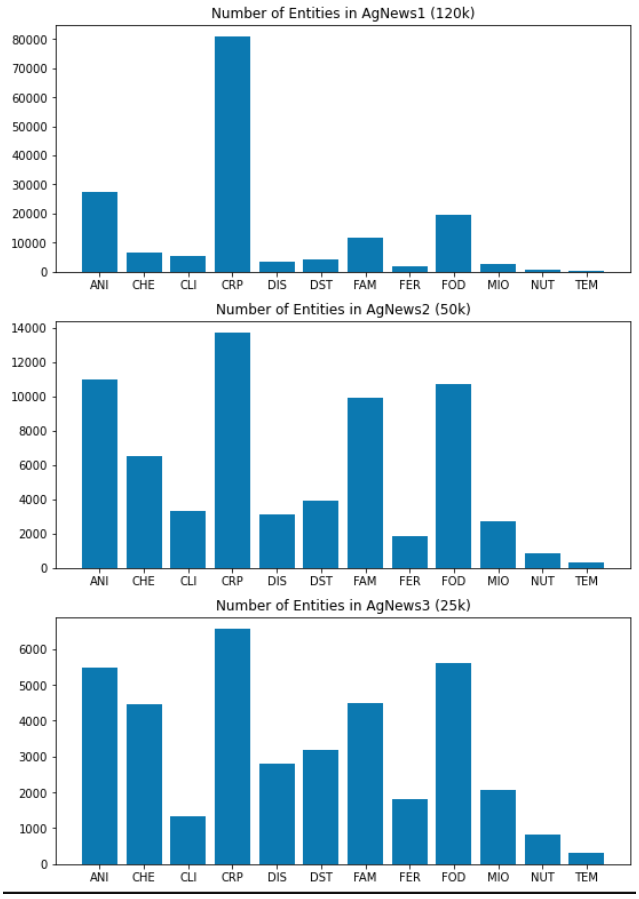

FIGURE 7. Number of agricultural entities in AgNews.

Table 9. As a result, individual models (including WordNet model and Semantic Class model) tag more entities, however, their performance is much lower than of the combined approach (Table 10). In fact, the WordNet model recognised over 65, 000 entities with F-core of 78.43% and the Semantic Class model recognised over 83, 673 entities with F-core of 68.14%, while the combined model recognised 50, 949

TABLE 9. Comparation of AGER semantic-based models on AgNews2.

Tag	Name	WordNet Tagging	SemClass Tagging	WN&SC Tagging
ANI	Animal	10,453	12,183	8,663
CHE	Chemical	3,105	3,218	3,097
CLI	Climate	3,233	3,548	3,182
CRP	Crops	9,665	9,643	5,441
DIS	Disease	3,277	4,710	1,972
DST	Disaster	4,828	4,697	3,720
FAM	Farm	12,991	17,538	10,797
FER	Fertilizer	1,782	1,787	1,782
FOD	Food	9,904	20,902	8,928
MIO	Microorganism	2,523	2,331	2,263
NUT	Nutrients	1,489	1,731	861
TEM	Temperature	1,812	1,385	243
	Total	65,062	83,673	50,949

entities and F-core of 86.19% (with a precision of 90.8% and a recall of 84.22% as shown in Table 10).

TABLE 10. Result of semantic-based AGER model (WN&SC model).

Method	Dataset	Data Size	Precision	Recall	F-score
WordNet	AgNews2	49,740	75.12	89.94	78.43
SemClass	AgNews2	49,740	62.73	84.28	68.14
WN & SC	AgNews2	49,740	90.8	84.22	86.19
WN & SC	AgNews1	120,023	89.0	82.94	84.39
WN & SC	AgNews3	25,288	91.63	83.64	86.26

The second experiences focus on evaluating two DL-based AGER models, including BiLSTM-CRF and BiLSTM-CNN-CRF on AGER corpus. In these experiences, we carry on two popular methods for word embedding, GloVe and BERT, and then combine BiLSTM, CNN, and CRF for hidden layers in the neural networks.

From Table 11 we noticed that the DL approaches with BiLSTM-CRF models have high accuracy with F-score of 95% and the BiLSTM-CRF model based on BERT embedding has a slightly higher F-score than the model based on GloVe word embedding in these experiences (Table 11). Moreover, when the size of the corpus increases, the accuracy of models based on DL approaches also increases.

Basically, the first stage (with semantic resources) is used as a baseline system to annotate agriculture entities from raw text for semi-automatically building an annotated corpus. The tagging data set is reviewed and corrected manually after tagging by the semantic-based model (using WordNet and Semantic classes). The semantic-based model has an accuracy of about 80%, which reduces a lot of work when compared to entirely manually building the annotated corpus.

B. RESULT OF ENTITY STEMMING

In another experiment, we evaluate unique entities and their root entities from tagged entities. In the experiment, the number of root agricultural entities is only less than 1% of the number of unique agricultural entities after the stemming

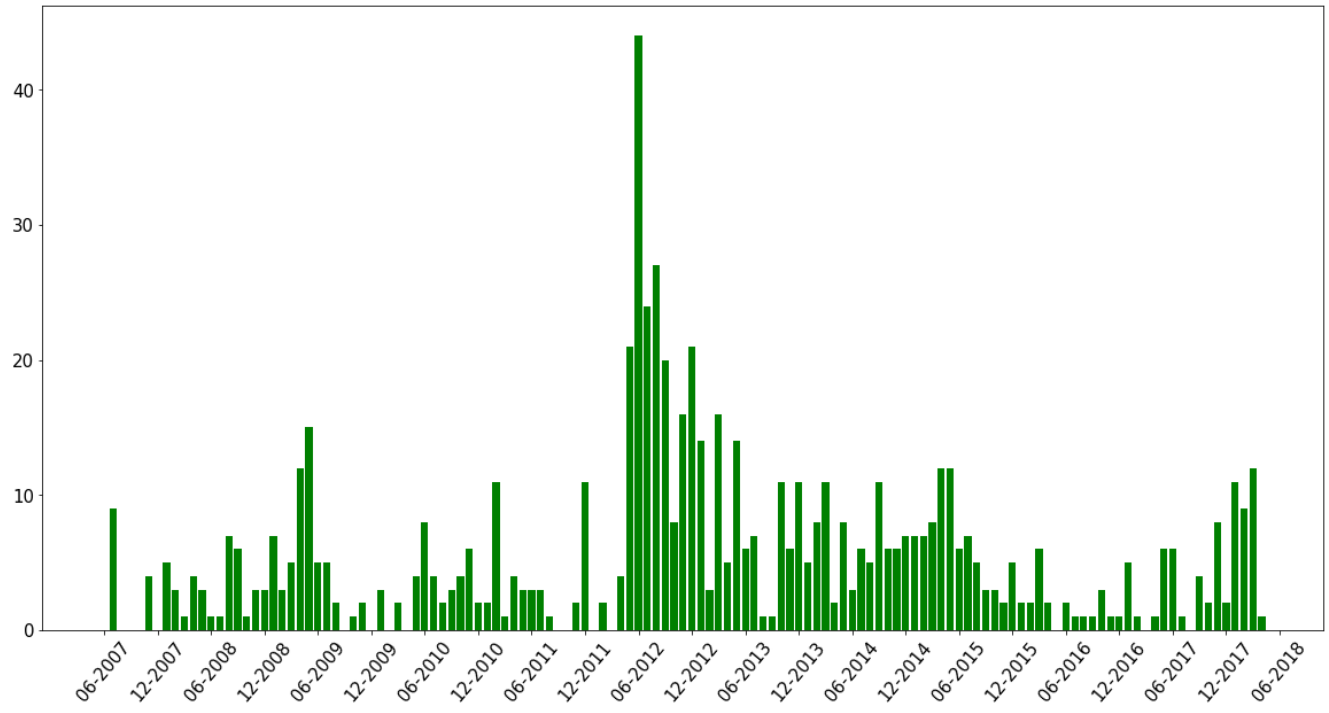

FIGURE 8. Number of news articles about entity DST:drought by time.

TABLE 11. Result of DL-based AGER models.

Model	Dataset	Precision	Recall	F-score
WN & SC	AgNews1	85.7	56.5	68.1
WN & SC	AgNews2	89.5	70.9	79.1
WN & SC	AgNews3	90.8	71.8	80.2
GloVe-BiLSTM-CNN-CRF	AgNews1	93.2	95.1	94.1
GloVe-BiLSTM-CNN-CRF	AgNews2	91.4	93.4	92.4
GloVe-BiLSTM-CNN-CRF	AgNews3	87.4	91.3	89.3
BERT-BiLSTM-CRF	AgNews1	95.3	95.5	95.4
BERT-BiLSTM-CRF	AgNews2	92.2	94.2	93.2
BERT-BiLSTM-CRF	AgNews3	91.0	90.4	90.7

step (as shown in Table 12). In our scenario, the extracted root entities have an important role in building agriculture linked datasets with an agricultural ontology, potential agriculture entities, and their relationships. Although the precision of the rule-based approach for entity stemming is only 73.9%, this step contributes great value in information analysis and retrieval. Because, with entity stemming, the number of unique entities decreases sharply from 21, 343 unique entities to only about 600 root entities (see Table 12). Moreover, there are differences in accuracy among agriculture labels. The result of entity stemming for *Farm* entities is only 21.9% because these entities often include proper names inside the entities and the proposed algorithm works ineffectively. In addition, these entities are also composed of common words, such as area, field. In contrast, groups of less common entities have high accuracies, such as *Fertilizer* and *Microorganism*. Finally, groups with a wide vocabulary and a large number of entities have low precision, for example, *Animals* and *Crops* have a precision of 59.0% and 67.5%, respectively.

TABLE 12. Result of agricultural entity recognition after entity stemming.

Tag	Name	Unique Entities	Root Entities	Precision	Recall	F-score
ANI	Animal	3,320	61	59.0%	70.6%	64.3%
CHE	Chemical	931	60	93.3%	94.9%	94.1%
CLI	Climate	1,243	47	80.9%	82.6%	81.7%
CRP	Crops	4,407	120	67.5%	66.9%	67.2%
DIS	Disease	913	49	91.8%	84.9%	88.2%
DST	Disaster	1,101	30	70.0%	72.4%	71.2%
FAM	Farm	5,938	92	16.3%	33.3%	21.9%
FER	Fertilizer	384	63	98.4%	98.4%	98.4%
FOD	Food	2,158	64	85.9%	84.6%	85.3%
MIO	Microorganism	477	11	100.0%	61.1%	75.9%
NUT	Nutrients	276	20	95.0%	95.0%	95.0%
TEM	Temperature	195	7	71.4%	71.4%	71.4%
	Total	**21,343**	**624**	**71.2%**	**76.9%**	**73.9%**

As a part of a news processing system, we carry out entity tags for collected agricultural news. The whole AgNews corpus is tagged AGER by our system and general NER by SpaCy toolkit (as mentioned in Section III-B). The tagged corpus can be used to extract and monitor agricultural entities by time and locations. For example, Figure 8 shows a number of news articles that have mentioned *Disaster:drought* entity from 2007 to 2018. As a result, documents with AGER tagging are useful for disaster management in agriculture.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a process for the agricultural entity recognition task with two stages, building an annotated dataset and then applying DL methods for identifying agricultural entities. Specifically, a hybrid approach in the first stage, which combines WordNet and Semantic Class methods to produce a highly accurate recognition system for building annotated corpus semi-automatically. This approach is suitable recognising entities in specific domains with a lack of resources. After identifying agricultural entities in the raw text, the corpus with 21, 000 articles has been reviewed and corrected manually to build an agricultural entity corpus for evaluation. Finally, our experiences based on DL approaches for the AGER task have shown high accuracy when evaluating on our AGER corpus.

In the future, we plan to collect a larger annotated agriculture corpus and design a general framework for information extraction and analysis, which includes agricultural entities and relationships between them. Moreover, the corpus also can be expanded with other resources in agriculture domain, such as scientific articles or farming manuals. This corpus will be made available for precision agriculture and ML communities. This will allow experts in the domain to enrich agricultural ontology, build linked data and knowledge maps, and provide precision agriculture with efficient management tools for crops, resources, etc.

ACKNOWLEDGMENT

This work forms part of CONSUS Project.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[2] A. Kumar, B. Payal, and A. Sharan, "Identifying crop specific named entities from agriculture domain using semantic vector," in *Proc. Inf. Syst. Design Intell. Appl.*, vol. 434. New Delhi, India: Springer, 2016, pp. 595–603.

[3] A. Rathod, N. Sinha, and M. P. Alappanavar, "Extraction of agricultural elements using unsupervised learning," *Imperial J. Interdiscipl. Res.*, vol. 2, no. 6, pp. 1–33 (2016).

[4] B. Magnini, M. Negri, R. Prevete, and H. Tanev, "A WordNet-based approach to named entities recognition," in *Proc. COLING SEMANET Building Using Semantic Netw.*, 2002, pp. 1–7.

[5] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, "NewsStand: A new view on news," in *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2008, p. 18.

[6] B. Payal, S. Aditi, and V. Sharad, "Named entity recognition for agriculture domain using WordNet," *Int. J. Comput. Math. Sci.*, vol. 5, no. 10, pp. 29–36, 2016.

[7] B. Payal, A. Sharan, and A. Kumar, "AGNER: Entity tagger in agriculture domain," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2015, pp. 1134–1138.

[8] C. S. Malarkodi, E. Lex, and S. L. Devi, "Named entity recognition for the agricultural domain," *Res. Comput. Sci.*, vol. 117, no. 1, pp. 121–132, Dec. 2016.

[9] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticæ Investigationes*, vol. 30, no. 1, pp. 3–26, Jan. 2007.

[10] E. F. T. Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, 2003, pp. 142–147.

[11] F. M. Alexander, "Information extraction—Lecture 3: Rule-based named entity recognition," in *Lecture Slides*, 2013, pp. 512–517. Accessed: Nov. 10, 2020. [Online]. Available: https://www.cis.uni-muenchen.de/~fraser/information_extraction_2015_lecture/

[12] J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva, "The Groningen meaning bank," in *Handbook Linguistic Annotation*. Dordrecht, The Netherlands: Springer, 2017, pp. 463–496.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[15] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. 1, pp. 180–182, 2003.

[16] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl. (JNLPBA)*, 2004, pp. 70–75.

[17] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.

[18] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 260–270.

[19] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2009, pp. 147–155.

[20] Y. Luo, F. Ying, and H. Zhao, "Hierarchical contextualized representation for named entity recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8441–8448.

[21] M. Song, W. C. Kim, D. Lee, G. E. Heo, and K. Y. Kang, "PKDE4J: Entity and relation extraction for public knowledge discovery," *J. Biomed. Informat.*, vol. 57, pp. 320–332, Oct. 2015.

[22] M. Tkachenko and A. Simanovsky, "Named entity recognition: Exploring features," in *Proc. 11th Conf. Natural Lang. Process. (KONVENS)*, 2012, pp. 118–127.

[23] M. K. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Mach. Learn. Appl. Int. J.*, vol. 3, no. 2, pp. 19–28, 2016.

[24] N. Chinchor and P. Robinson, "MUC-7 named entity task definition," in *Proc. 7th Conf. Message Understand.*, vol. 29, 1997, pp. 1–21.

[25] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q. H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, and M. Shigematsu, "BioCaster: Detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[26] N. Calzolari and S. Ananiadou, *Preliminary Recommendations on Lexical Semantic Encoding-Final Report*, document EAGLES LE3-4244, The EAGLES Lexicon Interest Group, 1999.

[27] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 2227–2237.

[28] Q. T. Tran, T. T. Pham, Q. H. Ngo, D. Dinh, and N. Collier, "Named entity recognition in Vietnamese documents," *Prog. Informat. J.*, vol. 5, pp. 14–17, Mar. 2007.

[29] Q. H. Ngo, N. A. Le-Khac, and T. Kechadi, "Ontology based approach for precision agriculture," in *Proc. Int. Conf. Multi-Disciplinary Trends Artif. Intell.*, Springer, 2018, pp. 175–186.

[30] S. Sekine, K. Sudo, and C. Nobata, "Extended named entity hierarchy," in *Proc. 3rd Int. Conf. Lang. Resour. Eval. (LREC)*, 2002, pp. 1–7.

[31] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 1064–1074.

[32] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6442–6454.

[33] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[34] W. Radford, X. Carreras, and J. Henderson, "Named entity recognition with document-specific KB tag gazetteers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 512–517.

QUOC HUNG NGO received the master's degree in computer science from the University of Science, Vietnam National University—Ho Chi Minh City (VNUHCM), Vietnam, in 2008. He is currently a Postgraduate Student in computer science at University College Dublin (UCD). He has involved in the BioCaster Project by building geographical ontology, integrating the geo-ontology into the Global Health Monitor system, and building the webpage for publishing project result.

TAHAR KECHADI received the master's and Ph.D. degrees in computer science from the University of Lille 1, France. He joined the UCD School of Computer Science (CS), in 1999. He is currently a Professor of computer science at CS, UCD. His research interests include data mining, distributed data mining heterogeneous distributed systems, grid and cloud computing, and digital forensics and cyber-crime investigations. He is a member of the *Communications of the ACM* journal and IEEE Computer Society. He is an Editorial Board Member of the journal of *Future Generation Computer Systems*.

NHIEN-AN LE-KHAC (Member, IEEE) received the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble (INPG), France, in 2006. He was a Research Fellow with Citibank, Ireland (Citi). He is currently a Lecturer with the School of Computer Science (CS), University College Dublin (UCD), Ireland. He is also the Programme Director of the UCD M.Sc. Programme in forensic computing and cybercrime investigation and an International Programme for the law enforcement officers specializing in cybercrime investigations. He is also the Co-Founder of the UCD-GNECB Postgraduate Certificate in fraud and e-crime investigation.