This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

Phenotype Extraction Based on Word Embedding to Sentence Embedding Cascaded Approach

Wenhui Xing, Xiaohui Yuan, Lin Li, Lun Hu, Jing Peng*
School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
*Corresponding Author: pengjing@whut.edu.cn

***Abstract*—As a significant determinant in the development of named entity recognition (NER), phenotypic descriptions are normally presented differently in biomedical literature with the use of complicated semantics. In this work, a novel approach has been proposed to identify plant phenotypes by adopting word embedding to sentence embedding cascaded approach. We make use of a word embedding method to find high-frequency phenotypes with original sentences used as input in a sentence embedding method. In doing so, a variety of complicated phenotypic expressions can be recognized accurately. Besides, the state-of-the-art word representation models have been compared and among them, Skip-Gram with Negative Sampling was selected with the best performance. To evaluate the performance of our approach, we applied it to the dataset composed of 56,748 PubMed abstracts of model organism *Arabidopsis thaliana*. The experiment results showed that our approach yielded the best performance, as it achieved a 2.588-fold increase in terms of the number of new phenotypic descriptions when compared to the original phenotype ontology.**

***Index Terms*—phenotype extraction, word embedding, sentence embedding**

I. Introduction

The volume of the biomedical literature is large and it continues to grow rapidly, for example, in 2016, a total of 869,666 citations were added to MEDLINE, and this is predicted to continue at a rate of 500,000 citations every year [1], thus making it important to mine the implicit knowledge in biomedical literature for creating research hypotheses. Recently, natural language processing (NLP) systems have been applied to various fields to extract relevant knowledge about such topics as protein-protein interactions [2]–[4], genes-diseases relationships [5], [6], and associations between genes, drugs and mutations [7], [8]. Named entity recognition (NER) is the foundation of all these applications. The goal of NER is simple: to find all the names and synonyms for a specific type of entity in a collection of texts. Because lexical features are relatively regular, the majority of research investigations have focused on gene or protein names [9], mutations [10], drugs, and diseases [11], [12]. Every NER system makes use of the following method: lexicon-based [13], rule-based [14], and machine learning [15]. Hybrid methods have also been used, such as ChemSpot [16] for recognizing chemicals in text; Settles [17] for identifying proteins, DNA, RNA, cell lines, and cell types and LSTM with CRF [18] system for extracting genes, chemicals and diseases.

A phenotype is the composite of an organism's observable morphologic, biochemical, physiological, and behavioral characteristics. Each phenotype is determined by the combined effects of genotype expression and environmental influences [19]. The fundamental goal of modern genetics is to establish genotype-phenotype correlations, which are often found in biomedical reports, but the sheer number of publications currently available warrants a reliable and automatic system to extract phenotypic information from the text.

Nevertheless, despite significant achievements in NER, the identification of all kinds of entities, i.e., phenotype, continues to be a challenge. First, a phenotype may be composed of multiple words of different lengths such as phenotype *'abiotic stress resistance'* or *'accumulation of anthocyanins in developing embryos'*. Thus, the problem is complicated by the need to draw a name boundary. To solve the problem, the *n-gram* model, which is based on the Markov hypothesis, may be used. This approach considers that the appearance of a word mainly depends on one or several words appearing in front of it. Actually, bigram and trigram are often used. However, parts of phenotypes require higher *n-gram* model to draw boundaries, which require larger corpora and higher time complexity, but with little improvement of accuracy. Another effective way is syntactic analysis. Because most of high-frequency phenotypic expressions are phrases, especially noun phrases. Syntactic analysis can annotate noun phrases and effectively solve the problem of dividing the boundaries of these specific phenotypic descriptions.

Second, due to the absence of standard expressions, phenotypic descriptions tend to be study- or author-specific, thereby making this task more difficult. For example, in the two sentences '...resulting in *root growth inhibition, smaller rosettes, and* ***leaf curling***.' (PMID: 26734017) and '...*forming a positive feedback loop with SEP3 and leading to early flowering and* ***curly leaves*** *phenotypes*.' (PMID: 25693187), the same leaf morphology uses two different descriptions, i.e., *'leaf curling'*, *'curly leaves'*.

Moreover, although many domain-specific lexicons are available, we are not aware of any lexicon that can be directly used in identifying comprehensive phenotypic descriptions in text, particularly for plants. For example, one famous vocabulary database is the Unified Medical Language System (UMLS) MetaThesaurus [20], which includes numerous semantic types, except for *Phenotype* type. In the plant domain, the controlled vocabulary Plant Trait Ontology (PTO)[1] is too

[1]http://bioportal.bioontology.org/ontologies/PTO

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

general and may not include all species traits. The Arabidopsis Information Resource (TAIR) [21] is manually curated by summarizing published reports and is thus limited and difficult to organize for future leverage. The AraPheno [22] database is an organization of the Genome-Wide Association Study (GWAS) phenotypic results in six published studies, and its amount of data is relatively small. These manual curation processes are time-consuming and do not account for the diversity in phenotypic expression that actually occurs in articles.

In the present study, we propose a novel approach using word embedding to sentence embedding cascaded approach to recognize various broad phenotypic information in the large-scale unlabeled biomolecular literature abstracts. We take advantage of word embedding to learn distributed representations for words or phrases, so we can locate sentences that include the phenotype. This method can extract some common phenotypic expressions that have been missing by ontology and expand the number of sentences that have the recognized phenotype. A better word representation model can effectively improves the phenotype recognition. The state-of-the-art word embedding models are "predict-based" models like Skip-Gram and CBOW using Negative Sampling or Hierarchical Softmax as optimized strategy, GloVe. We compared five types of word embedding models to choose the priority one i.e. Skip-Gram with Negative Sampling and embedded it into the cascaded approach.

Although we can extract most phenotypic descriptions using word embedding, it also has a limitation. For example:*'the floral meristems frequently develop additional whorls'* (PMID:20208065) is a sentence describing a phenotype. However, these sentence expressions can't be identified by the word embedding method which just encodes words or phrases. Thus, we proposed the use of the sentence embedding method for finding these specific phenotypic expressions. Although these phenotypes occur less frequently, their total number is large. We can further expand the number of phenotypic sentences as the word embedding supplement. To optimize the training model and enhance extraction of phenotypic diversity, additional input information is necessary, thereby we modified the unsupervised sentence embedding model to weakly supervised and used the proposed negative class label enhanced (NCLE) algorithm [41] to generate a different label for sentence to train.

Ultimately, we designed two baselines 1) dictionary-based method, 2) dictionary-based with word embedding method, to evaluate our cascaded approach performance. The results showed that our cascaded approach perform best at both integrity of recognition and recognition of the number, improving 2.588-fold compared with the original ontology dataset.

II. RELATED WORK

NLP researchers currently focus on identifying human phenotypes. Chen and Friedman [23] have transformed the MedLee system into BioMedLee to recognize phenotypic phrases as phenotypes in biomedical reports. The system used the NLP method to identify phenotypic information in article abstracts. As mentioned earlier, fewer lexicon terms can be used to describe phenotypes; therefore, they imported UMLS terms, Mammalian Ontology, and a few hundred manually settled terms. They tested 300 titles with 64% precision and a recall of 77.1%. Khordad et al. [24] employed a rule-based, semi-supervised machine learning method with two available resources, namely, MetaMap and Human Phenotype Ontology (HPO). The proposed system modified BANNER [25], an open-source biomedical NER system, by adding more features such as *phenotype, phenotype candidates, special modifier,* and *anatomy*. Even if their system had an F-Score of 92.25%, it was relatively time-consuming to conduct corpus annotation of 100 papers as a first step. Collier et al. [26] proposed a system that identifies human phenotypes using a hybrid approach of Hidden Markov Model (HMM), Conditional Random Fileds (CRFs), and knowledge-based methods but considering less semantics analysis with embedded entities. To our knowledge, no system is currently available for extracting plant phenotypic terms in the literature by text mining, thus preventing us from comparing this approach with other systems.

Word embedding collects name for a set of language modeling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers. One of the most important model is skip-gram model. This model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships [27]. Currently, this method has been used in the extraction of biological information, such as biomedical event trigger detection [28], drug-drug interaction extraction task and gene mention [29].

Sentence embedding [30] is proposed to give distributed representations for variable-length pieces of texts, ranging from sentences to documents. It has been widely and effectively used in various fields such as text classification [31], information retrieval [32], question answering [33], and query rewriting [34]. In addition, the paragraph vector is always used as a comparative approach [35], [36]. However, it has not been used in screening for phenotypic expressions.

To the best of our knowledge, the cascaded approach is the first method proposed for use in phenotype extraction. The approach considers both syntactic and semantic analysis of contextual structure without manual annotation and can identify more various phenotypic descriptions in texts.

III. METHOD

A. The Overview of Our Cascaded Approach

In this section, we introduce our word embedding to sentence embedding cascaded approach in extracting phenotypic descriptions from biomedical literature. Figure 1 shows the overview flowchart of our novel approach. We give a brief introduction as follow.

In Section III-B, as most phenotypic entities are noun phrases, while the general word vector representation methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

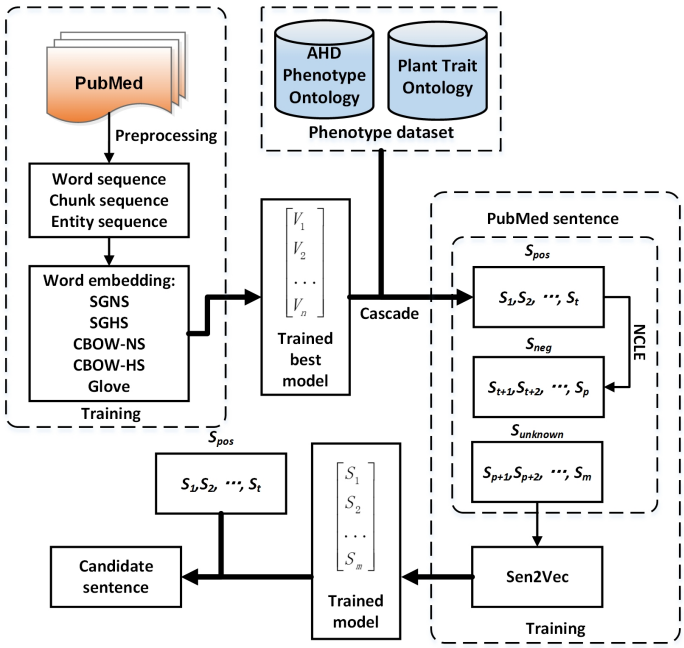

Fig. 1: Overview flowchart of our novel approach

always focus on words, which may miss some important information. Therefore, we make the chunks and ontology entities in phenotype dataset P as a parallel layer of word sequences to preprocess the corpus. Then we utilize five types of "count-based" word embedding models to explore the suitable one and then to find all phrases in texts that are similar to phrases in P. After calculating high similarity between vector of words in texts $\vec{v}$ and in P $\vec{v_p}$, we acquire new phenotypic descriptions to update the original P as P_{update}. Next, in Section III-C, to find more special phenotypic expressions that are not noun phrase, such as phenotypic description *'more root hair cells developing in positions normally occupied by non-hair cells.'*, we cascade output sentences of word embedding method as positive input S_{pos} to sentence embedding, while using proposed NCLE algorithm as negative input S_{neg}. $S_{unknown}$ are the rest of sentences to train Sen2Vec models. After training, we have all sentences vector $\vec{s}$, then calculate the most similar vector of sentences between vector in S_{pos} and $\vec{s}$ to find candidate phenotypic sentences. Sentences with similar template will aggregate together, thus by summarizing these templates, we can acquire more complicated and special phenotypic expressions, expanding the diversity of phenotypic recognition. The integrated cascaded approach is described in Algorithm 1.

Algorithm 1 Word embedding to sentence embedding cascaded approach

Input: S_{pub}: set of all sentences in PubMed, P: set of terms in original phenotypic dataset, W_{core}: set of high frequency words in P

Output: $P_{candidate}$: set of candidate phenotypes, P_{update}: set of updated original phenotypic dataset, $S_{candidate}$: set of candidate sentences, S_{update}: set of updated phenotypic sentences

$W = token(S_{pub}) = w_1, w_2, \ldots, w_n$
$P\_tree = parse(S_{pub}) = p_1, p_2, \ldots, p_n$
for w_i in W **do**
  if w_i *in* W_{core} **then**
    if *type of* $\{w_{i-j}, \ldots, w_i, \ldots, w_{i+k}\}$ *is* 'NP' **then**
      $joint(\{w_{i-j}, \ldots, w_i, \ldots, w_{i+k}\})$
    end if
  end if
  if $\{w_{i-n}, \ldots, w_i, \ldots, w_{i+t}\}$ *in* P **then**
    $joint(\{w_{i-n}, \ldots, w_i, \ldots, w_{i+t}\})$
  end if
end for
Train five word embedding models
Calculate similarity of each word in W *and each word in* P *as Similarity*
$Index = argsort(Similarity)$
$P_{candidate} = W[Index[n-9, n-8, .., n]]$
$P_{update} = P \cup evaluation(P_{candidate})$
Make sentences including term in P_{update} *as* S_{pos}
Use algorithm 2 to make S_{neg}
Train PV-DM/DBOW model using labeled S_{pub}
Calculate similarity of each sentence in S_{pos} *and each sentence in* S_{pub} *as* $Similarity'$
$Index' = argsort(Similarity')$
Select sentence in S_{pub} *as* $S_{candidate}$ *if* $Similarity' > Sim$
$S_{update} = S_{pos} \cup evaluation(S_{candidate})$

B. Word Embedding

In the cascaded approach, we first use the collected PubMed texts to train the word embedding model, which can give each word or phrase a distributed representation in dense and low dimensional vector space. By finding phrases that have high similarity to phenotypic entities, the original ontology of phenotype is expanded. Therefore, we can obtain larger numbers of sentences containing phenotypic information.

First, we make the domain ontology resources as phenotype dataset P. Many neural network embedding models like Skip-gram, CBOW use a sliding window of size k around the target word w, that means if $k = 3$, the contexts of the target word w are $w-3$, $w-2$, $w-1$, $w+1$, $w+2$, $w+3$. Note that a large number of phenotypes are phrases not single words in texts, a context window of tokens may miss some important information. Therefore we should preprocess the corpus before training:

Number and Character. Various types of numbers and punctuations in the text will bring noise to the training. So we replace all numbers with symbols "NBR". The punctuations like '( ) , ; . ! ? :' are replaced as space.

Chunk. A large number of phenotypes are made up of phrases, especially noun phrases, single tokens can not fully express the meaning of the phenotype. In order to enhance the ability of the model to identify phenotypes, we consider syntactic chunks (mainly focus on noun phrase) containing high frequency words (W_{core}) in P as a parallel layer of word sequences and joint these chunks as individual tokens.

Entity. The motivation of considering entity is similar with that of considering chunk. While the chunk in our model helps to understand the meaning of word sequences, ontology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

phenotype entities can provide fine-grained understanding of biomedical text. In the same way, we connect the phenotype entities in P together in the texts. An example of preprocessing is shown in Figure 2.

Then, we make tokenization of the processed texts, obtain a sequence of training words $w_1\ w_2\ w_3\ \ldots\ w_T$, to train word embedding model. The word embedding strategy can be divided two category: "count-based" representations like positive pointwise mutual information (PPMI), singular value decomposition (SVD); "prediction-based" embeddings like Skip-Gram, CBOW, GloVe models. The authors of [37] performed comprehensive evaluation , on a large number of lexical semantics tasks, proving that "prediction-based" models obtain better performance than "count-based" models. Thus, we consider to find the better one among three "prediction-based" word distributed representation models: Skip-Gram [38], CBOW [38], GloVe [39]. Because the first two models have two different optimization strategy 1) with Negative Sampling, 2) with Hierarchical Softmax [27], we train five different types of word embedding models to explore the most suitable one: Skip-Gram with Negative Sampling (SGNS), Skip-Gram with Hierarchical Softmax (SGHS), CBOW with Negative Sampling (CBOW-NS), CBOW with Hierarchical Softmax (CBOW-HS), GloVe.

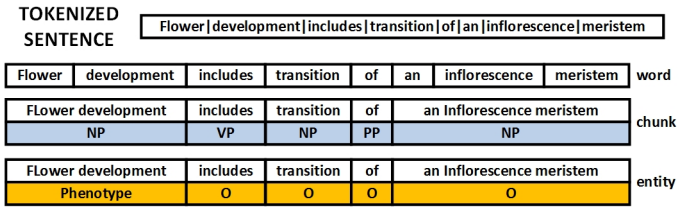

Fig. 2: An example (PMID: 1359429) of preprocessing the corpus. We make the words, chunks and entities as parallel sequence for word embedding training.

The **Skip-Gram and CBOW** model come from neural network language model which remove the non-linear hidden layer to minimize computational complexity. The architecture of these two models are very similar. CBOW predicts the current word based on the context as Eq. (1), Skip-Gram uses each current word as an input to the continuous projection layer, and predicts words within a window before and after the current word as Eq. (2). They can also use the Hierarchical Softmax and Negative Sampling as optimization of the probability formula Eq. (1), (2). The Hierarchical Softmax uses a binary Huffman tree representation of the output layer with the vocabulary words as its leaves and, for each node, explicitly expresses the probabilities of its child nodes. These assign probabilities to words by walk a random path.

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{t-k},\ldots,w_{t+k}) \tag{1}$$

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_{t-k},\ldots,w_{t+k}|w_t) \tag{2}$$

SGNS works by training two sets of embeddings: the word embeddings and context embeddings, in one certain k-dimensional vector space. The objective tries to make the co-occurrent word vectors closer than the randomly sampled "negative" word vectors [40]. The SGNS's objective is to maximize the objective function:

$$\sum_{t=1}^{|T|}\left[\sum_{c\in C(w_t)}\log p(c|w_t;\theta)\right] \tag{3}$$

where $C(w_t)=\{w_j, t-win\le j\le t+win\ and\ j\neq t\}$, win represents the window size, denotes the parameters to control. The Negative Sampling is used to approximate the conditional probability:

$$p(c|w)=\sigma(\vec{w}^T\vec{c})\prod_{i=1}^{k}E_{c_i\sim P_n(C)}\sigma(-\vec{w}^T\vec{c_i}) \tag{4}$$

where σ is sigmoid function, and k negative samples are for each data sample. $P_n(C)$ is the unigram distributions. Following this training, the word vectors are saved, and the context vectors are deleted.

GloVe learns word representation on the basis of $< word, context >$ co-occurrence matrix, which combines the advantages of the local context window method and global matrix factorization. GloVe seeks to best represent each word w_i and each context w_j as d-dimensional vectors $\vec{w_i}$ and $\vec{w_j}$ by minimizing the cost function of the weighted least squares regression model:

$$J=\sum_{i,j=1}^{V}f(X_{ij})(w_i^T\widetilde{w}_j+b_i+\widetilde{b}_j-\log X_{ij})^2 \tag{5}$$

where $b_i\ b_j$ are word/context-specific biases, the weighting function $f(X_{ij})$ can be set as:

$$f(x)=\begin{cases}(x/x_{max})^\alpha & \text{if } x<x_{max}\\ 1 & \text{otherwise}\end{cases} \tag{6}$$

All of these state-of-the-art embedding methods are applied for experiment and we can get different word representation vectors. Section IV-B will give comparison of the results. The better method is the first priority to embedded in the cascaded approach.

Next, we calculate the cosine similarity between vector of phenotype entities in P and vector of words or phrases in corpus as follows:

$$CosSim=\frac{\vec{P}\cdot\vec{Q}}{\|\vec{P}\|\cdot\|\vec{Q}\|} \tag{7}$$

where $\vec{P}$ and $\vec{Q}$ are phenotype entity vectors in the P and word vectors in corpus, respectively. $\|\vec{P}\|$ denotes the magnitude of vector $\vec{P}$ and "$\cdot$" is the vector dot product. After ranking the results, the top 10 similar phrases are considered as candidate phenotypes. After the evaluation by biomedical experts, we acquire new phenotypic descriptions to update the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

original P as P_{update}. Then we obtain phenotype-containing sentences as $S_{pos} = \{s_1, s_2, s_3, \ldots, s_i, \ldots, s_t\}, 1 \leq i \leq t$, and t is the amount of all phenotype sentences.

C. Sentence Embedding

Using the word embedding results, we automatically classify and tag PubMed texts as Sen2Vec input as shown in Figure 1. The trained model can find sentences containing phenotypic information, resulting in the acquisition of new phenotypic sentences. Because similar sentences have similar sentence templates, we can summarize the common sentence templates and extract new phenotypes again. More details are described in this section.

First, we map each word to a unique vector represented by a column in matrix W. We also map each sentence to a unique vector represented by a column in matrix D. Eq. (2) was used to train the Sen2Vec model. Because the Sen2Vec imports paragraph vector as input, we can give each sentence unique vector representation, which represents the missing information from the current context and acts as a memory of the topic of the paragraph.

To our knowledge, no positive and negative samples have been directly applied to phenotype extraction, particularly negative samples. Thus, to give more input information of training model, it was necessary to generate a large quantity of appropriate positive/negative samples. For the positive samples, we directly use the sentences resulting from the word embedding model as S_{pos}. For the negative samples, we adopt our proposed NCLE algorithm in [41], which can correct the similarity of the positive samples and generate the negative samples. More precisely, NCLE algorithm treats similarity score of positive samples as sentence score to label every positive sentences. Depending on the similarity mechanism, the higher scoring sentence is more likely to be judged by the model as phenotype-containing sentence. Therefore, the NCLE makes both the most dissimilar sentences of high score sentences and the medium similar sentences of relatively lower score sentences (controlled by a threshold) as negative samples.

Then, we use vector of S_{pos} and S_{pub} (all sentences) instead of $\vec{P}$ and $\vec{Q}$ in Eq. (7) for calculating the cosine similarity between S_{pos} and S_{pub} in corpus. After ranking the results, the sentences with a similarity of more than Sim are selected as candidate sentences for further expert verification. Sentence embedding is capable of aggregating similar expressions, thereby allowing us to further mine phenotypic sentence templates identify new phenotypes with the help of biologists.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Crawling PubMed Paper Abstracts

We conducted experiments on the model organism *Arabidopsis thaliana* for phenotype extraction. The experiment started with crawling all abstracts in PubMed with the keyword *'Arabidopsis thaliana'* using the Entrez Programming Utilities (E-utilities) web service[2]. We cleaned the irrelevant author information and ultimately acquire 56,748 abstracts as total cited literature evidence that mention *A. thaliana*.

Constructing the Phenotype Dataset

Two ontologies constituted our original phenotype dataset P, i.e., PTO and Arabidopsis Hormone Database 2.0 [42]. PTO is an important controlled vocabulary that describes phenotypic traits in plants. Each trait is a distinguishable feature, characteristic, quality, or phenotypic feature of a developing or mature plant or a plant part. Arabidopsis Hormone Database 2.0 is an updated version of Arabidopsis Hormone Database for plant systematic studies. It provides a systematic and comprehensive view of genes participating in plant hormonal regulation of the model organism *A. thaliana*. Its phenotype ontology is developed to precisely describe myriad hormone-regulated morphological processes with standardized vocabularies in *Arabidopsis*.

When processing PTO, we extracted 'name' and 'synonym' from every term in the ontology. Approximately 84% of all these names are associated with synonyms; on average, each of these names has 1.07 synonyms. For example, the phenotype 'nitrogen sensitivity' has two synonyms: 'NISN' and 'nitrogen deficiency'. Not all of terms in these ontologies appear in literature. A total of 805 terms were found in abstracts after removing duplicate entries. We combined these into a complete phenotype dataset P.

B. Experimental Results and Discussion

Word Embedding Results and Analysis

First, we preprocessed the corpus to join certain noun phrases as individual tokens using Stanford Parser[3] to create a parse tree of every sentence in all abstracts. After test, if a term in P appeared bigger than four times, we reserved it as vocabulary of W_{core}, which can retain sufficient information. Ultimately, W_{core} contained 1052 available words. Then, we compared the five model results and explored the suitable one to embedded in the whole system. We used Word2Vec[4] to train the skip-gram, CBOW methods,and set hyperparameter $window$ as 10, and the $negative$ as 10, used glove[5] to train GloVe method, and the dimensional size of all methods was set as 300.

In the biological text mining, due to the lack of complete golden-standards, researchers often need expert evaluation or annotation [1], [24], [43], [44]. The word embedding results of the semi-automatic classification and biological expert evaluation are shown in Table I. The "candidate phenotype" means the whole results of four dimensional model. The "matching dictionary phenotype" represents that phrases are the term in original ontology P. The "new phenotypic description" are the new phenotype not appearing in ontology P. The quantity of the above two constitutes the number of "all phenotype". "Sim.

[2]https://www.ncbi.nlm.nih.gov/books/NBK25501/
[3]http://nlp.stanford.edu/software/lex-parser.shtml
[4]https://code.google.com/p/word2vec/
[5]https://nlp.stanford.edu/projects/glove/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

Range" represents the similarity range compared original ontology terms with all phenotypic phrases.

From the results of Table 1, it can be seen that the SGNS model can identify the most new phenotypes (668), which was 303 more than the worst-performing CBOW-HS model. And the SGNS model obtained the largest number of total phenotypes (791) due to significant new phenotype recognition effects. From the range of similarity distribution, SGNS also had a relatively uniform and high similarity value (0.577 $\sim$ 0.923). However, the similarity value of the other models failed to exceed 0.900. GloVe also preformed well and was second only to SGNS, which can get 543 new phenotypes, but the similarity value was relatively low. In terms of the recognition effect, the length of the phrase identified by the SGNS was longer and averagely exceeds 2.5, such as 'broad-spectrum disease resistance'. However, GloVe and other models recognized an average phrase length no more than 2.1, which explained that SGNS can identify more complete phenotypic phrases.

From the table, we can also find that the model of using the Negative Sampling strategy was superior to the Hierarchical Softmax strategy in both the number of identification and similarity distribution. This may be due to that Negative Sampling defines negative samples and has better discrimination between similar words and noise words.

In general, the SGNS model is more suitable for phenotypic word vector representation. Therefore, we embedded SGNS into the entire approach as the preferred word embedding model. In order to maximize the phenotype recognition of the word embedding method, we trained two additional dimensional model, i.e. 500, 700 for the SGNS, to enhance the vector representation capabilities.

Ultimately, the word embedding method can extend original phenotype datasets, increasing the number of 899 new phenotype data by up to 1.117-fold. We used the extended dataset P_{update} to match the phenotypic descriptions in the abstracts. The total number of mapping sentences was 89,112.

The word embedding method can effectively find a phenotypic description according to the syntax and context of the text. Some examples of the results are given in Table II. Examples of 1-4 were type a, indicating that these candidates are new phenotypes and not in ontology. Ex. 1 and Ex. 2 come from the same ontology word *'leaf curling'* (TO:0002681). The word embedding method can extract similar words by not only considering syntax (*'leaf curling' - 'curly leaves'*), but also context semantics (*'leaf curling' - 'altered leaf shape'*).

Expert analysis showed that a large number of candidates can partly describe phenotypes, but not completely. We designated these as 'partial phenotypic descriptions' and were indicated as type c in Table II. For example, *'reduced size'* in Ex. 7 does not represent a complete phenotype. After adding description words, *'one or more petals with a reduced size'*, renders a more specific phenotype. This situation may be due to the fact that the dividing words boundary method does not include all cases. However, the sentence embedding method can find this whole sentence to locate the phenotype. Thus the complete description of Ex. 7 can be extracted using sentence embedding.

In addition, the effect of word embedding is barely satisfactory in dealing with long, non-phrase phenotype. For example, the phenotype in the sentence "Axillary meristems allow the ***production of secondary growth axes in the shoot systems of plants***"(PMID: 16724256), is more than a phrase and more like a short sentence. It is thus reasonable to use sentence level representation for this particular phenotype.

Sentence Embedding Results and Analysis

The sentence embedding method can be implemented by the PV-DW/DBOW model, thus we utilized Doc2Vec[6] to train the models, using labeled and pure abstracts to generate four models, namely:

1) Model A: PV-DM model with pure abstracts as input
2) Model B: PV-DM model with labeled abstracts as input
3) Model C: PV-DBOW model with pure abstracts as input
4) Model D: PV-DBOW model with labeled abstracts as input

Then, we used the results of word embedding S_{pos} as inputs and acquired candidate sentences with similarities greater than Sim after calculating for cosine distance with S_{pub}.

A reasonable Sim value has a great impact on the results. First, we analyzed the top 10 most similar sentences of $s_i \in S_{pos}$, which showed that the extracted sentence similarity was generally low, and the similarity between the sentence s_i and itself ranged from 0.45 to 0.72, mostly around 0.5. A similarity of the last sentence in the top 10 most similar sentences was about 0.2. Hence, it is necessary to set a fine-grained Sim, not just using top 10 as threshold.

Setting Sim as 0.4, we extracted the similar sentence of s_i. The number of extracted sentences substantially decreased. On average, each s_i had 1.2 candidate sentences, and the number of candidates was too few. Adjusting the Sim to 0.3, we can get 4.5 candidate sentences for each s_i, and the similarity of the last sentence in the ordered list was above 0.32, with most of them $>$ 0.35. Changing the Sim to 0.2, there were 7.4 candidate sentences for each s_i and the similarity became very low. Therefore, we set parameter Sim as 0.3, which can guarantee suitable similarity without substantially losing information.

After the experiment, the number of sentences in each model was more than 100,000 after removing the repeated sentences and some positive samples. Then, 700 sentences were randomly selected, and biological experts manually identified the phenotype. The precision of each model is shown in Table III.

From the data in the table, we can see that the C and D models were relatively better. However, model C only consisted of 7 sentences in S_{pos} and less than 4,284 sentences in S_{pos} of model D. It showed that the model with the labeled sentences can find more sentences of unknown types, thereby increasing the diversity of phenotypic sentences and identifying new phenotypic sentences. For example, model C could find three phenotypic sentences using the original sentence *"... those*

[6]http://radimrehurek.com/gensim/models/doc2vec.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

TABLE I. The five word embedding models results comparison for extracting *A. thaliana* phenotypes

Model	SGHS	CBOW-NS	CBOW-HS	GloVe	SGNS
No. of Candidate Phenotype	1609	1677	1626	1857	**1795**
No. of All Phenotype	535	586	491	669	**791**
No. of Matching Dictionary Phenotype	138	134	126	126	**123**
No. of New Phenotypic Description	397	452	365	543	**668**
Sim Range	.423 $\sim$ .772	.471 $\sim$ .875	.272 $\sim$ .700	.285 $\sim$ .637	**.577 $\sim$ .923**

TABLE II. Example of candidate phenotypes with their type and sentence. There are three types of candidates: *a* means the candidate is a phenotype and not an ontology term; *b* means the candidate is a phenotype and an ontology term; *c* is partial phenotypic description.

Candidate Phenotype	Type	Sentence	PMID
1. Curly leaves	*a*	...which exhibits constitutive BR response phenotypes including long and bending petioles, **curly leaves**, ...	12007405
2. Altered leaf shape	*a*	We showed that the **altered leaf shape** is caused by reduced cell proliferation ...	15125775
3. Pale green leaves	*a*	In contrast, weak (fc2-1) and null (fc2-2) mutants of FC2 showed **pale green leaves** ...	27630653
4. Increased anthocyanin accumulation	*a*	Over-expression of UGT78G1 in transgenic alfalfa resulted in **increased anthocyanin accumulation** when plants were exposed to abiotic stress.	19368693
5. Narrow leaves TO:0001013	*b*	...mutant yuc7-1D exhibited phenotypic changes ...such as tall, slender stems and curled, **narrow leaves**.	22109847
6. Floral organ size TO: 0002600	*b*	...caused significant increases in cell expansion that could explain the differences in **floral organ size**.	24985495
7. Reduced Size	*c*	In the mutant, the numbers of petals and stamens are reduced, and many flowers have **one or more petals with a reduced size.**	10528262

induced by other stimuli such as environmental or biotic stress." (PMID:10066585), whereas model D only detected the first one:

1) *"These analyses demonstrate that **ABA-related stress responses** are modulated..."* PMID: 18552355
2) *" **Plants regulate growth** and **respond to environmental stress through abscisic acid (ABA)** regulated pathways..."* PMID: 23290725
3) *"...**cellular differentiation** and **responses to different extracellular stress stimuli**."* PMID: 22631074

Some examples of specific phenotypes extracted by the sentence embedding method are given in Table IV. Similar to the examples in Type A, parts of phenotypes are not only described in simple phrase but also in short sentences, which probably include punctuations, prepositions, and conjunctions. In Type B, authors use a complete sentence to describe the phenotype, and the sentence may involve some environmental and temporal conditions that lead to the phenotype. In Ex.1, the phrase mixes up three phenotypes. The phenotypes in the *"response to ... stress"* template are usually described together in the literature, but the descriptions tend to be study-/author-specific. We further summarize these special descriptions of a sentence template to expand phenotype recognition.

With the help of biological experts, some sentence templates that have been used to describe phenotypes have been summed up, and five of these are shown in Table V. Some of the templates are not directly represented as phenotypes, but the expression is always used in describing phenotypes. For example, the *"Play ... role in + [phrase]"* template is always used to describe the event that some gene/protein/mutation has an effect on phenotype formation. Therefore, it contains phenotypic descriptions. After using these morphological patterns to match the phenotype in the literature, 28,159 phenotypic descriptions were obtained. For each sentence pattern, 200 samples were randomly selected for expert verification, and the results showed that precision of each template was $>$ 80%. At the same time, we also extracted some patterns that can be used to express relationship between genotype and phenotype such as *"overexpression of [genotype] resulted in [phenotype]"*. This is very useful for further identification of relationships between genotype and phenotype.

TABLE III. Number of sentences and precision of the four sentence embedding models

Model	Label Enhanced	No. of Sentences	Precision
(A) PV-DM	NCLE	186,346	23.5%
(B) PV-DM	-	183,923	22.1%
(C) PV-DBOW	**NCLE**	**145,287**	**45.5%**
(D) PV-DBOW	-	133,079	37.5%

The Cascade Approach Results and Analysis

In order to illustrate the effectiveness of our approach in phenotype identification, we designed two baselines. 1) dictionary-based traditional method. 2) dictionary-based combined with word embedding method. The comparison results of the three methods are shown in the Table VI.

From the table we can find that, the sentence embedding method can complement the shortcomings of word embedding method, both the number of recognition and phenotypic diversity have greatly improved. Due to the excessive number of candidate sentences from sentence embedding, we currently can only check some of the random samples. Nevertheless, the number of new phenotypic descriptions significantly increased by 2.588-fold compared to the original phenotype dataset after evaluating 1.88% of candidate phenotypes. However, the two baselines did not perform well for both the recognition numbers and the integrity of the recognition, so our cascaded approach achieved the best results.

V. CONCLUSIONS AND FUTURE WORK

A large quantity of plant phenotypic information currently exists in the biomedical literature and continues to grow. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2838137, IEEE Transactions on NanoBioscience

TABLE IV. Examples of sentences containing phenotypic information using the sentence embedding method. Types A and B stand for phenotypes described by short sentences and long sentences, respectively.

Example	Type	Sentence	PMID
1	A	GmaPHO1 genes had altered expression **in response to salt, osmotic, and inorganic phosphate stresses**.	23705930
2	A	Plants deficient in phytochrome B (phyB) exhibit a **constitutive shade avoidance syndrome including reduced branching**.	24492336
3	A	Transgenic plants overexpressing TCP2 displayed a **light-dependent short hypocotyl phenotype**, especially in response to blue light.	26596765
4	B	In strong ap3 and pi mutants, **petals and stamens are missing and sepals and carpels develop in their place**.	12943551
5	B	In gor, **the shoot meristem enlarges continuously during post-embryonic development and the floral meristems frequently develop additional whorls**.	20208065
6	B	**The leaf content of six free amino acids was reduced in the leaf tissue of diseased A**.	26381754

TABLE V. Examples of sentence template for phenotypic description

Sentence Template	Example	No. of sentence	Precision
Abnormal + [phrase]	Abnormal leaf size and shape Abnormal chloroplast phenotype	694	83%
Reduced + [phrase]	Reduced primary seed dormancy and thermoinhibition Reduced rate of flower formation on lateral	1629	89%
Play ...role in + [phrase]	Plays an important role in pollen tube tip growth Plays a major role in early seedling development	5886	80%
Essential for + [phrase]	Essential for pollen germination and pollen tube growth Essential for salt-stress signaling and tolerance in Arabidopsis	2393	87%
Development of + [phrase]	Development of the seed coat in plants Development of the early flower bud stages	2258	82%

TABLE VI. Performance of Different Method

Method	Number of Phenotype	Number of Sentence	Multiple of Improvement
Dictionary-based [13]	805	69,115	-
Word embedding [38]	1704	89,112	1.117-fold
Cascaded approach	**2083**	**93,432**	**2.588-fold**

propose a cascaded approach to extract phenotypic descriptions from the biomedical literature using *Arabidopsis thaliana* as an experimental object. We compare the state-of-the-art five word embedding methods and chose the best one i.e. Skip-Gram with Negative Sampling to embedded in cascaded approach. Using the proposed approach, we have extended the original phenotype ontology by 2.588-fold.

A possible future direction is the addition of larger corpora in our cascaded approach to identify more phenotypes as well as entities. If trained domain experts develop golden standards in the future, it will be more conducive to train and evaluate the model. Using advanced deep learning methods, such as the Recurrent Neural Networks (RNNs) method, with the results of the cascaded approach as annotation could further expand the recognition ability. The extraction of additional phenotypes may facilitate the establishment of genotype-phenotype relationships.

VI. Acknowledgements

This research project is supported by the National Natural Science Foundation of China (31701144), the Fundamental Research Funds for the Central Universities (WUT: 2017II39GX).

References

[1] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[2] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, and I. Iliopoulos, "Protein–protein interaction predictions using text mining methods," *Methods*, vol. 74, pp. 47–53, 2015.

[3] Z. Yang, N. Tang, X. Zhang, H. Lin, Y. Li, and Z. Yang, "Multiple kernel learning in protein–protein interaction extraction from biomedical literature," *Artificial intelligence in medicine*, vol. 51, no. 3, pp. 163–173, 2011.

[4] F. Zhu, Q. Liu, X. Zhang, and B. Shen, "Protein-protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer," in *Trustcom/BigDataSE/ISPA, 2015 IEEE*, vol. 1. IEEE, 2015, pp. 1306–1311.

[5] A. Coulet, N. H. Shah, Y. Garten, M. Musen, and R. B. Altman, "Using text to build semantic networks for pharmacogenomics," *Journal of biomedical informatics*, vol. 43, no. 6, pp. 1009–1019, 2010.

[6] J. Kim, J.-j. Kim, and H. Lee, "An analysis of disease-gene relationship from medline abstracts by digsee," *Scientific Reports*, vol. 7, 2017.

[7] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, "Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W399–W405, 2008.

[8] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, "Edgar: extraction of drugs, genes and relations from the biomedical literature," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 2000, p. 517.

[9] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, and P. Willett, "Protein structures and information extraction from biological texts: the pasta system," *Bioinformatics*, vol. 19, no. 1, pp. 135–143, 2003.

[10] F. Horn, A. L. Lau, and F. E. Cohen, "Automated extraction of mutation data from the literature: application of mutext to g protein-coupled receptors and nuclear hormone receptors," *Bioinformatics*, vol. 20, no. 4, pp. 557–568, 2004.

[11] I. Segura-Bedmar, P. Martínez, and M. Segura-Bedmar, "Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems," *Drug discovery today*, vol. 13, no. 17, pp. 816–823, 2008.

[12] R. Xu, K. S. Supekar, A. Morgan, A. K. Das, and A. M. Garber, "Unsupervised method for automatic construction of a disease dictionary from a large free text collection." in *AMIA*, 2008, pp. 820–824.

[13] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using blast for identifying gene and protein names in journal articles," *Gene*, vol. 259, no. 1, pp. 245–252, 2000.

[14] K.-i. Fukuda, T. Tsunoda, A. Tamura, T. Takagi *et al.*, "Toward information extraction: identifying protein names from biological papers," in *Pac symp biocomput*, vol. 707, no. 18, 1998, pp. 707–718.

[15] C. Nobata, N. Collier, and J.-i. Tsujii, "Automatic term identification and classification in biology texts," in *Proc. of the 5th NLPRS*, 1999, pp. 369–374.

[16] T. Rocktäschel, M. Weidlich, and U. Leser, "Chemspot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[17] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. Association for Computational Linguistics, 2004, pp. 104–107.

[18] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[19] N. Freimer and C. Sabatti, "The human phenome project," *Nature genetics*, vol. 34, no. 1, pp. 15–21, 2003.

[20] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett, "The unified medical language system," *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 1–11, 1998.

[21] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez *et al.*, "The arabidopsis information resource (tair): improved gene annotation and new tools," *Nucleic acids research*, vol. 40, no. D1, pp. D1202–D1210, 2012.

[22] Ü. Seren, D. Grimm, J. Fitz, D. Weigel, M. Nordborg, K. Borgwardt, and A. Korte, "Arapheno: a public database for arabidopsis thaliana phenotypes," *Nucleic Acids Research*, vol. 45, no. D1, pp. D1054–D1059, 2017.

[23] L. Chen and C. Friedman, "Extracting phenotypic information from the literature via natural language processing," *Medinfo*, vol. 11, no. Pt 2, pp. 758–762, 2004.

[24] M. Khordad, R. E. Mercer, and P. Rogan, "A machine learning approach for phenotype name recognition," in *Proceedings of COLING*, 2012, pp. 1425–1440.

[25] R. Leaman, G. Gonzalez *et al.*, "Banner: an executable survey of advances in biomedical named entity recognition." in *Pacific symposium on biocomputing*, vol. 13, 2008, pp. 652–663.

[26] N. Collier, M.-V. Tran, H.-Q. Le, A. Oellrich, A. Kawazoe, M. Hall-May, and D. Rebholz-Schuhmann, "A hybrid approach to finding phenotype candidates in genetic texts," *Proceedings of COLING 2012*, pp. 647–662, 2012.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[28] J. Wang, J. Zhang, Y. An, H. Lin, Z. Yang, Y. Zhang, and Y. Sun, "Biomedical event trigger detection by dependency-based word embedding," *BMC medical genomics*, vol. 9, no. 2, p. 45, 2016.

[29] Z. Jiang, L. Li, D. Huang, and L. Jin, "Training word embeddings for deep learning in biomedical text mining tasks," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 625–628.

[30] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[31] M. Iyyer, V. Manjunatha, J. L. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification." in *ACL (1)*, 2015, pp. 1681–1691.

[32] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 694–707, 2016.

[33] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4622–4630.

[34] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, and N. Bhamidipati, "Context-and content-aware embeddings for query rewriting in sponsored search," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 383–392.

[35] J. Tan, X. Wan, and J. Xiao, "A neural network approach to quote recommendation in writings," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 65–74.

[36] S. Wang, J. Tang, C. Aggarwal, and H. Liu, "Linked document embedding for classification," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 115–124.

[37] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 238–247.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[39] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[40] D. Mimno and L. Thompson, "The strange geometry of skip-gram with negative sampling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2873–2878.

[41] W. Xing, X. Yuan, L. Li, and J. Peng, "Cascade word embedding to sentence embedding: A class label enhanced approach to phenotype extraction," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 477–484.

[42] Z. Jiang, X. Liu, Z. Peng, Y. Wan, Y. Ji, W. He, W. Wan, J. Luo, and H. Guo, "Ahd2. 0: an update version of arabidopsis hormone database for plant systematic studies," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1123–D1129, 2011.

[43] J. Mao, L. R. Moore, C. E. Blank, E. H.-H. Wu, M. Ackerman, S. Ranade, and H. Cui, "Microbial phenomics information extractor (micropie): a natural language processing tool for the automated acquisition of prokaryotic phenotypic characters from text sources," *BMC bioinformatics*, vol. 17, no. 1, p. 528, 2016.

[44] S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg, "Brenda in 2017: new perspectives and new tools in brenda," *Nucleic acids research*, vol. 45, no. D1, pp. D380–D388, 2017.