SPRENO: A BioC Module for Recognizing and Normalizing Species and Their Model Organisms

A Species recognizer, identify mentioned species name in figure captions

Onkar Singh[1,2], Hong-Jie Dai[3,4]*

[1] Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan, R.O.C.

[2]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, 112, Taiwan, R.O.C.

[3]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C.

[4]Interdisciplinary Program of Green and Information Technology, National Taitung University, Taitung, Taiwan, R.O.C.

***Abstract*—**Name entity recognition is a key step in a biomedical text mining task. This becomes more critical and challenging due to the availability of huge amount of biomedical literature. To recognize and identify species become difficult for the domain experts due to the vagueness of the abbreviated term used for model organisms/strains. In this study, we present our species recognition tool−SPRENO (Species Recognition and Normalization) developed for recognizing organism terms mentioned in figure captions. SPRENO is an extension of our previous species recognition tool developed for the BioCreative V BioC task. We developed new algorithms optimized for normalizing organism terms mentioned in figure captions, which consider the contextual information from the corresponding full text. Furthermore, two disambiguation methods are developed to determine the ID of ambiguous organism mentions. One is based on the majority rule to select the ID that has been successfully linked to previously mentioned organism terms. Another is a convolutional neural network model trained by learning the context and the distance information of the target organism mention. We participated the BioCreative VI BioID task and submitted three runs for the assessment of the developed tool. The best micro F-scores achieved by SPRENO on the test set are 0.776 (entity recognition) and 0.755 (entity normalization).

Keywords—Named entity recognition, organism normalization, convolutional neural network

I. INTRODUCTION

The unprecedented growth in biomedical literature necessitates perpetual reformations of automated text mining tools which can correctly extract individual or multiple biomedical entities (e.g. gene/gene products, organism etc.) and transform orderly. However, the complexity of the dynamically changing terminology for the same bio-entity has emerged as a challenging task for the bio-curators. Recognizing bio-entities manually demonstrates high detection accuracy but is time-consuming and labor intensive. It provides us a lot of scope for the researchers to develop automated annotation tools.

The primary task of biomedical text mining is named entity recognition (NER) and normalization of the entity. NER tools are developed to ascertain biomedical entities such as the mentioned species, gene and gene products in biomedical literature. To facilitate downstream tasks, it is very important to accurately recognize those entities and associate them with their corresponding database/ontology IDs [1]. As one of the participants in the BioCreative VI Bio-ID assignment track, we extended our previous species recognition tool [2] for recognizing organism terms mentioned in figure captions. Comparing with the recognition of species terms mentioned in abstracts or full texts, which have been studied in previous works [3, 4], the process of recognizing terms described in figure captions is challenging owing to the absence of specific criterion, unique terminology, and unexplained abbreviated words. The ambiguous nature of the abbreviated terms requires a strategy to process full text to find the full term which makes exceedingly tough for the domain expert to identify organisms and link them to their unique taxonomy IDs. For example, the abbreviated term SIN may stand for Sindbis virus or the same term in gene ontology referring to Sex-lethal interactor gene (*Drosophila melanogaster*).

Another challenge that was encountered during the recognition of organism terms in the BioID task was the identification of strains. Authors use specialized terms of strains/model in figure captions to describe their experimental observations. For example, the terms of inbred strains of the mouse include C57BL/6J, R6/2, DBA/2J etc. Aside from the organism tagger developed by Naderi, Kappler, Baker and Witte [5], those strain mentions cannot be recognized by most of the current openly available tools.

In this paper, we present our new species recognition tool, SPRENO (Species Recognition and Normalization). We extended the lexicon used by our previous species recognition tool by including organism terms and common terms used to refer to strains or models. We also developed new algorithms optimized for normalizing organism terms mentioned in figure captions, which consider the contextual information from the corresponding full text. Finally, disambiguation methods based on the majority rule and the convolutional neural network (CNN) were developed to determine the ID of ambiguous organism mentions.

II. METHOD

A. Lexicon Extended with Terms of Strain/model

As mentioned in the previous section, the lexicon used by our previous species recognition tool only includes species terms, as well as prefixes in a gene name, which can refer to the species. At present work we extended the lexicon by adding the terms used for organisms, common terms and abbreviations used for strains/model organism. Table 1 summarizes the resources used in this work.

TABLE I. RESOURCES FOR STRAIN/MODEL TERMS

Source	Organism
http://www.informatics.jax.org/inbred_strains/mouse/STRAINS.shtml	Mouse
http://www.criver.com/find-a-model	Mouse
http://gcm.wfcc.info/speciesPage.jsp?strain_name=Lactobacillus%20acidophilus#specTopgcm.wfcc.info	Lactobacillus acidophilus
https://gold.jgi.doe.gov/organisms?Organism.Domain=BACTERIAL&Organism.Type%20Strain=Yes&Organism.Active=Yes	Bacteria
https://byo.com/resources/yeast	Yeast

B. Normalization and Disambiguation Approach

Fig. 1 demonstrates the workflow of the developed organism recognizer. We extended our previous BioC library[1] to support the process of figure captions represented in the Bio C format defined by the Bio-ID task. The developed library is then used to load the figure captions and their corresponding full-text article. Although the goal of Bio-ID task is to identify the organism terms in figure captions only, we still process both the full text and figure captions. Therefore both the full text and figure captions are preprocessed to detect sentence boundaries, tokens, part-of-speech (PoS) tags and full name-abbreviation pairs.

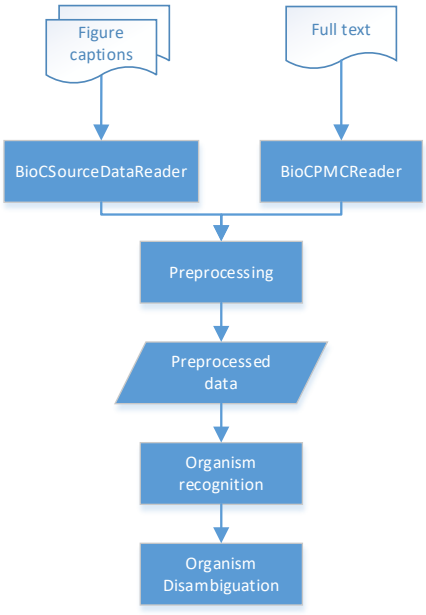

Fig. 1. Workflow of the developed organism recognizer.

When processing the full text, the base forms of the full names listed as the full name-abbreviation pairs found from the entire full-text are matched with our lexicon before performing the actual organism recognition process. If the full name is considered to be an organism term, both the name and its abbreviation are added to the organism lexicon for the one-off matching of the given full-text. Otherwise, the pairs are blacklisted for the current full-text. We then process the text of figure captions based on the updated lexicon. Through this way, we can reduce the ambiguity of abbreviated terms frequently used in figure captions. Finally, the algorithm developed in our previous work [2] was employed to recognize all organism terms from the full text by exploiting linguistic information and match against the extended dictionary. After identifying all of the organism candidates, the PoS information is used to filter out false-positive cases such as candidates with PoS as a verb.

In order to reduce the ambiguity of the recognized organisms in figure captions, we applied two disambiguation methods. The first is a rule-based approach which uses the normalization information from full text. The algorithm follows the similar idea of our multi-stage normalization algorithm [6] to disambiguate ambiguous organism terms by exploiting the normalization information from the entire article. The majority rule is used to select the ID that has been successfully linked to previously mentioned organism terms.

The second disambiguation method is a machine learning based method based on CNN. We formulated the disambiguation problem as a binary classification task and generated the training set based on the outcome of our organism recognizer. The generated training set includes both, the normalized terms and ambiguous terms along with the candidate IDs as well as their context in figure captions. Fig. 2 shows an overview of the developed CNN model for organism disambiguation.

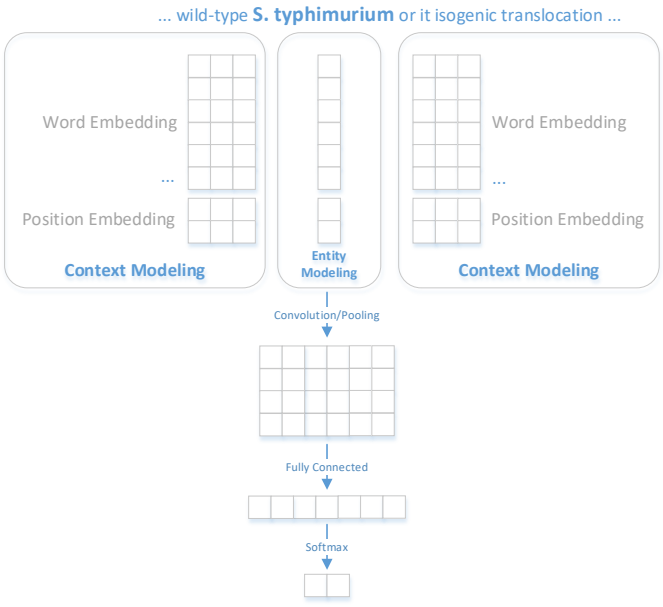

Fig. 2. The developed neural network model for organism disambiguation.

[1] https://www.nuget.org/packages/NTTU.BigODM.Bio.BioC/

The input of our model includes the context of the mention and the mention's candidate record from the NCBI taxonomy database. The output of our CNN is the probability distribution over two possible outcomes {yes, no}. The context is represented by word representations and the distance between a context word and the target mention based on the consideration that a closer context word might be more informative than a farther one for disambiguating that mention [7]. For the entity modeling, the target entity was represented as the official symbol recorded in the taxonomy database.

The pre-trained PubMed word embedding released by Moen and Ananiadou [8] was used for representing the words. Therefore, the dimension of word vector was set as 200. The learning rate and the window size of CNN were empirically set as 0.01 and 2, respectively.

C. Common Term Recognition and Normalization

We observed that in the Bio-ID corpus, the annotators tend to annotate common terms like larvae and embryos for organisms such as *Drosophila melanogaster* and *Danio rerio* (zebrafish) depending on the context mentioned in the full-text article or figure captions. For example, in Figure 1, the term "embryo" indicates the species zebrafish. We analyzed the training set of Bio-ID corpus to collect all possible grounding IDs for a given common term.

> (A) Yolk opaqueness and LC3 puncta formation in spns1-mutant **zebrafish embryos**. For EGFP-LC3 transgenic spns1-mutant [Tg(CMV:EGFP-LC3);… and spns1 mutant (spns1−/−) (lower) **embryos** at 84 hpf are shown.

Fig. 3. An example text indicates the common term (embryos) and its related species (zebrafish).

When processing a given article, if our tool detects an occurrence of the considered common terms, an algorithm was developed to select an ID from the term's possible IDs. The selection strategy is designed as a way to select the ID which appeared most frequently in the entire full text.

III. Results

We submitted three runs to assess the performance of the developed organism recognizer module. At first run, the rule-based normalization method utilizing full-text information only was applied. We set a threshold at two to filter those organism names assigned for more than two IDs after the disambiguation process. In the second run, the threshold for filtering increased to 10, i.e. those organism mentions having more than 10 IDs were filtered out. After that, we employed the developed CNN model to select the ID with the highest likelihood. At last, we performed third run where we used the original lexicon from our previous work along with the CNN-based disambiguation. The threshold for the last run was set to infinite.

Table II shows the recognition results for organism on the test set of the Bio-ID task. Two matching criteria are used. The strict span condition depicts that annotator annotates the term "zebrafish" and another annotator annotates "spns1-mutant zebrafish" then the match function will fail to identify. While in overlap situation the term is considered to be a match.

As shown in Table II, the first run achieves the best recall (R) and F-score among others. The second run with the developed CNN filter achieves better precision (P) under both matching criteria. However, as we increase the threshold the performance declining gradually.

TABLE II. Official Recognition Results on the Test Set

No of Run	NER			
	Criterion	Precision	Recall	F-measure
Run 1	Strict	0.662	0.873	0.753
	Overlap	0.681	0.910	0.776
Run 2	Strict	0.670	0.730	0.699
	Overlap	0.689	0.751	0.719
Run 3	Strict	0.516	0.250	0.337
	Overlap	0.535	0.260	0.350

In order to describe the normalization result, we used two methods to get average statistical scores i.e. micro average method and macro average method. Table III shows the official results on the test set. Comparing the normalization results with the recognition results we can observe that the developed CNN disambiguation method can improve the precision of the two tasks but reduce the recall. We believe that it may due to that the corpus released by the Bio-ID task does not exhaustively annotate all organism terms mentioned in figure captions.

TABLE III. Official Normalization Results on the Test Set

Run	Micro-P	Micro-R	Micro-F-score	Macro-P	Macro-R	Macro-F-score
1	0.660	0.882	0.755	0.709	0.924	0.685
2	0.668	0.760	0.711	0.732	0.824	0.611
3	0.525	0.327	0.403	0.766	0.462	0.273

IV. Conclusion

In summary, we introduce our new species recognition tool SPRENO which can recognize organism terms mentioned in figure captions. The developed tool will be released on https://www.nuget.org/packages/NTTU.BigODM.Bio.NER.Species/.

Acknowledgment

This work was supported by the Ministry of Science and Technology of Taiwan [grant numbers MOST-105-2221-E-143-003 and MOST-106-2221-E-143-007-MY3].

References

[1] A. Akkasi, and E. Varoglu, "Improving Biochemical Named Entity Recognition Performance Using PSO Classifier Selection and Bayesian Combination Method," *IEEE/ACM Trans Comput Biol Bioinform*, May 18, 2016.

[2] H.-J. Dai, O. Singh, J. Jonnagaddala, and E. C.-Y. Su, "NTTMUNSW BioC modules for recognizing and normalizing species and gene/protein mentions," *Database*, vol. 2016, January 1, 2016, 2016.

[3] S. Kim, R. Islamaj Doğan, A. Chatr-Aryamontri, C. S. Chang, R. Oughtred, J. Rust, R. Batista-Navarro, J. Carter, S. Ananiadou, S. Matos, A. Santos, D. Campos, J. L. Oliveira, O. Singh, J. Jonnagaddala, H.-J. Dai, E. C.-Y. Su, Y.-C. Chang, Y.-C. Su, C.-H. Chu, C. C. Chen, W.-L. Hsu, Y. Peng, C. Arighi, C. H. Wu, K. Vijay-Shanker, F. Aydın, Z. M. Hüsünbeyi, A. Özgür, S.-Y. Shin, D. Kwon, K. Dolinski, M. Tyers, W. J. Wilbur, and D. C. Comeau, "BioCreative V BioC track overview: collaborative biocurator assistant task for BioGRID," *Database: The Journal of Biological Databases and Curation*, vol. 2016, pp. baw121, 09/01

[4] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, "The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text," *PLoS ONE*, vol. 8, no. 6, pp. e65390, 2013.

[5] N. Naderi, T. Kappler, C. J. Baker, and R. Witte, "OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents," *Bioinformatics*, vol. 27, no. 19, pp. 2721-9, Oct 1, 2011.

[6] H.-J. Dai, P.-T. Lai, and R. T.-H. Tsai, "Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles," *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, vol. 7, no. 3, pp. 412-420, 2010.

[7] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang, "Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation." pp. 1333-1339.

[8] S. Moen, and T. S. S. Ananiadou, "Distributional semantics resources for biomedical text processing," LBM, 2013.