Organism named entity recognition on an enhanced CRF-based model and normalization for BioCreative VI bioentity normalization task

Fu-Chien Yang [1], Shiang-Chi Hsu [1], Hung-Yu Kao [1,2]
Institute of Medical Informatics, National Cheng Kung University, Tainan,
Taiwan, R.O.C. 701[1]
Department of Computer Science and Information Engineering, National Cheng Kung
University, Tainan, Taiwan, R.O.C. 701[2]

***Abstract*—Bioconcept curation is important to improve research on bio-literatures. Among bioconcepts, organism named entity recognition has been studied for a long time but without much success. According to this, we decided to develop a system for curating organisms and normalizing the entities of figure captions. In this work, we split the system in two parts, one is named entity recognition, and the other is entity normalization. In the first part, we leverage conditional random fields (CRFs) with several linguistic features that assist recognition, and in the second part, we utilize some heuristic ways to enhance the ability to identify the taxonomy ID for each entity. At the bioentity normalization task of BioCreative VI Bio-ID task, our system obtained 0.81 precision in the recognition stage and 0.61 precision in the normalization stage.**

Keywords—conditional random fields; taxonomy named entity recognition and normalization; Biomedical text mining

I. INTRODUCTION

In the biomedical field, biomedical publications are rapidly increasing. Text mining technology has become an integral part for biomedical literatures. Through natural language processing (NLP) techniques analysis, curating useful information within corpora becomes more accurate and lets us readily find relevant information. If there is too much irrelevant information in the corpus, existing techniques will struggle. As biomedical literature grows, so do the number of bioconcepts that require curation – e.g. chemicals, diseases, genes and organisms. There are many existing outstanding tools for curating bioconcepts, such as AuDis[1], OrganismTagger[2] and TaggerOne[3] etc. – but few address the need for curation of organisms. Therefore, we chose to participate in the BioCreative VI Task 1: Interactive Bio-ID Assignment, and we chose the bioentity normalization task to recognize organism named entities and normalize.

This task consists of a collection of figure captions from PubMed which are curated by the SourceData team. The format of the training data is BioC format, but to be easier to use in our system, we transform every caption into Pubtator format, and in the end of our system, we turn our result back into BioC format. After observing the captions in the training data, we found out that there are some words that are different from the original article, such as the word ‘Sykfl/flmice’ in corpus, but in original article it should be ‘Sykfl/fl mice’ that should have a space inside the word, and we developed some ways to overcome this problem. We also leveraged useful features and methods from other systems in our system, such as SR4GN[4] and AuDis[1]. Additionally, abbreviations of species names are widespread as well as the use of common English names instead of Latin names, which are easy to read but will cause difficulties for taxonomic identification of the organisms described in the captions. The use of acronyms, which can be both species specific and species independent, also poses challenges for recognition tasks. Lastly, incorrect spelling has created more ambiguity.

In this paper, in section II we will introduce the systems referenced; in section III we explain the method we use for developing our system; in section IV we will report the results of our experiments and discuss the differences between each experiment; finally, in section V we state our conclusions and future work.

II. RELATED WORK

SR4GN[4] identifies and disambiguates gene names, it also focus on species detection and recognition. The method to identify the main species mentioned in the article has been integrated into our system. The method gives more weight to species when the species mention occurs in the title as compared to the abstract. In particular, it will double counts on the frequency of the species mentions in the title. However, when multiple species have the same number of occurrences in a document, the author adopted a tie-breaking strategy that uses the global frequency of different species in the Linnaeus corpus. Another outstanding system is AuDis[1], which is for disease name entity recognition and normalization, but the method and features that the author used inspired us to develop this work.

Linnaeus[5] uses a dictionary-based approach to recognize species names and develops a set of heuristics to resolve ambiguous mentions. As a standalone and open source tool, Linnaeus has been widely used in many the biomedical literature. In LINNAEUS-species-corpus, it performs with 94% recall and 97% precision.

OrganismTagger[2] (OT) is another well-known system for curating organisms, it is a hybrid rule-based/machine learning system to extract organism mentions from the biomedical literature. OT addresses some challenges and makes some contributions, such as, a machine learning-based classifier for strain detection, and tools for automatically generating lexical and ontological resources from a copy of the NCBI Taxonomy database which allow the system to be updated by the users. On their manually annotated OT corpus, the OT achieves a precision of 95%.

In addition to the aforementioned systems, more recently, there is a new detection and recognition technology called TaggerOne[3], which is currently the state-of-the-art technology for tagging bioconcepts. The system is the first machine learning model for joint named entity recognition (NER) and normalization using semi-Markov models during both training and prediction. The result of NER F-measure of TaggerOne[3] in NCBI Disease corpus is 0.829 and in BioCreative V CDR corpus is 0.914.

Conditional Random Fields (CRFs)[6] is a popular probabilistic method for structured prediction. While most classifiers predict a label with just a single sample without considering other neighboring samples, CRF can take context into account which makes it perform better on predicting tokens' labels. For instance, the linear chain CRF which is famous in natural language processing and is used in our system, predicts sequences of labels for sequences of input samples.

III. Methods

To deal with the organism identification problem, we designed a semantic based recognition system which includes two modules as shown in Fig. 1. First, taxonomy named entity recognition. We utilize linear chain conditional random fields (CRFs)[7] as our recognition model based on semantic features and dictionaries (NCBI taxonomy dictionary and UniProt taxonomy identifier). Also, we expanded our dictionary in some ways to deal with the special cases that influence our system to recognize the taxonomy entities. Second, taxonomy normalization. In this stage, we not only use the dictionary to identify the entities but also find the target species of each article to enhance the ability to normalize the species mentions.

A. Named Entity Recognition

To train a taxonomy named entity recognition model, we leverage the CRF++ toolkit (https://taku910.github.io/crfpp/).

In this model, we utilized BIEO states (B: begin, I: inside, E: end, O: outside) to tag each word and using a template of CRF which assists our model to recognize a taxonomy named entity. For training the CRF model, we need to get all the features of each token in the preprocessing stage.

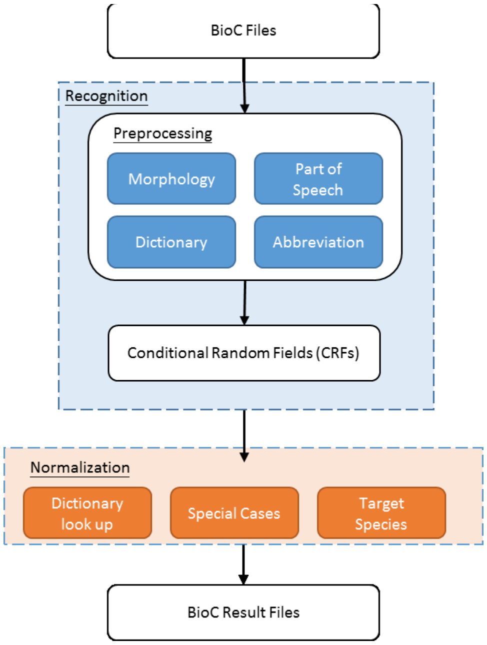

Fig. 1. *The architecture of our system, which can be divided into two stages, recognition stage and normalization stage.*

First, we divide tokens with not only spaces and punctuations but also letters and digits. After getting all the tokens, we transform them into lowercase, and get several features of them. In AuDis[1], the author uses features like morphology and part of speech, which are really useful to recognize the meaning of each word, so in our model, we utilized these as some of our important features. The significant features are describe as below:

- Morphology: this feature includes the original tokens, stemmed tokens which are extracted by the Snowball library, and prefixes/suffixes of the tokens whose lengths range from 1 to 5.
- Part of Speech: we use Stanford POS tagger to extract the part of speech of the token.
- Dictionary: we utilize the dictionary from NCBI taxonomy and UniProt Taxonomic identifier, to get the token's rank, like species, genus etc.
- Abbreviation: if the curated entity is an abbreviation, we label it as abbreviation, otherwise, we label it as normal.

For the dictionary feature, there are two dictionaries that we use. In the dictionaries, there are several columns and information that can be used in our features, such as taxonomy ids, taxonomy names, other names, rank of each mention, scientific name and common names. In particular, to strengthen our dictionary lookup feature, we have two possible ways; first, if the token is one of the taxonomy named entities but can't be found in these two dictionaries, we add this token into our dictionary. Second, we add to our dictionary if the token is an abbreviation of a taxonomy named entity or some common abbreviation.

In addition to the methods above, sometimes the taxonomy named entities will appear in a word which should have a space between it but missing. We store these tokens as special cases and using these at the normalization stage to avoid the situation where the correct entity is found in the text but can't be found in the dictionary.

B. Conditional Random Fields module

CRFs have been applied in many entity extraction studies in biomedical literature. And CRFs are a type of discriminative undirected probabilistic model for computing conditional probability distributions. Lafferty, McCallum and Pereira define the conditional probability distribution p(YX) of a random variable Y given the input X as follows:

$$p(YX) = \frac{1}{Z(X)} exp\big(f(Y,X)\big) = \frac{\exp(f(Y',X))}{\sum_{Y'} \exp(f(Y',X))} \qquad (1)$$

where Y = {y_1, y_2, …, y_n} is a label sequence from an observation sequence X = {x_1, x_2, …, x_n} which means a token sequence. Z(X) is the normalization term. To learn the feature weights in a CRF, we can use gradient ascent because it is memory efficient. The weighted feature function in equation 2 for deciding the label at position i is a function of the label at position i-1, and the entire observation sequence X, made up of all x_i, which are vectors of features.

$$f(Y,X) = \sum_{j=1}^{n} \sum_{i=1}^{m} w_i f_i(y_i y_{i-1}, X) \qquad (2)$$

C. Normalization

After passing the data through our CRF model, we obtain the label of each token. With these labels, we can extract the taxonomy entities. At the taxonomy normalization stage, we use NCBI Taxonomy dictionary to identify each taxonomy named entity. If the taxonomy named entity cannot be found in the dictionary, the entity might be one of the special cases or it might be a kind of species' embryo, larva etc.

If it is the latter, it is a bioconcept ambiguity problem. To address this problem, we developed an approach for looking up the articles' target species. At first, we break the original full text article into tokens, and get the features of each token as in the preprocessing stage; next, we utilize our CRF model to extract the taxonomy entities in the article; finally, we choose the taxonomy entity which appear the most frequently in the full text article as the target species for this article.

TABLE I. CAPTIONS STATISTICS

Task Dataset	Captions
Training data	13696
Caption contains Taxonomy	4206

TABLE II. RESULTS WITH AND WITHOUT SPECIAL CASES

RUN	SC-			SC+		
	P	**R**	**F**	**P**	**R**	**F**
1	0.571	0.373	0.451	0.548	0.403	0.464
2	0.486	0.315	0.382	0.52	0.359	0.424
3	0.594	0.273	0.374	0.625	0.315	0.419
4	0.799	0.481	0.6	0.81	0.533	0.642
5	0.778	0.485	0.598	0.793	0.548	0.648
Avg.	0.645	0.385	0.481	0.659	0.432	0.519

With the target species of all the articles, we can identify the entities that are extensions of the target species. As a result, we can address the problem of taxonomy ambiguity, and this approach improves our system for taxonomy normalization.

Some species' names can cause ambiguity (i.e. mouse, mice), because the species like "mouse" can be identified as taxonomy ID:10088 as genus and taxonomy ID:10090 as species. To address this problem, we normalize the mentions like mouse, mice, rat which IDs are 10088 into 10090 to make it consistent, which means that we identify IDs for species rank, not genus rank.

IV. EXPERIMENT AND RESULTS

To make the data easier to train and raise the accuracy of recognition, we transform all the training data from BioC format into Pubtator format, and extract the annotated captions that contains taxonomy entities. The statistics of the captions are shown in Table I.

In our evaluation, we use 5-fold cross-validation to evaluate our result on training data that only contains annotated captions with taxonomy entities. The results are shown in Table II. The SC- means normalization without checking if the entity is one of the special cases, and the SC+ is using the special cases to identify the unknown taxonomy named entities in the normalization stage. This approach raises the F-measure about 0.03%.

Due to the better result of using special cases, we go on to analyze the effect of identifying the unknown entities using original articles' target species. As shown in Table III, the F-measure raises about 0.12%.

Finally, we get our test result from BioID scorer, Table IV shows our score in each corpus that contains scores with and without normalization. The score without normalization counts extracted taxonomy named entities, and the score with normalization counts the correct taxonomy IDs.

V. DISCUSSION AND CONCLUSION

From our test results, we found out that our performance on taxonomy named entities recognition is effective, but we have more room to improve in our normalization stage. We now discuss some methods that might improve the performance. First, the way we find the target species in the original articles, we only use our CRF model to detect the mentions in the article. A better way is to not only use our model but also use some

TABLE III. PERFORMANCE WITH TARGET SPECIES

RUN	SC+			SC+_Target		
	P	**R**	**F**	**P**	**R**	**F**
Avg.	0.659	0.432	0.519	0.703	0.579	0.635

other taxonomy identifier to enhance the accuracy of getting the right target species. Besides, we only use two dictionaries (NCBI Taxonomy dictionary and UniPort Taxonomy identifier) and only use NCBI Taxonomy dictionary for the normalization stage. As the results show, we need more high quality dictionaries to improve our normalization stage. Lastly, if an entity is correctly mentioned once, if the same word appears again, the word should be marked same as the entity above, as in AuDis's[1] post-processing. We believe that integrating this step will significantly raise our performance.

Overall, we found several features which can assist our system in developing enhanced CRF model for taxonomy named entities recognition. With these features, our system sees increased performance in the recognition stage. In the normalization stage, we also use several methods to correctly identify the mentions' IDs, such as finding the target species and the special cases which significantly improved the F-measure. In our future work, we will focus on enhancing the performance of the normalization stage with the methods mentioned above, and by considering some other machine learning approaches.

TABLE IV. TEST RESULTS FROM BIOID SCORER

	precision	*F-measure*	*norm_ precision*	*norm_ F-measure*
Any_Strict	0.813	0.701	0.609	0.555
Any_Overlap	0.853	0.735	0.609	0.555
Normalized_Strict	0.812	0.701	0.609	0.555
Normalized_Overlap	0.852	0.735	0.609	0.555

REFERENCES

1. Lee H.C. et al. An Enhanced CRF-Based System for Disease Name Entity Recognition and Normalization on BioCreative V DNER Task. Proc BioCreative Workshop. Sevilla, Spain, pp. 226–233, 2016.

2. Naderi N, Kappler T, Baker CJO, Witte R, "OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents," Bioinformatics, vol. 27, pp.2721-2729, 2011.

3. R. Leaman et al. "TaggerOne: joint named entity recognition and normalization with semi-Markov models," Bioinformatics, vol. 32 pp. 2839–2846, 2016.

4. C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," Plos one, vol. 7, p. e38460, 2012.

5. M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature," BMC bioinformatics, vol. 11, p. 85, 2010.

6. J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

7. Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature, Database, vol. 2011, p. baq036, 2011.

8. Y.-Y. Hsu and H.-Y. Kao, "Curatable Named-entity Recognition using Semantic Relations," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, pp. 785-792, 2014.

9. Doğan R.I. et al. NCBI disease corpus: A resource for disease name recognition and concept normalization. J. Biomed. Inf ., 47, 1-10, 2014.

10. C.-H. Wei, H.-Y. Kao, and Z. Lu, "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains," BioMed Research International, vol. 2015, 2015.

11. R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," Journal of cheminformatics, vol. 7, 2015.

12. J. G. Caporaso, W. A. Baumgartner, D. A. Randolph, K. B. Cohen, and L. Hunter, "MutationFinder: a high-performance system for extracting point mutation mentions from text," Bioinformatics, vol. 23, pp. 1862-1865, 2007.