An ensemble CNN method for biomedical entity normalization

Pan Deng, Haipeng Chen, Mengyao Huang, Xiaowen Ruan, Liang Xu

Ping An Technology (Shenzhen) Co., Ltd, Shenzhen, People's Republic of China {dengpan749, chenhaipeng867, huangmengyao666, ruanxiaowen571, xuliang867}@pingan.com.cn

Abstract

Different representations of the same concept could often be seen in scientific reports and publications. Entity normalization (or entity linking) is the task to match the different representations to their standard concepts. In this paper, we present a two-step ensemble CNN method that normalizes microbiology-related entities in free text to concepts in standard dictionaries¹. The method is capable of linking entities when only a small microbiology-related biomedical corpus is available for training, and achieved reasonable performance in the online test of the BioNLP-OST19 shared task Bacteria Biotope.

1 Introduction

With over 500K papers in the biomedical field published on average every year², it is important to promote efficient information retrieval and knowledge processing from the literatures automatically. Named entity recognition (NER), which extracts meaningful real-world objects from free text, and entity normalization (entity linking), which links ambiguous or varied extracted objects to standard concepts, are two fundamental natural language processing (NLP) tasks to approach the goal.

With many attempts made for general entity normalization (Hachey, Radford et al. 2013, Luo, Huang et al. 2015, Wu, He et al. 2018, Aguilar, Maharjan et al. 2019), biomedical entity linking faces more challenges handling entity variations, making it an enthralling field to be explored. Many studies endeavored to solve biomedical entity normalization issues have been published (Hanisch, Fundel et al. 2005, Leaman and Lu 2016, Cho, Choi et al. 2017, Li, Chen et al. 2017, Luo, Song et al. 2018, Ji, Wei et al. 2019). Meanwhile, BioNLP Shared Tasks, one of the community-wide challenges that aim to find solutions for biomedical literature information retrieval, also addresses diverse tasks of entity linking (Bossy, Jourde et al. 2011, Bossy, Golik et al. 2013, Nédellec, Bossy et al. 2013, Chaix, Dubreucq et al. 2016, Deléger, Bossy et al. 2016). However, further investigations are required to improve the performance of the entity linking systems, especially when the available corpus is small.

Here, we present a two-step neural networkbased ensemble method that links free text preannotated microbiology-related entities to standard concepts using semantic information from pre-trained word vectors. By integrating a perfect match method with a shallow CNN, our model's performance is comparable to the SOTA methods' performance when trained with a small biomedical corpus (2258 microbiology-related entities, or 1248 after de-duplication, from 198 microbiology related publications and reports) provided by the BioNLP-19 task Bacteria Biotope challenge.³

We have compared our ensemble model to both a baseline method, of which we linked free text entities to the standard concepts by vector distance (Manning, Raghavan et al. 2010), and ABCNN, one of the SOTA models that could be used for entity normalization (Yin, Schütze et al. 2016). In addition, the method was tested online, and the results indicated that our model achieved a reasonable performance for microbiology-related entities linking tasks with small corpora.

2 Related work

Entity normalization is a rich research field where diverse approaches have been proposed.

¹ Our code is available at:

https://github.com/OXPHOS/BioNLP

² http://dan.corlan.net/medline-trend/language/absolute.html

³ https://sites.google.com/view/bb-2019/home



Figure 1 Model Architecture Overview. VSM: pre-trained word vector space model. Ab3P: an abbreviation detection tool developed specifically for biomedical concepts. CNN: convolutional neural network.

Most early studies utilized morphological similarity defined by editing distances between input terms and standard concepts to normalize the entities (Ristad and Yianilos 1998, Aronson 2001). Later, heuristic rules were incorporated to improve the performance (Hanisch, Fundel et al. 2005, Kang, Singh et al. 2012, Karadeniz and Özgür 2013, Tiftikci, Şahin et al. 2016, Cho, Choi et al. 2017). Among them, MaCallum et al. (2012), DNorm (Leaman, Islamaj Doğan et al. 2013) and LIMSI (Grouin and Moriceau 2016) introduced conditional random field (CRF) to the entity linking tasks, while TaggerOne (Leaman and Lu 2016) presented a NER and normalization joint system utilizing semi-markov models, and it has been adopted by an integrated bioconcept annotation and retrieval platform developed by NIH (Wei, Allot et al. 2019). However, many of the studies achieved good performance yet were limited for further improvements due to the common drawbacks of rule-based methods.

Approaches utilizing semantic information of the entities was made possible by the appearance of the word embedding technique. Word embedding projects words to vector spaces, where the cosine similarities between the vectors indicate their semantic similarities. The CONTES system (Ferré, Zweigenbaum et al. 2017) and the following HONOR system (Ferré, Deléger et al. 2018) performed entity linking tasks by minimizing the distances between embedded input terms and standard biomedical concepts. Karadeniz and Özgür (2019) proposed an unsupervised method for entity linking tasks using word embeddings and a syntactic parser.

Meanwhile, neural networks have been combined with word embeddings to normalize biomedical entities. Limsopatham and Collier (2016) applied convolutional neural network (CNN) and recurrent neural network (RNN) to pre-trained word embeddings to normalize medical concepts in social media texts, and achieved the SOTA performance on several datasets. Li et al. (2017) utilized a CNN structure to rank the candidates generated by rule-based methods. Deep neural networks such as multi-view CNN and BERT have also been proposed to normalize biomedical entities (Luo, Song et al. 2018, Ji, Wei et al. 2019). However, their applications might be limited due to the requirement of large amount of data.

3 Models

Our model architecture is shown in **Figure 1**, where our major work is highlighted in blue and further discussed in Section 3.1-3.3.

To process the entities from the standard dictionary, let \mathbf{Y}_i^d be the *i*-th entity from the dictionary, and $y_{ij} \in \mathbb{R}^k$ be the *k*-dimensional word vector of the *j*-th word in the entity \mathbf{Y}_i^d . the embedded vector \mathbf{y}_i of entity \mathbf{Y}_i^d is defined as

$$\mathbf{y}_i = \text{PCA}(\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \ n_{\text{component}} = 0.95),$$

where $n_i \in \mathbb{N}^+$ is the number of words present in a pre-trained vector space model (VSM) in the entity Y_i^d . The VSM was created from the biomedical scientific literature in the PubMed database and Wikipedia (Pyysalo 2013). k = 200. PCA was conducted to increase training efficiency, where $n_{\text{component}}$ is the reduction rate.

The processing of the entities from free text is described in the following section in detail. To be noted, abbreviations are commonly seen in free text from publications and reports. For example, CNS, standing for the central nervous system, might often be used in research literature over topics of neuroscience, and would be annotated as entities to be linked. However, abbreviations, mostly derived from phrases, are often absent from the pre-trained word vector spaces and would interfere with the model training. To solve this problem, we first converted potential abbreviations in free text pre-annotated entity list to their long forms with Ab3P. Ab3P is an abbreviation detection tool developed specifically for biomedical concepts. It reached 96.5% precision and 83.2% recall on 1250 randomly selected MED-LINE records as suggested by Sohn et al (2008).

The converted free text pre-annotated entities were then matched with dictionary-derived standard concepts by characters through a perfect match module (Section 3.1). The entities failed perfect matching were then fed to a set of shallow CNN models (Section 3.2) trained with bootstrap samples. Next, the outputs of the CNNs were mapped to standard entity vectors via cosine similarity. The standard entity vectors output from the voting classifier (Section 3.3) were predicted as the linked results of the input entities.

3.1 Perfect match

We noticed that some entities from the free text were able to match with the standard entities by characters after rule-based processing. These entities were then directly linked to the dictionary instead of being fed to the Word2Vector and CNN models. The rules we designed include:

• Hyphens were replaced with spaces.

• Characters except alphabetic letters and spaces were removed.

• Case-insensitive string matching was performed between the free text entities and standard entities.

3.2 Shallow CNN

The shallow CNN (Figure 1) was adapted from the previous ideas from Kim (2014) and Limsopatham and Collier (2016).

To start with, let X_i^t be the *i*-th input entity (which were provided by the task), and $x_{ij} \in \mathbb{R}^k$ be the *k*-dimensional word vector of the *j*-th word in the entity X_i^t , k = 200. The embedded matrix \mathbf{x}_i of entity X_i^t is defined as

$$\mathbf{x}_i = x_{i1} \oplus x_{i2} \oplus \dots x_{in_i}$$

Here $n_i \in \mathbb{N}^+$ is the number of words present in the pre-trained VSM (Pyysalo 2013) in the entity X_i^t . \oplus is the concatenation operator. \mathbf{x}_i is padded to length 8 as 98.8% of the input entities were composed of 8 or fewer words. For the entities with more than 8 words, average pooling was performed in prior with the pool size = (w, 200) and step= w, where w = [n/8]. In other words, simple average of the neighboring w words was calculated, so that the final embedded matrix would always have a length ≤ 8 .

A temporal convolution kernel followed by a max-over-time pooling operation and a fully connected layer were applied to each \mathbf{x}_i . The output $\overline{\mathbf{y}}_i$ was then passed to a cosine similarity function to calculate the similarity scores between $\overline{\mathbf{y}}_i$ and each standard entity vector respectively. The standard concept with the highest score was predicted as the linked entity $\overline{\mathbf{y}}_i^e$.

3.3 Ensemble mechanism with voting

To reduce overfitting, we designed an ensemble method that combined 5 shallow CNNs with the concept of boosting (Valiant 1984). The 5 CNNs shared the identical architecture, but their weights were randomly initialized respectively. To increase the generalization capability of our model, the CNNs were fed with training data randomly subsampled with bootstrap method (Efron 1982), with the out-of-bag samples used for cross-validation.

The final normalized results were achieved with a majority-vote classifier over the outputs from the 5 shallow CNNs. If no majority output was present, the output from the network with the best cross-validation estimates would be chosen.

3.4 Baseline model

For each entity E_i^d in the standard dictionary and E_i^t in free text, the corresponding embedded

		Total Number		Number after de-duplication			
	Article			Habitat+			Habitat+
	Number	Habitat	Phenotype	Phenotype	Habitat	Phenotype	Phenotype
Training	133	1118	369	1487	627	176	803
Development	66	610	161	771	348	97	445
Test	97	924	252	1176	596	148	744

Table 1. Data Statistics

vector \mathbf{v}_i^d and \mathbf{v}_i^t are defined, respectively, as:

$$\mathbf{v}_{i}^{d} = \frac{1}{n_{i}} \sum_{j=1}^{n_{i}} v_{ij}^{d}, \ \mathbf{v}_{i}^{t} = \frac{1}{n_{i}} \sum_{j=1}^{n_{i}} v_{ij}^{t}$$

where $n_i \in \mathbb{N}^+$ is the number of words present in the pre-trained VSM in the *i*-th entity.

Cosine similarity between each free text-dictionary entity pair was calculated. The free text entity was linked to the dictionary entity with the highest similarity score.

3.5 ABCNN

ABCNN (Yin, Schütze et al. 2016) is a state-ofthe-art deep learning model for text similarity learning, which could also be applied for entity linking tasks. The model introduced attention mechanism into a pair of siamese architecturebased weight sharing CNNs (Bromley, Guyon et al. 1994).

For our purpose, we used a slight variant of a published ABCNN model⁴. In addition, attention mechanism could be applied to different layers of the CNN pair according to the original publication. Considering the data volume and the model complexity, we applied the attention mechanism to the input layer.

4 Experiments and Results

4.1 Data and resources

The biomedical corpus and pre-annotated entities were provided by BioNLP-OST19 task Bacteria Biotope. **Table 1** shows the detailed data statistics provided by the task. Two types of entities were involved in the task: *phenotype*, which describes microbial characteristics, and *habitat*, which describes physical places where microorganisms could be observed. Dictionary with 3602 standard concepts was also provided by the task. In the original dictionary, each concept is assigned to a unique ID, while its hierarchical information of its direct parents is also listed. In our model, the hierarchical information is omitted.

Ab3P-detected abbreviations were provided as separate input files by the task organizers⁵.

The 4 GB word vector space model was downloaded in binary format⁶ and extracted with python package gensim.

4.2 Training

Our CNN model was trained using stochastic gradient descent optimizer with cosine proximity as the loss function. We randomly split 20% samples as validation dataset for each CNN and used early stopping criteria to determine the number of training epochs. The learning rate was fixed to 0.01. Batch size (2), kernel size (4) and filter number (5000) were determined by grid-search.

As expected with this small volume of data, extra convolution layers led to overfitting.

4.3 Held-out evaluation

We used precision metrics, the official metric of the challenge, to evaluate the performance of our model and the reference models on the held-out development dataset respectively.

	Habitat	Phenotype	Total
Ensemble CNN	0.624	0.615	0.622
Perfect match	0.863	0.937	0.869
Shallow CNN	0.526	0.572	0.538
ABCNN	0.244	0.134	0.221
baseline	0.207	0.101	0.184

 Table 2. Performance of different models on development dataset (Precision)

As shown in **Table 2**, non-supervised baseline model yielded a precision score of 0.184 on the development dataset, while ABCNN yielded 0.221, which might be attributed to the small

⁴ https://github.com/galsang/ABCNN

⁵ https://sites.google.com/view/bb-2019/supporting-resources

⁶ http://bio.nlplab.org/

training corpus. Our ensemble CNN model performed around 3 times better than both reference models with an average precision score of 0.622, indicating the efficiency of our model.

However, it should be noted that the perfect match module in our system had a remarkable higher precision score compared to the shallow CNN module, suggesting that the performance of the neural network could be further improved.

We analyzed the result of the shallow CNN module and concluded that 3 possible reasons might be associated with the performance: 1) Missing context. For example, our model normalized "children" to "child", while the provided label was "patient with infectious disease" in articles describing children with infectious disease. 2) Missing hierarchical information. For example, our model normalized "B cell" to "cell" instead of "lymphocyte", and the latter one was a more accurate description. Tackling the above two issues would require either the context or the hierarchical information of the standard concepts to be considered in the system. 3) Wrong match. For example, "cats" was normalized to "dog", suggesting that the networks were not trained well to normalize these words. However, we noticed that such errors mostly came with a majority vote of 2 or 1, which on the other hand demonstrated the power of the voting mechanism.

	Habitat	Phenotype
PADIA_BacReader	0.684	0.758
Challenge-provided	0.559	0.581
baseline		
AmritaCen	0.522	0.646
_healthcare		
BLAIR_GMU [#]	0.615	0.646
BOUN-ISIK [#]	0.687	0.566

Table 3. Performance of different models on online test dataset (Precision). The performance of our model (PADIA_BacReader) is bolded. [#]: Best run of the submissions is considered.

4.4 Online test

The ensemble CNN model was then evaluated through online testing⁷.

Our results showed a 12.5% and 17.7% precision increase in habitat and phenotype entity linking tasks respectively compared to the challenged-provided baseline model (**Table 3**), where case-insensitive string matching was applied for

5 Conclusions and Future direction

We introduced a two-step neural network-based ensemble method that linked microbiology-related biomedical entities extracted from free text to standard concepts. The shallow architecture and ensemble mechanism on top of a perfectmatch morphological similarity method achieved reasonable predictions with limited training samples. The comparison with reference models suggested the efficiency of our model. In addition, our approach could be applied to other scenarios where semantic linking between entities is required as well.

Further improvement might be achieved once more semantic clues are incorporated, as we briefly discussed at the end of Section 4.3. The normalization deviation due to missing context clues did not only affect the performance of shallow CNN, but also affected the performance of perfect math as well. For example, though entity 'cell' has a perfect match in standard dictionary, it might be referred to 'lymphocytes' specifically in a research paper discussing about immunity. While some efforts have been made to preserve hierarchical information between concepts during entity linking (Ferré, Deléger et al. 2018), It would be interesting to investigate if knowledge graphs derived from the standard dictionaries and input corpus could contribute to the semanticbased entity normalization.

In addition, our model assigned the same weight to all the words present in the VSM, which might compromise the performance of the system. For example, only the word "children" is informative in the entity "children less than five years of age", as the entity is normalized to "child". The presence of other words might interfere with the normalization. Regarding this issue, syntactic parsers might be adopted for performance improvement.

linking. In addition, it performed the best or among the best ones compared to models proposed by other participants, suggesting the advantages of our model. We did not test our own reference models online due to the limited number of submissions to the challenge.

⁷ https://sites.google.com/view/bb-2019/pre-

diction-submission

References

- Aguilar, G., Maharjan, S., López-Monroy, A. P., & Solorio, T. (2019). A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*.
- Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Paper presented at the Proceedings of the AMIA Symposium.
- Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., & Nédellec, C. (2013). Bionlp shared task 2013–an overview of the bacteria biotope task. Paper presented at the Proceedings of the BioNLP Shared Task 2013 Workshop.
- Bossy, R., Jourde, J., Bessieres, P., Van De Guchte, M., & Nédellec, C. (2011). BioNLP Shared Task 2011-Bacteria Biotope. Paper presented at the Proceedings of BioNLP Shared Task 2011 Workshop.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). *Signature verification using a" siamese" time delay neural network.* Paper presented at the Advances in neural information processing systems.
- Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M., . . . Lepiniec, L. (2016). Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. Paper presented at the Proceedings of the 4th bionlp shared task workshop.
- Cho, H., Choi, W., & Lee, H. (2017). A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC bioinformatics*, 18(1), 451.
- Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferre, A., Bessieres, P., & Nedellec, C. (2016). *Overview of the bacteria biotope task at bionlp shared task 2016.* Paper presented at the Proceedings of the 4th BioNLP shared task workshop.
- Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *Journal of the American Statistical Association*, 78(384), 987. doi:10.1137/1.9781611970319
- Ferré, A., Deléger, L., Zweigenbaum, P., & Nédellec, C. (2018). Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).
- Ferré, A., Zweigenbaum, P., & Nédellec, C. (2017). Representation of complex terms in a vector space structured by an ontology for a normalization task. Paper presented at the BioNLP 2017.
- Grouin, C., & Moriceau, V. (2016). LIMSI at SemEval-2016 Task 12: machine-learning

and temporal information to identify clinical events and time expressions. Paper presented at the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).

- Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with Wikipedia. Artificial intelligence, 194, 130-150.
- Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1), S14.
- Ji, Z., Wei, Q., & Xu, H. (2019). BERT-based Ranking for Biomedical Entity Normalization. *arXiv* preprint *arXiv:1908.03548*.
- Kang, N., Singh, B., Afzal, Z., van Mulligen, E. M., & Kors, J. A. (2012). Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5), 876-881.
- Karadeniz, I., & Özgür, A. (2013). Bacteria biotope detection, ontology-based normalization, and relation extraction using syntactic rules.
 Paper presented at the Proceedings of the BioNLP Shared Task 2013 Workshop.
- Karadeniz, I., & Özgür, A. (2019). Linking entities through an ontology using word embeddings and syntactic re-ranking. BMC bioinformatics, 20(1), 156.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22), 2909-2917.
- Leaman, R., & Lu, Z. (2016). TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18), 2839-2846.
- Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., & Huang, D. (2017). CNN-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11), 385.
- Limsopatham, N., & Collier, N. (2016). Normalising medical concepts in social media texts by learning semantic representation. Paper presented at the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
- Luo, G., Huang, X., Lin, C.-Y., & Nie, Z. (2015). Joint entity recognition and disambiguation. Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

- Luo, Y., Song, G., Li, P., & Qi, Z. (2018). *Multi-task medical concept normalization using multiview convolutional neural network.* Paper presented at the Thirty-Second AAAI Conference on Artificial Intelligence.
- Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100-103.
- McCallum, A., Bellare, K., & Pereira, F. (2012). A conditional random field for discriminatively-trained finite-state string edit distance. *arXiv preprint arXiv:1207.1406.*
- Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., & Zweigenbaum, P. (2013). *Overview of BioNLP shared task 2013*. Paper presented at the Proceedings of the BioNLP Shared Task 2013 Workshop.
- Pyysalo, S. G., Filip; Moen, Hans; Salakoski, Tapio; Ananiadou, Sophia. (2013). Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, 39-44.
- Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5), 522-532.
- Sohn, S., Comeau, D. C., Kim, W., & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1), 402.
- Tiftikci, M., Şahin, H., Büyüköz, B., Yayıkçı, A., & Özgür, A. (2016). Ontology-based categorization of bacteria and habitat entities using information retrieval techniques. Paper presented at the Proceedings of the 4th BioNLP Shared Task Workshop.
- Valiant, L. G. (1984). A theory of the learnable. symposium on the theory of computing, 27(11), 1134-1142. doi:10.1145/800057.808710
- Wei, C. H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res*, 47(W1), W587-W593. doi:10.1093/nar/gkz389
- Wu, G., He, Y., & Hu, X. (2018). Entity linking: an issue to extract corresponding entity with knowledge base. *IEEE Access*, 6, 6220-6231.
- Yin, W., Schütze, H., Xiang, B., & Zhou, B. (2016). Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4, 259-272.