# Integration of Deep Learning and Traditional Machine Learning for Knowledge Extraction from Biomedical Literature

**Jihang Mao[1], Wanli Liu[2]**
[1]Montgomery Blair High School, 51 University Blvd E, Silver Spring, MD 20901, USA
[2]TAJ Technologies, Inc., 7910 Woodmont Ave #1214, Bethesda, MD 20814, USA
jim-blair@hotmail.com, lwl75@live.com

## Abstract

In this paper, we present our participation in the Bacteria Biotope (BB) task at BioNLP-OST 2019. Our system utilizes fine-tuned language representation models and machine learning approaches based on word embedding and lexical features for entities recognition, normalization and relation extraction. It achieves the state-of-the-art performance and is among the top two systems in five of all six subtasks.

## 1 Introduction

With the rapid increasing volume of biomedical literature, finding useful knowledge from large amount of scientific papers, databases or web pages has become more and more difficult. Knowledge about microbial diversity is crucial for the study of the microbiome and the interaction mechanisms of bacteria with their environment, as well as phylogenetic and ecology perspectives. Such knowledge has been produced by biology and bioinformatics projects in the microbiology domain, including food safety, health sciences and waste processing. However, a significant portion of this information is expressed in free text, e.g., the microbial strains experimentally identified in a given environment (habitat), and theirs properties (phenotype). Given such information, there is no comprehensive resource gathering the knowledge (Deléger et al., 2016).

It is crucial to automatically extract information from heterogeneous resources as it can help with reaching the desired information efficiently for fundamental research and applications, especially in biomedical fields (Cohen and Hersh, 2005). Not only is extracting the relationships between biomedical terms necessary, normalizing them with respect to common references is equally important (Floyd et al., 2005; Buttigieg et al., 2013). However, despite the recent progress in machine learning, text mining and natural language processing, automating the knowledge extraction pipeline is rather challenging. A system must first identify entities (e.g. Microorganisms or Habitats names) in the document through a named entity recognition method. Next, linguistic cues within the document are used to predict whether a relationship between each pair or group of entities exists and which type of relationship it is. The entities are normalized according to domain knowledge resources, so that they can be represented in a formal and structured way by using concepts from an ontology or a taxonomy. Scientific literature mining challenges have been organized to address the need of knowledge extraction. For instance, BioNLP Shared Task is a community-wide effort on the development of fine-grained information extraction methods in biomedicine since 2009.

The Bacteria Biotope (BB) task is part of the BioNLP Open Shared Tasks, and has been previously conducted in 2016 (Deléger et al., 2016), 2013 (Bossy et al., 2013) and 2011 (Bossy et al., 2011). The goal of the BB task is to provide a framework for the evaluation and comparison of automatic information extraction methods for Bacteria organism habitats. The 2019 BB task (Bossy et al., 2019) consisting of three subtasks: named entity recognition and normalization (BB-norm and BB-norm+ner), entity and relation extraction (BB-rel and BB-rel+ner) and knowledge base extraction (BB-kb and BB-kb+ner). The representation scheme of the BB task contains four entity types: *Microorganisms*, *Habitats*, *Geographical places* and *Phenotypes*. The normalization subtask focuses on normalizing the entities with taxa from NCBI Taxonomy (for *Microorganism*) and concepts from OntoBiotope ontology (for *Habitat* and *Phenotype*). The relation extraction subtask focuses on extracting *Lives_In*



*Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 168–173

relations between *Microorganism*, *Habitat* and *Geographical* entities, and *Exhibits* relations between *Microorganism* and *Phenotype* entities. The knowledge base extraction subtask can be viewed as a combination of the first two subtasks, aggregating their results at the corpus level. We participated in all subtasks in this challenge.

A brief description of our method for the 2019 BB task is presented in Section 2. In Section 3 we show the results of our method on the official BB test datasets and a brief discussion of the results. In sections 4 we conclude our participation in the BB task.

## 2 Methods

In this section, we present the methods we used while participating in the 2019 BB task. We build our system upon methods from successful tools in previous BioNLP Shared Task (Lever and Jones, 2016; Mehryary et al., 2016), and partially reuse the method we designed while participating in other recent natural language processing challenges (Mao and Liu, 2019).

Given the main purposes of the three subtasks of the BB task, we design three corpus-level components in our system: named entity recognition, normalization, and relation extraction. We do not use any additional or customized training data besides the BB corpus provided by the organizers.

### 2.1 Named Entity Recognition

The first step in the knowledge extraction process is to accurately recognize the names of entities in text. Our NER component is based on most recent advances in deep learning for NLP applications: pre-trained language representation model and transfer learning.

The BB corpus is provided in the BioNLP-ST standoff annotation format. After the input text is loaded, it is converted to the CoNLL IOB (Inside, Outside, Beginning, respectively) format for NER processing. For discontinuous entities, multiple annotations will be tagged. Since there are only a small number of such entities in the corpus, we expect a minimal effect on the accuracy.

Our first method builds on BERT, which was proposed in October 2018, and obtained state-of-the-art performance on NLP tasks (Devlin et al., 2018). BERT utilizes a multilayer bidirectional transformer encoder which can learn deep bi-directional representations and can be later fine-tuned for a variety of tasks such as NER. Before BERT, deep learning models, such as Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) have greatly improved the performance in NER over the last few years (Huang et al., 2015). OpenAI GPT (Radford et al., 2018) has proved the effectiveness of generative pre-training a language model and subsequent discriminative fine-tuning it on a specific natural language understanding task.

For each sentence from the BB corpus, this method first obtains its token representation from the pre-trained BERT model using a case-preserving WordPiece model, including the maximal document context provided by the data. Next, we formulate this task as a tagging task by feeding the representation into a CRF (Lafferty et al., 2001) output layer, which is a token-level classifier over the NER label set.

The pre-trained BERT models were trained on a large corpus (Wikipedia + BookCorpus). There are several pre-trained models released. In the BB task, we choose BERT-Large, Cased (Whole Word Masking, WWM) model for the following reasons: 1) The BB corpus is in English, and for high-resource languages, a single-language model is better than the multilingual model[1]; 2) The BERT-Large model generally outperforms the BERT-Base model in most NLP tasks (Tenney et al., 2019); 3) The cased model is better than uncased model because the case information is important for the NER task[2]; 4) The recently released WWM variant of BERT-Large[3] yields improvements on various NLP tasks by masking whole words instead of random masking in original BERT in pre-processing. The variant of BERT model that trained on biomedical text, such as BioBERT (Lee et al., 2019), is more helpful for biomedical text mining tasks. However, BioBERT is based on the same vocabulary as the BERT-Base model, and it does not outperform the BERT-Large (WWM) model in our experiments.

In the BB task, we represent the input passage as a single packed sequence using BERT embedding, then use a CRF layer as the tag

[1] https://github.com/google-research/bert/blob/master/multilingual.md

[2] https://github.com/google-research/bert#pre-trained-models

[3] https://github.com/google-research/bert (5/31/2019 notes)

decoder. We set the maximum sequence length to 512 in order to avoid missing entities in long sentences.

Our second method builds on XLNET, which was proposed in June 2019, also achieved state-of-the-art performance on various NLP tasks (Yang et al., 2016). XLNET is similar to BERT, but it overcomes the limitations of BERT. It enables learning bidirectional contexts using Permutation Language Modeling as the training objective and integrates ideas from the autoregressive model Transformer-XL to model long text.

While the input to XLNET is similar to BERT, XLNET uses relative segment encoding instead of adding an absolute segment embedding to the word embedding at each position. Due to the time constraint, we only fine-tuned the XLNet model by adding a dense and softmax layer for NER on top of the last layer. We use the pre-trained XLNet-Large, Cased model in the BB task.

The result of NER is converted back to the standoff annotation format for normalization and relation extraction.

### 2.2 Normalization

In the BB normalization subtasks, our method is based on the vector representations of entities and identifiers.

For *Microorganism* entities that are normalized to taxa from the NCBI taxonomy, we apply the common TFIDF weighted sparse vector space representations (Salton and Buckley, 1988). This method treats each identifier as well as its curated classification and nomenclature information in the taxonomy as a document and gets the IDF weights based on such content. After that, each identifier and each entity is represented with a TFIDF weighted vector. According to the cosine distance between the vectors of identifiers and a given entity, the identifier with the highest cosine similarity will be assigned for the given entity. The scikit-learn library (Pedregosa et al., 2011) is used for TFIDF vectorization implementation.

For *Habitat* and *Phenotype* entities that are normalized to concepts from the OntoBiotope ontology, we use word embedding to represent both entity mentions and the ontology in a vector space.

There are several pre-trained biomedical word embeddings, such as PubMed-w2v (Pyysalo et al., 2013) and BioWordVec (Zhang et al., 2019). Based on the tests with the BioNLP-ST 2016 Evaluation Service (Deléger et al., 2016), we select the pubmed2018_w2v (McDonald, et al., 2018) 400-dimensional embeddings for the output vectors, which is the English word embeddings pre-trained on biomedical texts from MEDLINE/PubMed.

We then train a regression model to determine the similarity between the vectors of entities and the vectors of concepts. The model creates two training matrices for the vectors of entities and associated concepts respectively. After training with the BB corpus, the model will learn regression variables for predicting the similarity between new entities and concepts. We select the nearest concept as the ontology identifier for a given entity according to the cosine distance between the vectors of the concepts and the entity.

### 2.3 Relation Extraction

In the BB relation extraction subtasks, our method is based on the vector of a set of lexical features for classifying the relation types.

We use the Stanford CoreNLP toolkit (Manning et al., 2014) for sentence splitting and tokenization, as well as dependency parsing for each sentence. After parsing, the entity information is associated with the corresponding sentence. Since inter-sentence events still remain a challenge (Deléger et al., 2016), we focus on relations contained within a sentence. Only relations that occur entirely within a sentence will be associated with that sentence. For discontinuous entities in the BB corpus, we link each token overlapping with an entity's annotation to that entity. In addition, the sentence is also parsed to generate a dependency graph, which is represented as a set of two nodes and a dependency.

For every possible pair of entities within each sentence, we identify a possible relation with a class label. The relations annotated in the training data are tagged with the label "1" (denoting the *Lives_in* relation) or "2" (denoting the *Exhibits* relation). Other relations are tagged with the label "0" (denoting no relation). For each possible relation within a sentence, our method generates a vector from the features extracted, including the entity types, the unigrams between entities, the bigrams for the full sentence, and the edges in the dependency path.

We use the scikit-learn library to implement two multiclass classifiers: the support vector machine (SVM) and the logistic regression classifiers. For the SVM classifier, we use the linear kernel as it is

fast to train and has shown good performance. The set of relations in the training data is used to infer the possible argument types for each relation, and to filter the predicted set of relations.

### 2.4 Knowledge Base Extraction

In the BB knowledge base subtask, we use the above methods to recognize mentions from the given corpus, normalize the mentions according to domain knowledge resources, and extract relations between these mentions. The results are combined to build a knowledge base, which is the set of *Lives_in* and *Exhibits* relations with the concepts of their *Microorganism*, *Habitat* and *Phenotype* arguments.

## 3 Results & Discussion

The BB corpus contains PubMed references related to microorganisms and extracts from full-text articles related to microorganisms living in food products. In each subtask, it has been divided into three subsets for training, development and testing.

In BB subtasks, the official evaluation and the ranking of the submitted systems will be based on *Precision* for BB-norm, *F1* for BB-rel, *Slot Error Rate (SER)* for BB-norm+ner and BB-rel+ner, and *Mean References* for BB-kb and BB-kb+ner. Here we present the official results on the test sets. We submitted two runs for each subtask. For NER subtasks, the first run is based on the BERT+CRF model, fine-tuned using the hyperparameter values suggested in (Devlin et al., 2018): learning rate=2e-5, number of epochs=3, max sequence length=512, and batch size=8; the second run is based on the XLNET model with setting: batch size = 8, max length = 512, learning rate = 2e-5, num steps = 4,000. For normalization subtasks, the first run trains the regression model only with the training set of the normalization subtask while the second run trains the model with all training and development sets. For relation extraction subtasks, the first run uses the SVM classifier while the second run uses the logistic regression classifier.

As shown in Table 1, while the performance of our system is average compared to those of other teams in the BB-rel subtask, we ranked second among all participants in the BB-rel+ner, BB-norm and BB-norm+ner subtasks. Since no other teams participated in both normalization and relation extraction subtasks, we are the only team that can finish the knowledge base extraction subtasks and outperforms the baselines.

Our best runs also significantly outperformed the baselines in the BB-rel+ner and BB-norm subtasks, while the *Precision* of our best run in the BB-norm subtask is very close to the highest score (-0.0006). In addition, our system achieved the best *SER* for boundary accuracy of all three types of entities in the BB-norm+ner subtask, which demonstrates a good performance of our system in recognizing names of entities in a corpus for automatic knowledge extraction. However, our system performed poorly on entities new in test, which might be caused by the lack of generalization of the method or over-fitting of the machine learning model. After the release of golden standard results, we will conduct detailed error analysis to find out the actual reason and how each component variant contributes to the overall system performance.

| **Subtasks**<br>**Submissions** | **BB-rel**<br>**F1** | **BB-rel+ner**<br>**SER** |
|---|---|---|
| Our 1st run | 0.5495 | 1.0128 |
| Our 2nd run | 0.5943 | 1.0587 |
| 1st place system | 0.6639 | 0.9539 |
| Baseline | 0.6347 | 1.2109 |
| **Subtasks**<br>**Submissions** | **BB-norm**<br>**Precision** | **BB-norm+ner**<br>**SER** |
| Our 1st run | 0.6609 | 0.7931 |
| Our 2nd run | 0.6782 | 0.8059 |
| 1st place system | 0.6788 | 0.7160 |
| Baseline | 0.5310 | 0.8234 |
| **Subtasks**<br>**Submissions** | **BB-kb** | **BB-kb+ner** |
| | **Mean References** | |
| Our 1st run | 0.2907 | 0.2589 |
| Our 2nd run | 0.3077 | 0.2688 |
| Baseline | 0.2160 | 0.2642 |
| **Subtasks**<br>**Submissions** | **Habitats**<br>**NER** | **Microorganisms**<br>**NER** |
| | **SER** | |
| Our 1st run | 0.4787 | 0.3036 |
| Our 2nd run | 0.4639 | 0.3147 |
| 2nd place system | 0.5701 | 0.3428 |
| Baseline | 0.7702 | 0.6765 |
| **Subtasks**<br>**Submissions** | **Phenotypes NER**<br>**SER** | |
| Our 1st run | 0.4955 | |
| Our 2nd run | 0.6515 | |
| 2nd place system | 0.6378 | |
| Baseline | 0.8536 | |

Table 1: The BB task results comparison.

## 4 Conclusions

We described our system that participated in the Bacteria Biotope (BB) Task at BioNLP-OST 2019. Compared to previous works, our system has some significant differences from fundamental basis to the actual implementation of the model. It is comprehensive and has showed competitive performance among all participating systems during the BB evaluations. In future work, we will attempt supplemental approaches to tune our system to improve the robustness for unseen data and explore its use in practical applications such as biomedical knowledge bases construction. We also plan to make the codes available as open source.

## Acknowledgments

The authors would like to thank Dr. Yutao Zhong for providing Jihang Mao the summer research intern opportunity at George Mason University and valuable suggestions and comments on the manuscript.

## References

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. Bionlp shared task 2013–an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169.

Robert Bossy, Julien Jourde, Philippe Bessières, Maarten Van De Guchte, and Claire Nédellec. 2011. BioNLP Shared Task 2011: Bacteria Biotope. In *Proceedings of the BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics, pages 56–64.

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. 2019. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *Proceedings of the BioNLP Open Shared Task 2019 Workshop*. Association for Computational Linguistics.

Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4(1):1.

Aaron M Cohen and William R Hersh. 2005. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, 6(1):57–71.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop.* Association for Computational Linguistics, Berlin, Germany, pages: 12-22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, Minneapolis, Minnesota, USA, pages 4171–4186.

Melissa Merrill Floyd, Jane Tang, Matthew Kane, and David Emerson. 2005. Captured diversity in a culture collection: case study of the geographic and habitat distributions of environmental isolates held at the American type culture collection. *Applied and Environmental Microbiology*, 71(6):2813–2823.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Jake Lever and Steven JM Jones. 2016. VERSE: Event and relation extraction in the BioNLP 2016 Shared Task. In *Proceedings of the 4th BioNLP shared task workshop*, pages 42-49.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *In ACL (System Demonstrations)*. pages 55–60.

Jihang Mao, Wanli Liu. 2019. Factuality Classification using the Pre-trained Language Representation Model BERT. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings*, CEUR-WS, Bilbao, Spain, pages 126-131.

Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1849–1860.

Farrokh Mehryary, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2016. Deep learning with minimal training data: TurkuNLP entry in the

BioNLP shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 73-81.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM* (2013), pages 39-44.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report*, OpenAI.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24(5):513–523.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data* 6, no. 1 (2019): 52.