Bacteria Biotope at BioNLP Open Shared Tasks 2019

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba and **Claire Nédellec**
MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
robert.bossy@inra.fr louise.deleger@inra.fr
estelle.chaix@gmail.com mouhamadou.ba@inra.fr
claire.nedellec@inra.fr

Abstract

This paper presents the fourth edition of the Bacteria Biotope task at BioNLP Open Shared Tasks 2019. The task focuses on the extraction of the locations and phenotypes of microorganisms from PubMed abstracts and full-text excerpts, and the characterization of these entities with respect to reference knowledge sources (NCBI taxonomy, OntoBiotope ontology). The task is motivated by the importance of the knowledge on biodiversity for fundamental research and applications in microbiology. The paper describes the different proposed subtasks, the corpus characteristics, and the challenge organization. We also provide an analysis of the results obtained by participants, and inspect the evolution of the results since the last edition in 2016.

1 Introduction

In this paper, we present the fourth edition[1] of the Bacteria Biotope (BB) task. The task was introduced in 2011. It has the ambition of promoting large-scale information extraction (IE) from scientific documents in order to automatically fill knowledge bases in the microbial diversity field (Bossy et al., 2012). BB 2019 is part of BioNLP Open Shared Tasks 2019[2]. BioNLP-OST is a community-wide effort for the comparison and evaluation of biomedical text mining technologies on manually curated benchmarks.

A large amount of information about microbes and their properties that is critical for microbiology research and development is scattered among millions of publications and databases (Chaix et al., 2019). Information extraction as framed by the Bacteria Biotope task identifies relevant entities and interrelationships in the text and map them to reference categories from existing knowledge resources. This information can thus be combined with information from other sources referring to the same knowledge resources. The knowledge resources used in the BB task are the NCBI taxonomy[3] (Federhen, 2011) for microbial taxa and the OntoBiotope ontology[4] (Nédellec et al., 2018) for microbial habitats and phenotypes. The large size of these resources relative to the small number of training examples reflects the real conditions of IE application development, whilst it challenges current IE methods. The lexical richness of the two resources partially offsets the difficulty.

Compared to the 2016 corpus that contained only scientific paper abstracts from the PubMed database (Deléger et al., 2016), the 2019 corpus is enriched with extracts from full-text articles. We introduced a new entity type (phenotype) and a new relation type (linking microorganisms and phenotypes). Phenotypes are observable characteristics such as morphology, or environment requirement (e.g. acidity, oxygen). It is very valuable information for studying the ability of a given microbe to adapt to an environment (Brbić et al., 2016). The definition of microorganism phenotype in the OntoBiotope ontology includes host interaction characteristics (e.g. symbiont) and community behavior and growth habit (e.g. epilithic). The task organization and the evaluation metrics remain unchanged.

2 Task Description

The representation scheme of the Bacteria Biotope task contains four entity types:

- *Microorganism*: names denoting microorganism taxa. These taxa correspond to microorganism branches of the NCBI taxon-

[1] https://sites.google.com/view/bb-2019
[2] https://2019.bionlp-ost.org/
[3] https://www.ncbi.nlm.nih.gov/taxonomy
[4] https://tinyurl.com/OntoBiotope2019

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 121–131
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

omy. The set of relevant taxa is given on the BB task website.

- *Habitat*: phrases denoting physical places where microorganisms may be observed;
- *Geographical*: names of geographical places;
- *Phenotype*: expressions describing microbial characteristics.

The scheme defines two relation types:

- *Lives_in* relations which link a microorganism entity to its location (either a habitat or a geographical entity, or in few rare cases a microorganism entity);
- *Exhibits* relations which link a microorganism entity to a phenotype entity.

Arguments of relations may occur in different sentences. In addition, microorganisms are normalized to taxa from the NCBI taxonomy. Habitat and phenotype entities are normalized to concepts from the OntoBiotope ontology. We used the BioNLP-OST-2019 version of OntoBiotope available on AgroPortal [5]. We used the NCBI Taxonomy version as available on February 2, 2019 from NCBI website [6]. Copies of both resources can be downloaded from the task website. The microorganism part of the taxonomy contains 903,191 taxa plus synonyms, while the OntoBiotope ontology includes 3,601 concepts plus synonyms (3,172 for the Habitat branch and 429 for the Phenotype branch of the ontology).

Geographical entities are not normalized.

Figure 1 shows an example of a sentence annotated with normalized entities and relations.

As in the 2016 edition, we designed three tasks, each including two modalities, one where entity annotations are provided and one where they are not and have to be predicted.

2.1 Entity Normalization

The first task focused on entity normalization.

In the **BB-norm** modality of this task, participant systems had to normalize textual entity mentions according to the NCBI taxonomy for microorganisms and to the OntoBiotope ontology for habitats and phenotypes.

In the **BB-norm+ner** modality, systems had to recognize the mentions before normalizing them.

2.2 Relation Extraction

The second task focused on the extraction of the two types of relations— *Lives_in* relations among microorganism, habitat and geographical entities, and *Exhibits* relations between microorganism and phenotype entities.

In the **BB-rel** modality, participant systems only had to extract the relations, while in the **BB-rel+ner** modality they had to perform entity recognition in addition to relation extraction.

2.3 Knowledge Base Extraction

The goal of the third task is to build a knowledge base using the entities and relations extracted from the corpus. It can be viewed as the combination of the previous tasks, followed by a merging step. Participant systems must normalize entities and extract relations.

In the **BB-kb** modality, participant systems had to perform normalization and relation extraction with entity mentions being provided. In the **BB-kb+ner** modality, they had to perform entity recognition as well.

3 Corpus Description

3.1 Document Selection

The BB task corpus consists of two types of documents: PubMed references (titles and abstracts) related to microorganisms, and extracts from full-text articles related to beneficial microorganisms living in food products.

The PubMed references are the same as the 215 references of the Bacteria Biotope 2016 corpus. They were sampled from all PubMed entries indexed with a term from the Organisms/Bacteria subtree of the MeSH thesaurus. The full selection process is described in Deléger et al. (2016).

Full-text extracts were selected from scientific articles about microorganisms of food interest and annotated by microbiologist experts in the context of the Florilege project (Falentin et al., 2017). We reused and complemented this corpus for the BB task.

Because manual annotation is time-consuming and experts have limited time to dedicate to this task, they did not annotate the full articles. Instead, they chose the paragraphs and sentences they found the most informative in the articles. Thus, this part of the BB corpus is composed of 177 extracts of variable lengths (from one single

[5] http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE

[6] ftp://ftp.ncbi.nih.gov/pub/taxonomy

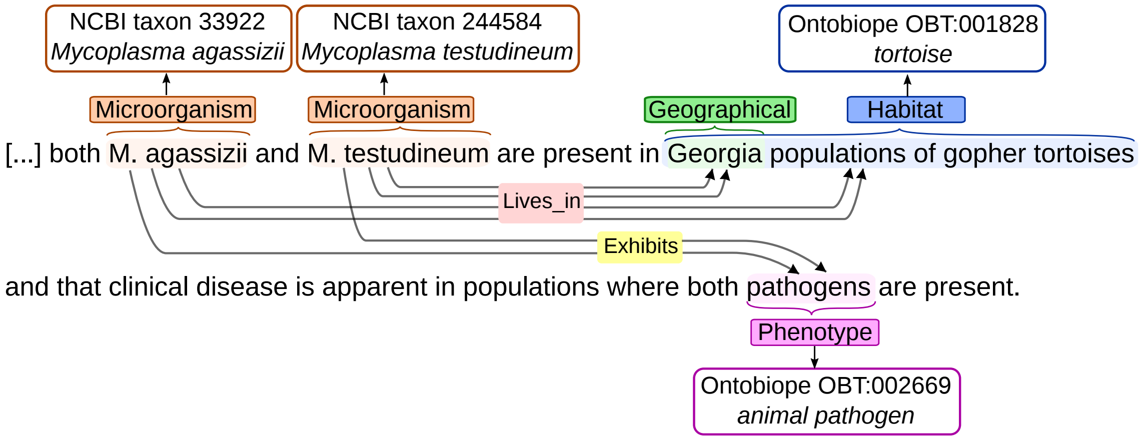

Figure 1: Annotation example

sentence to a few paragraphs) selected from 20 articles.

3.2 Annotation

The PubMed references were already annotated as part of the 2016 edition. We revised these annotations to add phenotype entities with their concept normalization and *Exhibits* relations. Habitat annotations were also revised to take into account the new and enriched version of the OntoBiotope ontology (compared to the 2016 version[7]).

We also extended the existing annotations of the full-text extracts of the Florilege project by assigning normalized concepts to the entities.

Annotation revision was performed by six annotators with backgrounds in biology, computer science and natural language processing. All documents were annotated independently by two annotators and disagreements were resolved through an adjudication phase. Detailed annotation guidelines (Bossy et al., 2019) were provided to the annotators and were regularly updated following issues raised during the annotation or adjudication phases.

The inter-annotator agreement was computed by evaluating one of the two annotations before adjudication against the other. Table 1 summarizes the inter-annotator agreement for named entities, normalization and relations. The metrics used for inter-agreement are the same as for the evaluation of predictions and thus are described below (5.1).

3.3 Descriptive Statistics

Table 2 gives the size of the corpus, in terms of documents, words, sentences and annotated elements. The last row shows the number of unique relations in the whole corpus, i.e. the unique pairs of microorganism and habitat/phenotype concepts that are in a relation. The proportion is rather high (1,931 out of a total of 3,578 occurrences), which reflects the rich information content of the corpus.

Named-entities (F1)	0.893
Normalization (semantic similarity)	0.974
Relations (F1)	0.786
BB-norm+ner evaluation (SER)	0.322
BB-norm+ner evaluation (F1)	0.823
BB-rel+ner evaluation (SER)	0.448
BB-rel+ner evaluation (F1)	0.765
BB-kb+ner evaluation	0.723

Table 1: Inter-annotator agreement metrics (SER stands for Slot Error Rate).

Documents	392
Words	60,402
Unique words	12,566
Sentences	2,646
Entity mentions	7,232
Unique entity mentions	3,300
Concepts	1,072
Relations	3,578
Unique relations between concepts	1,931

Table 2: Global statistics of the corpus

In the following, we present more detailed statistics and highlight corpus characteristics that may be challenging for the participants.

[7] http://2016.bionlp-st.org/tasks/bb2/OntoBiotope_BioNLP-ST-2016.obo

3.3.1 Entities and Concepts

Table 3 shows the number of mentions, unique (lemmatized) mentions, concepts and average number of mentions per concept for each entity type. Habitat entities are the most frequent, followed by Microorganism entities. Geographical entities are very scarce.

There is much more variation in the expression of habitats and phenotypes than in that of microorganisms. There is an average of respectively 4 and 3.5 unique mentions per habitat and phenotype concept while microorganisms only have 1.9. Their proportion of unique entities out of all mentions is also higher (respectively 50.6% and 45.2% vs. 38.2% for microorganisms).

The proportion of direct mappings (i.e., exact string matches, taking into account lemmatization) between entity mentions and labels of concepts (from the NCBI taxonomy or the OntoBiotope ontology) is displayed on Figure 2. It emphasizes once more the variability of Habitat and Phenotype entity expressions, with respectively 72.5% and 91.2% mentions that do not exactly match a concept label or synonym. Among exact matches, a small proportion of mentions are not actually normalized with the concept whose label they match. These are "contextual normalization" cases, i.e. entities are normalized with a more specific concept which can be inferred from the context. These often correspond to lexical coreference cases.

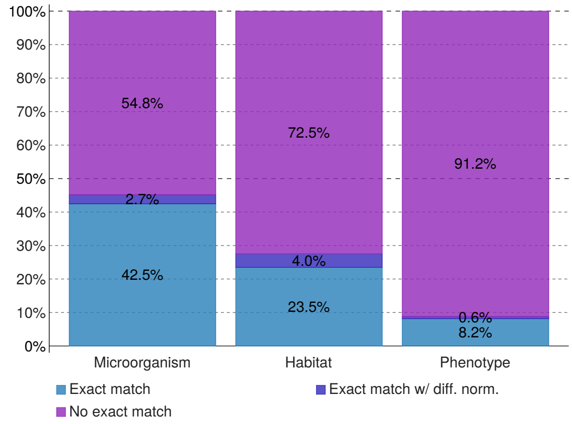

Figure 2: Exact matches between entity mentions and concepts. The *exact match* label refers to entities exactly matching the concept they are normalized with; the *exact match w/ diff. norm.* label refers to entities exactly matching a concept but normalized with a different one; the *no exact match* label refers to entities that do not match exactly a concept.

A distinctive feature of the BB task is that multiple concepts may be assigned to a given entity mention. Multiple normalization happens when two (or more) concepts can describe an entity and are all deemed necessary because each concept corresponds to a different aspect of the entity. An example of such a case is the Habitat entity *"diseased cow"* which is normalized by both the <*cow*> and <*animal with disease*> concepts. This is the case mainly for Habitat entities (8.7%), and rarely happens for Phenotype entities (0.6%) and Microorganism entities (only one occurrence).

Another characteristic of the corpus is the presence of nested entities (entities embedded in another larger entity) and discontinuous entities (entities split in several fragments). Both phenomena can be challenging for machine-learning methods and are often ignored. The proportion of discontinuous entities in the corpus is limited, with a total of 3.7%. Nested entities are more frequent (17.8% in total), especially for habitats. For instance, the Habitat entity *"cheese making factory"* also contains the smaller Habitat entity *"cheese"*.

3.3.2 Relations

Table 4 shows the number of relations for both *Lives_in* and *Exhibits* types, including intra-sentence and inter-sentence relations. Intra-sentence relations involve entities occurring in the same sentence while inter-sentence relations involve entities occurring in different sentences, not necessarily contiguous. Inter-sentence relations are known to be challenging for automatic methods. Their proportion in the corpus is not negligible (17.5% in total). An example can be seen in the following extract: *Vibrios [. . .] are ubiquitous to oceans, coastal waters, and estuaries. [. . .] The bacterial pathogen is a growing concern in North America*. There is an inter-sentence relation between the two underlined entities.

3.3.3 Training, Development and Test Sets

The BB corpus is split into training, development and test sets. In practice, there are two test sets, one for the modalities involving entity recognition (the "+ner" sub-tasks) and one for the modalities where entity annotations are given. We kept the corpus division of the 2016 edition for the PubMed references. This was possible because the gold annotations of the test set were never released to the public. Then we split the Florilege full-text extracts using the same proportions as for

	Microorganism	Habitat	Phenotype	Geographical
Entity mentions	2,487	3,506	1,102	137
Unique entity mentions	950	1,774	498	78
Concepts	491	440	141	N/A
Unique mentions per concept (average)	1.9	4.0	3.5	N/A

Table 3: Statistics for each entity type

	Intra-sent.		Inter-sent.		Total
Lives_In	2,099	(79.8%)	532	(20.2%)	2,631
Exhibits	852	(90.0%)	95	(10.0%)	947
Total	2,951	(82.5%)	627	(17.5%)	3,578

Table 4: Statistics for each relation type

the PubMed references. Figure 3 shows the distribution of documents, entities, concepts and relations in the training, development and test sets of the BB-kb+ner task, as an example. The proportions are similar in all sub-tasks. Details for each sub-task can be found on the task website[8].

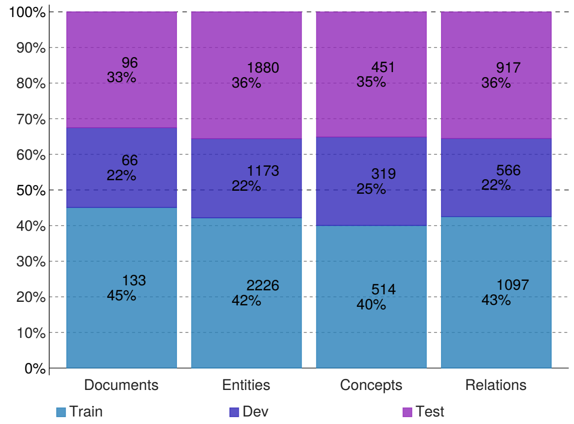

Figure 3: Distribution of documents, entities, concepts and relations in the training, development and test sets (BB-kb+ner task)

The proportion of concepts seen in the training set out of all concepts present in the knowledge resources is low for all entity types, which means that there is a large number of unseen examples (0.02% for microorganisms, 7.3% for habitats, and 15.6% for phenotypes). It emphasizes the need for methods that handle few-shot and zero-shot learning. Microorganisms have the lowest proportion, due to the large size of the microorganism taxonomies. However, the names of the microorganism entities show little variation in the corpus compared to habitat and phenotype types, and should be easier to recognize.

4 Supporting Resources

Supporting resources were made available to participants. They consist of outputs from state-of-the-art tools applied to the BB data sets (e.g., POS tagging, syntactic parsing, NER, word embeddings). We proposed in-house embeddings trained on selected relevant PubMed abstracts, and links to external embeddings (Pyysalo et al., 2013; Li et al., 2017) trained on PubMed and Wikipedia. The full list of tools and resources is available on the website.

5 Evaluation

5.1 Metrics

We used the same evaluation metrics as in the 2016 edition. The underlying rationale and formula of each score is detailed in Deléger et al. (2016); Bossy et al. (2013). Additionally we compute a variety of alternate scorings in order to distinguish the strengths of each submission. The evaluation tool was provided to participants[9].

Normalization accuracy is measured through a semantic similarity metric, and micro-averaging across entities. Relation extraction is measured with Recall, Precision, and F_1.

However for tasks where systems must recognize entities, we used the Slot Error Rate (SER) instead of F_1 in order to avoid sanctioning twice the inaccuracy of boundaries. The SER measures the amount of errors according to three types: insertions (false positives), deletions (false negatives), and substitutions (partial matches). The SER is normalized by the number of reference items. The higher the value the worse is the prediction, and there is no upper bound since insertions can exceed the number of items in the reference.

[8] https://sites.google.com/view/bb-2019/dataset/corpus-statistics

[9] https://github.com/Bibliome/bionlp-st

Confidence intervals were computed for each metric with the bootstrap resampling method (90%, n=100).

5.2 Baseline

We designed simple baselines for each sub-task in order to provide a comparison reference. We preprocessed the corpus with the AlvisNLP[10] engine, that performs tokenization, sentence splitting, and lemmatization using the GENIA tagger (Tsuruoka et al., 2005).

- BB-norm: we performed exact matching between lemmatized entities and the knowledge resources. When no match was found, we normalized habitats and phenotypes with the top-level concept of the Habitat and Phenotype ontology branches, and microorganisms with the high-level <*Bacteria*> taxon.
- BB-norm+ner: we used our exact matching approach on the lemmatized text of the documents instead of on given entity mentions.
- BB-rel: we used a simple co-occurrence approach, linking pairs of entities occurring in the same sentences.
- BB-rel+ner: we first detected entities using our exact matching strategy for microorganisms, habitats and phenotypes. For geographical entities, we used the Stanford Named Entity Recognition tool (Finkel et al., 2005). Then we linked entities occurring in the same sentences, as for the BB-rel task.
- BB-kb: we combined the BB-norm and BB-rel approaches.
- BB-kb+ner: we combined our BB-norm+ner method with our co-occurrence approach.

6 Outcome

6.1 Participation

The blind test data was released on the 22^{nd} of July 2019 and participants were given until the 31^{st} of July to submit their predictions. Each team was allowed two submissions to each sub-task.

Ten teams participated to all six sub-tasks and submitted a total of 31 runs. Table 5 details team affiliations. Teams are from five different countries in Europe, Asia, and North America. Six of the teams are affiliated to universities, three to industry companies, and one has a mixed university-industry affiliation.

Team	Affiliation
AliAI (Zhang et al., 2019)	Alibaba
Amrita_Cen	Amrita Vishwa Vidyapeetham
AmritaCen_healthcare	Amrita Vishwa Vidyapeetham
BLAIR_GMU (Mao and Liu, 2019)	George Mason University
BOUN-ISIK (Karadeniz et al., 2019)	Boğaziçi University & Işık University
MIC-CIS (Gupta et al., 2019)	Siemens AG & Ludwig Maximilian University of Munich
PADIA_BacReader (Deng et al., 2019)	Ping An Technology
UTU	University of Turku
whunlp (Xiong et al., 2019)	Wuhan University
Yuhang_Wu	Yunnan University

Table 5: Participating teams and their affiliations.

6.2 Participants' Methods and Resources

As in 2016, most methods are based on Machine Learning algorithms.

For named entity recognition, the CRF algorithm is still the most used (BLAIR_GMU), though sometimes combined with a neural network (MIC-CIS).

In 2016, the majority of participants used SVMs for relation extraction. In this edition nearly all participants used neural networks in a diversity of architectures: multi-layer perceptron (Yuhang_Wu), bi-LSTM (whunlp), AGCNN (whunlp). One participant predicted relations through filtered co-occurrences (BOUN-ISIK), and another by bagging SVM and Logistic Regression (BLAIR_GMU). Note that AliAI employed a multi-task architecture similar to BERT (Devlin

[10] https://bibliome.github.io/alvisnlp/

et al., 2019) to perform both named-entity recognition and relation extraction.

The normalization task was addressed in a more diverse manner. On one hand several distinct ML algorithms were used to discriminate entity categories: ensemble CNNs (PADIA_BacReader), kNN with reranking (BOUN-ISIK), or Linear Regression (BLAIR_GMU). On the other hand MIC-CIS employed an exact and an approximate matching algorithm.

Word embeddings trained with Word2Vec (Mikolov et al., 2013) on a domain-specific corpus (PubMed abstract, PMC articles) seem to be an universal resource since all but one submissions for any task used them. BLAIR_GMU used contextual embeddings based on BERT and XLNet (Yang et al., 2019).

Dependency parsing was used in every relation extraction submission, and also for normalization (BOUN-ISIK).

The most popular NLP tool libraries are Stanford CoreNLP (Manning et al., 2014) and NLTK (Bird et al., 2009). We also note that the WordPiece segmentation is used even in systems that do not use BERT.

6.3 Results

In this section we report the results for all subtasks, and highlight notable results as well as a comparison with results obtained in 2016 in the third edition of the Bacteria Biotope task in BioNLP-ST 2016. The task site presents detailed results, including main and alternate metrics, as well as confidence intervals.

However comparison with 2016 is limited by the evolution of the task. On one hand the data set has increased approximately by 50%, and the annotations were revised and their quality improved. On the other hand the tasks were made harder because the schema was enriched with an entity type and a relation type, and the target taxa have been extended from *Bacteria* only to all microorganisms.

6.3.1 BB-norm and BB-norm+ner

The main results as well as the results for each entity type are shown in Tables 6 and 7. BOUN-ISIK and BLAIR_GMU obtained the best overall results for BB-norm, and MIC-CIS for BB-norm+ner.

The results for each entity type highlight different profiles. While BOUN-ISIK predicts accurate normalizations for habitat entities for BB-norm, BLAIR_GMU predicts better normalizations for microorganism entities. PADIA_BacReader's predictions for habitats is on par with BOUN-ISIK, and their normalization of phenotype entities is outstanding.

As for BB-norm+ner, MIC-CIS consistently predicts the best entity boundaries and normalizations for all types.

In comparison to 2016, the state of the art for multi-word entity recognition and normalization, like habitats and phenotypes, has improved. We note that with the introduction of new taxa the recognition and normalization of taxa may have been rendered more difficult than anticipated since the results are lower than obtained in 2016.

6.3.2 BB-rel and BB-rel+ner

The results of BB-rel and BB-rel+ner are given in Tables 8 and 9 respectively. The table includes the scores obtained for each relation type, as well as the best results obtained in 2016.

The highest F-score for BB-rel was obtained by the whunlp submission, with AliAI as a very close contender. UTU, and very closely behind AliAI, obtained the highest Precision, whereas BOUN-ISIK the highest Recall. The Recall of the baseline prediction indicates the highest recall possible for relations contained in a single sentence. No participating system addresses cross-sentence relations, which appears to be the most productive lead to increase performance.

Most submissions outperform the best predictions of 2016 in at least one score, and five of the eleven submissions obtain a significantly higher F-score.

For BB-rel+ner, AliAI obtains the highest recall and precision, consistently for *Lives_In* and *Exhibits* relations. This submission also outperforms significantly the state of the art set in 2016.

6.3.3 BB-kb and BB-kb+ner

BLAIR_GMU is the only team to submit to the BB-kb and BB-kb+ner tasks, their results are shown in Table 10. The knowledge-base task and evaluation necessarily require end-to-end prediction systems that must perform named-entity recognition, entity normalization, relation extraction, as well as contributory tasks like POS-tagging, or coreference resolution. The limited scores obtained might be explained by the accumulation of errors by successive prediction steps.

Since the data of all sub-tasks comes from the

Team	All types	Habitats	Phenotypes	Microorganisms
Baseline	0.531	0.559	0.581	0.470
Best 2016	0.679	0.620		0.801
BOUN-ISIK-2	**0.679**	**0.687**	0.566	0.711
BLAIR_GMU-2	**0.678**	0.615	0.646	**0.783**
BOUN-ISIK-1	**0.675**	**0.687**	0.566	0.700
BLAIR_GMU-1	0.661	0.586	0.628	**0.783**
PADIA_BacReader-1	0.633	**0.684**	**0.758**	0.511
AmritaCen_healthcare-1	0.514	0.522	0.646	0.450

Table 6: Results for the BB-norm sub-task. The metric is the average of the semantic similarity between the reference and the predicted normalizations. Best scores are in bold font, several scores are in bold if their difference is not significant.

Team	All types	Habitat	Phenotype	Microorganism
Baseline	0.823	0.830	0.872	0.790
Best 2016	0.628	0.775		0.399
MIC-CIS-1	**0.716**	**0.728**	**0.747**	**0.686**
MIC-CIS-2	0.787	0.855	**0.759**	**0.715**
BLAIR_GMU-1	0.793	0.785	**0.775**	0.810
BLAIR_GMU-2	0.806	**0.722**	0.894	0.865
AmritaCen_healthcare-1	2.571	3.626	1.597	

Table 7: Results for the BB-norm+ner sub-task. The metric is the Slot Error Rate (lower is better) and takes into account false positives and negatives, entity boundary accuracy, and normalization accuracy. Best scores are in bold font, several scores are in bold if their difference is not significant.

	Average			Lives_In			Exhibits		
Team	F1	Recall	Precision	F1	Recall	Prec.	F1	Recall	Prec.
Baseline	0.635	0.801	0.525	0.621	0.767	0.521	0.677	0.915	0.538
Best 2016				0.558	0.646	0.623			
whunlp-1	**0.664**	0.702	0.629	**0.643**	0.664	0.624	**0.725**	**0.829**	0.644
AliAI-1	**0.650**	0.620	**0.682**	**0.648**	0.606	**0.697**	0.654	0.667	0.642
Yuhang_Wu-1	0.605	0.670	0.551	0.593	0.645	0.549	0.640	0.752	0.556
BOUN-ISIK-1	0.603	**0.731**	0.514	0.592	**0.709**	0.508	0.640	**0.808**	0.530
BLAIR_GMU-2	0.594	0.650	0.548	0.578	0.618	0.543	0.642	0.752	0.560
BOUN-ISIK-2	0.575	0.601	0.552	0.562	0.562	0.561	0.613	0.729	0.529
UTU-2	0.550	0.474	0.655	0.495	0.417	0.610	**0.715**	0.662	**0.777**
BLAIR_GMU-1	0.549	0.496	0.617	0.526	0.463	0.609	0.619	0.603	0.636
UTU-1	0.529	0.428	**0.694**	0.505	0.403	**0.679**	0.603	0.510	**0.738**
Amrita_Cen-1	0.500	0.617	0.420	0.499	0.643	0.407	0.503	0.531	0.478
Amrita_Cen-2	0.493	0.610	0.414	0.491	0.642	0.397	0.505	0.502	0.507

Table 8: Results for the BB-rel sub-task. Best scores are in bold font, several scores are in bold if their difference is not significant.

	Average			Lives_In			Exhibits		
Team	SER	Recall	Prec.	SER	Recall	Prec.	SER	Recall	Prec.
Baseline	1.211	0.134	0.229	1.266	0.171	0.228	1.211	0.134	0.229
Best 2016				0.984	0.111	0.498			
AliAI-1	**0.954**	**0.351**	**0.509**	**0.941**	**0.309**	**0.520**	**0.982**	**0.449**	**0.492**
BLAIR_GMU-1	1.013	**0.330**	0.456	1.020	**0.325**	0.451	**0.996**	0.339	**0.468**
BLAIR_GMU-2	1.059	**0.331**	0.425	1.046	**0.320**	0.435	1.086	0.358	0.406
UTU-1	1.085	0.209	0.332	1.091	0.182	0.307	1.069	0.272	0.382
UTU-2	1.227	0.182	0.267	1.169	0.168	0.279	1.362	0.217	0.249

Table 9: Results for the BB-rel+ner sub-task (Prec. = Precision). Best scores are in bold font, several scores are in bold if their difference is not significant.

same pool of annotated documents, we were able to build a BB-kb prediction by combining the best predictions for the BB-norm and BB-rel tasks. The combination of the microorganism normalization by BLAIR_GMU, the habitat and phenotype normalization by PADIA_BacReader, and relations by whunlp yield a much higher precision. The best result for BB-kb+ner was obtained by combining the relation extraction of BLAIR_GMU and the normalization of MICCIS. The named entities concurrently predicted by the BB-norm+ner and BB-rel+ner systems were matched by maximizing the overlap segment.

Team	BB-kb	BB-kb+ner
Baseline	0.216	0.264
Combined	0.505	0.290
BLAIR_GMU-2	**0.308**	**0.269**
BLAIR_GMU-1	**0.291**	**0.259**

Table 10: Results for the BB-kb and BB-kb+ner sub-tasks. The metric is the average of the semantic similarity between the reference and the predicted normalizations for all relation arguments after removing duplicates at the corpus level. Best scores are in bold font, several scores are in bold if their difference is not significant.

7 Conclusion

The Bacteria Biotope Task arouses sustained interest with a total of 10 teams participating in the fourth edition. As usual, the relation extraction sub-tasks (BB-rel and BB-rel+ner) were the most popular, demonstrating that this task is still a scientific and technical challenge. The most notable evolution of participating systems since the last edition is the pervasiveness of methods based on neural networks and word embeddings. These systems yielded superior predictions compared to those in 2016. As mentioned previously, there is still much room for improvement in addressing cross-sentence relation extraction.

We also note a growing interest in the normalization sub-tasks (BB-norm and BB-norm+ner). The predictions improved for habitat entities, and are very promising for phenotype entities. However the generalization from bacteria-only taxa in 2016 to all microorganisms in this edition proved to pose an unexpected challenge.

Knowledge base population (BB-kb and BB-kb+ner) is the most challenging task, since it requires a wider set of capabilities. Nevertheless we demonstrated that the combination of other sub-task predictions allows to produce better quality knowledge bases.

To help participants, supporting resources were provided. The most used resources were pre-trained word embeddings, and general-domain named entities.

The evaluation on the test set will be maintained online[11] in order for future experiments to compare with the current state of the art.

Acknowledgments

The authors thank Léonard Zweigenbaum for his contribution to the manual annotation of the corpus.

The current affiliation of Estelle Chaix is the French Agency for Food, Environmental and Occupational Health & Safety (ANSES), 14 rue Pierre et Marie Curie, 94701 Maisons Alfort Cedex, France.

[11] http://bibliome.jouy.inra.fr/demo/BioNLP-OST-2019-Evaluation/index.html

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media.

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, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten Van De Guchte, Philippe Bessières, and Claire Nédellec. 2012. BioNLP shared task–the bacteria track. In *BMC bioinformatics*, volume 13, page S3. BioMed Central.

Robert Bossy, Claire Nédellec, Julien Jourde, Mouhamadou Ba, Estelle Chaix, and Louise Deléger. 2019. Bacteria biotope annotation guidelines. Technical report, INRA.

Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. 2016. The landscape of microbial phenotypic traits and associated genes. *Nucleic acids research*, page gkw964.

Estelle Chaix, Louise Deléger, Robert Bossy, and Claire Nédellec. 2019. Text mining tools for extracting information about microbial biodiversity in food. *Food microbiology*, 81:63–75.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferre, Philippe Bessieres, and Claire Nedellec. 2016. Overview of the bacteria biotope task at BioNLP shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 12–22.

Pan Deng, Haipeng Chen, Mengyao Huang, Xiaowen Ruan, and Liang Xu. 2019. An ensemble CNN method for biomedical ontology alignment. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hélène Falentin, Estelle Chaix, Sandra Derozier, Magalie Weber, Solange Buchin, Bedis Dridi, Stéphanie-Marie Deutsch, Florence Valence-Bertel, Serge Casaregola, Pierre Renault, Marie-Christine Champomier-Verges, Anne Thierry, Monique Zagorec, Francoise Irlinger, Céline Delbes, Sophie Aubin, Philippe Bessieres, Valentin Loux, Robert Bossy, Juliette Dibie, Delphine Sicard, and Claire Nédellec. 2017. Florilege : a database gathering microbial phenotypes of food interest. In *Proceedings of the 4th International Conference on Microbial Diversity 2017*, pages 221–227, Bari, Italy. Poster.

Scott Federhen. 2011. The NCBI taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Pankaj Gupta, Usama Yaseen, and Hinrich Schütze. 2019. Linguistically informed relation extraction and neural architectures for nested named entity recognition in BioNLP-OST 2019. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

İlknur Karadeniz, Ömer Faruk Tuna, and Arzucan Özgür. 2019. BOUN-ISIK participation: An unsupervised approach for the named entity normalization and relation extraction of bacteria biotopes. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating different syntactic context types and context representations for learning word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2421–2431.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: system demonstrations*, pages 55–60.

Jihang Mao and Wanli Liu. 2019. Integration of deep learning and traditional machine learning for knowledge extraction from biomedical literature. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Claire Nédellec, Robert Bossy, Estelle Chaix, and Louise Deléger. 2018. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. *arXiv preprint arXiv:1805.04107*.

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

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*, pages 382–392. Springer.

Wuti Xiong, Fei Li, Ming Cheng, Hong Yu, and Donghong Ji. 2019. Bacteria biotope relation extraction via lexical chains and dependency graphs. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

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*.

Qi Zhang, Chao Liu, Ying Chi, Xuansong Xie, and Xiansheng Hua. 2019. A multi-task learning framework for extracting bacteria biotope information. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.