Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016

[1]Louise Deléger, [1]Robert Bossy, [1]Estelle Chaix, [1]Mouhamadou Ba, [1,2]Arnaud Ferré,
[1]Philippe Bessières, [1]Claire Nédellec
[1]MaIAGE, INRA, Universit Paris-Saclay, 78350 Jouy-en-Josas, France
[2]LIMSI, CNRS, Universit Paris-Saclay, 91405 Orsay, France
firstname.lastname@jouy.inra.fr

Abstract

This paper presents the Bacteria Biotope task of the BioNLP Shared Task 2016, which follows the previous 2013 and 2011 editions. The task focuses on the extraction of the locations (biotopes and geographical places) of bacteria from PubMed abstracts and the characterization of bacteria and their associated habitats with respect to reference knowledge sources (NCBI taxonomy, OntoBiotope ontology). The task is motivated by the importance of the knowledge on bacteria habitats for fundamental research and applications in microbiology. The paper describes the different proposed subtasks, the corpus characteristics, the challenge organization, and the evaluation metrics. We also provide an analysis of the results obtained by participants.

1 Introduction

Since 2009, BioNLP Shared Task is a community-wide effort on the development of fine-grained information extraction methods in biomedicine (Kim et al., 2009; Kim et al., 2011; Nédellec et al., 2013). The tasks provide a sound framework for the comparison and evaluation of the technologies on a manually curated benchmark with the aim to contribute to progress by drawing general lessons from the individual contributions and assessment of the participants. In this paper, we present the third edition of the Bacteria Biotope task that has been first introduced in 2011 with the ambition to use information extraction from scientific documents at a large scale in order to automatically fill knowledge bases (Bossy et al., 2012).

Information about bacteria biotopes (*e.g.,* habitats of bacteria) is critical for studying the interaction and association mechanisms between organisms and their environments from genetic, phylogenetic and ecological points of view. This information is not only highly useful in all fields of applied microbiology such as food processing and safety, health sciences and waste processing, but also in fundamental research (*e.g.*, metagenomics, phylogeography, phyloecology).

Currently, there is no centralized resource gathering the state of knowledge on habitats of bacteria in a comprehensive and normalized way. A large part of this knowledge is scattered in numerous scientific papers and databases, such as genomics databases (*e.g.*, GenBank[1], GOLD[2]), international microorganism culture collections (*e.g.*, ATCC[3], DSMZ[4]), and biodiversity surveys (*e.g.* , GBIF[5]). The information on bacteria biotopes is mostly expressed in free text (*e.g.*, articles or free-text fields of databases) describing very diverse locations (any physical location may be a bacteria habitat) in many different ways. The need for information processing is not only the extraction of habitats and microorganisms relationships from text, but also their normalization with respect to a common referential so that they can be integrated and compared. This need has been acknowledged by previous work on habitat classifications for metagenomic samples (Ivanova et al., 2010), microorganisms (Floyd et al., 2005) and other living organisms (Buttigieg et al., 2013) and text-mining tools for mapping textual descriptions to habitat classification (Pignatelli et al., 2009).

The aim of Bacteria Biotope (BB) task is to provide a framework for the evaluation and compari-

[1]http://www.ncbi.nlm.nih.gov/genbank
[2]https://gold.jgi.doe.gov/ (Genomes Online Database)
[3]http://www.atcc.org/ (American Type Culture Collection)
[4]https://www.dsmz.de/ (Deutsche Sammlung von Mikroorganismen und Zellkulturen)
[5]http://www.gbif.org/ (Global Biodiversity Information Facility)

Proceedings of the 4th BioNLP Shared Task Workshop, pages 12–22,
Berlin, Germany, August 13, 2016.

son of such methods for Bacteria organism habitats. More specifically, the BB task consists in the extraction of bacteria and their locations (habitats or geographical places) from the text, their categorization according to dedicated knowledge sources, and the linking of bacteria to their locations through so-called localization events named "Lives_in". The widely used NCBI taxonomy[6] (Federhen, 2012) is the resource used for Bacteria entity categorization. The OntoBiotope ontology[7], which is dedicated to the description of microorganism habitats, is used for biotope categorization. Previous work has shown the relevance of OntoBiotope for bacteria habitat detection (Ratkovic et al., 2012). The first two editions of the task (Bossy et al., 2012; Bossy et al., 2015) used general-purpose documents, mostly web pages of genomics projects that can be understood by non-specialists. However, scientific literature is the major source of detailed and accurate information on bacteria for biologists. This edition focuses then on scientific paper abstracts from the PubMed database, which offers a twofold advantage, open access and easier readability than full-text. We also introduce this year a new subtask of knowledge base extraction, in which systems are evaluated by measuring how much information content can be extracted from the corpus.

2 Task Description

The BB task involves three types of entities, Bacteria, Habitats and Geographical places. It also involves a single type of event, the *Lives_in* event, which is a relation between two mandatory arguments, the bacterium and the location where it lives, either a *Habitat* or a *Geographical* entity. Figure 1 displays an example of entities and events in the BB task.

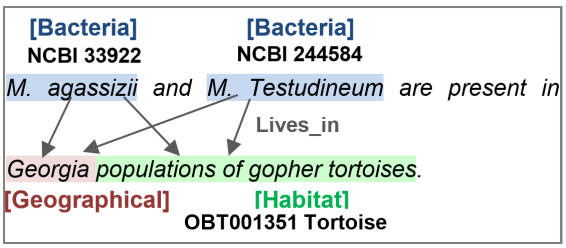

Figure 1: Example of entities and *Lives_in* events in the BB task

[6] http://www.ncbi.nlm.nih.gov/taxonomy

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

We proposed three subtasks with two modalities each. Each subtask had a plain modality where named entities were given as input, thus participants were not required to perform entity recognition. In the second modality, entities were *not* provided, thus methods had to perform named entity recognition and submissions are partly evaluated on the accuracy of entity boundaries. Our purpose is to assess independently the quality of the methods when dealing with different sub-goals and to assess the impact of predictions made at a given step on the predictions made at the next steps.

2.1 Bacteria and Habitat Categorization

The first subtask focused on the categorization of *Bacteria* and *Habitat* entity occurrences in the text with categories from the NCBI Taxonomy for *Bacteria* and from the OntoBiotope ontology for *Habitat* entities. In the first modality of the subtask (referred to as BB-cat), entity mentions were given and participants had only to perform categorization. In the second modality (BB-cat+ner), systems had to perform bacteria and habitat entity detection as well as categorization.

2.2 Entity and Event Extraction

The second subtask consists in the extraction of *Lives_in* events among *Bacteria*, *Habitat* and *Geographical* entities. In the BB-event modality, entity mentions were given and participant systems only had to perform event extraction. In the BB-event+ner modality, systems had to perform *Bacteria*, *Habitat* and *Geographical* entity recognition as well as event extraction.

2.3 Knowledge Base Extraction

The third subtask aims at building a knowledge base using information extracted from the corpus. The knowledge base is composed of the set of distinct *Lives_in* events between categorized *Bacteria* and *Habitats*. This subtask can be seen as a combination of the entity categorization and event extraction subtasks. In contrast with the two previously described subtasks, this task does not evaluate text-bound annotations. All pieces of information extracted from the text are gathered and merged into a single knowledge base, without duplicate events. The focus of this task is the knowledge itself (which types of bacteria and habitat are linked through a *Lives_In* event) and not the individual text-bound annotations (where *Lives_In* events are marked precisely in each text segment).

In the first modality, BB-kb, entity mentions were given and participating systems perform categorization and event extraction. In the second modality, BB-kb+ner, systems had to perform *Bacteria* and *Habitat* entity detection and categorization as well as event extraction.

3 Corpus Construction

3.1 Corpus Selection

The BB corpus consists of titles and abstracts of PubMed entries. We followed a four step procedure to build a representative reference corpus for the task from the whole PubMed database. It started from the set of all PubMed references and successively selected a subset of references while preserving the distribution of bacteria and habitat categories.

In the first step, we selected PubMed entries relevant to bacteria, relying on the MeSH index provided by the NLM. We selected all entries that were indexed by any term of the Organisms/Bacteria subtree (B03). PubMed contained 27,872,481 entries, of which 1,156,824 indexed by a term in the Bacteria subtree (4%).

In the second step, we automatically annotated *Bacteria*, *Geographical* and *Habitat* entities in the title and abstract of these selected entries (see the corpus annotation subsection for details about this automatic approach). We found 6.8 million habitat occurrences, 3.7 million occurrences of bacteria taxon names, and 374 thousand geographical names. This gave us a broad idea of the quantity and diversity of the entries in terms of bacterial taxa and bacterial habitats.

However this collection is too large to be manageable by human annotators. Therefore in the third step, we built a sub-collection of 1,000 entries. We selected the most representative in 2,000 random samples of 1,000 entries. The representativeness was evaluated by the mean squared error (MSE) between the sample and the original collection. We selected the sample with the lowest MSE. The observations from which we computed the MSE included the number of words, the number of occurrences of taxon names for each bacterial family and the number of occurrences of habitat mentions for each top-level concept of OntoBiotope. As expected from a PubMed sample, the majority of entries were biomedical studies. Even though habitats related to human health and welfare are important, the sample does not convey the full diversity of bacteria habitats.

In the fourth step we manually annotated the title and abstract of references from the sample (see section 3.2). As it would require too much human resources the manual annotation of 1,000 PubMed entries is not an option. We randomly picked entries as we finished annotating the previous ones in order to preserve the distribution. The random selection used the same method as the sampling, however we deliberately biased against clinical habitats in order to leave room for more diverse and less frequent habitats.

3.2 Corpus Annotation

Manual annotation was performed by seven annotators with diverse backgrounds: biology, computer science, linguistics, and bioinformatics. Three annotators had annotated documents in the previous editions of the BB task. Each document was annotated by two annotators in a double-blind manner and an adjudication phase resolved disagreements. Annotators relied on detailed guidelines which were revised and clarified when questions arose during the annotation process. The guideline document is available on the BB task website[8].

Annotators used the AlvisAE annotation editor (Papazian et al., 2012). In order to speed up the annotation process, we used Alvis Suite (Ba and Bossy, 2016) to automatically pre-annotate the corpus. It included the Stanford NER tool (Finkel et al., 2005) to annotate *Geographical* locations and the ToMap method (Golik et al., 2011) to detect and categorize *Habitat* entities. *Bacteria* entity automatic recognition and categorization were performed with a rule-based approach relying on a customized dictionary of taxon names, *i.e.* NCBI taxonomy names augmented with typographical variations. Events were extracted using manually defined trigger words and rules in a similar way as Ratkovic et al. (2012). Table 1 gives pre-annotation performance for habitat and bacteria recognition and categorization (cat+ner) and for entity and *Lives_In* event extraction (event+ner). Performance is low, especially for event extraction, which calls into question the benefit of using automatic pre-annotation for these tasks. The low performance of pre-annotation compared to the final gold standard is also an indication that text pre-annotation did not much bias manual an-

[8] http://2016.bionlp-st.org/tasks/bb2

notation, since the annotators did not hesitate to make extensive changes in the pre-annotation. We computed the inter-annotator agreement by comparing the individual manual annotations with the consensus, using the same evaluation framework as for the evaluation of participant systems (Section 5). We did not compute any Kappa statistics, since this type of metric is not well-suited for the annotation of textual entities (Hripcsak and Rothschild, 2005). Moreover, even in the case of event annotation, computing Kappa would have been difficult, because event annotation is based on entity annotation. Table 2 shows the agreement of entity boundaries and categorization computed with BB-cat+ner scores and the agreement of entity boundaries and *Lives_In* events computed with BB-event+ner scores. The high precision demonstrates that there was not much disagreement among annotators on the entity boundaries and categorization, or in the *Lives_In* events. The consensus consisted mostly in annotation merging. The lower recall stresses the necessity of multiple annotators in order to ensure that the reference is complete.

	SER	Recall	Precision	F1
cat+ner	1.167	0.287	0.341	0.312
event+ner	1.749	0.187	0.158	0.171

Table 1: Pre-annotation performance (SER = Standard Error Rate)

	Recall	Precision	F1
Entity recog.	0.621	0.955	0.753
Lives_In event	0.311	0.952	0.468

Table 2: Inter-annotator agreement

3.3 Corpus Statistics

Tables 3, 4 and 5 provide descriptive statistics of the corpus for the three subtasks respectively.

They show the distributions of entities, categories, and events among the different datasets (training, development and test) of each subtask. We analyzed these statistics in order to study the characteristics of the BB corpora with respect to the tasks. Each distinct entity surface form has only two occurrences on average in the corpus, which makes the recognition task more difficult than with highly repeated mentions: there are 1,489 and 1,466 distinct entity mentions (*i.e.*, strings or surface forms) out of a total 2,887 and 2,842 annotated entity mentions in the BB-cat and BB-cat+ner datasets, respectively (see Table 3). In comparison, there is less variety in entity categories, since the number of distinct categories is only 519 out of a total of 3,189 occurrences. The combination of these two observations indicates that there is quite a lot of variation in the surface forms of entities, *i.e.,* the same category can be expressed in several different ways in the text. This is particularly true for *Habitat* entities for which there is a higher proportion of distinct surface forms than for *Bacteria* names (59% *vs.* 38% in the combined BB-cat datasets).

Additionally, we computed the proportion of direct mappings (*i.e.*, exact string matches) between *Habitat* surface forms from the training and development datasets of BB-cat and BB-cat+ner and the ontology labels. We found that respectively 24% and 27% Habitat entity occurrences exactly matched with an ontology label. As expected, proportions were similar in the test sets of these two tasks, with respectively 25% and 27% exact matches. This finding emphasizes the fact that there is much variation in the expression of *Habitat* entities, and thus simple methods based on exact string matching are not sufficient to automatically categorize entities with high quality.

Multiple categories may be assigned to a given entity mention, as can be seen in Table 3, which is more challenging than single categories. This is the case mainly for *Habitat* entities, since there is a total of 1,921 distinct *Habitat* entities for a total of 2,221 assigned *Habitat* categories in the BB-cat datasets.

The number of *Geographical* entities in the BB-event+ner sets is much lower than the other entity types with 101 *Geographical* entities only in total, which may make machine-learning approaches less efficient for this type of entity.

Not surprisingly, the majority of *Lives_in* events links *Bacteria* entities to *Habitat* entities and only a small number of events involves *Geographical* entities in the BB-event datasets (*e.g.*, 98 out of a total of 890 events (11%)).

Table 4 also shows the number of intra-sentence *vs.* inter-sentence events, *i.e.* events that involve entity arguments occurring in the same sentence *vs.* events that involve entities occurring in different sentences. The proportion of inter-sentence events is still significant (27%). Methods restricted to the extraction of sentence-level events would suffer from a serious disadvantage.

However the extraction of inter-sentence events is a major challenge, since they are notably more difficult to predict and may require co-reference resolution.

Table 5 details statistics for the knowledge base extraction subtask (BB-kb and BB-kb+ner). Its goal is to build a knowledge base composed of all distinct pairs of Bacteria and Habitat categories linked through the Lives_in relation that can be extracted from the corpus. The number of linked pairs of distinct categories is high with respect to the total number of pairs. There are 185 distinct events out of a total of 312 events in the test set of the BB-kb task (last row of Table 5). This reflects the richness of the information content of the corpus.

4 Shared Task Organization

The BB task schedule was divided in a training period of two months and a test period of twelve days. After the test, detailed evaluation of the system performances was provided to the participating teams and published on the BioNLP-ST 2016 website.

Supporting resources were made available to the participants. These resources are the output of state-of-the-art automated corpus analysis tools applied to the BB datasets. They were generated in the same way as for the SeeDev task of BioNLP-ST (see Chaix et al. (2016) for further details). In addition to the information available on the website, we maintained a set of community web tools. They included a dedicated forum that allowed participants to interact directly with each other and with the organizers, and an online evaluation service the participants could use to evaluate their predictions during the training phase. This service also keeps track of multiple runs allowing participants to monitor their experiments and to compare their predictions to other participant predictions in an anonymous way.

5 Evaluation

The metrics used to evaluate systems depend on the subtasks. When possible we reused metrics from the previous editions so that the results remain comparable.

5.1 BB-cat and BB-cat+ner

BB-cat. For each entity the metrics measures the similarity between the reference category and the predicted category. The overall score is equal to the mean of the similarities for all entities. For *Bacteria* entities the similarity is defined as follows, if the predicted taxon identifier is identical to the reference taxon identifier, then it is set to 1, otherwise 0. For *Habitat* entities we used the same similarity measure as for the 2013 edition of the BB task (Bossy et al., 2013): it is the semantic similarity defined by Wang et al. (2007) with the weight parameter set to 0.65.

BB-cat+ner. The BB subtask was evaluated using the Slot Error Rate (SER), the same method as BioNLP-ST 2013 BB task 1 (Bossy et al., 2013) since the two tasks are the same.

5.2 BB-event and BB-event+ner

The metrics for the evaluation of the BB-event and BB-event+ner subtasks are recall, precision and F-score as for BioNLP-ST 2013 BB task 2 and 3 for the same reasons (Bossy et al., 2013).

5.3 BB-kb and BB-kb+ner

The evaluation of BB-kb submissions is based on the comparison of the reference knowledge base to the one that each participant system has built. The knowledge base associates bacterial taxa with habitat categories. The taxon-habitat category associations are obtained from text-bound *Lives_In* event arguments assigned to taxa and habitat categories. Duplicate associations are removed to generate the knowledge base so that a single association remains between a given taxon and a given habitat category. We applied this procedure to the set of reference events and categories to generate the reference knowledge base and to the events and categories predicted by the participant systems in the same way.

The goal of the BB-kb is to assess how much knowledge a system can extract from a collection of documents. The measure of the exact match between the predicted knowledge base and the reference knowledge base would be too strict and would not satisfy this goal. Thus we designed a measure that evaluate the *similarity* between the two knowledge bases

Each predicted association is paired to the closest reference association using the similarity functions of BB-cat. This process results in each reference association paired to zero (false negative), one, or several predicted associations. Then we can measure the accuracy by which each reference association was found. If the association is not

paired to any prediction, then its accuracy is zero, otherwise the accuracy is the mean of the similarity to each prediction. The submissions are evaluated by the mean accuracy for each reference association (mean references). The " mean references" score computes how much the predicted knowledge base maps into the reference knowledge base.

Since the evaluation does not rely on text-bound annotations, the BB-kb+ner was evaluated with the same metrics as BB-kb.

6 Results

A total of 14 teams participated in Bacteria Biotope 2016. They were from several countries: Turkey (BOUN), France (LIMSI), Denmark (TagIt), Canada (VERSE), Finland (TurkuNLP, UTS, UMS) and China (DUTIR, WhuNlpRE, HK, whunlp, WXU). Two participants retracted their submissions (they correspond to blank lines in result tables). We present the results obtained by the participating teams. Detailed results are available on the task page[9].

6.1 Performance on BB-cat / BB-cat+ner

The results of systems that participated to BB-cat (2 teams) and BB-cat+ner subtasks (3 teams) are given in Table 6 and 7, respectively.

Team	Prec. *all*	Prec. *Bacteria*	Prec. *Habitat*	Prec. *Multi cat.*
BOUN	**0.679**	**0.801**	**0.620**	0.486
LIMSI	0.503	0.637	0.438	**0.516**

Table 6: Team results for the BB-cat task (" Prec." = " Precision"; " Multi cat." = " Multiple categorizations")

BOUN achieved the best performance for the categorization task (BB-cat) with 0.679 precision. As expected, performance was much higher for the categorization of Bacteria entities (0.801 for the best precision) than for that of Habitat entities (0.62). Bacteria are usually referred to using names from the NCBI taxonomy with a few variations, while Habitats are mainly noun and adjectival phrases that are expressed in many ways and may be very different from their concept label form. Moreover, Habitats may be categorized using several ontology concepts, which creates an additional difficulty. The last column of Table 6 shows results for multiple categorization cases. The LIMSI team obtained stable performance while the BOUN team performed significantly lower than for all entities (0.486 *vs*. 0.679).

When taking into account entity recognition in addition to categorization (BB-cat+ner, Table 7), TagIt achieved the best SER (0.628), and the difference between the top and last teams is significant (0.27 points). As for the BB-cat task, systems performed better on Bacteria entities than on Habitat entities. We also assessed the performance of entity recognition (without taking into account categorization), *i.e.*, systems are evaluated for their ability to predict entity boundaries in the text (see the bottom part of Table 7). The results of boundary detection also reflect the difference in difficulty between *Habitat* and *Bacteria* entities.

Compared to the Bacteria Biotope 2013 edition, the performance seems to have dropped. The best SER for *Habitat* entity recognition and categorization was 0.661 (Bossy et al., 2015), while it is 0.775 this year. This may be due to the change of document source, *i.e.,* scientific dense documents instead of general purpose web pages. It may also be due to the higher proportion of cases of multiple category assignments, while these cases remained marginal in the 2013 edition. Another reason may be the high number of clinical studies where the distinction between categories (*e.g.*, treated and non-treated patients, pediatric and adult patients) may require a more thorough analysis of the event context. Therefore the task also entails co-reference resolution.

		TagIt	LIMSI	whunlp
Overall	SER	**0.628**	0.827	0.901
	Recall	**0.456**	0.361	0.273
	Precision	**0.612**	0.486	0.407
Bacteria	SER	**0.399**	0.771	0.823
	Recall	**0.692**	0.539	0.397
	Precision	**0.857**	0.623	0.637
Habitat	SER	**0.775**	0.862	0.950
	Recall	**0.303**	0.246	0.193
	Precision	**0.430**	0.371	0.275
Bacteria boundaries	SER	**0.236**	0.277	0.436
	Recall	**0.772**	0.751	0.565
	Precision	**0.954**	0.903	0.893
Habitat boundaries	SER	0.599	**0.597**	0.627
	Recall	0.476	**0.504**	0.493
	Precision	0.675	**0.728**	0.690

Table 7: Team results for the BB-cat+ner task

6.2 Performance on BB-event / BB-event+ner

Among subtasks, the event extraction subtask (more specifically the BB-event task) attracted the most participants, with a total of eleven differ-

[9] http://2016.bionlp-st.org/tasks/bb2/bb3-evaluation

ent teams, three of which participated in the BB-event+ner subtask and eleven in the BB-event subtask. Tables 8 and 9 show team performances on BB-event and BB-event+ner tasks respectively.

VERSE obtained the highest F1 score for the BB-event task (0.558). The difference between the top and last teams is only 0.10 points and participants ranked 4th to 11th obtained very similar results (ranging from 0.474 to 0.455 F1 score). All participants achieved better performance when predicting *Lives_in* events with *Geographical* arguments than events with *Habitat* arguments (5th and 6th columns of Table 8), although events with *Geographical* arguments are less frequent. The reason could be that most of *Geographical* entities are linked to a *Bacteria* entity, which makes the decision easier than for *Habitat* entities, for which there are many occurrences that are not involved in any *Lives_in* event.

Not surprisingly systems had less trouble predicting intra-sentence events than inter-sentence events, as all yielded significantly higher F1 score on intra-sentence events (see last column of Table 8). Detailed analysis of the predictions made by the systems shows that LIMSI was the only team to consistently predict inter-sentence events. Other systems predicted roughly the same number of events when considering only intra-sentence events or all events together in the evaluation.

There is a drastic drop in performance when adding entity recognition to the event extraction task (BB-event+ner task, see Table 9). All three participating teams obtained very similar results in terms of F1 score, although the balance between precision and recall differs. The LIMSI team (ranked 1st) achieved a perfect balance between precision and recall, while UTS and the WhuNlpRE team obtained much higher precision but lower recall. As for the BB-event task, performances are significantly higher for *Lives_in* events involving *Geographical* entities, and intra-sentence events.

For both tasks, systems performed better in average than in the 2013 edition. Indeed, the best F1 scores (Bossy et al., 2015) were 0.49 for the detection of localization events (*vs.* 0.558 for *Lives_in* events in this edition) and 0.14 for the combination of entity recognition and event extraction (vs. 0.19). This suggests that participant methods have improved and become more accurate. However, the F1-score for BB-event+ner remains relatively low, which directly results from the combined complexity of the two sub-problems in the same task.

6.3 Performance on BB-kb / BB-kb+ner

Only the LIMSI team participated in the knowledge base extraction subtask. Results are given in Table 10 for both the BB-kb and BB-kb+ner tasks. The LIMSI system for BB-cat (Table 6) and BB-event (Table 8) provides a good reconstruction of the knowledge base (BB-kb) which highlights the fact that automatic categorization and event extraction methods are already efficient for the task of knowledge base construction. However, the performance is significantly lower when reference entities are not provided. This large gap in performance may be explained by the difficulty of recognizing entities (as also shown in the BB-cat+ner task), and the fact that a fair amount of entities is not repeated in the corpus. Consequently the false negatives in entity detection have a strong impact on the end-to-end task of knowledge base construction.

	LIMSI	UTS	WhuNlpRE
F1	**0.192**	0.190	0.182
Recall	**0.191**	0.133	0.111
Precision	0.193	0.331	**0.498**
F1 (Habitat)	0.186	0.174	**0.196**
F1 (Geographical)	0.283	**0.350**	NA
F1 (Intra-sentence)	**0.286**	0.234	0.232

Table 9: Team results for the BB-event+ner task

	BB-kb	BB-kb+ner
LIMSI	0.771	0.202

Table 10: Results for BB-kb and BB-kb+ner (mean-references measure)

6.4 Systems

Systems used different resources and methods depending on the sub-tasks.

Entity Detection and Categorization. Systems used dictionary-based (TagIt) and machine-learning based (LIMSI, WhuNlpRE, UTS) methods to detect entity mentions in text in the BB-cat+ner and BB-event+ner subtasks. All relied on existing terminology and ontology resources, including the NCBI Taxonomy, the List of Prokaryotic Names with Standing in Nomenclature (Parte, 2013), the Brenda Tissue Ontology (Gremse et al., 2011), the Environment Ontology (Buttigieg et al.,

2013), the OntoBiotope ontology, and WordNet (Fellbaum, 1998). The TagIt system performed dictionary matching coupled with acronym detection and heuristic rules to adjust entity boundaries. The LIMSI team used conditional random fields (CRFs) and the WhuNlpRE team used neural networks. Both these teams generated rich features for their machine-learning algorithms: lexical, morpho-syntactic, dictionary projection, existing named entity recognition tools, Brown clustering, and word embeddings. The UTS team relied on Support Vector Machines (SVM) with features based on the output of existing NER tools provided by the organizers as supporting resources. The rule-based approach of the TagIt system achieved the highest performance in entity detection and categorization (BB-cat+ner), although the CRF approach of the LIMSI system was the most accurate in Habitat boundary detection.

Teams relied on rule-based (TagIt, LIMSI) and similarity-based (BOUN) approaches to categorize entities in the BB-cat and BB-cat+ner subtasks. The TagIt system performed entity categorization jointly with entity detection using dictionaries and rules. The BOUN team combined approximate string matching (edit distance) with an Information Retrieval based bag-of-word approach (cosine similarity of word vectors weighted with the tf-idf). This approach was the most successful in the BB-cat.

Prediction of Events. All systems used machine-learning to predict *Lives_in* events. The most popular algorithms are SVM (VERSE, HK, UTS, LIMSI) and neural networks (TurkuNLP, WhuNlpRE, DUTIR). UMS combined predictions from a SVM and a neural network. Most systems rely on syntactic parsing to generate features (VERSE, TurkuNLP, UMS, HK, DUTIR, UTS). Other common features included part-of-speech tags, word embeddings (trained on large corpora, *e.g.*, large sets of PubMed abstracts), and entity recognition. Rankings do not show any correlation to the machine learning algorithm, for instance the top ranking is based on SVM and the second is based on neural networks. Therefore, no conclusion can be drawn on the most appropriate class of methods. The quality of the predictions seems to rely mainly on the feature design, *i.e.*, what types of feature were used by the systems. To this respect the two top ranking systems have syntactic parsing-based features. More specifically, they both generate features based on the dependency path between entities.

7 Conclusion

The interest for the Bacteria Biotope Task keeps growing with a total of 14 teams participating in this third edition, and showing very promising results. 11 teams participated in the event extraction task (BB-event), demonstrating the interest of the NLP community for this challenging subject. For this event detection task, the most commonly used methods were SVMs and neural networks, and they yielded higher performance than during the 2013 edition of the task. However, a detailed analysis of the results showed that inter-sentence events still remain a challenge and are ignored by most systems. The other BB tasks, *i.e.* entity detection and categorization and knowledge base extraction, attracted fewer participants in comparison to event extraction. Knowledge base population was the most challenging task, since it required a large range of skills.

To help participants, supporting resources were provided but they were not much used. A more thorough investigation is needed to better understand the needs of participants in terms of external resources. The introduction of the online evaluation service with detailed metrics appears to have facilitated the development cycle of predictive systems. This service will be maintained online allowing for future experiments and comparisons with BB'16 data.

	BB-cat				BB-cat+ner			
	Train	Dev	Test	**Total**	Train	Dev	Test	**Total**
Documents	71	36	54	***161***	71	36	54	***161***
Words	16,295	8,890	13,797	***38,982***	16,295	8,890	13,933	***39,118***
Bacteria	375	244	347	***966***	375	244	401	***1,020***
Habitat	747	454	720	***1,921***	747	454	621	***1,822***
Total entities	*1,122*	*698*	*1,067*	***2,887***	*1,122*	*698*	*1,022*	***2,842***
Distinct *Bacteria*	167	111	146	***364***	167	111	181	***393***
Distinct *Habitat*	476	267	478	***1,125***	476	267	416	***1,073***
Total distinct entities	*643*	*378*	*624*	***1,489***	*643*	*378*	*597*	***1,466***
Bacteria categories	376	245	347	***968***	376	245	401	***1,022***
Habitat categories	825	535	861	***2,221***	825	535	681	***2,041***
Total categories	*1,201*	*780*	*1,208*	***3,189***	*1,201*	*780*	*1,082*	***3,063***
Distinct *Bacteria* categories	85	70	80	***190***	85	70	87	***193***
Distinct *Habitat* categories	210	122	177	***329***	210	122	168	***341***
Total distinct categories	*295*	*192*	*257*	***519***	*295*	*192*	*255*	***534***

Table 3: Descriptive statistics of the corpus for BB-cat and BB-cat+ner

	BB-event				BB-event+ner			
	Train	Dev	Test	**Total**	Train	Dev	Test	**Total**
Documents	61	34	51	***146***	71	36	54	***161***
Words	13,850	8,491	13,039	***35,380***	16,295	8,890	13,933	***39,118***
Bacteria	358	238	336	***932***	375	244	401	***1,020***
Habitat	687	454	720	***1,861***	747	454	621	***1,822***
Geographical	35	38	37	***110***	36	38	27	***101***
Total entities	*1,080*	*730*	*1,093*	***2,903***	*1,158*	*736*	*1,049*	***2,943***
Lives_in events (*Habitat*)	294	186	312	***792***	294	186	288	***768***
Lives_in events (*Geog.*)	33	37	28	***98***	33	37	26	***96***
Intra-sentence events	240	165	248	***653***	240	165	231	***636***
Inter-sentence events	87	58	92	***237***	87	58	83	***228***
Total Lives_in events	*327*	*223*	*340*	***890***	*327*	*223*	*314*	***864***

Table 4: Descriptive statistics of the corpus for BB-event and BB-event+ner

	BB-kb				BB-kb+ner			
	Train	Dev	Test	**Total**	Train	Dev	Test	**Total**
Documents	61	34	50	***145***	71	36	54	***161***
Words	13,850	8,491	12,758	***35,099***	16,295	8,890	13,933	***39,118***
Bacteria	358	238	330	***926***	375	244	401	***1,020***
Habitat	687	454	720	***1,861***	747	454	621	***1,822***
Total entities	*1,045*	*692*	*1,050*	***2,787***	*1,122*	*698*	*1,022*	***2,842***
Bacteria categories	359	239	330	***928***	376	245	401	***1,022***
Habitat categories	765	535	861	***2,161***	825	535	681	***2,041***
Total categories	*1,124*	*774*	*1,191*	***3,089***	*1,201*	*780*	*1,082*	***3,063***
Distinct *Bacteria* categories	81	69	77	***183***	85	70	87	***193***
Distinct *Habitat* categories	197	122	177	***317***	210	122	168	***341***
Total distinct categories	*278*	*191*	*254*	***500***	*295*	*192*	*255*	***534***
Lives_in events	294	186	312	***792***	294	186	288	***768***
Distinct Lives_in events	*204*	*156*	*185*	***522***	*204*	*156*	*183*	***524***

Table 5: Descriptive statistics of the corpus for BB-kb and BB-kb+ner

Team	F1	Recall	Precision	F1 (*Habitat*)	F1 (*Geo.*)	F1 (Intra-sentence)
VERSE	**0.558**	0.615	0.510	**0.545**	0.714	0.634
TurkuNLP	0.521	0.448	**0.623**	0.499	**0.755**	0.620
LIMSI	0.485	**0.646**	0.388	0.482	0.525	**0.636**
HK	0.474	0.392	0.599	0.452	0.708	0.567
WhuNlpRE	0.471	0.407	0.559	0.471	0.465	0.561
UMS	0.463	0.399	0.551	0.439	0.704	0.550
DUTIR	0.456	0.382	0.566	0.451	0.512	0.544
WXU	0.455	0.383	0.560	0.445	0.578	0.540
-	-	-	-	-	-	-
UTS	0.451	0.382	0.551	0.425	0.704	0.537
-	-	-	-	-	-	-

Table 8: Team results for the BB-event task

References

Mouhamadou Ba and Robert Bossy. 2016. Interoperability of corpus processing work-flow engines: the case of alvisnlp/ml in openminted. In Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, and Stelios Piperidis, editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 15–18, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

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. *BMC bioinformatics*, 13(11):1.

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, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC bioinformatics*, 16(10):1.

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.

Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessières, Loic Lepiniec, and Claire Nédellec. 2016. Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task workshop*, Berlin, Germany, August. Association for Computational Linguistics.

Scott Federhen. 2012. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

C Fellbaum. 1998. Wordnet: An on-line lexical database and some of its applications.

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.

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.

Wiktoria Golik, Pierre Warnier, and Claire Nédellec. 2011. Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 37–39.

Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. 2011. The brenda tissue ontology (bto): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39(suppl 1):D507–D513.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Natalia Ivanova, Susannah G Tringe, Konstantinos Liolios, Wen-Tso Liu, Norman Morrison, Philip Hugenholtz, and Nikos C Kyrpides. 2010. A call for standardized classification of metagenome projects. *Environmental microbiology*, 12(7):1803–1805.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Frédéric Papazian, Robert Bossy, and Claire Nédellec. 2012. Alvisae: a collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 149–152. Association for Computational Linguistics.

Aidan C Parte. 2013. Lpsnlist of prokaryotic names with standing in nomenclature. *Nucleic acids research*, page gkt1111.

Miguel Pignatelli, Andrés Moya, and Javier Tamames. 2009. Envdb, a database for describing the environmental distribution of prokaryotic taxa. *Environmental Microbiology Reports*, 1(3):191–197.

Zorana Ratkovic, Wiktoria Golik, and Pierre Warnier. 2012. Event extraction of bacteria biotopes: a knowledge-intensive nlp-based approach. *BMC bioinformatics*, 13(11):1.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. 2007. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281.