BioNLP shared Task 2011 - Bacteria Biotope

Robert Bossy[1], **Julien Jourde**[1], **Philippe Bessières**[1], **Maarten van de Guchte**[2], **Claire Nédellec**[1]

[1]MIG UR1077 [2]Micalis UMR 1319
INRA, Domaine de Vilvert
78352 Jouy-en-Josas, France
forename.name@jouy.inra.fr

Abstract

This paper presents the Bacteria Biotope task as part of the BioNLP Shared Tasks 2011. The Bacteria Biotope task aims at extracting the location of bacteria from scientific Web pages. Bacteria location is a crucial knowledge in biology for phenotype studies. The paper details the corpus specification, the evaluation metrics, summarizes and discusses the participant results.

1 Introduction

The Bacteria Biotope (BB) task is one of the five main tasks of the BioNLP Shared Tasks 2011. The BB task consists of extracting bacteria location events from Web pages, in other words, citations of places where a given species lives. Bacteria locations range from plant or animal hosts for pathogenic or symbiotic bacteria, to natural environments like soil or water. Challenges for Information Extraction (IE) of relations in Biology are mostly devoted to the identification of bio-molecular events in scientific papers where the events are described by relations between named entities, *e.g.* genic interactions (Nédellec, 2005), protein-protein interactions (Pyysalo et al., 2008), and more complex molecular events (Kim et al., 2011). However, this far from reflects the diversity of the potential applications of text mining to biology. The objective of previous challenges has mostly been focused on modeling biological functions and processes using the information on elementary molecular events extracted from text. The BB task is the first step towards linking information on bacteria at the molecular level to ecological information. The information on bacterial habitats and properties of these habitats is very abundant in literature, in particular in Systematics literature (*e.g. International Journal of Systematic and Evolutionary Microbiology*), however it is rarely available in a structured way (Hirschman et al., 2008; Tamames and de Lorenzo, 2009). The NCBI GenBank nucleotide *isolation source* field (GenBank) and the JGI Genome OnLine Database (GOLD) *isolation site* field are incomplete with respect to the microbial diversity and are expressed in natural language. The two critical missing steps in terms of biotope knowledge modeling are (1) the automatic population of databases with organism/location pairs that are extracted from text, and (2) the normalization of the habitat name with respect to biotope ontologies. The BB task mainly aims at solving the first information extraction issue. The second classification issue is handled through the categorization of locations into eight types.

2 Context

According to NCBI statistics there are nearly 900 bacteria with complete genomes, which account for more than 87% of total complete genomes. Consequently, molecular studies in bacteriology are shifting from species-centered to full diversity investigation. The current trend in high-throughput experiments targets diversity related fields, typically phylogeny or ecology. In this context, adaptation properties, biotopes and biotope properties become critical information. Illustrative questions are:

Proceedings of BioNLP Shared Task 2011 Workshop, pages 56–64,
Portland, Oregon, USA, 24 June, 2011.

- Is there a phylogenetic correlation between species that share the same biotope?
- What are common metabolic pathways of species that live in given conditions, especially species that survive in extreme conditions?
- What are the molecular signaling patterns in host relationships or population relationships (*e.g.* in biofilms)?

Recent metagenomic experiments produce molecular data associated with a habitat rather than a single species. This raises new challenges in computational biology and data integration, such as identifying known and new species that belong to a metagenome.

Not only will these studies require comprehensive databases that associate bacterial species to their habitat, but they also require a formal description of habitats for property inference. The bacteria biotope description is potentially very rich since any physical object, from a cell to a continent, can be a bacterial habitat. However these relations are much simpler to model than with general formal spatial ontologies. A given place is a bacterial habitat if the bacteria and the habitat are physically in contact, while the relative position of the bacteria and its dissemination are not part of the BB task model.

The BB Task requires the locations to be assigned different types (*e.g.* soil, water). We view location typing as a preliminary step of more fine-grained modeling in location ontologies. Some classifications for bacteria biotopes have been proposed by some groups (Floyd et al., 2005; Hirschman et al., 2008; Field et al., 2008; Pignatelli et al., 2009). The Environment Ontology project (EnvO) is developing an ambitious detailed environment ontology for supporting standard manual annotation of environments of all types of organisms and biological samples (Field et al., 2008). In a similar way, the GOLD group at JGI defined a standard classification for bacteria population metagenome projects. Developing methods for the association of such biotope classes to organisms remains an open question. EnvDB (Pignatelli et al., 2009) is an attempt to inventory isolation sources of bacteria as recorded in GenBank and to map them to a three level hierarchy of 71 biotope classes. The assignment of bacterial samples in one of the EnvDB classes is supported by a text-mining tool based on a Naïve Bayes (NB) classifier applied to a bag of words representing the associated reference title and abstract. Unfortunately, the low number of paper references associated with the isolation source field (46 %) limits the scope of the method.

The BB task has a similar goal, but directly applies to natural language texts thus avoiding the issue of database incompleteness. As opposed to database-based approaches, biotope information density is higher but the task has to include bacteria and location identification, as well as information extraction to relate them.

The eight types of locations in the BB task capture high-level information for further ontology mappings. The location types are *Host*, *HostPart*, *Geographical* and *Environmental*. *Environmental* is broadly defined to qualify locations that are not associated to hosts, in a similar way to what was described by Floyd et al. (Floyd et al., 2005). In addition, the BB task types exclude artificially constructed biotopes (*e.g.* bacteria growing in labs on a specific medium) and laboratory mutant bacteria. The *Environmental* class is divided into *Food*, *Medical*, *Soil* and *Water*. Locations that are none of these subtypes are classified as *Environmental*.

The exact geographical location (*e.g.* latitude and longitude coordinates) has less importance here than in eukaryote ecology because most of the biotope properties vary along distances smaller than the precision of the current positioning technologies. Geographical names are only useful in bacteria biotope studies when the physico-chemical properties of the location can be inferred. For the sake of simplicity, the locations of bacteria host (*e.g.* the stall of the infected cow) are not taken into account despite their richness (Floyd et al., 2005).

The important information conveyed by the locations, especially of *Environment* type, is the function of the bacterium in its ecosystem rather than the substance of the habitat. Indeed the final goal is to extract habitat properties and bacteria phenotypes. Beyond the identification of locations, their properties (*e.g.* temperature, pH, salinity, oxygen) are of high interest for phenotypes (*e.g.* thermophily, acidophily, halophily) and trophism studies. This information is difficult to extract, and is often incomplete or even not available in papers (Tamames and de Lorenzo., 2009). Hopefully, some properties can be automatically retrieved

with the help of specialized databases, which give the physico-chemical properties of locations, such as hosts (plant, animal, human organs), soils (see WebSoilSurvey, Corine Land Cover), water, or chemical pollutants.

From a linguistic point of view, the BB task differs from other IE molecular biology tasks while it raises some issues common to biomedicine and more general IE tasks. The documents are scientific Web pages intended for non-experts such as encyclopedia notices. The information is dense compared to scientific papers. Documents are structured as encyclopedia pages, with the main focus on a single species or a few species of the same genus or family. The frequency of anaphora and coreferences is unusually high. The location entities are denoted by complex expressions with semantic boundaries instead of rigid designators.

3 Task description

The goal of the BB task is illustrated in Figure 1.

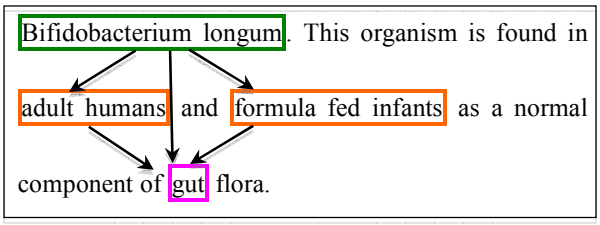

Figure 1. Example of information to be extracted in the BB Task.

The entities to be extracted are of two main types: bacteria and locations. They are text-bound and their position has to be predicted. Relations are of type *Localization* between bacteria and locations, and *PartOf* between hosts and host parts. In the example in Figure 1, *Bifidobacterium longum* is a bacterium. *adult humans* and *formula fed infants* denote host locations for the bacteria. *gut* is also a bacteria location, part of the two hosts and thus of type host part.

Coreference relations between entities denoting the same information represent valid alternatives for the relation arguments. For example, the three taxon names in Figure 2 are equivalent.

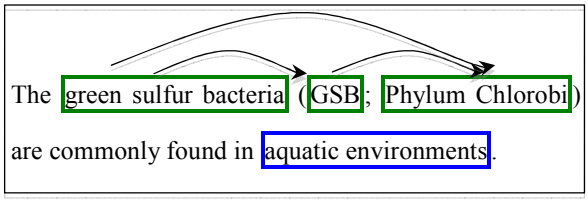

Figure 2. Coreference example.

The coreference relation between pairs of entities is binary, symmetric and transitive. Coreference sets are equivalence sets defined as the transitive closure of the binary coreference relation. Their annotation is provided in the training and development sets, but it does *not* have to be predicted in the test set.

4 Corpus description

The corpus sources are the following bacteria sequencing project Web pages:

- Genome Projects referenced at NCBI;
- Microbial Genomics Program at JGI;
- Bacteria Genomes at EBI;
- Microorganisms sequenced at Genoscope;
- Encyclopedia pages from MicrobeWiki.

The documents are publicly available and quite easy to understand by non-experts compared to scientific papers on similar topics. From the 2,086 downloaded documents, 105 were randomly selected for the BB task. A quarter of the corpus was retained for test evaluation. The rest was split into train and development sets. Table 1 gives the distribution of the entities and relations per corpus. The distribution of the five document sources in the test corpus reflects the distribution of the training set and no other criteria. Food is therefore underrepresented.

	Training+Dev	**Test**
Document	78 (65 + 13)	27 (26 %)
Bacteria	538	121 (18 %)
Environment	62	16 (21 %)
Host	486	101 (17 %)
HostPart	217	84 (28 %)
Geographical	111	25 (18 %)
Water	70	21 (23 %)
Food	46	0 (0 %)
Medical	24	2 (8 %)
Soil	26	20 (43 %)
Coreference	484	100 (17 %)
Total entities	1,580	390

	Training+Dev	Test
Localization	998	250 (20 %)
Part of Host	204	78 (28 %)
Total relations	1,202	328

Table 1. Corpus Figures.

5 Annotation methodology

HTML tags and irrelevant metadata were stripped from the corpus. The Alvis pipeline (Nédellec et al., 2009) pre-annotated the species names that are potential bacteria and host names. A team of 7 scientists manually annotated the entities, coreferences and relations using the Cadixe XML editor (Cadixe). Each document was processed by two independent annotators in a double-blind manner. Conflicts were automatically detected, resolved by annotator negotiation and irrelevant documents (*e.g.* without bacterial location) were removed. The remaining inconsistencies among documents were resolved by the two annotators assisted by a third person acting as an arbitrator.

The annotator group designed the detailed annotation guidelines in two phases. First, they annotated a set of 10 documents, discussed the options and wrote detailed guidelines with representative and illustrative examples. During the annotation of the rest of the documents, new cases were discussed by email and the guidelines amended accordingly.

Location types. The main issues under debate were the definition of location types, boundaries of annotations and coreferences. Additional annotation specifications concerned the exclusion of overly general locations (*e.g. environment*, *zone*), artificially constructed biotopes and indirect effects of bacteria on distant places. For instance, a disease symptom occurring in a given host part does not imply the presence of the bacteria in this place, whereas infection does. Boundaries of types were also an important point of discussion since the definite formalization of habitat categories was at stake. For instance we decided to exclude land environment citations (*fields*, *deserts*, *savannah*, *etc.*) from the type *Soil*, and thus enforced a strict definition of soil bacteria. The most controversial type was host parts. We decided to include fluids, secretions and excretions (which are not strictly organs). Therefore, the host parts category required specifications to determine at which point of dissociation from the original host is a habitat not a host part anymore (e.g. *mother's milk* vs. *industrial milk*, *rhizosphere* as host part instead of soil).

Boundaries. The bacteria name boundaries do not include any external modifiers (*e.g. two A. baumannii strains*). Irrelevant modifiers of locations are considered outside the annotation boundaries (*e.g. responsible for a hospital epidemic*). All annotations are contiguous and span on a single fragment in the same way as the other BioNLP Shared Tasks. This constraint led us to consider cases where several annotations occur side by side. The preferred approach was to have one distinct annotation for each different location (*e.g. contact with infected animal products or through the air*). In the case of head or modifier factorization, the annotation depends on the information conveyed by the factorized part. If the head is not relevant to determine the location type, then each term is annotated separately (*e.g. tropical and temperate zones*). Conversely, if the head is the most informative with regards to the location type, a single annotation spans the whole fragment (*fresh and salt water*).

Coreferences. Two expressions are considered as coreferential and thus valid solution alternatives, if they convey the same information. For instance, complete taxon names and non-ambiguous abbreviations are valid alternatives (*e.g. Borrelia garinii vs. B. garinii*), while ambiguous anaphora ellipses are not (*e.g.* as in "*[..] infected with Borrelia duttonii. Borrelia then multiplies [..]*"). The ellipsis of the omitted specific name (*dutotonii*) leaves the ambiguous generic name (*Borrelia*).

The full guidelines document is available for download on the BioNLP Shared Task Bacteria Biotope page[1].

6 Evaluation procedure

6.1 Campaign organization

The training and development corpora with the reference annotations were made available to the participants by December 1st 2010 on the BioNLP Shared Tasks pages together with the evaluation software. The test corpus, which does not contain

[1] https://sites.google.com/site/bionlpst/home/bacteria-biotopes/BioNLP-ST_2011_Bacteria_Biotopes_Guidelines.pdf

any annotation, was made available by March, 1st 2011. The participants sent the predicted annotations to the BioNLP Shared Task organizers by March 10th. Each participant submitted a single final prediction set. The detailed evaluation results were computed, provided to the participants and published on the BioNLP website by March, 11th.

6.2 Evaluation metrics

The evaluation metrics are based on precision, recall and the F-measure. In the following section, the *PartOf* and *Localization* relations will both be referred to as events. The metrics measure the accuracy of the participant prediction of events with respect to the reference annotation of the test corpus. Predicted entities that are not event arguments are ignored and they do not penalize the score. Each event E_r in the reference set is matched to the predicted event E_p that maximizes the event similarity function *S*. The recall is the sum of the *S* results divided by the number of events in the reference set. Each event E_p in the predicted set is matched to the reference event E_r that maximizes *S*. The precision is the sum of the *S* results divided by the number of events in the predicted set. Participants were ranked by the F-score defined as the harmonic mean between precision and recall.

E_{ab}, the event similarity between a reference *Localization* event *a* and a predicted *Localization* event *b*, is defined as:

$$E_{ab} = B_{ab} \,.\, T_{ab} \,.\, J_{ab}$$

- B_{ab} is the bacteria boundary component defined as: if the *Bacterium* arguments of both the predicted and reference events have exactly the same boundaries, then B_{ab} = 1, otherwise B_{ab} = 0. Bacteria name boundary matching is strict since boundary mistakes usually yield a different taxon.
- T_{ab} is the location type prediction component defined as: if the *Location* arguments of both the predicted and reference events are of the same type, then T_{ab} = 1, otherwise T_{ab} = 0.5. Thus type errors divide the score by two.
- J_{ab} is the location boundary component defined as: if the *Location* arguments of the predicted and reference events overlap, then

$$J_{ab} = \frac{LEN_a + LEN_b}{OV_{ab}} - 1$$

where LEN_a and LEN_b are the length of the *Localization* arguments of predicted and reference events, and OV_{ab} is the length of the overlapping segment between the *Localization* arguments of the predicted and reference events. If the arguments do not overlap, then J_{ab} is 0. This formula is a Jaccard index applied to overlapping segments. Location boundary matching is relaxed, though the Jaccard index rewards predictions that approach the reference.

For *PartOf* events between *Host*s and *HostPart*s, the matching score P_{ab} is defined as: if the *Host* arguments of the reference and predicted events overlap and the *Part* arguments of the reference and predicted events overlap, then P_{ab} = 1, otherwise P_{ab} = 0. Boundary matching of *PartOf* arguments is relaxed, since boundary mistakes are already penalized in E_{ab}.

Arguments belonging to the same coreference set are strictly equivalent. In other words, the argument in the predicted event is correct if it is equal to the reference entity or to any item in the reference entity coreference set.

7 Results

7.1 Participating systems

Three teams submitted predictions to the BB task. The first team is from the University of Turku (UTurku); their system is generic and produced predictions for every BioNLP Shared Task. This system uses ML intensely, especially SVMs, for entity recognition, entity typing and event extraction. UTurku adapted their system for the BB task by using specific NER patterns and external resources (Björne and Salakoski, 2011).

The second team is from the Japan Advanced Institute of Science and Technology (JAIST); their system was specifically designed for this task. They used CRF for entity recognition and typing, and classifiers for coreference resolution and event extraction (Nguyen and Tsuruoka, 2011).

The third team is from Bibliome INRA; their system was specifically designed for this task (Ratkovik et al., 2011). This team has the same affiliation as the BB Task authors, however great care was taken to prevent communication on the subject between task participants and the test set annotators.

The results of the three submissions according to the official metrics are shown in Table 2. The scores are micro-averaged: *Localization* and *PartOf* relations have the same weight. Given the novelty and the complexity of the task, these first results are quite encouraging. Almost half of the relations are correctly predicted. The Bibliome team achieved the highest F-measure with a balanced recall and precision (45%).

	Recall	**Precision**	**F-score**
Bibliome	**45**	45	**45**
JAIST	27	42	33
UTurku	17	**52**	26

Table 2. Bacteria Biotope Task results.

7.2 Systems description and result analysis

All three systems perform the same distinct sub-tasks: bacteria name detection, detection and typing of locations, coreference resolution and event extraction. The following description of the approaches used by the three systems in each subtask will be supported by intermediate results.

Bacteria name detection. Interestingly the three participants used three different resources for the detection of bacteria names: the List of Prokaryotic Names with Standing in Nomenclature (LPNSN) by UTurku, names in the genomic BLAST page of NCBI by JAIST and the NCBI Taxonomy by Bibliome.

Bibliome	84
JAIST	55
UTurku	16

Table 3. Bacteria entity recall.

Table 3 shows a disparity in the bacteria entity recall of participants. The merits of each resource cannot be deduced directly from these figures since they have been exploited in different manners. UTurku and JAIST systems injected the resource as features in a ML algorithm, whereas Bibliome directly projected the resource on the corpus with additional rule-based abbreviation detection.

However there is some evidence that the resources have a major impact on the result. According to Sneath and Brenner (1992) LPNSN is necessarily incomplete. NCBI BLAST only contains names of species for which a complete genome has been published. The NCBI Taxonomy used by INRA only contains names of taxa for which some sequence was published. It appears that all the lists are incomplete. However, the bacteria referenced by the sequencing projects, which are mentioned in the corpus should all be recorded by the NCBI Taxonomy.

Location detection and typing. As stated before, locations are not necessarily denoted by rigid designators. This was an interesting challenge that called for the use of external resources and linguistic analysis with a broad scope.

UTurku and JAIST both used WordNet, a sensible choice since it encompasses a wide vocabulary and is also structured with synsets and hyperonymy relations. The WordNet entries were injected as features in the participant ML-based entity recognition and typing subsystems.

It is worth noting that JAIST also used word clustering based on MEMM for entity detection. This method has things in common with distributional semantics. JAIST experiments demonstrated a slight improvement using word clustering, but further exploration of this idea may prove to be valuable.

Alternatively, the Bibliome system extracted terms from the corpus using linguistic criteria classified them as locations and predicted their type, by comparing them to classes in a habitat-specific ontology. This prediction uses both linguistic analysis of terms and the hierarchical structure of the ontology. Bibliome also used additional resources for specific types: the NCBI Taxonomy for type *Host* and Agrovoc countries for type *Geographical*.

	Bibliome	**JAIST**	**UTurku**
Host	**82**	49	28
Host part	**72**	36	28
Geo.	29	**60**	53
Environment	**53**	10	11
Water	**83**	32	2
Soil	**86**	37	34

Table 4. Location entity recall by type. The number of entities of type *Food* and *Medical* in the test set is too low to be significant. The scores are computed using T_{ab} and J_{ab}.

The location entity recall in Table 4 shows that Bibliome consistently outperformed the other groups for all types except for *Geographical*. This demonstrates the strength of exploiting a resource with strong semantics (ontology vs. lexicon) and with mixed semantic and linguistic rules.

In order to evaluate the impact of *Location* entity boundaries and types, we computed the final score by relaxing T_{ab} and J_{ab} measures. We re-defined T_{ab} as always equal to 1, in other words the type of the localization was not evaluated. We also re-defined J_{ab} as: if the *Location* arguments overlap, then J_{ab} = 1, otherwise J_{ab} = 0. This means that boundaries were relaxed. The relaxed scores are shown in Table 5. While the difference is not significant for JAIST and UTurku, the Bibliome results exhibit a 9 point increase. This demonstrates that the Bibliome system is efficient at predicting which entities are locations, while the other participants predict more accurately the boundaries and types.

	Recall	Prec.	F-score	Diff.
Bibliome	54	54	54	**+9**
JAIST	29	45	35	+2
UTurku	19	56	28	+2

Table 5. Participants score using relaxed location boundaries and types.

Coreference resolution. The corpus exhibits an unusual number of anaphora, especially bacteria coreferences since a single bacterium species is generally the central topic of a document. The Bibliome submission is the only one that performed bacteria coreference resolution. Their system is rule-based and dealt with referential "it", bi-antecedent anaphora and more importantly sortal anaphora. The JAIST system has a bacteria coreference module based on ML. However the submission was done without coreference resolution since their experiments did not show any performance improvement.

Event extraction. Both UTurku and JAIST approached the event extraction as a classification task using ML (SVM). Bibliome exploited the co-occurrence of arguments and the presence of trigger words from a predefined list. Both UTurku and Bibliome generate events in the scope of a sentence, whereas JAIST generates events in the scope of a paragraph.

As shown in Table 6, UTurku achieved the best score for *PartOf* events. For all participants, the prediction is often correct (between 60 and 80%) while the recall is rather low (20 to 32%).

		Recall	Precis.	F-score
	Host	**61**	48	**53**
	Host part	**53**	42	**47**
	Geo.	13	38	19
B.	Env.	**29**	24	**26**
	Water	**60**	**55**	**57**
	Soil	**69**	**59**	**63**
	Part-of	23	79	36
	Host	30	43	36
	Host part	18	**68**	28
	Geo.	**52**	35	**42**
J.	Env.	5	0	0
	Water	19	27	23
	Soil	21	42	28
	Part-of	31	61	41
	Host	15	**51**	23
	Host part	9	40	15
	Geo.	32	**40**	36
U.	Env.	6	50	11
	Water	1	7	2
	Soil	12	21	15
	Part-of	**32**	**83**	**46**

Table 6. Event extraction results per type.

Conversely, the score of the *Localization* relation by UTurku has been penalized by its low recognition of bacteria names (16%). This strongly affects the score of *Localizations* since the bacterium is the only expected agent argument. The good results of Bibliome are partly explained by its high bacteria name recall of 84%.

The lack of coreference resolution might penalize the event extraction recall. To test this hypothesis, we computed the recall by taking only into account events where both arguments occur in the same sentence. The goal of this selection is to remove most events denoted through a coreference. The recall difference was not significant for Bibliome and JAIST, however UTurku recall raised by 12 points (29%). That experiment confirms that UTurku low recall is explained by coreferences

rather than the quality of event extraction. The paragraph scope chosen by JAIST probably compensates the lack of coreference resolution.

As opposed to Bibliome, the precision of the *Localization* relation prediction by JAIST and UTurku, is high compared to the recall, with a noticeable exception of geographical locations. The difference between participants seems to be caused by the geographical entity recognition step more than the relation itself. This is shown by the difference between the entity and the event recall (Table 4 and 6 respectively).. The worst predicted type is *Environment,* which includes diverse locations, such as agricultural, natural and industrial sites and residues. This reveals significant room for improvement for *Water*, *Soil* and *Environment* entity recognition.

8 Discussion

The participant papers describe complementary methods for tackling BB Task's new goals. The novelty of the task prevents participants from deeply investing in all of the issues together. Depending on the participants, the effort was focused on different issues with various approaches: entity recognition and anaphora resolution based on extensive use of background knowledge, and relation prediction based on linguistic analysis of syntactic dependencies. Moreover, these different approaches revealed to be complementary with distinct strengths and limitations. In the future, one may expect that the integration of these promising approaches will improve the current score.

The corpus of BioNLP BB Task 2011 consists of a set of Web pages that were selected for their readability. However, some corpus traits make the IE task more difficult compared to scientific papers. For example, the relaxed style of some pages tolerates some typographic errors (*e.g. morrow* instead of *marrow*) and ambiguous anaphora. The genome sequencing project documents aim at justifying the sequencing of bacteria. This results in abundant descriptions of potential uses and locations that should not be predicted as actual locations. Their correct prediction requires complex analysis of modalities (possibility, probability, negation). Some pages describe the action of hosted bacteria at the molecular level, such as cellular infection. Terms related to the cell are ambiguous locations because they may refer to either bacteria or host cells.

Scientific papers form a much richer source of bacterial location information that is exempt from such flaws. However, as opposed to Web pages, most of them are not publicly available and they are in PDF format.

The typology of locations was designed according to the BB Task corpus with a strong bias towards natural environments since bioremediation and plant growth factor are important motivations for bacteria sequencing. It could be necessary to revise it according to a broader view of bacterial studies where pathogenicity and more generally human and animal health are central issues.

9 Conclusion

The Bacteria Biotope Task corpus and objectives differ from molecular biology text-mining of scientific papers. The annotation strategy and the analysis of the participant results contributed to the construction of a preliminary review of the nature and the richness of its linguistic specificities. The participant results are encouraging for the future of the Bacteria Biotope issue. The degree of sophistication of participating systems shows that the community has technologies, which are mature enough to address this crucial biology question. However, the results leave a large room for improvement.

The Bacteria Biotope Task was an opportunity to extend molecular biology text-mining goals towards the support of bacteria biodiversity studies such as metagenomics, ecology and phylogeny. The prediction of bacterial location information is the very first step in this direction. The abundance of scientific papers dealing with this issue and describing location properties form a potentially rich source for further extensions.

Acknowledgments

The authors thank Valentin Loux for his valuable contribution to the definition of the Bacteria Biotope task. This work was partially supported by the French Quaero project.

References

Jari Björne and Taio Salakoski. 2011. Generalizing Biomedical Event Extraction. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

Cadixe. http://caderige.imag.fr/Articles/CADIXE-XML-Annotation.pdf

Corine Land Cover. http://www.eea.europa.eu/themes/landuse/interactive/clc-download

EnvDB database. http://metagenomics.uv.es/envDB/

EnvO Project. http://gensc.org/gc_wiki/index.php/EnvO_Project

Dawn Field [et al]. 2008. Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nature Biotechnology*. 26: 541-547.

Melissa M. 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):2813-23.

GenBank. http://www.ncbi.nlm.nih.gov/

GOLD. http://www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi

Lynette Hirschman, Cheryl Clark, K. Bretonnel Cohen, Scott Mardis, Joanne Luciano, Renzo Kottmann, James Cole, Victor Markowitz, Nikos Kyrpides, Norman Morrison, Lynn M. Schriml, Dawn Field. 2008. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *Omics*. 12(2):129-136.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, Jun'ichi Tsujii. 2010. Extracting bio-molecular events from literature - the BioNLP'09 shared task. *Special issue of the International Journal of Computational Intelligence*.

MicrobeWiki. http://microbewiki.kenyon.edu/index.php/MicrobeWiki

Microbial Genomics Program at JGI. http://genome.jgi-psf.org/programs/bacteria-archaea/index.jsf

Microorganisms sequenced at Genoscope. http://www.genoscope.cns.fr/spip/Microorganisms-sequenced-at.html

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge" in *Proceedings of the Learning Language in Logic (LLL05)* workshop joint to *ICML'05*. Cussens J. and Nédellec C. (eds). Bonn.

Claire Nédellec, Adeline Nazarenko, Robert Bossy. 2008..Information Extraction. *Ontology Handbook*. S. Staab, R. Studer (eds.), Springer Verlag, 2008.

Nhung T. H. Nguyen and Yoshimasa Tsuruoka. 2011. Extracting Bacteria Biotopes with Semi-supervised Named Entity Recognition and Coreference Resolution. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

Miguel Pignatelli, Andrés Moya, Javier Tamames. (2009). EnvDB, a database for describing the environmental distribution of prokaryotic taxa. *Environmental Microbiology Reports*. 1:198-207.

Prokaryote Genome Projects at NCBI. http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*. vol 9. Suppl 3. S6.

Zorana Ratkovic, Wiktoria Golik, Pierre Warnier, Philippe Veber, Claire Nédellec. 2011. BioNLP 2011 Task Bacteria Biotope – The Alvis System. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

Peter H. A. Sneath and Don J. Brenner. 1992. "Official" Nomenclature Lists. *American Society for Microbioloy News.* 58, 175.

Javier Tamames and Victor de Lorenzo. 2010. EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*. 11:294.

Web Soil Survey. http://websoilsurvey.nrcs.usda.gov/