Bacteria Biotope (BB) task at BioNLP Shared Task 2013
Task Proposal

Robert Bossy, Claire Ndellec, Julien Jourde

October 19, 2012

Useful links

- BioNLP 2013 Shared Task
- Bacteria Biotopes at BioNLP 2013 Shared Task
- Specification of file formats
- MBTO Ontology

Task Summary

Task name Bacteria Biotope (BB)

Goal in IE

1. Promote information extraction and entity categorization on microorganisms ecology.
2. Assess the performance of automatic categorization systems in this subject.

Motivation in biology There is surprisingly no comprehensive database of natural environment location of bacteria, although this is a critical information for studying the interaction mechanisms of the bacteria with its environment at a molecular level. As in 2011, the extraction of habitat mentions and their attachment to bacteria would fill this need. Moreover, once the habitats are identified, they must be normalized to be compared. Ontologies of habitats such as EnvO and MBTO are available. There is a nice opportunity to develop and evaluate methods for normalizing the habitat descriptions with these ontologies. This task is more generally related to semantic annotation of entities with ontologies.

Description of the entities and events Bacteria taxon names are annotated as text-bound entities. The definition of the *Bacteria* type may be at any taxonomic level from phylum (*Eubacteria*) to strain. Habitats have also been annotated and typed with one or several concepts in the MBTO-Habitat ontology. This ontology contains more than 1500 concepts describing different aspects of bacteria habitats (hosts, natural habitats like soil and water, artificial habitats like processed food or man-made instruments, etc).

Localization relation relate a bacterium to the place where it lives. This type of relation has two mandatory arguments: the first is of type *Bacterium* and the second is of type *Habitat*.

Concrete examples in BioNLP format Available in attached data.

Evaluation and criteria We propose three sub-tasks:

1. Habitat categorization: participant systems are evaluated for their capacity to detect bacteria and habitat entities and categorize them with MBTO-Habitat concepts.
2. Relation extraction: participant systems are evaluated for their capacity to extract localization relations between bacteria and habitats (entities are given).
3. Full relation extraction: participant systems are evaluated for their capacity to detect bacteria and habitats, then to extract localization relations between them.

Task corpus The BB 2011 data set re-annotated following improved guidelines, complemented with the annotation of habitats with ontology concepts, with all habitat annotations (for the event prediction) and with new documents for a better distribution of categories.

Existing corpus The BB 2011 data set: a set of encyclopedia-like entries that give general information about bacteria species in common language. These documents were taken from relevant public web sites.

Feasibility of task

y recognition and categorization) Experiments with this corpus were carried out for a low number of categories (9). One of the goals of this task is to assess the feasibility of categorization with a larger number of categories.

sub-task 2 (relation extraction) A baseline method based on arguments co-occurrence and trigger words was conducted on this corpus yielding .56 F-measure.

cognition and relation extraction) This is the same task definition as BioNLP Shared Task 2011 Bacteria Biotope task; participants had promising results but there was a lot of room for improvement. No participating system had used state of the art techniques for every aspect of the task (entity recognition, coreference resolution, relation extraction).

Annotation process The annotation is a revision of the BioNLP-ST 2011 BB annotation, by 9 annotators in double-blind.

1 Introduction

This document describes the BioNLP Shared Task 2013 Bacteria Biotope task (BB). The BB task consists in:

1. Entity recognition of bacteria taxa and bacteria habitats.
2. Bacteria habitat categorization through the MBTO-Habitat ontology.
3. Extraction of localization relations between bacteria and habitats.

The goal of this task is to:

1. Promote Information Extraction on the subject of microorganisms ecology.
2. Assess the performance of automatic categorization systems.
3. Assess the performance of relation extraction on this subject by different methods.

The knowledge tackled by this task is the habitats where bacteria live, and the environment properties of bacteria. This information is a particularly interesting in the fields of food processing and safety, health sciences and waste processing. There are also fundamental research that requires this knowledge like metagenomics or phylogeography/phyloecology. There is currently no database that supply the habitats of bacteria in a comprehensive way. Moreover the efforts for normalizing the habitats are just beginning. The diversity of habitats is such that several ongoing projects aim at building habitat ontologies (EnvO, MBTO).

In this context, we anticipate the need of IE systems able to:

1. detect mentions habitats;
2. categorize them with concepts of large ontologies;
3. extract localization relations between bacteria and habitats.

2 Representation and Task Setting

The BB corpus annotation follows the BioNLP-ST 2013 frame of representation with a few modifications in order to cope with:

- Discontinuous entities, and
- Categorization with concepts from an ontology.

The ontology of habitats is provided to participants in OBO format. It contains approximately 1500 concepts organized in a hierarchy of *is-a* relations. Each concept has one or several parents (the root has no parent), as well as several synonyms.

We propose three sub-tasks:

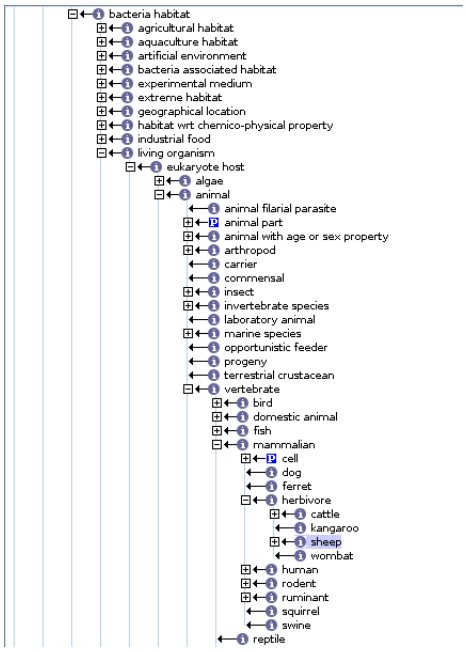

Figure 1: Excerpt of the MBTO ontology showing the "*living organism*" subtree.

2.1 Sub-task 1: entity detection and categorization

Train input	**Test**
document contents	input
habitat ontology	input
entities	output
entity category	output

Evaluation The evaluation will focus on the accuracy of predicted categorization compared to gold category(ies). We will design a concept distance measure in order to sanction over-generalization or overspecialization with a fair penalty. Note that if an entity has several categories, then it is a conjunction: all categories must be predicted.

Boundary accuracy will be factored in the evaluation since the inclusion or exclusion of modifiers can change the meaning and the categorization of phrases.

We anticipate the need for adjudication of the test set.

2.2 Sub-task 2: localization relation extraction

Train input	**Test**
document contents	input
entities	input
coreferences	evaluation
relations	output

Evaluation Recall/Precision of predicted relations against gold relations. Gold coreferences will be used to factor equivalent relations.

2.3 Sub-task 3: relation extraction without gold entities

Train input	Test
document contents	input
entities	**output**
coreferences	evaluation
relations	output

Evaluation Recall/Precision of predicted relations against gold relations. The evaluation of predicted entities boundaries against gold entities should be relaxed.

3 Document selection

The corpus is an extension of the BioNLP-ST 2011 BB corpus. It consists in a collection of web pages of the textbook genre from the following sources:

- Genome Project overview pages from NCBI
- MicrobeWiki articles
- Microbial Genomics Program pages from JGI
- 2Can bacteria pages at EBI
- Genome overview pages at Genoscope

Each document is centered around a species, a genre or a family of bacteria; it contains general information about their classification, ecology and interest in human activities.

2040 documents were extracted and 85 were randomly selected for the BioNLP-ST 2011. The test set of BioNLP-ST 2011 BB was not fully representative of the train set, additional documents will be selected in order to overcome this shortcoming.

The three sub-tasks will share the train set. Sub-tasks 1 and 3 will share the test set.

4 Entity and relation Annotation

4.1 Annotation Overview

Bacteria entities are annotated as contiguous spans of text that contains a full unambiguous prokaryote taxon name, the type label is *Bacteria*. Habitat entities are annotated as spans of text that contains a complete mention of a potential habitat for bacteria, the type label is *Habitat*. Additionally geographical and administrative entities are annotated with the label *Geographical*.

Bacteria and *Geographical* entities are all referential entities, their name is unambiguous and refer to a strongly controlled nomenclature. On the other

hand *Habitat* entities are rarely referential entities, they are usually noun phrases including properties and modifiers. There are rare cases of habitats referred with adjectives or verbs. The spans are generally contiguous but a few of them are discontinuous in order to cope with conjunctions.

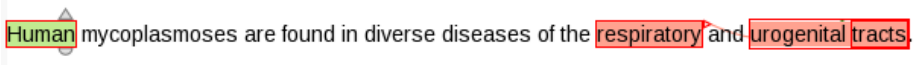

Figure 2: Example of discontinuous annotation due to a conjunction: "*respiratory tracts*".

Habitat entities are assigned to one or several concepts in the MBTO-Habitat ontology. This ontology contains The assigned concepts are as specific as possible.

Entity	**MBTO concept**	**Concept label**
"*sheep in Europe*"	MBTO:00001178	*sheep*
"*sequencing centers*"	MBTO:00001990	*research and study center*
"*milk industry*"	MBTO:00000324	*dairy industry*
"*cutaneous*"	MBTO:00000497	*skin*

Table 1: Examples of entity categorizations with MBTO concepts.

Localization relations are annotated as pairs of a *Bacteria* and a *Habitat* or a *Geographical* entities.

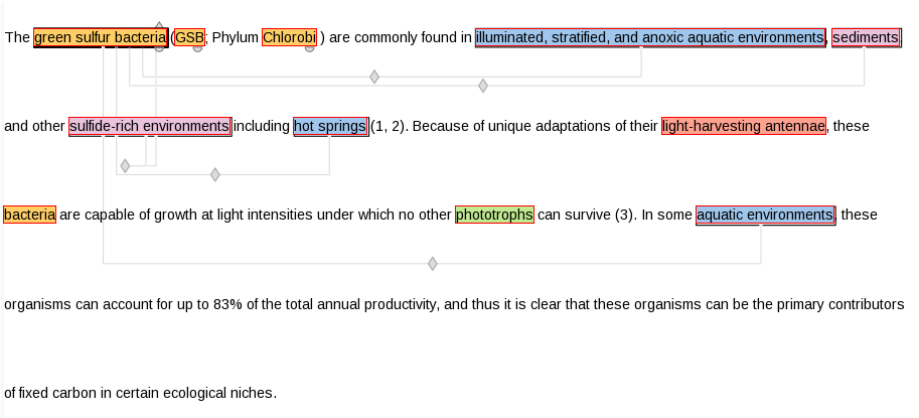

Figure 3: Example of entity and relation annotations in a text fragment

Details of the annotation will be available in the annotation guidelines.

4.2 Annotation Process

The annotation is build on top of the BioNLP-ST 2011 BB corpus. The whole corpus was re-annotated by seven annotators and two senior annotators. Their background was diverse and included: computer science, biology, bioinformatics, linguistics and knowledge engineering. However all annotators are quite familiar

with the corpus and the target domain. All annotators are very familiar with the MBTO-Habitat ontology, some of them have taken part in its design.

The annotation of entities was performed in a double-blind fashion, and the pairs have build a consensus gold annotation. Annotators have required meetings regularly in order to clarify or amend the guidelines. The MBTO-Habitat ontology was completed and modified during the annotation process.

The annotation of relations was revised by a single annotator with a biology and bioinformatics background. This annotator was very familiar with the corpus and guidelines since they have already annotated the BioNLP-ST 2011 BB corpus.