Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts

Cyril Grouin
LIMSI, CNRS, Université Paris-Saclay
Bât 508, Campus Universitaire, F-91405 Orsay
cyril.grouin@limsi.fr

Abstract

This paper presents our participation in the Bacteria/Biotope track from the 2016 BioNLP Shared-Task. Our methods rely on a combination of distinct machine-learning and rule-based systems. We used CRF and post-processing rules to identify mentions of bacteria and biotopes, a rule-based approach to normalize the concepts in the ontology and the taxonomy, and SVM to identify relations between bacteria and biotopes. On the test datasets, we achieved similar results to those obtained on the development datasets: on the categorization task, precision of 0.503 (gold standard entities) and SER of 0.827 (both NER and categorization); on the event relation task, F-measure of 0.485 (gold standard entities, ranking third out of 11) and of 0.192 (both NER and event relation, ranking first); on the knowledge-based task, mean references of 0.771 (gold standard entities) and of 0.202 (both NER, categorization and event relation).

1 Introduction

In this paper, we present the methods we used while participating in the Bacteria/Biotope track from the 2016 BioNLP Shared-Task. We partially reused the method we designed while participating in the previous edition of the challenge (Grouin, 2013), and we updated afterwards while designing new experiments (Lavergne et al., 2015).

2 Background

Four teams participated in the Bacteria/Biotope track (Bossy et al., 2015) from the 2013 BioNLP Shared-Task.

On the entity detection and categorization task, the best results were obtained using either machine-learning approaches, as done by Bannour et al. (2013) who ranked first (Slot Error Rate (SER) of 0.661), or using syntactic hand-coded rules, as done by Karadeniz and Özgür (2013) who ranked second (SER=0.676). We ranked third (SER=0.678) using CRF and normalization rules.

On the localization relation extraction task, the best results were obtained through machine-learning approaches. Björne and Salakoski (2013) ranked first (F=0.42), using a system based on Support Vector Machine (SVM), while Claveau (2013) ranked second (F=0.40) using a lazy machine learning (kNN) approach.

3 Task description

3.1 Presentation

The 2016 Bacteria/Biotope track[1] (Deléger et al., 2016) consists in three main objectives: (i) named entity recognition (NER) to identify mentions of bacteria and biotopes from scientific abstracts, (ii) categorization to normalize mentions of bacteria in the NCBI taxonomy and mentions of biotopes in the OntoBiotope ontology, and (iii) event extraction to identify relations of localization between a bacteria and a biotope.

The track is organized into three main tasks, based on gold standard annotations of entities: a categorization task (cat), an event extraction task (event), and a knowledge-base population task (kb) which combines categorization and relation identification. Additionally, each task is composed of a named entity recognition sub-task: categorization and relation identification are based on predictions of entities (cat+ner, event+ner, and kb+ner tasks) instead of gold standard annotations.

[1] http://2016.bionlp-st.org/

Proceedings of the 4th BioNLP Shared Task Workshop, pages 64–72,
Berlin, Germany, August 13, 2016.

3.2 Material

3.2.1 Corpus

The corpus is composed of 215 scientific texts (title and abstract) focusing on bacteria, extracted from the Medline database. This corpus is split into three datasets: training (71 texts), development (36 texts), and test (108 texts).[2] We used the train dataset to develop our systems and to tune our models while results produced by those systems were evaluated on the dev dataset. The test datasets were used for the official evaluation.

3.2.2 Annotations

Bossy et al. (2016) defined three kinds of entities (bacteria, habitat, geographical) and one type of relation (lives in) between a bacteria and a biotope.

Entities Annotations of entities imply three kinds of annotations: (i) single entities, (ii) embedded entities, in case of different meanings, and (iii) discontinuous entities, to deal with coordination. Figure 1 highlights discontinuous annotations (*throat cultures*) and embedded annotations (*throat* within *throat cultures*, and *nasopharyngeal* within *nasopharyngeal cultures*).

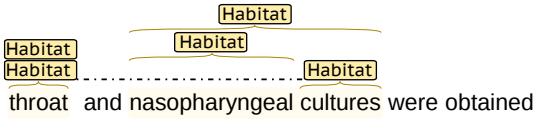

Figure 1: Discontinuous and embedded annotations of entities

Specific annotation rules apply for classifiers (*genus, species, strain*) and generic classes (*bacteria, cohort, in vivo, microbe, suspension*) which must not be annotated, except for specified strain (*mutants, serotypes, serovars*).

Categorization The categorization focuses on two types of entity (bacteria, habitat). Annotations provide the ID for each mention to be normalized, based on the NCBI taxonomy[3] (Federhen, 2002) for mentions of bacteria and the OntoBiotope ontology[4] (Nédellec, 2016) for mentions of habitat.

Mentions of bacteria are normalized into only one category while mentions of habitat can be normalized into several categories. The categorization into one or several categories for habitat mentions is dependent on the structure of the ontology, whether an "is a" relation between category candidates exists in the ontology or not (see figure 2). As an example, the mention *chicks* is normalized into three categories ("laboratory animal—000323", "infant–002177", "chicken–002229") while all mentions of *mice* are normalized into one category ("laboratory mice—002153") since this category is related with the category "laboratory animal—000323".

```
[Term]
id: OBT:000323
name: laboratory animal
is_a: OBT:000218 ! animal

[Term]
id: OBT:002153
name: laboratory mice
is_a: OBT:001865 ! mouse
is_a: OBT:000323 ! laboratory animal

[Term]
id: OBT:002229
name: chicken
is_a: OBT:002165 ! poultry
```

Figure 2: Extract from the OntoBiotope ontology

Relations Annotations of relations always imply one bacteria with one or several biotopes (habitat, geographical). Figure 3 shows relations between a bacteria and two biotopes, a geographical unit (*UK*) and a habitat (*UK retail poultry*). According to the guidelines, even if arguments from a relation must be as close as possible, one can find a few cases of relations between two distant entities. The longest distance is of 1868 characters, 276 words, implying 10 sentences.

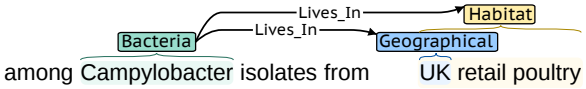

Figure 3: Example of relation annotations

3.2.3 Statistics

We present in table 1 the number of annotations for each category of entities (bacteria, habitat, geographical) and relations (lives in), as well as the number of categorizations performed in the associated resource (OntoBiotope ontology or NCBI

[2]The test dataset is split into two datasets: one set of 54 files for all tasks implying a named entity recognition (NER) process (cat+ner, event+ner, kb+ner) and a second set of 54 files giving gold standard annotations of bacteria and biotope for tasks without NER (cat, event, kb).

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

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

taxonomy) in each dataset (train, dev, and test[5]). The figures presented in cells with a grey background refer to the number of predictions to be made during the challenge. While annotations of entities are found in almost all files (one file from the train dataset does not propose any annotation), relations are found in about 80% of files (i.e., 84 files out of 107 files from the train+dev datasets). The number of annotated entities per file is quite unbalanced, from 1 to 69 entities.

Annotations		Train	Dev	Test	
				#1	#2
Number of files		71	36	54	54
Entities	Bacteria	375	244	341	401
	Habitat	747	454	720	621
	Geographical	36	38	37	27
Category	NCBI Taxon	376	245	347	401
	OntoBiotope	825	535	861	681
Relations	Lives in	327	223	340	314

Table 1: Number of annotations per category in each dataset (test #1=dataset with reference annotations of entities, #2=dataset without annotations). Grey background refers to the number of predictions to be made during the challenge

We observed that discontinuous entities: (i) mainly concern habitat entities (87.0%), (ii) generally involve two entities, more rarely three entities, and that (iii) the pivot shared by discontinuous and continuous entities is generally at the end of the portion (e.g., "cultures" in *throat and nasopharyngeal cultures*). In the training and development datasets (107 files), out of 1894 annotations of entities, we only found 46 discontinuous entities (i.e., 2.4% of annotations are discontinuous entities).

4 Methods

Based on the three main objectives of the track and the previous observations, we considered distinct systems (cf. figure 4): named entity recognition, categorization, and relation identification. We did not use any of the provided supporting resources. Due to the low number of discontinuous entities, we decided not to process this type of annotation.

[5]Test #1 refers to the test dataset with gold standard annotations of entities (cat+ner, event+ner, kb+ner tasks) while test #2 refers to the test dataset without annotations of entities (cat, event, kb tasks).

4.1 Additional data

Presentation In order to improve the robustness of our systems, we annotated a new set of 22 files.[6] To produce this new set, we queried PubMed with names of bacteria we randomly selected from the train and development datasets: *Francisella, Lactobacillus, LVS, Mycoplasma, Rickettsia, Trichomonas vaginalis* and *Vibro parahaemolyticus*. Among all results returned by PubMed, we kept abstracts published in 2016 we found interesting.

Annotations We used our systems (see sections 4.2 and 4.4) to automatically pre-annotate this dataset. One human annotator corrected and completed the automatic pre-annotations in one hour using the BRAT annotation tool (Stenetorp et al., 2012). Since we were not trained to annotate such files, even if we tried to follow the guidelines (Bossy et al., 2016), we hope our annotations are not too much inconsistent with annotations provided by the organizers. Our dataset includes 252 annotations of bacteria, 176 habitat, 31 geographical and 130 relations. Except for habitat and relations, this distribution is consistent with statistics presented in table 1.

4.2 Named Entity Recognition

4.2.1 Presentation

We considered the named entity recognition (NER) issue as a classification task, where tokens from a text should be classified into three categories (bacteria, habitat, geographical). Our NER system relies both on machine-learning approach and post-processing rules.

Machine-learning Conditional Random Fields (CRF) (Lafferty et al., 2001) are widely used for sequence labeling tasks. Our experiments rely on the Wapiti system (Lavergne et al., 2010), based on the linear-chain CRFs framework.

The feature sets are: (i) the token itself, (ii) token typographic case, presence of punctuation marks in the token, presence of digits in the token, token length, (iii) identification of the token in the OntoBiotope ontology or in the NCBI taxonomy, (iv) semantic class of the token among 37

[6]Our additional dataset, annotated before the release of the test datasets, is composed of files (title and abstract) corresponding to the following PMIDs: 1262454, 21624472, 26358917, 26510639, 26678135, 26709916, 26773254, 26901499, 26902724, 26919818, 26941131, 26941728, 26942354, 26950451, 26951983, 26961264, 26962869, 26964722, 26965788, 26965874, 26968160, 26968657. None of our additional data are also part of the test datasets.

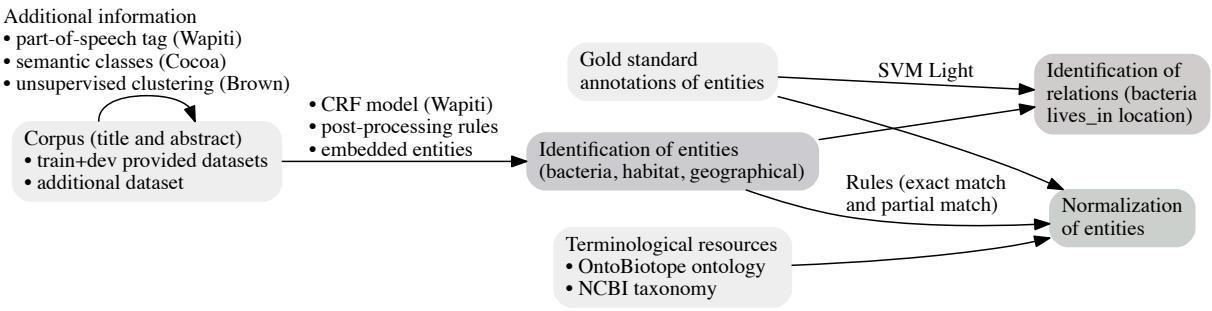

Figure 4: Systems used to identify entities, normalize entities, and identify relations

pre-defined classes *(Body part, Chemical, Food, Habitat, Organism, Physiology, etc.)*, provided by the Cocoa web API,[7] (v) part-of-speech tag[8] of the token, and (vi) cluster ID of each token through an automatic unsupervised clustering of all tokens from the train and dev datasets into 120 clusters, using the algorithm designed by Brown et al. (1992) and implemented by Liang (2005).

Since a lot of tokens from texts are not mentions of bacteria, habitat and geographical,[9] those unannotated tokens lead to an unbalanced distribution of data. This may imply an over-training of the CRF system of the unannotated tokens. In order to reduce this over-training issue, we deleted portions of unannotated tokens. Specifically, we deleted parts of text composed of unannotated tokens, if those parts are distant of more than 16 tokens[10] from the closest annotated token. As a consequence, we kept the wholeness of the context of annotated parts and we reduced the number of unannotated tokens in our training set.

We tuned our system to predict widest entities since we considered that shorter entities can easily be identified through post-processing rules. Because embedded entities only concern habitats, this strategy does not concern bacteria and geographical units. So that the CRF produces widest entities, in case of embedded annotations, we only kept the widest entities in the sample file given as input to train the CRF model.

Post-processing In order to improve the predictions we made in the previous step and to deal with some of the specific annotation rules defined in the guidelines (Bossy et al., 2016), we designed a few post-processing rules:

- annotation of abbreviations *(EHEC, EPEC, LVS, MRSA, etc.)*, generic classes with an initial upper case *(Bacteria, Bacterium)*, some nomenclatural suffixes *(sp., spp.)*, adjectives for habitat *(aquatic, nosocomial, saprophyte)* and geographical *(northern, southern, etc.)*;

- deletion of annotations for generic classes *(bacteria, bacterial, bacterium)*, modifiers *(methicillin-resistant, pathogenic)*, some nomenclatural suffixes *(gen. nov., sp. nov.)*, and 34 generic habitat terms *(antibiotic, ecosystem, world, etc.)*.

Embedded entities Since our CRF predicted widest entities, we processed embedded habitat entities through a post-processing system. For all predictions of mentions of habitat, we searched for shortened entities within widest entities. As an example, based on the prediction *gastric mucosa-associated lymphoma*, this simple rule allows us to identify the single mention *gastric*. We thus increased the coverage of the habitat mentions.

4.2.2 Design of experiments

We designed several experiments, depending on the size of the training corpus and whether we used or not post-processing rules and embedded entities processing. Results are presented in section 5.1.1. The configuration we used on the test dataset is the following one: we trained the final CRF model on all available annotated files (193 files),[11] we

[7] Cocoa: compact cover annotator for biological noun phrases, `http://npjoint.com/annotate.php`

[8] POS tagging was performed using an English POS CRF model for Wapiti: `https://wapiti.limsi.fr/`

[9] Based on our tokenization, among 15 530 tokens from the training dataset, only 2 110 of them (i.e., 13.59%) are part of bacteria, habitat and geographical mentions.

[10] This distance of 16 tokens has been chosen empirically. This threshold reduced by 23.1% the number of unannotated tokens in the training dataset. From now on, the 2 110 annotated tokens represent 20.45% of all tokens.

[11] Those annotated files came from the training dataset (71 files), the development dataset (36 files), the additional dataset we manually annotated (22 files), and the test #1

applied post-processing rules to correct the CRF outputs, and we processed the embedded entities through a last script.

4.3 Categorization

Exact match We performed the categorization task using a basic rule-based approach. We searched the mention to normalize in the Onto-Biotope ontology (habitat) or in the NCBI taxonomy (bacteria), through an exact match search, and returned the corresponding numeric identifier.

Partial match Additionally, we searched for partial matching of mentions of bacteria in the taxonomy: (i) shortened versions: *H. pylori* vs. *Helicobacter pylori*, (ii) specified versions: *bacillus intermedius s3-19* vs. *bacillus intermedius*, and (iii) linguistics variations: plural form (*lactobacilli* vs. *lactobacillus*) or adjectival derivation (*mycobacterial* vs. *mycobacteria*). Similarly, we searched for partial matching of mentions of habitat in the ontology: (i) linguistic variations: plural forms (*patients* vs. *patient*), hand-coded nominalization of adjectives (*clinical* vs. *clinic*), (ii) split of multi-terms into single terms (*human* and *blood* vs. *human blood*), and (iii) hand-coded transformation of specific cases (*adult* is replaced by *human adult*; *children* is replaced by *child*).

Default value At last, we defined default values for all unmatched mentions of bacteria and habitat, based on the most used values in the training and development datasets (this choice is not relevant for all unmatched mentions but it allows us to slightly improve our results). We used the taxonomy entry #210 (i.e., *Campilobacter pylori* and *Helicobacter pilori*) for bacteria, and the Onto-Biotope entry #002216 *patient with infectious disease* (the second most used category) for habitat.

4.4 Relation Extraction

In order to identify relations between bacteria and biotope, we designed experiments based on the SVM framework (Vapnik, 1995), as done by Björne and Salakoski (2013). Our experiments rely on the SVM Light implementation proposed by Joachims (1999). Since a few long distance relations exist, in order to ensure the robustness of our system, we decided to remove all relations implying a distance higher than 80 tokens between both entities from our training set. This threshold produced the best results. It allows us to keep the shortest relations from the training dataset (i.e., 60% of all positive relations). We strictly balanced positive and negative examples to train our model.

The feature sets are: (i) a bag of words of all tokens from both entities to be linked, and (ii) the distance in characters between those entities.

5 Results

5.1 Development dataset

In this section, we present the results we achieved on the development dataset. Since we produced outputs compatible with the BRAT annotation tool, results were computed using the BRATeval evaluation tool developed by Verspoor et al. (2013) and updated by Deléger et al. (2014). This evaluation tool allows us to evaluate all kinds of entities (single, embedded and discontinuous entities) as well as relations between entities.

5.1.1 Named entity recognition

Table 2 presents the results we achieved on the development dataset in the named entity recognition sub-task. We give both the F-measure we achieved on each category (bacteria, habitat, geographical) and the detailed overall results (exact match). We designed five experiments:

1. CRF model trained on the train dataset (71 files);
2. CRF model trained on the train+additional datasets (93 files);
3. CRF model trained on the train+additional datasets (93 files) using an over-training reduction function (we reduced the number of tokens which must not be annotated);
4. CRF model trained on the train+additional datasets (93 files) using an over-training reduction function, and post-processing rules were applied (all categories);
5. CRF model trained on the train+additional datasets (93 files) using an over-training reduction function, post-processing rules were applied (all categories), and embedded entities (habitat) were processed.

dataset (54 files). For clarification, the named entity recognition evaluation (cat+ner, event+ner, kb+ner tasks) is performed on the test #2 dataset, composed of different files than the test #1 dataset. As a consequence, since there is no common files between test datasets #1 and #2, the use of the annotated files from the test #1 dataset to train the final CRF model does not hedge the official evaluation.

#	Entity F-measures			Overall results		
	Bact	Hab	Geo	P	R	F
1	0.668	0.470	0.727	0.721	0.452	0.556
2	0.769	0.462	0.739	0.753	0.488	0.592
3	0.772	0.469	0.739	0.740	0.500	0.597
4	0.785	0.469	0.739	0.747	0.504	0.602
5	0.785	0.523	0.739	0.737	0.548	0.628

Table 2: Results on the development dataset, F-measure for each category (Bact=Bacteria, Hab=Habitat, Geo=Geographical) and overall results (P=Precision, R=Recall, F=F-measure) depending on the experiment

5.1.2 Categorization

Table 3 presents the results we achieved on the development dataset for the categorization task. Our evaluation only computes an exact match between the IDs from the taxonomy and the ontology provided in the hypothesis and the reference. This evaluation does not compute any similarity distance within the hypothesis and reference categories. We give the overall and detailed results for both the OntoBiotope ontology and the NCBI taxonomy. Results are provided for two tasks:

1. categorization performed on the entities identified by our CRF system, configuration #5 (cat+ner task);
2. categorization performed on the gold standard annotations of entities (cat task).

#	Evaluation	P	R	F
1	Overall	0.404	0.286	0.335
	OntoBiotope	0.509	0.360	0.422
	NCBI taxonomy	0.621	0.457	0.527
2	Overall	0.456	0.412	0.433
	OntoBiotope	0.570	0.515	0.541
	NCBI taxonomy	0.886	0.885	0.886

Table 3: Results (exact match) on the development dataset on the categorization tasks (P=Precision, R=Recall, F=F-measure)

5.1.3 Relations

Table 4 presents the results we achieved (exact match) on the development dataset in the relation identification task. We designed four experiments:

1. SVM model trained on the train dataset (71 files), prediction of entities from the CRF system (event+ner task);
2. SVM model trained on the train+additional dataset (93 files), prediction of entities from the CRF system (event+ner task);
3. SVM model trained on the train dataset (71 files), gold standard annotations of entities (event task);
4. SVM model trained on the train+additional dataset (93 files), gold standard annotations of entities (event task).

#	Evaluation	P	R	F
Entities from the CRF system (event+ner task)				
1	Overall	0.171	0.213	0.189
	Bacteria-Habitat	0.162	0.235	0.192
	Bacteria-Geographical	0.364	0.111	0.170
2	Overall	0.190	0.213	0.201
	Bacteria-Habitat	0.181	0.235	0.204
	Bacteria-Geographical	0.400	0.111	0.174
Entities from the gold standard (event task)				
3	Overall	0.381	0.652	0.480
	Bacteria-Habitat	0.355	0.658	0.461
	Bacteria-Geographical	0.622	0.622	0.622
4	Overall	0.385	0.652	0.484
	Bacteria-Habitat	0.357	0.658	0.463
	Bacteria-Geographical	0.657	0.627	0.639

Table 4: Results on the development dataset on the relation identification tasks (P=Precision, R=Recall, F=F-measure)

5.1.4 Online evaluation service

Since the online evaluation service provides a distinct evaluation (giving final scores and using different metrics), in order to compare the results we achieved on both the development and the test datasets, we present in table 5 the results we achieved on all tasks on the development datasets using our last configuration, as computed by the evaluation service.

5.2 Test dataset (official results)

Table 6 presents the results we achieved on the test dataset. Our results are similar to results obtained on the development datasets. This observation highlights the robustness of our methods.

We ranked second (out of 2) on all categorization tasks. We ranked third (out of 11) on the event task, and first (out of 3) on the event+ner task. At last, we were the only participant on all knowledge-based tasks.

task	Official results (dev)			
cat	Precision			
	1.000			
cat+ner	SER	Mism	Ins	Del
	0.702	49.85	127	314
event	Precision	Recall	F-measure	
	0.389	0.644	0.485	
event+ner	SER	P	R	F
	1.486	0.216	0.201	0.208
kb	Mean references			
	0.7861			
kb+ner	Mean references			
	0.2074			

Table 5: Official results computed on the development datasets (SER=Slot Error Rate, Mism=Mismatch, Ins=Insertion, Del=Deletion, P=Precision, R=Recall, F=F-measure)

task	Official results (test)			
cat	Precision			
	0.503			
cat+ner	SER	Mism	Ins	Del
	0.827	198.16	192	455
event	Precision	Recall	F-measure	
	0.388	0.646	0.485	
event+ner	SER	P	R	F
	1.558	0.193	0.192	0.192
kb	Mean references			
	0.7714			
kb+ner	Mean references			
	0.2024			

Table 6: Official results computed on the test dataset (SER=Slot Error Rate, Mism=Mismatch, Ins=Insertion, Del=Deletion, P=Precision, R=Recall, F=F-measure)

6 Discussion

6.1 Observations

Additional data A first observation concerns the use of additional data. Increasing the number of annotated files proved to be useful for all machine-learning approaches. In the named entity recognition task—using a CRF system—we gained +3.6 points of F-measure (see table 2). In the relation identification task—using a SVM system—we gained +1.2 points of F-measure for relations based on entities predicted by the CRF, and +0.4 point for relations based on gold standard entities annotations (see table 4). The advantage of using more annotated data is real for all tasks.

Post-processing rules Despite the use of both additional data and over-training reduction function, the CRF model achieved moderate results (F=0.597, see table 2). The use of post-processing rules to refine the CRF outputs slightly increased the overall results (+0.5 points, F=0.602) and mainly impacted the bacteria category (+1.3 points). At last, processing embedded habitat entities with rules improved the overall results (+2.6 points, F=0.628). Using a few post-processing rules increased by +3.1 points the overall results achieved through the CRF model.

Named entity recognition Our strategy based on four steps (additional annotated data, over-training reduction function, post-processing rules, and embedded entities processing) allows us to achieve quite moderate results (F=0.628, see table 2). We failed to identify correctly entities of habitat (F=0.523) while results are higher for both bacteria (F=0.785) and geographical (F=0.739).

Nevertheless, when annotating additional data, we experienced harder work for habitat than for bacteria or geographical. As a consequence, this type of entities is complex for both human annotators and automatic systems.

Categorization The rule-based approach we designed to categorize entities in both the OntoBiotope ontology and the NCBI taxonomy is quite simple. Since our named entity recognition system obtained moderate results (overall F-measure of 0.628, see table 2), on the categorization task, we achieved better results on the gold standard annotations of entities (overall F-measure of 0.446) than on predictions of entities made by our CRF system (overall F-measure of 0.338, see table 3).

Since we failed to categorize more habitat than bacteria, using default categorization values (see section 4.3) led us to obtain lower precision values for habitat, on both cat+ner (P_{hab}=0.482 vs. P_{bact}=0.714) and cat (P_{hab}=0.518 vs. P_{bact}=0.983) tasks. Moreover, the lowest recall values are also obtained on the categorization of habitat.

6.2 Error analysis

We give in figure 5 a sample of annotations performed by our system on the development dataset (event+ner task).

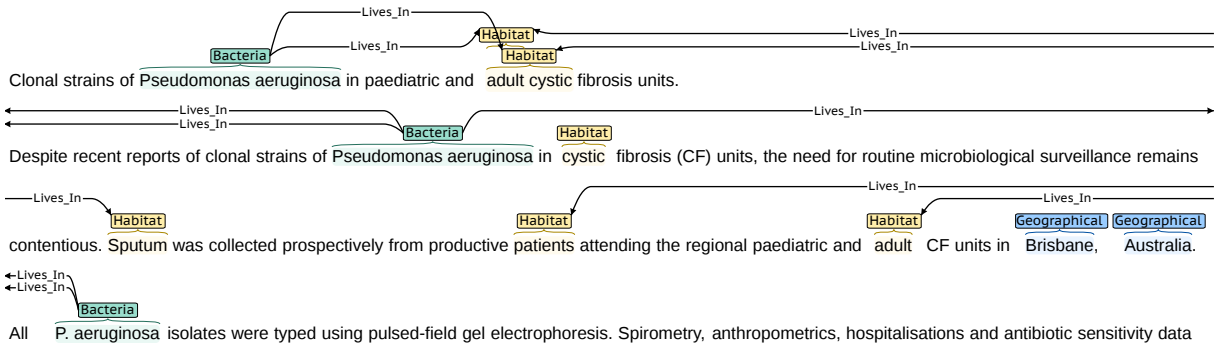

Figure 5: Sample of entities and relations predicted by our system on the development dataset (event+ner task). The first line is the title of the scientific text while other lines are part of the first paragraph

On the NER tasks, our system failed to identify acronyms (*HMDM, HMDMs, PMN, PMNs*), and all discontinuous entities since we chose to not process this kind of entity. False negatives mainly concern habitats: (i) single entities (*paediatric*), (ii) discontinuous entities (*paediatric ... cystic fibrosis units*, *regional ... adult CF units*, and *regional paediatric ... CF units*), and (iii) frontiers errors for which annotations depend on the context (*adult cystic fibrosis units* vs. only *adult cystic* in our sample, *cystic fibrosis (CF) units* vs. *cystic*, or *productive patients* vs. *patients*).

On the categorization tasks, the main errors concern all entities we failed to categorize and for which we gave a default value. Those entities refer to adjectives the system did not process (*pulmonary, duodenal,* etc.) and complex entities (*vacuum- and modified-atmosphere-packed cold-smoked salmon stored at 5 degrees C*, categorized as "vacuum-packed meat" in the reference). As a consequence, each category used as a default value obtained bad results on the development dataset: the NCBI taxonomy entry #210 achieved 34 true positives and 69 false positives while the OntoBiotope entry #002216 achieved 14 true positives, 207 false positives and 11 false negatives.

At last, on the event identification tasks, since there is only one type of relation to identify, the errors concern missing relations and too much relations. Missing relations concern geographical entities (cf. missing relations between *P. aeruginosa* and geographical entities *Brisbane* and *Australia* on figure 5): due to a low number of entities in this category (see table 1), our SVM model failed to learn relations with geographical entities. False positives concern cases where the context between entities prohibits relations (*Neutrophils are resistant to Yersinia*), and annotations done on several lines, including between the content of the title and the content of the other paragraphs (cf. relations between *Pseudomonas aeruginosa* from the first paragraph and habitats *adult cystic* and *adult* from the title).

7 Conclusion

In this paper, we presented the experiments we made while participating in the Bacteria/Biotope track from the 2016 BioNLP Shared-Task. We combined CRF and post-processing rules to identify entities (bacteria, habitat, geographical), including embedded entities, and we used rules based on exact and partial match to normalize the entities in the NCBI taxonomy (bacteria) and the OntoBiotope ontology (habitat). For relation extraction, we used a SVM system based on a basic set of features.

As future work, we plan to deal with discontinuous entities. To process this issue, we consider that a CRF model making the distinction between the pivot and tokens specific to each entity would be useful. As an example, in *throat and nasopharyngeal cultures*, the pivot is *cultures* while specific tokens are *throat* and *nasopharyngeal*. Post-processing rules would bring together tokens so as to produce the final entities (*throat cultures* and *nasopharyngeal cultures*). Our categorization approach to search for partial matches is relatively simple. Future work is needed to provide a better processing of the OntoBiotope ontology, namely, in order to take into account the "is a" relations.

At last, we estimate that using unsupervised learning of relations may provide interesting results, especially to improve the features set used in the SVM model.

References

Sondes Bannour, Laurent Audibert, and Henry Soldano. 2013. Ontology-based semantic annotation: an automatic hybrid rule-based method. In *BioNLP-ST Work Proc*, pages 139–43, Sofia, Bulgaria.

Jari Björne and Tapio Salakoski. 2013. TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. In *BioNLP-ST Work Proc*, pages 16–25, Sofia, Bulgaria.

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(Suppl 10):S1.

Robert Bossy, Claire Nédellec, Julien Jourde, and Mouhammadou Ba, 2016. *Guidelines for Annotation of Bacteria Biotopes*. INRA.

Peter F Brown, Vincent J Della Pietra, Peter V de Souza, Jenifer C Lai, and Robert L Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–79.

Vincent Claveau. 2013. IRISA participation to BioNLP-ST 2013: lazy-learning and information retrieval for information extraction tasks. In *BioNLP-ST Work Proc*, pages 188–96, Sofia, Bulgaria.

Louise Deléger, Anne-Laure Ligozat, Cyril Grouin, Pierre Zweigenbaum, and Aurélie Névéol. 2014. Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proc of LREC*, pages 1267–74, Reykjavik, Iceland.

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

Scott Federhen. 2002. The Taxonomy Project. In Johanna McEntyre and Jim Ostell, editors, *The NCBI Handbook*, chapter 4. National Center for Biotechnology Information, Bethesda, MD, 2nd edition.

Cyril Grouin. 2013. Building A Contrasting Taxa Extractor for Relation Identification from Assertions: BIOlogical Taxonomy & Ontology Phrase Extraction System. In *BioNLP-ST Workshop Proc*, pages 144–52, Sofia, Bulgaria.

Thorsten Joachims. 1999. Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

İlknur Karadeniz and Arzucan Özgür. 2013. Bacteria Biotope Detection, Ontology-based Normalization, and Relation Extraction using Syntactic Rules. In *BioNLP-ST Work Proc*, pages 170–7, Sofia, Bulgaria.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden.

Thomas Lavergne, Cyril Grouin, and Pierre Zweigenbaum. 2015. The contribution of co-reference resolution to supervised relation detection between bacteria and biotopes entities. *BMC Bioinformatics*, 16(Suppl 10):S6.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Claire Nédellec. 2016. The OntoBiotope Ontology. Institut National de la Recherche Agronomique.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Juni'chi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proc of EACL Demonstrations*, pages 102–7, Avignon, France. ACL.

Vladimir N Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zo Thomas, and John-Paul Plazzer. 2013. Annotating the Biomedical Literature for the Human Variome. In *Database: The Journal of Biological Databases and Curation*.