Building A Contrasting Taxa Extractor for Relation Identification from Assertions: BIOlogical Taxonomy & Ontology Phrase Extraction System

Cyril Grouin
LIMSI–CNRS, Orsay, France
cyril.grouin@limsi.fr

Abstract

In this paper, we present the methods we used to extract bacteria and biotopes names and then to identify the relation between those entities while participating to the BioNLP'13 Bacteria and Biotopes Shared Task. We used machine-learning based approaches for this task, namely a CRF to extract bacteria and biotopes names and a simple matching algorithm to predict the relations. We achieved poor results: an SER of 0.66 in sub-task 1, and a 0.06 F-measure in both sub-tasks 2 and 3.

1 Introduction

The BioNLP'13 Bacteria and Biotopes shared task aims at extracting bacteria names *(bacterial taxa)* and biotopes names *(bacteria habitats; geographical and organization entities)*. The task comprises three sub-tasks (Bossy et al., 2012b).

- Sub-task 1 aims at extracting habitat names and linking those names to the relevant concept from the OntoBiotope ontology.
- Sub-task 2 aims at identifying relations between bacteria and habitats among two kinds of relations *(localization, part-of)* based on a ground truth corpus of bacteria and habitat names. The "localization" relation is the link between a bacterium and the place where it lives while the "part-of" relation is the relation between hosts and host parts *(bacteria)* (Bossy et al., 2012a).
- Sub-task 3 aims at extracting all bacteria and biotopes names *(including both habitat and geographical names)*, and then identifying relations between these concepts.

In this paper, we present the methods we designed as first time participant to the BioNLP Bacteria Biotopes Shared Task.

2 Background

Scientific documents provide useful information in many domains. Because processing those documents is time-consuming for a human, NLP techniques have been designed to process a huge amount of documents quickly. The microorganisms ecology domain involves a lot of microorganisms *(bacteria, living and dead cells, etc.)* and habitats *(food, medical, soil, water, hosts, etc.)* that have been described in details in the literature. NLP techniques would facilitate the access to information from scientific texts and make it available for further studies.

Bacteria and biotopes identification has been addressed for the first time during the BioNLP 2011 Bacteria Biotopes shared task (Bossy et al., 2012a; Kim et al., 2011). This task consisted in extracting bacteria location events from texts among eight categories *(Host, HostPart, Geographical, Environment, Food, Medical, Water and Soil)*.

Three teams participated in this task. All systems followed the same process: in a first stage, they detected bacteria names, detected and typed locations; then, they used co-reference to link the extracted entities; the last stage focused on the event extraction.

Björne et al. (2012) adapted an SVM-based Named Entity Recognition system and used the list of Prokaryotic Names with Standing in Nomenclature. Nguyen and Tsuruoka (2011) used a CRF-based system and used the NCBI web page about the genomic BLAST. Ratkovic et al. (2012) designed an *ad hoc* rule-based system based on the NCBI Taxonomy. The participants obtained poor results (Table 1) which underlines the complexity of this task.

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 144–152,
Sofia, Bulgaria, August 9 2013. ©2013 Association for Computational Linguistics

Team	R	P	F
Ratkovic et al. (2012)	**0.45**	0.45	**0.45**
Nguyen and Tsuruoka (2011)	0.27	0.42	0.33
Björne et al. (2012)	0.17	**0.52**	0.26

Table 1: Recall, Precision and F-measure at BioNLP 2011 Bacteria and Biotopes Shared Task

3 Corpus

3.1 Presentation

The corpus comprises web pages about bacterial species written for non-experts. Each text consists of a description of individual bacterium and groups of bacteria, in terms of first observation, characteristics, evolution and biotopes. Two corpora have been released including both raw textual documents and external reference annotations. The training corpus contains 52 textual documents while the development corpus contains 26 documents. No tokenization has been performed over the documents. In Table 2, we provide some statistics on the annotations performed over both corpora for each type of entity to be annotated *(bacteria, habitat,* and *geographical).*

Corpus	Training	Development
# Documents	52	26
# Words	16,294	9,534
Avg # words/doc	313.3	366.7
# Bacteria	832	515
# Habitat	934	611
# Geographical	91	77

Table 2: Annotation statistics on both corpora

3.2 Corpus analysis

The bacteria names appear in the texts, either in their longer form *(Xanthomonas axonopodis pv. citri)*, in a partial form *(Xanthomonas)* or in their abbreviated form *(Xac)*. The abbreviations are case-sensitives since they follow the original form: *MmmSC* is derived from *M. mycoides ssp mycoides SC*.[1] A few bacteria names can appear in the text followed by a trigger word: *Spirillum bacteria*, but it will be abbreviated in the remainder of the text, sometimes with a higher degree of specificity: *S. volutans* standing for *Spirillum volutans*.

4 Methods

This year, the BioNLP organizers encouraged the participants to use supporting resources in order to reduce the time-investment in the challenge. Those resources encompass sentence splitting, tokenization, syntactic parsing, and biological annotations. Moreover, a specific ontology has been released for the Bacteria Biotopes task.

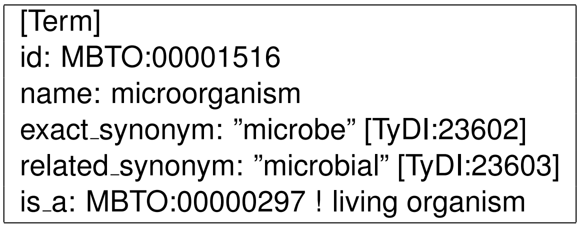
We used some of the resources provided and combined them with additional resources, in a machine-learning framework we specifically designed for this task.

4.1 Linguistic resources

4.1.1 The OntoBiotope Ontology

OntoBiotope[2] is an ontology tailored for the biotopes domain. The BioNLP-ST 2013 version has been released in the OBO format. This ontology integrates 1,756 concepts. Each concept has been given a unique ID and is associated with exact terms and related synonyms. The concept is also defined in a “is_a” relation. The normalization of the habitat names in the first sub-task must be based on this ontology.

For example, the concept *microorganism* (unique id MBTO:00001516) is a *living organism* which unique id is MBTO:00000297. For this concept, *microbe* is an exact synonym while *microbial* is a related synonym (see Figure 1).



```
[Term]
id: MBTO:00001516
name: microorganism
exact_synonym: ”microbe” [TyDI:23602]
related_synonym: ”microbial” [TyDI:23603]
is_a: MBTO:00000297 ! living organism
```

Figure 1: The concept *microorganism* in the OntoBiotope ontology

4.1.2 The NCBI taxonomy

In order to help our system to identify the bacteria names, we built a list of 357,387 bacteria taxa based on the NCBI taxonomy database[3] (Federhen, 2012). This taxonomy describes a small part *(about 10%)* of the living species on earth, based on public sequence databases.

[1] *Mycoplasma mycoides subspecies mycoides Small Colony* in its longer form.

[2] `http://bibliome.jouy.inra.fr/MEM-OntoBiotope/OntoBiotope_BioNLP-ST13.obo`

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

It includes twelve categories of information from the biological domain *(bacteria, invertebrates, mammals, phages, plants, primates, rodents, synthetics, unassigned, viruses, vertebrates* and *environmental samples).*

We extracted from this taxonomy all names belonging to the *Bacteria* category, which represent 24.3% of the content. This output includes a few variants of bacteria names (see Table 3).

tax_id	name_txt	name class
346	Xanthomonas citri (ex Hasse 1915) Gabriel et al. 1989	authority
346	Xanthomonas citri	scientific name
346	Xanthomonas axonopodis pv. citri	synonym
346	Xanthomonas campestris (pv. citri)	synonym
346	Xanthomonas campestris pv. Citri (A group)	synonym

Table 3: Bacteria names from the NCBI taxonomy

4.1.3 The Cocoa annotations

Cocoa is a WebAPI annotator tool for biological text.[4] We used the Cocoa annotations provided by the organizers as part of the supporting resources. These annotations emphasize 37 pre-defined categories. We noticed a few categories are often tied with one of the three kinds of entities we have to process:

- Bacteria: *Cell, Chemical, Mutant_Organism, Organism, Protein, Unknown*;
- Habitat: *Body_part, Cell, Cellular_component, Chemical, Disease, Food, Geometrical_part, Habitat, Location, Multi-tissue_structure, Organism, Organism_subdivision, Pathological_formation, Tissue*;
- Geographical: *Company, Habitat, Technique, Unknown*.

We believe these categories should be useful to identify bacteria and biotopes entities in the texts, and we used them as features in the CRF model (see column #10 in Table 4).

4.2 System

4.2.1 Formalisms

Depending on the sub-task to process, we used two distinct formalisms implemented in the Wapiti tool (Lavergne et al., 2010) to build our models:

- Conditional Random Fields *(CRF)* (Lafferty et al., 2001; Sutton and McCallum, 2006) to identify bacteria and biotopes names *(subtasks 1 and 3).*
- Maximum Entropy *(MaxEnt)* (Guiasu and Shenitzer, 1985; Berger et al., 1996) to process the relationships between entities *(subtasks 2 and 3).*

4.2.2 Bacteria biotopes features set

We used several sets of features, including “classical” internal features *(columns #4 to #7 in Table 4: typographic, digit, punctuation, length)* and a few semantic features. In table 4, we present a sample tabular file produced in order to train the CRF model.

- Presence of the token in the NCBI taxonomy (column #9);
- Presence of the token in the OntoBiotope ontology (column #8);
- Category of the token based on the Cocoa annotations (column #10);
- Unsupervised clusters (column #11) created using Brown's algorithm (Brown et al., 1992) with Liang's code[5] (Liang, 2005).

Taxonomy feature. We noticed that 1,169 tokens out of 1,229 (95.1%) tokens we identified in the NCBI taxonomy in both corpora correspond to a Bacteria name in the reference (Table 5). This characteristic should be useful to identify the bacteria names.

OntoBiotope feature. Regarding the presence of the token in the OntoBiotope ontology, we noticed that 1,487 tokens out of 1,906 (78.0%) from both corpora correspond to a habitat name in the reference (Table 6). The identification of habitat names will benefit from this characteristic.

[4] Compact cover annotator for biological noun phrases, `http://npjoint.com/annotate.php`

[5] `http://www.cs.berkeley.edu/~pliang/software/`

1	2	3	4	5	6	7	8	9	10	11	12
33	8	Borrelia	Mm	O	O	7	O	NCBI	Organism	11101010	B-Bacteria
42	7	afzelii	mm	O	O	7	O	NCBI	Organism	O	I-Bacteria
49	1	.	O	Punct	O	1	O	O	O	0010	O
51	4	This	Mm	O	O	4	O	O	O	1001000	O
56	7	species	mm	O	O	7	O	O	Organism1	100101100	O
64	3	was	mm	O	O	3	O	O	O	0101000	O
68	8	isolated	mm	O	O	7	O	O	O	1100100	O
77	4	from	mm	O	O	4	O	O	O	011110110	O
82	1	a	mm	O	O	1	O	O	O	1011000	O
84	4	skin	mm	O	O	4	MBTO	O	Pathological _formation	110111011	B-Habitat
89	6	lesion	mm	O	O	6	MBTO	O	Pathological _formation	111101100	I-Habitat
96	4	from	mm	O	O	4	O	O	O	011110110	I-Habitat
101	1	a	mm	O	O	1	O	O	O	1011000	I-Habitat
103	4	Lyme	Mm	O	O	4	O	O	Disease	100010	I-Habitat
108	7	disease	mm	O	O	7	O	O	Disease	110111101	I-Habitat
116	7	patient	mm	O	O	7	MBTO	O	Organism2	1100110	I-Habitat
124	2	in	mm	O	O	2	O	O	O	0111100	O
127	6	Europe	Mm	O	O	6	MBTO	O	Habitat	111101101	B-Geographical
134	2	in	mm	O	O	2	O	O	O	0111100	O
137	4	1993	O	O	Digit	4	O	O	O	111101101	O
141	1	.	O	Punct	O	1	O	O	O	0010	O

Table 4: Tabular used for training the CRF model. Column 1: character offset; 2: length in characters; 3: token; 4: typographic features; 5: presence of punctuation; 6: presence of digit; 7: length in characters (with a generic '7' category for length higher than seven characters); 8: presence of the token in the OntoBiotope ontology; 9: presence of the token in the NCBI taxonomy; 10: category of the token from the Cocoa annotations; 11: cluster identifier; 12: expected answer

Reference annotation	Token in the NCBI	
	Present	Absent
Bacteria	**1,169**	1,543
Geographical	0	276
Habitat	2	2,466
O *(out of annotation)*	58	25,060

Table 5: Correspondence between the reference annotation and the token based on the presence of the token in the NCBI taxonomy

Reference annotation	Token in OntoBiotope	
	Present	Absent
Bacteria	1	2,711
Geographical	156	120
Habitat	**1,487**	981
O *(out of annotation)*	262	24,856

Table 6: Correspondence between the reference annotation and the token based on the presence of the token in the OntoBiotope ontology

4.2.3 Normalization with OntoBiotope

Habitat names normalization consisted in linking the habitat names to the relevant concept in the OntoBiotope ontology using an exact match of the phrase to be normalized. This exact match is based on both singular and plural forms of the phrase to normalize, using a home-made function that includes regular and irregular plural forms. Nevertheless, we did not manage discontinuous entities.

4.2.4 Relationships approaches

Relationships features set. Our MaxEnt model only relies on the kind of entities that can be linked together:

- *Bacterium* and *Localization* (Habitat) for a "localization" relation,
- *Host* and *Part* for a "PartOf" relation (between two entities being of the same type).

For example, *Bifidobacterium* is a bacteria name, *human* and *human gastrointestinal tract* are two habitats (localizations). A “localization” relation can occur between *Bifidobacterium* and *human* while a “PartOf” relation occurs between *human* and *human gastrointestinal tract*.

Basic approach. For the official submission, we did not use this model because of the following remaining problems: (i) a few relations we produced were not limited to the *habitat* category but also involved the *geographical* category, (ii) we did not manage the relations we produced in duplicate, and (iii) the weight our CRF system gave to each relation was not relevant enough to be used (for a relation involving A with B, C, and D, the same weight was given in each relation).

All of those problems led us to process the relations between entities using a too much simple approach: we only considered if the relation between two entities from the test exists in the training corpus. This approach is not robust as it does not consider unknown relations.

5 Results and Discussion

5.1 Identification of bacteria and biotopes

In this subsection, we present the results we achieved on the development corpus (Table 2) to identify bacteria and biotopes names without linking those names to the concept in the OntoBiotope ontology. We built the model on the training corpus and applied it on the development corpus. The evaluation has been done using the *conlleval.pl* script[6] (Tjong Kim Sang and Buchholz, 2000) that has been created to evaluate the results in the CoNLL-2000 Shared Task. We chose this script because it takes as input a tabular file which is commonly used in the machine-learning process. Nevertheless, the script does not take into account the offsets to evaluate the annotations, which is the official way to evaluate the results. We give in Table 7 the results we achieved. Those results show our system succeed to correctly identify the bacteria and biotopes names. Nevertheless, the biotopes names are more difficult to process than the bacteria names. Similarly, Kolluru et al. (2011) achieved better results on the *bacteria* category rather than on the *habitat*, confirming this last category is more difficult to process.

Category	R	P	F
Bacteria	0.8794	0.9397	0.9085
Geographical	0.6533	0.7903	0.7153
Habitat	0.6951	0.8102	0.7482
Overall	0.7771	0.8715	0.8216

Table 7: Results on the bacteria biotopes identification (development corpus)

There is still room for improvement, especially in order to improve the recall in each category. We plan to define some post-treatments so as to identify new entities and thus, increase the recall in those three categories.

5.2 Official results

	SER			
Sub-task 1	0.66			4th/4
	R	P	F	
Sub-task 2	0.04	0.19	0.06	4th/4
Sub-task 3	0.04	0.12	0.06	2nd/2

Table 8: Official results and rank for LIMSI

5.2.1 Habitat entities normalization

General results. The first sub-task is evaluated using the Slot Error Rate (Makhoul et al., 1999), based on the exact boundaries of the entity to be detected and the semantic similarity of the concept from the ontology between reference and hypothesis (Bossy et al., 2012b). This semantic similarity is based on the “is_a” relation between two concepts.

We achieved a 0.66 SER which places us 4th out of four participants. Other participants obtained SERs ranging from 0.46 to 0.49. Our system achieved high precision (0.62) but low recall (0.35). It produced two false positives and 144 false negatives. Out of 283 predicted habitats, 175.34 are correct. There was also a high number of substitutions (187.66).

Correct entity, incorrect categorization. On the entity boundaries evaluation, our system SER (0.45) was similar to that of the other participants (from 0.46 to 0.42). We achieved a 1.00 precision, a 0.56 recall and a 0.71 F-measure (the best from all participants). Those results are consistent with those we achieved on the development corpus (Table 7) and confirm the benefit of using a CRF-based system for entity detection.

[6] http://www.clips.ua.ac.be/conll2000/chunking/

While we correctly identified the *habitat* entities, the ontology categorization proved difficult: we achieved an SER of 0.62 while other participants obtained SERs ranging from 0.38 to 0.35. For this task, we relied on exact match for mapping the concept to be categorized and the concepts from the ontology, including both singular and plural forms match. When no match was found, because the categorization was mandatory, we provided a default identifier—the first identifier from the ontology—which is rarely correct.[7]

5.2.2 Relationships between entities

General results. The relation sub-task is evaluated in terms of recall and precision for the predicted relations. On both second and third subtasks, due to our too basic approach, we only achieved a 0.06 F-measure. Obviously, because considering only existing relations is not a robust approach, the recall is very low (R=0.04). The precision is not as high as we expected (P=0.19), which indicates that if a relation exists in the training corpus for two entities, this relation does not necessarily occur within the test for the two same entities *(two entities can occur in the same text without any relation to be find between them).* On the second sub-task, other participants obtained F-measures ranging from 0.42 to 0.27, while on the third sub-task, the other participants obtained a 0.14 F-measure, which underlines the difficulty of the relation task.

Out of the two types of relation to be found, this simple approach yielded better results for the *Localization* relation (F=0.07) than for the *PartOf* relation (F=0.02). While our results are probably too bad to yield a definite conclusion, the results of other participant also reflect a difference in performance for relation *Localization* and *PartOf.*

Improvements. After fixing the technical problems we encountered, we plan to test other algorithms such as SVM, which may be more adapted for this kind of task.

6 Additional experiments

After the official submission, we carried out additional experiments.

6.1 Habitat entities normalization

6.1.1 Beyond exact match

The improvements we made on the *habitat* entities normalization are only based on the mapping between the predicted concept and the ontology. In our official submission, we only used an exact match. We tried to produce a more flexible mapping in several ways.

First, we tried to normalize the mention gathering all words from the mention into a single word. Indeed, the concept *"rain forest"* is not found in the ontology while the concept *"rainforest"* in one word exists.

Second, we split the mention into single words and tried matching based on the features listed below, in order to manage the subsumption of concepts.

- all words except the first one: *"savannah"* instead of *"brazilian savannah"*,
- all words except the last one: *"glossina"* instead of *"glossina brevipalpis"*,
- the last three words *(we did not find example in the corpus)*,
- the first three words: *"sugar cane fields"* instead of *"sugar cane fields freshly planted with healthy plants"*,
- the last two words: *"tsetse fly"* instead of *"blood-sucking tsetse fly"*,
- and the first two words: *"tuberculoid granulomas"* instead of *"tuberculoid granulomas with caseous lesions"*.

If two parts of a mention can be mapped to two concepts in the ontology, we added both concepts in the output.

We also extended the coverage of the ontology using the reference normalization from both training and development corpora, adding 316 entries in the ontology. Those new concepts can be considered either as synonyms or as hyponyms:

- synonyms: *"root zone"* is a synonym of *"rhizosphere"*. While only the second one occurs in the ontology, we added the first concept with the identifier from the second concept;
- hyponyms: *"bacteriologist"* and *"entomologist"* are both hyponyms of *"researcher"*. We gave the hypernym identifier to the hyponym concepts.

[7] We gave the MBTO:00000001 identifier which is the id for the concept *"gaz seep"*.

At last, if no concept was found in the ontology, instead of using the identifier of the first concept in the ontology, we gave as a default identifier the one of the more frequent concept in the corpora.[8] This strategy improves system performance.

6.1.2 Results

The improvements we made allowed us to achieved better results on the test corpus (table 9). While on the official submission we achieved a 0.66 Slot Error Rate, we obtained a 0.53 SER thanks to the improvements we made. This new result does not lead us to obtain a better rank, but it is closer to the ones the other participants achieved (from 0.49 to 0.46).

Category	Official Evaluation	Additional Experiments
Substitution	187.66	121.99
Insertion	2	2
Deletion	144	144
Matches	175.34	241.01
Predicted	283	283
SER	**0.66**	**0.53**
Recall	0.35	0.48
Precision	0.62	0.85
F-measure	0.44	0.61

Table 9: Results on sub-task 1 on both the official submission and the additional experiments

These improvements led us to obtain better recall, precision and F-measure. While our recall is still the lowest of all participants (0.48 vs. [0.60;0.72]), our precision is the highest (0.85 vs. [0.48;0.61]) and our F-measure is equal to the highest one (0.61 vs. [0.57;0.61]).

6.2 Relationships between entities

6.2.1 Processing

On the relationships, as a first step, we fixed the problems that prevented us to use the MaxEnt model during the submission stage: (i) we produced correct files for the algorithm, removing the *geographical* entities from our processing accordingly with the guidelines, (ii) when dealing with all possible combinations of entities that can be linked together, we managed the relations so as not to produce those relations in duplicate, and (iii) we better managed the confidence score given by the CRF on each relation.

6.2.2 Results

We produced new models on the training corpus based on the following features: entities to be linked, category of each entity, and whether a relation between those entities exists in the training corpus. We performed two evaluations of those models: (i) on the development corpus, using the official evaluation script, and (ii) on the test corpus via the evaluation server.[9] As presented in Table 10, we achieved worse results (F=0.02 and F=0.03) than our official submission (F=0.06) on the test corpus.

#		Sub-task 2		Sub-task 3
		Dev	Test	Test
1	R	0.18	0.11	0.06
	P	0.49	0.01	0.01
	F	0.26	**0.02**	0.01
2	R	0.58	0.02	0.02
	P	0.77	0.16	0.33
	F	0.66	**0.03**	0.04

Table 10: Results on sub-tasks 2 and 3 based on the additional experiments (#1 and #2)

We also noticed that we achieved very poor results on the test corpus while the evaluation on the development corpus provided promising results, with a F-measure decreasing from 0.26 to 0.02 on the first experiment, and from 0.66 to 0.04 on the second one. The difference between the results from both development and test corpora is hard to understand. We have to perform additional analyses on the outputs we produced to identify the problem that occurred.

Moreover, we plan to use more contextual features *(specific words that indicate the relation, distance between two entities, presence of relative pronouns, etc.)* to improve the model. Indeed, in relations between concepts, not only the concepts must be studied but also the context in which they occur as well as the linguistic features used in the neighborhood of those concepts.

[8] The concept *"human"* with identifier MBTO:00001402 is the more frequent concept in all corpora while the concept *"gaz seep"* with identifier MBTO:00000001 was never used.

[9] The reference annotations from the test corpus will not be released to the participants. Instead of those relations, an evaluation server has been opened after the official evaluation took place.

7 Conclusion

In this paper, we presented the methods we used as first time participant to the BioNLP Bacteria Biotopes Shared Task.

To detect bacteria and biotopes names, we used a machine-learning approach based on CRFs. We used several resources to build the model, among them the NCBI taxonomy, the OntoBiotope ontology, the Cocoa annotations, and unsupervised clusters created through Brown's algorithm. The normalization of the *habitat* names with the concepts in the OntoBiotope ontology was performed with a Perl script based on exact match of the entity to be found, taking into account its plural form. On this sub-task, we achieved a 0.66 Slot Error Rate.

In order to process the relationships between entities, our MaxEnt model was not ready for the official submission. The simple approach we used relies on the identification of the relation between entities only if the relation exists in the training corpus. This simple approach is not robust enough to correctly process new data. On the relation subtasks, due to the approach we used, we achieved a 0.06 F-measure.

On the first sub-task, we enhanced our habitat entities normalization process, which led us to improve our Slot Error Rate from 0.66 *(official submission)* to 0.53 *(additional experiments)*.

On the relation detection, first, we plan to make new tests with more features, including contextual features. Second, we plan to test new algorithms, such as SVM which seems to be relevant to process relationships between entities.

Acknowledgments

This work has been done as part of the Quaero program, funded by Oseo, French State Agency for Innovation. I would like to thank the organizers for their work and Aurélie Névéol for the proofread of this paper.

References

Adam L Berger, Stephen Della Pietra, and Vincent J Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 shared task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Marteen van de Guchte, Philippe Bessières, and Claire Nédellec. 2012a. BioNLP shared task – the bacteria track. *BMC Bioinformatics*, 13(Suppl 11):S3.

Robert Bossy, Claire Nédellec, and Julien Jourde, 2012b. *Bacteria Biotope (BB) task at BioNLP Shared Task 2013. Task proposal.* INRA, Jouy-en-Josas, France.

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.

Scott Federhen. 2012. The NCBI taxonomy database. *Nucleic Acids Res*, 40(Database issue):D136–43.

Silviu Guiasu and Abe Shenitzer. 1985. The principle of maximum entropy. *The Mathematical Intelligence*, 7(1).

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Juni'chi Tsujii. 2011. Overview of BioNLP shared task 2011. In *BioNLP Shared Task 2011 Workshop Proc*, pages 1–6, Portland, OR. ACL.

BalaKrishna Kolluru, Sirintra Nakjang, Robert P Hirt, Anil Wipat, and Sophia Ananiadou. 2011. Automatic extraction of microorganisms and their habitats from free text using text mining workflows. *J Integr Bioinform*, 8(2):184.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. *Proc of ACL*, pages 504–13, July.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, MIT.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, pages 249–52.

Nhung T. H. Nguyen and Yoshimasa Tsuruoka. 2011. Extracting bacteria biotopes with semi-supervised named entity recognition and coreference resolution. In *BioNLP Shared Task 2011 Workshop Proc*, pages 94–101, Portland, OR. ACL.

Zorana Ratkovic, Wiktoria Golik, and Pierre Warnier. 2012. Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics*, 13(Suppl 11):S8.

Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared-task: Chunking. In *Proc of CoNLL-2000 and LLL-2000*, pages 127–32, Lisbon, Portugal.