Bacteria Biotope Detection, Ontology-based Normalization, and Relation Extraction using Syntactic Rules

İlknur Karadeniz
Department of Computer Engineering
Boğaziçi University
34342, Bebek, İstanbul, Turkey
ilknur.karadeniz@boun.edu.tr

Arzucan Özgür
Department of Computer Engineering
Boğaziçi University
34342, Bebek, İstanbul, Turkey
arzucan.ozgur@boun.edu.tr

Abstract

The absence of a comprehensive database of locations where bacteria live is an important obstacle for biologists to understand and study the interactions between bacteria and their habitats. This paper reports the results to a challenge, set forth by the Bacteria Biotopes Task of the BioNLP Shared Task 2013. Two systems are explained: Sub-task 1 system for identifying habitat mentions in unstructured biomedical text and normalizing them through the OntoBiotope ontology and Sub-task 2 system for extracting localization and part-of relations between bacteria and habitats. Both approaches rely on syntactic rules designed by considering the shallow linguistic analysis of the text. Sub-task 2 system also makes use of discourse-based rules. The two systems achieve promising results on the shared task test data set.

1 Introduction

As the number of publications in the biomedical domain continues to increase rapidly, information retrieval systems which extract valuable information from these publications have become more important for scientists to access and utilize the knowledge contained in them.

Most previous tasks on biomedical information extraction focus on identifying interactions and events among bio-molecules (Krallinger et al., 2008; Kim et al., 2009). The Bacteria Biotope Task (Bossy et al., 2011; Bossy et al., 2012) is one of the new challenges in this domain, which was firstly presented in the BioNLP 2011 Shared Task. The main goals of the Bacteria Biotope Task were to extract bacteria locations, categorize them into one of the eight types *(Environment, Host, Host-Part, Geographical, Water, Food, Medical, Soil)*, and detect Localization and PartOf events between bacteria and habitats. Automatically extracting this information from textual sources is crucial for creating a comprehensive database of bacteria and habitat relations. Such a resource would be of great value for research studies and applications in several fields such as microbiology, health sciences, and food processing.

Three teams participated in the Bacteria Biotope Task using different methodologies (Bossy et al., 2011; Bossy et al., 2012). Bibliome INRA (Ratkovic et al., 2012), which achieved the best F-score (45%) among these teams, implemented a system which used both linguistic features and reasoning over an ontology to predict location boundaries and types. Bibliome also utilized some resources such as NCBI Taxonomy[1], list of Agrovoc geographical names[2], and an in-house developed ontology for specific location types. UTurku (Björne et al., 2012), presented a machine-learning based system which can be used to find solutions for all main tasks with a few alteration in the system. UTurku used this generic system with additional named entity recognition patterns and external resources, whereas JAIST (Nguyen and Tsuruoka, 2011) used CRFs in order to recognize entities and their types.

UTurku and JAIST treated event extraction as a classification problem by using machine learning approaches, while Bibliome created and utilized a trigger-word list. Bibliome tried to find events by checking if a trigger-word and entities co-occur in the scope of the same sentence. Bibliome was the only team that considered coreference resolution. Not considering coreference resolution deteriorated the performance of JAIST's system less than that of UTurku's system, since JAIST's system operated in the scope of a paragraph, while UTurku's system operated in the scope of a sen-

[1] http://www.ncbi.nlm.nih.gov/Taxonomy/

[2] http://aims.fao.org/standards/agrovoc/about

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 170–177,
Sofia, Bulgaria, August 9 2013. ©2013 Association for Computational Linguistics

tence.

The Bacteria Biotope Task (BB) in the BioNLP 2013 Shared Task (Bossy et al., 2013) gives another opportunity to scientists to develop and compare their systems on a reliable platform. This task contains three subtasks. For **Sub-task 1**, participants are expected to detect the names and positions of habitat entities, as well as to normalize these habitats through the OntoBiotope (MBTO) Ontology concepts. For **Sub-task 2**, when the names, types, and positions of the entities (*bacteria, habitat, geographical*) are given, participants are expected to extract relations which can be either between bacteria and habitat pairs (Localization event) or between host and host part pairs (PartOf event). **Sub-task 3** is the same as Sub-task 2, except that the gold standard entities are not provided to the participants.

In this paper, we present two systems, one for Sub-task 1 (Entity Detection and Categorization) and one for Sub-task 2 (Localization Relation Extraction) of the Bacteria Biotope Task in the BioNLP 2013 Shared Task. Both systems are rule-based and utilize the shallow syntactic analysis of the documents. The Sub-task 2 system also makes use of the discourse of the documents. The technical details of our systems are explained in the following sections.

2 Data Set

The corpus provided by the organizers was created by collecting documents from many different web sites, which contain general information about bacteria and habitats. The data set, consisting of 52 training, 26 development, and 26 test documents, was annotated by the bioinformaticians of the Bibliome team of MIG Laboratory at the Institut National de Recherche Agronomique (INRA).

For the training and development phases of Sub-task 1, document texts with manually annotated habitat entities and the concepts assigned to them through the OntoBiotope ontology were provided, while in the test phase, only the unannotated document texts were given by the task organizers. The OntoBiotope ontology which contains 1,700 concepts organized in a hierarchy of is-a relations was also provided by the organizers for this task.

For the training and development phases of Sub-task 2, document texts with manually annotated bacteria, habitat and geographical entities, as well as the localization and part-of relations were provided, while in the test phase, document texts annotated only for bacteria, habitat and geographical entities were given.

3 Bacteria Biotope Detection and Ontology-based Normalizaton

For Sub-task 1 (Entity Detection and Categorization), we implemented a system which applies syntactic rules to biomedical text after a pre-processing phase, where a given text is split into sentences and parsed using a shallow parser. The workflow of our Sub-task 1 system is shown in Figure 1. Firstly, each input file is split into sentences using the Genia Sentence Splitter (GeniaSS) (Saetre et al., 2007). The outputs of the splitter are given to the Genia Tagger (Tsuruoka et al., 2005; Tsuruoka and Tsujii, 2005) as input files with the aim of obtaining the lemmas, the part-of-speech (POS) tags, and the constituent categories of the words in the given biomedical text *(e.g., surface form: ticks; lemma: tick; POS tag: NNS; phrase structure: I-NP)*. We utilized these syntactic information at the following steps of our system.

In the following subsections, a detailed explanation for the detection of habitat boundaries and their normalization through the OntoBiotope Ontology concepts is provided.

3.1 Entity Boundary Detection

Entity boundary detection, which is the first step of Sub-task 1, includes automatic extraction of habitat entities from a given natural language text, and detection of the entity boundaries precisely. In other words, the habitat boundaries that are retrieved from the texts should not include any unnecessary and non-informative words. In order to achieve this goal, we assume that bacteria habitats are embedded in text as noun phrases, and all noun phrases are possible candidates for habitat entities. Based on this assumption, our system follows the steps that are explained below by using the modules that are shown in Figure 1.

As explained before, the **Sentence Splitter**, **POS Tagger**, and **Shallow Parser** are the modules that are utilized in the pre-processing phase.

The **Noun Phrase Extractor & Simplifier** module firstly detects the noun phrases in the text by using the Genia Tagger and then post-processes these noun phrases by using some syn-

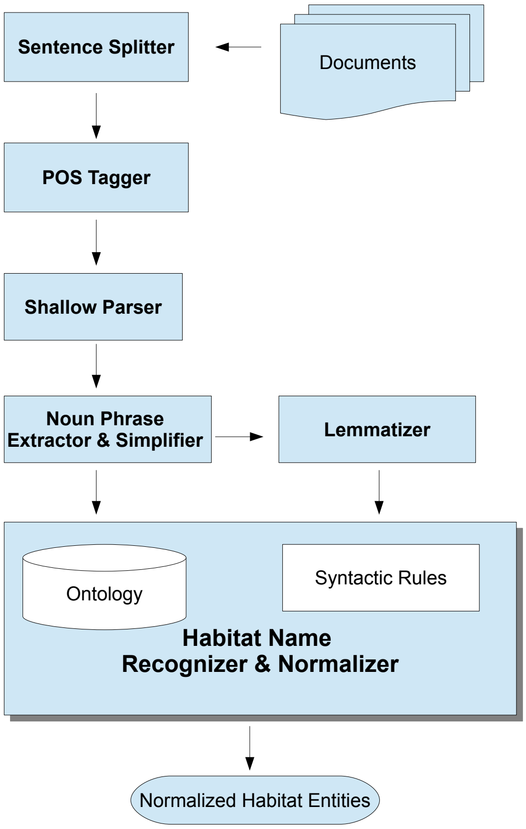
Sentence Splitter
Documents
POS Tagger
Shallow Parser
Noun Phrase Extractor & Simplifier
Lemmatizer
Ontology
Syntactic Rules
Habitat Name Recognizer & Normalizer
Normalized Habitat Entities

Figure 1: Workflow of the Sub-task 1 System

tactic rules. The functions of this module include the removal of some unnecessary words from the noun phrases, which are not informative for environmental locations of bacteria. To distinguish informative words from non-informative ones, our system utilizes the POS Tags of each word that compose the noun phrases in question. For example, words that have determiners or possessive pronouns as their POS Tags should not be included to the boundaries of the candidate habitat entities. For example, *the* in the noun phrase "*the soybean plant Glycine max*" and *its* in the noun phrase "*its infectious saliva*" are eliminated from the candidate noun phrases, restricting the habitat boundary, and creating new candidate noun phrases.

The Noun Phrase Extractor & Simplifier module also includes a mechanism to handle noun phrases that contain the conjunction *"and"*. First, such noun phrases are separated from the conjunction *"and"* into two sub-phrases. Next, each sub-phrase is searched in the OntoBiotope ontology. If the ontology entries matched for the two sub-phrases have the same direct ancestor (i.e., the two ontology entries have a common is-a relation), then the noun phrase consisting of the two sub-phrases connected with the conjunction *"and"* is identified as a single habitat entity. On the other hand, if the ontology entries matched for the two sub-phrases don't have a common direct ancestor, then each sub-phrase is identified as a separate habitat entity. For example, each of the entity boundaries of the phrases *"nasal and oral cavity"* , *"fresh and salt water"*, and *"human and sheep"* are handled differently from each other as described below.

- For the first phrase, *"nasal"* is the first sub-phrase and *"oral cavity"* is the second sub-phrase. The direct ancestor (i.e., the first level is-a concept) of the first sub-phrase *"nasal"* is *"respiratory tract part"* and that of the second sub-phrase *"oral cavity"* is *"buccal"*. Since *"respiratory tract part"* and *"buccal"* is-a concepts are not the same, *"nasal cavity"* and *"oral cavity"* are generated as two separate habitats. In other words, if there is not a direct common *is-a* concept between the matching terms for the sub-phrases in the OntoBiotope ontology, then one habitat entity *"nasal cavity"* is generated from the noun phrase by adding the second part of the second sub-phrase *"cavity"* to the first sub-phrase *"nasal"* and another entity is generated by taking the second sub-phrase as a whole *"oral cavity"*.

- For the second sample phrase, *"fresh"* is the first sub-phrase and *"salt water"* is the second sub-phrase. The first sub-phrase *"fresh"* matches with an ontology entry whose direct ancestor is *"environmental water with chemical property"* and the second sub-phrase *"salt water"* matches with an ontology entry that has two different direct ancestors *"environmental water with chemical property"* and *"saline water"*. Since *"environmental water with chemical property"* is a common ancestor for both sub-phrases in the ontology, a single habitat entity *"fresh and salt water"* is generated. In other words, if there is a direct common ancestor between the matching terms for the sub-phrases in the OntoBiotope ontology, then only one habitat entity that is composed of the whole noun phrase is generated.

- For the third phrase, *"human"* is the first sub-phrase and *"sheep"* is the second sub-phrase. In this case, two separate habitat entities *'human"* and *"sheep"* are generated directly from the two sub-phrases since they don't have a common ancestor in the ontology.

At the end of these phases, purified sub-noun phrases, which are habitat entity candidates whose boundaries are roughly determined by the deletion of non-informative modifiers from noun phrases, are obtained.

To determine whether a candidate noun phrase is a habitat entity or not, the **Habitat Name Recognizer & Normalizer** module searches all ontology entries, which compose the OntoBiotope Ontology, to find an exact match with the candidate noun phrase or with parts of it. In this step, the names, exact synonyms, and related synonyms of ontology entries (ontology entry features) are compared with the candidate noun phrase.

[Term]		
id:	MBTO:00001828	
name:	digestive tract	
related_synonym:	"gastrointestinal tract"	[TyDI:23802]
exact_synonym:	"GI tract"	[TyDI:23803]
related_synonym:	"intestinal region"	[TyDI:23805]
related_synonym:	"gastrointestinal"	[TyDI:23806]
exact_synonym:	"GIT"	[TyDI:23807]
related_synonym:	"alimentary canal"	[TyDI:24621]
is_a:	MBTO:00000797	! organ

Table 1: First ontology entity match for *human gastrointestinal tract.*

For example, if our candidate noun phrase is *"the human gastrointestinal tract"*, after the post-processing phase, the purified candidate phrase will be *"human gastrointestinal tract"*. When the search step for this simplified candidate entity is handled, two different ontology entries are returned by our system as matches (see Table 1 for the first ontology entry match and Table 2 for the second one). These two ontology entries are returned as results by our system because the first one contains the *related_synonym: "gastrointestinal tract"* and the second one contains the *name: human*. Since the system returns matches for the candidate noun phrase *"human gastrointestinal tract"*, it is verified that one or more habitat entities can be extracted from this phrase.

[Term]		
id:	MBTO:00001402	
name:	human	
related_synonym:	"person"	[TyDI:25453]
related_synonym:	"individual"	[TyDI:25454]
exact_synonym:	"subject"	[TyDI:25374]
exact_synonym:	"homo sapiens"	[TyDI:26681]
related_synonym:	"people"	[TyDI:25455]
is_a:	MBTO:00001514	! mammalian

Table 2: Second ontology entity match for *human gastrointestinal tract.*

To detect the exact habitat boundaries, manually developed syntactic rules are utilized in addition to the ontology entry matching algorithm, which is used for entity verification of a candidate phrase. Our system determines the boundaries according to the following syntactic rules:

- If an ontology entry matches exactly with the noun phrase, take the boundaries of the noun phrase as the boundaries of the habitat, and use the whole phrase to create a new habitat entity.
- If an ontology entry matches beginning from the first word of the noun phrase, but does not match totally, take the boundaries of the matched parts of the phrase, and create a new habitat entity using the partial phrase.
- If an ontology entry matches beginning from an internal word of the noun phrase, take the boundaries of the noun phrase as the boundaries of the habitat, and use the whole phrase to create a new habitat entity. For example, in Table 1, the match of the noun phrase *"human gastrointestinal tract"* with the *related_synonym: "gastrointestinal tract"* generates *"human gastrointestinal tract"* as a habitat entity.

In many cases habitat entity names occur in different inflected forms in text. For example, the habitat name *"human"*, can occur in text in its plural form as *"humans"*. We used the **Lemmatizer** module in order to be able to match the different inflected forms of habitat names occurring in text against the corresponding entires in the OntoBiotope ontology. This module applies the rules described above to the lemmatized forms of the candidate noun phrases, which are obtained using the Genia Tagger.

After running the same algorithm also for lemmatized forms of the noun phrase, a merging algorithm is used for the matching results of the surface and lemmatized forms of the noun phrases in order to create an output file, which contains the predicted habitat entities and their positions in the input text.

3.2 Ontology Categorization

For Sub-task 1, detection of the entities and their boundaries is not sufficient. In order to obtain normalized entity names, participants are also expected to assign at least one ontology concept from the OntoBiotope Ontology to all habitat entities, which are automatically extracted by their systems from the input text.

While our system detects entities and their boundaries (as explained in detail in Section 3.1), it also assigns ontology concepts to the retrieved entities. All assigned concepts are referenced by the MBTO-IDs of the matched ontology entries (e.g, *MBTO:00001402* for *human* and *MBTO:00001828* for *human gastrointestinal tract*) (see Table 3).

4 Event Extraction

For Sub-task 2 (Localization Event Extraction Task), we used different methods according to the relation type that we are trying to extract. The workflow of our system is shown in Figure 2. The details of our approach are explained in the following sub-sections.

4.1 Localization Event Extraction

In order to extract localization relations, we assume that discourse changes with the beginning of a new paragraph. Our system firstly splits the input text into paragraphs. Next, the entities (bacteria and habitats) that occur in the given paragraph are identified. We assume that the paragraph is about the bacterium whose name occurs first in the paragraph. Therefore, we assign all the habitat entities to that bacterium. If the name of this bacterium occurs in previous paragraphs as well, then the boundary of the bacterium entity is set to its first occurrence in the document.

We also have a special case for boundary determination of bacteria in the localization relation. If a bacterium name contains the word *"strain"* , we assign the first occurrence of its name without the word *"strain" (e.g, Bifidobacterium longum NCC2705* instead of *Bifidobacterium longum strain NCC2705).*

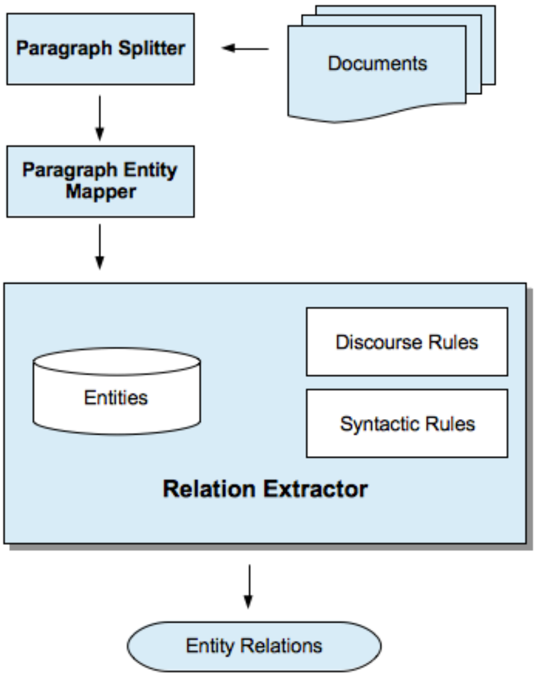

Figure 2: Workflow of the Sub-task 2 System

4.2 PartOf Event Extraction

In order to detect partOf relations between hosts and host parts in a given biomedical text, we assumed that such relations can only exist if the host and the host part entities occur in the same paragraph. Based on this assumption, we propose that if a habitat name is a subunit of the term which identifies another habitat that passes in the same discourse, then they are likely to be related through a partOf relation. In other words, if one habitat contains the other one, and obeys some syntactic rules, then there is a relation. For example, *"respiratory track of animals"* is a habitat and *"animals"* is another habitat, both of which are in the same paragraph. Since the *"respiratory track of animals"* phrase contains the *"animals"* phrase and the word *"of"*, and the *"animals"* phrase is on the right hand side of the *"respiratory track of animals"* phrase, our system detects a partOf relation between them.

5 Evaluation

The official evaluation results on the test set are provided using different criteria for the two subtasks by the task organizers[3].

[3] http://2013.bionlp-st.org/tasks/bacteria-biotopes/test-results

EntityID	**Boundary**	**Entity**
T1 Habitat	113 118	human
T2 Habitat	113 141	human gastrointestinal tract
ID	**EntityID**	**Reference**
N1	OntoBiotope Annotation:T1	Referent:MBTO:00001402
N2	OntoBiotope Annotation:T2	Referent:MBTO:00001828

Table 3: Detected entities and boundaries from the *human gastrointestinal tract* noun phrase

For Sub-task 1, submissions are evaluated considering the Slot Error Rate *(SER)*, which depends on the number of substitutions *S*, deletions *D*, insertions *I*, and *N*. *N* is the number of habitats in the reference, while *D* and *I* are the number of reference and predicted entities that could not be paired, respectively.

$$SER = \frac{S + D + I}{N} \tag{1}$$

The number of substitutions *S* is calculated by using Equation 2. Here *J* is the Jaccard index between the reference and the predicted entity, which measures the accuracy of the boundaries of the predicted entity (Bossy et al., 2012). *W* is a parameter that defines the semantic similarity between the ontology concepts related to the reference entity and to the predicted entity (Wang et al., 2007). This similarity is based on the is-a relationships between concepts, and used for penalizing ancestor/descendent predictions more compared to sibling predictions as it approaches to 1.

$$S = J \cdot W \tag{2}$$

For Sub-task 2, precision, recall, and f-score metrics are used for evaluation. In the following subsections, our official evaluation results for Sub-task 1 and Sub-task 2 are given.

5.1 Results of Sub-task 1

Our official evaluation results on test set are shown in Table 4. Our system ranked second according to the *SER* value among four participating systems in the shared task.

The official results of our system on the test set for entity boundary detection are shown in Table 5. Our system obtained the smallest *SER* value for detecting the entity boundaries (i.e., the best performance) among the other participating systems.

Main Results	
S	112.70
I	43
D	89
M	305.30
P	520
SER	0.48
Recall	0.60
Precision	0.59
F1	0.59

Table 4: Main results on test set for Sub-task 1*(Entity Boundary Detection & Ontology Categorization)*

Entity Boundary Evaluation	
S	82.71
M	335.29
SER	0.42
Recall	0.66
Precision	0.64
F1	0.65

Table 5: Entity boundary detection results on the test set for Sub-task 1

Our ontology categorization evaluation results on the test set, which do not take into account the entities' boundaries are shown in Table 6. Our system ranked second on the main evaluation where the parameter *w* (described in Section 5) was set to *0.65*. As shown in the table, as the *w* value increases, our results get better. According to the official results, our system ranked first for *w = 1* with the highest f-score, and our *SER* result is same as the best system for *w = 0.8*.

The parameter *w* can can be seen as a penalization value for the false concept references. As *w* increases, the false references to distant ancestors and descendants of the true reference concepts are penalized more, whereas as *w* decreases the false references to the siblings are penalized more severely.

The results also show that our system is able to achieve balanced precision and recall values. In other words, the recall and precision values are close to each other.

w	**S**	**M**	**SER**	**Recall**	**Precision**	**F**
1	38.64	379.36	0.34	0.75	0.73	0.74
0.8	44.90	373.10	0.35	0.74	0.72	0.73
0.65	50.95	367.05	0.36	0.72	0.71	0.71
0.1	70.78	347.22	0.40	0.68	0.67	0.68

Table 6: Ontology Categorization results for Sub-task 1 on the test set

5.2 Results of Sub-task 2

The precision, recall, and f-measure metrics are used to evaluate the Sub-task 2 results on the test set. Our main evaluation results, which consider detection of both *Localization* and *PartOf* event relations for Sub-task 2 are shown in the first row of Table 7, whereas our results that are calculated for the two event types separately are shown in the *Localization* and *PartOf* rows of the table. According to the official results, our system ranked third for detecting all event types. On the other hand, it achieved the best results for detecting the *PartOf* events.

	Recall	**Precision**	**F**
All	0.21	0.38	0.27
Localization	0.23	0.38	0.29
PartOf	0.15	0.40	0.22

Table 7: Main results on test set for Sub-task 2

6 Conclusion

In this study, we presented two systems that are implemented in the scope of the BioNLP Shared Task 2013 - Bacteria Biotope Task. The aim of the Sub-task 1 system is the identification of habitat mentions in unstructured biomedical text and their normalization through the OntoBiotope ontology, whereas the goal of the Sub-task 2 system is the extraction of localization and part-of relations between bacteria and habitats when the entities are given. Both systems are based on syntactic rules designed by considering the shallow syntactic analysis of the text, while the Sub-task 2 system also makes use of discourse-based rules.

According to the official evaluation, both of our systems achieved promising results on the shared task test data set. Based on the main evaluation where the parameter *w* is set to *0.65*, our Sub-task 1 system ranked second among four participating systems and it ranked first for predicting the entity boundaries when ontology categorization outputs are not considered. The results show that our system performs better as w increases and achieves the best performance when *w = 1* and *w = 0.8*. Our Sub-task 2 system achieved encouraging results by ranking first in predicting the *PartOf* events, and ranking third when all event types are considered.

The proposed systems can be enhanced by incorporating a stemming module and including more syntax and discourse based rules.

Acknowledgments

This work has been supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme.

References

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, Philippe Bessières, Maarten van de Guchte, and Claire Nédellec. 2011. Bionlp shared task 2011: bacteria biotope. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 56–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Bossy, Julien Jourde, Alain P. Manine, Philippe Veber, Erick Alphonse, Maarten van de Guchte, Philippe Bessieres, and Claire Nedellec. 2012. BioNLP Shared Task - The Bacteria Track. *BMC Bioinformatics*, 13(Suppl 11):S3+.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessir̀es, and Claire Nédellec. 2013. BioNLP shared task 2013 - an overview of the bacteria biotope task. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, AUG. Association for Computational Linguistics.

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

Martin Krallinger, Florian Leitner, Carlos Rodriguez-penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome Biology*, pages 2–4.

Nhung T. H. Nguyen and Yoshimasa Tsuruoka. 2011. Extracting bacteria biotopes with semi-supervised named entity recognition and coreference resolution. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 94–101, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Rune Saetre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. AKANE System: Protein-Protein Interaction Pairs in the BioCreAtIvE2 Challenge, PPI-IPS subtask. In Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, editors, *Proceedings of the Second BioCreative Challenge Workshop*.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 467–474. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In Panayiotis Bozanis and Elias N. Houstis, editors, *Advances in Informatics*, volume 3746, chapter 36, pages 382–392. Springer Berlin Heidelberg.

J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, May.