BioNLP 2011 Task Bacteria Biotope – The Alvis system

Zorana Ratkovic[1,2] Wiktoria Golik[1] Pierre Warnier[1] Philippe Veber[1] Claire Nédellec[1]

1 MIG INRA UR1077, Domaine de Vilvert
F-850 Jouy-en-Josas, France
forename.name@jouy.inra.fr

2 LaTTiCe UMR 8094 CNRS Univ. Paris 3
1 rue Maurice Arnoux
F-92120 MONTROUGE

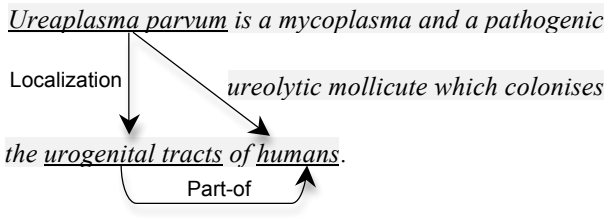
Abstract

This paper describes the system of the INRA Bibliome research group applied to the Bacteria Biotope (BB) task of the BioNLP 2011 shared tasks. Bacteria, geographical locations and host entities were processed by a pattern-based approach and domain lexical resources. For the extraction of environment locations, we propose a framework based on semantic analysis supported by an ontology of the biotope domain. Domain-specific rules were developed for dealing with Bacteria anaphora. Official results show that our Alvis system achieves the best performance of participating systems.

1 Introduction

Given a set of Web pages, the information extraction goal of the Bacteria Biotope (BB) task is to precisely identify bacteria and their locations and to relate them. The type of the predicted locations has to be selected among eight types. Among them the host and host-part locations have to be related by the part-of relation. Three teams participated in the challenge.

BB task example



One of the specificities of the BB task is that the bacteria location vocabulary is very large and various as opposed to protein subcellular locations in biology challenges (Kim et al., 2010) and geographical locations (Zhou et al., 2005). Locations include natural environments and hosts as well as food and medical locations. In order to deal with this heterogeneity, we propose a framework based on a term analysis of the test corpus and a shallow mapping of these terms to a bacteria biotope (BB) termino-ontology. This mapping derives the type of location terms and filters out non-location terms. Large external dictionaries of host names (*i.e.* NCBI taxonomy) and geographical names (*i.e.* Agrovoc thesaurus) complete the lexical resources.

The high frequency of bacteria anaphora and ambiguous antecedent candidates in the corpus was also a difficulty. Our Alvis system implements an anaphora resolution algorithm that takes into consideration the anaphoric distance and the position of the antecedent in the sentence. Alvis predicts the bacteria names and their relation to the locations with the help of hand-made patterns based on linguistic analysis and lexical resources.

The methods for predicting and typing locations (section 2) and bacteria (section 3) are first described. Section 4 details the method for relating them. Section 5 comments the experimental results.

2 Location

Our system handles separately the recognition of host and geographical names by dictionary mappings, while the recognition of locations of the environment and host part types is based on linguistic analysis and ontology inference.

Host names and geographical names appeared to be easier to predict by using a named-entity recognition strategy than the other types of location. They are less subject to variation than environmental locations, which can include any physical feature. For host name extraction, we used the NCBI taxonomy as the major source. Only the eukaryote subtree was considered for host detection.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 102–111,
Portland, Oregon, USA, 24 June, 2011.

Our system filters out the ambiguous names such as Indicator (honeyguides) or Dialysis (xylophage insect) by comparing them to a list of common words in English. The host name list was enriched with additional common names including non-taxonomic host groups (*e.g.* herbivores), progeny names (*e.g.* calf) and human categories (*e.g.* patient). The resulting host name list contains more than 1,800,000 scientific names and 60,000 common names. The geographical name recognition component uses a small dictionary of all geographic terms from the Agrovoc thesaurus sub-vocabularies. At first, we considered using the very rich resource GeoNames. However, it contains too many ambiguous names to be directly usable by short-term development.

2.1 Location of Environment type

The identification of environment locations is done in two steps. First, the automatic extraction of all candidate terms from the test corpus, then the assignment of a location type to these terms with the help of the Bacteria Biotope (BB) termino-ontology. The type assigned to a given term is the type of the closest concept label in the ontology. Since the BB termino-ontology was originally not structured according to the eight types, in order to be usable it first had to be enriched by the new concepts and then mapped to this topology.

Corpus term extraction. The corpus terms were automatically extracted by the AlvisNLP/ML pipeline (Nedellec et al., 2008) with BioYatea (Nedellec et al., 2010). BioYatea is the version of Yatea (Hamon & Aubin, 2006) adapted to the biology domain. We modified BioYatea setting according to the training dataset study. We observed that most of the location terms in the training dataset are noun phrases with adjective modifiers (*e.g. rodent nests*) while prepositional phrases are rather rare (*e.g. breaks in the skin*). We set the term boundaries of BioYatea to include all prepositions except the *of* preposition. Considering other prepositions such as *with* may yield syntactic attachment errors, thus we prefer the risk of incomplete terms to incorrect prepositional attachments.

Bacteria Biotope ontology. We used the Bacteria Biotope (BB) termino-ontology for typing the extracted terms. It is under development for the study of bacteria phenotypes and habitats. The high level of the habitat part is structured in a manner similar to that proposed by the one level classification by Floyd (Floyd et al., 2005). It has a fine-grained structure with the same goal as the generalist EnvO habitat ontology (Field et al., 2008), but it focuses on bacteria phenotype and biotope modeling. It includes a terminological level that records lexical forms of the concepts including terms, synonyms and variations.

For the purpose of the challenge, the initial ontology was manually completed using location concepts. The training corpus, as well as the habitat and isolation site fields of the GOLD database on sequenced prokaryotes (Liolios et al., 2009) are the main sources of location terms and synonyms. The analysis of the training corpus mainly led to the addition of adjectival forms of host parts (*e.g. lymphatic, intracellular*) and human references (*e.g. patient, infant, progeny*).

The GOLD database isolation site field is a very rich source of bacteria location terms. It is filled by natural language descriptions of matters, natural habitats, hosts and geographical locations. For instance, the isolation site of *Anoxybacillus flavithermus* bacterium is *waste water drain at the Wairakei geothermal power station in New Zealand*. The term analysis of GOLD isolation site entries yielded 3,415 location terms including 1,050 geographical names. Hundreds of these terms were manually added to the BB termino-ontology. The lack of time as well as the full sentence structure of the GOLD resource prevented us from correctly handling them in a fully automatic way. We are currently developing a method for the automatic alignment of the terms extracted from GOLD to the BB termino-ontology. Additionally, the GOLD habitat field provided around a hundred different terms that have been directly integrated into the BB termino-ontology.

The current version of the habitat subpart of the BB termino-ontology contains 1,247 concepts and 266 synonyms.

Location types in Bacteria Biotope ontology. The BB termino-ontology has been developed previous to the BB task and the structure of its habitat subpart does not reflect the eight location types of the task. In order to reuse the ontology for the BB task, we assigned types to each location concept. We manually associated the high level nodes of the location hierarchies to the eight BB task types. The types of the lower level concepts were then automatically inferred. For instance, the concept *aquatic environment* is tagged Water in the ontol-

ogy and all of its descendants *lake*, *sea*, *ocean* are of type Water as well. Local type exceptions were manually tagged. For instance, the *waste* tree includes water-carried wastes of type Water and solid industrial residues of type Environment. This way all concepts in the resulting typed ontology were assigned a unique type. The concept types are then propagated to their associated term classes at the terminological level. For instance, *underground water* and its synonym *subterranean water* are both typed as Water. The resulting typed BB termino-ontology is then usable for deriving the types of the terms extracted from the test corpus.

Derivation of location type. The BB termino-ontology scope is too limited for the correct prediction of all candidate term types by Boolean and exact comparison. From the 2,290 candidate terms of the test corpus, only 152 belong as such to the BB termino-ontology. We propose a method based on the head comparison of the candidate and BB terms for the derivation of the candidate term type.

The quality of the ontology-based annotation depends to a large extent on an accurate match between the resource and the terms extracted from the corpus. Our method targets the syntactic structure of terms (candidate and BB terms) in order to gather the most of semantically similar terms. This approach differs from the ontology alignment and population methods that also use the information from the ontology structure in order to infer semantic relationships (*e.g.* hyponyms, meronyms) (Euzenat, 2007). It also differs from semantic annotation supported by context analysis such as distributional semantics (Grefenstette, 1994) or Hearst patterns (Hearst, 1992). It belongs to the class of methods that focus on the morphology of the corpus terms, which use string-based (Levensthein, 1966, Jaro, 1989) or linguistic-based methods (Jacquemin & Tzoukermann, 1999).

Even though the context-based approach should produce very good results, we chose a less time-consuming method that is easier and faster to set up, which is based on morphosyntactic analysis. In our case, string similarity measures turn out to be irrelevant (*laboratory rat* does not mean *rat laboratory*). We observed that in candidate and BB terms, the head is very often the most informative element. Thus, the linguistic-based analysis of terms, in particular the head-similarity analysis (Hamon & Nazarenko, 2001), represents a promising alternative. Our method is inspired by MetaMap (Aronson, 2001). MetaMap tags biomedical corpora with the UMLS Metathesaurus by syntactic analysis that takes into account lexical heads of terms. The similarity scores computed by linguistically-based metrics are higher for terms whose heads have previously been analyzed.

The MetaMap method includes a variant computation that maps acronyms, abbreviations, synonyms as well as derivational, inflectional and spelling variants. Our term typing method is less sophisticated and uses a few lexical variants due to the lack of a complete resource. Some ontology enrichment applications also use head-supported term matching, as in Desmontils (Desmontils et all, 2003). In Desmontils, new concepts belonging to WordNet (Fellbaum, 1998) are automatically added to the ontology in order to improve the indexing process. However, the analysis of the results shows that a great number of concepts found in the texts are not considered because they do not exist in WordNet. Our typing task uses a similar head-based method, but only for type derivation.

Our system derives the location type of candidate terms in several steps. First, if there is a term in the BB termino-ontology that is strictly equal to the candidate term, it is assigned the same type. Then, the other candidate terms are assigned types according to the comparison of their heads to the BB term heads. We assume that in most of the cases the term head conveys the information about the type and is non-ambiguous. A given head *H* is non-ambiguous if all BB terms with head *H* are of the same type. The location term head set is the set of all habitat term heads found in the BB termino-ontology. The current version contains 693 different heads. Let T_e denote the extracted term to be typed. If the head of T_e does not belong to the BB term head set, then the type of T_e is simply *not* Location (e.g. *high metabolic diversity*). If T_e head *does* belong to the BB term head set and the head is non-ambiguous, then T_e is assigned the associated type. For instance, the head of the extracted term *stratified lake* is *lake*. The type of all the BB terms with *lake* head is *Water* (e.g. *meromictic lake*). *Stratified lake* is therefore typed as *Water*.

Specific processing is applied to terms with ambiguous heads. The associative set of BB term heads and types exhibits some cases of ambiguous heads with multiple types that we analyzed in detail. There are two kinds of ambiguities that were

processed in different ways. In the first, multiple types reflect different roles of the same object. In the second, the head is non-informative with respect to the type. In the latter case the type is conveyed by the subterm (term after head removal). We qualify non-informative BB term heads as *neutral*. They mainly denote habitats (*habitat*, *environment*, *medium, zone*) and extracts (*sample*, *surface*, *isolate, material, content*). In this case, the type is derived from the subterm. For instance, the head *isolate* of the extracted term *marine isolate* is neutral. After head removal, it is assigned the type Water since *marine* is of type Water. *Freshwater* has the same type as *freshwater medium* or *freshwater environment* since *medium* and *environment* are neutral heads.

Some heads have more than one type although they denote specific locations. Their multiple types reflect different uses or states. For instance, the head *bottle* has two types: Food and Medical. The type Food is derived from the BB concept *water bottle* and the type Medical is derived from *bedside water bottles* in a hospital environment. The correct type for the extracted terms is then selected by a set of patterns based on the context of the term in the document. For instance, many vegetables and meats could be either of type Host or Food. The type is Host by default. One pattern states that if a term includes or is preceded by a food processing-related word (*e.g. cooked*, *grilled*, *fermented*), then the term is reassigned the type Food. Another pattern states that if a host is preceded by a death-related adjective (*dead*, *decaying*), then its type should be revised as Environment.

Our system currently includes nine disambiguation/retyping patterns. The first version of the type derivation method was automatically applied to the 1,263 GOLD terms after head analysis. Manual examination of the results yielded an extension of the two lists of neutral heads and heads with ambiguous types. There are 20 neutral heads and 21 ambiguous heads in the current version of the BB termino-ontology. The head-matching algorithm appears to be quite productive for the biotope terms. The procedure applied to the test corpus yielded the following figures: BioYatea extracted 2,290 terms. 416 terms matching the post-processing filters were discarded. This includes terms which are too general (i.e. *approach, diversity*), terms containing irrelevant or non desirable adjectives (i.e. *numerous deficiencies, known species*) and terms containing forbidden words according to the annotation location rules (i.e. *bacteria, pathogen, contaminated*, *parasite*). Finally, 1,873 candidate terms were kept.

Among these figures:

- 152 terms belong to the BB termino-ontology
- 90 terms were typed using the ontology heads
- 6 terms with several types were handled by disambiguation patterns.

We plan to extend the list of neutral heads and discriminate adjectives for type disambiguation by machine learning classification applied to the BB termino-ontology modifiers.

Location entity boundary. The analysis of term extraction result from the training corpus shows that the predicted boundaries of locations were not fully consistent with the task annotation guidelines. Post-processing adjusts incorrect boundaries by filtering irrelevant words, packing and merging terms. Irrelevant words (*e.g. contaminated, infected*, *host species*, *disease*, *inflammation*) were removed from the location candidate terms independently of their types (e.g. *contaminated Bachman Road site* vs. *Bachman Road* ; *host plant* vs. *plant*). Note that BioYatea extracts not only the maximum terms (e.g. *contaminated Bachman Road site*), but also their constituents (*Bachman Road site*, *Bachman Road* and *site*). Boundary adjustment often consists in selecting the relevant alternative among the subterms.

Other boundary issues are handled by several patterns, which are applied after the typing stage. These patterns are type-dependent: each pattern only applies to one type or a subset of location types. When necessary, they shift the boundaries in order to include relevant modifiers. They also split location terms or join adjacent location terms. BioYatea may have missed relevant modifiers because of POS-tagging errors. For instance, if a nationality name precedes a location, then it is included (*e.g. German oil field*). Also, it frequently happens that hosts are modifiers of host parts (*e.g. insect gut*). BioYatea extracts the whole term and its constituents. The term is correctly typed as *Host-part* and the host modifier as *Host*. In order to avoid embedded locations, a specific pattern is devoted to the splitting of these terms. In this way *insect gut* (Host-part) becomes *insect* (Host) and *gut* (Host-part).

Most of these patterns involve several specific lexicons, including cardinal directions, relevant and

irrelevant modifiers for each type of location, as well as types, which can be merged and split. The current resources were manually built by examining the location terms of the training set and GOLD isolation fields. The acquisition of relevant and irrelevant modifiers could be automated by machine learning. Some linguistic phenomena could be better handled by the customization of BioYatea. For instance BioYatea considers the preposition *with* as a term boundary so it cannot extract terms containing *with*, like *areas with high sulfur and salt concentrations*.

3 Extraction of Bacteria names

We observed in the training corpus that not only were bacteria names tagged, but also higher level taxa (families) and lower level taxa (strains). We used the NCBI taxonomy as the main bacteria taxon resource since it includes all organism levels and is kept up-to-date. This bacteria dictionary was enriched by taxa from the training corpus, in particular by non standard abbreviations (*e.g. Chl.* = *Chlorobium*, *ssp.* = *subsp*) and plurals, (*Vibrios* as the plural for *Vibrio*) that were hopefully rather rare.

Determining the boundaries of the bacteria names was one of the main issues because corpus strain names do not always follow conventional nomenclature rules. Also, the recognition of bacteria name is evaluated using a strict exact match. Patterns were developed to account for such cases. They handle inversion (*LB400 of Burkholderia xenovorans* instead of *Burkholderia xenovorans LB400*) and parenthesis (*Tropheryma whipplei (the Twist strain)* instead of *Tropheryma whipplei strain* Twist). The corpus also mentions names of bacteria that contain modifiers not found in the NCBI dictionary, such as *antimicrobial-resistant C. coli* or *L. pneumophila serogroup 1*. Such cases, as well as abbreviations (e.g. *GSB* for *green sulfur bacteria*) and partial strain names (*e.g. strain DSMZ 245 T* for *Chlorobium limicola strain DSMZ 245 T*) were also specifically handled.

The main source of error in bacteria name prediction is due to the mixture of family names and strain name abbreviations in the same text. It frequently happens that the strain name is abbreviated into the first word of the name. For instance *Bartonella henselae* is abbreviated as *Bartonella*. Unfortunately, *Bartonella* is a genus mentioned in the same text, thus yielding ambiguities between the anaphora and the family name, which are identical.

3.1 Bacteria anaphora resolution

Anaphors are frequent in the text, especially for bacteria reference and to a smaller extent for host reference. Our effort focused on bacteria anaphora resolution ignoring host anaphora. The extraction method of location relations (section 4) assumes that the relation arguments, location and bacterium (or anaphora of the bacterium) occur in the same sentence. From a total of 2,296 sentences in the training corpus, only 363 sentences contain both the location and the explicit bacterium, while 574 mention only the location. Two thirds of the locations do not co-occur with bacteria. This demonstrates the importance of recovering the bacteria for these cases, which is potentially referred to by an explicit anaphora.

The manual examination of the training corpus showed that the most frequent anaphora of bacteria are not pronouns but higher level taxa, often preceded by a demonstrative determinant, (i.e. *This bacteria, This Clostridium*) and sortal anaphora (i.e. *genus*, *organism*, *species* and *strain*), both of which are commonly found in biological texts (Torri & Vijay-Shanker, 2007). The style of some of the documents is rather relaxed and the antecedent may be ambiguous even for a human reader. We observed three types of anaphora in the corpus. First, the standard anaphora which includes both pronouns and sortal anaphora, which requires a unique bacterial antecedent. Second, *bi-anaphora* or an anaphora that requires two bacteria antecedents. This happens when the properties of two strains are compared in the document. Finally, the case of a higher taxon being used to refer to a lower taxon, which we named *name taxon anaphora*.

Anaphora with a unique antecedent

C. coli is pathogenic in animals and humans. People usually get infected by eating poultry that contained the bacteria, eating raw food, drinking raw milk, and drinking bottle water […].

Anaphora with two antecedents

C. coli is usually found hand in hand with its bacteria relative, C. jejuni. These two organisms are recognized as the two most leading causes of acute inflammation of intestine in the United States and other nations.

Name taxon anaphora

Ticks become infected with Borrelia duttonii while feeding on an infected rodent. Borrelia then multiplies rapidly, causing a generalized infection throughout the tick.

For anaphora detection and resolution a pattern-based approach was preferred to machine learning because the constraints for relating anaphora to antecedent candidates of the same taxonomy level were mainly semantic and domain-dependent and the annotation of anaphora was not provided in the training corpus.

Anaphora detection consists of identifying potential anaphora in the corpus, given a list of pronouns, sortal anaphora and taxa and then filtering out irrelevant cases (Segura-Bedmar *et al.*, 2010, Lin & Lian, 2004) before anaphora resolution. Not all the pronouns, sortal anaphora terms and higher taxon bacteria are anaphoric. For example, if a higher taxon is preceded or followed by the word *genus*, this signals that it is not anaphoric but that the text is actually about the higher taxon.

Non-anaphoric higher taxon

Burkholderia cenocepacia HI2424[…]
The genus Burkholderia consists of some 35 bacterial species, most of which are soil saprophytes and phytopathogens that occupy a wide range of environmental niches.

The anaphora resolution algorithm takes into account two features: the distance to the antecedent candidate and its position in the sentence. The antecedent is usually found in proximity to the anaphora, in order to maintain the coherence of the text. Therefore, our method ranks the antecedent candidates according to the anaphoric distance counted in sentences.

If more than one bacterium is found in a given sentence, their position is discriminate. Centering theory states that in a sentence the most prominent entities and therefore the most probable antecedent candidates are in the order: subject > object > other position (Grosz *et al.*, 1995). In English, due to the SVO order of the language the subject is most often found at the beginning of the sentence, followed by the object and the others. Therefore, the method retains the leftmost bacterium in the sentence when searching for the best antecedent candidate.

More precisely, the method selects the first antecedent that it finds according to the following precedence list:

- First bacterium in the current sentence (s)
- First bacterium in the previous sentence (s-1)
- First bacterium in sentence s-2
- First bacterium in sentence s-3
- First bacterium in the current paragraph
- Last bacterium in the previous paragraph
- First bacterium in the first sentence of the document
- The first bacterium ever mentioned.

The method only relates anaphora to antecedents that are found before. It does not handle cataphors since they are rarely found in the corpus. For anaphors that require two antecedents we use the same criteria but search for two bacteria in each sentence or paragraph, instead of one. For taxon anaphora we look for the presence of a lower taxon in the document found before the anaphora that is compatible according to the species taxonomy.
The counts of anaphora detected by the patterns are given in Table 1.

Corpus	Single ante	Bi ante	Taxon ante
Train	933	4	129
Dev	204	3	22
Test	240	0	18
Total	1,377	7	169

Table 1. The count of the types of anaphora per corpus.

The anaphora resolution algorithm allowed us to retrieve more sentences that contain both a bacterium and a location. Out of the 574 sentences that contain only a location, 436 were found to contain an anaphora related to at least one bacterium. The remaining 138 sentences are cases where there is no bacterial anaphora or the bacterium name is implicit. It frequently happens that the bacterium is referred to through its action. For example in the sentence below, the bacterium name could be derived from the name of the disease that it causes.

In the 1600s anthrax was known as the "Black bane" and killed over 60,000 cows.

One of the questions we had about the resolution of anaphora is whether anaphora that are found in the same sentence together with a bacterium (therefore potentially its antecedent) should be consid-

ered or not. We tested this on the development set. We found that removing such anaphora from consideration improved the overall score. It yielded an F-score of 53.22% (precision: 46.17%, recall: 62.81%), compared to the original F-score of 50.15% (precision: 41.06%, recall: 64.44%). This improvement in F-score is solely due to an increase in precision, which shows that while resolving anaphora is important and required, the incorrect recognition of terms as anaphora and incorrect anaphora resolution can introduce noise.

4 Relation extraction

In this work we concentrated most of our effort on the prediction of entities. For the prediction of events we used a strategy based on the co-occurrence of arguments and trigger words within a sentence:

- If a bacteria name, a location and a trigger word are present in a sentence, then the system predicts a Localization event between the bacterium and the location.
- If a bacteria anaphora, a location and a trigger word are present in a sentence, then the system predicts a Localization event between each anaphora antecedent and the location.
- If a host, a host part, a bacterium and at least one trigger word are present in a sentence, then the system predicts a *PartOf* event between the host and the host part.

The list of trigger words contains 20 verbs (*e.g. inhabit*, *colonize*, but also *discover*, *isolate*), 16 disease markers (*e.g. chronic*, *pathogen*) and 19 other relevant words (*e.g. ingest*, *environment*, *niche*). This list was designed by ranking words in the sentences of the training corpus containing both a bacteria name and a location. The ranking criterion used was the information gain with respect to whether the sentence contained an event or not. The ranked list was adjusted by removing spurious words and adding domain knowledge words.

By removing the constraint of the occurrence of a trigger word in the sentence, we can determine that the maximum recall the method can achieve with this strategy is 47% (precision: 41%, F-score: 44%). The selected trigger word list yielded a recall close to the maximum, thus it seems that the trigger words do not affect the recall and are suitable for the task.

5 Results

Table 2 summarizes the official scores that the Bibliome Alvis system achieved for the Bacteria Biotope Task. It ranked first among three participants. The first column gives the recall of entity prediction. The prediction of hosts and bacteria named-entities achieved a good recall of 84 and 82, respectively.

	Entity recall	**Event recall**	**Event Precis.**	**F-score**
Bacteria	84	-	-	-
Host	82	61	48	53
Host part	72	53	42	47
Env.	53	29	24	26
Geo.	29	13	38	19
Food	-	-	29	41
Medical	100	50	33	40
Water	83	60	55	57
Soil	86	69	59	63
Total		**45**	**45**	**45**

Table 2. Bibliome system scores at Bacteria Biotope Task in BioNLP shared tasks 2011.

However, geographical locations based on a similar strategy were poorly predicted (29%). Our system predicted only 15 countries. A more appropriate resource of geographical names than the Agrovoc thesaurus would certainly increase the recall of geographical locations.

The host parts, medical, water and soil locations predicted with the same ontology-based method were surprisingly good with a recall of 72, 100, 83 and 86, respectively. The small size of the ontology and the small number of different term heads (*i.e.* 51 different heads) initially appeared as a limitation factor for reuse on new corpora. The good recall shows that the location vocabulary of the test set has similarities with the training set compared to potential space of location names. The potential space is reflected by the richness of the GOLD isolation site field. This demonstrates the robustness of the type derivation approach based on term heads. The correctness of the derivation type cannot be calculated without a corpus where all the locations and not only bacteria ones are annotated. The recall of the environment location prediction is a little bit lower, 53%. The environment type in-

cludes many different types that cannot all be anticipated. Therefore the coverage of the BB termino-ontology environment part is limited except for water and soil, which are more focused topics.

The localization event recall (column 2) is on average 20% lower for all types than the location entity recall. The regularity of the difference may suggest that once the argument is identified, the localization relation is equally harder to find by our method independently of the type. The localization event precision (column 3) is more difficult to analyze because many sources of error may be involved, such as an incorrect arguments, incorrect anaphora resolution, relation to the wrong bacterium among several or the absence of a relation.

The prediction precision of localization events involving soil, water and host is better than environment and food. The manual analysis of the test corpus shows that in some cases environmental locations were mentioned as potential sources of industrial applications without actually being bacteria isolation places. For instance, in *Other fields of application for thermostable enzymes are starch-processing, organic synthesis, diagnostics, waste treatment, pulp and paper manufacture, and animal feed and human food,* the Alvis system erroneously predicted *waste treatment*, *paper manufacture*, *animal feed* and *human food*. This is due to the fact that the system does not handle modalities. Such hypotheses are specific to the BB task text genre, *i.e.* Bacteria sequencing projects. Such projects contain details for potential industrial applications, which are absent from academic literature.

Ambiguous types are also a source of error. Despite the host dictionary cleaning, some ambiguities remained. For example, the head *canal* in *tooth root canal* is erroneously typed as water and should be disambiguated with its *tooth* host-part modifier.

After test publication we measured the gain of anaphora resolution by using the on-line service. The anaphora resolution algorithm was found to have a strong impact on the final result. Running the test set using all of the modules *except for* the anaphora resolution algorithm yielded a decrease in the F-score by almost 13% (F-score: 32.5%, precision: 48.5%, 24.4%). This shows that the addition of an anaphora resolution algorithm significantly increases the precision and that a resolution algorithm adapted to the Bacteria domain is necessary for the Biotope corpus.

The part-of event prediction relies on the strict co-occurrence of a bacterium, trigger word, host and host part within a sentence. An additional run with the more relaxed constraint where the bacterium can be denoted by an anaphora as well yielded a gain of 6 recall points, a loss of 5 precision points with a net benefit of 1 F-measure point.

6 Discussion

The use of trigger words for the selection of sentences for relation extraction does not take into account the structure or syntax of the sentence for the prediction of relation arguments. The system predicts all combinations of bacteria and locations as localization events and all combination of host and host parts as part-of event. This has a negative effect on the precision measure since some pairs are irrelevant as in the sentence below.

Baumannia cicadellinicola. This newly discovered organism is an obligate endosymbiont of the leafhopper insect Homalodisca coagulata (Say), also known as the Glassy-Winged Sharpshooter, which feeds on the xylem of plants.

It has been shown that the use of syntactic dependencies to extract biological events (such as protein-protein interactions) improves the results of such systems (Erkan *et al.*, 2007, Manine *et al*., 2008, Airola *et al*. 2008). The use of syntactic dependencies could offer a more in depth examination of the syntax and the semantics and therefore allow for a more refined extraction of bacteria-localization and host-host part relations.

Term extraction appears to be a good method for predicting locations including unseen terms, but it is limited by the typing strategy that filters out all terms with unknown heads (with respect to the BB termino-ontology). In the future, we will study the effect of linguistic markers such as enumeration and exemplification structures for recovering additional location terms. For instance, in *heated organic materials such as compost heaps, rotting hay, manure piles or mushroom growth medium*, our system has correctly typed *heated organic materials* as environment but not the other examples because of their unknown heads.

The promising performance of the Alvis system on the BB task shows that a combination of semantic analysis and domain-adapted resources is a good strategy for information extraction in the biology domain.

References

Agrovoc: http://aims.fao.org/website/AGROVOC-Thesaurus

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahnikkala, Filip Ginter, and Tapio Salakoski. 2008. A Graph Kernel for Protein-Protein Interaction Extraction. BioNLP2008: Current Trends in Biomedical Natural Language Processing, pages 1-9.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proceedings of AMIA Symposium 2001, pages 17-21.

Emmanuel Desmontils, Christine Jacquin and Laurent Simon. 2003. Ontology enrichment and indexing process. Research report RR-IRIN-03.05, Institut de Recherche en Informatique de Nantes, Nantes, France.

Jérôme Euzenat and Pavel Shvaiko. 2007. Ontology matching, Springer Verlag, Heidelberg (DE),page 333.

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

GeoNames: http://www.geonames.org/

Gregory Grefenstette. 1994. Explorations in Automatic Thesaurus Discovery. Natural Language Processing and Machine Translation. London: Kluwer Academic Publishers.

Barbara J. Grosz, Araving K. Joshi and Scott Weinstein. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse. University of Pennsylvania Institute for Research in Cognitive Science Technical Reports Series.

Güneş Erkan, Arzucan Özgür and Dragomir R. Radev. 2007. Semi-Supervised Classification for Extracting Protein Interaction Sentences using Dependency Parsing. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 228-237.

Thierry Hamon and Sophie Aubin. 2006. Improving term extraction with terminological resources. In Salakoski, T. et al., editors, Advances in Natural Language Processing 5th International Conference on NLP (Fin- TAL'06), pages 380–387. Springer.

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: experiment and results, Recent Advances in Computational Terminology. Pages 185-208. John Benjamins.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Zampolli, A.(ed.), Proceedings of the 14 th COLING, pages 539–545, Nantes, France.

Christian Jacquemin and Evelyne Tzoukermann. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In Strzalkowski, T. (ed.), Natural language information retrieval, volume 7 of Text, speech and language technology, chapter 2, pages 25–74. Dordrecht & Boston: Kluwer Academic Publishers.

Matthew A. Jaro. 1989. Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84(406), pages 414-20.

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

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. Doklady akademii nauk SSSR, 163(4):845-848, 1965. In Russian. English translation in Soviet Physics Doklady, 10(8), pages 707-710.

Yu-Hsiang Lin and Tyne Liang. 2004. Pronomial and Sortal Anaphora Resolution for Biomedical Literature. In Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing.

Konstantinos Liolios, I-Min A. Chen., Konstantinos Mavromatis, Nektarios Tavernarakis, Philip Hugenholtz, Victor M. Markowitz and Nikos C. Kyrpides. 2009. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. NAR Epub.

Alain-Pierre Manine, Erick Alphonse and Philippe Bessières. 2008. Information extraction as an ontology population task and its application to genic interactions, 20th IEEE Intl. Conf. Tools with Artificial Intelligence, ICTAI'08., vol. II, pp. 74-81.

NCBI taxonomy: http://www.ncbi.nlm.nih.gov/Taxonomy/

Claire Nédellec, Wiktoria Golik, Sophie Aubin and Robert Bossy. 2010. Building Large Lexicalized Ontologies from Text: a Use Case in Indexing Biotechnology Patents, International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010), Lisbon, Portugal.

Isabel Segura-Bedmar, Mario Crespo, César de De Pablo-Sánchez and Paloma Martínez. 2010. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics 11(Supl 2):S1.

Manabu Torii and K. Vijay-Shanker. 2007. Sortal Anaphora Resolution in Medline Abstracts. Computational Intelligence 23, pages 15-27.

Zhou GuoDong, Su Jian, Zhang Jie and Zhang Min. 2005. Exploring Various Knowledge in Relation Extraction. In Proceedings of the 43rd Annual Meeting of the ACL, pages 427-434, Ann Arbor. Association for Computational Linguistics.