Automatic extraction of microorganisms and their habitats from free text using text mining workflows

BalaKrishna Kolluru[1,2*], Sirintra Nakjang[3,4],Robert P. Hirt[3], Anil Wipat[3,4], Sophia Ananiadou[1,2]

[1]National Centre for Text Mining, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

[2]School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL, UK

[3]Institute for Cell and Molecular Biosciences, University of Newcastle, Newcastle upon Tyne, NE2 4HH, UK

[4]School of Computing Science, University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK

Summary

In this paper we illustrate the usage of text mining workflows to automatically extract instances of microorganisms and their habitats from free text; these entries can then be curated and added to different databases. To this end, we use a Conditional Random Field (CRF) based classifier, as part of the workflows, to extract the mention of microorganisms, habitats and the inter-relation between organisms and their habitats.
Results indicate a good performance for extraction of microorganisms and the relation extraction aspects of the task (with a precision of over 80%), while habitat recognition is only moderate (a precision of about 65%). We also conjecture that pdf-to-text conversion can be quite noisy and this implicitly affects any sentence-based relation extraction algorithms.

1 Introduction

Microorganisms play a significant role in symbiotic relationships with animal hosts ranging from mutualism, commensalism to parasitism. To gain more insight into the mechanisms involved in the host-microbe interactions, it is essential to be able to contrast genotypic features of microorganisms from various sources where microbes live, including both host-associated (from a range of hosts-microbes contexts) and various environmental niches [7]. To date, there is no detailed data source for this information regarding habitat or isolation source of microorganisms whose genome sequence data are available. GOLD [12] and NCBI [13] databases are some of the most popular public resources where information describing taxa can be obtained in a form of flat files.

Due to the large numbers of taxa for which genome sequence data are available, and their increase on daily basis, there is an urgent need to be able to describe the habitat or isolation source

*To whom correspondence should be addressed. Email: kollurub@cs.man.ac.uk

for each taxon in an automated and consistent fashion. This issue was recently specifically recognised by the "Minimal information about a Genome Sequence" (MIGS) specifications [8] and a few papers have discussed this further or applied some initial approaches to address this issue [9, 10]. In order to gather this scattered information, text-mining approaches are employed to extract metadata from the published literatures. Here, we present a biological use case investigating the aspects of mapping habitat to microorganisms via the use of text-mining techniques.

Extracting microorganisms and their habitats are just a part of the information-processing paradigm that could involve several other stages.

We address two related issues in this paper:

1. How effective are the statistical approaches for extracting microorganisms and their habitats?
2. Develop workflows to combine text processing, named entity recognition and relation-mining

2 Related work

A number of investigators have worked on automatic entity extraction for the Biology domain. Hanisch *et al.* [1] have developed a system called ProMiner for identifying entities from scientific literature. They have subsequently expanded this work for identifying proteins, genes and diseases. Sasaki *et al.* developed a CRF based Named Entity Recogniser (NER), Nemine that identifies genes and proteins using domain-specific dictionaries [4]. Using features such as orthographic features, Part-of-Speech (POS) tags, dictionaries and contextual information an F-score of 78.72% has been achieved. The same method has been adapted for the recognition of metabolites (using the dictionary ChemSpider) achieving an F-score of 78.49% (precision of 83.02%) [15]. Similar NER techniques have been deployed within the advanced search service, KLEIO [22], at the National Centre for Text Mining, UK. KLEIO uses NER boosted by including term variation (acronym detection) and normalisation (spelling variants). Collier *et al.* have developed a system called Biocaster [5], which employs ontologies and rules to support text mining to track public health rumours. Biocaster reported an F-score of 76.97%. Ananiadou *et al.* have employed named entity recognisers for type IV secretion systems and shown the applicability of statistical approaches for this domain [21]. They reported an F-score of about 90% for named entity recognition for identification of bacteria.

3 Data

Since this task represents novel challenges for application of text mining, there were no prior standard annotated corpora to serve as training data for machine learning algorithms or to provide a gold standard for evaluation. Furthermore, the types of relations and patterns of term occurrence, in which we are interested, are not typically attested in the abstracts of the papers, but only appear as part of the full text of the articles. Therefore, we developed new training and evaluation corpus material for these concepts of interest, based on annotation of full papers. In

the subsequent sections, we describe the criteria used while creating the corpus. In both classes of entity, i.e., microorganisms and habitats, we followed a similar paradigm, employing independent manual annotation of some amount of text and augmenting that material as necessary through the use of an "accelerated annotation" (Acela) interface [18].

Acela iteratively and interactively trains a machine learning classifier to recognise a specific concept or entity class, based on the current set of labelled examples, labels new example sentences. These examples are then given to the human annotator for verification or correction. In previous work [18], this approach has been shown to achieve full annotation coverage with roughly 50% of the annotator effort, by focusing on only those sentences, which are most likely to include items that need to be annotated.

3.1 Organism-habitat corpus

For this corpus, two classes of entities were annotated: microorganisms and habitats.

- Microorganisms: Scientic names of microorganisms, including bacteria, archaea and microbial eukaryotes, are annotated
    1. If they are specified at least to the genus level of precision. Species, strain, and serovar entries are also tagged, if they are present. Typical examples of microorganisms are *Campylobacter spp.*, *Escherichia coli K12* and *Trichomonas vaginalis*.
    2. If they are in sentences which contain habitat or isolation source information for the organism.
- Habitats: Habitats or isolation sources of organisms, such as those stored in the GOLD database [12], are tagged
    1. If they are context-related or can be referred to as a habitat or isolation source of an organism. For example, if they refer to a host organism (*human*, *cow*), a body part or organ of a host organism (*lung*, *gut*,*lung abscess*), refer to an environmental habitat (*mine tailing*, *wastewater*), or employ the adjectival forms of habitats listed above (*bovine*, *pulmonary*, *rumen*).
    2. If they are not associated directly with disease e.g. diarrhea, respiratory tract infection.
    3. If they are in sentences, which contain the organism associated with the habitat.

A fully annotated sentence with organism and habitat information is shown here

> *Bacteroides salyersae sp. nov.* isolated from clinical specimens of *human intestinal origin*.

An example of the exclusion criterion for organism annotation appears in the sentence below, where *Campylobacter* is not tagged because the sentence lacks correct context; there is no habitat information.

> The *Campylobacter* species were all isolated anaerobically and identified by sequencing analysis of the 16S rRNA gene.

Table 1: Corpus statistics for Newcastle Organism-Habitat data

	Microorganisms	Habitats
Annotated sentences	1418	1609
Tokens (words)	57020	57020
Tagged Tokens	40349	47393
Tokens Entities	921	875
Tagged tokens (within the entities)	1951	1201
Estimated coverage	99.8%	99.4%

3.2 Corpus statistics

Here, the corpus was seeded with a set of annotated excerpts from 22 full papers that specified organisms and habitats, drawn from a bibliography provided by a domain expert. An additional set of 10 full text documents from the same expert was employed in the accelerated annotation process. Annotator instances for the two classes, microorganism and habitat, were created. An expert annotator then interactively applied the annotation interface to label instances until an estimated coverage over 99% was achieved. Estimated coverage is a ratio of number of manually annotated entities to the total number of entities as expected by the CRF classifier. See [18] for details. Detailed statistics for the resulting corpus are presented in Table 1.

4 Our approach

In this paper, we address the issue of automatically extracting different organisms, their habitats and the inter-relation from free literature. We employ workflows using text-mining tools for identification of organisms and habitats. We have used U-Compare [3] to design a workflow to build a named entity recogniser to accept plain text or a pdf. This workflow can be run both in the U-Compare environment as a stand-alone or as part of Taverna workflow management system [2]. Figure 1 shows the schematics view of our workflow.

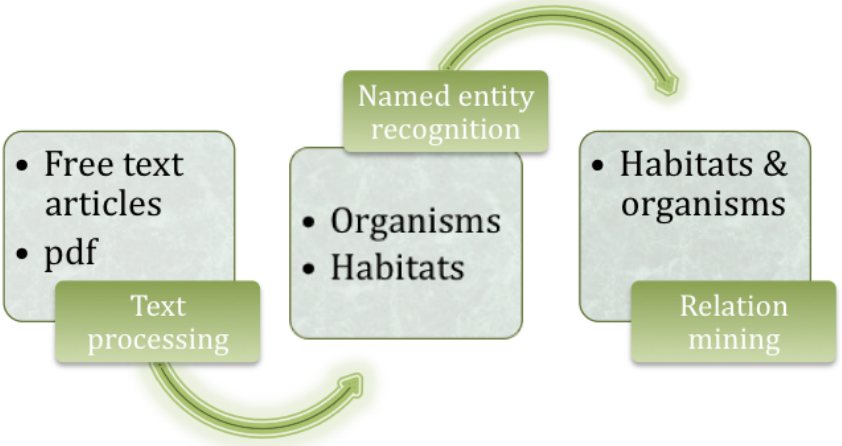

Figure 1: Overview of our approach.

The pdf-to-text converter component of the workflow was built on the application programming

interface (API) provided by the Apache group [19]. The component reads individual file and converts into to raw text.

The workflow implementing the named entity recogniser employs conditional random fields, CRFs [6] using a combination of dictionary (NCBI/habitat list) features, lexical features, orthographic features and contextual features. CRFs are a type of discriminative and undirected probabilities graphical models often used for tagging sequential data and in named entity recognition in natural language processing and biology domains [16, 17]. In our implementation, we use the Mallet [23] implementation of CRFs with lexical and orthographic features to train the CRF model. We have also employed two dictionaries that were tailored to the task from a combination of established and curated domain ontologies and term lists provided by domain experts.

4.1 Resources

- **Microorganism Resources** Two large-scale resources for scientic names for microorganisms were used:

  1. Microorganisms' scientific names from NCBI taxonomy including bacteria, archaea, microbial eukaryote (`http://ncbi.nlm.nih.gov/taxonomy`)
  2. List of Prokaryotic names and microbial eukaryotes with Standing in Nomenclature (LPSN) (`http://www.bacterio.net`)

  The bacterial names from these resources were converted to a set of standardised forms to cover typical variability for these terms including the common abbreviation of the genus term, removal of species, strain, or serovar components, removal of tags such as '*subsp.*', '*str.*', '*strain*', etc., and generation of plural forms for genus terms. The resulting term list comprises 52715 entries for 12256 distinct organisms.

- **Habitat Resources** A key source of habitat terms is the GOLD [12] database, which includes 135 habitat types. After some simple normalisation to enhance matching between the terms and running text, the specific dictionary was further enhanced with names of animals (169306 entries, 166244 head terms) and body parts and organs (120668 entries, 56578 head terms) extracted from the UMLS Metathesaurus [14].

4.2 Features for the CRF model

These experiments incorporate three main sets of base features, inspired by previous research in biomedical NER [4].

- Lexical features are current word, the root form of the current word, and the part-of-speech tag of the current word, computed by the Genia tagger [24].
- Orthographic features are made up of substring and word form features. In the word form features, all uppercase letters are converted to 'A', lowercase to 'a', and all numbers to '0'. The first two & four characters and the last two & four characters of the original word and the word form are chosen as features.

- Dictionary features are binary features to indicate the presence of the word in the dictionary and the position of the word within any dictionary entries.

For each of the base features, corresponding features for words within a context window are added to the representation. The window ranges from 1-3 words preceding and following the current word.

4.3 Hybrid Dictionary-Machine Learning Based Approach

For entity recognition, we used the CRF based approach. The current approach employs a sequence classifier, trained on a hand-annotated corpus, which was converted to a standard BIO (Begin of a sequence, Inside a sequence, Outside a sequence) format. This corpus consisted of 32 full papers from various journals and was specifically annotated for microorganisms and habitats. The CRFs were employed with a linear chain model.

4.4 Machine Learning Approach for Relation mining workflows

The main focus of this experiment was to elicit the sentences containing a microorganism-habitat relation from free-text, typically in a pdf-format.

The workflow had three principal components, as shown in Figure 1:

1. PDF-to-text convertor: this component was based on Apache's pdfbox API. It converts each pdf into a stream of text into a Java String object.
2. Named entity recogniser, as described in the earlier section: input from this component was used to populate the feature-space for CRF component.
3. CRF-component for building a bespoke relation mining apparatus

Relation mining workflow was based on a CRF model trained on co-occurrence of microorganisms and their habitats, thus modelling the relation between a microorganism and its habitat. Just as for the named entity recognition experiments described in earlier sections, Mallet [23] implementation was used for these experiments with a feature-space designed for sentences extraction.

4.5 Features for Relation mining CRF component

These experiments incorporate two main sets of base features

- Entity occurrence in a sentence i.e. if a sentence had any organism or habitat
- Contextual information (the word preceding the entity and the word following it) around the occurred entities

Table 2: Performance of CRF and dictionary-based approaches

	Microorganisms			Habitats		
	P (precision)	R (recall)	F(F-score)	P	R	F
Dictionary	54	75	63	58	55	56
CRF	84	79	81	68	50	57

5 Results & Discussion

5.1 Dictionary-based & CRF approaches

First, we consider the contrast between recognition accuracy for the systems employing the CRF with all features trained on the annotated corpora with the recognition accuracy obtained using a simple, longest match strategy using dictionary resources. The results are shown in Table 2.

We can observe that the results range from F-score of 56% to 63% for dictionary-based approaches and from 57% to 81% for machine learning based methods, using all features. These figures show some interesting contrasts. For microorganism tagging, the CRF approach outperforms the dictionary. While recall remains good for the dictionary-based approach, the precision is on 54%. In other words, the dictionary-based approach had a high number of false positives. The precision improved to 84% in the CRF set up; even though the recall is not very different from the dictionary-based approach.

Furthermore, the habitat annotation task is shown to be particularly challenging. Several factors such as a very broad class definition spanning animals, anatomy, environment and adjectival forms thereof, in conjunction with the restriction to organismal contexts and a lack of explicit word-morphologies restrict the CRF-based approach.

5.2 Hybrid CRF

As can be seen from Table 3, the hybrid classifier has achieved an F-score of about 80% and 60% for Organisms and Habitats respectively on a 9-fold cross validation.

As a general observation, microorganisms had a distinct attributes: word shape features, such as upper case letters, a subtle pattern in names such as "*ccus*" or "*cci*" etc. The current features that we have used, could model the microorganisms well and an F-Score of 81 is an indicative of this success. Typical false negatives include hyphenated words such as "*B. taylorii-like*" and partial names such as "*M. succiniciproducens*" of the complete name "*M. succiniciproducens MBEL55E (KCTC 0769BP)*".

Habitats, on the other hand, seemed to lack any such explicit attributes making them more difficult to model. There were quite a few false negatives such as *water*, *skin*, *abcess* etc. We conjecture that perhaps more training data (we used only 850 training instances of habitats for the experiments discussed here) could certainly help alleviate such false negatives. We are also currently exploring widening our feature-space by including deeper semantic information such as parse-tree information to enhance the performance of habitat recognition.

Table 3: Performance of our approach

Class of entities	Precision (%)	Recall (%)	F-score (%)
Microorganism	84	79	81
Habitats	68	50	57

Table 4: Performance of relation extraction approach

	Precision (%)	Recall (%)	F-score(%)
Relations	85	49	57

6 Relation mining

The relation mining classifier achieved a precision of about 85% and a recall of less than 50% (Table 4). We conjecture that slightly below par performance of habitats was partly accounted for by the contextual features as a result of which the precision was not proportional to the individual precisions of microorganisms and habitats. One of the important factors for the number of false negatives is the noise generated from pdf-text conversion. Typical examples of "corrupted" sentences include those where sentence boundaries could not be identified automatically and therefore had more noisy features. This could have had an impact on the performance. A typical example of corrupted sentence is a journal title concatenating with the title of the paper is shown here (taken from [25])

> "*Biometrika 40:237–264 Hallberg KB, Johnson DB (2003) Novel acidophiles isolated from moderately acidic mine drainage waters.*"

And another example where the caption of a figure and some text from the figure itself are combined with the text from [26] is shown here:

> "These mechanisms may have evolved in bacterial pathogens to increase the frequency of phenotypic variation in genes involved in 1 100,000 200,000 300,000 1,600,00 Figure 2 Circular representation of the H. pylori 26695 chromosome."

Clearly, any paradigm that extracts sentences will be affected by such erroneous conversion. There are several softwares which convert a pdf-document to text such as Apache's pdfbox [19], Utopia [20] and Unix command "pdftotext". But all of them have the same problems in dealing with tables and figures in pdfs, which result in such noise. We are currently working eliminating some of the erroneous sentences using statistical methods such as language modelling. We are also working on using parsers to eliminate some of HTML-related noise such as captions and headers etc.

Another factor that we could identify for the lower recall was the below par performance of the habitat recognition by the automatic named entity recognisers.

6.1 A toy example

Consider a small excerpt about Acinetobacter species [27]

"***Acinetobacter spp*** are widely distributed in nature and soil. Most infections occur in immunocompromised individuals, and the strain ***Acinebacter baumannii*** is the second most commonly isolated nonfermenting bacteria in human specimens. It can survive on the ***human skin*** or dry surfaces for weeks."

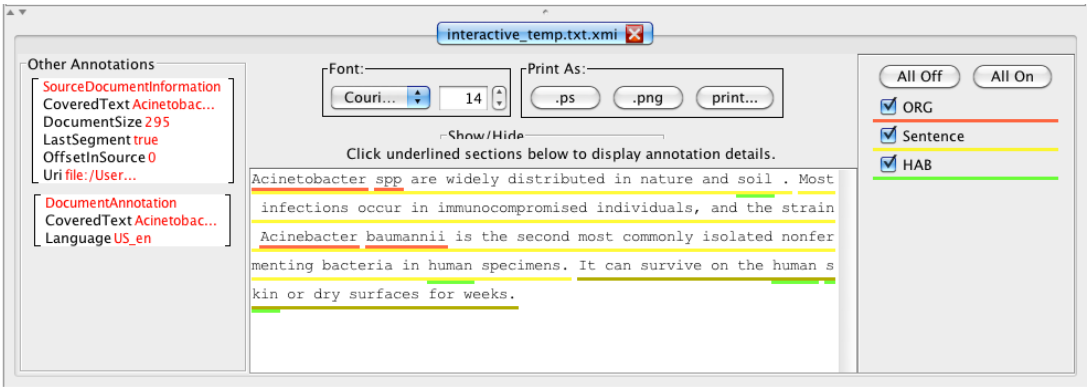

Figure 2: The result of our workflow on extraction of microorganisms and habitats in U-Compare

Figure 2 shows the result of our text-mining workflow on this excerpt. All the microorganisms are underlined in red, the habitats are underlined in green and the relation-indicating sentences are underlined in yellow. The workflow underlines sentences in yellow if and only if it contains both microorganisms and their habitats, thus highlighting host-associated habitats from non-host-associated ones.

7 Conclusion

As can be inferred from the results, the workflows that we have developed achieve a reasonable accuracy that makes them likely candidates to be ported for other databases and similar applications. We are planning to implement this workflow to curate a microorganism-habitat database. As the workflows are inter-operable as sub-workflows in Taverna [2], we hope the scientific community at large, and biologists in particular, can make use of them for their respective research projects.

Acknowledgements

This research has been supported by the Biotechnology & Biological Sciences Research Council (ONDEX project, BB/F006039/1). The National Centre for Text Mining is funded by the Joint Information Systems Committee. Sirintra Nakjang was supported by the Faculty of Medical Sciences and the School of Computing Science at Newcastle University and an Overseas Research Students Scheme.

References

[1] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer and J. Fluck. ProMiner: Organism specific protein name detection using approximate string matching. *EMBO Workshop*, Granada, Spain, March 28–31, 2004.

[2] Taverna, `http://www.taverna.org.uk`, last accessed on 08 July 2011.

[3] Y. Kano, W. A. Baumgartner, L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunder and J. Tsujii. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, 25(15):1997–1998, 2009.

[4] Y. Sasaki, Y. Tsurouka, J. McNaught and S. Ananiadou. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9(Suppl 11):S5, 2008.

[5] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, *et al.* BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941, 2008.

[6] J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of 18th International Conference on Machine Learning*, 2001.

[7] O. O'Sullivan, J. O'Callaghan, A. Sangrador-Vegas, O. Auliffe, L. Slattery, P. Kaleta, M. Callanan, G. F. Fitzgerald, R.P. Ross, and T. Beresford. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC Microbiology*, 9(1):50, 2009.

[8] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, *et al.* The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–547, 2008.

[9] L. Hirschman, C. Clark, K. B. Cohen, S. Mardis, J. Luciano, R. Kottman, J. Cole, V. Markowitz, N. Kyrpides, N. Morrison, *et al.* Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS*, 12(2):129–136, 2008.

[10] C. von Mering, P. Hugenholtz, J. Raes., S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–1130, 2007.

[11] J. H. Lee, V. N. Karamychev, S. A. Kozyavkin, D. Mills, A. R. Pavlov, N. V. Pavlova, N. N. Polouchine, P. M. Richardson, V. V. Shakhova, A. I. Slesarev, *et al.* Comparative genomic analysis of the gut bacterium longum reveals loci susceptible to deletion during pure culture growth. *BMC Genomics*, 9(1):247, 2008.

[12] K. Liolois, I. A. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 38(Database issue):D346–D354, 2010.

[13] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(Database issue):D38–D51, 2011.

[14] S. J. Nelson, T. Powell and B. L. Humphreys. The United Medical Language System (UMLS) Project, 2006. `http://www.nlm.nih.gov/mesh/umlsforelis.html`, last accessed 08 July 2011.

[15] C. Nobata, P. D. Dobson, S. A. Iqbal, P. Mendes, J. Tsujii, D. B. Kell, and S. Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature . *Metabolomics*, 7:94–101, 2011.

[16] R. McDonald, and F. Pereira. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics*, 6(Suppl 1):S6, 2005.

[17] J. R. Finkel, and C. D. Manning. Hierarchial Bayesian domain adaptation. *In Proceedings of Empirical Methods in Natural Language Processing*, 2009.

[18] Y. Tsuruoka, J. Tsujii and S. Ananiadou. Accelerating the annotation of sparse named entities by dynamic sentence selection. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, ACL, 2008.

[19] Apache, PDFBox tool `http://pdfbox.apache.org`, last accessed on 26 April 2011.

[20] S. Pettifer, J. R. Sinnott, and T. K. Attwood. UTOPIA – user friendly tools for operating informatics applications. *Comparative and Functional Genomics*, 5(1):56–60, 2004.

[21] S. Ananiadou, D. Sullivan, W. J. Black, G. Levow, J. Gillespie, C. Mao, S. Pyysalo, B. Kolluru, J. Tsujii, and B. Sobral. Named Entity Recognition for Bacterial Type IV Secretion Systems. *PLoS ONE*, 6(3):e14780, 2011.

[22] C. Nobata, Y. Sasaki, N. Okazaki, C. J. Rupp, J. Tsujii, and S. Ananiadou. Semantic Search on Digital Document Repositories based on Text Mining Results. *In International Conferences on Digital Libraries and the Semantic Web*, 2009.

[23] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`, last accessed 08 July 2011.

[24] Y. Tsuruoka, and J. Tsujii. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. *In Proceedings of HLT/EMNLP*, 2005.

[25] G.-L. Tan, W.-S. Shu, K. B. Hallberg, F. Li, C.-Y. Lan, W.-H. Zhou and L.-N. Huang. Culturable and molecular phylogenetic diversity of microorganisms in an open-dumped, extremely acidic Pb/Zn mine tailings. *Extremophiles*, 12(5):657–664, 2008.

[26] J-F. Tomb, O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, *et al.* The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*, 388(389):539–547, 1997.

[27] Wikipedia on Acinetobacter, `http://en.wikipedia.org/wiki/Acinetobacter`, last accessed on 08 July 2011.