Applying a text mining framework to the extraction of numerical parameters from scientific literature in the biotechnology domain

André Santos, Regina Nogueira, Anália Lourenço
andresantos@deb.uminho.pt, nogueira@isah-uni.hannover.de, analia@deb.uminho.pt
IBB - Institute for Biotechnology and Bioengineering, Centre of Biological Engineering,
University of Minho, Portugal

KEYWORD

Text mining
Biotechnology applications
Procedure optimization

ABSTRACT

Scientific publications are the main vehicle to disseminate information in the field of biotechnology for wastewater treatment. Indeed, the new research paradigms and the application of high-throughput technologies have increased the rate of publication considerably. The problem is that manual curation becomes harder, prone-to-errors and time-consuming, leading to a probable loss of information and inefficient knowledge acquisition. As a result, research outputs are hardly reaching engineers, hampering the calibration of mathematical models used to optimize the stability and performance of biotechnological systems. In this context, we have developed a data curation workflow, based on text mining techniques, to extract numerical parameters from scientific literature, and applied it to the biotechnology domain. A workflow was built to process wastewater-related articles with the main goal of identifying physico-chemical parameters mentioned in the text. This work describes the implementation of the workflow, identifies achievements and current limitations in the overall process, and presents the results obtained for a corpus of 50 full-text documents.

1. Introduction

The field of ecotechnologies for wastewater treatment is being challenged by a gap of communication between wastewater scientists and engineers concerning the way to integrate the knowledge about microbial ecology acquired by the first ones in the design and optimization of the work that the latest perform in wastewater treatment plants (WWTPs).

New insights into the composition of the bioceosis present in these systems have been gathered during the last decades by researchers and engineers, especially after the “revolution” introduced by the use of molecular biology methods for the detection, identification and quantification of microorganisms. Nevertheless, the use of these data in the optimization of WWTPs is hampered by the current inability to relate them with conventional wastewater operating parameters [Hamouda et al., 2009].

Indeed, civil engineers are still designing and operating WWTPs mainly based on the characteristics of the wastewater and on "rules of thumb" obtained from practical experience. This gap is largely motivated by the nonexistence of a comprehensive and consistent knowledge base on the huge amount of observations available on microbial composition, structure and activity, and their relationship with operation/performance parameters of WWTPs. Data are widespread across an ever increasing number of scientific publications and researchers have to undertake a very tedious and error-prone process of manual curation to find and relate relevant data [Nogueira and Melo, 2006, Nogueira et al., 2002]. Researchers have

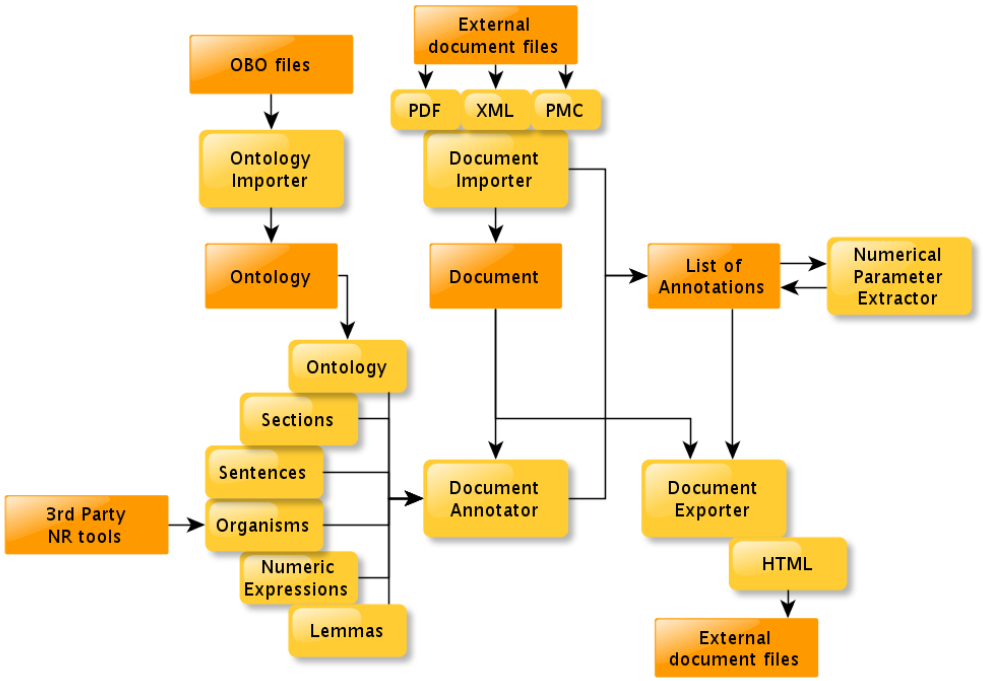

Figure 1: Generic T-Eng system architecture.

to cross references from multiple Web services (e.g. PubMed[1] and ScienceDirect[2]), inspect documents thoroughly and filter out irrelevant documents; then, they have to scrutinize the potentially relevant documents, managing all sorts of terminological inconsistencies (e.g. the same variable may be referred by different names, and numerical values may be expressed in a myriad of different units, often non-comparable directly); and, at the end, they have to summarize the bibliographic review and their own findings into a new publication. Data are thus being lost during compilation and processing but also at dissemination.

Given all the effort invested in using molecular biology tools to acquire as much data as possible about the microbiological communities and the associated environments, it is only logical that similar efforts should be put in incorporating these data into environmental decision support systems. Specifically, the development of suitable text mining systems, i.e. systems addressing the computer-assisted harvest and extraction of relevant information from the literature [Krallinger et al., 2010] , is considered the key for the large-scale, systematic and standardized compilation and curation of data. There is some recent work on the extraction of con-textual information for ecological analyses (e.g. physico-chemical variables and geographical locations) [Tamames and De Lorenzo, 2010] and some ontologies have been proposed in support of a consensual and unambiguous interpretation of wastewater vocabulary [Koegst et al., 2007, Ceccaroni et al., 2004].

We aim to go a step forward and delineate an automatic text mining workflow that looks for textual evidences of physical, chemical and microbiological related-parameters of WWTPs in scientific literature and constructs a network of information hyperlinked by wastewater concepts. This work outlines the architecture of such workflow, identifying the resources and tools in use as well as the interface created to interact with researchers, namely to assist on literature search and contents comparison. A first prototype of the knowledge base is available at:

http://stardust.deb.uminho.pt/t-eng-ww.

1 http://www.ncbi.nlm.nih.gov/pubmed/

2 http://www.sciencedirect.com

2. Literature curation workflow for microbial ecology

The project development was divided in several tasks:

- development of a domain specific ontology for wastewater
- tuning T-Eng for the project's needs
- processing

2.1. Building a wastewater ontology

In order to fine tune the text mining framework to the wastewater treatment domain, an ontology was built comprising the biological terms and units of measure commonly found in scientific literature. In this ontology, which has been designed, populated and curated by an expert, each entry -- a concept -- includes a common name and a number of synonyms, i.e. terms which may alternatively be used to refer to the same concept.

This vocabulary is used to generate the code responsible for annotating parameter names and units of measure in the documents, and, additionally, provides a code-independent platform to assist in the interchange of information between researchers and engineers.

Examples 1 and 2 present entries contained in the wastewater ontology:

- the first entry classifies *activated sludge* as a domain entity and indicates *nitrifying activated sludge* as a (narrower) synonym;
- the second entry refers *cells/l* as a unit of measure, listing *cells(-1)*, *cells-1* and *copies/l* as equivalent synonyms.

```
1 [Term]
2 id: ID:0000007
3 name: activated sludge
4 synonym: "nitrifying activated
     sludge" NARROW []
5 is_a: ID:0000026 ! domain
     entities
```

Example 1: Examples of entries (***activated sludge***) in the wastewater ontology.

```
1 [Term]
2 id: ID:0000014
3 name: cells/l
4 synonym: "cells(-1)" RELATED []
5 synonym: "cells-1" RELATED []
6 synonym: "copies/l" RELATED []
7 is_a: ID:0000002 !
     units_of_measure
```

Example 2: Examples of entries (***cells/l***) in the wastewater ontology.

2.2. Literature mining annotation process

The data curation workflow was set on top of a text mining framework developed in-house and named T-Eng.

T-Eng consists of a set of Perl libraries and command-line tools developed to perform text mining-related activities in scientific literature. Its main goal is the extraction of relevant information, based on the recognition of physic, chemical and microbiological parameters.

In order to process the wastewater papers, T-Eng components were used to build a processing workflow, starting with the document importation, the generation of structure-related annotations and content-related annotations, the extraction of knowledge from the annotated elements and the exportation of the annotated documents. The whole workflow can be found in Figure 1: Generic T-Eng system architecture.

2.2.1. Importing documents and ontologies

The importation of abstracts and/or full-text documents in a number of file formats (e.g. PDF and commonly used XML schemas) and the delimitation of document sections are the priority in terms of document compilation and pre-processing. They grant the ability to process most of the documents available online. Despite the importation process being slightly different for distinct formats and sources, it usually consists in converting the original file to plain text while trying to extract information about the document structure (e.g. sections titles and boundaries),

which can be later used to restrict or focus the processing on specific sections.

2.2.2. Annotating documents

After being imported, documents are submitted to several parallel processes of annotation which are responsible for the recognition of different types of data from the document. These data will later be used in further classification and relation inference mechanisms which will ultimately lead to knowledge extraction.

The annotation process consists of marking a segment (e.g. one or more words, a sentence, a paragraph) of the document with some kind of information. Each annotation has to be somehow linked to a specific "zone" of the document. Typically, there are two ways of doing this: either by inserting predefined marks on the boundaries of the segmented being annotated, marking the beginning and the end of the annotation and other relevant information, or by maintaining the annotations in a standalone file which contains the offsets of the beginning and end characters of the annotation. The latter form of annotation does not require the modification of the original document and makes it easier to cope with the possible overlap of different types of annotations. This is the form used in T-Eng, allowing the documents to be preserved all the way since they are imported till they are exported.

```
1 Total microbes measured by dot-blot
2 hybridizations, ranged from 3*10^11
3 to 3*10^12 cells/L.
```

Example 3: Example of a sentence before annotation [Dionisi et al., 2002].

```
1 <entity class="variable_name"
2 ont_id="10">Total microbes</entity>,
3 measured by dot-blot hybridizations,
4 ranged from <numexp type="range">
5 3*10^11 to 3*10^12<type> <entity
6 class="unit" ont_id="14"> cells/L
7                </unit>.
```

Example 4: Example of a sentence annotated with an inline format.

```
1 1, 14, Total microbes ,
        variable_name, 10
2 65, 82, 3*10^11 to 3*10^12, numexp ,
3 84, 90, cells/L , unit , 14
```

Example 5: Example of a stand-off annotation format.

Determining sentence and word boundaries

Being able to distinguish and operate over the different sentences which compose a document is a crucial feature in many text mining engines. Given that annotated elements are more likely related if they appear mentioned in the same sentence than if they are mentioned in separate sentences, knowing the sentences boundaries allows determining confidence levels on the relationships between annotated elements. Word tokenization, on the other hand, is required for the named entity recognizers which will look for biological entities and other terms of interest.

Given the previously mentioned goal of keeping the original text of the document unchanged, both the sentence splitter and the word tokenizer used by T-Eng provide a list with the offsets of the start and end positions of, respectively, the sentences and words in the document.

Annotating organisms and other biological entities

Identifying the biological entities and concepts being mentioned in the document is useful not only to infer what the document is about and to classify it, but also for the identification of the domain-specific variables.

T-Eng includes wrappers for third-party NER tools which can be used in the detection of relevant biological entities. Specifically, given the importance of extracting mentions to specific organisms in wastewater-related literature, T-Eng was extended to use Linnaeus, an open-source software tool for recognizing and normalizing species names [Gerner et al., 2010].

The annotation of other biological concepts of interest, perhaps more specific to the domain (e.g. in wastewater, nitrification processes, molecular methods), is supported by pattern-matching based in the

contents of the domain-specific ontology previously mentioned.

Annotating numerical parameters

Recognizing numerical expressions and units of measure is required to be able to identify numerical parameters (see Section 2.3for specifics). The recognition of each of these elements of information presents *per se* several challenges and demands for specific approaches of recognition.

T-Eng is able to recognize numerical expressions even when they are represented in uncommon notations, using UTF-8 characters or other types of "noise". When documents are converted from PDF to plain text, mathematical formulas or complex numerical expressions are often disfigured. Also, due to the lack of a true standard representation of mathematical formulas in plain text, a wide range of notations are used.

2.3. Extracting numerical parameters

Usually, the description of a numerical parameter comprises a triplet composed by the parameter short or extended name (e.g. *height*), a numerical value or expression (e.g. *5* or *3.4E-03*) and the units of measure (e.g. *meter*). The annotation steps described in Section 2.2.2 are meant to pave the way for the extraction of these parameters. In fact, the annotation process generates lists containing the occurrences of each of these three components, and the following challenge is to infer which ones are related: which variable name is connected with which numerical expression and which unit of measure.

```
1 The detection limit of each primer
2 set was in the range of 3x10^1 to
3 6x10^2 genes/reaction. Reliable
4 quantification of the target AOB DNA
5 was obtained when the target AOB DNA
6 comprised more than 0.1% of total
7 AOB DNA in the sample.
```

Example 6: Annotated elements in two consecutive sentences [Limpiyakorn et al., 2006].

Currently, T-Eng used the sentence delimitation to help extracting the relations, based on the idea that terms mentioned in the same sentence have a higher probability of being related. An example of this is presented in Example 6, were the number *0.1*, the unit *%*[3] and the variable name *total AOB* are all related and in the same sentence, while the number *3x10*[1] *to 6x10*[2] [4] in the previous sentence is unrelated.

This relation extraction mechanism is not yet fully implemented, and several improvements and extensions are already planned. See Section 4 for more details.

2.4. Exporting annotated documents

After annotating the documents and extracting the desired information, the results must be exported with two main goals:

- **Visualization**: Graphical and intuitive visualization of the results obtained, including elements such as graphs, tables, coloring the elements of interest and linking them to external resources.
- **Further processing**: Making all the information gathered available in convenient formats suitable for being processed by other tools. This includes exporting the documents and annotations to formats such as CSV, JSON, XML or other commonly used formats.

3. Results

The ontology created for the wastewater domain was built in the OBO file format, and is meant to be used not only to generate the code for the recognition of the relevant elements in the text, but also as a code-agnostic platform for discussing details of this work with people without requiring them to be familiar with the details of the implementation. Details about the number of entries in the ontology can be found in Table 1.

3 Despite % not being an actual unit of measurement, in this context, it is considered as one given the fact that, like a proper unit, it provides context to the numeric expression.

4 In a similar way, $3x10^1$ to $6x10^2$, being a *range*, contains two numbers instead of only one; however, T-Eng handles ranges as a single numerical expression.

Type of entity	Number of entities
Biological entities	24
Units of measure	6
Variable names	6

Table 1: Number of entries for each type of entity in the wastewater ontology.

In order to test the described tool, a set of nearly 50 scientific articles from the wastewater domain were selected by an expert, based on the importance and diversity of the parameters previously referred.

Given the interest in processing the articles' full-text, they were retrieved in PDF format. The examples presented in this article, however, are extracted from the abstracts only, due to the journal copyright laws to which the majority of the articles are subject to.

Example 7 presents the abstract of an article with the annotated elements highlighted.

In this example, entities like *industrial WWTP* and *municipal WWTP* were only partially recognized. In order to improve this, either new elements have to be inserted in the ontology, or the rules for the matching will have to be improved (as it already contains entries for *industrial wastewater* and *municipal wastewater*. Also, the word *copies* (line 8), together with *per cell* (lines 8 and 9) should have been considered a unit, but the recognition failed due to the words in between.

T-Eng also includes a web front-end which allows viewing the annotated abstracts with highlighted annotated terms, and presents table to navigate on the extracted entities. Currently being implemented is the automatic generation of graphics representing metrics about the entities annotated. Examples of all these visual features can be found in Figure 2.

4. Discussion and Future Work

This article presented an application of a text mining workflow system, T-Eng, to the task of inferring knowledge from articles belonging to the wastewater domain. T-Eng currently includes module for importing the articles from different formats and sources, extracting their structure, annotating the relevant elements and extracting information from them.

Regarding the importation of articles, PDF has proved to be a problematic format to handle, because it was not originally designed to be automatically processed. The elements in PDF documents are dis-

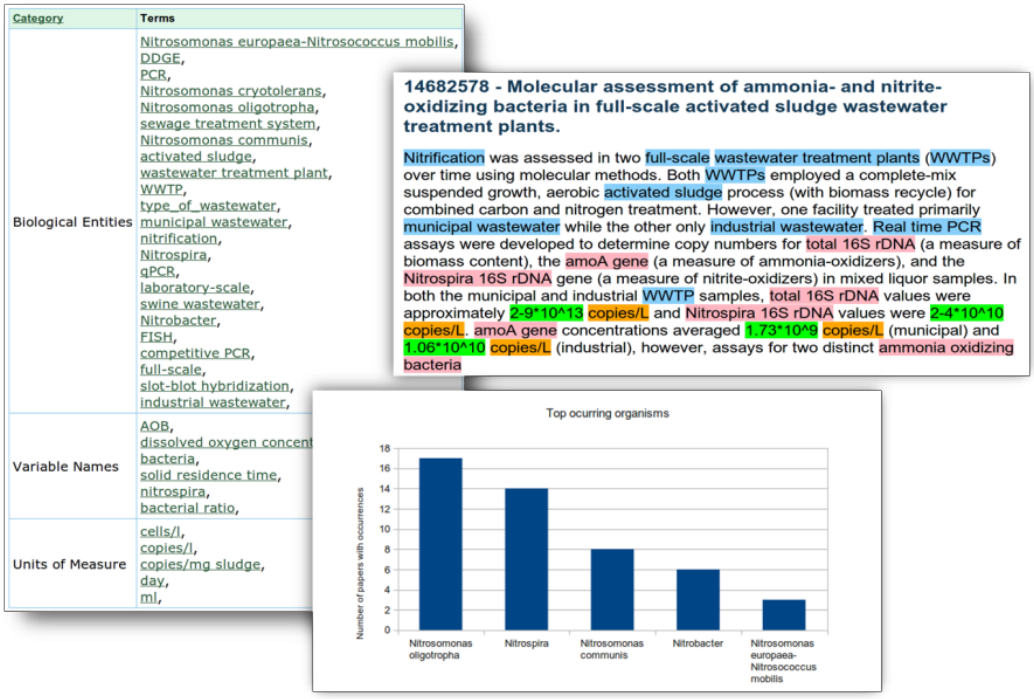

Figure 2: An example of the semantic navigation of annotated abstracts.

posed according to graphical directives, with little to no semantics involved. However, given that the majority of articles are available precisely in this format, there is the need to develop tools to process them as best as possible, and features such as the embedded table of contents which some PDF contain can help.

Most of the available tools for sentence splitting and work tokenization return the split sentences or the tokenized versions of the words without detailing where they were found in the original text. This motivated the development of in-house tools for these tasks whose output includes the character offsets of the boundaries of the sentences and words, which made possible working with stand-off annotation formats.

The annotation of numerical expressions in the text could not be performed using a simple regular expression due to the complexity and multiplicity of the notations used for representing mathematical formulas (which are made even worse in by the format conversion tools). As such, an in house tool was developed for this purpose. This tool has increased the recall of the recognition of numerical expressions, but simultaneously lowered its recall. Further adjustments will be made, trying to achieve a better point in this trade-off.

The annotation of units of measure, variable names and other relevant biological entities is performed with the help of a domain-specific ontology. The wastewater ontology will be improved to include the expected relations between variable names and units of measure.

The identification of variable names, numerical expressions and unit of measures occurring in the same sentence is a starting point for the extraction of numerical parameters. Next steps include implementing mechanisms of guessing which ones match based on typical sentence patterns.

Currently the annotated documents can be exported to HTML with the relevant elements highlighted in distinct colors.

```
1  Utilizing the principle of competitive PCR, we developed two assays to enumerate
2  Nitrosomonas oligotropha-like ammonia-oxidizing bacteria and nitrite-oxidizing
3  bacteria belonging to the genus NITROSPIRA: The specificities of two primer sets,
4  which were designed for two target regions, the amoA gene and Nitrospira 16S
5  ribosomal DNA (rDNA), were verified by DNA sequencing. Both assays were optimized
6  and applied to full-scale, activated sludge wastewater treatment plant (WWTP)
7  samples. If it was assumed that there was an average of 3.6 copies of 16S rDNA per
8  cell in the total population and two copies of the amoA gene per ammonia-oxidizing
9  bacterial cell, the ammonia oxidizers examined represented 0.0033% +/- 0.0022% of
10 the total bacterial population in a municipal WWTP. N. oligotropha-like ammonia-
11 oxidizing bacteria were not detected in an industrial WWTP. If it was assumed that
12 there was one copy of the 16S rDNA gene per nitrite-oxidizing bacterial cell,
13 Nitrospira spp. represented 0.39% +/- 0.28% of the biosludge population in the
14 municipal WWTP and 0.37% +/- 0.23% of the population in the industrial WWTP. The
15 number of Nitrospira sp. cells in the municipal WWTP was more than 62 times greater
16 than the number of N.oligotropha-like cells, based on a competitive PCR analysis.
17 The results of this study extended our knowledge of the comparative compositions of
18 nitrifying bacterial populations in wastewater treatment systems. Importantly, they
19 also demonstrated that we were able to quantify these populations, which ultimately
20 will be required for accurate prediction of process performance and stability for
21 cost-effective design and operation of e and stability for cost-effective design
22 and operation of WWTPs.
```

Example 7: Annotated abstract from [Dionisi et al., 2002].

5. References

[Ceccaroni et al., 2004] Ceccaroni, L., Cortés, U., and Sànchez-Marrè , M. OntoWEDSS: augmenting environmental decision-support systems with ontologies . Environmental Modelling & Software, 19(9)(2004) 785–797.

[Dionisi et al., 2002] Dionisi, H., Layton, A., Robinson, K., Brown, J., Gregory, I., Parker, J., and Sayler, G. Quantification of nitrosomonas oligotropha and nitrospira spp. using competitive

polymerase chain reaction in bench-scale wastewater treatment reactors operating at different solids retention times. Water environmental research, 2002, pp. 462–469 .

[Gerner et al., 2010] Gerner, M., Nenadic, G., and Bergman, C. Linnaeus: a species name identification system for biomedical literature. BMC Bioinformatics, 11(1)(2010) 85.

[Hamouda et al., 2009] Hamouda, M., Anderson, W., Huck, P., et al. Decision support systems in water and wastewater treatment process selection and design: a review. Water Science and Technology, 60(7)(2009)1757–1770.

[Koegst et al., 2007] Koegst, T., Tränckner, J., Blumensaat, F., Eichhorn, J., and Mayer-Eichberger, V. On the use of an ontology for the identification of degrees of freedom in urban wastewater systems. Water science and technology: a journal of the International Association on Water Pollution Research, 55(4)(2007) 155

[Krallinger et al., 2010] Krallinger, M., Leitner, F., and Valencia, A. Analysis of biological processes and diseases using text mining approaches. Methods in Molecular Biology, 593(2010) 341–382.

[Limpiyakorn et al., 2006] Limpiyakorn, T., Kurisu, F., and Yagi, O. Development and application of real-time pcr for quantification of specific ammonia-oxidizing bacteria in activated sludge of sewage treatment systems. Applied microbiology and biotechnology, 72(5)(2006) 1004–1013.

[Nogueira and Melo, 2006] Nogueira, R. and Melo, L. Competition between nitrospira spp. and nitrobacter spp. in nitrite-oxidizing bioreactors. Biotechnology and bioengineering, 95(1)(2006) 169–175.

[Nogueira et al., 2002] Nogueira, R., Melo, L., Purkhold, U., Wuertz, S., and Wagner, M. Nitrifying and heterotrophic population dynamics in biofilm reactors: effects of hydraulic retention time and the presence of organic carbon. Water research, 36(2)(2002) 469–481.

[Tamames and De Lorenzo, 2010] Tamames, J. and De Lorenzo, V. Envmine: A text-mining system for the automatic extraction of contextual information. BMC bioinformatics, 11(1)(2010) 294.