pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210589

A NLP Pipeline for the Automatic Extraction of Microorganisms Names from Microbiological Notes

Sara MORA[a, 1], Jacopo ATTENE[a], Roberta GAZZARATA[b], Giustino PARRUTI[c] and Mauro GIACOMINI[a]
[a] *Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy*
[b] *Healthropy, Corso Italia 15/6, 17100 Savona, Italy*
[c] *Department of Infectious Diseases, AUSL Pescara, Pescara, Italy*

Abstract. According to the "Istituto Superiore di Sanita`" (ISS), hospital infections are the most frequent and serious complication of health care. This constitutes a real health emergency which requires incisive and joint action at all levels of the local and national health organization. Most of the valuable information related to the presence of a specific microorganism in the blood are written into the notes field of the laboratory exams results. The main objective of this work is to build a Natural Language Processing (NLP) pipeline for the automatic extraction of the names of microorganisms present in the clinical texts. A sample of 499 microbiological notes have been analysed with the developed system and all the microorganisms names have been extracted correctly, according to the labels given by the expert.

Keywords. Microbiological infections, natural language processing, automatic extraction, standard coding system, laboratory information systems

1. Introduction

After the beginning of the well-known worldwide pandemic of COVID-19, it has become even more evident that hospitals, assisted residences and shelters for the elderly nowadays represent areas where the circulation of pathogenic microorganisms is increasingly worrying and widespread. According to the *"Istituto Superiore di Sanità"* (ISS) [1], hospital infections are the most frequent and serious complication of health care. They can be defined as infections that arose during hospitalization or after the patient's discharge, which at the time of admission were not clinically manifest, nor were they incubating. This constitutes a real health emergency which requires incisive and joint action at all levels of local and national health organization. The goal is to uniformly activate stable and automatic systems of reporting and epidemiological surveillance capable of promptly identifying pathogenic microorganisms, multi-resistant or not, responsible for infections and to allow the immediate adoption of specific control measures.

[1] Corresponding Author: Sara Mora, Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy; E-mail: sara.mora@edu.unige.it

This kind of information can be found into the results of laboratory exams, in particular when microbiological cultures are executed on the blood sample. It is more than 10 years that laboratory analyses have been managed in a computerized manner through the use of *Laboratory Information Systems* (LISs). The non-contemporary and highly localized development of this type of systems led each center to create its own vocabulary for the coding of individual laboratory analyses. In order to make the results coming out of the single laboratories comparable, appropriate standard systems have been devised for the management of terminology, for example the *Common Terminology Services re- lease 2* (CTS2) [2], whose specifications derive from the synergy between the *Object Management Group* (OMG) [3] and *Health Level 7* (HL7) [4].

The main problem that computerized systems have faced is precisely the management of microbiology. This is because it is a highly variable discipline and linked to the habits of individual laboratories (for example which coding system is used for the nomenclature of bacteria, how sensitivity analyses are performed, ...). Therefore, the management of this type of laboratory analysis should be more varied. On the other hand, national laws imposed the mandatory use of LISs, whose structure in some specific cases may be too stiff. This problem has been overcome during the years with the simple trick of writing lots of natural text in the clinical notes, probably because clinicians could not find more appropriate fields for that kind of information. So, clinical notes became a great source of valuable information both for patient care and biomedical research, but they require manual inspection which is very expensive from an effort and timing point of view.

To tackle the problem, artificial intelligence tools can be used such as *Natural Language Processing* (NLP), a branch of computer science that deals with the interactions between computers and natural human language; it studies the problems connected with the automatic generation and understanding of human language, written or spoken [5,6,7].

The goal of this work is to create a pipeline for the automatic extraction of specific information from microbiological reports through the use of NLP techniques. This work is a first step for the development of a hospital and territorial antibiotic prescription monitoring system in the Abruzzo Region [8].

2. Materials & Methods

2.1. Characteristics of the Sample

The clinical notes used in this study are extracted from 1 month of anonymized laboratory referral from the main hospital of Pescara in Abruzzo Region. The sample is composed by 499 texts, 276 of them containing the name of a microorganism. The presence of the specific microorganism was confirmed by an expert from the hospital.

2.2. Environment & Libraries

The pipeline was completely developed in Python and the environment used is Jupyter Notebook. The libraries used for the specific tasks of this project are:

1. *Pandas*: it is a Python library containing open source data analysis and manipulation tools [9].

2. *Natural Language Toolkit* (NLTK): it is the most used library to perform text analysis in multiple languages, in fact it is very popular in academia and for research [10]. Some examples of the supported operations are: tokenization, stemming, part of speech tagging and disambiguation.
3. *SpaCy*: is an open source library for NLP in Python. It supports different languages and it is particularly suitable for the creation of software applications in- tended for production.
4. *FuzzyWuzzy*: is a Python library that supports the comparison of strings with each other. In particular, its main functions compute the distance in different cases: strings with the same length or not, taking into account the order in which the words are arranged and how many times a string can be repeated. The comparison between strings is based on the Levenshtein distance:

$$lev_{a,b}(i,j) = \begin{cases} max(i,j), & \text{if min(i,j) = 0} \\ min \begin{cases} lev_{a,b}(i-1,j)+1, \\ lev_{a,b}(i,j-1)+1, \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq a_j)} \end{cases} & \text{otherwise} \end{cases}$$

where i and j are the indexes of the last character of the substring.

2.3. Steps of the Pipeline

1. **Vocabulary building:** a vocabulary was created containing the names of microorganisms (bacteria, fungi, yeasts, viruses) respecting the current taxonomic subdivision proposed by Carl Woase in 1990. Together with the name of the family, genus and species, the microorganism has been mapped into 3 standard coding systems, at national or international level: *Italian Clinical Microbiologists Association* (AMCLI), *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMEDCT) and *National Healthcare Safety Network* (NHSN).
2. **Data acquisition:** the data were imported together with the vocabulary through the use of pandas library.
3. **Tokenization, stopwords removal and genus extension:** in this phase, clinical notes have been divided into tokens. Then stopwords longer than 1 character have been deleted in order not to compare strings that are prepositions, articles and adverbs.

   Starting from the Linnaeus classification [11] the binomial nomenclature is used. It is formed by the name of the genus with the first capital letter and the name of the species in lowercase. Often, after a species name is introduced in a text, the genus is abbreviated to the first letter in subsequent mentions (followed by a fullstop). Unfortunately, however, since the notes are very short, it is a shared agreement to always use the abbreviated form, even without having first specified the entire genus once. In this binomial nomenclature, the use of a two-letter abbreviation for the genus has not been introduced. So, words composed by only one character haven't been deleted.

This check could be done also through the use of regular expressions. However, this choice was made in order to consider that abbreviations could be spelled incorrectly. For example abbreviations not followed by a fullstop or letters followed by a fullstop but lowercase. The last step of this phase was the extension of the microorganism genus, in particular the "n+1" token have been compared with each species of the vocabulary. If the two tokens had a very high similarity wuzzy index (greater than or equal to 98) the token "n" was checked. In the event that the token "n" began with the same letter of the genus belonging to the species found in position "n+1", the extension of the genus was made.

4. **Extraction of the desired microorganism from the clinical notes:** initially, an attempt was made to carry out a morphological and lexical analysis. However this approach did not produce any good results due to the lack of morphological structure of the reports. The extraction of the microorganism was done by comparing the tokens present within the report and the vocabulary using the library FuzzyWuzzy. As regards the extraction of the genus, the threshold on the similarity index is set at 75, while the threshold for the species is set at 85 (they were more frequently written correctly).

3. Results

3.1. Genus Extension

There were 107 abbreviated genera followed by species in the notes. Once all the notes have been elaborated by the system the abbreviation extension from all 107 genera have been extended. The extended genera completely matched with the indications of the expert.

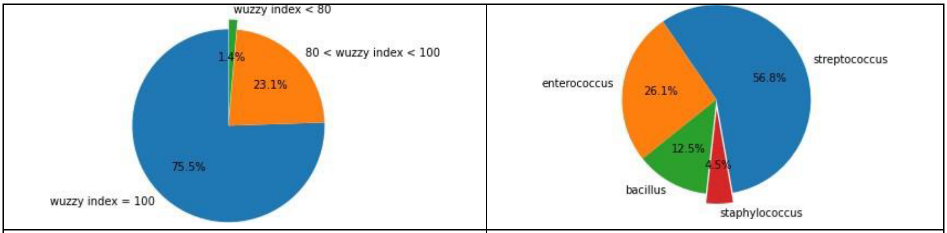
3.2. Microorganism Detection

The total number of available clinical notes was 499, 276 of them actually contained the name of a microorganism while 223 did not contain any microorganisms name. Two tests were carried out:

1. First all 499 notes were introduced into the microorganism extraction module.
2. Then only the notes that actually contained the microorganisms were introduced.

In both cases the system extracted all the microorganisms names, this suggests that it is not necessary a pre-processing phase that filters the data in some way. In particular 416 genera of microorganisms were found, most of them (321) with a wuzzy index of 100, also thanks to the process genus extension.

As can be seen in figure 1b below, the microorganism with the lowest score is Staphylococcus. If a species is not specified, Staphylococcus tends to have a very low similarity index, between 76 and 80. This is because Staphylococcus and 'stafilococco/stafilococchi' (term referable to the Staphylococcus microorganism that is tran- scribed in the notes) have respectively 14 and 12 letters. In addition, only 9 letters coincide, so they have a Levesthein distance of 5, as it takes 5 changes to transform the first word into the other. This is due to the fact that the Staphylococcus is one of the most widespread bacteria, therefore it is common to mention it in the natural discourse and therefore it is frequent to find the Italian term and not only the strictly scientific term in the notes.



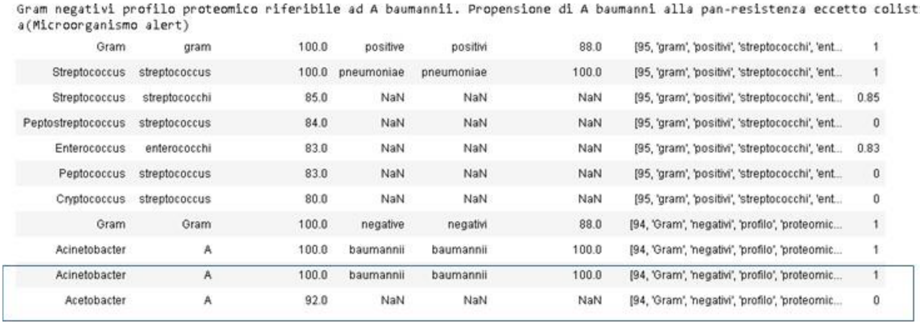
a) Wuzzy indexes percentage distribution of genera.

b) Percentage distribution of genera with lower index.

Figure 1. System performance on genera extraction.

Species had a wuzzy index always greater than 88.

Finally, a weight parameter was introduced. It was a decimal value, between 0 and 1, associated with each couple composed by genus and species, or only with the genus if present alone, in order to highlight the maximum wuzzy indexes. This because the same word (genus and/or species) could be associated with more than one genus/species. For example, if the genera *Acetobacter* and *Acinetobacter* were compared, these would have a similarity index of 92, therefore quite high. In order to confirm the selected genus the following token was compared to the species of that specific genus in the vocabulary. If a matching was found (wuzzy index over 98) than that genus assumed weight equal to 1 and the other 0.



```
Gram negativi profilo proteomico riferibile ad A baumannii. Propensione di A baumanni alla pan-resistenza eccetto colistin
a(Microorganismo alert)
```

Gram	gram	100.0	positive	positivi	88.0	[95, 'gram', 'positivi', 'streptococchi', 'ent...	1
Streptococcus	streptococcus	100.0	pneumoniae	pneumoniae	100.0	[95, 'gram', 'positivi', 'streptococchi', 'ent...	1
Streptococcus	streptococchi	85.0	NaN	NaN	NaN	[95, 'gram', 'positivi', 'streptococchi', 'ent...	0.85
Peptostreptococcus	streptococcus	84.0	NaN	NaN	NaN	[95, 'gram', 'positivi', 'streptococchi', 'ent...	0
Enterococcus	enterococchi	83.0	NaN	NaN	NaN	[95, 'gram', 'positivi', 'streptococchi', 'ent...	0.83
Peptococcus	streptococcus	83.0	NaN	NaN	NaN	[95, 'gram', 'positivi', 'streptococchi', 'ent...	0
Cryptococcus	streptococcus	80.0	NaN	NaN	NaN	[95, 'gram', 'positivi', 'streptococchi', 'ent...	0
Gram	Gram	100.0	negative	negativi	88.0	[94, 'Gram', 'negativi', 'profilo', 'proteomic...	1
Acinetobacter	A	100.0	baumannii	baumannii	100.0	[94, 'Gram', 'negativi', 'profilo', 'proteomic...	1
Acinetobacter	A	100.0	baumannii	baumannii	100.0	[94, 'Gram', 'negativi', 'profilo', 'proteomic...	1
Acetobacter	A	92.0	NaN	NaN	NaN	[94, 'Gram', 'negativi', 'profilo', 'proteomic...	0

Figure 2. Columns are: genus, word in text matching the genus, genus wuzzy index, species, word in text matching the species, species wuzzy index, tokenized clinical note, weight. In the upper part of the figure the full text of the highlighted clinical note is available.

Otherwise, if there was no species present and the two words had the same wuzzy index, for example due to a spelling error, then the algorithm would return both genera but with a weight of 0.5.

4. Discussion

In general the genus extension led to good results in the phase of microorganisms names extraction compared to the labels given by the expert. Anyway some ambiguities can be found during this phase. In fact, there are several microorganisms that have identical species and the genus begins with the same letter. Among these it is worth mentioning the *intermedius* species as it is the most probable among the ambiguous ones. It has both *Streptococcus* and *Staphylococcus* as genus. *Staphylococcus intermedius* is a very rare human pathogen. There are very few cases in the literature describing *S. intermedius* as a cause of infection in humans. Most of these cases have been described in association with exposure to animals, mainly dogs. While *Streptococcus intermedius* is the major cause of brain abscesses (70%). In Italy, however, there are very few cases of brain abscesses per year, in fact the incidence is less than 0.1% per year. In the event that an extreme case similar to the one mentioned above appears, the clinical note will be duplicated and both the microorganisms will be extracted, but the weight of the single couple is 0.5.

5. Conclusions

The main objective of this work was building a NLP pipeline that supported the automatic extraction of the names of microorganisms contained in microbiological notes. All the microorganisms present were extracted correctly, so the main goal was achieved. The next step of the aforementioned project will be to proceed with the automatic extraction of the antibiotic prescription from the same clinical notes. In particular, a key information will be the specific sensitivity of the microorganism to each single antibiotic tested.

Finally, being based on international nomenclature standards, this pipeline can be applied with microbiology notes in Italian from hospitals of all over the national territory.

References

[1] https://www.epicentro.iss.it/

[2] Gazzarata R, et al. *A terminology service compliant to CTS2 to manage semantics within the regional HIE.*, European Journal of Biomedical Informatics 13.1 (2017).

[3] https://www.omg.org/

[4] https://www.hl7.org/

[5] Matheny, Michael E., et al. *Detection of blood culture bacterial contamination using natural language processing.*, AMIA Annual Symposium Proceedings. Vol. 2009. American Medical Informatics Association, 2009.

[6] Maganti N, et al. *Natural language processing to quantify microbial keratitis measurements.*, Ophthalmology 126.12 (2019): 1722-1724.

[7] Fu S, et al. *Automated detection of periprosthetic joint infections and data elements using natural language processing.*, The Journal of Arthroplasty 36.2 (2021): 688-692.

[8] Gazzarata R, et al. *A SOA based solution for MDRO surveillance and improved antibiotic pre- scription in the Abruzzo region.*, pHealth 2019. IOS Press, 2019. 49-54.

[9] https://pandas.pydata.org/

[10] Bird S. *NLTK: the natural language toolkit.*, Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. 2006.

[11] Linné Cv. *Systema naturae.* Ed 10 (1758): 551.