# IMS-UNIPD @ CLEF eHealth Task 1: A Memory Based Reproducible Baseline

Giorgio Maria Di Nunzio<sup>1,2</sup>

<sup>1</sup>Department of Information Engineering, University of Padova, Italy <sup>2</sup>Department of Mathematics, University of Padova, Italy

#### Abstract

In this paper, we report the results of our participation to the CLEF eHealth 2021 Task on "Multilingual Information Extraction". This year, this task focuses on Named Entity Recognition from Spanish clinical text in the domain of radiology reports. In particular, the main objective is to classify entities into seven different classes as well as hedge cues.

Our main contribution can be summarized as follows: 1) continue the study of minimal/reproducible pipeline for text analysis baselines using a tidyverse approach in the R language; 2) evaluate the simplest memory based classifiers without optimization.

### **Keywords**

classification, memory based classifier, R tidyverse

# 1. Introduction

In this paper, we report the results of our participation to the CLEF eHealth [1] Task 1 "Multilingual Information Extraction" [2]. The 2021 task focuses on the Named Entity classification of clinical textual data which consists in 513 ultrasonography reports. The unstructured text and the number of orthographic and grammatical errors make this tasks challenging for automatic approaches.

The contribution of our experiments to this task can be summarized as follows:

- The implementation of a reproducible pipeline for text analysis;
- An evaluation of a simple memory based classifier.

The remainder of the paper will introduce the methodology and a brief summary of the experimental settings that we used in order to create the run that we submitted for the task.

# 2. Method

In this section, we summarize the pipeline for text pre-processing which has been developed in the last three years [3, 4, 5] and has been simplified and the source code will be made available.<sup>1</sup>

CLEF 2021 - Conference and Labs of the Evaluation Forum, September 21-24, 2021, Bucharest, Romania

<sup>☆</sup> giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

http://github.com/gmdn (G. M. Di Nunzio)

D 0000-0001-9709-6392 (G. M. Di Nunzio)

<sup>© 0 2021</sup> Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CEUR Workshop Proceedings (CEUR-WS.org)

<sup>&</sup>lt;sup>1</sup>https://github.com/gmdn

## 2.1. Pipeline for Data Cleaning

In order to produce a dataset ready for training a classifier, we followed the same pipeline for data ingestion and preparation for all the experiments. We used the *tidytext* approach to automatically parse and extract the text.<sup>2</sup>

The following code summarizes the initial steps of the analysis of the documents:

With just two lines of code we separate each token of each document and extract the location in the text. As an additional example, with the following line, we tried to reduce the possibility to match smaller sequences by adding spaces around the text (even though we may lose some matches with this extra characters)

```
ta$text_lower[nchar(ta$text_lower) < 3] <- paste0(" ", short_text, " ")</pre>
```

### 2.2. Classification

The main idea of a memory based classifier follows the idea presented in [3]:

- Choose the morphological level (in our experiments token level);
- Given a (multi-word) token in a sentence, search for any previously classified documents that contains that sentence;
- Add the classification label to the document.

We built the rules for the memory based system by looking at all the documents provided in the training and validation set. No optimization was performed at any steps and only one run was submitted.

# 3. Preliminary Results

In this section, we briefly comment the official results sent by the organizers before the workshop.

<sup>&</sup>lt;sup>2</sup>https://bnosac.github.io/udpipe/en/index.html

### 3.1. Considerations before the official results

Our initial goal with this approach was to build a memory based approach that could capture with high precision (only known sequences) and low recall (all the sequences that are not previously seen are not recognized) some entities that were labeled by the experts in the training and validation set.

Without any optimization or evaluation on the validation set, our initial guess was a recall around 50% (we suppose that at least half of the entities of the test set are not in the training/-validation set), and a precision around 70-80% (when a sequence previously labeled is found, we suppose that there is a low chance that it is categorized wrongly).

### 3.2. Considerations after the official results

Compared to the same approach used in the past years, the results on this task were surprisingly low: on one hand, recall was around the figure we expected for most of the categories; on the other hand, precision was extremely low.

These results, despite being negative, open interesting questions about what went wrong in the implementation of rules of the classifier. In particular, we have started to analyze the runs for the first time since the runs were submitted, and we found some odd classifications of one or two characters entities. For example, for document 4901 we have

```
T5 Abbreviation 34 35 c
T6 Abbreviation 38 39 c
T7 Abbreviation 46 47 c
T8 Abbreviation 51 52 c
T9 Abbreviation 58 59 c
T10 Abbreviation 96 97 c
...
T212 Uncertainty 12 13 o
T213 Uncertainty 15 16 o
T214 Uncertainty 20 21 o
T215 Uncertainty 29 30 o
T216 Uncertainty 52 53 o
...
```

Therefore, we believe that in some parts of the source code (see for example the line in Section 2.1 where we try to find a solution for smaller sequences) we introduced errors that could and should be avoided.

### 3.3. Source code debugging

After a careful analysis of the code, we found that we unintentionally removed the trailing white spaces introduced for shorter character sequences. In particular, when we extract the pattern to find in the text, we also "squish" any multiple white spaces including the trailing

white spaces. In doing so, single characters like "m", "o", "s", are included as patterns; thus, the number of patterns wrognly assigned to each document increases and the precision for that category decreases.

For this reason, we modify the code in order to avoid this passage when shorter sequences are involved and run again the classification of the validation sets.

We provide the source code to replicate these results;<sup>3</sup> in particular, with this simple modification we could recover most of the precision initially lost. For example, for the category "Abbreviation", the initial precision around 7% increases up to 40%, for "Anatomical Entity" from 25-30% up to 50%, for "Negation" from 10% up to 60%.

# 4. Final remarks and Future Work

The aim of our participation to the CLEF 2021 eHealth Task 1 was to test the effectiveness of a simple textual pipeline implemented in R with the 'tidyverse' approach for the problem of classification of Spanish clinical textual data. A preliminary failure analysis showed an anomaly in the values of precision, too low compared to the expected effectiveness. After a careful analysis, we found a mistake in the source code and, after we fixed the error, the performances increased significantly. In order to make this study reproducible, we will make the source code available. Additional analyses will be carried out to find patterns that can be easily used as a kind of knowledge base to support more advanced systems. We will provide a finer analysis on the test set when the ground truth will be made available.

# References

- H. Suominen, L. Goeuriot, L. Kelly, L. A. Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, R. Upadhyay, J. Vivaldi, M. Viviani, C. Xu, Overview of the CLEF eHealth evaluation lab 2021, in: CLEF 2021 - 12th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, 2021.
- [2] V. Cotik, L. A. Alemany, F. Luque, R. Roller, H. Vivaldi, A. Ayach, F. Carranza, L. D. Francesca, A. Dellanzo, M. F. Urquiza, Overview of CLEF eHealth task 1 - spradie: A challenge on information extraction from spanish radiology reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR Workshop Proceedings, 2021.
- [3] G. Di Nunzio, As simple as possible: Using the R tidyverse for multilingual information extraction. IMS unipd ad CLEF ehealth 2020 task 1, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of CEUR Workshop Proceedings, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper\_137.pdf.
- [4] G. Di Nunzio, Classification of animal experiments: A reproducible study. IMS unipd at CLEF ehealth task 1, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland,

<sup>&</sup>lt;sup>3</sup>https://github.com/gmdn

September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper\_104.pdf.

[5] G. Di Nunzio, Classification of ICD10 codes with no resources but reproducible code. IMS unipd at CLEF ehealth task 1, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018. URL: http: //ceur-ws.org/Vol-2125/paper\_180.pdf.