Prototype of an Entity Recognition System for Antimicrobial Resistance Data Management

Francisco Prado-Valiño, Roi Santos-Ríos, Carlos Gómez-Rodríguez, and Jesús Vilares

Universidade da Coruña, Centro de Investigación CITIC - Grupo LYS, Facultad de Informática, Campus de Elviña, 15071 - A Coruña (España)
Correspondence: jesus.vilares@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.36*

Abstract: Often, a study or research process requires the analysis of large volumes of information in the form of unstructured text. This task consumes a large amount of time and resources of the human experts in charge of it. For this reason, there is great interest in developing automatic systems to support these activities by applying Text Mining techniques. A key task in this type of process is Entity Recognition: extracting the entities of interest contained in a text and classifying them into pre-established categories.
The work presented here is part of the GRALENIA project, which seeks to improve antimicrobial resistance data management in the hospital setting. The main goal of our work is the development of an entity recognizer prototype for the identification and labeling of indicators of interest present in clinical reports.

1 Introduction

In recent years, thanks to the progressive digital conversion of our health systems, the Biomedical field has experienced a significant increase in the amount of information available for research and analysis. In this context, electronic health records (EHR) constitute a major information source, since they contain a large amount and variety of information about our health (Fleuren and Alkema, 2015). However, although part of these data may be structured (laboratory tests, prescriptions, etc.), many other is formed by unstructured text (nursing notes, admission reports, etc.), making it difficult to analyze said data.

The study of the huge amount of information contained in these unstructured sources involves a great consumption of time and resources of trained personnel. Due to the complexity and cost of these analyses, there is great interest in developing support systems capable of automating these tasks as much as possible.

2 Main Goals

Our work consists of the application of Text Mining (TM) (Hotho et al., 2005) techniques for Entity Recognition (ER) in order to automate the detection of entities on unstructured clinical text. In our case, we will focus on the field of antimicrobial resistance (AMR), specifically on the resistance of bacteria to antibiotics. This work is part of the GRALENIA project,[1] which seeks to improve AMR data management in the hospital setting, specifically focusing on three bacteria: *MRSA*, *Klebsiella ESBL* and *Klebsiella carbapenemase*. For this purpose, the project proposes to

[1] *http://gralenia.es* (visited on Sept. 2023).

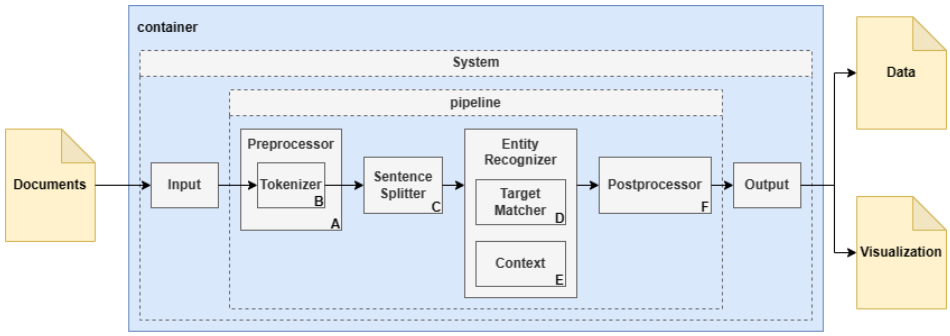

Figure 1: General architecture of the ER system: rule-based approaches.

create a digital platform that, using Natural Language Processing (NLP) techniques, analyzes patients' clinical reports in search of AMR-related indicators.

These indicators belong to one of the categories pre-defined by our experts: BACTEREMIA, CENTRAL NERVOUS SYSTEM (CNS) INFECTION, FEVER, GASTROENTERITIS, GENITAL INFECTION, ORAL INFECTION, PHLEBITIS, RESPIRATORY INFECTION, SEPSIS, SURGICAL WOUND INFECTION, and URINARY INFECTION. Within the framework of the parent project, and in a later phase, the detected indicators will serve as input features to an AI system for the detection of resistances and possible sources of bacterial contagion.

In this work, we intend to develop a prototype ER system to identify possible AMR indicators (named *symptomatic expressions*) in patients' clinical reports.

3 Development

At the time of writing this article, a suitable *dataset* for analysis and system evaluation in Spanish was not yet available. As a temporary solution, we decided to work with the MIMIC-III (Johnson et al., 2016) English database in the meantime, since the nature of its content adapts to the needs of our project. This database contains anonymized information on medical data belonging to more than 40,000 patients at *Beth Israel Deaconess Medical Center* (*BIDMC*).

SpaCy[2] open-source library has been used for building the core of our Natural Language Processing (NLP) system. SpaCy provides us with a framework in the form of a modular pipeline, this being the architecture that we will use for our ER system. As shown in Figure 1, it consists of the following modules:

1. **Tokenizer:** Segments the text into tokens. It also generates the *Doc* object on which the other SpaCy components will work.
2. **Preprocessor:** Preprocesses the tokenized text prior to creating the *Doc*.
3. **Sentence Splitter:** Segments the text into sentences, marking them in the *Doc*.
4. **Entity Recognizer:** Performs the ER task, properly speaking, on the *Doc*.
5. **Postprocessor:** Postprocesses the *Doc*.

3.1 First approach: Baseline rule-based system

Instead of generating the analysis rules of the system manually, we have automated their creation through code. In this way, it is enough to add, delete or modify any of the symptomatic expressions and the system will generate the new updated rules automatically. In addition,

[2] *http://spacy.io* (visited on Sept. 2023)

this allows us to easily add new categories by simply adding new sets. These rules are applied through a pattern matching process, labeling the detected category on the text, as can be seen in Example 1.

The following sentence, extracted from a MIMIC-III document, presents several indicators that we are interested in identifying and labeling. Specifically, it contains the symptomatic expressions *"abdominal pain"*, *"vomit"* and *"diarrhea"*, belonging to the category Gastroenteritis, along with *"dysuria"*, belonging to Urinary Infection. Once we apply the rules on this phrase, we obtain the following output:

"Denies specific complaints when asked, including chest pain, cough, ***abdominal pain GASTRO-*** ***ENTERITIS*** *, nausea and vomits,* ***diarrhea GASTROENTERITIS*** *, and* ***dysuria URINARY_*** ***INFECTION*** *."*

(Example 1)

Preprocessing

To facilitate establishing correspondences, those abbreviations and acronyms present in the input text are replaced by their components: *"rx'd"* for *"prescribed"* or *"tº"* for *"temperature"*, for instance. This process is carried out during the *Preprocessing* phase, before applying the rules.

Results

The results for this first approach (*base*) are shown in Table 1, which details the number of labeled entities (i.e. that matched some pattern), the number of documents with at least one label, and the percentage that these documents represent over the total. Total coverage was 41.01%, so there is still ample room for improvement.

3.2 Second approach: Fuzzy matching and term expansion

One of the main problems with our initial approach is the negative impact of spelling errors, since they prevent matching: *"fevre"* instead of *"fever"* or *"taychardia"* instead of *"tachycardia"*, for example. To reduce this problem as much as possible, we will introduce *fuzzy matching* mechanisms in the rules in order to make the patterns more flexible. Specifically, we will use the *Levenshtein editing distance* (Levenshtein et al., 1966) on the words contained in a symptomatic expression. Given two words, the Levenshtein edit distance is the minimum number of single-character edits (insertions, deletions, and substitutions) required to convert one word into the other.

Similarly, another of the main problems detected is the use of morphological variants (e.g. changing from singular to plural) and semantic variants (e.g. synonyms) of the original symptomatic expressions. As a solution, inflectional morphemes accounting for number variation (e.g. *"expectoration"*-*"expectorations"*) and verbal forms (e.g. *"vomit"*-*"vomiting"*) have been integrated. Synonyms commonly used in the application domain have also been added (e.g. *"secretion"*-*"discharge"*) after being obtained from the OpenMD online medical dictionary.[3]

Rules for complex symptomatic expressions

As explained before, our automatically generated rules provide flexibility when it comes to expanding the set of symptomatic expressions to be detected by the system but, at the same time, they are currently limited in terms of their complexity. Some of these expressions, such as measurements, present more complex structures than those initially proposed by experts and, then, automatically encoded into the rules. To define rules capable of correctly identifying

[3] *http://openmd.com/dictionary/* (visited on Sept. 2023).

and labeling this kind of complex structures, we have chosen to manually refine the patterns involved. Example 2 includes one of these complex expressions.

> In order to label expressions related to blood pressure along with their associated numerical value, we defined *ad hoc* rules with more elaborate patterns. These were based on a list of dictionaries containing (*characteristic*, *value*) pairs. Once these new rules are applied, the following output is obtained:
>
> *"His* ***blood pressure was 223/125 SEPSIS*** *, otherwise Hemodialysis stable."*
>
> (Example 2)

Postprocessing

By allowing matches with words within edit distance 1 of those considered within a rule, it is inevitable that expressions containing correct terms that are very similar to those of a pattern, may be erroneously labeled as entities (e.g. *"fever"-"never"*).

To avoid this, those entities identified by the rules are then checked by the system using a dictionary of the language from which we have previously removed our symptomatic expressions. Furthermore, frequent erroneous expressions detected in previous analyses of the system output have been added, as in the case of that one shown in Example 3.

> *"We have* ***not found SURGICAL_WOUND_INFECTION*** *a particular bacteria that is causing your recurrent infections."*
>
> (Example 3)

Results

As we can see in Table 1, the coverage for our second approximation (*fuzzy*) has increased to 58.41%, with the categories FEVER and RESPIRATORY INFECTION standing out. The use of fuzzy matching rules now make it possible to capture symptomatic expressions of these categories containing variants and spelling errors. Simultaneously, other categories, such as CNS INFECTION and GASTRENTERITIS, have lost labeled entities, but this is because they were redistributed to other categories. GENITAL INFECTION continues to be the category with the fewest entities, due to the small number of cases of this type in the corpus.

3.3 Third approach: Transformers

Despite the limitations described, the performance achieved through the previous improvements made it possible to generate a set of automatically labeled examples of sufficient size. The availability of this new *silver corpus*, enabled us to implement a new approach based on *transformers* (Vaswani et al., 2017). To do this, we have maintained our *pipeline* architecture, although modifying some of its modules. Specifically, the *Sentence Splitter* and *Entity Recognizer* modules have been removed, since they are not necessary in this new approach, while otheres have been introduced, as shown in Figure 2:

- **Transformer**: Contains the pretrained *language model* used by the system to obtain highly contextualized *embeddings* of the input words.
- **Perceptron**: Responsible for assigning labels from the output of the previous module.
- **Remove Transformer Data**: Responsible for optimizing the memory consumption by removing unnecessary data from the *embeddings*.

Table 1: Analysis of results for the base system (*base*) and its improved version (*fuzzy*) using fuzzy correspondence and term expansion.

Entity type	**No. Tagged Entities**		**No. Documents with tags**		**% Documents with tags**	
	base	*fuzzy*	*base*	*fuzzy*	*base*	*fuzzy*
Bacteremia	6,016	7,488	4,069	4,704	6.80	7.86
CNS Infection	17,999	11,464	9,656	6,426	16.64	10.74
Fever	4,054	34,638	2,628	15,954	4.39	26.67
Gastroenteritis	13,630	10,483	6,730	5,798	11.25	9.69
Genital Infection	3	9	3	9	0.01	0.02
Oral Infection	991	1,108	771	852	1.29	1.42
Phlebitis	100	269	80	231	0.13	0.39
Respiratory Infection	997	29,854	946	15,585	1.58	26.05
Sepsis	31,347	36,045	15,549	16,958	25.99	28.35
Surgical Wound Infection	26	3,334	25	2,374	0.04	3.97
Urinary Infection	1,246	1,298	1,044	1,088	1.75	1.82
Total	76,409	135,990	24,534	34,945	41.01	58.41

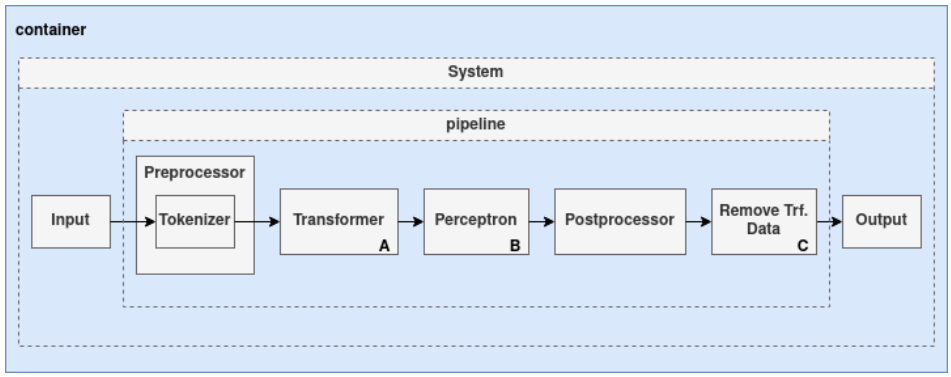

Figure 2: General architecture of the ER system: *transformer*-based approach.

Language models and training

As previously mentioned, we have used the output of our improved rule-based system to automatically generate a *silver corpus*, taking as training examples those sentences that contained labeled cues. Next, the resulting examples set was shuffled in order to balance its distribution and, then, splitted into three (sub)sets: 70% for training, 20% for validation and 10% for testing purposes.

When analyzing the performance of this last approach, several language models have been considered:[4]

- **DistilBERT** (Sanh et al., 2019): *Distilled* version of the base BERT model Devlin et al. (2018). It is smaller (reduced from 100M to 67M parameters) and faster (60%) than the original model, while maintaining a similar linguistic understanding capacity.
- **Bio-DistilBERT** (Rohanian et al., 2022): Model obtained by training DistilBERT on the PubMed dataset[5] in *batches* of size 192 for 200k training cycles.

[4] Downloaded from HuggingFace: *http://huggingface.co* (visited on Sept. 2023).
[5] *http://pubmed.ncbi.nlm.nih.gov* (visited on Sept. 2023).

Table 2: Global comparison of the outputs for our different approaches.

Approach	No. Tagged Entities	No. Documents with tag	% Documents with tag
Base rules	76,409	24,534	41.01
Improved rules	135,990	34,945	58.41
DistilBERT	170,503	41,456	69.30
Bio-DistilBERT	183,968	**44,148**	**73.80**
Bio-ClinicalBERT	166,081	39,386	65.84
PubMedBERT	**192,652**	42,860	71.65

- **Bio-ClinicalBERT** (Alsentzer et al., 2019): Model obtained by training BioBert (Lee et al., 2020) on the MIMIC-III dataset in *batches* of size 32 for 150k cycles.
- **PubMedBERT** (Gu et al., 2020): Model obtained by training BERT on the PubMed dataset in *batches* of size 8,192 for 62.5k cycles.

Systems Comparison

As shown in Table 2, we can see a significant increase in performance with respect to our previous rule-based approaches. In the case of our *transformer*-based approach, the PubMedBERT model obtained a higher number of labeled entities, while Bio-DistilBERT achieved higher document coverage. This increase in performance is mainly due to the generalization capacity of the language models with respect to the use of rules, which is a great advantage when working with unstructured free text content, as in the case of the documents managed by our system.

Regarding the type of symptomatic expressions detected, the new *transformer*-based system is able to identify expressions with large variations and errors, although it might also identify words incorrectly, thus resulting in incorrect labelling, as shown in Example 4. However, some of these incorrectly labeled expressions may also be of interest to the task, thus making the work of annotators easier for the construction of a future *gold standard*: replacing its original label by the correct one should be enough.

> For the symptomatic expression *"hypotension"*, with category SEPSIS, the PubMedBERT model also allows variants such as *"hypotensive"* to be identified. However, it also incorrectly labels terms such as *"hyponatremia"* or *"hypoglycemia"*, since they also contain the prefix *"hypo-"*.
>
> (Example 4)

Acknowledgements

This work was funded by the GRALENIA project (Ref. 2021/C005/00150055) supported by the Spanish Ministry of Economic Affairs and Digital Transformation, the Spanish Secretariat of State for Digitization and Artificial Intelligence, Red.es and by the NextGenerationEU funding. We also acknowledge the European Research Council (ERC), that has also partially funded this research under the Horizon Europe research and innovation programme (SALSA, grant agreement No. 101100615), ERDF/MICINN-AEI (SCANNER-UDC, Ref. PID2020-113230RB-C21), Xunta de Galicia (ED431C 2020/11), and Centro de Investigación del Sistema Universitario de Galicia (SUG) CITIC, funded through the collaboration agreement for the reinforcement of the research centres of the Galician University System (CIGUS) between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities.

Bibliography

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly available clinical bert embeddings, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

W. W. Fleuren and W. Alkema. Application of text mining in the biomedical domain. *Methods*, 74:97–106, 2015. Text mining of biomedical literature.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

A. Hotho, A. Nürnberger, and G. Paas. A brief survey of text mining. *Journal for Language Technology and Computational Linguistics*, 20(1):19–62, 2005.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

V. I. Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.

O. Rohanian, M. Nouriborji, S. Kouchaki, and D. A. Clifton. On the effectiveness of compact biomedical transformers, 2022.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.