Polimi at CLinkaRT: a Conditional Random Field vs a BERT-based approach

Vittorio Torri[1], Francesca Ieva[1,2]

[1]MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy
[2]HDS - Health Data Science Centre, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, Italy

Abstract

In the context of the EVALITA 2023 challenge, we present the models we have developed for the CLinkaRT task, which aims to identify medical examinations and their corresponding results in Italian clinical documents. We propose two distinct approaches: one utilising a Conditional Random Field (CRF), a probabilistic graphical model traditionally used for Named Entity Recognition, and the other based on BERT, the transformer-based model that is currently state-of-the-art for many Natural Language Processing tasks. Both models incorporate external knowledge from publicly available medical resources and are enhanced with heuristic rules to establish associations between exams and results. Our comparative analysis elects the CRF-based model as the winner, securing the third position in the competition ranking, but the BERT-based model demonstrated competitive performance.

Keywords

Natural Language Processing, Named Entity Recognition, Clinical documents,

1. Introduction

The widespread adoption of Electronic Health Records (EHR) has led to a significant transformation in healthcare data collection, allowing for the accumulation of extensive patient information. However, a considerable portion of this data remains unstructured, posing challenges to its utilisation in statistical analyses. Within EHR systems, vast amounts of textual data, such as clinical notes, reports, and discharge summaries, are stored, containing valuable patient history that often lacks in traditional databases. In recent years, Natural Language Processing (NLP) advancements have opened up possibilities for extracting structured data from text. However, numerous challenges persist, particularly in specialised domains like medicine [1] and when dealing with languages other than English [2].

This paper presents the models we have developed for the CLinkaRT task [3] as part of the EVALITA 2023 challenge [4]. The task entails identifying pairs of medical examinations and their corresponding results within Italian clinical documents. To accomplish this, a subset of the Italian section of the E3C corpus [5], annotated by the task organisers, was provided as the training set.

Our first system is based on a Conditional Random Field (CRF), a probabilistic graphical model that has been widely used for Named Entity Recognition (NER) [6]. NER is an NLP task that involves identifying and categorizing specific types of entities within a given text. In our case, we apply NER to recognize examination names and their corresponding results. This model is enhanced by incorporating external knowledge from additional resources and employing rules to associate each examination with its result. We compare it with an approach based on BERT, the more recent transformer-based neural network that is currently state-of-the-art for many NLP tasks [7]. In this case, we fine-tune the latest Italian version of BERT, Umberto [8], using the E3C corpus. To exploit the entire corpus, we automatically translated documents that are in languages other than Italian. Subsequently, this fine-tuned BERT model undergoes training for token classification using the annotated training set provided for the challenge, incorporating a linear classification layer.

Both models demonstrated discrete performances in the NER tasks of identifying examinations and results, while the figures were lower for the actual CLinkaRT task, which involves associating examinations with their corresponding results. The CRF-based model achieved the best results, particularly due to higher recall on examination names and higher precision on examination results, achieving the third position in the final ranking.

The code of our models is available on GitHub[1].

The rest of this paper is structured as follows: Section 2 provides an analysis of related works, Section 3 discusses the dataset used for the task, Section 4 presents a detailed description of our system, Section 5 reports the obtained results, and Section 6 provides a comprehensive discussion on our findings.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
vittorio.torri@polimi.it (V. Torri); francesca.ieva@polimi.it (F. Ieva)
© 2023 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)

[1]https://github.com/vittot/CLinkaRT-2023-Polimi

2. Related works

The application of Natural Language Processing (NLP) techniques to Italian medical documents has been relatively limited. However, a few studies have addressed tasks relevant to this challenge. Viani et al. [9] focused on identifying various entities within Italian cardiology reports, including specific examination results and drug prescriptions. Their approach involved a pipeline utilising dictionary lookup and an ontology with regular expressions linked to concepts. They developed and evaluated their methodology using a dataset of 5400 reports. In a subsequent study [10], a supervised learning approach based on recurrent neural networks was employed to extract events from a smaller dataset of 75 cardiology reports, encompassing 4300 event occurrences.

Chiaramello et al. [11] explored the mapping between relevant terms in Italian clinical notes and concepts in the Italian version of the Unified Medical Language System (UMLS) [12], including the use of the MetaMap tool [13] on Italian documents.

Another example of NER on Italian clinical data, based on recurrent neural network architecture, is [14], even if the goal, in this case, was the de-identification of clinical notes and not information extraction.

While the number of works specifically focusing on Italian documents remains limited, a more extensive body of literature exists concerning English documents. These studies predominantly employ rule-based and dictionary lookup approaches [15], conditional random fields [16], recurrent neural networks [17] and, more recently, transformer-based neural networks [18].

3. Data

The training set provided by the task organisers consists of 83 documents extracted from the Italian subset of the E3C corpus. These documents have been annotated with pairs of examination mentions and corresponding results. In particular, there are 658 pairs in the dataset, among which there are 367 unique examination names and 395 unique examination values.

The challenge ranking is based on the performance of the models on a test set consisting of 80 documents. The test set was initially released to participants without annotations.

The documents in the E3C corpus are clinical narratives originating from different sources: journal papers, admission tests for specialities in medicine, patient information leaflets for medicines, and abstracts of theses in medical science.

The CLinkaRT task poses several difficulties due to the heterogeneity of the documents and the small size of the training set. Previous works in related areas often had access to datasets comprising thousands of documents or annotations, typically from a single source and within a specific medical domain (e.g., cardiology). In our case, the documents can cover any medical area, and the concept of examinations and their results have to be intended in a broad sense.

Table 1 provides examples of annotated sentences from the training set. Sentence #1 has been annotated with two examinations: *"fluenza"* (*"fluency"*) and *"memoria"* (*"memory"*), both with the value *"valori ai limiti della norma"* (*"values at normal limits"*). This example demonstrates that the task involves not only identifying laboratory examinations with precise numerical results but also encompasses various types of examinations where results can be expressed qualitatively. Sentence #2 has been annotated as containing an examination *"calo"* (*"loss"*) with a result of *"4 kg circa"* (*"about 4 kg"*), although it can be debated whether this qualifies as an examination.

Another element of uncertainty relates to the annotations boundaries, particularly for examination results. For instance, Sentence #3 has been annotated as having the result *"della positività"* (*"of the positivity"*) for the examination *"asCa"* while the proposition *"della"* (*"of the"*) could have been excluded from the result.

It is important to note that no specific annotation guidelines have been released, at least at the present time.

The complexities arising from document heterogeneity, different possible interpretations of examination results, and the absence of comprehensive annotation guidelines highlight the challenges involved in the CLinkaRT task.

4. Description of the system

We decomposed the task problem into three subproblems:

1. NER of examination names
2. NER of examination results
3. Linking between examination names and results

For the NER subproblems, we propose the two alternative approaches of CRF and BERT in Subsection 4.1 and 4.2, respectively, while for the linking, we propose an approach based on heuristic rules in Subsection 4.3

4.1. CRF model

The primary model we developed and used for the results submission is a Conditional Random Field (CRF). A Conditional Random Field is an undirected probabilistic graphical model widely used for Named Entity Recognition (NER). The model's random variables are divided between the observed variables $\mathbf{X}$ and the output variables $\mathbf{Y}$, and the graph models the conditional probability

ID	Sentence (ITA)	Sentence (ENG)	Annotations (ITA)	Annotations (ENG)
1	La valutazione neuropsicologica ha evidenziato deficit della memoria verbale a breve e a lungo termine, della memoria di prosa e delle funzioni prassiche e valori ai limiti della norma per la memoria visuo-spaziale e per la fluenza verbale	Neuropsychological evaluation showed deficits in short- and long-term verbal memory, prose memory, and praxic functions, and values at normal limits for visuospatial memory and verbal fluency	(fluenza,valori ai limiti nella norma) (memoria,valori ai limiti nella norma)	(fluency, values at normal limits) (memory, values at normal limits)
2	Il ragazzo manifestava da circa una settimana vomiti ripetuti accompagnati da coliche addominali, inappetenza e vistoso calo ponderale (4 kg circa in una settimana)	The boy had been experiencing repeated vomiting accompanied by abdominal colic, loss of appetite, and significant weight loss (about 4 kg in a week)	(calo, 4 kg circa)	(loss, about 4 kg)
3	Alla luce della positività degli asCa	In light of the positivity of the asCa	(asCa, della positività)	(asCa, of the positivity)

Table 1
Examples of annotated documents from the training set

$P(\mathbf{Y}|\mathbf{X})$. This conditional probability is modelled as

$$P(\mathbf{y}|\mathbf{x}) \propto exp(\sum_j \lambda_j \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i))$$

where $\mathbf{x}$ is the vector of tokens (observations) that form the sequence, $\mathbf{y}$ is the vector of labels (states) over the tokens, i is an index over the sequence tokens, n is the length of the sequence, j indexes the feature functions f_j and λ_j are the parameters to be learnt. Multiple feature functions f_j can be defined, both as state feature functions or as transition feature functions. While the first ones depend on the current label y_i and on the observed sequence $\mathbf{x}$, the latter also depends on the previous label y_{i-1}.

This task has two types of entities: *examination names* and *examination results*. It is possible to use two distinct CRFs for the two types of entities or a single one, which might be preferable as it can leverage the information obtained from predicting an *examination name* label to predict an *examination result* label, and vice versa. We considered an extensive set of internal features for the CRF model, as listed in Table 2. Different combinations of them have been tested, but the best results have been achieved with the complete set of features.

All these features are computed on the current token, the previous, and the next token.

Additionally, we incorporated features related to external knowledge sources. The first source is the UMLS vocabulary. We translated each token in the training set to English and queried the English UMLS vocabulary to obtain the list of concepts corresponding to the token, with their associated semantic types. We considered a set of binary features for the presence of the 50 most relevant semantic types and a more restricted set of features only for the presence of three specific semantic types (*Laboratory or Test Result*, *Laboratory Procedure*, *Amino Acid, Peptide, or Protein*) that are most likely associated with examination names, particularly for laboratory examinations.

The second external knowledge base we used is the official medical procedures nomenclature in Lombardy Region[2]. It is a list containing the names of all medical procedures provided by the Regional Health System in Lombardy. We considered only the categories primarily related to examinations: *Anatomy-Pathological Histology-Genetics*, *Immunohematology-Transfusion*, *Clinical Chemistry*, *Laboratory in general*, *Microbiology-Virology*. We extracted a binary feature indicating if a token is present in a processed version of this list, where we removed the most frequent words (frequency > 5).

The CRF was trained with the lbgfs gradient descent algorithm, 200 maximum iterations and regularisation coefficients $c_1 = 0.03$ and $c_2 = 0.02$.

4.2. BERT-based model

BERT is a transformer-based neural network that has achieved state-of-the-art performances in many NLP tasks. Although there are no domain-specific versions of BERT for the medical domain in Italian, there are general-domain versions, the most recent of which is Umberto.

[2]https://www.dati.lombardia.it/Sanit-/Transcodifica-Codici-prestazioni/7ugz-vcug

Feature	Details
Lowercased value of the current token	
Lemmatized lowered value of the current token	Lemmas are computed with Spacy Lemmatizer
Prefix of the current token	First three characters
Suffix of the current token	Last three characters
Upper token flag	True if the current token is uppercase
Title token flag	True if the current token is lowercase, beginning with an uppercase letter
Digit flag	True if the current token is composed of digits
Math symbol flag	True if the token is a mathematical symbol
Part of speech tag	Computed with Spacy POS Tagger
Exam abbreviation flag	True if it is an acronym of two or three letters present in one of the examination names mentioned in the training set

Table 2
Set of features considered for the CRF model (excluding those related to external knowledge sources)

	B-value	I-value	B-exam	I-exam
CRF				
Precision	0.8006 (0.107)	0.8025 (0.131)	0.6501 (0.143)	0.8127 (0.192)
Recall	0.6993 (0.121)	0.7435 (0.186)	0.3991 (0.103)	0.4398 (0.232)
F1-score	0.7424 (0.100)	0.7572 (0.147)	0.4867 (0.107)	0.5245 (0.198)
BERT				
Precision	0.7256 (0.110)	0.7826 (0.126)	0.6864 (0.077)	0.6823 (0.183)
Recall	0.7204 (0.094)	0.7779 (0.147)	0.3248 (0.076)	0.4720 (0.119)
F1-score	0.7176 (0.088)	0.7748 (0.120)	0.4354 (0.079)	0.5521 (0.130)

Table 3
10-fold CV results for NER (std dev in parenthesis)

We fine-tuned Umberto using the entire E3C corpus, including labelled and unlabelled documents in all E3C languages. Non-Italian documents were automatically translated into Italian using Google Translate's APIs. A linear token-level classification layer was added to this BERT version and trained on the annotated dataset provided for the challenge while keeping the other layers frozen.

Fine-tuning of the Umberto model over the E3C corpus involved 3 epochs of training with a learning rate of $2 \cdot 10^{-5}$ and weight decay of 0.01. The last layer was trained for 50 epochs with a learning rate of 10^{-3} and weight decay of 0.01.

4.3. Linking between exams and results

We employed the following heuristic rules to link pairs of examinations and results: each exam/result is paired with the nearest result/exam within the same sentence. If there are no available elements to pair with it, it is discarded.

5. Results

Table 3 reports 10-fold cross-validation results for both models. These results are related to the NER subtask only, and they are reported for the B-Exam, I-Exam, B-Value, and I-Value, even if there is no distinction between B and I tags in the annotations, to verify if longer entities show different performances. The NER results of the two models are comparable. They show higher precision, in particular for the examination names, for which the recall is very low. The CRF has higher precision than BERT on examination results and it has higher recall on examination names. These results are not surprising, given the limited amount of training data and the large number of possible examinations that can exist in this type of data. NER results on the test set are comparable to those obtained via cross-validation over the training set (we do not report them here due to space constraints).

The CLinkaRT task evaluation is based only on recognising pairs of examinations and results. Only the pairs that precisely matched the gold standard annotations were considered for ranking and evaluation. Precision, recall, and F1-score were computed based on this precise matching. The results on the final test set for both systems, computed with the official evaluation script, are shown in Table 5. Both are aligned with the NER results in terms of precision and lower in terms of recall. The CRF model results are the best, for both precision and recall.

A manual analysis of the results highlighted that in

System	Precision	Recall	F1-Score
CRF	70.34	27.12	39.15
BERT	68.69	22.22	33.58

Table 4
Results of the two systems on the test set for the (exam, result) pairs recognition

some cases the BERT model is capturing only part of the value, while the CRF model is typically capturing it entirely or not capturing it at all. Some examples of values captured by BERT vs the gold standard: *"39"* vs *"39%"*, *"10 ng/L"* vs *"inferiori a 10 ng/L"*, *"3.6"* vs *"3.6-0.9mg/dL"*.

Another observed aspect is that the BERT model seems to be based more on the position of the words in the sentence than on the words themselves. While this is positive, sometimes it leads to recognizing as examination names words that are nearer to the value but are not the actual name (e.g.: in *"bilirubina diretta 1,8 mg/dL" ("direct bilirubin 1,8 mg/dL")* it takes *"diretta" ("direct")* as name instead of *"bilirubina" ("bilirubin")*). On the contrary, the CRF model is more based on the words themselves, at least for exam names, as it is shown by the fact that among the features with the highest weight it has many specific exam names (7 out of the first 10 features).

6. Discussion

Our two systems performed similarly on the Named Entity Recognition (NER) task. They demonstrated reasonable results for identifying *examination results*, although there is room for improvement. However, the identification of *examination names* proved to be more challenging for both systems. This can be attributed to the limited size of the training data, which made it difficult for the models to generalise to a larger set of previously unseen examination names. Despite incorporating external knowledge resources directly into the CRF model and indirectly into the BERT-based model, they were insufficient to enhance the performance in recognising a broader range of examinations. Further investigation is necessary to explore how other data sources can be utilised for this purpose. While we were unable to find an Italian dataset specifically annotated for examination names, it may be worthwhile to investigate the use of existing annotations for *clinical entities* in the E3C corpus, selecting a subset that closely aligns with the concept of examinations. Another element which is worth to further investigating is the tokenizer: for the BERT-based model, we utilised the Umberto tokenizer, but its limitations in dealing with medical terminology might have negatively affected the performances.

Regarding linking examinations to results, our naive approach based on heuristics did not appear to be a limiting factor, considering the F1-score achieved on the (exam, result) pairs compared to the F1-score for the NER of exam names. However, it is possible to explore data-driven models for this subtask, even though the scarcity of available data presents challenges.

We strongly believe in the application of Natural Language Processing techniques to the medical domain and recognise the huge need for developing models that can effectively process Italian healthcare data. Simultaneously, it is crucial to improve the quantity and quality of annotated datasets to drive the development of such models. Challenges like this are a valuable tool for motivating the academic community to contribute to this field.

References

[1] O. G. Iroju, J. O. Olaleke, A systematic review of natural language processing in healthcare, International Journal of Information Technology and Computer Science 8 (2015) 44–50.

[2] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than English: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[3] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain , in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, 2020.

[6] H. M. Wallach, Conditional random fields: An introduction, Technical Report, Department of CIS, University of Pennsylvania (2004) 22.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[8] F. Tamburini, How "BERTology" changed the state-of-the-art also for Italian NLP, Computational Linguistics CLiC-it 2020 (2020) 415.

[9] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, L. Sacchi, Information extraction from Italian medical reports: An ontology-driven approach, International journal of medical informatics 111 (2018) 140–148.

[10] N. Viani, T. A. Miller, C. Napolitano, S. G. Priori, G. K. Savova, R. Bellazzi, L. Sacchi, Supervised methods to extract clinical events from cardiology reports in italian, Journal of biomedical informatics 95 (2019) 103219.

[11] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli, G. Tognola, Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes, Journal of biomedical informatics 63 (2016) 22–32.

[12] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[13] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[14] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A novel Covid-19 data set and an effective deep learning approach for the de-identification of Italian medical records, Ieee Access 9 (2021) 19097–19110.

[15] M. A. Tanenblatt, A. Coden, I. L. Sominsky, The ConceptMapper Approach to Named Entity Recognition, in: LREC, 2010, pp. 546–51.

[16] H. U. Rahman, N. Chowk, T. Hahn, R. Segall, Disease named entity recognition using conditional random fields, in: Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine, 2016.

[17] A. Magge, M. Scotch, G. Gonzalez-Hernandez, Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers, in: International workshop on medication and adverse drug event detection, PMLR, 2018, pp. 25–30.

[18] M. Abadeer, Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts, in: Proceedings of the 3rd clinical natural language processing workshop, 2020, pp. 158–167.