BERT Representations to Identify Professions and Employment Statuses in Health data

José-Alberto Mesa-Murgado[1], Pilar López-Úbeda[1], Manuel-Carlos Díaz-Galiano[1], M. Teresa Martín-Valdivia[1], and L. Alfonso Ureña-López[1]

Departamento de Informática, CEATIC, Universidad de Jaén, España
{jmurgado,plubeda,mcdiaz,maite,laurena}@ujaen.es

Abstract. In this paper we detail our submission for the Medical Documents Profession Recognition (MEDDOPROF) shared task by submitting three different proposals for each subtask which involves: Named Entity Recognition of mentions regarding professions, occupations and employment statuses within clinical reports. Classify those mentions whether they refer to the visiting patient, a relative, medical staff or others and lastly, mapping those mentions with respect to the European Skills, Competences, Qualifications and Occupations (ESCO) classification and relevant SNOMED-CT terms. Our proposals are based on BERT representations and the similarity between the resulting word embeddings and its corresponding classes.

Keywords: Named Entity Recognition · BERT Representations · Professions and Employment Statuses · Natural Language Processing · Spanish

1 Introduction

This paper describes the systems we have developed as our participation in the Medical Documents Profession Recognition (MEDDOPROF) shared task [7] focused on the Detection and Recognition of Professions and Employment statuses in Health data.

This challenge has been organised by the Barcelona Supercomputing Center as part of the 2021 Iberian Languages Evaluation Forum (IberLEF) evaluative initiative co-located with the XXXVII Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2021).

Detecting professions and employment statuses in these resources could help public and private organizations to evaluate and predict the impact of future pandemic diseases spreading in any location using the reports written after every medical visit.

IberLEF 2021, September 2021, Málaga, Spain.

These medical reports are a very valuable resource, as well as strictly restricted, whose study leads to the development of highly useful applications for the general public such as the detection of COVID-19 using radiological textual reports [9] or the detection of tumor morphology within medical reports as seen in the CANcer TExt Mining Shared Task – tumor named entity recognition (NER) (CANTEMIST) [11] shared task, another challenge proposed by the Barcelona Supercomputing Center in which our research group also participated [8].

The idea behind this study was introduced in the ProfNER shared task as part of the Social Media Mining for Health Applications or SMM4H, co-located with the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021) [12] using data retrieved from social media sources instead of clinical reports. Our research group also participated [10] in this initiative using a Bidirectional Long Short Term Memory (BiLSTM) Recurrent neural network (RNN) approach.

The development of such applications tackle tasks that are commonly attributed to the the Natural Language Processing (NLP) field, aimed at providing techniques to process textual information written in natural language with the purpose of enabling a computer to understand such data in order to fulfill an specific purpose such as Question Answering [1], Automated Translation [3], among others [4]. These tasks are solved through the use of Machine Learning and Deep Learning approaches.

In particular, the latest state of the art approach revolves around the use of BERT (Bidirectional Encoder Representations from Transformers) [5] which has presented good grounding results in a wide variety of the stated NLP common tasks [13].

Our team proposes three systems based on BETO [2], the BERT Transformer language model trained on a Spanish Corpus. On the one hand, a multiclass approach and, on the other, two approaches relying on procedures similar to the one of binary classification, meaning, masking all classes under the same label in order to discriminate between tokens as either entities or non-entities.

Considering this information, this paper is structured as follows: Section 2 introduces the aim of this challenge as well as the description and characteristics of the provided dataset. Section 3 describes each of the three systems proposed to tackle each task, including what hyperparameters we have explored, the preprocessing applied to the input data and the results we have obtained on our development set. Section 4 exhibits the results obtained by our systems on the evaluation set, subsequently, this data is analyzed in Section 5. Lastly, we state our conclusions in Section 6 concluding our submission.

2 MEDDOPROF Tasks Description

The MEDDOPROF Shared Task proposed three tasks in which to participate although participants were not obliged to provide a system for each one of them as these tasks were not dependent on each other.

1. Task 1, namely "MEDDOPROF-NER", asks participants to automatically identify mentions of professions or occupations and employment statuses out of plain text medical documents and tag each one of those entities under the label of "PROFESION" or "SITUACION_LABORAL, there is also one last label, "ACTIVIDAD" but it seems to be scarce label among the given training documents.
2. Task 2, namely "MEDDOPROF-CLASS", requires participants to once again identify mentions of professions or occupations and employment statuses from a given set of plain text documents and label each one of those mentions according to the person to whom they refer to, considering whether the mention is associated to the patient "PACIENTE", to a relative "FAMILIAR", to medical staff "SANITARIO" or any other as "OTROS".
3. Lastly, task 3 or "MEDDOPROF-NORM" requires participants to catalogue those same mentions as per the European Skills, Competences, Qualifications and Occupations (ESCO) classification and relevant SNOMED-CT terms.

2.1 Dataset provided

The MEDDOPROF-NER and MEDDOPROF-CLASS corpus consists of 1844 documents; these documents have been transformed by the organizers into BRAT format to extract the entities associated with each document. Below, we examine each one of the datasets provided: training and evaluation.

Training dataset contains 1500 documents, out of them we have computed the length of their content (number of characters in the document), the number of tokens as well as entities contained in them. This information can be found detailed in Table 1.

Table 1. Metrics associated with the Training dataset.

	Sum	Minimum	Maximum	Average
Lengths	6,239,588	184	27,529	4,159.72
Tokens	1,114,919	29	4,807	743.28
Entities	9,217	1	86	7.44

Development dataset In order to evaluate the performance of our system we had to divide the training corpus, therefore we decided to use the 80% of the given data as training data (1191 documents) while the 20% remaining was used as our development set (309 documents). This division was random but consistent with the number of entities of the corpus. Therefore, the number of entities in the development set is 80 times lower than those in the training set. Its characteristics are detailed in Table 2.

Table 2. Metrics associated with the Development dataset.

	Sum	Minimum	Maximum	Average
Lengths	1,339,910	274	27,529	4,336.27
Tokens	239,033	45	4,807	773.57
Entities	1,762	36	1	5.7

Evaluation dataset consists of 344 documents and we have also retrieved the associated document's length (number of characters contained) as well as the number of tokens and entities in them, this information is available in Table 3.

Table 3. Metrics associated with the Evaluation dataset.

	Sum	Minimum	Maximum	Average
Lengths	1,240,562	228	23,446	3,606.29
Tokens	222,744	37	4,376	647.51
Entities	2,786	1	79	9.05

3 System Description

As stated in the previous section, we have implemented three systems based on BERT language model trained using Spanish Corpus or BETO for which we used an implementation that uses Pytorch library for Python. However, our approaches differ one from the other in the labels used to discriminate tokens:

1. The first approach is a multiclass NER system trained to detect and differentiate between professions and occupations, employment statuses and activities as per their corresponding labels: "PROFESION", "SITUACION_LABORAL" and "ACTIVIDAD", respectively. Throughout this document we will refer to this particular system as "Multiclass BETO".
2. The second one discern tokens that are entities from those that are not, tagging them as either 'ENTITY' or 'O', respectively. This approach follows the IOB (Inside-Outside-Beginning) [14] format. After retrieving the text spans identified as entities, we map each one of them to their corresponding word embedding using BETO to, subsequently, calculate the distance between them and the word embeddings related to each tag. Afterwards, our system associates the label with the smallest distance value (in other words, with the highest similarity) to the identified entity. This distance is measured using the Cosine similarity method as per the Equation 1, being x and y, two word distinct embeddings.

$$k(x, y) = \frac{xy^\top}{\|x\|\|y\|} \tag{1}$$

3. The last of our systems is an extension of the one presented in the previous point. This expansion is produced on the training set through the usage of the corpus provided by ProfNER shared task, this data provides our system with more training data that results in better predictions.

3.1 Preprocessing

We use the spaCy [6] library for Python in order to create a text file containing each sequence contained in our dataset, their start and end positions with respect to the document in which they are found, as well as the original token and its normalized counterpart.

For this normalization we have decided to transform each token to lowercase, converted accented characters into their non accented shapes and removed all characters differing from numbers or special characters such as dots, commas, semicolons, etc.

Also, considering the fact that BERT allows a maximum of 512 tokens per sequence and the given dataset contains sentences above that range, we have decided to establish a maximum of 215 tokens per sequence. This decision was made after noticing the memory limitations of our machine. We experimented with distinct configurations of this value (256 and 300 to be precise), however, the improvement of our results was not noticeable or, results were unobtainable.

Consequently, our strategy involved the division of sequences into smaller ones after normalization. This separation took place right after the detection of a dot character or until conforming a sequence whose length was equal or greater than our established maximum number of tokens per sequence

3.2 Hyperparameters tuning

Training was carried out using 6 RTX 2080Ti, each one containing 11 GB of memory built under Nvidia's Turing architecture. For this purpose we decided to use different batch sizes: 8 and 16, and different epochs: 3, 4 and 5. We have used the development dataset in order to test the performance of our systems.

The combination of these parameters (epochs and batch sizes) produced similar scores between them, therefore we have decided to include only the top performers for each task and system displayed as follows:

- Table 4 shows top combinations for the MEDDOPROF-NER task using the multiclass approach.
- Table 5 does the same for the MEDDOPROF-CLASS task using the multiclass approach.
- Table 6 displays the scores applied to both tasks MEDDOPROF-NER and MEDDOPROF-CLASS using the binary approach.
- Lastly, Table 7 shows the results retrieved from the binary approaches used for the MEDDOPROF-NORM task.

The metrics displayed on these tables are expressed in terms of micro precision, micro recall and micro F1 score.

Table 4. Metrics associated with the Training dataset using Multiclass BETO on MEDDOPROF-NER.

	Batch Size	Epochs	Precision	Recall	F1
Multiclass BETO	8	3	0.83	0.78	0.80
Multiclass BETO	16	4	0.82	0.79	0.80

Table 5. Metrics associated with the Training dataset using Multiclass BETO on MEDDOPROF-CLASS.

	Batch Size	Epochs	Precision	Recall	F1
Multiclass BETO	16	3	0.79	0.81	0.80
Multiclass BETO	16	4	0.82	0.79	0.80

Table 6. Metrics associated with the Training dataset using Binary BETO on MEDDOPROF-NER and MEDDOPROF-CLASS.

	Batch Size	Epochs	Precision	Recall	F1
Binary BETO	16	3	0.83	0.82	0.82
Binary BETO	16	5	0.85	0.8	0.82
Binary BETO + ProfNER	16	3	0.83	0.81	0.82

Table 7. Metrics associated with the Training dataset using Binary BETO on MEDDOPROF-NORM.

	Batch Size	Epochs	Precision	Recall	F1
Binary BETO	16	3	0.859	0.637	0.732
Binary BETO	16	5	0.866	0.628	0.728

4 Results

Organizers have provided participants with a baseline from which to compare our results. These comparisons can be found in detail in Table 8 for the NER task, Table 9 for the CLASS task and lastly, Table 10 for the NORM task. Metrics used are common in NLP tasks to compare the performance of different systems and are expressed in terms of micro-precision (Precision), micro-recall (Recall) and micro-F1 scoring (F1).

Table 8. Results on MEDDOPROF-NER.

System	Precision	Recall	F1
Multiclass BETO	**0.821**	**0.740**	**0.778**
Binary BETO	0.787	0.707	0.745
Binary BETO + ProfNER	0.788	0.720	0.752
Baseline	0.465	0.508	0.486

Our results show an increase of 54% (on average) with respect to those given by the organization as baseline. The Multiclass BETO approach is the one that shows the highest gain, an increase of 58% compared against the baseline, along with the Binary approach added to the ProfNER dataset, showing an increase of 52%. In this task, all of our proposed systems have obtained similar results exceeding 78% accuracy, 70% recall and more than 71% on F1.

Table 9. Results on MEDDOPROF-CLASS.

System	Precision	Recall	F1
Multiclass BETO	**0.775**	**0.69**	**0.730**
Binary BETO	0.592	0.531	0.560
Binary BETO + ProfNER	0.593	0.541	0.566
Baseline	0.391	0.377	0.384

In the MEDDOPROF-CLASS task, we obtained an increase of 61% (on average) with our systems, being the Multiclass BETO approach the one which, once again, holds the highest gain with respect to the baseline (90% better than the organizers given score). Our Binary approach trained along with the ProfNER data showed an increase of 47% when compared to the baseline score. In this case, the results obtained by our binary approach are worse than those of the multiclass.

Table 10. Results on MEDDOPROF-NORM.

System	Precision	Recall	F1
Binary BETO	0.685	0.404	0.508
Binary BETO + ProfNER	0.679	0.404	0.507
Baseline	0.502	0.533	0.517

Lastly, for this task we did not provide a Multiclass BETO approach and only provided results for the systems using the binary approaches. As mentioned, these models managed to tackle all three tasks at the same time as they compared the word embeddings of each text span with those of their respective classes. Our results for this task fall below the baseline given by the organizers by 8%.

5 Results Analysis

To assess the problems that our models have encountered for each task, we implemented our own evaluation software. This program compares each document in the golden test with the results obtained by our systems and retrieves the entities that have not been properly identified and labelled as per MEDDOPROF-NER and MEDDOPROF-CLASS taks, as well as those that have not been correctly normalized as per the MEDDOPROF-NORM taks.

We have applied this evaluation to the top performer systems of every task and created one single evaluation summary, this information can be appreciated in Figure 1 with respect to every task (coloured as blue, orange and gray for MEDDOPROF-NER, MEDDOPROF-CLASS and MEDDOPROF-NORM, respectively). The data we wanted to extract and highlight consisted of:

- Entities that our system detected and labelled correctly with respect to the golden test, namely, True Positives (TP).
- Entities detected by our systems not found in the golden test, namely, False Positives (FP).
- Lastly, entities included in the golden test that were not detected by our systems, namely, Missed Entities.

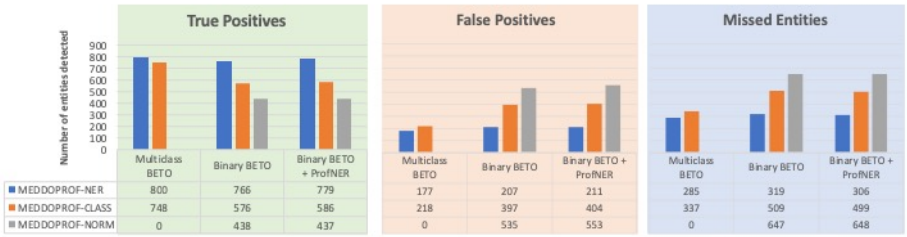

Fig. 1. Number of entities mislabeled by our systems.

To comprehend why our models were unable to correctly identify some entities in the evaluation dataset, we decided to retrieve the text spans concerning FP as well as Missed Entities. Considering that this information is too large to be included in this document, we decided to only display the top 10 most frequent entities included in each of the previously mentioned sets (FP and Missed Entities) along with their frequency (number of occurrences within those documents) , respectively in Table 11 and Table 12.

After retrieving the results from our error assessment software, we have noticed that there are certain entities that have been conceived as false positives while also appearing as missed entities.

From the list of entities contained in Table 11 and Table 12, we observe that:

- For the MEDDOPROF-NER, the number of errors in the label assignment was 0 out of the 10 top most frequent.
- For the MEDDOPROF-CLASS, this event happened in 4 entities and what we have noticed is that our system has incorrectly tagged with a different label (e.g: “compañero” as “OTROS” instead of “PACIENTE”, or “psicóloga” as “OTROS” instead of “SANITARIO”) although the entity has been span has been identified.

Table 11. False Positives per task along with their Frequencies (Freq).

NER Task with Multiclass BETO	**Freq**	**CLASS Task with Multiclass BETO**	**Freq**	**NORM Task with Binary BETO + ProfNER**	**Freq**
compañeros	6	compañero	5	trabajadora	26
equipo	3	trabajador	5	trabajador	23
terapeuta	3	psicóloga	3	compañeros	21
en prisión	2	profesionales	3	ama de casa	9
estudia	2	en prisión	3	estudiante	7
tutora	2	terapeuta	3	en prisión	6
médico de atención primaria	2	compañeros	3	jubilado	5
músico	2	de baja	2	Jubilado	5
compañera de la peluquería	1	fábrica de aluminio	2	profesionales	4
especialista en otro centro	1	compañeras de piso	2	Incapacidad Temporal	4

– Lastly, it is in MEDDOPROF-NORM where we find more matching terms as 9 entities of the 10 most frequent appear in both tables. After analyzing these matching results we have concluded that this is due to the fact that some codes included the term "SCTID: " in their labelling while our output only included the number. After repeating the experiment and applying the official evaluation software provided by the organization, our scores seem to have outperformed the baseline as seen in Table 13.

6 Conclusions

For our participation on the MEDDOPROF shared task challenge we have proposed three systems based on the BERT representations to tackle all three tasks: MEDDOPROF-NER, MEDDOPROF-CLASS and MEDDOPROF-NORM, respectively. Our systems managed to outperform the given baselines with an increase of 54% and 61% for the MEDDOPROF-NER and MEDDOPROF-CLASS. The exception of this performance is the MEDDOPROF-NORM in which our performance has fallen below the baseline, although close to it.

For future work our plan is to implement another Deep Learning model, a BiLSTM approach in combination with a Conditional Random Field (CRF) to tackle all of the proposed tasks and therefore, evaluate whether this system would have outperformed our proposed solutions or not. In addition, we also want to implement the first of our approaches to be able to tackle the MEDDOPROF-NORM for which our premise is that it will outperform the baseline.

Table 12. Misses per task along with their Frequency (Freq).

NER Task with Multiclass BETO	**Freq**	**CLASS Task with Multiclass BETO**	**Freq**	**NORM Task with Binary BETO + ProfNER**	**Freq**
IT	10	IT	10	trabajadora	29
en el colegio	4	trabajador	7	trabajador	23
AXE	4	compañeros	6	compañeros	16
cuidadores	3	en el colegio	4	IT	10
especialista	3	cuidadores	4	ama de casa	9
militar	2	AXE	4	estudiante	7
robo de dinero	2	médico	3	incapacidad temporal	6
cuidado de su madre	2	psicóloga	3	profesionales	5
excursionistas	2	bajas laborales	2	en el colegio	5
despido	2	compañero de viaje	2	de baja laboral	5

Table 13. Results on MEDDOPROF-NORM after fixing the problem regarding codes.

System	Precision	Recall	F1
Binary BETO	0.687	0.613	0.647
Binary BETO + ProfNER	**0.688**	**0.624**	**0.654**
Baseline	0.502	0.533	0.517

Acknowledgements

This work has been partially supported by the LIVING-LANG project [RTI2018-094653-B-C21] of the Spanish Government and the Fondo Europeo de Desarrollo Regional (FEDER).

References

1. Bouziane, A., Bouchiha, D., Doumi, N., Malki, M.: Question answering systems: survey and trends. Procedia Computer Science 73 pp. 366–375 (2015)
2. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
3. Chowdhury, G.G.: Natural language processing. Annual review of information science and technology 37 (1), 51–89 (2003)
4. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in bioinformatics 6 (1), 57–71 (2005)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018), http://arxiv.org/abs/1810.04805, cite arxiv:1810.04805Comment: 13 pages
6. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303, https://doi.org/10.5281/zenodo.1212303

7. Lima-López, S., Farré-Maduell, E., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural **67** (2021)
8. López-Úbeda, P., Díaz-Galiano, M.C., Martín-Valdivia, M.T., López, L.: Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. In: IberLEF@SEPLN (2020)
9. López-Úbeda, P., Díaz-Galiano, M.C., Martín-Noguerol, T., Luna, A., Ureña-López, L.A., Martín-Valdivia, M.T.: Covid-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine 127 p. 104066 (2020)
10. Mesa Murgado, J.A., Parras Portillo, A.B., López-Úbeda, P., Martín-Valdivia, M.T., Lopez, L.A.U.: Identifying professions & occupations in health-related social media using natural language processing. In: Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task (2021)
11. Miranda-Escalada, A., Farré, E., Krallinger, M.: Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings (2020)
12. Miranda-Escalada, A., Farré-Maduell, E., López, S.L., Briva-Iglesias, V., Agüero-Torales, M., Gascó-Sánchez, L., Krallinger, M.: The profner shared task on automatic recognition of professions and occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In: Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task (2021)
13. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22 (10), 1345–1359 (2009)
14. Schneider, N., Danchik, E., Dyer, C., Smith, N.A.: Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. Transactions of the Association for Computational Linguistics 2 pp. 193–206 (2014), https://www.aclweb.org/anthology/Q14-1016