PUCRJ-PUCPR-UFMG at eHealth-KD Challenge 2021: A Multilingual BERT-based System for Joint Entity Recognition and Relation Extraction

Lucas Pavanelli[1][0000−0003−2228−7965], Elisa Terumi Rubel Schneider[3][0000−0002−8921−5598], Yohan Bonescki Gumiel[2,3][000−0001−8239−2930], Thiago Castro Ferreira[2][0000−0003−0200−3646], Lucas Ferro Antunes de Oliveira[3][0000−0003−4052−7993], João Vitor Andrioli de Souza[3][0000−0002−8950−0890], Giovanni Pazini Meneghel Paiva[3][0000−0002−9789−9547], Lucas Emanuel Silva e Oliveira[3][0000−0003−1811−5087], Claudia Maria Cabral Moro[3][0000−0003−2637−3086], Emerson Cabrera Paraiso[3][0000−0002−6740−7855], and Adriana Pagano[2][0000−0002−3150−3503]

[1] Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil
lpavanelli@inf.puc-rio.br
[2] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil thiagocf05@ufmg.br
[3] Pontifícia Universidade Católica do Paraná, Curitiba, Brazil c.moro@pucpr.br

Abstract. This study introduces the system submitted to the eHealth-KD Challenge 2021 by the PUCRJ-PUCPR-UFMG team. We proposed a multilingual BERT-based system for joint entity recognition and relation extraction in multidomain texts. Our end-to-end multitasking model benefits from the transformer architecture, which has proved to capture better the global dependencies of the input text. Also, the use of a multilingual model contributed to our system to perform well even in the set of tests containing non-Spanish sentences. Our system ranked first in the entity recognition task and second in the Main scenario, where both tasks of entity recognition and relation extraction had to be solved. The full code of our approach and more details of the implementation are publicly available.

Keywords: eHealth · Entity Recognition · Relation Extraction · BERT · Deep Learning

1 Introduction

Recent advances in Natural Language Processing (NLP) allow the extraction of relevant information from clinical and biomedical texts, automatically acquiring

a wide variety of knowledge from unstructured health documents [10]. Tasks such as named entity recognition (NER) and extracting relations between entities can support other tasks and assist in healthcare decision-making. In the clinical domain, NER can identify clinical concepts such as symptoms, diseases and procedures, extracting valuable information about the patients. The extraction of relations between entities allows finding information, such as interactions between drugs, which can aid healthcare professionals by providing information to enhance patient care.

The IberLEF eHealth Knowledge Discovery Challenge (eHealthKD) 2021 [18] targets the recognition of entities and their relations in the clinical domain, encouraging researchers and scientists to discover new knowledge through text mining and NLP in the health domain. The challenge involves the use of human language modeling in electronic health documents in Spanish with semantic interpretation, in the tasks of entity recognition and relation extraction. The semantic structure has four types of information units, which can have relationships among themselves (13 types of semantic relations). In addition to having a larger dataset, compared to the previous challenge [17], the 2021 edition also contains cross-domain and multi-language features, encouraging the development of more generic and adaptive systems, which can be applied readily in several languages and domains. In this respect, the released dataset comprised texts in Spanish and English and covered both healthcare and news domains.

In our method, we use a transformer-based model, as large pre-trained language models based on the transformer architecture [20] have reached the state-of-the-art in various NLP tasks. We employ the multilingual version of Bidirectional Encoder Representations from Transformers (BERT) [5], which supports 104 languages including Spanish. We implemented an end-to-end multitasking BERT-based model fine-tuned to extract the entities from text and classify relations between them. Our approach is based on the 2020 Vicomtech method [7], one of the best approaches in the 2020 version of the challenge [17]. In order to contribute research in the tasks of extracting entities and relations, we also make our code available in a public repository [16], allowing it to be easily reproduced.

The paper is organised as follows: Section 2 describes the proposed method with architecture and implementation details; Section 3 presents the results of all competing systems; and Sections 4 and 5 provide some discussions and conclusions obtained from the observed results.

2   System Description

Based on the 2020 Vicomtech approach [7], our method consists in an end-to-end multilingual BERT-based system that jointly predicts both entities and relations. During training, the proposed multi-task system is fine-tuned in 3 sequential steps: the first one prioritizes the entity recognition task whereas the second gives precedence to the relation extraction one. Finally, the third and last step is trained for both tasks using a multi-task strategy. In this section, we detail

the system's architecture, how we handle the inputs and the outputs and, finally, present the parameters and training setup.
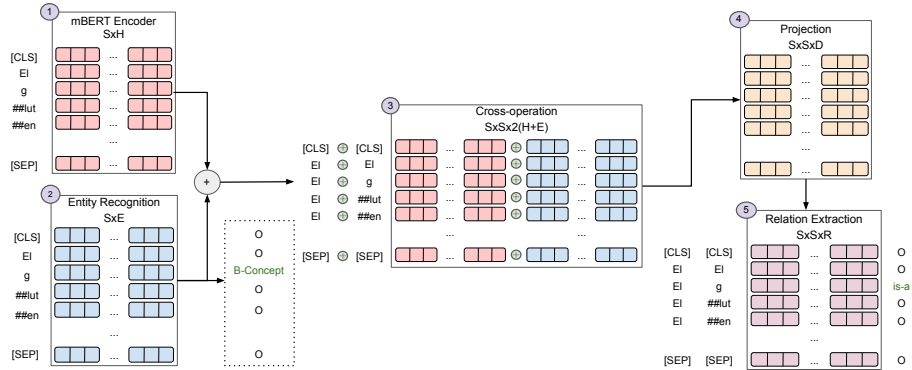

Fig. 1. Architecture of the approach submitted to the eHealth-KD Challenge 2021 by the PUCRJ-PUCPR-UFMG team.

2.1 Architecture

The architecture of our model is presented in Figure 1. In the following paragraphs we will explain each of its components.

Encoder Like 2020 Vicomtech, our approach first tokenizes the input text and encodes its tokens (words or subwords) into vector representations (Step 1 in Figure 1). Unlike Vicomtech's previous approach, which used BETO [3], a BERT version trained on the Spanish language, our approach uses mBERT, a multilingual version of BERT pretrained in texts of 104 languages [5]. We used the `bert-base-multilingual-cased` setting, with 12 self-attention heads, 12 layers (transformer blocks), and embedding length of 768 dimension, which encodes multilingual cased texts.

Entity Recognition Once the input text is encoded, as depicted in Step 2 in Figure 1, the encoded vector representations are fed into a softmax classifier for entity recognition. In the provided dataset, each entity can be classified into 4 categories: `Concept`, `Action`, `Predicate` and `Reference`. In order to know when a token is part of an entity mention and where each of these mentions starts and ends in terms of tokens, we used the `IOB2` format, popular in Named Entity Recognition applications, so that each token of the text could be labeled by the classifier according to 9 categories: `O`, `B-Concept`, `I-Concept`, `B-Action`, `I-Action`, `B-Predicate`, `I-Predicate`, `B-Reference` and `I-Reference`. The `O` label is used to mark tokens which are not part of an entity mention, whereas the ones starting with `B-` and `I-` indicate the beginning and subsequent tokens of a mention, respectively.

Relation Extraction Like the 2020 Vicomtech approach, we concatenate the logits of the entity recognition classifier with the vector representations related to the respective tokens. A cross-operation is then performed by concatenating each pair of vector representations among the tokens, resulting in a tensor of dimension $SxSx2(H + E)$, being S the sequence size of the input text, H the 768 dimensions of the vector representations and E the 9 dimensions of the logits (Step 3 of Figure 1). The matrix is further fed into a projection layer with an `Tanh` activation function, which maps the input onto a $SxSxD$, where $D = 768$ (Step 4 of Figure 1). Finally, the output of the previous operation is given as input to a classifier which predicts the relation of each pair of tokens according to 13 categories of the dataset (`is-a`, `part-of`, `has-property`, `causes`, `entails`, `in-context`, `in-place`, `in-time`, `subject`, `target`, `domain`, `arg` and `same-as`), plus a `O` one, which indicates there is no relation among the target pair.

Classifiers Both entity recognition and relation extraction classifiers consist in a projection layer with a `Mish` [13] activation function and dropout of 0.2, followed by a softmax layer.
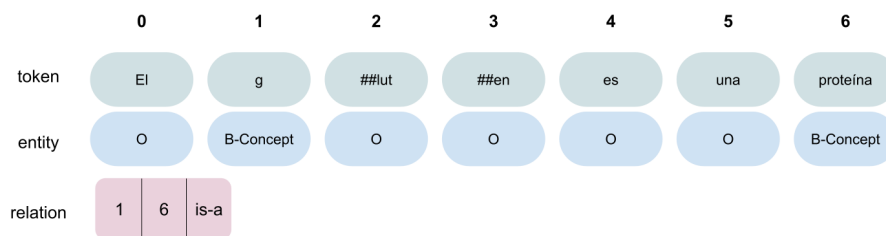

Fig. 2. Example of model's input.

2.2 Input handling

Since the corpora have been provided in a character span-based format and our network works at token level, we tokenize the sentence text using BERT default tokenizer [5], resulting in WordPiece information.

Next, for each token, we assign Begin and Inside tags (IOB2 format), if it is part of an entity, and `O` otherwise. By using this approach, we can represent consecutive entities with more than one token. However, this prevents us from representing discontinuous entities e.g. considering the text span "uno o dos días", we cannot represent the entity ("un día") using the IOB2 format. In this case, we only consider the first entity ("un"). We opt for this simple approach instead of a complex representation, because we value building a more simple and efficient model.

As for relations, we represent them as triples, containing the first token of each entity in the relation and the relation type. We use these triples to fill the relation matrix. Figure 2 shows an example input of our model.

2.3 Output handling

The output of the model needs to be converted back to a character span-based format. So we implement a postprocessing module that is responsible for this conversion. The model's output contains a sequence of tokens, each one assigned to an entity tag and a SxS matrix informing the relation between each pair of tokens.

For each token, if it is the beginning of an entity, we identify the character range that it spans and add this span list to the result. Next, we discard entities that are entirely contained within another one and which start with a stopword. Lastly, we construct the relations by linking entities that contain at least one token in the model's relations output.

2.4 Parameters and Training Setup

Our neural network approach was trained using the `AdamW` [11] optimizer combined with a linear scheduler which warms-up the training process from an initial learning rate of 2e-6 up to 2e-5 over the first 10 epochs. Using a batch of size 1, we train the approach in 3 sequential steps.

In the first step, all the training parameters of the network are frozen except for the ones from mBERT and the entity recognition classifier. The model is then trained for 50 epochs with early stopping of patience 15 (i.e., the training algorithm waits 15 epochs before early stop if no progress on the validation set), computing the loss only based on the entity recognition task:

$$J_{ent}(x^{(ent)}, y^{(ent)}) = \frac{1}{N} \sum_{n=1}^{N} x_{y_n^{(ent)}}^{(ent)} \tag{1}$$

where $x^{(ent)}$ is the likelihood computed by the entity recognition classifier, $y^{(ent)}$ is the gold-standards and N is the size of the batch.

For the second step, which focuses on the relation extraction task, we only freeze the training parameters of the entity recognition classifier. The approach is also trained for 50 epochs with early stopping of patience 15, though unlike the previous step, the loss is computed based on the relation extraction task:

$$J_{rel}(x^{(rel)}, y^{(rel)}) = \frac{1}{N} \sum_{n=1}^{N} x_{y_n^{(rel)}}^{(rel)} \tag{2}$$

where $x^{(rel)}$ is the likelihood computed by the relation extraction classifier, $y^{(rel)}$ the gold-standards and N is the size of the batch.

Finally, we perform a third training step with 100 epochs and early stopping of patience 15 in order to fine-tune the model for both tasks. None of the training parameters are frozen and the loss is computed based on [9] in the following way:

$$J = e^{-\alpha_{ent}} \times J_{ent}(x^{(ent)}, y^{(ent)}) + \alpha_{ent} + e^{-\alpha_{rel}} \times J_{rel}(x^{(rel)}, y^{(rel)}) + \alpha_{rel} \quad (3)$$

being α_{ent} and α_{rel} training parameters as well.

Our approach was trained using the training and development sets released for the shared-task. Although the third step was scheduled to run for 100 epochs, due to time constraints regarding the submission deadline, we reported results for 67 epochs in this step. After submission, we performed the third training process with a 100 epochs run. In the following section we present the results of our officially submitted approach as well as the results obtained in our subsequent 100 epochs run in the third step.

3   Results

Table 1 displays the results (precision, recall and F1-score) reported by the participating systems in the eHealth-KD Challenge 2021 [18]. For the entity recognition task (Task A), our official approach, which ran for 67 epochs in the third training step and is labeled as `Our Approach` in the table, ranked first with an F-Score of 70.60 outperforming `Vicomtech`, the second best in the task and developed by the winning team of the 2020 version of the challenge [17]. On the other hand, for the relation extraction task (Task B) our system had a significant drop, ranking 4th in the task with an F-Score of 26.32 behind the `IXA`, `Vicomtech` and `uhKD4` systems. The intermediate performance of our approach in task B was made up for by its good performance for the entity recognition task so that our approach ranked second in the Main task (which combines both tasks), just behind `Vicomtech`.

Table 1. Participating systems' reported results for the eHealth-KD Challenge 2021

	Main Entity + Relation				Task A Entity R.				Task B Relation E.			
	#R	P	R	F1	#R	P	R	F1	#R	P	R	F1
Our Approach	2	**56.85**	50.28	52.84	1	**71.49**	69.73	**70.60**	4	36.66	20.54	26.32
Our Approach 100 epcs	-	53.63	49.39	51.42	-	**71.49**	69.20	70.33	-	32.31	22.96	26.85
Vicomtech [6]	1	54.08	53.46	**53.11**	2	69.99	**74.71**	68.41	2	54.19	28.31	37.19
IXA [2]	3	46.46	**53.86**	49.89	3	61.37	69.8	65.33	1	45.36	**40.95**	**43.04**
uhKD4 [1]	4	48.53	37.43	42.26	5	51.75	53.74	52.73	3	**55.62**	22.24	31.77
UH-MMM [14]	5	29.16	40.37	33.87	4	54.60	68.50	60.77	5	07.73	4.13	05.38
CodestrangeTeam [12]	6	33.70	17.69	23.20	10	41.50	4.44	8.02	6	43.75	1.70	3.28
baseline [18]	7	33.70	17.69	23.20	7	35.03	27.17	30.60	7	43.75	1.70	3.28
JAD [15]	8	10.95	23.44	7.14	8	31.58	22.46	26.25	8	37.50	0.365	0.722
Yunnan-Deep [8]	-	-	-	-	6	52.04	24.60	33.41	-	-	-	-
Yunnan-1 [21]	-	-	-	-	9	27.11	12.73	17.32	-	-	-	-

As mentioned in Section 2.4, due to time constraints, we ran the third step of training for 67 epochs instead of the originally scheduled 100 ones. We initially

hypothesized this to be the reason for the significant drop in performance of our model in the Relation Extraction task. However, this does not seem to have been the case as evidenced by the results of our system obtained after subsequently training for 100 epochs in the third step, depicted as `Our Approach 100 epcs`. In terms of F-Score, the 100 epochs version of our approach showed a slight improvement compared to our official results in the task to the cost of a slight drop in performance in the entity recognition task.

4   Discussion

Our approach is a simplified version of 2020 `Vicomtech2020`, the winner of the 2020 challenge. After encoding a sentence using a BERT-based method, the original system uses two classifiers for entity recognition: the first one predicts whether each token is part or not of an entity of a certain type; and the second one, a multiword classifier, predicts whether each pair of tokens is part of a same entity mention. Unlike 2020, adopting the `IOB2` format, we used a single classifier for the entity recognition task. Besides simpler, our approach outperformed the original, ranking first for the entity recognition task.

A further distinction between the original 2020 `Vicomtech2020` approach and ours is that the former uses a DistilBERT [19] module to map the tensor with pairs of token representations, whereas ours uses a linear projection with a `Tanh` activation function, which makes our system much less computationally intensive. Moreover, the original approach uses three classifiers in order to solve the relation extraction task. The first predicts a bidirectional `same-as` relation between pairs of tokens. The second aims to predict whether there is or not another kind of relation between tokens. If a relation is predicted, a third classifier is used to predict the relation type. Unlike 2020 `Vicomtech2020`, our approach is built with only one classifier for relation extraction, which predicts whether there is a relation and its type. As an impact of our simplification, the method was outperformed and ranked 4th in task B, even after training the method for a total of 100 epochs.

5   Conclusions

This study has introduced the approach developed by the `PUCRJ-PUCPR-UFMG` team for the eHealth-KD Challenge 2021. The approach is able to jointly solve the entity recognition and relation extraction tasks in multilingual texts, using a fine-tuned version of mBERT, a BERT version which supports 104 languages. Our system ranked first in the entity recognition task and second in the Main scenario, with a simple and little computationally-intensive approach. As the trained model can benefit several downstream NLP tasks, we have publicly released [16] our method for researchers and community in general. As a further step in our work, we intend to explore relation classification heuristics in order to improve our results for entity relation extraction.

Acknowledgments

References

1. Alfaro-González, D., Pérez-Perera, D., González-Rodríguez, G., Otaño-Barrera, A.J.: uhKD4 at eHealth-KD Challenge 2021: Deep Learning Approaches for Knowledge Discovery from Spanish Biomedical Documents. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
2. Andrés, E.: IXA at eHealth-KD Challenge 2021: Generic Sequence Labeling as Relation Extraction Approach. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
3. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
4. 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)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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) pp. 4171–4186 (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
6. García-Pablos, A., Pérez, N., Cuadros, M.: Vicomtech at eHealth-KD Challenge 2021: Deep Learning Approaches to Model Health-related Text in Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
7. García-Pablos, A., Perez, N., Cuadros, M., Zotova, E.: Vicomtech at eHealth-KD Challenge 2020: Deep End-to-End Model for Entity and Relation Extraction in Medical Text. In: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@ SEPLN. vol. 2020 (2020)
8. Guan, Z., Liu, R.: Yunnan-Deep at eHealth-KD Challenge 2021: Deep Learning Model for Entity Recognition in Spanish Documents. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
9. Kendall, A., Gal, Y., Cipolla, R.: Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
10. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., Botsis, T.: Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics **73**, 14–29 (2017)
11. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
12. Marti, R., Bermudez, C., García, L., Gutiérrez, L.: CodeStrange at eHealth-KD Challenge 2021. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)

13. Misra, D.: Mish: A Self Regularized Non-Monotonic Activation Function. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press (2020), https://www.bmvc2020-conference.com/assets/papers/0928.pdf

14. Monteagudo-García, L., Marrero-Santos, A., Fernández-Arias, M.S., Cañizares-Díaz, H.: UH-MMM at eHealth-KD Challenge 2021. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)

15. Navarro Comabella, J.G., Valle Diaz, J.D., Helguera Fleitas, A.: JAD at eHealth-KD Challenge 2021: Simple Neural Network with BERT for Joint Classification of Key-Phrases and Relations. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)

16. Pavanelli, L., Schneider, E.T.R., Gumiel, Y.B., Ferreira, T.C., Oliveira, L.F.A., De Souza, J.V.A., Paiva, G.P.M., Oliveira, L.E.S., Moro, C.M.C., Paraiso, E.C., Pagano, A.: PUCRJ-PUCPR-UFMG. https://github.com/eHealth-KD-PUCs-UFMG/pucrj-pucpr-ufmg (2021)

17. Piad-Morffis, A., Gutiérrez, Y., Cañizares-Diaz, S., Estevez-Velarde, S., Almeida-Cruz, Y., Muñoz, R., Montoyo, A.: Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020. In: Proceedings of the Iberian Languages Evaluation Forum (2020)

18. Piad-Morffis, A., Gutiérrez, Y., Estevez-Velarde, S., Almeida-Cruz, Y., Muñoz, R., Montoyo, A.: Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)

19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), http://arxiv.org/abs/1910.01108

20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, u., Polosukhin, I.: Attention is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems p. 6000–6010 (2017)

21. Yang, M.: Yunnan-1 at eHealth-KD Challenge 2021: Deep-Learning Methods for Entity Recognition in Medical Text. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)