Leveraging Biomedical Ontologies for Clinical Procedures Recognition in Spanish at BioASQ MedProcNER

Notebook for the BioASQ Lab at CLEF 2023

Petar Ivanov[1], Anna Aksenova[1], Tsvetan Asamov[1] and Svetla Boytcheva[1]

[1]*Ontotext, Bulgaria*

Abstract

This paper presents a hybrid approach for name entity recognition (NER) of procedures in unstructured medical texts in Spanish. Our approach combines fine-tuned transformers with domain-specific corpora for Spanish clinical texts and a dictionary-based solution. Specifically, we focus on the MedProcNER CLEF (Conference and Labs of the Evaluation Forum) 2023 challenge subtask 1 "Clinical Procedure Recognition" and provide a light and efficient solution. As a result, our best-performing system showed 0.5505 f1-score on the test set of the challenge. The results on the challenge test set demonstrate the effectiveness of our approach.

Keywords

Name Entity Recognition (NER), Clinical procedure recognition, Spanish Clinical Transformers, Biomedical ontologies, Biomedical NLP

1. Introduction

In this paper we present a hybrid approach for name entity recognition (NER) of procedures in unstructured medical texts based on fine-tuned transformers with domain-specific corpora for Spanish clinical texts combined with a dictionary-based solution. The proposed approach is a light and efficient solution for MedProcNER[1] [1] track at BioASQ CLEF 2023 challenge [2] subtask 1 "Clinical Procedure Recognition".

Our approach makes use of SNOMED CT[2] which is one of the most comprehensive medical ontologies, containing over 360 thousand classes. Due to the large number of classes text-based classification for SNOMED CT is considered an extreme-scale classification task. This is also true for the smaller task of classification of procedures only as SNOMED CT contains 7 groups of procedures with over 86 thousand labels. Furthermore, SNOMED CT is designed such that concepts can be used in postcoordinated manner, meaning that multiple codes can

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

petar.ivanov@ontotext.com (P. Ivanov); anna.aksenova@ontotext.com (A. Aksenova); tsvetan.asamov@ontotext.com (T. Asamov); svetla.boytcheva@ontotext.com (S. Boytcheva)

project/647468604806634069 21f5c5 (P. Ivanov); 0000-0002-3489-874X (A. Aksenova); 0000-0002-7556-1350 (T. Asamov); 0000-0002-5542-9168 (S. Boytcheva)

[1]https://temu.bsc.es/MedProcNER/

[2]https://www.snomed.org/

be combined to form a new code in order to represent additional clinical detail. This makes the scale of the classification task even greater, as the total number of viable SNOMED CT codes pertaining to medical procedures is hard to quantify. Finally, clinicians often use various linguistic descriptions and abbreviations which makes the task for NER even more challenging.

The paper is structured as follows: Section 2 briefs the task formulation; Section 3 presents some related work; Section 4 describes the data used for experiments; The backbone models and their combination with a simple gazetteer method are presented in Section 5; Section 6 reports the experiment settings; Section 7 discusses the results; Section 8 highlights the conclusion and sketches some directions for further work.

2. Task formulation

The aim of the MedProcNER 2023 challenge was to stimulate the development of AI systems that can identify medical procedures within unstructured medical texts in languages other than English, more specifically - Spanish. The task is composed of 3 subtasks - clinical procedure recognition (named-entity recognition), clinical procedure normalization (entitiy linking) and clinical procedure-based document indexing (document classification).

Each subtask is dependent on the outputs of the previous task, i.e. procedure normalization can only be performed on the recognised procedures, while document indexing with SNOMED CT codes requires that procedures identified within the document are themselves normalized to SNOMED CT codes.

In this paper, we describe our approach to the first subtask, namely clinical procedure recognition.

3. Related work

The popularity and importance of NER tasks for procedures in the medical text accelerated during the COVID-19 pandemic. The most recent state-of-the-art approaches include deep learning methods like word embeddings models for SNOMED post coordination [3], pre-trained transformers for biomedical domain like BioBERT [4] and ClinicalBERT [5] for English and the Spanish versions mBERT-Galén, BETO-Galén, and XLM-R-Galén [6]. Even showing great performance for NER tasks in clinical texts, the fine-tuning of transformers is a computationally and resources expensive process. The limited resources for training such models force scientists to turn to alternative approaches involving few-shot learning models based on transformers [7]. Large Language Models (LLM) were game-changers and introduces zero-shot learning based on prompt engineering for GPT-4 [8]. Despite the advancement of LLM still the classical combination between the transformers and dictionary-based approaches [9] shows significantly high performance.

4. Data

The data used for our system can be divided into two parts: data provided by the challenge organizers and additional data that we collected to expand the gazetteer of medical procedures.

4.1. MedProcNER corpus

The MedProcNER corpus [10] is the same subset of the SPACC corpus [11] that is used for the DisTEMIST challenge[3]. More specifically, the corpus consists of 1000 clinical case reports in Spanish which are annotated with clinical procedure mentions (as opposed to disease mentions in DisTEMIST) and normalized to SNOMED CT codes. The corpus contains a total of 16,504 sentences, averaging 16.5 sentences per clinical case.

The annotation procedure was performed by clinical experts using the Brat annotation tool[4] and following annotation guidelines [12] defined by the challenge organizers. This resulted in almost 10,000 annotations over the whole corpus, averaging 10 annotations per clinical case report and 0.6 annotations per sentence.

4.2. Data preprocessing

We used the brat2CoNLL tool[5] to convert the provided training data to CoNLL format [13]. CoNLL is the common format for fine-tuning BERT-like models in token classification setting. The format splits text into individual words and their corresponding labels (classes), each line containing one word separated by its label with an empty space.

4.3. Expanded gazetteer data

The gazetteer [10] provided by the organizers of the MedProcNER challenge contains approximately 89,000 procedure labels in Spanish.

We used external linked open data resources to enlarge the gazetteer with additional procedure names. Namely, we added:

- Spanish labels of SNOMED CT codes for procedures;
- CIE9[6] & CIE10[7] labels for procedures mapped to SNOMED CT codes;
- Machine translated (from English to Spanish) labels for procedures from UMLS[8] mapped to SNOMED CT codes. Procedures from the following UMLS categories were used:
  - PROC|Procedures|T060|Diagnostic Procedure
  - PROC|Procedures|T058|Health Care Activity
  - PROC|Procedures|T059|Laboratory Procedure
  - PROC|Procedures|T063|Molecular Biology Research Technique
  - PROC|Procedures|T061|Therapeutic or Preventive Procedure

[3] https://temu.bsc.es/distemist/
[4] https://brat.nlplab.org/
[5] https://github.com/pranav-s/brat2CoNLL
[6] Spanish version of ICD-9 (International Classification of Diseases 9th edition) https://www.cdc.gov/nchs/icd/icd9cm.htm maintained by the Spanish Ministry of Health https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_9_mc.html
[7] Spanish version of ICD-10 (International Classification of Diseases 10th edition) https://www.cdc.gov/nchs/icd/icd10.htm maintained by the Spanish Ministry of Health https://eciemaps.mscbs.gob.es/ecieMaps/browser/metabuscador.html
[8] https://www.nlm.nih.gov/research/umls/index.html

Our expanded gazetteer contained approximately 292 thousand labels (in Spanish) corresponding to SNOMED CT procedure codes, a over 3-fold increase compared to the original gazetteer.

5. Methodology

5.1. Backbone models

Our system was built upon biomedical transformer models that were presented as top-rated submission in the DisTEMIST (DISease TExt Mining Shared Task) track of a previous iteration of the BioASQ challenge [14]. We experimented with 3 versions of Spanish clinical RoBERTa model:

- *bsc-bio-ehr-es*[9]: RoBERTa-based model pre-trained on biomedical documents, clinical cases and Electronic Health Record (EHR) documents. This model was used as a base model for the next two.
- *bsc-bio-ehr-es-pharmaconer*[10]: RoBERTa-based model pre-trained on biomedical documents, clinical cases and EHR documents and PharmeCoNER data [15].
- *bsc-bio-ehr-es-cantemist*[11]: RoBERTa-based model pre-trained on biomedical documents, clinical cases and EHR documents and CANTEMIST data [16].

5.2. Named Entity Recognition (NER) task

For the Named Entity Recognition task we employed simple yet efficient transformer fine-tuning approach for token classification task. To simplify the setting and to avoid the input sequence length limits, we split the texts into sentences before passing them through transformer model while training. Standard token classification pipeline from Huggingface Transformers was applied [17].

As described earlier, there are approximately 0.6 annotated entities per sentence. Hence the majority of tokens to be evaluated by the model will be negative examples. We therefore fine-tuned our models with different class weights ratios (positive to negative samples) ranging from 3:1 to 30:1.

To improve the recall of our system, we also used the CLEF 2023 expanded gazetteer described in 4.3 in order to find all the mentioned procedures with regular expressions search.

6. Experiments

6.1. Fine-tuning experiments

All the models were fine-tuned on a single NVIDIA A5000 GPU.

For fine-tuning the models we used the following hyperparameter settings:

[9] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es
[10] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer
[11] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-cantemist

- `learning rate`: We used AdamW optimizer [18] with 1-e5 default learning rate.
- `number of epochs`: We experimented with 1, 3, 5, 10, 15 and 20 epochs.
- `batch size`: Initialized to 32 due to GPU memory limitations.
- `class weights in loss`: As the data is highly imbalanced, we applied the weighted loss function manually choosing the following combinations of class weights: 3:1, 5:1, 10:1 and 30:1. In all cases the first number corresponds to the weight of minor class (procedure entity).

This resulted in 24 hyperparameter combinations to fine-tune for each of the 3 model versions (pre-trained on EHR, PharmaCoNER and CANTEMIST datasets).

We split the provided training set such that we used 80% to fine-tune our models, leaving 20% for validation. Using this validation set, we compared the recall, precision and f1-score of all fine-tuned models.

Overall, for all 3 model versions recall was higher than precision for all hyperparameter combinations. We also observed a few other trends in how the scores change with the hyperparameter values.

First, as the class weight ratio increased, meaning tokens corresponding to procedures weight more, the recall improved, but precision decreased, leading to a slight decrease of f1-score. On the other hand, for each class weight ratio as the number of fine-tuning epochs increased recall decreased, but precision improved much faster, leading to higher f1-scores.

Graphs representing the change of precision, recall and f1-score with different hyperparameter values for the PharmaCoNER and CANTEMIST versions of the model are provided in the Appendix.

We selected the 3 final models by the highest f1-score. Those were the EHR, CANTEMIST and PharmaCoNER RoBERTa models fine-tuned for 10, 15 and 20 epochs respectively. Should one chose to optimize for recall instead, models fine-tuned for 3 to 5 epochs would be a more suited choice, as those achieve up to 10% higher recall than their counterparts fine-tuned for longer.

6.2. Voting ensemble for noise removal

Upon visual examination of the predictions of our models we observed that a non-negligible number of our models' predictions were short single words, who's neighbouring words were not labeled as procedures. We deduced that this was likely the result of such words being observed in the training data as part of a procedure label. However, such words on their own rarely reflected a medical procedure.

However, some short words did correspond to medical procedures, for example medical abbreviations, though these were most often expressed using capital letters (e.g. CIN3). We therefore concluded, that we should aim to remove only lowercase words of length 5 or less, thus preserving the abbreviations of procedures and other medical terms.

To achieve this, for each lowercase word of length 5 or less that was labeled a procedure, we combined the predictions of our three best models by using a majority voting rule. When 2 out of the 3 models agreed on whether that particular token is a named entity, that was reflected in our voting ensemble classifier.

6.3. Regular expressions search

We also employed a simple, yet very effective approach of using regular expressions to match snippets of the data to the named entities provided in the gazetteer. We expected that named entities identified through this approach would improve both recall and precision. This was indeed confirmed when comparing the results on the validation set (discussed in the next section).

As initial tests with the gazetteer provided in the challenge resulted in significant number of matches, we decided to expand the gazetteer with additional labels of procedures as described in section 4.3.

On the test set, combining the outputs of the regular expressions search using the expanded procedures gazetteer with the outputs of the best performing RoBERTa models resulted in more than 3-fold increase in the number of named entities recognised compared to the best RoBERTa model alone.

7. Results

Table 1 shows the precision, recall and f1-score for our best 3 models without the gazetteer, as well as the results of our voting ensemble classifier with and without the gazetteer exact match.

We observe that the voting ensemble results in only a minor improvement over our best individual model, which is the RoBERTa model pre-trained for the PharmaCoNER challenge. However, combining this with the results form our regular expressions search results in a significant improvement of precision (+4%), recall (+6%) and f1-score (+4%).

Table 1
Experimental results of systems on development set. Model name is followed by class weight ratio, followed by number of epochs (e.g. _10_1_20 stands for 10:1 ratio for 20 epochs).

Model	Precision	Recall	F1
MedProcNER_esp_ehr_pharmaconer_10_1_20	0,51	0,62	0,56
MedProcNER_esp_ehr_cantemist_10_1_15	0,49	0,63	0,55
MedProcNER_esp_ehr_5_1_10	0,46	0,61	0,53
MedProcNER_esp_voting	0,52	0,62	0,57
MedProcNER_esp_voting_gazetteer	0,56	0,68	**0,61**

As for the test set, our results differ from those on the validation set. As mentioned in the previous section and also shown in the Appendix, on the validation set we consistently observed that all model variations with all hyperparameter combinations produced results with higher recall and lower precision.

However, on the test set, our best individual model, as well as the one using a voting ensemble to filter out short words, had a precision score of 0.74 and recall of 0.43, as can be seen in Table 2. The specific evaluation script for the challenge has not been released as of writing this paper, hence we are not able to conclusively reason on the contributing factors to this difference.

Furthermore, adding the extended gazetteer exact match outputs to the voting ensemble classifier resulted in a significant drop in precision (from 0.74 down to 0.32) and, although recall

improved (from 0.43 to 0.61), overall f1-score dropped from 0.55 to 0.43. Post factum we tend to explain this perplexing result by the nature of entities presented in the gazetteer: not full text entities are annotated in the text, but rather parts of them, therefore greedy match extracts more tokens then needed.

Table 2
Experimental results of submitted systems on the test set

Model	Precision	Recall	F1
MedProcNER_esp_ehr_pharmaconer	0,74	0,43	0,54
MedProcNER_esp_voting	0,74	0,43	**0,55**
MedProcNER_esp_voting_gazetteer	0,32	0,61	0.43

8. Conclusion and Further Work

To summarize, this paper introduces a robust baseline for clinical procedure recognition with biomedical Spanish language models by leveraging a pre-trained transformer network and fine-tuning it for Named Entity Recognition (NER) coupled with ontology-based exact match. Our NER approach provides a solid solution for the task, as demonstrated by the official results of MedProcNER track at BioASQ 2023 challenge.

In future work, an interesting path to explore is the utilization of a gazetteer for synthetic data generation in the context of our paper. By incorporating a comprehensive biomedical ontology-based terms, we can augment the existing dataset with additional annotated entities, thereby expanding the training data for our NER models. This approach has the potential to enhance the generalization ability of the approach, enabling it to handle a wider range of clinical cases. Moreover, by leveraging the gazetteer to generate synthetic data, we can potentially address the limited availability of annotated clinical text in Spanish, which is often a limiting factor for model performance.

References

[1] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[2] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), ????

[3] J. Castell-Díaz, J. A. Miñarro-Giménez, C. Martínez-Costa, Supporting snomed ct post-coordination with knowledge graph embeddings, Journal of Biomedical Informatics 139 (2023) 104297.

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

[5] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342 (2019).

[6] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical nlp in spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 193–199.

[7] S. Amin, N. P. Goldstein, M. K. Wixted, A. García-Rudolph, C. Martínez-Costa, G. Neumann, Few-shot cross-lingual transfer for coarse-grained de-identification of code-mixed clinical texts, arXiv preprint arXiv:2204.04775 (2022).

[8] Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, et al., Deid-gpt: Zero-shot medical text de-identification by gpt-4, arXiv preprint arXiv:2303.11032 (2023).

[9] R. Ahmed, P. Berntsson, A. Skafte, S. K. Rashed, M. Klang, A. Barvesten, O. Olde, W. Lindholm, A. L. Arrizabalaga, P. Nugues, et al., Easyner: A customizable easy-to-use pipeline for deep learning-and dictionary-based named entity recognition from medical text, arXiv preprint arXiv:2304.07805 (2023).

[10] S. L. López, E. F. Maduell, L. G. Sánchez, M. Krallinger, MedProcNER/ProcTEMIST Corpus: Gold Standard annotations for Clinical Procedures Information Extraction, 2023. URL: https://doi.org/10.5281/zenodo.7886453. doi:10.5281/zenodo.7886453.

[11] A. Intxaurrondo, Spaccc, 2018. URL: https://doi.org/10.5281/zenodo.2560316. doi:10.5281/zenodo.2560316, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[12] S. L. López, E. F. Maduell, L. G. Sánchez, M. Krallinger, MedProcNER/ProcTEMIST Guidelines: Annotation and Normalization of Clinical Procedures in Medical Documents, 2023. URL: https://doi.org/10.5281/zenodo.7817667. doi:10.5281/zenodo.7817667.

[13] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, in: W. Daelemans, M. Osborne (Eds.), Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 142–147.

[14] V. Moscato, M. Postiglione, G. Sperlí, Biomedical spanish language models for entity recognition and linking at bioasq distemist (2022).

[15] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: https://aclanthology.org/D19-5701. doi:10.18653/v1/D19-5701.

[16] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.

[18] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.

9. Appendix

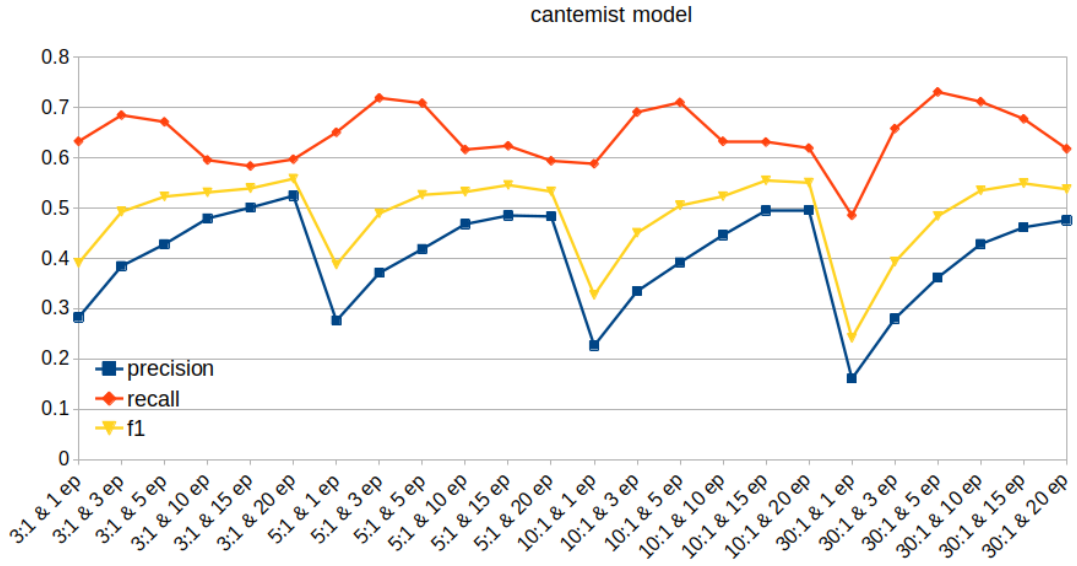

Figure 1: Precision (blue), recall (red) and f1-score (yellow) for all combinations of class weight ratios & number of epochs tested with the CANTEMIST model. X-axis runs from smallest class weight ratio (3:1) and smallest number of epochs (1 ep) to largest (30:1 & 20 ep).

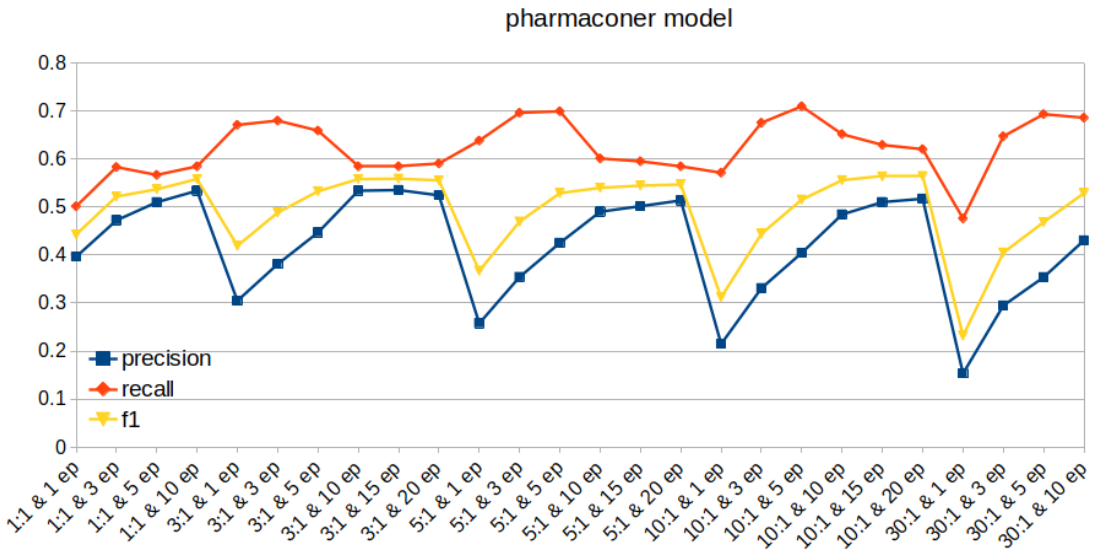

Figure 2: Precision (blue), recall (red) and f1-score (yellow) for all combinations of class weight ratios & number of epochs tested with the PharmaCoNER model. X-axis runs from smallest class weight ratio (1:1) and smallest number of epochs (1 ep) to largest (30:1 & 10 ep).