Evaluation of Named Entity Recognition in Historical Argentinian Documents

Facundo Darfe[1], Eduardo Xamena[1,2], and Carlos I. Orozco[1]

[1] Departamento de Informática - Fc. Cs. Exactas - Universidad Nacional de Salta (UNSa)
[2] Inst. de Investigación en Cs. Sociales y Humanidades, ICSOH - CONICET - UNSa, Av. Bolivia 5150, Salta, CP 4400 examena@di.unsa.edu.ar

Abstract. Research over historical text volumes can be performed by means of automatic tools that help historians achieve more abstract and aggregated points of view. Tasks such as Information Extraction or Text Mining can be performed more efficiently if Machine Learning models are employed. We propose the evaluation of different state-of-the-art models over a new dataset for Named Entity Recognition. The dataset was built over a History texts volume about General Güemes, a national Argentinian independence hero. The results show that some models perform better in terms of precision, recall and f1-score for most types of entities. Specifically, pretrained language models fine-tuned for this particular task show considerably higher performance than classical models based on word embeddings and other kinds of representations and models. Besides, statistical tests are provided to ensure the significance in the differences of the performance values attained. Hence, the contribution of this work is twofold, on the one hand a new corpus and dataset for Named Entity Recognition and a complete statistical assessment of performance values of state-of-the-art models over the generated dataset.

Keywords: Named Entity Recognition and Classification · Argentinian History · Pretrained Language Models.

1 Introduction

Information retrieval tasks have gained importance in the efficient formulation of hypothesis and conclusions arising from rigorous research over different scopes of history, social sciences and related disciplines. The different characters, facts and dates of interest can be identified and retrieved from the documents that make up documentary funds. This is a first step for building knowledge bases that allow mining relationships among the found entities. This is the context where Named Entity Recognition (NER) becomes fundamental [14]. There are different approaches for performing a NER task in an automated way: Searching for regular expressions matches into the original text strings that identify the presence of typical text structures of entity references, or using Machine Learning (ML) tools to detect the type of element (entity or not) in each position of the original sequence. While methods based on regular expressions can provide very good results in general [4], a lot of important information can be lost due to the multiple ways of writing a name or an address. Even if we can typify all possible ways of writing for references to entities of different classes, the algorithm that finds the matches with corresponding regular expressions will become highly inefficient as the number of possibilities grows. There may also be small typing errors or subtle differences in the typing that cause the loss of relevant information, and/or the need for a constant update on the algorithm's regular expressions list.

The use of ML tools to perform NER tasks allows the generation of software pieces adapted to the original documents' context [5]. The associated cost –as in any task involving ML– is the necessity of training data volumes and the corresponding training process for each context. However, depending on the accuracy (and other ML metrics) of the developed models for the different types of entities and documents, they can perform at an acceptable level when employed in similar contexts [8]. For instance, in the context of text volumes that share syntactic and grammatical styles, an ML model trained on one of these contexts could likely achieve similar performance on the other volumes, or could be efficiently fine-tuned by adjusting the weights of its outer layers.

This NER task is closely related to the objectives of the associated research project, as it solves an immediate, natural and fundamental information demand for different stages in data analysis, search and visualization, among others [20]. The search for interactions between characters, dates, facts and other types of entities in historical documentary collections cannot be carried out without previously identifying all entity references. This process is linked with other works developed under our project, such as the automated transcription of manuscripts [18] and the OCR post-processing of the available printed documents [19, 12], as a consecutive step after such tasks. In addition, the subsequent phase of information extraction by recognizing relations between entities presents a total dependence on the availability of a NER mechanism. In this sense, NER plays a central role in the complete workflow.

The starting point for building NER models and software tools for our purpose is supported by the most promising and relevant software and results in this task, namely the state of the art in NER, both for production and research instances. As a baseline we take the models and software developed by Honnibal and Montani [6] (Spacy software), Manning et al [11] (Stanford CoreNLP toolkit), and Transformer-based models as Multilingual BERT [3], BETO [2] (Spanish language pretrained version of BERT from the Computer Science Department - Universidad de Chile) and SpanBERTa [3].

This work has two main contributions: First, a new dataset has been released for the evaluation of NER models in a historical context, specifically for Argentinian texts of the Revolution period (19th century); and second, a set of models and scripts have been made available to the community, for performing the task of NER in documentary volumes of the same period and location. Such

[3] Available at: https://skimai.com/roberta-language-model-for-spanish/

scripts and models can be reused for different tasks, hopefully with high levels of precision, recall and f1-score.

Next sections describe the following topics. The Related work section shows approaches to tackle the NER problem using ML tools, from Conditional Random Fields to Transformer-based models. Next, the Methodology section illustrates how the dataset was built and the process of implementation of the different tools employed for the NER task. Then, the Results section shows the key aspects of the experiments carried out. Finally, the conclusions about results and future research lines are expressed.

2 Related work

The NER task is particularly hard to approach by means of non-ML techniques, given the stochastic nature of language in general. A NER tool should be aware of too many complex rules that change all the time, with every new language, dialect or speaker/writer in general. Many ML techniques have been employed for solving this sequential tagging task. For instance, Huang et al [7], employ a combination of Convolutional and Recurrent Neural Networks (CNN-LSTM), representing texts with byte-pair encoding and achieving interesting results in popular corpora such as CONLL 2003. Another more recent approach to NER [22] takes care of the possibility of existing unlabeled entities in the datasets, tuning some specific parameters over ML models, also keeping and enhancing performance values.

In the context of software packages for production environments, many projects have been developed making use of ML tools such as pre-trained language models. Spacy [6] is a software project that makes use of different word embedding architectures and state-of-the-art language models for Natural Language Processing (NLP) tasks in its last versions. Stanford CoreNLP is another software developed for NLP tasks, that implements ML models trained over big corpora in different languages.

Recent research works put the focus on trending kinds of ML instances, mostly employing Transformer-based language models [16]. Yamada et al [21] employ entity embeddings based on knowledge bases, combining it with pre-trained language models. They show different experiments performed on known datasets from the NER literature. Their results are part of the state of the art in the discipline. Besides, models for more specific tasks can be achieved by means of these language models, as the case of the work of Alkomah (2022) [1], where the goal is the detection of fraud by means of NER in clinical texts.

3 Methodology

Given the final goal of developing an ML platform for NER in historical texts of the Revolution period in Argentina, the complete task involved the phases of corpus building, state-of-the-art tools evaluation, fine-tuning of pre-trained language models and a test phase for the tuned models. Next subsections explain the NER task and each enumerated phase for this project.

3.1 The NER task explained

NER is essentially a sequence labelling problem, where the objective is to predict the class each token belongs to. The corresponding token concept for this problem is associated with the words of the text sequence, i.e. each word in the text represents a token of the related sequence. In the first place, the detection task consists of a binary classification problem for each token, determining if it belongs to an entity or not. On the other hand, the recognition task involves stating the class an entity belongs to, among an arbitrary number of possible entity types.

According to the context of the particular NER task performed, there is a set of classes previously identified for the recognition task. For instance, in most information retrieval systems it is useful to look for people, dates or places, but in more specific contexts such as medical reports, the entities could be diseases or medical studies.

Given the natural class imbalance that exists in this kind of problems (many non-entity tokens compared to entity tokens), the accuracy metric is not suitable for measuring performance. Instead of that, metrics like precision, recall and f1-score provide realistic performance results, avoiding the possibility of getting good performance by only returning 'non-entity' for every token.

3.2 Building a corpus for NER in historical documents of Argentina

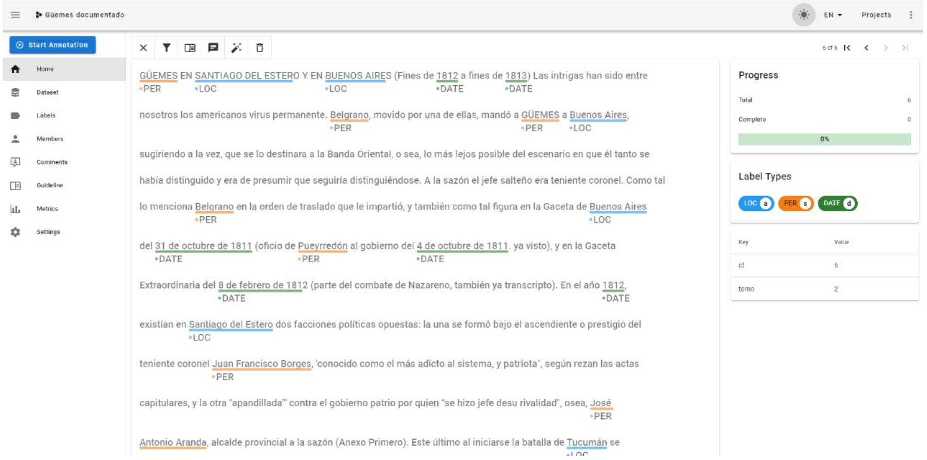
The texts volume employed for this task is "Güemes Documentado"[4] (GD). It consists of a set of letters and legal documents related to General Güemes, and a narrative text that accompanies them. In this work, a subset of the first and second volumes of GD (specifically 30 pages of volume 1 and 90 of volume 2) have been annotated for NER. Table 1 summarizes the composition of the generated dataset. The entity types annotated are Date (DATE), Location (LOC) and Person (PER). Most entities belong to the class PER, and there are 94.49% of Outer tokens (i.e. text that does not belong to any type of entity). This number denotes the classical imbalance problem present in corpora annotated for NER.

Table 1. Number and percentage of different types of entities in the text.

Type of entity	Number of tokens	Percentage
OUTER	45415	94.49%
PER	1254	2.61%
LOC	837	1.74%
DATE	559	1.16%

[4] http://www.portaldesalta.gov.ar/documentado.html

The annotation tool adopted for this task was Doccano [5], given the ease of use and installation it provides. Fig. 1 shows an instance of this tool for the annotation process of GD. In that figure the different visualization features of the Doccano tool can be appreciated.



GÜEMES EN SANTIAGO DEL ESTERO Y EN BUENOS AIRES (Fines de 1812 a fines de 1813) Las intrigas han sido entre
•PER •LOC •LOC •DATE •DATE

nosotros los americanos virus permanente. Belgrano, movido por una de ellas, mandó a GÜEMES a Buenos Aires,
•PER •PER •LOC

sugiriendo a la vez, que se lo destinara a la Banda Oriental, o sea, lo más lejos posible del escenario en que él tanto se

había distinguido y era de presumir que seguiría distinguiéndose. A la sazón el jefe salteño era teniente coronel. Como tal

lo menciona Belgrano en la orden de traslado que le impartió, y también como tal figura en la Gaceta de Buenos Aires
•PER •LOC

del 31 de octubre de 1811 (oficio de Pueyrredón al gobierno del 4 de octubre de 1811. ya visto), y en la Gaceta
•DATE •PER •DATE

Extraordinaria del 8 de febrero de 1812 (parte del combate de Nazareno, también ya transcripto). En el año 1812,
•DATE •DATE

existían en Santiago del Estero dos facciones políticas opuestas: la una se formó bajo el ascendiente o prestigio del
•LOC

teniente coronel Juan Francisco Borges, 'conocido como el más adicto al sistema, y patriota", según rezan las actas
•PER

capitulares, y la otra "apandillada"' contra el gobierno patrio por quien "se hizo jefe desu rivalidad", osea, José
•PER

Antonio Aranda, alcalde provincial a la sazón (Anexo Primero). Este último al iniciarse la batalla de Tucumán se

Progress

Total 6

Complete 0

0%

Key	Value
id	6
tomo	2

Fig. 1. Visualization of the Doccano tool for the GD annotation process.

Prior to the annotation phase with Doccano, a cleaning phase was carried on with the covered GD texts. This phase is known as post-OCR correction, and is necessary due to the pervasive amount of errors that optical text acquisition tools often have. As part of the same research project, post-OCR correction has approached from an ML perspective [13] in parallel, and we expect to use the developed tools on the rest of the GD texts. Once the corpus was cleaned and annotated, the next phase was a transcription of the Doccano format to the corresponding format of each technological tool employed. In the case of Stanford NLP we did not train nor fine-tune new models. Instead, we evaluated this out of the box. Hence, the format migration consisted only in capturing the tokens from the annotation source and tagging them with the tool. For the case of Spacy and pre-trained models based on Transformers, a conversion from the original Doccano format to the CoNLL format [15] was required. Then, each token and its related tags are identified with a line in the corpus file, separated by a blank space (one-token-per-line format).

[5] https://doccano.herokuapp.com/

3.3 Software packages for the NER task

In order to test the performance of available software tools for NER, two packages were evaluated: Spacy [6] from the Explosion group, and Stanford Core-NLP [7] from the Stanford NLP Group. For the Stanford case, special tokenization routines were necessary for adapting the tokens to the Ground Truth (GT) volume. This is caused by the specific tokenization routines in this case, that do not match exactly with the words tagged in the GT, avoiding the calculation of ML metrics for NER performance. After including the custom tokens, such performance metrics were computed for this software tool in the present task.

Spacy is a complete platform for NLP tasks designed for production environments. In it's last versions, Transformer-based models such as BERT [3] and derivatives were included given the high performance they achieve in many cases. Formerly, this suite only included classical models based on the Common Random Fields (CRF) or Neural Networks techniques, using specific word embeddings for each language. In the present work we report very promising results of this tool for GD texts.

Stanford CoreNLP is a classical library with CRF models trained over Ancora and CoNLL datasets. The different models available for this library allow alternatives for each task to be carried on. For this task, a CRF model based on the Ancora corpus was used. Resutls do not show higher performance with this model than those attained with Spacy or Transformer models, but it is surely due to the advances in the techniques employed. CRF comprises an older methodology than Transformer-based models.

3.4 ML Transformer-based models

The current state of the art in most NLP tasks has been governed by Transformer-based models, starting with the appearance of BERT [3]. The ease of use derived from libraries and frameworks such as HuggingFace Transformers [17] allowed the availability and reproducibility of popular baselines for the complete NLP community. Besides, more than fine-tunning efforts for transfer learning tasks, some authors performed complete training tasks from scratch with the goal of generating brand new specific NLP models for other languages, with Spanish among them. Even though BERT Miltilingual is one of the variants of BERT and actually includes Spanish, in this work two specific Spanish language models based on BERT derivatives were fine-tuned and evaluated for our task: SpanBERTa and BETO, in order to compare performances.

SpanBERTa is a RoBERTa-based [9] model trained from scratch for Spanish language by the Skim AI Company [8]. The corpus used for this purpose is the Spanish fraction of OSCAR [9], a huge multilingual corpus obtained by language

[6] https://spacy.io/

[7] https://stanfordnlp.github.io/CoreNLP/

[8] https://skimai.com/

[9] https://oscar-corpus.com/

classification and filtering tasks. In this work we employ the SpanBERTa original model and fine-tune it for the NER task on GD corpus.

Another Spanish language model based on a BERT architecture is BETO [2] from DCC-UChile. BETO has been built and trained with the BERT basic architecture, and has 110M parameters. The data employed for the training process was acquired from Wikipedia, Spanish blogs, Subtitles and many other sources, making up a comprehensive repository for the task.

4 Results

This section summarizes the results obtained with each one of the described tools and models. As the NER task inherently includes imbalanced datasets, plain accuracy metric is not suitable for the evaluation of models. Then, Precision, Recall and F1-score are reported for each case.

For the case of the Stanford NLP tool, we only report the performance in the complete corpus, using an Ancora CRF model out of the box. Hence, the use of such tool and model on the described dataset resulted in a Precision of 70%, a Recall value of 52.33% and hence the F1-score was 59.89%.

Spacy offers a complete suite of tools for carrying out training procedures for it's ML NLP models. The dataset was split into three parts, 60% for training the models for the task, 20% for validation set and 20% for the test set. The fine-tuning and assessment of an existing Spacy model for this task resulted in considerably higher levels of F1-score, Precision and Recall, compared to Stanford tool. Even though, the achieved values do not outperform the results of using state-of-the-art Transformer-based models, as reported in the next paragraphs.

Apart from testing the Spacy and Stanford tools for the present NER task, we report the results of three state-of-the-art Transformer-based models, one multilingual and two specifically pre-trained for Spanish language. The complete process of training and the new annotated corpus are accessible through the Github site of one of the authors [10]. In order to have statistical significant volumes of data, each model has been fine-tuned and evaluated with three different numbers of training epochs, and each configuration had 30 independent runs. As done for the Spacy suite, 60% of the tokens of the dataset were used for training, 20% for the validation set and 20% for the test phase. Table 2 and figure 2 show the mean values of performance on the test set for each metric and training setup. A first trend can be observed having BETO as the best performing model for most metrics and numbers of training epochs. Such trend is an effect of higher values in Recall for BETO, but the Precision metric is only higher in BETO for 30 epochs, while BERT-base Multilingual outperforms BETO in 10 and 20 epochs. This fact could indicate that BETO is the best model for the current NER task, but requires more fine-tuning resources. Another direct consequence of this analysis suggests that maybe SpanBERTa is not a very good model for this task, despite the high performance values.

[10] https://github.com/facudarfe/NERC-GuemesDocumentado

Table 2. Mean values for Precision, Recall and F1-score for 30 runs of each model and training epochs configuration.

Model		BERT-base Multilingual	SpanBERTa	BETO
10 epochs	Precision	**81,95%**	78,69%	81,57%
	Recall	82,50%	74,78%	86,41%
	F1	82,21%	76,98%	83,91%
20 epochs	Precision	81,37%	77,12%	81,04%
	Recall	82,89%	75,03%	86,94%
	F1	82,11%	76,03%	83,88%
30 epochs	Precision	80,92%	76,53%	81,42%
	Recall	82,71%	75,29%	**87,16%**
	F1	81,80%	75,87%	**84,19%**

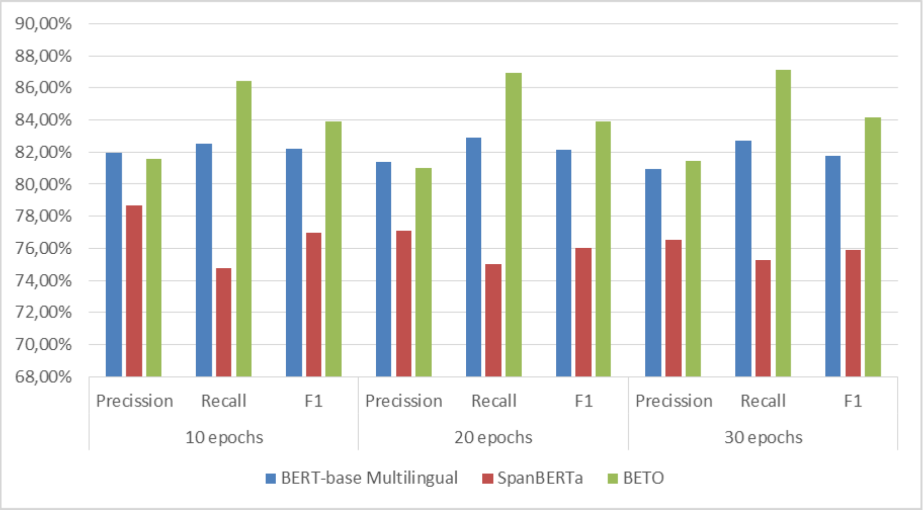


Fig. 2. Average values of metrics for each model and training configuration.

Figure 3 shows three graphs of the performance of each run in the fine-tuning experiments for the different models. As can be seen there, BETO outperforms other models with F1 values that reach more than 86.15% and always achieves more than 81%. BERT-base Multilingual F1-score moves over the range between 79%-85%, a few steps lower than BETO. Finally, SpanBERTa shows values between 75% and 80% of F1, but sometimes this value is even lower than 60%. In terms of the observation of these graphs, BETO seems to be the best contender for this task, but the statistical significance of this difference should be assessed.

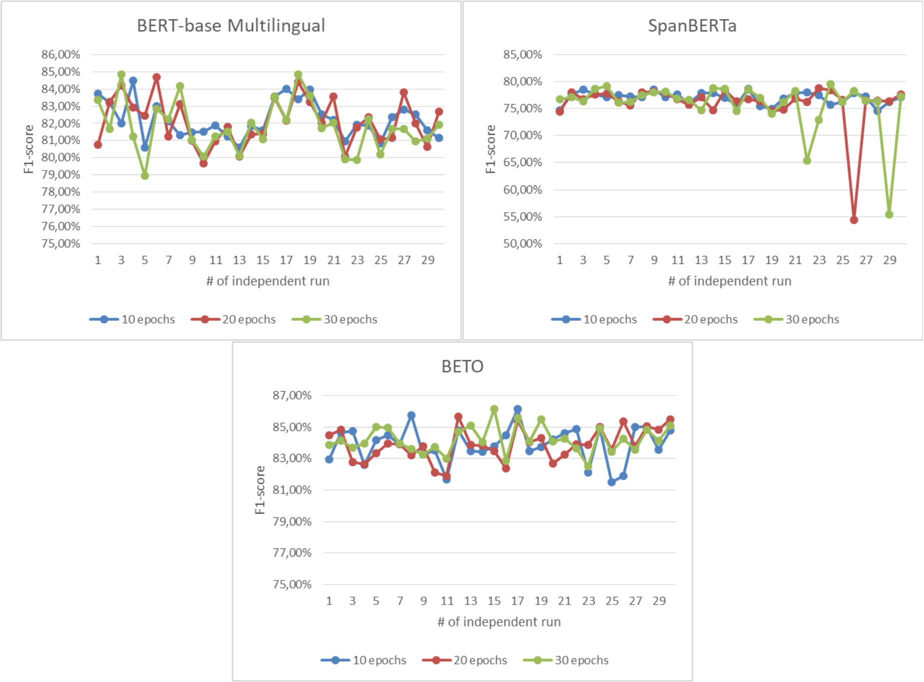
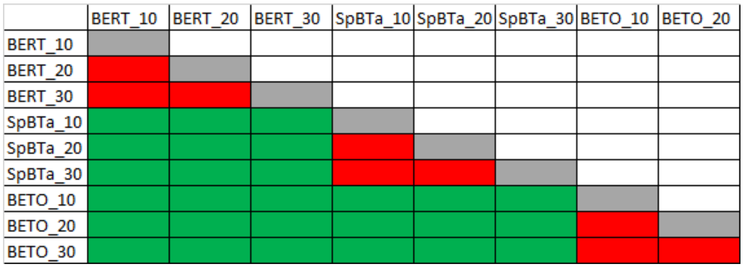


Fig. 3. F1-score for independent training and evaluation runs on the models BERT-base Multilingual, SpanBERTa and BETO for the NER task.

In order to state the statistical significance of the difference between configurations in the experiments performed, the Mann-Whitney U rank test [10] was taken for each pair of samples. This test assumes as null hypothesis that the two corresponding samples come from the same population. By rejecting the null hypothesis, the difference between media and standard deviation values turns to be statistically significant.

Figure 4 shows the results obtained for the Mann-Whitney U rank test for every combination of models and configurations. In this figure, red boxes reflect non-significant differences on the training configurations represented by the corresponding row and column in the table. On the other side, green boxes stand for significant p-values, hence denoting instances that are statistically different in F1-score. For instance, it is remarkable that the differences between different configurations of the same model have no significant p-values for the test, taking as an example BERT_10, BERT_20 and BERT_30, while every comparison between two different models yielded significant p-values as well. This results confirm the statements about which model should be considered the best choice for the current NER task.



	BERT_10	BERT_20	BERT_30	SpBTa_10	SpBTa_20	SpBTa_30	BETO_10	BETO_20
BERT_10	▨							
BERT_20	🟥	▨						
BERT_30	🟥	🟥	▨					
SpBTa_10	🟩	🟩	🟩	▨				
SpBTa_20	🟩	🟩	🟩	🟥	▨			
SpBTa_30	🟩	🟩	🟩	🟥	🟥	▨		
BETO_10	🟩	🟩	🟩	🟩	🟩	🟩	▨	
BETO_20	🟩	🟩	🟩	🟩	🟩	🟩	🟥	▨
BETO_30	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥

Fig. 4. Results of Mann-Whitney U rank test over F1-score of pairs of samples for each configuration of model and training epochs. Red-coloured boxes indicate non-significant p-values and green-coloured boxes indicate the pairs that exhibit a significant difference of F1-score.

5 Conclusions and Future work

In this work a complete NER task has been reported. First, a corpus for NER has been generated and disposed for the use of the worldwide scientific community, for Spanish language. This corpus consists of historical texts from the Argentinian Revolution period, regarding the life of General Güemes. For the annotation process the Doccano tool was employed, and the entity types worked were people, places and dates. The new corpus represents a very valuable contribution of this work.

We report results for the NER task over this new corpus by means of two different types of software tools: Classical production NLP tools and state of the art Transformer-based pre-trained models. For the case of the former production tools, the results achieved are not as high as the performance of Transformer-based models. Besides, a statistical test was carried on to enforce the difference among median values of performance of Transformer-based models. Hence, we can conclude that the pre-trained model BETO achieved the best performance values in the related experiments.

With the purpose of enhancing the results obtained so far, one prospect of future work will be the exploration of data augmentation techniques on the dataset for acquiring better performance values. Besides, another important work will be the inclusion of previously studied post-OCR processing models for the semi-automation of the text acquisition process. This way, a very modern pipeline for historical texts processing could be available for historians and humanities researchers.

Acknowledgements

This research has been developed with funds from projects B-2825 and A-2735 (Consejo de Investigación de la Universidad Nacional de Salta - CIUNSa) and Project PUE 22920160100056CO (CONICET - Proyecto de Unidad Ejecutora, ICSOH), and Universidad Nacional del Sur (PGI-UNS 24/N051). Besides, this research was enabled in part by Compute Canada (www.computecanada.ca) and the Natural Sciences and Engineering Research Council of Canada.

References

1. Alkomah, B.: Abuse Detection in Medical Claims Using NLP and Deep Learning Techniques. Ph.D. thesis, University of Idaho (2022)
2. Canete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. Practical Machine Learning for Developing Countries (PML4DC) workshop - ICLR **2020**, 2020 (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Ek, T., Kirkegaard, C., Jonsson, H., Nugues, P.: Named entity recognition for short text messages. Procedia-Social and Behavioral Sciences **27**, 178–187 (2011)
5. Emelyanov, A.A., Artemova, E.: Multilingual named entity recognition using pretrained embeddings, attention mechanism and ncrf. arXiv preprint arXiv:1906.09978 (2019)
6. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
7. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
8. Lee, J.Y., Dernoncourt, F., Szolovits, P.: Transfer learning for named-entity recognition with neural networks. arXiv preprint arXiv:1705.06273 (2017)
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
10. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics pp. 50–60 (1947)
11. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)

12. Mechaca, A.L., Marmanillo, W.G., Xamena, E., Ramirez-Orta, J., Maguitman, A.G., Milios, E.E.: A web platform for collaborative semi-automatic ocr post-processing. In: VII Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2021)-JAIIO 50 (Modalidad virtual) (2021)
13. Ramirez-Orta, J., Xamena, E., Maguitman, A., Milios, E., Soto, A.J.: Post-ocr document correction with large ensembles of character sequence-to-sequence models. ArXiv (2022)
14. Ruiz, P.: Concept-based and Relation-based Corpus Navigation: Applications of Natural Language Processing in Digital Humanities.(Navigation en corpus fondée sur les concepts et les relations: Applications du Traitement automatique des langues aux Humanités numériques). Ph.D. thesis, École Normale Supérieure, Paris, France (2017)
15. Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
18. Xamena, E., Barbosa, H., Orozco, C.I.: Evaluación de una plataforma completa para reconocimiento de textos manuscritos en español. Ciencia y tecnología **2021**(21), 6 (2021)
19. Xamena, E., Maguitman, A.G.: Language modeling tools for massive historical ocr post-processing. In: VI Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2020)-JAIIO 49 (Modalidad virtual) (2020)
20. Xamena, E., Marmanillo, W.G., Mechaca, A.L.: Rebuilding the story of a hero: Information extraction in ancient argentinian texts. In: V Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2019)-JAIIO 48 (Salta) (2019)
21. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: Luke: deep contextualized entity representations with entity-aware self-attention. arXiv preprint arXiv:2010.01057 (2020)
22. Zhang, F., Ma, L., Wang, J., Cheng, J.: An mrc and adaptive positive-unlabeled learning framework for incompletely labeled named entity recognition. International Journal of Intelligent Systems (2022)