Pre-trained language models to extract information from radiological reports

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

[1]*Universidad de Jaén, Campus Las Lagunillas s/n, E-23071, Jaén, Spain*

Abstract

This paper describes the participation of the SINAI team in the SpRadIE challenge: Information Extraction from Spanish radiology reports which consists of identifying biomedical entities related to the radiological domain. There have been many tasks focused on extracting relevant information from clinical texts, however, no previous task has been centered on radiology using Spanish as the main language. Detecting relevant information automatically in biomedical texts is a crucial task because current health information systems are not prepared to analyze and extract this knowledge due to the time and cost involved in processing it manually. To accomplish this task, we propose two approaches based on pre-trained models using the BERT architecture. Specifically, we use a multi-class classification model, a binary classification model and a pipeline model for entity identification. The results are encouraging since we improved the average of the participants by obtaining a 73.7% F1-score using the binary system.

Keywords

Biomedical information extraction, Radiological domain, Spanish clinical reports, Pre-trained language models, BERT,

1. Introduction

Medical texts such as radiology reports or Electronic Health Records (EHR) are a powerful source of data for researchers. These data sources contain relevant information that can help in clinical decision-making and report structuring, among other benefits. However, current health information systems are not prepared to analyze and extract knowledge due to the time and cost involved in processing it manually. The field of artificial intelligence known as Natural Language Processing (NLP) is being applied to medical documents to build applications that can understand and analyze this huge amount of textual information automatically [1].

This paper describes the system presented by the SINAI team for the SpRadIE (Information extraction from Spanish Clinical Reports) challenge [2] at CLEF eHealth 2021 [3] (Task 1). SpRadIE challenge focuses on information extraction from Spanish biomedical texts, more specifically, on the NER (Named Entity Recognition) task. Spanish has more than 480 million native speakers[1]

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

plubeda@ujaen.es (P. López-Úbeda); mcdiaz@ujaen.es (M. C. Díaz-Galiano); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia)

0000-0003-0478-743X (P. López-Úbeda); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. Martín-Valdivia)

[1]Spanish language: https://en.wikipedia.org/wiki/Spanish_language

and nowadays there is a worldwide interest in processing medical texts in this language. For this challenge, our proposal is focused on pre-trained models based on the transformer architecture using BERT (Bidirectional Encoder Representations from Transformers). More specifically, we employ the BETO model trained on a large corpus in Spanish to address the task. For this purpose, we submitted three different systems: a BERT model with multi-class classification to perform entity extraction, BERT using binary classification, and a combined model.

The rest of the paper is structured as follows: In Section 2 we present some previous studies related to information extraction in radiology reports. The dataset, the pre-processing carried out and the descriptions of the implemented systems are presented in Section 3. Section 4 provides the results achieved for the SpRadIE challenge. Finally, conclusions and future work are presented in Section 5.

2. Related Work

In the NLP literature, studies related to the biomedical domain are focused on specific subdomains such as radiology. NLP techniques can be an interesting tool for successfully analyzing radiology reports to extract clinically important findings [4]. In order to be effective in the treatment of significant information, it is necessary to have specialized and domain-focused corpora. On the one hand, PadChest [5] dataset contains 27,593 reports in Spanish that were manually annotated by trained physicians. The reports included in PadChest were labeled with 174 different radiographic findings, 19 differential diagnoses, and 104 anatomic locations. On the other hand, there are other datasets available to the scientific community such as Chestx-ray8 [6], PLCO [7], CheXpert [8], and MIMIC-CXR [9], which use English as the principal language.

The NLP clinical community organized a series of open challenges to identify and extract relevant information included in clinical reports. These challenges highlight the importance of tackling this type of task, offering participants the opportunity to submit novel systems and compare their results using the same dataset. Some examples of these challenges are PharmaCoNER [10], eHealth-KD [11], CLEF eHealth [12], MEDDOCAN [13], CHEMDNER [14], and I2B2 [15].

Researchers interested in information extraction tasks have explored a variety of Machine Learning (ML) approaches [16]. Previous studies applied the CRF method [17] in order to perform the identification and subsequent classification of entities. CRF is the most popular solution among conventional ML algorithms. Recently, deep learning has become prevalent in the ML research community to improve biomedical named entity recognition using different neural networks such as BiLSTM-CRF [18, 19]. Although Recurrent Neural Networks (RNNs) have obtained high results and a wide range of related literature on the NER task in recent years, the pre-training of Transformer-based language models such as BERT [20] has also led to impressive gains in NER systems.

Based on the idea of using the latest methods employed in the area of information extraction, our study uses BERT as a model to identify named entities in radiological reports written in Spanish.

3. System Overview

3.1. Dataset

SpRadIE dataset consists of 513 ultrasonography reports provided by a pediatric hospital in Argentina. Because the reports are written by specialists, they are unstructured and have abundant spelling and grammatical errors. For confidentiality, each report has been anonymized by removing patient information, names, and physician registration numbers.

The dataset was annotated by clinical experts [21] and then revised by linguists and it is composed of 175 reports in the training set, 92 reports in the development set and 207 for the test set.

Seven entity types and three radiological concept hedging cues are distinguished.These entities may be very long, sometimes even spanning over sentence boundaries, embedded within other entities of different types and may be discontinuous. Moreover, different text strings may be used to refer to the same entity, including abbreviations and typos. Some of the most relevant statistics of the types of entities in each dataset along with some examples are shown in Table 1.

Table 1
Statistics on the number of entities in each dataset.

	Training		Development		Example
	# entities	# uniques	# entities	# uniques	
Anatomical Entity	1335	179	868	167	*ambos riñones* (both kidneys)
Finding	825	289	698	295	*líquido libre* (free fluid)
Location	529	104	286	93	*cavidad* (cavity)
Measure	596	389	369	268	10 cm
Type of Measure	357	36	163	30	*diametro longitudinal* (longitudinal diameter)
Degree	41	19	35	22	*leve* (slight)
Abbreviation	903	57	538	59	cm
Negation	496	29	282	36	*sin* (without)
Uncertainty	55	16	22	12	*compatible* (compatible)
Conditional Temporal	12	8	7	4	*antecedente* (antecedent)

3.2. Pre-processing

The initial step in data science is data preparation or text pre-processing. In our particular case, we work with texts written in Spanish and related to the radiology domain. The pre-processing carried out in all the texts is the following:

- **Sentence tokenization**. This process consists of splitting the text into individual sentences. For this purpose, we use the FreeLing library [22] that incorporates analysis functionalities for a variety of languages, including Spanish.
- **Word tokenization**. This step converts text strings to streams of token objects, where each token object is a separate word, punctuation sign, number/amount, date, etc. In this

step, we also use the FreeLing library. This library offers an optimal result for clinical texts since in some cases, e.g. the sentence "*Compatibles con hipertrofia pilorica*" is separated into the following tokens: "*Compatibles*", "*con*", "*hipertrofia_pilorica*". Where the token *hipertrofia_pilorica* is annotated with the category Findings.

- **Lowercase**. The texts have been converted to lowercase.
- **BIO tagging scheme**. The final step in pre-processing the reports involves converting to CoNLL format using the BIO tagging scheme [23]. Thus each token in a sentence was labeled with B (beginning token of an entity), I (inside token of an entity), and O (non-entity).

3.3. Methodology

The methodology employed for the achievement of the task is based on BERT architecture. BERT [20] uses a Transformer [24] architecture and is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. Moreover, BERT proposes a Masked Language Modelling (MLM) objective, where some of the tokens of an input sequence are randomly masked, and the goal is to predict these masked positions taking the corrupted sequence as input. Specifically, for our experimental design, we use the pre-trained BERT model named BETO [25] because it is trained on a big Spanish corpus.

The following sections describe the models and parameters used for each system submitted to the SpRadIE challenge.

3.3.1. BERT Multi-class Entity

The first approach is to use BERT to detect all possible entities found in a text. For this purpose, we have taken into account all types of entities in each sentence of the dataset. In this case, BERT is trained to assign an entity type to each token. Figure 1 shows an example of the sentence "*Ecografía Inguinal: Aumento de partes blandas en región inguinal*" (Inguinal ultrasound: soft tissue enlargement in the inguinal region) where each token can have up to ten different entity types or the label non-entity (O), i.e., this approach can be considered as a multi-class classification for each token.

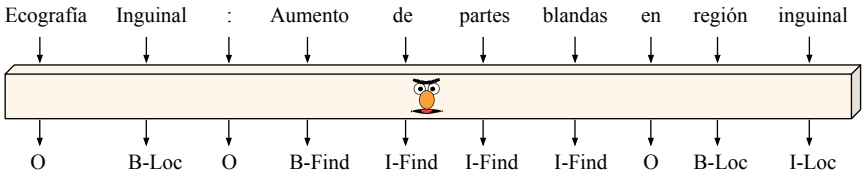

Figure 1: BERT architecture for the multi-class classification of entities approach.

3.3.2. BERT Binary Class Entity

The second approach used in the SpRadIE challenge is to perform a binary classification for each entity, instead of multi-class classification. To carry out this experimentation, we have developed ten BERT models (one for each type of entity). Figures 2 and 3 show examples using a BERT system for each entity independently. On the one hand, Figure 2 shows the same sentence as the previous approach training the dataset with the Finding entity. On the other hand, Figure 3 shows an example of the methodology followed using only the Location entity.

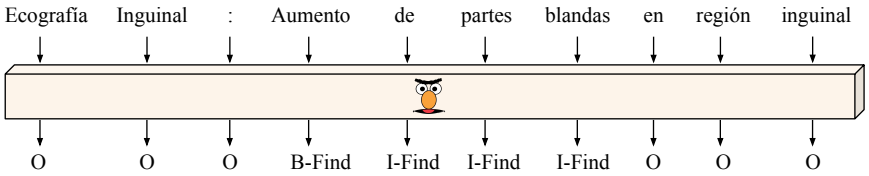

Figure 2: BERT architecture used to train *Finding* entities.

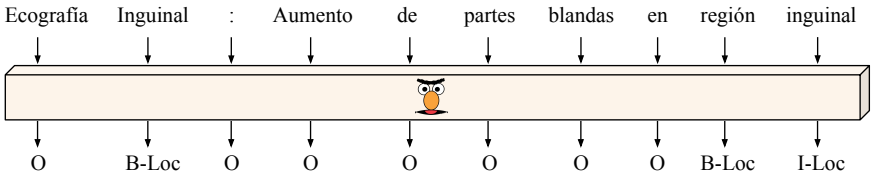

Figure 3: BERT architecture used to train *Location* entities.

3.3.3. BERT Pipeline

Following the two experiments proposed above using the development corpus, we observed that some entities performed better with the multi-class method and others with the binary method (using the development dataset). For this reason, on the one hand, we use the output of the multi-class approach for the following entities: *conditional temporal*, *finding*, *location*, *type of measure*, and *uncertainty*. On the other hand, the outputs of the *abbreviation*, *anatomical entity*, *degree*, *measure*, and *negation*, we use the output obtained from the binary approach.

Briefly, as shown in Figure 4, the BERT pipeline consists of combining the outputs of the previously proposed systems according to the type of entity in order to obtain the results.

For all our experiments, we fine-tune our models using the following hyperparameters: the BETO model used is "*bert-base-spanish-wwm-cased*" according to the Huggingface library [26] used, the maximum sequence is set to 150 and we use 5 epoch with a batch size of 32.

Finally, all experiments (training and evaluation) were performed on a node equipped with two Intel Xeon Silver 4208 CPU at 2.10 GHz, 192 GB RAM, as main processors, and six GPUs NVIDIA GeForce RTX 2080Ti (with 11 GB each).

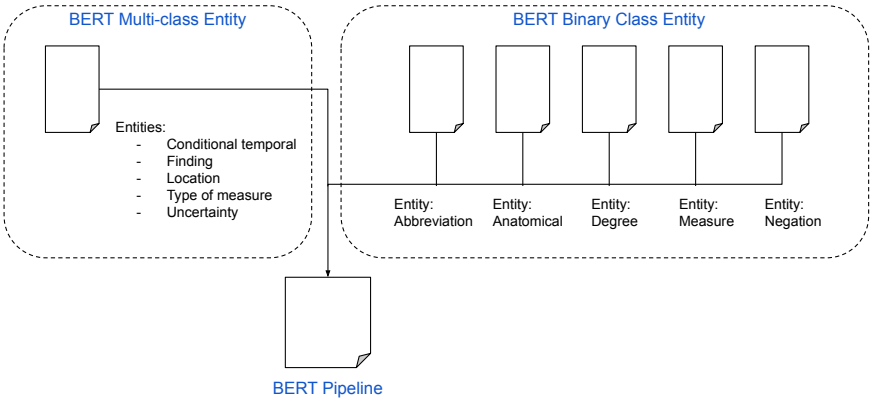

Figure 4: BERT pipeline architecture.

4. Results

The metrics defined by the SpRadIE challenge to evaluate the submitted experiments are those commonly used for some NLP tasks such as NER or text classification, namely precision (P), recall (R), and F1-score (F1) considering exact and lenient match. Table 2 shows the results obtained by the SINAI team for each run submitted.

Table 2
Systems test results achieved by the SINAI group in SpaRadIE Task 5.

	Lenient			**Exact**		
	P(%)	**R(%)**	**F1(%)**	**P(%)**	**R(%)**	**F1(%)**
BERT Multi-class Entity	83.58	45.13	58.61	76.53	41.33	53.67
BERT Binary Class Entity	**86.07**	**64.43**	**73.70**	**79.37**	**59.42**	**67.96**
BERT Pipeline	83.94	61.19	70.78	77.95	56.82	65.73
Mean participant	75.90	66.31	68.97	68.66	59.32	62.41

On the one hand, in the first approximation and taking into account both lenient and exact matching, we obtain high precision although we drop in the recall measurement. Specifically, we achieved 76.53% precision, 41.33% recall, and 53.67% F1-score by performing exact matching. On the other hand, using *BERT Binary Class Entity* approach, we improved in all measurements compared to the results of *BERT Multi-class Entity*, which means that by performing binary classification, the system can detect entities more accurately. Following this methodology, the system does not have to choose between ten different entity types, but between 0 and 1 for each entity annotated in the dataset. With this methodology, the system obtains a 14% F1 improvement reaching 67.96% (above the average of the participants). It is important to highlight that although we improved in all measures using the *BERT Multi-class Entity* method, we especially increased the recall. Finally, the results of the last approach (*BERT Pipeline*) do not obtain improvements in evaluation compared to the binary method.

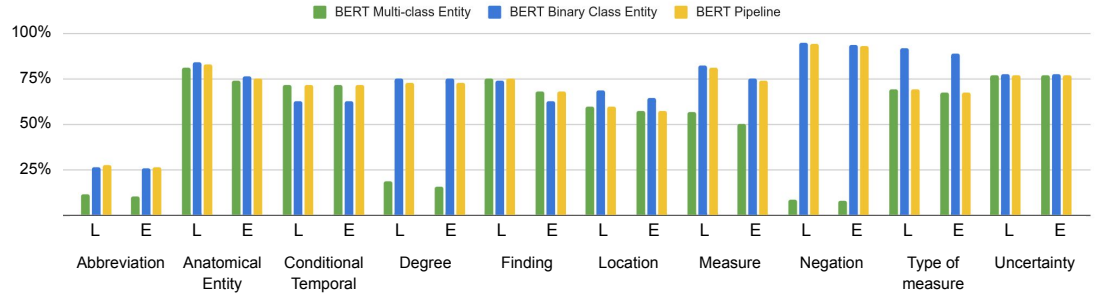

Figure 5: SINAI group results for each type of entity by using the F1 metric. L: Lenient. E: Exact.

To perform a specific evaluation for each entity, Figure 5 shows the obtained F1-score value taking into account lenient (L) and exact (E) matching for each submitted system. As we can see, there is a significant difference between the systems *BERT Multi-class Entity* and *BERT Binary Class Entity* in entities such as *degree*, *measure*, *negation* and *type of measure*. These entities have been detected more accurately using the binary entity classification system. We should highlight the recognition of negated entities because we achieved an 85% F1 improvement and the entity degree that we also improved by 59% using the *BERT Multi-class Entity* system.

5. Conclusion and Future Work

This paper presents the participation of the SINAI research group in the SpRadIE challenge at CLEF 2021. This challenge aims to extract relevant information related to the radiological domain in Spanish. Specifically, the collection is composed of reports manually annotated by specialists with ten different types of entities such as findings, anatomical entities, and abbreviations, among others.

Our proposal follows a pre-trained model-based approach using the Transformer architecture for the NER task on Spanish health documents. The proposed methods uses the BERT model trained on a large Spanish corpus called BETO. First, we performed a pre-processing step using the annotated datasets provided by the organization, which were previously tokenized and labeled using the BIO scheme. Subsequently, we propose three evaluation methodologies: using BERT to label entities as a multi-class classification system, using BERT to extract entities in a binary approach and a method that combines the outputs of the two previous ones.

Using BERT's binary system, the results obtained were better than the average of the challenge participants, achieving 67.96% F1-score, 79.37% precision, and 59.42% recall using the exact matching evaluation and 73.70% F1 with the lenient evaluation. Moreover, we found that this binary architecture for entity extraction that we propose provides more information of each entity individually for the learning phase of the model achieving better results than using the multi-class model of entity detection.

For future work, we plan to perform an in-depth error analysis once the gold annotation of the test is released. Moreover, we will study the performance of using linguistic features such

as Part-Of-Speech tags as an input in the BERT model, as well as the use of ontologies related to the radiological domain such as RadLex.

Acknowledgments

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] C. Friedman, S. B. Johnson, Natural language and text processing in biomedicine, in: Biomedical Informatics, Springer, 2006, pp. 312–343.

[2] V. Cotik, L. A. Alemany, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. D. Francesca, A. Dellanzo, M. F. Urquiza, Overview of CLEF eHealth Task 1 - SpRadIE: A challenge on information extraction from Spanish Radiology Reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2021.

[3] H. Suominen, L. Goeuriot, L. Kelly, L. Alonso Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, R. Upadhyay, J. Vivaldi, M. Viviani, C. Xu, Overview of the CLEF eHealth Evaluation Lab 2021, in: CLEF 2021 - 11th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2021.

[4] E. Pons, L. M. Braun, M. M. Hunink, J. A. Kors, Natural language processing in radiology: a systematic review, Radiology 279 (2016) 329–343.

[5] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Medical image analysis 66 (2020) 101797.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[7] G. L. Andriole, E. D. Crawford, R. L. Grubb III, S. S. Buys, D. Chia, T. R. Church, M. N. Fouad, C. Isaacs, P. A. Kvale, D. J. Reding, et al., Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up, Journal of the National Cancer Institute 104 (2012) 125–132.

[8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 590–597.

[9] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, arXiv preprint arXiv:1901.07042 (2019).

[10] A. G. 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, 2019, pp. 1–10.

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

[12] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, et al., Overview of the CLEF eHealth evaluation lab 2019, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 322–339.

[13] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), volume TBA, CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain, 2019, p. TBA. URL: TBA.

[14] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, CHEMDNER: The drugs and chemical names extraction challenge, Journal of cheminformatics 7 (2015) S1.

[15] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, Journal of the American Medical Informatics Association 20 (2013) 806–813. URL: https://doi.org/10.1136/amiajnl-2013-001628. doi:10.1136/amiajnl-2013-001628.

[16] P. L. Úbeda, M. C. D. Galiano, M. T. Martín-Valdivia, L. A. U. Lopez, Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media, in: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, 2019, pp. 102–106.

[17] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[18] P. López-Úbedaa, M. Dıaz-Galianoa, M. Martın-Valdiviaa, L. A. Ureña-Lópeza, Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings, Proceedings of IberLEF (2020).

[19] P. López-Úbedaa, J. M. Perea-Ortegab, M. C. Díaz-Galianoa, M. T. Martín-Valdiviaa, L. A. Ureña-Lópeza, SINAI at eHealth-KD Challenge 2020: Combining Word Embeddings for Named Entity Recognition in Spanish Medical Records (2020).

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805.

[21] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of Entities and Relations in Spanish Radiology Reports, in: RANLP, 2017, pp. 177–184.

[22] L. Padró, E. Stanilovsky, Freeling 3.0: Towards wider multilinguality, in: LREC2012, 2012.

[23] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), 2009, pp. 147–155.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[25] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: to appear in PML4DC at ICLR 2020, 2020.

[26] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.