Extracting information from radiology reports by Natural Language Processing and Deep Learning

Miguel Ángel Martín-Caro García-Largo[1], Isabel Segura-Bedmar[1]

[1]Universidad Carlos III de Madrid, Madrid 28911, Spain

Abstract

Radiology reports are texts that include the description and interpretation of ultrasound images. The automatic processing of these texts, if well performed, can help healthcare professionals and the diagnosis. This work is part of the Information Extraction from Spanish radiology reports task (SpRadIE) of CLEF eHealth 2021. Regarding the case of study of this work, it is remarkable the correct detection of unusual findings because they can affect the patient's health. Furthermore, it can help health professionals and researchers to be focused on problematic cases. Three different models have been proposed to face that task, evaluating and comparing their performance. Conditional Random Field (CRF), Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) and Bidirectional Encoders Representation from Transformers(BERT). With BiLSTM-CRF, two different approaches have been used: the use of randomized initialized vectors and the use of a pre-trained word embedding in Spanish. Both will appear more detailed in the paper. The task is complex and some of the reasons are: reports are written in Spanish, the extensive number of types of entities and the ambiguity in the language used by the doctors in the reports. The best results have been obtained by the CRF model, which has obtained a Lenient F1 score of 77% for a dataset that contained most of the words in the training dataset and a 67% of Lenient score for the dataset with words that are not present in the dataset used to train the model.

Keywords

Natural Language Processing, Named-entity recognition, Radiology reports, Information Extraction, CRF, BERT, BiLSTM-CRF

1. Introduction

1.1. Motivation

Artificial Intelligence (AI) aims to create algorithms to make computers think smarter but a computer is still being a machine of computing with a memory. Natural Language Processing (NLP), a field of a AI, is a set of techniques to automatically 'understand' and create human language. Nonetheless, it has to be clear that today computers can not think as humans do. NLP it is a multidisciplinary field that involves many areas including Linguistics, Computer Science and Psychology among others. NLP is an ambitious area of research and it includes many tasks like: Information Retrieval, Information Extraction, Question-Answering, summarization, Machine Translation, dialogue systems, among others .[1]

Information Extraction consists of structuring information from the texts. IE comprises thre

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

100363781@alumnos.uc3m.es (M. Á. M. García-Largo); isegura@inf.uc3m.es (I. Segura-Bedmar)
https://hulat.inf.uc3m.es/ (I. Segura-Bedmar)

main subtasks: Named Entity Recognition (NER), relation extraction and co-reference resolution. NER consists of identifying the type of word or expressions that appear within a text. It is a crucial task for many NLP applications such as relation extraction, information retrieval, machine translation, text simplification, text summarization.

SpRadIE task[2] (included in the CLEF eHealth 2021)[3] targets the detection of seven different entities as well as hedge cues.[2] This paper describes the models that we have presented for the task. The goal of this work is to explore different NER approaches for detecting findings from radiology reports written in Spanish. Reports about patients in health services are essential as they collect the necessary information to understand the diseases or abnormalities of a patient. The correct detection of unusual findings from radiology reports could help health professionals and researchers to easily interpret these reports and accelerate the diagnostic of possible conditions.[4]

Most existing NLP for extracting information are based on supervised machine learning, which require annotated corpora for training and evaluating their models. Information extraction techniques allows to effectively transform unstructured text to structured data, which could bring benefits to the the radiology area. Nevertheless, it remains a complex task so that collaboration between radiologists, data scientists and engineers will be a key point to achieve optimal results.[5] Unfortunately, there is a lack of these resources for other languages than English. There have already been some research initiatives to promote research on information extraction from clinical texts written in Spanish, such as the Cantemist Track for Cancer Text Mining in Spanish.[6] or the ehealth knowledge discovery challenge[7].

SpRadIE 2021 aims to promote NLP research applied to the extraction of information from radiology reports. Several challenges must be addressed in the task: the complexity of the entities, the language in which reports are written (Spanish) and the fact that some reports do not contain the same words than reports in the training dataset.

1.2. Objectives

The general objective of this project appears next:

To build a Named Entity Recognition (NER) system able to recognise entities in the domain of radiology (ultrasound images) clinical reports written in Spanish.

That objective is decomposed into four more specific objectives:

- Review the main approaches used to extract information from clinical text, particularly, from radiology reports.
- Explore different NER approaches for detecting the entities of the dataset provided by the organisers of SpRadIE 2021.
- Evaluate and compare the performance of the proposed models to determine the weak and strong points for each of them.
- Participate in the SpRadiE 2021 competition and submit the results of our three best models.

2. Related work

Recently, the use of NLP in biomedical texts has increased [8]. It presents challenging problems to deal with like discontinuous entities, misspellings, abbreviations, the low presence of some entity types and the existence of ambiguity.

In the last years, many competitions of NLP applied to biomedical texts have tried to look for good approaches to overcome these problems. These competitions are described in in [9]. The task of NER has been used to solve different problems in the biomedical domain, like the anonymization of clinical reports[10] or the extraction of information in various subfields such as pharmacovigilance[11, 12] or oncology[13]. This chapter is focused in the carried out works performed in texts about clinical reports or more specifically in radiology reports.

Clinic report anonymization is necessary to ensure the protection of data from the patients and during last years that task has been developed using NER systems. In [14], a NER system was developed for anonymizing private data in radiology reports written in Spanish. The used approach is focused on neural networks with different architectures (LSTM-CRF, BiLSTM-CRF, Convolutional-BiLSTM-CRF). The alternative was to use pattern matching but that implies some negative aspects. For example: as it only considers exact patterns the context is not taken into account and thus, plenty of information would be lost. Furthermore, mispelled words could have a bad influence on results. For example: An example which is very explanatory is the Spanish surname 'Cabeza' (head). Approaches based on pattern matching would not distinguish between the surname and this part of the body. The best results were obtained by an BiLSTM-CRF: Bi-LSTM initialized with character embeddings and embeddings pre-trained with Glove vectors and a CRF in its last layer.[14] This model obtained an F1 score of 92.63% on the test data using their own radiology report dataset. It consists of 7,848 brain radiology reports taken from the Medical Imaging Databank of the Valencia Region (BIMCV). To ensure that retrieved reports contained personal information, only texts with more than two name tags were selected. Only a third part was annotated. This model was also tested on the MEDDOCAN challenge dataset, achieving an F1 score of 81%.

Another recent research [15] has tried to use NER for radiological reports in Japanese using deep learning. The goal was to recognise mentions of the following entity types: observations, clinical findings, anatomical location modifiers, certainty modifiers, change modifiers, characteristic modifiers and size modifiers. Different neural networks were taken into account, among them: BiLSTM-CRF, BERT and BERT-CRF. BiLSTM was the network that achieved the highest F1 score with an F1 score of 95.36% using an in-house dataset. The in-house dataset was built using 118,078 reports from the radiology information system at Osaka University Hospital. The model was evaluated on external reports, providing an F1 score of 94.62% using. These external reports were 77 chest CT reports from the Osaka International Cancer Institute (OICI) [15]. For the optimization of the networks hyper-parameters, Optuna[16] which is a recent software for optimization was used.

The main goal of the paper [17] was to develop a NER system to extract information from clinical reports written in Chinese. The authors explored a deep learning approach and compared it with a more traditional algorithm such as Condition Random Field (CRF). Two corpus were used: one for training and evaluating the models and another for creating a word embedding used later in the deep learning developed model. The dataset used for the training and evaluation of the models consists of 400 reports from the EHR database of Peking Union Medical Collegue Hospital. The dataset contains four types of entities: medications, procedures, problems and lab tests. For the creation of the embedding, 36,828 reports were selected from the same institute of China. The deep learning network consisted of a convolutional layer, followed by a non-linear layer and several linear layers. Two different approaches were taken: one where the input vectors were randomly initializated and another where the input vectors are taken from the word embedding model trained on the second corpus. The results were: 91.9% F1-score for the CRF, 90.7% for the CNN with random initialization and 92.8% for the CNN with word embeddings.

BERT is a general purpose language representation based on deep learning which takes into account the context of a word by looking to the left and the right of the word at the same time. It only needs a fine-tuning to adapt to a new particular dataset. BERT has widely used in NER taks with very successful results. We now describe some of the most important works on the clinical domain.

Kim and Lee proposed a slightly better modified version of BERT. This version consists of modifying the labelling strategy. The proposed labeling strategy for BERT was carried out due to some peculiarities of the Korean language so it is not guarantee that in other languages the use of that labelling system would improve the results. Furthermore, results are not much different from the default BERT labels. The corpus was created by extracting texts from the biggest questions and answers platform in Korean (Kin Naver) and a total of 536 answers were retrieved. These texts were annotated with the following three entity types:diseases, symptoms and body parts. The authors also used another external dataset, the Exo Brain Korean dataset. It consists of 10,000 sentences with five different entity types: person, location, time, date and organization. Tokens were represented by using the IOB standard format. BERT and BiLSTM-CRF methods were compared. Macro-averages for the created dataset were 83% of F1 score for BERT and 79% for BiLSTM-CRF. For the Exo brain dataset, the macro-averages of F1 score were 94% for BERT and 89% for BiLSTM-CRF.

The scope of the NER tasks comprises different approaches such as rule-based and deep learning. In Gorinski et al. addressed the NER task by appliying thre different approaches: a rule-based method, a transfer learning approach and a deep learning model. Transfer learning consists of training a model for a specific task and then using that model with some fine hyper-parameter tuning to another problem. The authors used a Scottish radiology reports, the TayExt dataset. This dataset contains brain images reports. Some of the types of entities which appeared were: ischaemic strokes, hemorrhagic strokes, strokes, tumours, etc. The best results were obtained by the hand-written rule-based approach with a 93% of total F1 score. The second place was for the transfer learning approach using the SemEHR tool [20], with an F1 score of 89. The third place was for the deep learning approach (LSTM-CRF) with an 80% of

F1-score. The rule-based approach achieved the best results, however this kind of approaches needs from domain experts and takes longer to developed the rules. Moreover, these rules cannot be used to recognize entity types of other subfields. On the other hand, one of the main advantages of deep learning is that it does not requires domain experts.

SpRadIE proposes a challenging NER task since its dataset contains up to ten named entity types. As it has been explained in the beginning of this section, it presents a challenging scenario for developing a NER system. It is due to discontinuities, embedded entities, ambiguities, abbreviations and imbalanced datasets among others. Moreover, radiology reports are written in Spanish that although it is a language with presence in the domain of this task, it has been less studied in NLP than English[21].

3. Methods

3.1. Data preprocessing

Texts were represented as vectors to feed the models with data. First, texts are split into sentences and each of those sentences is divided into tokens(words). Then, those tokens were assigned with a part of speech tag (PoS). We use Spacy, a python library for NLP. Moreover, each token is represented following the BIOES-V token annotation format, where 'B' is used for the beginning token of an entity, 'I' is used for tokens in the middle of entities, 'E' is a tag used for indicating a token is the end of an entity, 'S' is used when the entity is composed only by one token, 'V' is used when the token is part of a nested entity, and 'O' is used for those tokens that do not belong to an entity. Moreover, the position of the token in the sentence is also considered. These BIOES-V labels are obtained by using the information from the annotation files, which were provided in BRAT format.

In the case of the first approach for BiLSTM which uses random initialization of vectors, a vocabulary is created using the words of the texts. It is done by assigning to each of the words present in the texts a random vector. The counter part of random initialization, is that no relation between words is captured. Furthermore, the sentences for this model must have the same length. In the case of longer sentences than the fixed one, truncation is applied to remove the rest of tokens. In the case of sentences shorter than the fixed length, padding is applied.

The preprocessing of the BiLSTM network with the second approach (using a clinical Spanish embedding) is similar to the BiLSTM approach 1. The difference is that the vocabulary is created by assigning to each word a vector from the word embedding model. This model can capture the semantic relationships of the words in a corpus. As in the case of the BiLSTM approach 1, the length of the sentences are fixed but no much values were checked.

BERT is a model that does not need the PoS tag, so it is not necessary to take it into account. The way data are preprocessed to prepare them for the model is by tokenizing them with the BERT tokenizer. That tokenizer divides the tokens of the input texts into smaller ones. In that way, the length of wordpieces is larger than the list of tokens. BERT assigns a vector to each token, similar to what was done for the BiLSTM networks and all the sentences are set to a fixed length.

3.2. CRF

Our first approach is based on CRF [22]. This model was state-of-art for NER [23] before the appearance of deep learning. It fits perfectly for the NER task and it is a good baseline model to compare its results with more new approaches. Technically, it is a discriminative probabilistic non directed graph model based on the maximum entropy and Hidden Markov models. [24] The difference between discriminative and generative models is that generative models learn the joint probability distribution $p(x, y)$, while discriminative models learn the conditional probability distribution $p(y|x)$ (where x is the sequence of observations and y the sequence of output labels). In other words, generative models model the distribution of each class and discriminative models the frontiers between them. [25]

Before exposing the formula of the model, some simplified notation is presented to facilitate the labour of representation of the formula:

$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i) \tag{1}$$

The probability of a label sequence y given an observation sequence x, can be written as:

$$p(y|x, \lambda) = \frac{1}{Z(x)} exp(\sum_{j} \lambda_j F_j(y, x)) \tag{2}$$

where each $f_j(y_{i-1}, y_i, x, i)$ is a state function or a transition function and Z(x) is a normalization factor. [26] λ parameters are estimated by the model by using an optimization algorithm. [24]

The features set used for the CRF classifier takes into account the following features: word, the previous, the next one tokens, lemmas of the word, as well as their PoS tags. As it was mentioned previously, these elements were obtained by using Spacy.

3.3. BiLSTM-CRF

Our second approach is a deep learning architecture, which is composed of two parts: a BiLSTM network and a CRF, as the last layer (BiLSTM-CRF) [27]. BiLSTM is a kind of recurrent neural network (RNN) which are able to consider past observations. Moreover, BiLSTM takes into account the tokens around the current token to predict its label. Therefore, this is a effective approach for NER. [28]

Moreover, among recurrent neural networks, BiLSTM are a good candidate because they deal with both the gradient vanishing problem and the explosion problem. The vanishing gradient problem appears after multiplying the gradient many times by a number lower than one. On the other hand, the explosion gradient problem appears after multiplying many times the gradient by a number greater than one. This kind of networks deal with those problems. To deal with these problems RNN propose the definition of cell. A cell is an operation with two inputs (the sequence values and a given past state) and two outputs (the current state and the sequence of values computed for the current state).

The solution is to set cells that link the past state with the current output. [29] The characteristics which allow BiLSTM networks to consider past observations and control the importance that is given to them, resides in the interior of the cells. Figure **??** shows the structure that a cell usually has. Forget, input and output gates play a key role in that task of extracting the insights from data that are of interest. Oblivion gate controls when some part of the information is forgotten, input gate controls when new information should enter the cell and the output gate controls when the information stored in the cell is used in the result of that cell.[29]

It has been said that as BiLSTM considers the past, it is appropriate for the NER task but not only because of that. As it is bidirectional, it considers more information making it even more appropriate. Bidirectionality was introduced in [30] and it can be considered as if two independent agents were extracting information from the same problem (each one in one direction) to finish concatenating the information and passing it to another layer. [29] The output of the BiLSTM are two vectors (one per direction) and they are received by the CRF. Typically the output function in many neural network tasks is the softmax function but as CRF considers context, it is preferred.[23]

Two approaches have been used with the BiLSTM model: the use of randomized initializated vectors and the use of vectors from a pre-trained word embedding model which was trained on a collection of texts in Spanish[31]. In the first case, each sentence is divided into tokens and a random vector is assigned to each token. The problem with that is that words that do not appear in the training set and the relationships with the words in the training set, could be not correctly identified. Word embeddings can capture syntactic and semantic relationships between words. [32] The texts[31] that were used to create the word embedding model contain information from many different biomedical areas and although radiology reports are not the main content, they are present in the corpus. The method which was applied to create the embedding was Word2Vec with an Skip-gram architecture. [31] The SkipGram architecture receives a vector representing a word and returns another vector representing the probabilities of other words in the vocabulary to be near to the given word.

3.4. BERT

BERT is a bidirectional self-attention (which is the ability to understand the context) transformer composed of two unsupervised steps: the masked language model (MLM) and the next sentence prediction (NSP). The first is essential to achieve bidirectional training avoiding the model to 'see' the target and the second, to understand sentence relationships. [33]

BERT was developed in 2019. The power of that model was the wide range of applications to which it can be used. The employed model in this work uses a pre-trained BERT and then performs fine-tuning over it to find the optimal combination of parameters for the task of NER.

The pre-trained BERT model uses the following parameters: L=12, H=768, A=12 and Total Parameters=110M, where L denotes the number of layers (or transformed blocks), H is the hidden size and A is the number of self-attention heads. This fixed establishment of the parameters in the pre-training BERT, allows focusing on the parameters of fine-tuning (maximum length, batch size, learning rate and the number of epochs). [33] In the work of this master thesis the chosen values for those parameters have been: maximum sequence length=75, batch size=32, learning rate=3e-5 and number of epochs=3. They were chosen based on the values which were used in [23].

4. Experiments

4.1. Dataset description

From the website of the competition the dataset is presented: 'The data consists of ultrasonography reports provided by a pediatric hospital in Argentina. Reports are unstructured, have an abundance of orthographical grammatical errors and have been anonymized in order to remove patient IDs, and names and the enrollment numbers of the physicians. Reports were annotated by clinical experts and then revised by linguists. Annotation guidelines and training were provided for both rounds of annotations. Automatic classifiers will be expected to perform well in those cases where human annotators have a strong agreement, and worse in cases that are difficult for human annotators to identify consistently. Annotations are provided in brat format'.[2]

The proposed task is ambitious and not at all easy. It can be confirmed by checking the numerous entities which have been considered in the clinical reports and the strict annotation process which has had to be taken in order to get a set of annotations of quality. All of that supports the idea of being a complex task. The entities which appear in the dataset are: findings, anatomical entities, location, measure, type of measure, texture, negation, uncertainty terms, abbreviations and temporal terms. The meaning for each of them is described in [34]. An interesting fact is the huge quantity of negations that this dataset contains. It would be useful in order to evaluate the performance of methods which pretend to detect negations.

4.2. Global results and confusion matrices for the best model in development and validation datasets

In this subsection the results are divided for each of the data partitions described in the previous section: development and held-out datasets. The vocabulary of the entities in the development set, is the same with which the algorithms have been trained. In contrast, the held-out dataset, contains vocabulary which has not been present during the training phase of the algorithms. In order to not extend a lot this paper, only the confusion matrices for the best model (CRF) have been included.

The obtained global results for the development dataset have been:
It can be seen that in the case of the development dataset, the best method has been CRF with a significant difference. In the second position, can be found BERT and the BILSTM2 (BERT has obtained slightly better results but the difference is not significant). The method with the worst results has been the BILSTM1.

Although the dataset is not exactly the same with which the methods have been trained, it can be seen that CRF is able of identifying the entitites in a correct manner in most of the occasions.

The obtained global results for the held-out dataset have been:
The results are worse than in the case of the development set, which was something expected. Furthermore, the performance of the models follows the same order: the best model is CRF followed by BERT, BILSTM2 and BILSTM1.

The confusion matrix of Table 4 despite being less accurate than the matrix in Table 2, contains more interesting insights.

In the description of the task [2], it is stated that there are challenging situations like the fact that there exists regular polysemy between anatomical entities and locations. The irregular use of abbreviations also increases the difficulty of the task. That a priori description of problems related to the task, has been confirmed after looking at the results of Table 4. The anatomical entities, locations and findings are easily confused between them. Apart from this, the irregular abbreviations also cause many confusions with the anatomical entities. In the case of the held-out dataset confusion matrix, it is also remarkable the fact that the number of entities not classified as any type of the present entities, is high (see last column in Table 4). In both of the cases, by looking at both confusion matrices, it is confirmed the high presence of negations in the reports. This reinforces the idea of using the datasets as examples to evaluate entity recognition models build to detect negations.

4.3. Results for test dataset in the competition

The global obtained Lenient F1 score in the competition has been of 75.64% whereas the winner result has been 85.51%. Dividing the result into the different entities, the following results have

Table 1
Global results for development dataset

	CRF	BILSTM1	BILSTM2	BERT
Lenient F1	0.77658	0.46019	0.68015	0.68155
Exact F1	0.69430	0.38515	0.56708	0.58847

Table 2
CRF confusion matrix for entities in development dataset

		PREDICTED										
		Ana	Fin	Unc	Neg	Loc	Con	Typ	Mea	Abb	Deg	Oth
REAL	Ana	479	1	0	0	5	0	0	0	9	0	4
	Fin	2	483	0	1	3	0	0	0	0	1	10
	Unc	0	0	18	0	0	0	0	0	0	0	0
	Neg	0	0	0	227	0	0	0	0	0	0	0
	Loc	4	1	0	0	208	0	0	0	0	0	3
	Con	0	0	0	0	0	3	0	0	0	0	0
	Typ	0	0	0	0	0	0	102	0	0	0	0
	Mea	0	7	0	0	0	0	0	216	0	0	1
	Abb	1	3	0	0	0	0	5	1	193	0	0
	Deg	0	0	0	0	0	0	0	0	0	13	0
	Oth	10	2	0	5	2	0	0	5	0	0	2213

Table 3
Global results for held-out dataset

	CRF	BILSTM1	BILSTM2	BERT
Lenient F1	0.67134	0.20137	0.62514	0.63036
Exact F1	0.60776	0.13658	0.56126	0.56308

Table 4
CRF confusion matrix for entities in held-out dataset

		PREDICTED										
		Ana	Fin	Unc	Neg	Loc	Con	Typ	Mea	Abb	Deg	Oth
REAL	Ana	507	17	0	0	25	0	1	0	8	0	44
	Fin	46	451	1	0	37	0	7	13	5	0	268
	Unc	0	0	6	0	0	0	0	0	0	0	1
	Neg	0	9	1	234	1	0	0	0	0	0	26
	Loc	19	16	0	0	107	0	2	0	0	0	57
	Con	0	0	0	0	0	0	0	0	0	0	2
	Typ	0	0	0	0	5	0	106	0	0	0	0
	Mea	1	23	0	0	1	0	0	208	0	0	17
	Abb	12	2	0	0	2	0	1	1	196	0	16
	Deg	0	1	0	0	0	0	0	0	0	12	8
	Oth	34	46	1	0	19	0	1	11	1	0	2295

been obtained: 83.72% for abbreviations, 70.07% for anatomical entities, 61.54% for conditional temporal, 53.44% for degree, 69.18% for findings, 68.35% for locations, 62.50% for measures, 93.78% for negations, 86.28% for types of measures and 73.26% for uncertainties.

5. Conclusions and future work

Best results were obtained by CRF despite the implementation of other models like BiLSTM-CRF and BERT, which are state-of-the-art in the named entity recognition area. This means that not always an approach based on neural networks is the best. It is true that they can get good results like the shown in Section 2 of this paper but if the selection of hyperparameters is not good, it is going to be difficult to obtain the best results.

A great improvement (18% of F1 score) has been seen in the BiLSTM-CRF network when using a Spanish biomedical word embedding versus when not using it. This suggests, that as future work, maybe the use of BETO [35] (Spanish BERT) could improve the performance of BERT until the point of surpassing the CRF model.

This task has used Spanish data from Argentina. It could also be interesting to compare the behaviour of the models in radiological texts in Spanish but from different areas like Spain and South America.

6. Acknowledgments

This work was supported by the NLP4RARE-CM-UC3M, which was developed under the Interdisciplinary Projects Program for Young Researchers at University Carlos III of Madrid. The work was also supported by the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

References

[1] E. D. Liddy, Natural language processing (2001).

[2] V. Cotik, L. Alonso Alemany, R. Roller, F. Luque, H. Vivaldi, D. Filippo, A. Ayach, F. Carranza, L. De Francesca, A. Dellanzo, M. Fernández 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, 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 - 12th conference and labs of the evaluation forum, lecture notes in computer science (lncs), Springer (2021).

[4] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Ureña-López, M.-T. Martín-Valdivia, A. Luna, Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0957417420304711.

[5] T. Martín-Noguerol, F. Paulano-Godino, R. López-Ortega, J. Górriz, R. Riascos, A. Luna, Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team, 2020. URL: https://n9.cl/gdxlc.

[6] 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, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020. URL: https://www.bsc.es/research-and-development/publications/named-entity-recognition-concept-normalization-and-clinical.

[7] A. Piad-Morffis, Y. Gutiérrez, H. Canizares-Diaz, S. Estevez-Velarde, R. Muñoz, A. Montoyo, Y. Almeida-Cruz, et al., Overview of the ehealth knowledge discovery challenge at iberlef 2020, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, CEUR, 2020.

[8] V. Gopalakrishnan, K. Jha, W. Jin, A. Zhang, A survey on literature based discovery approaches in biomedical domain, Journal of biomedical informatics 93 (2019) 103141.

[9] C.-C. Huang, Z. Lu, Community challenges in biomedical text mining over 10 years: success, failure and the future, Briefings in bioinformatics 17 (2016) 132–144.

[10] C. Colón-Ruiz, I. Segura-Bedmar, Protected health information recognition bybilstm-crf (2019).

[11] I. Segura-Bedmar, P. Martinez, C. de Pablo-Sánchez, Using a shallow linguistic kernel for drug–drug interaction extraction, Journal of biomedical informatics 44 (2011) 789–804.

[12] I. Segura-Bedmar, S. de la Peña González, P. Martínez, Extracting drug indications and adverse drug reactions from spanish health social media, in: Proceedings of BioNLP 2014, 2014, pp. 98–106.

[13] G. de Vargas Romero, I. Segura-Bedmar, Exploring deep learning for named entity recognition of tumor morphology mentions, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[14] I. Perez-Diez, R. Perez-Moraga, A. Lopez-Cerdan, J.-M. Salinas-Serrano, M. de la Iglesia-Vaya, De-identifying spanish medical texts-named entity recognition applied to radiology reports, Journal of Biomedical Semantics 12 (2021) 1–13.

[15] K. Sugimoto, T. Takeda, J.-H. Oh, S. Wada, S. Konishi, A. Yamahata, S. Manabe, N. Tomiyama, T. Matsunaga, K. Nakanishi, et al., Extracting clinical terms from radiology reports with deep learning, Journal of Biomedical Informatics 116 (2021) 103729.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[17] Y. Wu, M. Jiang, J. Lei, H. Xu, Named entity recognition in chinese clinical text using deep neural network, Studies in health technology and informatics 216 (2015) 624.

[18] Y.-M. Kim, T.-H. Lee, Korean clinical entity recognition from diagnosis text using bert, BMC Medical Informatics and Decision Making 20 (2020) 1–9.

[19] P. J. Gorinski, H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, B. Alex, Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches, arXiv preprint arXiv:1903.03985 (2019).

[20] H. Wu, G. Toti, K. I. Morley, Z. M. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale, et al., Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research, Journal of the American Medical Informatics Association 25 (2018) 530–537.

[21] L. Campos, V. Pedro, F. Couto, Impact of translation on named-entity recognition in radiology texts, Database 2017 (2017).

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

[23] G. D. V. Romero, Development of a named entity recognition system to automatically assign tumor morphology entity mentions to health-related documents in Spanish, Master Thesis dissertation, Universidad Carlos III de Madrid, 2019-2020.

[24] S. Song, N. Zhang, H. Huang, Named entity recognition based on conditional random fields, Cluster Computing 22 (2019) 5195–5206.

[25] P. M. Joshi, Generative vs discriminative models, 2018. URL: shorturl.at/yEH26.

[26] H. M. Wallach, Conditional random fields: An introduction, Technical Reports (CIS) (2004) 22.

[27] Z. Zhai, D. Q. Nguyen, K. Verspoor, Comparing cnn and lstm character-level embeddings in bilstm-crf models for chemical and disease named entity recognition, arXiv preprint arXiv:1808.08450 (2018).

[28] D. S. Sachan, P. Xie, M. Sachan, E. P. Xing, Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition, in: Machine learning for healthcare conference, PMLR, 2018, pp. 383–402.

[29] L. Bagén, B. R. Toni, C. R. Anna, et al., Deep learning: principios y fundamentos, Deep learning (2019) 1–260.

[30] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (1997) 2673–2681.

[31] A. Gutiérrez-Fandiño, J. Armengol-Estapé, C. P. Carrino, O. D. Gibert, A. Gonzalez-Agirre, M. Villegas, Spanish biomedical and clinical language embeddings, 2021. arXiv:2102.12843.

[32] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[34] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of entities and relations in Spanish radiology reports, in: Proceedings of the International Conference Recent

Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 177–184. URL: https://doi.org/10.26615/978-954-452-049-6_025. doi:10.26615/978-954-452-049-6_025.

[35] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.