

Information Extraction from Spanish Radiology Reports using multilingual BERT

Oswaldo Solarte-Pabón^{1,2}, Orlando Montenegro², Alberto Blazquez-Herranz¹, Hadi Saputro¹, Alejandro Rodriguez-González¹ and Ernestina Menasalvas¹

¹Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Spain

²Universidad del Valle, Cali, Colombia

Abstract

This paper describes our team's participation in Task 1 of the Conference and Labs of the Evaluation Forum (CLEF eHealth 2021). The Task 1 challenge targets Named Entity Recognition (NER) from radiology reports written in Spanish. Our approach addresses this challenge as a sequence labeling task and is based on multilingual BERT with a classification layer on top. Three BERT-based models were trained to support overlapping entities extraction: the first model predicts the first specific label annotated in the corpus; the second predicts the second label for tokens that have two different annotations; and the third is used for tokens annotated with a third label in the corpus. Our approach obtained 78.47% and 73.27% for a Lenient and the exact F1 score, respectively.

Keywords

Information Extraction, Named Entity Recognition (NER), Multilingual BERT, Radiology Reports

1. Introduction

Radiology reports are one of the most important sources of clinical imaging information. They document critical information about the patient's health and the radiologist's interpretation of medical findings [1]. Extracted information from Radiology reports can be used to support clinical research, quality improvement, and evidence-based medicine [2]. However, the information in radiology reports is presented in free text format, which makes the task of structuring the data particularly challenging [3]. Extracting this information manually is not a viable task as it is costly and time-consuming [4].

Several studies have been proposed to extract information from radiology reports [5, 6, 7]. Most of these proposals use machine learning methods, mainly based on the Conditional Random Fields (CRF) algorithm [8], in which entity extraction is considered as a sequence labeling task. Recently, deep learning approaches have been shown to improve performance at processing natural language texts [9, 10, 11]. Previous studies have shown significant progress in extracting information from clinical reports, but most efforts have focused only on the English language. However, the Task 1 challenge of CLEF 2021 targets Named Entity Recognition (NER) and

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

✉ oswaldo.solartep@alumnos.upm.es (O. Solarte-Pabón); orlando.montenegro@correounivalle.edu.co (O. Montenegro); alberto.bherranz@upm.es (A. Blazquez-Herranz); H.saputro@alumnos.upm.es (H. Saputro); alejandro.rg@upm.es (A. Rodriguez-González); ernestina.menasalvas@upm.es (E. Menasalvas)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Classification from radiology reports written in Spanish, and specifically ultrasounds reports. [12, 13].

In this paper, we describe the approach we have followed to deal with Task 1 of the Conference and Labs of the Evaluation Forum (CLEF 2021). Our approach takes advantage of the transfer learning technique, which aims to learn new and specific tasks by transferring knowledge from pre-trained models [14, 15, 16]. Specifically, we use contextualized pre-trained embeddings from multilingual BERT [17] with a classification layer to perform NER from radiology reports written in Spanish. Three classifiers were trained to predict overlapping entities: i) one to predict the first specific label in the corpus; ii) another classifier to predict the second label for those tokens that have two different annotations; iii) and the third is used for those tokens that have a third annotated label.

The rest of the paper has been organized as follows: Section 2 reviews the most recent works on extracting information from radiology reports. Section 3 describes the components of the proposed approach. Section 4 presents the results of the experiments, and Section 5 presents the main conclusions and outlook for future work.

2. Related works

Computational analysis of radiology reports has recently gained greater attention. Most of the research focuses on identifying specific medical conditions contained in a given report, and they generally deal with the task of classification. Another generic area of research is developing information extraction models, which try to extract specific information such as medical recommendations from radiology reports [18].

A system for identifying named entities from the radiology reports was proposed by Hassanpour [1]. The objective of this model was to identify the specific named entities from the radiology reports based on their information extraction model. They used a CRF-based model with several auxiliary features, including POS tags and Radlex Lexicons [19] to identify entities of five different classes: Anatomy, Anatomy Modifier, Observation, Modifier, and Uncertainty. In [20], an approach was proposed to extract useful information in abdominopelvic radiology reports. This approach combines Glove word embeddings with a Recurrent Neural Network [21] to extract named entities from radiology reports. This approach shows the feasibility of neural nets for extracting information from a small corpus of 120 abdominopelvic cross-sectional imaging reports.

Most of the proposals mentioned above have been focused on the English language. According to [22], information extraction in the medical domain also represents its own challenges in other languages. In the case of the Spanish language, [23] proposed a manually annotated corpus using text data from radiology reports. This corpus aims to support NER in clinical text written in Spanish.

3. Data and Methods

The CLEF eHealth 2021 Task 1 (Multilingual Information Extraction) focuses on NER from radiology reports written in Spanish, and specifically from ultrasounds reports. Our approach addresses this challenge as a sequence labeling task, in which each token in a sentence is classified using the BIO tagging format. In this format, each token is labeled with B (at the beginning of the entity), I (inside the entity), or O (Outside the entity).

3.1. Data

The CLEF eHealth 2021 Task 1 organizers provided an annotated corpus in BRAT format [24]. This corpus consists of ultrasonography reports provided by a pediatric hospital in Argentina. Reports were annotated by clinical experts and then revised by linguists. The corpus contains three sets: i) Training set (175 reports), ii) Development set (92 reports), and iii) Test set (207 reports). Figure 1 shows an example of annotations provided in the corpus for Task 1.

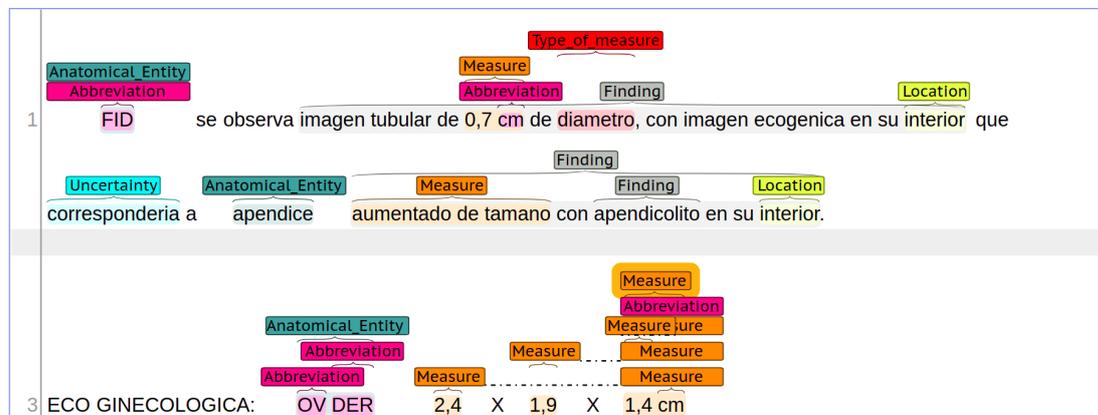


Figure 1: An example of annotated sentences in the corpus

The corpus provided was annotated using ten different labels: "Abbreviation, Anatomical_Entity, ConditionalTemporal, Degree, Finding, Location, Measure, Negation, Type_of_measure, and Uncertainty". In this corpus, annotations can be represented by a word or a sequence of words (not necessarily continuous). Moreover, a same token can be annotated with different labels, this implies that overlapping entities can be presented. From Figure 1, it can be seen that extracting information from radiology reports written in Spanish has several challenges:

- **Overlapping entities:** some tokens may have more than one annotated label. For instance, the token "FID" is annotated with the label "Anatomical_Entity" and, with the label "Abbreviation". Moreover, overlapping entities can also mean that a token can belong to more than one entity type. For instance, in the first sentence (see Figure 1), the token "cm" is embedded into the labels "Finding" and "Measure". It causes that this token contains three annotated labels: "Finding, Measure, and Abbreviation".

- **Lengthy entities:** these entities are composed of many tokens, which is common in the corpus. One can see that annotations with the label *"Finding"* in Figure 1 are composed of many tokens.
- **Discontinuities:** these occur when an entity contains annotated tokens that are not continuous in the text document. This is the case of annotations with the label *"Measure"* in Figure 1.
- **Polysemy:** this is also a common linguistic phenomenon in the Task 1 corpus provided. It frequently occurs between labels such as *"Location"* and *"Anatomical Entity"*.

3.2. Methods

Our approach consists of three steps: Text pre-processing, Entity Extraction, and Post-Processing (See Figure 2). In the first step, files in BRAT format are converted into BIO format. In the second, NER models are trained to perform entity extraction, and in the third, predicted entities from NER models are assembled in BRAT format again as output. Details of each step are described as follows.

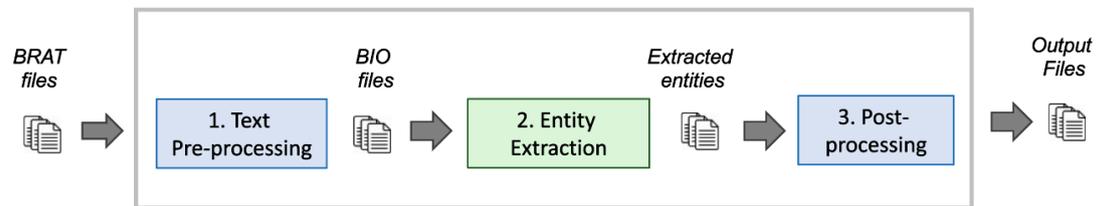


Figure 2: The proposed approach

3.2.1. Text Pre-processing

In this step, each file in the corpus is divided into a set of sentences, and each sentence is tokenized. Next, the sentences are converted into a BIO file format. As the corpus contains annotated tokens of up to three different labels; the generated BIO files also have three columns that describe the annotations for each token. Table 1 shows an example of the generated BIO file for a specific text sentence taken from the corpus. The column "Label 1" represents the first annotated label in the corpus for a specific token. The column "Label 2" is used to represent a second annotated label in the corpus. Those tokens that do not have a second annotation are assigned the "O" (Outside) tag by default. Lastly, the column "Label 3" represents a third annotated label in the corpus. For instance, the token *"cm"* contains three different annotated labels in the corpus; therefore, this token has three different BIO labels (see Table 1). The columns "Label 2" and "Label 3" are used to represent overlapped entities for the tokens that have more than one annotated label.

Table 1
A BIO file example

Token	Label 1	Label 2	Label 3
FID	B-Abbreviation	B-Anatomical_Entity	O
se	O	O	O
observa	O	O	O
imagen	B-Finding	O	O
tabular	I-Finding	O	O
de	I-Finding	O	O
0.7	I-Finding	B-Measure	O
cm	I-Finding	I-Measure	B-Abbreviation
de	I-Finding	O	O
diametro	I-Finding	B-Type_of_measure	O
,	I-Finding	O	O
con	I-Finding	O	O
imagen	I-Finding	O	O
ecogenica	I-Finding	O	O
en	I-Finding	O	O
su	I-Finding	O	O
interior	I-Finding	B-LOCATION	O
.	O	O	O

3.2.2. Entity Extraction

The purpose of this step is to perform NER from radiology reports using a transformers-based model[25] with a classification layer on top of it. Specifically, we use the *bert-base-multilingual-cased* model, which has been pre-trained on data in 104 languages. The input for the BERT model is a sequence of tokenized and encoded tokens of a sentence extracted previously from the BIO files created.

Figure 3 shows the entity extraction process in which three BERT-based models (M_1, M_2, M_3) were trained to support overlapping entities extraction. Each NER model is trained using a different input data column (I_1, I_2, I_3). The first model is trained with annotations obtained from the "Label 1" column in the BIO file (see Table 1), the second is trained with annotations from the "Label 2" column, and the third is trained using annotations from the "Label 3" column. Each NER model M_n operates independently and predicts a different set of entities (E_1, E_2, E_3), depending on the annotations with which it was trained.

As can be seen in Table 1, there are three data columns containing different annotated labels for a given token, and we trained each NER model using a different input data column. Therefore, each NER model M_n predicts different entities. On the one hand, the first model is trained to predict all ten labels annotated in the dataset. On the other hand, the second and third models are used to predict overlapped entities learned from the tokens that have a second and a third annotated label respectively. Although these models work independently, the fact that they predict different labels helps to solve the task of predicting overlapped labels.

The BERT-based models are specialized for NER, with a fully connected layer on top of

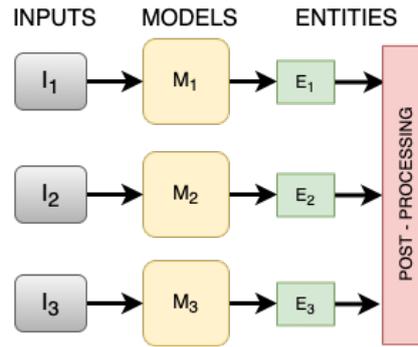


Figure 3: Entity extraction process

the hidden states of each token. These models are fine-tuned on the corpus provided. The training and development sets of the corpus were mixed, and then we trained the models using a cross-validation strategy with a value of $k = 10$. The fine-tuning was performed with a sequence length of 256 tokens, a batch size of 32, and the number of epochs is 10. The original BERT tokenizer model drove the tokenization process.

3.2.3. Post-processing

The goal in this step is to take as input the predicted entities (E_1, E_2, E_3) from the BERT-based NER models (Figure 3), and to create files in BRAT format with their annotations and corresponding offsets.

Algorithm 1 shows the procedure used to create annotations in BRAT format. This procedure receives a sentence text, the text file to which the sentence belongs, and the entities extracted from the NER models. The NER models work at the sentence level. However, to create a file in BRAT format, it is important to have the original text file in order to obtain the annotations' offsets.

Algorithm 1 Procedure to create BRAT files

Require: sentence, text_file, NER-Model-Results

```

1: for each model in NER-Model-Results do
2:   for each entity in model do
3:      $tokens \leftarrow getTokens(entity)$ 
4:      $label \leftarrow getLabel(entity)$ 
5:     if label  $\neq$  'O' then
6:        $offsets \leftarrow getOffsets(text\_file, sentence, tokens, label)$ 
7:       addAnnFile(tokens, label, offsets, text_file)
8:     end if
9:   end for
10: end for
  
```

4. Results

In the official test phase, 207 files with unannotated radiology reports were available for evaluating the models. Table 2 shows the official performance of our proposal for Lenient and the exact F1 score. We obtained 78.47% and 73.27% for a Lenient and the exact F1 score, respectively. The proposed approach shows promising results and suggests the feasibility of multilingual BERT to perform NER from radiology reports written in Spanish.

Table 2
Results obtained for Task 1

	Lenient (%)			Exact (%)		
	P	R	F1	P	R	F1
Abbreviation	85.39	81.20	83.25	82.05	78.03	79.99
Anatomical-Entity	84.01	83.63	83.82	80.33	79.97	80.15
ConditionalTemporal	71.43	62.50	66.67	71.43	62.50	66.67
Degree	74.93	53.00	62.09	74.14	52.44	61.43
Finding	64.76	73.69	68.94	55.54	63.20	59.12
Location	67.92	64.24	66.03	66.34	59.91	61.58
Measure	73.53	76.16	74.83	64.08	66.38	65.21
Negation	89.32	88.53	88.92	87.97	87.20	87.58
Type-of-measure	85.30	81.89	83.56	81.80	78.54	80.14
Uncertainty	89.48	46.58	61.26	84.21	43.84	57.66
Total	78.62	78.32	78.47	73.27	72.99	73.13

Taking into account the performance for each label, Table 2 shows that those with the best performance were: *Abbreviation, Anatomical_Entity, Negation, and Type_of_measure*. These labels obtained a performance above 80%. We found that the competitive performance of these labels can be due to:

- These labels do not commonly contain overlapping entities which increases the performance of entity extraction in comparison with those labels containing overlapping entities.
- These labels contain a high number of annotations in the corpus in comparison with other labels. For instance, the label "Anatomical_Entity" has more than 1700 annotations. This fact suggests that having a corpus with more annotated data is useful to improve NER performance using deep learning methods such as multilingual BERT.
- In the case of the negation label, note that the proposed approach shows a very good performance. The "negation" label has a high number of annotations in comparison with other labels. We found more than 750 annotations with this label.

In contrast, the lowest-performing labels were *Finding, Uncertainty, ConditionalTemporal, Degree, Location and Uncertainty*, with a performance below 70%. We identified several reasons for this behavior:

- In the case of the label "Finding", it is used to represent lengthy entities that contain a long number of tokens. It is clear that the extraction of lengthy entities affects the performance of the proposed NER models. In this case, the Recall measure is higher than Precision measure (see Table 2). This fact can be explained by two reasons: The first reason is because this label frequently contains annotations with a high number of tokens, which makes extracting all annotated tokens more difficult than in those entities that only have one annotated token. The second is because the "Finding" label contains annotations with discontinuous tokens which also reduces the Precision measure.
- The high variability of words annotated with the label "Uncertainty" together with the low number of annotations, when compared with the "Negation" label, can be the cause of the low performance in predicting the "Uncertainty" label.
- The low performance of the "ConditionalTemporal" and "Degree" labels is caused by the low number of annotations in the corpus compared to other labels. We found 75 annotations with the "Degree" label and only 25 with the "ConditionalTemporal" label. This number of annotations are much lower in comparison with those labels which obtained better results.

4.1. Error Analysis

We identified three more common sources of error which could affect the performance of our approach. These error sources are related to: i) annotations with discontinuous tokens; ii) the high number of annotations with the label "O" (Outside) in the BIO file; and iii) the high variability of annotated tokens in those labels with a low number of annotations.

- The first source of error identified is when an annotation contains discontinuous tokens. In these cases, predicting the correct annotated label for each token, and obtaining the correct offsets, can be difficult. For instance, in the text "*VIA BILIAR intra y extrahepatica*", there are two annotations with the label "*Anatomical_Entity*" in the provided dataset: "VIA BILIAR extrahepatica" and "VIA BILIAR intra hepatica". However, these two annotations contain discontinuous tokens and predicting labels for each token fails because the automatic text tokenizer being used produces different tokens to those annotated in the input text. In this case, our approach takes the string "extrahepatica" as one token, but in the annotated corpus, it is used as two sub-words: "extra" and "hepatica". A similar problem occurs when processing the text "*Quiste en epididimo izq de: 3,6 x 1,4 x 1,9mm*", where there are three discontinuous annotations with the label "*Measure*": "3,6 mm", "1,4 mm" and "1,9 mm". In the above examples, the main issue was calculating the exact token offsets which match the original text in the dataset.
- The second source of error is the high number of annotated tokens with the label "O" (Outside) in the BIO file. This problem mainly affects the third NER model (M_3), which has been trained with annotations from the column "Label 3" in the BIO file (see Table 1). This column is used to represent a third annotation for those tokens which have overlapped entities. In this case, the number of annotations with the label "O" is much higher than other labels. This fact causes a reduction in the performance of the third NER

model in comparison with the first and the second model.

- The third source of error is caused by the low number of annotations on certain labels combined by the high variability of annotated tokens in these labels. This occurs in labels such as "Uncertainty", "Degree", and "ConditionalTemporal". In these cases, the Precision measure is higher than the Recall measure. This fact can be explained by the high variability of tokens used to annotate these labels, which increases the difficulty of predicting all possible tokens in the extraction process. For instance, comparing the behavior of the "Uncertainty" label with the "Negation" label, one can see that "Negation" has less variability in the annotated corpus. In the case of the "Negation" label, tokens like *"no"*, *"sin"*, *"no se observa"* are frequently used. On the other hand, the number of words expressing "Uncertainty" are more dispersed which causes a reduction in the Recall measure. The fact that the "Uncertainty" label has less annotations and more variability in annotated tokens than the "Negation" label, could be the reason why the "Uncertainty" label does not perform as well as the "Negation" label.

Finally, in the proposed approach, the biggest challenge was being able to predict labels that have overlapping entities. For this purpose, a three-parallel NER architecture was chosen as explained in Section 3.2.2, which showed promising results. However, before reaching this solution we tested two other configurations. The first solution was to combine entities with one annotation with all overlapping entities into a single column in the BIO file. This single column represented all the annotations in the corpus. However, the main disadvantage of this solution was the large number of labels that need to be predicted. The second solution was to create a different annotation column for each annotated label. In this case, the BIO file would need ten columns with annotations, one column for each label in the corpus. The main drawback of this solution is that for each column of the BIO file, the number of annotations with the label "O" (Outside), was greater than the label we wanted to predict. The performance of this solution was much lower than our current proposed approach, so we rejected that solution. After carrying out those experiments, the approach that obtained the best performance was to create a combination of three NER models, as we described in section 3.2.2. This proposed solution allows extracting labels that have up to three overlapping entities, just as the annotated entities are found in the corpus provided.

5. Conclusions

The proposed approach to perform NER from radiology reports has shown promising results, suggesting that the multilingual BERT is feasible to extract information from clinical resources written in Spanish. The main advantage of the proposed approach is that it can automatically extract features from text data, avoiding the time-consuming feature engineering process.

Extracting information from radiology reports written in Spanish is a complex task. It poses several linguistic challenges, such as overlapping entities, lengthy entities, and discontinuities. We deal with these challenges by training three BERT-based models. Although our approach has shown promising results, some challenges still have to be addressed, such as extracting

lengthy entities, improving the performance to extract labels with a low number of annotations, and identifying uncertainty in radiology reports.

References

- [1] S. Hassanpour, C. P. Langlotz, Information extraction from multi-institutional radiology reports, *Artificial Intelligence in Medicine* 66 (2016) 29–39. URL: <http://dx.doi.org/10.1016/j.artmed.2015.09.007>. doi:10.1016/j.artmed.2015.09.007.
- [2] W. W. Yim, M. Yetisgen, W. P. Harris, W. K. Sharon, Natural Language Processing in Oncology Review, *JAMA Oncology* 2 (2016) 797–804. doi:10.1001/jamaoncol.2016.0213.
- [3] O. Solarte-Pabon, M. Torrente, A. Rodriguez-Gonzalez, M. Provencio, E. Menasalvas, J. M. Tunas, Lung cancer diagnosis extraction from clinical notes written in spanish, *Proceedings - IEEE Symposium on Computer-Based Medical Systems 2020-July (2020)* 492–497. doi:10.1109/CBMS49503.2020.00099.
- [4] O. Solarte Pabón, M. Torrente, M. Provencio, A. Rodríguez-Gonzalez, E. Menasalvas, Integrating Speculation Detection and Deep Learning to Extract Lung Cancer Diagnosis from Clinical Notes, *Applied Sciences* 11 (2021). URL: <https://www.mdpi.com/2076-3417/11/2/865>. doi:10.3390/app11020865.
- [5] S. Hassanpour, C. P. Langlotz, Information extraction from multi-institutional radiology reports, *Artificial Intelligence in Medicine* 66 (2016) 29–39. URL: <http://dx.doi.org/10.1016/j.artmed.2015.09.007>. doi:10.1016/j.artmed.2015.09.007.
- [6] N. Nandhakumar, E. Sherkat, E. E. Milios, H. Gu, M. Butler, Clinically significant information extraction from radiology reports, *DocEng 2017 - Proceedings of the 2017 ACM Symposium on Document Engineering (2017)* 153–162. doi:10.1145/3103010.3103023.
- [7] J. M. Steinkamp, C. Chambers, D. Lalevic, H. M. Zafar, T. S. Cook, Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning, *Journal of Digital Imaging* 32 (2019) 554–564. doi:10.1007/s10278-019-00234-y.
- [8] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, Morgan Kaufmann Publishers Inc., Williamstown, MA, USA, 2001, pp. 282–289. URL: <http://dl.acm.org/citation.cfm?id=645530.655813>.
- [9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: <https://www.aclweb.org/anthology/N16-1030>. doi:10.18653/v1/N16-1030.
- [10] L. Wang, L. Luo, Y. Wang, J. Wampfler, P. Yang, H. Liu, Natural language processing for populating lung cancer clinical research data, *BMC Medical Informatics and Decision Making* 19 (2019) 1–10. URL: <http://dx.doi.org/10.1186/s12911-019-0931-8>. doi:10.1186/s12911-019-0931-8.
- [11] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional

- transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (2019) 4171–4186. arXiv:1810.04805.
- [12] H. Suominen, L. Goeuriot, L. Kelly, L. A. 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: 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.
- [13] V. Cotik, L. A. Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. D. Francesca, A. Dellanzo, M. F. Urquiza, Overview of the clef ehealth evaluation lab 2021, in: 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, Springer, 2021.
- [14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A Comprehensive Survey on Transfer Learning, Proceedings of the IEEE 109 (2021) 43–76. doi:10.1109/JPROC.2020.3004555. arXiv:1911.02685.
- [15] C. Sun, Z. Yang, Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 100–104. URL: <https://www.aclweb.org/anthology/D19-5715>. doi:10.18653/v1/D19-5715.
- [16] Y. Peng, S. Yan, Z. Lu, Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets, arXiv (2019) 58–65. doi:10.18653/v1/w19-5006. arXiv:1906.05474.
- [17] C.-L. Liu, T.-Y. Hsu, Y.-S. Chuang, H.-y. Lee, What makes multilingual BERT multilingual?, ArXiv (2020). URL: <http://arxiv.org/abs/2010.10938>. arXiv:2010.10938.
- [18] N. Nandhakumar, E. Sherkat, E. E. Milios, H. Gu, M. Butler, Clinically significant information extraction from radiology reports, DocEng 2017 - Proceedings of the 2017 ACM Symposium on Document Engineering (2017) 153–162. doi:10.1145/3103010.3103023.
- [19] C. P. Langlotz, RadLex: A new method for indexing online educational materials, Radiographics 26 (2006) 1595–1597. doi:10.1148/rg.266065168.
- [20] J. M. Steinkamp, C. Chambers, D. Lalevic, H. M. Zafar, T. S. Cook, Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning, Journal of Digital Imaging 32 (2019) 554–564. doi:10.1007/s10278-019-00234-y.
- [21] S. Hochreiter, J. Schmidhuber, Lstm can solve hard long time lag problems, in: Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96, MIT Press, Cambridge, MA, USA, 1996, p. 473–479.
- [22] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: Opportunities and challenges, Journal of Biomedical Semantics 9 (2018) 1–13. doi:10.1186/s13326-018-0179-8.
- [23] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of entities and relations in Spanish radiology reports, International Conference Recent Advances in Natural Language Processing, RANLP 2017-Septe (2017) 177–184. doi:10.26615/978-954-452-049-6-025.

- [24] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.
- [25] H. Adel, H. Schütze, Exploring different dimensions of attention for uncertainty detection, 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference 1 (2017) 22–34. doi:10.18653/v1/e17-1003. arXiv:1612.06549.