A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports

Víctor Suárez-Paniagua[1,3], Hang Dong[1,3] and Arlene Casey[2]

[1]*Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.*
[2]*Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.*
[3]*Health Data Research UK, London, United Kingdom.*

Abstract

The present work describes the proposed methods by the EdIE-KnowLab team in Information Extraction Task of CLEF eHealth 2021, SpRadIE Task 1. This task focuses on detecting and classifying relevant mentions in ultrasonography reports. The architecture developed is an ensemble of multiple BERT (multi-BERT) systems, one per each entity type, together with a generated dictionary and available off-the-shelf tools, Google Healthcare Natural Language API and GATECloud's Measurement Expression Annotator system, applied to the documents translated into English with word alignment from the neural machine translation tool, Microsoft Translator API. Our best system configuration (multi-BERT with a dictionary) achieves 85.51% and 80.04% F1 for Lenient and Exact metrics, respectively. Thus, the system ranked first out of 17 submissions from 7 teams that participated in this shared task. Our system also achieved the best Recall merging the previous predictions to the results given by English-translated texts and cross-lingual word alignment (83.87% Lenient match and 78.71% Exact match). The overall results demonstrate the potential of pre-trained language models and cross-lingual word alignment for limited corpus and low-resource NER in the clinical domain.

Keywords

Named Entity Recognition, Radiology Reports, Deep Learning, BERT, Machine Translation

1. Introduction

Medical imaging reports are interpretations of diagnostic images written by radiologists. Whilst radiology reports contain a relatively restricted vocabulary compared to other electronic health records they are still unstructured, and this makes it difficult to extract meaningful data. However, being able to effectively extract information from these narratives has the potential to quickly and accurately identify abnormalities supporting clinical decision in a timely manner. The application of Natural Language Processing (NLP) to radiology reports is a growing area such as shown in a recent systematic review [1].

The SpRadIE Task 1 [2] was the first challenge to deal with Named Entity Recognition (NER) in the domain of Spanish radiology reports. Concretely, the target is to detect and classify relevant mentions in the ultrasonography reports produced by physicians during their clinical practice. These documents cover different domains such as heart and liver related reports.

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

vsuarez@ed.ac.uk (V. Suárez-Paniagua); hang.dong@ed.ac.uk (H. Dong); arlene.casey@ed.ac.uk (A. Casey)

CEUR Workshop Proceedings (CEUR-WS.org)

Currently, large pre-trained language models with layers of multi-head self-attention architectures [3], specifically Bidirectional Encoder Representations from Transformers (BERT) [4], outperform other machine learning systems for the task of NER [5, 6] particularly in the biomedical domain [7, 8]. BERT based models have been successful applied to NLP tasks in radiology, such as Smit et al. [9] who label findings in chest radiology reports, Wood et al.[10] who explore document level labels at a coarse and finer grained level from free-text, and Schrmepf et al. [11] who use BERT with a per-label attention mechanism. However, BERT models do not always outperform more traditional methods [12]. In this paper our main approach is to use BETO [13], BERT pre-trained models for Spanish NLP tasks. There are existing works that perform Spanish NER for radiology reports [14, 15]. However, the use of a BERT based model is still largely unexplored in Spanish radiology report named entity recognition.

This work describes the participation of the team EdIE-KnowLab in the CLEF 2021 eHealth Task 1 [16] that involves the recognition of named entities in Spanish radiology reports. The proposed method, which ranked first in the task, is a hybrid system that combines multiple Spanish BERT classifiers (BETO) that were fine-tuned independently for each entity type, and the use of a dictionary extracted with the annotations from the training set. The cloud-based machine translation service, Microsoft Translator API, was used to translate the documents to English. Once the documents were translated, available off-the-shelf tools Google Healthcare Natural Language API (GHNL) and GATECloud's Measurement Expression Annotator (MEA) system predicted the entities in the documents which were then traced back into Spanish using the translation alignment.

2. Dataset

The dataset for the CLEF 2021 eHealth Task 1 contains 513 anonymized radiology reports from a major pediatric hospital in Buenos Aires. Clinical experts and linguists annotated 17,000 mentions using Brat Standoff format [17] following an annotation guideline. The organizers provided this dataset split in three different sets annotated: 174 reports for training the models, 92 documents to validate the systems and 207 reports without annotations to test the predictions submitted by the participants. In addition, the development set was divided into 47 documents with the same vocabulary as the training set (*same-sample*) and 45 documents containing words that are not in the training set (*held-out*). The entities are divided into seven entity types: *Anatomical Entity*, *Abbreviation*, *Finding*, *Location*, *Measure*, *Type of Measure*, and *Degree*, and three hedge cues: *Negation*, *Uncertainty*, and *Conditional Temporal*.

Table 1 presents the number of annotations for each entity type in the three different sets. It can be observed that this task presents a highly unbalanced problem with the greatest represented class (1,292) differing in two orders of magnitude with respect to the lowest (11). A more detailed description of the dataset and its annotations can be found in [18].

2.1. Data preprocessing

The annotated reports show multiple linguistic challenges, such as orthographic and grammatical errors, very long entities, discontinuous and embedded entities, subordination and coordination, and systematic polysemy. Thus, our team carried out a cleaning step of the data using simple

Table 1
Number of instances for each class in the training and development sets from the highest represented to the lowest.

Entity type	training set (174 reports)	*same-sample* set (47 reports)	*held-out* set (45 reports)
Anatomical Entity	1292	365	476
Abbreviation	877	251	269
Finding	795	230	450
Measure	579	148	211
Location	517	121	156
Negation	484	121	156
Type of measure	346	80	79
Uncertainty	52	15	6
Degree	41	13	22
Conditional Temporal	11	3	1

rules and regular expressions in order to solve some of these challenges and prepare the documents for the classifier.

There are some tokens whose spans were not completely annotated, like the mention "venas femorale" in the sentence "Ambas arterias y venas femorales permeables". In these cases, we utilized regular expressions to redefine the token offsets until the span covers the mention. Concretely, the regular expressions "*[a-zA-Z]+$*" and "*^[a-zA-Z]+*" were used from the given span to complete the mention backward and forward, respectively. For the discontinuous mentions, we followed a naïve approach that takes the minimum and the maximum offsets as the span of these split mentions.. The embedded entities were de-overlaped obtaining all the possible paths without overlapping walking recursively through a graph representation of the sentence, where the nodes are the entities and the edges are the links to the other mentions which overlap. We automatically solve the problem of the annotations with multiple types using one classifier for each class and then merging all their outputs into a prediction file.

Documents were transformed into lower case and some special characters, like the escape sequence "*\n*" for newline, were replaced by a white space and the sentences were tokenized using the Spanish transformer pipeline of spaCy [19]. Finally, the Brat annotations were tagged with the BIOES schema, an extension to the BIO-encoding [20], where the B tag indicates the beginning token of an entity, the I tag indicates the inside token of an entity, the E tag indicates the ending token of an entity, the S tag indicates a single entity token, and the O tag represents other tokens that do not belong to any entity.

3. Methods

This section presents the different approaches used for the SpRadIE 2021 Shared Task. Figure 1 shows the whole system with the four proposed methods integrated together merging their predictions.

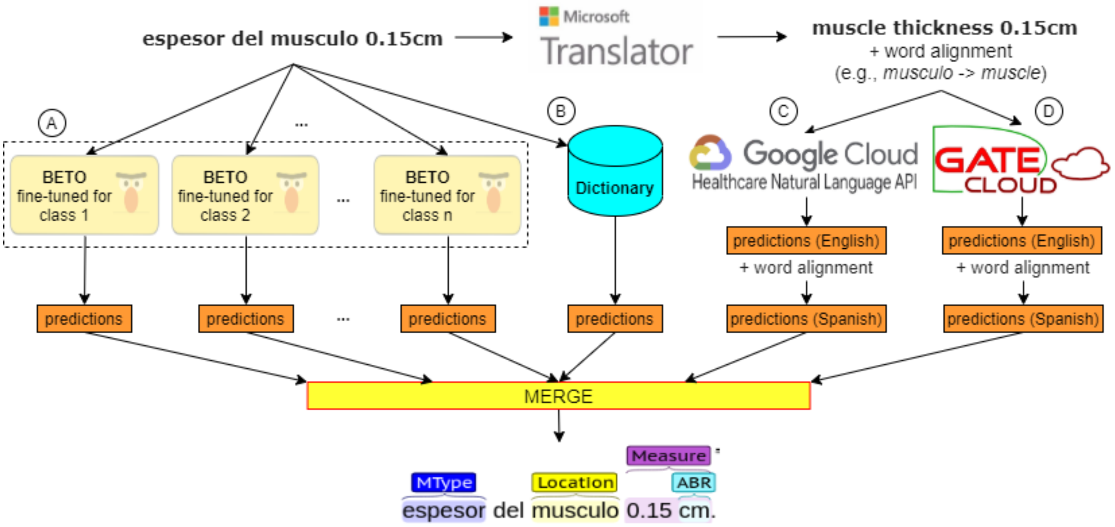

Figure 1: The proposed NER system that recognizes and classifies the mentions in a given Spanish radiology report. The whole hybrid system contains four different methods (from A to D): A. multi-BERT classifier; B. dictionary-based approach; C-D: English translation and cross-lingual word alignment (all obtained using Microsoft Translator API) with Google Healthcare Natural Language API (C) and GateCloud's Measurement Expression Annotator (D). All predictions from the different methods are integrated into a final prediction. Submissions are a cumulative combination of the four methods A-D. The system A+B (multi-BERT with dictionary) obtained the best F1 and A+B+C (further using word alignment with English NER) obtained the best Recall in the final evaluation in SpRadIE.

3.1. Multi-BERT classifier

Once the Spanish radiology reports were preprocessed, all the annotations in the datasets were divided by classes. Thus, a single BETO classifier [13] was fine-tuned for each class using the corresponding named entities in the training set independently. Then, each model was validated with the development set to get the best performance for each entity type. Figure 1A shows the merging of individual predictions in order to generate the multi-BERT final annotation.

3.2. Dictionary based

We observed that at a word-level, annotations for entities are highly repetitive and thus, entity vocabulary is similar across the reports. For this reason, we created a dictionary using the vocabulary from the named entities mapped to their corresponding classes in the training set. We use the generated vocabulary to find exact string matches in the reports and classify them with the labels given by the dictionary (see Figure 1B).

3.3. Cross-lingual word alignment with English NER tools

Since the most advanced NER tools are usually tailored for texts in English, we used machine translation with cross-lingual word alignment to leverage results from these tools. We used

Table 2
Matching between the medical knowledge categories in Google Healthcare Natural Language (GHNL) API and the entity types in SpRadIE

GHNL API category	SpRadIE entity type
ANATOMICAL_STRUCTURE	*Anatomical_Entity*
BODY_MEASUREMENT	*Type_of_measure*
BM_UNIT	*Abbreviation*
BM_VALUE	*Measure*
LAB_VALUE	*Measure*
LABORATORY_DATA	*Measure*
MED_STRENGTH	*Measure*
MED_UNIT	*Abbreviation*
PROBLEM	*Finding*
SEVERITY	*Degree*

Microsoft Translator API[1], a neural machine translation (NMT) tool, to translate reports from source language (Spanish) to target language (English). The key reason to use Microsoft Translator API is that it allows for the output of word alignment between the original and the translated texts[2]. Each word alignment is represented as a pair of mention spans, where the span in the source language is aligned to the one in the target language and each span includes a start and an end index. However, one major limitation is that NMT methods can produce erroneous and unreliable word alignment as we see in the experiments, potentially because the widely used attention-based alignment [21] is not accurate enough, e.g. in the sentence in Figure 1, "muscule" in English was actually aligned to "espesor" in Spanish (rather than "musculo") despite the correct translation at the sentence level. Nevertheless, the word alignment enables leveraging of NER tools for other languages (e.g. English), which are presented below.

Google Healthcare Natural Language (GHNL) API The GHNL API[3], released in Nov 2020 [22], allows for the extraction and matching of mentions in clinical texts into medical terminologies and classify the mentions into a set of "medical knowledge categories" (See Figure 1C). We matched some of these categories into the entity types for this task (e.g. "SEVERITY" to *Degree*). The matching dictionary from the categories in GHNL API to the entity types in the SpRadIE task is presented in Table 2.

GATECloud's Measurement Expression Annotator (MEA) GATE is open source free software that performs text analysis with a multitude of applications [23]. We used MEA through the GATE cloud service[4] (See Figure 1D). This tool annotates numbers and measurement expressions in text. We map the indices of numeric measurements and their units returned by MEA to *Measure* entities.

[1] https://docs.microsoft.com/en-us/azure/cognitive-services/translator/translator-info-overview
[2] https://docs.microsoft.com/en-us/azure/cognitive-services/translator/word-alignment
[3] https://cloud.google.com/healthcare/docs/concepts/nlp
[4] https://cloud.gate.ac.uk/shopfront/displayItem/measurement-expression-annotator

After obtaining the English entities identified from the two off-the-shelf tools, we converted these back into the entities and offsets in the original Spanish texts based on the word alignment. A tolerance value (from 0 to 3) of number of characters was allowed in matching the mention spans from either the source or the target language to the indexes in an alignment word pair. We selected the best tolerance value (2 for GHNL API and 1 for MEA) according to the results in the development set.

4. Results

Each BETO classifier was trained for 8 epochs independently with the training set of each class. Early stopping criteria is applied to each model taking the best performance over both the development sets together (*same-sample* and the *held-out*). We applied the uncased model of BETO because this achieved better results than the cased model during the validation. In addition, the maximum length of a sentence was fixed to 300 and the remaining hyper-parameters to their default values for fine-tuning each BETO classifier.

The submissions were evaluated using two metrics:

- Lenient F1: is computed using the Precision and Recall of the Jaccard Index that measures the coefficient between the intersection and the union of the reference entities and the predicted entities.
- Exact F1: is calculated with the Jaccard Index scores when the reference and predicted entities have a perfect match.

In order to choose the methods to be used for each submission, we evaluated each system independently with the two development sets (*same-sample* and *held-out*). The multi-BERT approach achieved very good results in the majority of the classes, but the F1 score was 0% for *Degree* and *Conditional Temporal*, likely due to the low number of entities in the training set (submission 1). Thus, we decided to create a hybrid approach and include the predictions of the two methods (A+B), A the multi-BERT approach and B the dictionary based approach (submission 2). While the micro-F1 of GHNL API was low (26.29%) due to the inaccurate cross-lingual alignment, we observed that GHNL API performed better in the class *Degree* than BETO (30.43% vs. 0% F1) and we aggregate its predictions to the hybrid system (submission 3). Finally, the GATECloud's MEA obtained higher Precision for the class *Measure* than BETO (84.29% vs 81.78%) and we merge the prediction for this entity type to the previous methods having the complete system (submission 4). We also experimented with adding results from the GATECloud's BioYODIE[5] [24]. This tool applies a gazetter-based approach for named entity recognition and disambiguation to identify various biomedical named entities and tries to link entities to concept labels in UMLS. We mapped entities from the BioYODIE results to *Anatomical Entity*, *Location* and *Findings* but as it did not improve performance we left this out of submission 4. Doing this ablation study, we can evaluate the contribution of each method to the whole system.

Table 3 presents the final results over the test set for each submission. The best performance is obtained by submission 2 which is the hybrid approach of the multi-BERT and the dictionary

[5] https://cloud.gate.ac.uk/shopfront/displayItem/bio-yodie

based. This system ranked first in the SpRadIE Shared Task with a 85.51% and 80.26% in F1 for Lenient and Exact metrics, respectively. However, the best Precision is obtained only using the multi-BERT approach because the dictionary based approach introduces False Positives. Moreover, submission 3 achieved the best Recall results in the task, meaning that it helps to recognize some missing entities of the *Degree* type, but subsequently it marginally drops the Precision measure by introducing spurious predictions. Submission 4 slightly dropped the results in all the metrics, suggesting that aggregating the MEA predictions for the *Measure* class did not improve the BETO classification for this class.

Table 3
SpRadIE 2021 official results for the EdIE-KnowLab team submissions. Best performance for each column is marked in bold.

Submission	**Lenient**			**Exact**		
	Precision	Recall	F1	Precision	Recall	F1
EdIE-KnowLab 1	**87.81%**	82.99%	85.33%	**82.36%**	77.84%	80.04%
EdIE-KnowLab 2	87.24%	83.85%	**85.51%**	81.88%	78.70%	**80.26%**
EdIE-KnowLab 3	87.15%	**83.87%**	85.48%	81.79%	**78.71%**	80.22%
EdIE-KnowLab 4	85.67%	83.75%	84.70%	80.17%	78.37%	79.26%

Table 4
The multi-BERT hybrid system results for each entity type in the SpRadIE 2021 Shared Task.

Entity type	**Lenient**			**Exact**		
	Precision	Recall	F1	Precision	Recall	F1
Anatomical Entity	87.51%	84.35%	85.90%	80.91%	77.99%	79.43%
Abbreviation	95.93%	94.83%	95.38%	95.35%	94.25%	94.79%
Finding	72.65%	75.63%	74.11%	59.94%	62.40%	61.15%
Measure	90.15%	85.85%	87.94%	83.06%	79.09%	81.03%
Location	75.56%	62.70%	68.53%	71.83%	59.60%	65.14%
Negation	93.52%	94.97%	94.24%	92.39%	93.82%	93.10%
Type of measure	97.77%	82.80%	89.66%	94.50%	80.03%	86.67%
Uncertainty	68.44%	43.13%	52.91%	52.17%	32.88%	40.34%
Degree	53.72%	79.27%	64.04%	53.72%	79.27%	64.04%
Conditional Temporal	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4 shows the submission 2 results for each entity type, which is the model that performs the best overall. Surprisingly, this system achieves very good results in some classes with a low number of instances such as *Negation*, *Type of measure* or *Uncertainty*. This is due to the fact that each BERT classifier is forced to be trained for a specific category, so the model optimization is done independently. Submission 1, which is only the multi-BERT, obtained 0% in Lenient F1 and Exact F1 for the classes *Conditional Temporal* and *Degree* and we increased the results of the *Degree* class to 64.04% in both metrics using the dictionary. However, it could not be increased for the *Conditional Temporal* which we believe is due to it being one of the most linguistically complex entity types. The F1 scores of the submission 3 and 4 were slightly lower than the multi-BERT hybrid system in the classes that were applied, *Degree* and *Measure*, respectively. Thus, we conclude that more exploration about the translation, their alignment and the mapping of the classes for these tools is required to enhance the overall predictions.

5. Conclusions

This work describes the multi-BERT hybrid system presented by the EdIE-KnowLab for the CLEF 2021 eHealth Task 1, SpRadIE. This model ranked first using multiple BETO classifiers, one for each named entity, in Spanish radiology reports together with a dictionary extracted from the training set. The proposed NER method shows very promising results achieving a 85.51% in Lenient F1 and 80.04% in Exact F1. The main advantage of this approach is that it does not require any expert domain knowledge or external resources for classifying the mentions. The multi-classifier approach deals with the problem of class imbalance in some entities within this task and it can recognize the overlapped entities with different classes, but it is not able to predict embedded mentions with the same class.

In addition, the method combining the multi-BERT hybrid system and the GHNL over the translated radiology reports with cross-lingual word alignment obtained the best Recall in the task. While the improvement is marginal due to the low quality of the word alignment, the approach provides a framework to leverage English NER tools for texts in relatively low-resource languages and domains with limited corpus (e.g. Spanish radiology reports in this task). This approach is practical as it does not require training data and has the potential to be improved with more accurate word alignment.

We will explore fine-tuning new pre-trained language models with larger Spanish corpus such as PadChest [25] in the same way as [9] and extend it to the multi-language classification of radiology reports from [26]. We also suggest to further study using translation and cross-lingual word alignment to leverage English NER tools for Spanish clinical texts. The current performance using the neural machine translation (NMT) tool, Microsoft Translator, with GHNL API is poor due to the inaccurate alignment. Jointly generating accurate alignment with translations in NMT is an open question being addressed (e.g. in [27, 28]) and to be applied for low-resource NER in future studies.

Acknowledgments

The authors would like to thank to members in the Clinical Natural Language Processing Research Group and KnowLab in the University of Edinburgh and University College London for their valuable discussion and comments. This work was supported by the HDR UK National Text Analytics Implementation Project, the HDR UK National Phenomics Resource Project, Wellcome Institutional Translation Partnership Awards (PIII032, PIII029, PIII009), the Alan Turing Institute via Turing Fellowships and Turing project funding (ESPRC grant EP/N510129/1), a Legal and General PLC (research grant to establish the independent Advanced Care Research Centre at University of Edinburgh). Legal and General PLC had no role in conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of Legal and General PLC.

References

[1] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, H. Wu, B. Alex, A systematic review of natural language processing applied to radiology reports, BMC Medical Informatics and Decision Making 21 (2021) 179. URL: https://doi.org/10.1186/s12911-021-01533-7. doi:10.1186/s12911-021-01533-7.

[2] 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 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 Workshop Proceedings, CEUR-WS.org, 2021.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] 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. doi:10.18653/v1/N19-1423.

[5] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 465–476. doi:10.18653/v1/2020.acl-main.45.

[6] M. Eberts, A. Ulges, Span-based joint entity and relation extraction with transformer pre-training, in: ECAI, 2020.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2021. arXiv:2007.15779.

[9] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, M. Lungren, Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1500–1519. doi:10.18653/v1/2020.emnlp-main.117.

[10] D. A. Wood, J. Lynch, S. Kafiabadi, E. Guilhem, A. Al Busaidi, A. Montvila, T. Varsavsky, J. Siddiqui, N. Gadapa, M. Townend, M. Kiik, K. Patel, G. Barker, S. Ourselin, J. H. Cole, T. C. Booth, Automated labelling using an attention model for radiology reports of mri scans (alarm), in: T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, C. Pal (Eds.), Proceedings of the Third Conference on Medical Imaging with Deep Learning,

volume 121 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 811–826. URL: http://proceedings.mlr.press/v121/wood20a.html.

[11] P. Schrempf, H. Watson, E. Park, M. Pajak, H. MacKinnon, K. Muir, D. Harris-Birtill, A. O'Neil, Templated text synthesis for expert-guided multi-label extraction from radiology reports, Machine Learning and Knowledge Extraction 3 (2021) 299–317. doi:10.3390/make3020015.

[12] A. Grivas, B. Alex, C. Grover, Tobin, R., Whiteley, W., Not a cute stroke: Analysis of Rule- and Neural Network-Based Information Extraction Systems for Brain Radiology Reports, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020.

[13] 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.

[14] V. Cotik, H. Rodríguez, J. Vivaldi, Spanish Named Entity Recognition in the Biomedical Domain, in: J. A. Lossio-Ventura, D. Muñante, H. Alatrista-Salas (Eds.), Information Management and Big Data, volume 898 of *Communications in Computer and Information Science*, Springer International Publishing, Lima, Peru, 2018, pp. 233–248. doi:10.1007/978-3-030-11680-4-23.

[15] V. Cotik, D. Filippo, J. Castaño, An Approach for Automatic Classification of Radiology Reports in Spanish., Studies in Health Technology and Informatics 216 (2015) 634–638. URL: https://europepmc.org/article/med/26262128.

[16] L. Goeuriot, H. Suominen, L. Kelly, L. A. Alemany, N. Brew-Sam, V. Cotik, D. Filippo, G. Gonzalez Saez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, J. Vivaldi, M. Viviani, C. Xu, Clef ehealth evaluation lab 2021, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2021, pp. 593–600.

[17] 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.

[18] 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.

[19] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, 2020. doi:10.5281/zenodo.1212303.

[20] J. Turian, L. Ratinov, Y. Bengio, Word representations: A simple and general method for semi-supervised learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 384–394. URL: http://dl.acm.org/citation.cfm?id=1858681.1858721.

[21] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[22] A. Bodnari, Healthcare gets more productive with new industry-specific AI tools, 2020. https://cloud.google.com/blog/topics/healthcare-life-sciences/now-in-preview-healthcare-natural-language-api-and-automl-entity-extraction-for-healthcare, accessed

15 Mar, 2021.

[23] V. Tablan, I. Roberts, H. Cunningham, K. Bontcheva, Gatecloud.net: a platform for large-scale, open-source text processing on the cloud, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371 (2013) 20120071. doi:10.1098/rsta.2012.0071.

[24] G. Gorrell, X. Song, A. Roberts, Bio-yodie: A named entity linking system for biomedical text, arXiv preprint arXiv:1811.04860 (2018).

[25] 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. doi:https://doi.org/10.1016/j.media.2020.101797.

[26] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data 6 (2019) 317. doi:10.1038/s41597-019-0322-0.

[27] K. Song, X. Zhou, H. Yu, Z. Huang, Y. Zhang, W. Luo, X. Duan, M. Zhang, Towards better word alignment in transformer, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1801–1812. doi:10.1109/TASLP.2020.2998278.

[28] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, Y. Liu, Neural machine translation with explicit phrase alignment, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1001–1010. doi:10.1109/TASLP.2021.3057831.