A Hierarchical Approach for Extracting and Displaying Entities and Relations from Radiology Medical Reports

Gabriel Toyoda[1], **Yunevda Rojas**[1], **Juan G. Colonna**[1], **Joao Gama**[2]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Caixa Postal 69077-470 – Manaus – AM – Brasil

[2]Faculdade de Economia – Universidade do Porto (U. Porto)
Código Postal 4200-464 – Porto – Portugal
{gabriel.toyoda, yunevda, juancolonna}@icomp.ufam.edu.br
jgama@fep.up.pt

Abstract. *Extracting information from medical reports can be challenging due to the large volume of data. Therefore, this study proposes a method that uses a hierarchical classification approach with two levels, each consisting of a neural network instance. One for extracting clinical anatomical or observational entities along with their levels of uncertainty, and another for classifying the relations that exist between these entities. For this research, 600 radiological reports from the RadGraph dataset were used. The entity extraction task achieved an F1-score of 91%, while the entity classification and relation classification tasks achieved 88% each. Our hierarchical method enhances entity and relation classification performance by filtering and double checking classified entries.*

1. Introduction

A medical report is a record that gathers information about a patient's health. These kinds of reports generally lack specific writing standards, leading to a wide variety of formats, presentations, or unstructured data [Jain et al., 2021, Jensen et al., 2012]. Furthermore, the large volume of reports that exist, on account of their preservation over time, presents an additional challenge when attempting to retrieve data and extract information. To address these challenges, automated data extraction methods have been developed, as proposed by Jain et al. [2021], Solarte-Pabón et al. [2021], Sugimoto et al. [2021], Yim et al. [2016]. These methods are capable of taking unstructured data as input and providing entities, relations, or both as output data.

Hierarchical structures provide a natural and convenient way to organize a dataset and extract relevant information, as each element within it is a generic type of its subordinate elements while simultaneously being a specific subset of the main element [Naik and Rangwala, 2018]. Therefore, based on free texts from unstructured radiology reports, this study aims to develop a hierarchical classification method with two levels: a higher level entity recognition task and a lower level relation classification; each employing a Convolutional Neural Network (CNN) capable of extracting anatomical and observational entities, along with their levels of uncertainty, and identifying the relations between them.

This study utilizes the structured data from the RadGraph dataset, as provided by Jain et al. [2021]. The radiology reports in RadGraph dataset were originally from Irvin et al. [2019](CheXpert) and Johnson et al. [2019] (MIMIC-CXR) datasets, where three board-certified radiologists manually labeled 600 reports.

The importance of the information extracted from radiology reports through the developed method lies in providing data that can be reused for other complex tasks, such as classification or automatic report generation. Additionally, the extracted entities and relations can be used to retrieve information from previous patient reports for comparison with current patients. In this article, we present the extracted entities and relations highlighted in the text, which enhances their visibility.

2. Related Works

Over the past decade, significant strides in Machine Learning have opened doors to effectively tackle numerous Natural Language Understanding (NLU) tasks within the realm of Natural Language Processing (NLP) [Hapke et al., 2023]. One illustrative example lies in healthcare applications, where the objective might revolve around extracting pertinent information from radiology reports [Casey et al., 2021, Landolsi et al., 2023]. These reports, often presented in free text format, serve as critical documentation to convey diagnostic imaging findings [Pons et al., 2016].

Faced with the diverse array of NLP tasks, particularly those centered on named entity recognition (NER) and relation extraction (RE), convolutional neural networks (CNNs) have consistently demonstrated reliable performance. For instance, in the domain of RE, the Att-Pooling-CNN model, as proposed by Wang et al. [2016], achieves an F-score of 88.0%. Similarly, Santos et al. [2015] and Liu et al. [2013] utilized CNNs for relation classification through ranking (CR-CNN), achieving an F-score of 84.1%. However, the most notable advancement has been with Bidirectional Encoder Representations from Transformers (BERT). BERT, based on a transformer architecture, is specifically designed to comprehend text bidirectionally, incorporating contextual information from both left and right contexts, as demonstrated by Devlin et al. [2018].

BERT has served as the foundation for various research endeavors targeting NER and RE tasks Solarte-Pabón et al. [2021]. For instance, Wu and He [2019] proposed three BERT-based models for NER in Spanish radiology reports, achieving exact F1 and lenient F1 scores of 73.27% and 78.47%, respectively. In the realm of RE, Zeng et al. [2014] introduced R-BERT, a model that leverages pre-trained BERT with entity information, outperforming other approaches such as Deep Neural Network (DNN) - Softmax, with an F1 score of 89.25% compared to 82.7%.

Joint extraction models, like those combining BERT and R-BERT, have significantly improved performance in both NER and RE tasks, yielding noteworthy F1 scores. In our study, we follow a similar strategy, presenting a hierarchical model comprising two classifiers. The first classifier identifies entities, while the second classifier predicts the type of relation between two recognized entities. Our neural networks are constructed using Keras Dense Layers, Embedding Layers, and leverage the RadGraph database obtained from PhysioNet, enabling the simultaneous extraction of entities and relations.

3. Materials and Methods

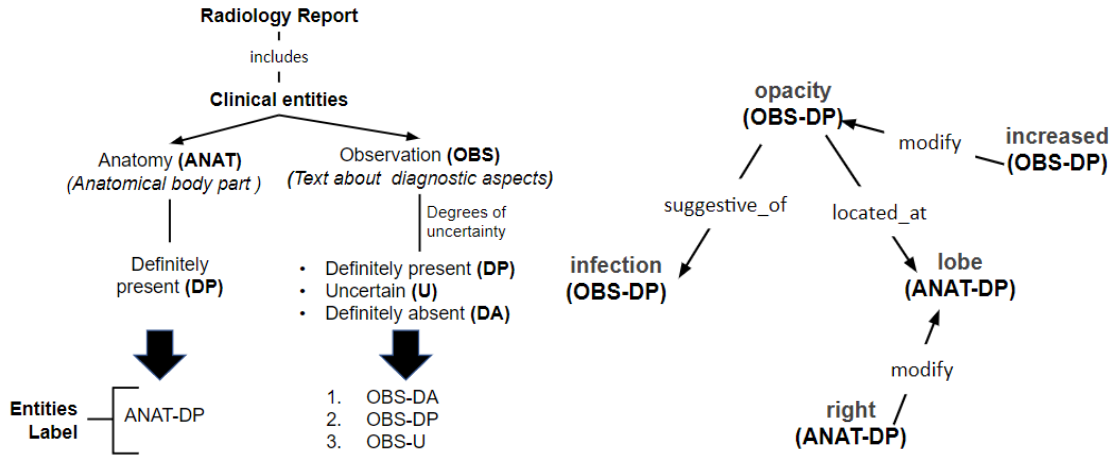
3.1. Dataset

RadGraph is a dataset available on the PhysioNet platform, was created to train and evaluate Natural Language Processing (NLP) models focused on the extraction of clinical

entities and relations in radiology reports [Jain et al., 2021]. This dataset, which contains reports in English from MIMIC-CXR (MIMIC Chest X-ray) and CheXpert (Chest eXpert), has manually annotated details by radiology and linguistics experts about the examination performed, radiological findings and patient diagnostic impressions [Jain et al., 2021].

For this research, a total of 600 reports were sourced from the RadGraph dataset. These reports were then segregated into two distinct datasets for different purposes. The first dataset comprised 500 reports (*train* and *dev* dataset partitions) exclusively from the MIMIC-CXR subset. These reports were utilized for tasks such as entity and relations recognition. Specifically, this subset contained 14,579 entities and 10,889 relations. Among these reports, 320 were allocated for training the models. The remaining reports were split into 100 for testing and 80 for validation purposes. On the other hand, the second dataset (*test*), comprising 100 reports, including 50 reports each from MIMIC-CXR and CheXpert datasets. These reports were employed to highlight the identification of entities and relations, potentially offering insights into how the trained models performs.

It is important to emphasize that each radiology report encapsulates clinical entities (one or more adjacent words), categorized as anatomical body parts (ANAT) or observations (OBS), as depicted in Figure 1a. These observations can encompass visual characteristics, identifiable pathophysiological processes, or a definitive disease diagnosis [Jain et al., 2021]. Furthermore, each observation has a graded level of uncertainty, ranging from definitely present (OBS-DP), uncertain (OBS-U), to definitely absent (OBS-DA). Correspondingly, ANAT-DP signifies the unambiguous presence of an anatomical concept.



(a) Radiology Report Structure. **(b) The relations between entities, adapted from Jain et al. [2021].**

Figure 1. Ontology of entities and relations used to label anatomic and observation entities.

The relations between entities, Figure 1b, can be categorized into three types: "suggestive of", representing the relation between two observation entities; "located at", indicating the relation between observation and anatomy entities; and "modify", corresponding to the relation between two observations or two anatomical entities [Jain et al., 2021]. Furthermore, each report data is stored in a dictionary, each of them being acessed

with the highest level key. All the reports contains a key named "text" containing the report descriptions, a "data_split" key, "data_source" and a dictionary of labeled "entities". Each "entity id" key within the dictionary contains the following fields: *"tokens"* are tokens of any entity, *"labels"* are anatomy or observations with the uncertainty level, *"start_ix"* and *"end_ix"* are the position of the first and last token and *"relations"* is a list of relation labels and the related entity ID.

3.2. Pre-processing

Our experiments were conducted using the Python programming language. In the preprocessing phase, we began by merging the *train* and *dev* files. The resulting file named *merged* contained 500 reports, which were then iterated through to extract annotated entities and their corresponding labels. These annotated report strings were saved in a Python list. Additionally, each entity was processed to extract its token and label, along with any relations it had with other entities.

We utilized the SpaCy library to tokenize both entities and non-entities within the report texts. These texts, alongside their corresponding entities and non-entity tokens, underwent preprocessing to eliminate non-alphanumeric characters, redundant whitespaces, and to normalize all characters to lowercase. The same preprocessing procedures were applied to the *test* data, resulting in a total of 933 final tokens. Subsequently, we employed the Keras Tokenizer to map the vocabulary of entity tokens to integer values, a necessary format for inputting into the embedding layer of our proposed Neural Network. In cases where tokens were out-of-vocabulary, they were represented by "1".

RadGraph text reports consist of two primary sections: the first contains technical details regarding image acquisition or patient-specific information, while the second presents specialized findings in free text format. Typically, these sections are delineated by one of three keywords: FINDINGS or IMPRESSION. Nonetheless, not all reports adhere to this pattern; occasionally, the FINDINGS keyword may be absent, or entity labeling may occur before the IMPRESSION keyword, resulting in missed labeling before this point. To address this inconsistency, we conduct a token count within the delimited report. If FINDINGS is absent but IMPRESSION is present, the number of entities in the report must match the number of labeled entities. Consequently, we define that following any of these keywords, any unclassified text should be designated as a non-entity. By implementing this method, 439 out of 500 reports from *merged* file and 96 out of 100 reports from *test* file were accurately segmented and are now prepared for tokenization. Additionally, for each unlabeled token in these JSON files, a new "non-entity" label was created as explained in the following section.

3.2.1. The non-entity Class

The MIMIC-CXR dataset comprises a total of 14,579 entities, with 339 entities spanning multiple tokens, which were subsequently excluded, resulting in 14,240 entities along with their corresponding labels. It is crucial to note that each radiology report consists of a sequence of tokens, some of which are labeled while others remain unlabeled. In order to equip our proposed model with the capability to handle unlabeled tokens within the unstructured text reports, we made a strategic decision to augment the dataset by

introducing a negative class. This entailed automatically labeling each token that was not assigned a specific class label within the reports as "non-entity". Tokens that were already labeled retained their original classifications. This enhancement to the dataset forms a key component of our contribution.

Given that the total number of tokens significantly exceeds the count of labeled entities, simply selecting all non-labeled tokens to form the negative class would lead to an imbalanced dataset, with a much larger proportion of data in this new non-entity class. To address this imbalance, a sample equivalent to the number of occurrences of the most frequent entity label was randomly selected from the non-entity data. This resulted in 6,216 non-entity tokens being added to the 14,240 already labeled entities, yielding a total of 20,456 tokens for training our entity recognition model. These sampled non-entities were labeled as "N-ENT", as illustrated in Figure 2a.

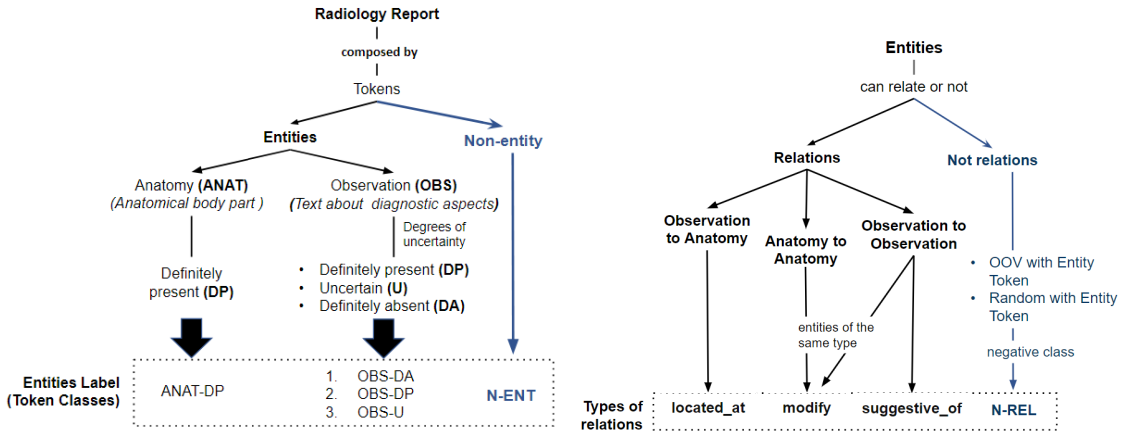
3.2.2. The non-relation Class

As elaborated in the preceding section, the incorporation of a non-entity class serves a crucial purpose in our hierarchical methodology. This addition prevents the initial classifier from forwarding pairs of tokens to the second level of classification if one or both tokens are not genuine entities. This cautionary step is essential since the secondary classifier in the hierarchy is tasked with identifying relationships between pairs of entities. In simpler terms, if there is a misidentification of at least one of the two entities, the response from the second classifier becomes inaccurate. This is because the second classifier relies on accurate entity identification to correctly label relationships between the two entities. More details about our hierarchical classification model are given in section 3.3.

In real-world scenarios, the risk of misclassifying a non-entity as an entity introduces the possibility of a false positive, subsequently passed on to the second level of classification. Consequently, the second classifier independently assigns a label to the relationship, making it challenging to rectify errors originating at the first level. To improve the resilience of the second classifier against false positives from the initial classification stage, we have chosen to introduce a new label for relationships involving either an entity and a non-entity or between two non-entities. This additional label aligns with the negative class concept discussed earlier but now encompasses pairs of tokens labeled as "non-relation". Figure 1b depicts this new relation ontology.

To enhance training for the relation classifier, we collected a total of 10,598 examples featuring positively related pairs of entities along with their corresponding relation labels. Subsequently, akin to the entity pre-processing phase, we generated examples of non-related pairs and labeled them as "N-REL", effectively creating a new negative class. The negative relation dataset comprises two types of pairs: an entity token paired with an Out-Of-Vocabulary token, and an entity token paired with another entity token that is not related. This introduction of a new label constitutes another aspect of our contribution to the research. The composition of the Relation Classifier data is illustrated in Figure 2b.

In summary, two types of entity pairs are now labeled as non-relation: the first type consists of an entity token paired with an Out-Of-Vocabulary token (representing unknown words), while the second type involves a pair of entities that are unrelated. This



(a) New ontology of the radiology report dataset.

(b) New relation data composition schema.

Figure 2. Figures a and b show the incorporation of a new negative class. The blue branches represent the extension achieved through our methodology.

latter case helps prevent misclassification of erroneous relations between pairs of true positive entities. All pairs were generated through random sampling. However, to ensure that randomly formed pairs do not overlap with the list of positively related entities, we conduct a thorough check. Ultimately, the number of non-relations equals the count of the most frequently occurring class of true relations, which is 6,426 labeled as "modify", to keep the balance between classes.

3.3. Hierarchical classifier

Our method employs a hierarchical classification approach comprising two levels. In the first level, all tokens are processed to discern their entity types, while the second level analyzes only pairs of tokens positively classified by the first level, identifying the relation between them. This hierarchical setup, known as "hierarchical classification", offers a key advantage: the initial level acts as a filter for the subsequent classifier, enhancing the likelihood of achieving a more precise final outcome; the last level acts as a double checker, negatively classifying a relation that includes a unfiltered non-entity. Figure 3 visually depicts this structure, with each level consisting of a neural network instance.

The Entity Classifier is the initial model in the hierarchical process. It takes token integers as inputs and outputs tokens classified into four positive classes (ANAT-DP, OBS-DA, OBS-DP, OBS-U). Hyperparameters include an input dimension of 934 (vocabulary size), an output dimension of 96 (dense vector size), and an input length of 1 (handling one token at a time). The output of the Embedding Layer connects to a Flatten Layer, then to a Dense Layer with 5 output neurons and a softmax activation function for class probabilities. The model compiles using Adam optimizer, Categorical Crossentropy loss function, and accuracy metrics.

The second model in the hierarchy is the Relation Classifier, which takes pairs of token integers as inputs and predicts the presence or absence of a relation between them. Its architecture includes two Input layers, receiving two integers for each pair of tokens. The Embedding Layer converts these integers into dense vectors, with hyperparameters as: 934 for input dimension, an output dimension of 96, and an input length of 2. The

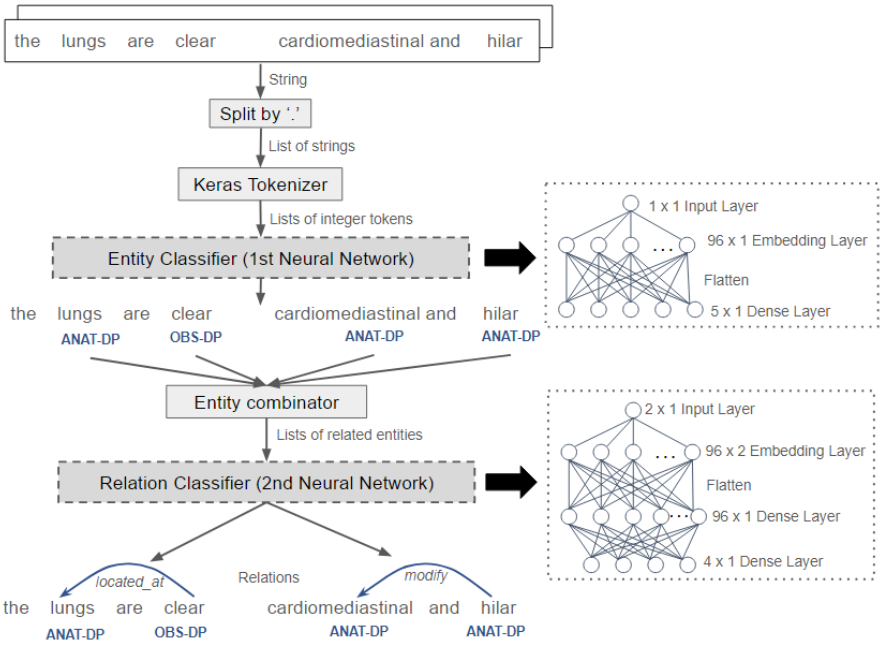


Figure 3. Architecture of the Hierarchical Model.

output vectors are then flattened and connected to Dense Layers. The final Dense Layer has 4 output neurons corresponding to the three types of relations and non-relation, each with a softmax activation function for class probabilities. The model is compiled using the Adam optimizer, Categorical Crossentropy loss function, and accuracy as the evaluation metric.

3.4. Proof of concept

In this section, we aim to apply the model trained with *merged* data in *test* data. With this, we aspire to evaluate the Entity Classifier on performing entity extraction on a unseen free-text scenario. We will apply the entity model to every token in the texts, defining which are entities (ANAT-DP, OBS-DU, OBS-DP, OBS-U) and non-entities.

The process starts when the report text is split into sentence fragments separated by periods, Figure 3. Then, using the Keras *texts_to_sequences* tokenizer function along with spaCy tokenizer, we transform each word of the fragmented sentences into integers. The Keras tokenizer is responsible of transforming every token obtained by spaCy tokenizer into an integer. Thus, entities are designated by values greater than 1, while non-entities are represented by the value 1. Afterward, every obtained token is passed to the Entity Recognizer model, each returning the probability for each of the five classes (ANAT-DP, N-ENT, OBS-DA, OBS-DP, OBS-U). Subsequently, the tokens are stored in arrays along with the most probable label.

Afterward, every obtained token is passed to the Entity Recognizer model, each returning the probability for each of the five classes (ANAT-DP, N-ENT, OBS-DA, OBS-DP, OBS-U). Subsequently, the tokens are stored in arrays along with the most probable label. It is important to note that entities that are from the same fragmented sentences are stored together. For the relations between the extracted entities, a function that receives the list of entities and generates all possible combinations up to a certain range in the

same section of text, which in this case is divided by periods. Then, the pairs are passed to Relation Classifier that returns the probability for each class. In the end, the pair is stored along with the most probable label.

3.5. Evaluation methodology and highlighting entities and relations process

The evaluation methodology involves comparing entities identified by the first classification level in each report to form an expected list, with alphabetic entities recognized and classified by the developed Entity model. An identified entity is considered correct only if both the token and its label have a corresponding match in the expected list. If an identified entity lacks a counterpart in the expected list of labels, it is counted as a false negative. Conversely, if an entity is identified that is not present in the expected list, it is counted as a false positive. Finally, if an identified entity token lacks an expected corresponding entity token, it means that the token is not an entity. In this case, it is recorded as a false negative for the "N-ENT" label and a false positive for the incorrectly predicted label.

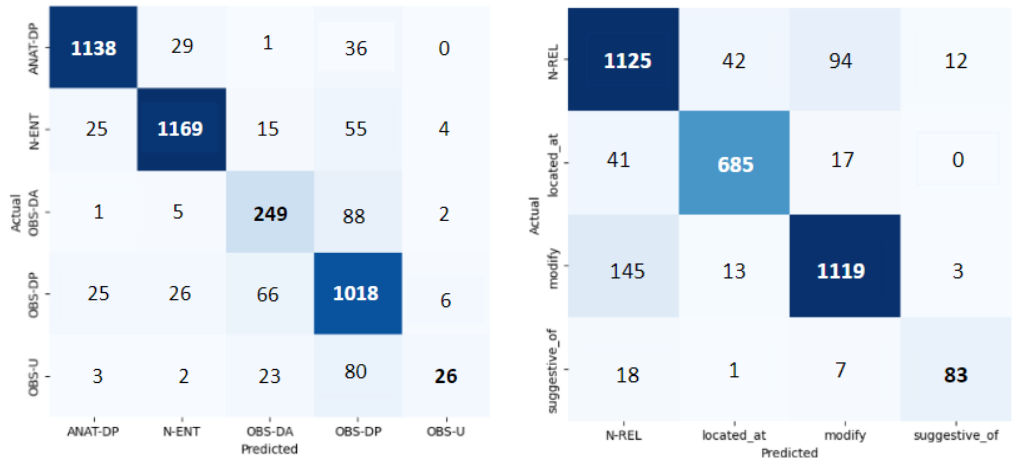
We will use displaCy dependency visualizer along with the extracted data of the radiology reports to finally represent and better emphasize the relevant information found in each medical report. SpaCy is used to tokenize the report text, transforming it into a Doc object. Then, each token in the obtained Doc is checked if it is present in the extracted entity list. If yes, the entity token is tagged with its label, and its index is saved for the arcs elaboration. Using the entity token indexes obtained and the pairs of entities obtained in previous sections, arcs are defined and labeled to represent the dependencies (relations) between words (entities).

4. Results

4.1. Entity model results (First level)

The model's results are available in both Figure 4a and Table 1, where P stands for precision, R for recall and F1 for F-score. These results indicate that the model accurately predicted 1,138 instances for the ANAT-DP class, with 53 false positives and 66 false negatives, resulting in a precision of 96%, a recall of 95%, and an F1-score of 95%. For the N-ENT class, similar metrics were obtained, with 1,196 instances correctly predicted, 62 false positives, and 99 false negatives, resulting in 95% precision, 93% recall, and 95% F1-score. Nevertheless, a less favorable scenario is observed for observational entities. The model shows preference towards the OBS-DP class, which is the class with the largest number of instances, obtaining 1,018 true positives, 259 false positives, and 123 false negatives, resulting in a precision of 80%, a recall of 89% and an F1-score of 84%.

The OBS-DA is the second most frequent Observational class in the training data, comprising 345 test instances. Out of these, 249 were correctly labeled, 96 were false negatives (with 88 predicted as OBS-DP), and 104 were false positives (66 of which actually belonged to OBS-DP). As a consequence, the precision was 70%, the recall of 72%, and an F1-score of 71%. The OBS-U is the least represented class in the database with 134 test instances. Among these, only 26 instances were correctly classified, 108 were false negatives (80 of these being classified as OBS-DP), and 12 false positives, resulted in a precision of 68%, a recall of 19%, and an F1-score of 30%. Despite its low F1 score, its accuracy is enough to filter out erroneous tokens and not pass them to the



(a) Entity recognition.

(b) Relation Classifier.

Figure 4. Confusion matrices of the every levels of the classifier.

Class	P	R	F1	Support
ANAT-DP	0.96	0.95	0.95	1204
N-ENT	0.95	0.93	0.94	1268
OBS-DA	0.70	0.72	0.71	345
OBS-DP	0.80	0.89	0.84	1141
OBS-U	0.68	0.19	0.30	134
Accuracy			**0.88**	4092
Macro Avg	0.82	0.74	0.75	4092

Table 1. Classification Metrics for Entity Model.

Class	P	R	F1	Support
N-REL	0.84	0.88	0.86	1273
located_at	0.94	0.92	0.93	743
modify	0.90	0.87	0.88	1280
suggestive_of	0.90	0.74	0.81	109
Accuracy			**0.88**	3405
Macro Avg	0.89	0.85	0.87	3405

Table 2. Classification Metrics for Relation Model

second-level classifier. Overall, these results present an accuracy of 88%, justifying the excellent performance of the first level of the proposed hierarchical classifier.

4.2. Relation model results (Second level)

The results of the relation classifier can be found in Figure 4b and in Table 2. The "modify" class contains 1,119 correctly predicted test instances, 118 false positives, and 161 false negatives, resulting in 90% precision, 87% recall, and 88% F1-score. For the N-REL negative class, the second most frequent, 1,125 instances were correctly predicted by the model, with 204 false positives (145 are 'modify' entities) and 148 false negatives (94 are "modify" entities), resulting in a precision of 84%, a recall of 88% and an F1-score of 86%.

Similar metrics were obtained for the other two classes, even though they contain fewer training instances. The "located_at" relation class, obtained 685 true positives, 56 false positives (42 of which were N-REL), and 58 false negatives (41 of which were classified as N-REL), resulting in a precision of 94%, a recall of 92% and an F1-score of 93%. The "suggestive_of" is the least represented relation class in the test data with 109 test instances. In this case, 83 instances were correctly classified, 26 false negatives, and 15 false positives, resulting in a precision of 90%, a recall of 74%, and an F1-score of 81%. For the developed solution, similar metrics were obtained for all four classes, although the "suggestive_of" class has a slightly lower number of test and train instances, as well as a lower Recall and F1-score compared to the others.

4.3. Proof of concept results

The hierarchical evaluation method results are shown in Table 3. As showcased, the most represented positive class, ANAT-DP, is correctly predicted 1,040 times, with 49 false negatives and 101 false positives, resulting in 91% of precision, 96% of recall and 93% F1-score. For the N-ENT class, our negative class, 4,102 instances were correctly predicted, with 154 false positives, and 217 false negatives, resulting in 96% precision, 94% recall, and 95% F1-score. We would like to emphasize that the incorporation of this negative class constitutes a significant aspect of our contribution. Its inclusion greatly enhances the performance of the hierarchical classifier, leading to exceptionally favorable results.

Class	**P**	**R**	**F1**	**Support**
ANAT-DP	0.91	0.96	0.93	1089
N-ENT	0.96	0.94	0.95	4319
OBS-DA	0.79	0.64	0.70	378
OBS-DP	0.77	0.87	0.82	983
OBS-U	0.60	0.24	0.34	83
Accuracy			0.91	6852
Macro Avg	0.80	0.73	0.75	6852

Table 3. Classification Metrics for Entity Model in test data.

Similarly as seen in Section 4.1, the observational entities got a worse result compared to the others. Again, the model tends towards OBS-DP class, which had more instances for training, obtaining 860 true positives, 123 false negatives, and 256 false positives, resulting in a precision of 77%, a recall of 87% and an F1-score of 82%. The OBS-DA class had 378 instances. Out of these, 242 were correctly labeled, 64 were false positives (interestingly, none of it was ANAT-DP), and 136 were false negatives. For this class, the precision is 79%, recall of 64%, and an F1-score of 70%. The OBS-U is the least frequent class evaluated, with only 83 instances. Of these, 20 instances were true positives, 63 were false negatives, and 13 were false positives, resulting in a precision of 60%, a recall of 24%, and an F1-score of 34%.

Entities extracted from the radiological report text can be effectively highlighted and integrated within the report itself using the displaCy visualizer, as illustrated in Figure 5 with two examples of reports: "*The lungs are clear. No acute process. Aorta shows no abnormalities.*" and "*Cardiomegaly is stable . No acute intrathoracic process.*". This exemplifies our final summarized report, as it will be presented to healthcare professionals.

5. Conclusion

The proposed design scheme for entity recognition in radiological reports seamlessly integrates entity classes (ANAT-DP, OBS-DA, OBS-DP, OBS-U) with a dedicated negative class (N-ENT). This approach simplifies the recognition task by enabling both entity identification (determining if it is an entity or not) and classification into one of the four specified classes. Additionally, the relation model addresses the classification of relationships between entities by incorporating a negative class (N-REL) to identify non-related entities. This inclusion facilitates relation classification in a manner akin to the entity

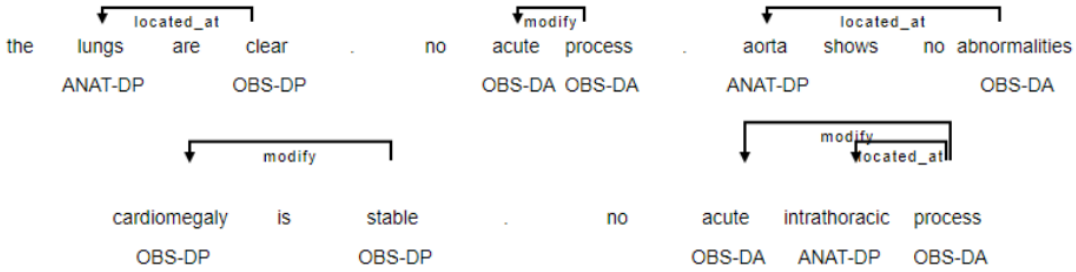


Figure 5. Used displaCy dependency viewer to highlight entities and relationships after employing our trained hierarchical classifier.

model. The incorporation of these two new negative classes significantly enhances the learning capabilities of our model, effectively preventing trivial mistakes.

Our hierarchical method, employing two levels of classification, offers a distinct advantage. By initially categorizing all tokens to discern their entity types and subsequently focusing on positively classified pairs for relation identification, we significantly enhance the precision of our final outcomes through meticulous filtering and double-checking.

The entity recognition and classification model proposed enabled the extraction of 2,162 entities and 4,102 non-entities from a total of 96 reports in the test dataset, yielding a micro F1-score of 0.91. Furthermore, when solely performing entity classification on the *merged* dataset without the need to tokenize the text, as done in the test dataset, the model achieved a micro F1-score of 0.88%. Regarding the relation model, conducting the relation classification task on the *merged* dataset resulted in a micro F1-score of 0.88%.

Our proof of concept effectively showcases the capability of the proposed models to extract entities from text reports not seen during training. Moreover, the extracted data can be seamlessly integrated with the displaCy dependency visualizer to emphasize relevant information from each report. Ultimately, this innovative scheme holds promise for various medical applications related to natural language processing (NLP), including automated report generation.

In future work, we could leverage Large Language Model (LLM) architectures to improve the contextual capabilities of our model and achieve better entity recognition. Furthermore, we could explore the effectiveness of Med-Gemini – specialized in medical tasks – in carrying out the proposed tasks.

6. Acknowledgements

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 39 of Decree n°10.521/2020, was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n°8.387/1991 through agreement 001/2020, signed with UFAM and FAEPI, Brazil. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001. This work was partially supported by Amazonas State Research Support Foundation - FAPEAM - through the POSGRAD 2024 project.

References

A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, et al. A systematic review of natural language processing applied to radiology reports. *BMC medical informatics and decision making*, 21(1):179, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

H. M. Hapke, H. Lane, and C. Howard. Natural language processing in action: Understanding, analyzing, and generating text with python, 2023.

S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516, 2023.

C. Liu, W. Sun, W. Chao, and W. Che. Convolution neural network for relation extraction. In *International conference on advanced data mining and applications*, pages 231–242. Springer, 2013.

A. Naik and H. Rangwala. *Introduction*, pages 1–11. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01620-3. doi: 10.1007/978-3-030-01620-3_1. URL https://doi.org/10.1007/978-3-030-01620-3_1.

E. Pons, L. M. Braun, M. M. Hunink, and J. A. Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016.

C. N. d. Santos, B. Xiang, and B. Zhou. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*, 2015.

O. Solarte-Pabón, O. Montenegro, A. Blazquez-Herranz, H. Saputro, A. Rodriguez-González, and E. Menasalvas. Information extraction from spanish radiology reports using multilingual bert. *CLEF eHealth*, 2021.

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*, 116:103729, 2021.

L. Wang, Z. Cao, G. De Melo, and Z. Liu. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, 2016.

S. Wu and Y. He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364, 2019.

W.-w. Yim, T. Denman, S. W. Kwan, and M. Yetisgen. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science Proceedings*, 2016:455, 2016.

D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344, 2014.