Information Extraction from Polish Radiology Reports using Language Models

Aleksander Obuchowski **Barbara Klaudel** **Patryk Jasik**
Gdańsk University of Technology, TheLion.ai
obuchowskialeksander@gmail.com
barbara.klaudel@student.pg.edu.pl
partyk.jasik@pg.edu.pl

Abstract

Radiology reports are vital elements of directing patient care. They are usually delivered in free text form, which makes them prone to errors, such as omission in reporting radiological findings and using difficult-to-comprehend mental shortcuts. Although structured reporting is the recommended method, its adoption continues to be limited. Radiologists find structured reports too limiting and burdensome. In this paper, we propose the model, which is meant to preserve the benefits of free text, while moving towards a structured report. The model automatically parametrizes Polish radiology reports based on language models. The models are trained on a large dataset of 1200 chest computed tomography (CT) reports annotated by multiple medical experts reports with 44 observation tags. Experimental analysis shows that models based on language models are able to achieve satisfactory results despite being pretrained on general domain corpora. Overall, the model achieves an F1 score of 81% and is able to successfully parametrize the most common radiological observations, allowing for potential adaptation in clinical practice. Our model is publicly available [1].

1 Introduction

A radiology report is the most important product radiologists generate to help direct patient care. They are vital to the referring physicians that depend upon them while making a decision about further treatment of a patient. It represents the highest level of radiologists' synthesis and insight into a patient's condition. However, radiology reports are almost always formulated in natural language. Natural language is flexible and enables the writer to express the same idea in a variety of different ways with varied complexity. As a result, the style, length, and level of detail vary among the radiologists, even among those coming from the same institution. Moreover, the reports often contain misspellings and mental shortcuts. Such properties make them difficult to analyze for referring physicians and incomprehensible to patients.

The well-known initiative of the American College of Radiology – Imaging 3.0 introduced a roadmap to transition radiological practice from volumed-based care to value-based care. The critical element of the roadmap was the adoption of structured reporting. A structured report (SR) is a report generated from a predefined, standardized format. The SR is considered a better strategy in terms of reduction in diagnostic error, comprehensiveness, adherence to consensus guidelines, and reduction in the omission of findings and other preventable errors. The negative effects of medical errors were publicized by the report of the Institute of Medicine "To Err is Human" (Donaldson et al., 2000). The report highlighted the importance of limiting preventable medical errors, such as omission in reporting radiological findings.

The adoption of SR was defined as a critical step to provide the best quality of service to referring physicians and patients by both the European Society of Radiology (ESR) and Radiological Society of North America (RSNA) (European Society of Radiology (ESR), 2018). The SR is believed to improve the quality of reports by providing a checklist to ensure that all relevant points were addressed. Moreover, the SR is easier to integrate with tools helping radiologists express relevant information, e.g., CO-RADS classification (Prokop et al., 2020). Lastly, they could facilitate the adoption of value-based healthcare – a new healthcare delivery model in which healthcare providers are paid based on patient outcomes, not the number of performed procedures.

[1] github.com/AleksanderObuchowski/PLRadIE

Although structured reports have many benefits, their acceptance among radiologists is still limited (Faggioni et al., 2017). They require radiologists to change their habits which they often practiced for many years. The radiologists may be reluctant to change for many reasons, including the limited scope of expression resulting in the downgrade of quality, the feeling that there is no clinical necessity to change, and even because they perceive it as an attack on the art of medicine (Ganeshan et al., 2018). With SR, the structure of a report would also have to be manually updated with the changes in classification ontology, possibly resulting in discrepancies between the latest state of knowledge and clinical practice. Moreover, while the proposed structured reports schema could be introduced in clinical practice, it does not solve the problem of already generated reports, where the clinical observations may need to be rewritten to follow the parameterized structure, therefore resulting in additional labor. Although those older reports might not be used in further clinical practice due to being outdated, their parametrization could still be beneficial for data analysis and training of machine learning models.

To bring the most out of both structured reporting and free-texts, in this paper we propose a model for the automatic parametrization of Polish radiology reports based on language models. The model's role is to assign one of 44 labels to each radiological observation. Example texts with extracted radiological observations are shown in Figure 1. Formally, our task falls under the information extraction category, as the goal of the model is to detect spans corresponding to specific radiological findings rather than detect a broader set of entities. This was motivated by the fact, that as shown in (Steinkamp et al., 2019) systems that strictly perform named entity recognition-level tasks are insufficient for answering clinical queries. For example, in the sentence "No lesion observed," a NER-only system could (correctly) identify "lesion" as an entity, but cannot correctly answer the intended question. Moreover, we decided to model this task as sequence labeling rather than multi-class sequence text classification, as not only more informative to the end user by also previous work has shown that token-level labeling can result in improved accuracy (Lew et al., 2021). To the best of our knowledge, this is the first model for information extraction from radiology reports in the Polish language.

2 Related work

2.1 Structured Reporting

Structured reporting in radiology has been a subject of debate in the last decade. Even though free text is still the dominant report format, there have been several approaches that received some attention. The most widely-spread form of structured reporting are disease-specific templates, such as BI-RADS (Liberman and Menell, 2002) and CO-RADS (Prokop et al., 2020) schemes. Such templates provide a guideline with a list of features, which presence or absence should e.g. indicate that the disease has greater progression. An important step towards SR was DICOM Structured Reporting (DICOM SR) (Hussein et al., 2004). It is a standard developed to store structured data and clinical observations along with the images. Medical images are usually stored in a Digital Imaging and Communications in Medicine (DICOM) format. DICOM format was created to enable the interoperability of medical images. The standard was widely adopted in any field of medicine where medical images play a significant role. DICOM SR was developed to link the clinical notes to the images within the same format.

RadLex (Datta et al., 2020) is a radiology lexicon produced by the Radiological Society of North America. It contains an ontology of radiology terms for use in radiology reporting, decision support, data mining, data registries, education, and research. It defines standard names and codes for radiology findings.

The idea of unifying terminology and linking the reports to the images was combined in the Annotation and Image Markup (AIM) project (Channin et al., 2010) of the National Institutes of Health Cancer Biomedical Informatics Grid. AIM was created to develop a uniform machine-readable format for storing both the image and a radiology report. It enables the description of an image using common data elements and controlled terminologies, such as RadLex. The usage of ontology enables easy queries and retrieval of information. The annotations and measurements made with AIM can be serialized as XML or DICOM SR.

Another approach was the RSNA's radreport.org reporting templates. The templates for various clinical scenarios provide a standardized radiology lexicon with the terms defined in Web Ontology Lan-

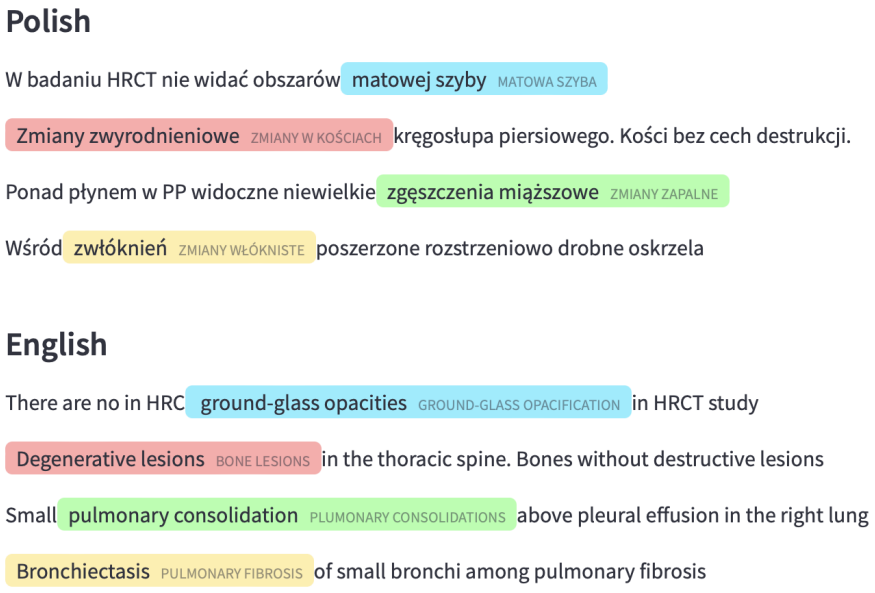

Figure 1: Sample of the annotated data. The report was stripped of the sentences without entities for visualization purposes.

guage (Bechhofer et al., 2009).

Although there have been some important attempts to make SR feasible, it is still at the early stage of adoption.

2.2 Clinical IE and NER

(Solarte-Pabón et al., 2021) proposed an information extraction model for Spanish radiology reports using a multilingual BERT (Devlin et al., 2018) model. The model's role was to parametrize ultrasonography reports. The corpus was annotated using ten different labels: Abbreviation, Anatomical Entity, Conditional Temporal, Degree, Finding, Location, Measure, Negation, Type of measure, and Uncertainty, and was split into a Training set (175 reports), Development set (92 reports), Test set (207 reports). Similar to our work the authors have also used BIO annotation schema, however, in our work, we focus solely on radiological findings but use much more detailed annotations with 44 different possible findings.

The dataset development by Jain et al. (2021) includes annotations for 500 radiology reports taken from the MIMIC-CXR dataset (Johnson et al., 2019), which comprises 14,579 entities and 10,889 relations. Additionally, the test dataset consisted of two independent sets of annotations for 100 radiology reports, sourced from both the MIMIC-CXR and the CheXpert dataset (Irvin et al., 2019). The authors evaluated the performance of several clinical language models, including BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), PubMedBERT (Gu et al., 2021), and BlueBERT (Peng et al., 2019), on this dataset.

(Sugimoto et al., 2021) proposed an information model comprising three groups of entities: observations, clinical findings entity, and modifiers entity. The model was trained and evaluated using 540 in-house chest CT reports. The authors have tested two types of models: BiLSTM-CRF and BERT and different pretraining datasets: Wikipedia articles (12 million sentences) and CR reports (118 thousand sentences).

CNNs have also been used in NER for the medical domain, for example in (Kong et al., 2021) where authors use a multi-level CNN layer to capture the information of neighboring characters and integrate them to generate a new embedding with context information for each character. An interesting approach can also be seen in (van de Kerkhof, 2016) where the authors use CNN for medical NER in the context of computer vision where the network is fed an image representing a medical document and its goal is to extract bounding boxes of the named entities. Zhang et al. (2022) use dilated convolutional neural networks (Akbik et al., 2018) to capture global information with fast computing speed.

Florez et al. (2018) use both character-based and word-based LSTM for clinical NER. LSTM layer is followed by a conditional random field (CRF) (Lafferty et al., 2001) to predict the most probable label sequence. Tang et al. (2019) also use the BiLSTM-CRF network for the identification of clinical texts that are modeled as a specific example of NER task.

Mykowiecka et al. (2009) presented a rule-based information extraction system developed for Polish medical texts, focusing on mammography reports and hospital records of diabetic patients. The system uses a special ontology and two separate models represented as typed feature structure hierarchies to extract data from documents. The system also addresses linguistic issues such as ambiguous keywords, negation, coordination, and anaphoric expressions.

2.3 Medical language models

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) (Lee et al., 2020) was the first domain-specific language model trained for the biomedical domain. It shares the architecture of the original BERT model and uses its weights as a starting point for further pretraining. The model uses PubMed abstracts PubMed Central and full text for further pre-training and domain adaptation. BioBERT obtained higher F1 scores in biomedical NER than the SOTA models at the time, achieving much better results than the standard BERT model.

ClinicalBERT (Huang et al., 2019) is a language model designed for the analysis of clinical narratives (e.g. physicians' notes) that are known to have differences in linguistic characteristics from both general texts and non-clinical biomedical texts (such as the ones used for training of BioBERT). The model was trained on 2 million discharge summaries and clinical notes and discharge summaries from the MIMIC-III database (Johnson et al., 2016). The authors showed that using clinical-specific contextual embeddings improves both general domain results and BioBERT results across 2 well-established clinical NER tasks and one medical natural language inference task.

BlueBERT (Peng et al., 2019) is a benchmark for evaluating medical language models based on 5 NLU tasks including Sentence Similarity, NER, Relation Extraction, Document Multilabel Classification, and Inference. The total model score is calculated as the macro-average of F1 scores and Pearson scores. The authors also share a dataset for pre-training medical language models based on PubMed abstracts and MIMIC-III, as well as two language models pre-trained on these datasets as baselines – one based on BERT and the other based on ELMo (Peters et al., 2018).

2.4 Polish Language Models

Unfortunately, at the time of writing this paper, there are no dedicated Polish Language Models for the medical domain. There are, however, several general domain models available:

Polbert (Kłeczek, 2020) is a Polish BERT-based language model trained on the Polish subset of Open Subtitles, ParaCrawl, Polish Parliamentary Corpus, and Polish Wikipedia with almost 2 billion words in total;

Polish RoBERTa (Dadas et al., 2020a) is a RoBERTa-based (Liu et al., 2019) language model trained on the Polish subset of the Common Crawl dataset;

PoLitBERT (Sopyła and Sawaniewski, 2021) is a Polish Roberta model trained on Polish Wikipedia, Polish literature and Oscar. The major assumption is that high-quality text will give a high performance model;

plT5 (Chrabrowa et al., 2022) is a set of T5-based language models trained on Polish corpora. The models were optimized for the original T5 denoising target. plT5 was trained on six different corpora available for the Polish language: CCNet Middle, CCNet Head, National Corpus of Polish, Open Subtitles, Wikipedia, Wolne Lektury;

papuGaPT2 (Wojczulis and Kłeczek, 2021) is a Polish version of the GPT-2 model trained on the Polish subset of multilingual Oscar corpus;

HerBERT (Mroczkowski et al., 2021) is a Polish BERT based model trained on NKJP, Wikipedia, and Wolne Lektury as well as CCNet and Open Subtitles. The model weights were initialized using weights from the multilingual XLM-RoBERTa model. The model was trained using only MLM objective with dynamic masking of whole words. The authors also introduced the KLEJ benchmark for evaluating Polish language models (Rybak et al., 2020) on which HerBERT is at the time of writing this work a state-of-the-art solution.

Table 1: Overview of the dataset

Entity (PL)	Entity (EN)	Train	Test
płyn w jamie opłucnowej	pleural effusion	722	184
zmiany włókniste	pulmonary fibrosis	631	165
zmiany w kościach	bone lesions	619	156
zmiany zapalne/niedodmowo-zapalne	pulmonary consolidation	543	143
matowa szyba	ground-glass opacities	482	141
rozedma	pulmonary emphysema	422	110
pojedyncze guzki	single nodules	384	95
rurka intubacyjna/wkłucie	endotracheal tube/venous line	254	71
rozstrzenie oskrzeli	bronchiectasis	253	62
konsolidacje w płucach	pulmonary consolidations	248	62
liczne guzki	numerous nodules	223	57
niedodma	atelectasis	202	57
adenopatia śródpiersia	mediastinal lymphadenopathy	202	50
przepuklina rozworu przełykowego	hiatal hernia	198	49
powiększenie serca	cardiomegaly	197	47
płyn w worku osierdziowym	pericardial effusion	175	41
zmiany o typie pączkującego drzewa	tree-in-bud pattern	164	40
patologie opłucnej	pleural disorders	162	39
odma opłucnowa	pneumothorax	156	34
jamy opłucnowe	pleural cavities	126	33
złamanie żeber	broken ribs	117	29
zwapnienia w naczyniach wieńcowych	coronary artery calcification	117	28
plaster miodu	honeycombing	117	26
zmiany w tarczycy	changes in the thyroid gland	94	20
pogrubienie ścian oskrzeli	bronchial wall thickening	83	19
zmiany w tkankach miękkich	soft tissue changes	81	19
poszerzenie pnia płucnego lub tt płucnych	pulmonary trunk dilatation	74	18
odma podskórna	subcutaneous emphysema	73	18
radiologiczne podejrzenie covid	radiological findings of COVID-19 infection	71	17
zwapnienia w miąższu	soft-tissue calcifications	68	17
wydzielina w oskrzelach	bronchial secretions	67	17
patologie nadnerczy	adrenal disorders	66	15
zmiany miażdżycowe aorty	atherosclerosis of the aorta	65	15
urządzenia kardiologiczne	cardiac devices	63	15
tętniak aorty poszerzenie aorty	aortic aneurysm	56	10
zastój w krążeniu płucnym	pulmonary congestion	46	9
adenopatia wnęk	hilar lymphadenopathy	39	9
odma śródpiersia	pneumomediastinum	35	9
kostka brukowa	crazy paving	17	6
patologie przewodu pokarmowego	gastrointestinal disorders	33	6
zatorowość płucna	pulmonary embolism	13	1
rozwarstwienie aorty	aortic dissection	11	1

3 Our Solution

3.1 Dataset

3.1.1 Collection and annotation

For our dataset, we used a real-life collection of 1200 randomly-selected radiological reports describing chest X-ray images. The data used was obtained from historical radiology reports collected at University Clinical Centre in Gdańsk, Poland. The annotation was modeled as a sequence labeling task, where each annotator was tasked with selecting spans in the report that corresponded to the specific tag. The words were labeled as entities following the Inside–Outside–Beginning (IOB) annotation schema (Ramshaw and Marcus, 1999) where the first token of each entity is labeled with the prefix "B-" standing for "Beginning" and each consecutive token of the same entity is labeled with the prefix "I-" standing for "Inside". The tokens not belonging to any entity are labeled as "O" standing for "Outside". The annotations were performed using lighttag annotation tool.

The annotation guidelines for observation tags were created out by radiologists, who selected 44 tags representing the most common radiological observations in the chest x-ray. However, we emphasize keeping annotation classes as general as possible so that the task of information extraction

can be easily transferred to other clinical domains. The dataset was annotated by 2 clinical experts with each annotator being resposible for half of the dataset.

The dataset and annotations guidelines are availabe upon resanable request.

3.2 Models

3.2.1 Pre-processing

The reports were anonymized by replacing occurrences of patients and radiologists names with empty strings. They were then split into sentences and tokenized using the Stanza NLP tool (Qi et al., 2020). This step was performed as the reports themselves were longer than the maximum number of tokens allowed for model inputs.

3.2.2 Train/Test split

The sentences were then split into training and test sets using the 80/20 ratio. The distribution of entities in the training and test set are shown in tables 1. From the initial dataset, 2 tags having fewer than 8 occurrences ("krwiak śródścienny aorty" and "zwężenie/koarktacja aorty") were removed due to insufficient number examples to perform the split.

In our implementation, we used 4 openly available Polish language models:

Polish-roberta-base-v2 – trained using Sentencepiece Unigram tokenization model and whole-word masking objective instead of classic token masking, the model also utilized the full context of 512 tokens and was retrained for 400k steps;

Polish-distilroberta – trained using knowledge distillation with RoBERTa-v2 base as a teacher model;

Polish-longformer – initialized with Polish RoBERTa (v2) weights and then fine-tuned on a corpus of long documents, ranging from 1024 to 4096 tokens.

All the models were pre-trained using a Polish subset of the Common Crawl corpus. The model's pre-training details are shown in (Dadas et al., 2020b).

We also used **HerBERT** (Mroczkowski et al., 2021).

In addition to Polish language models, we have also tested the performance of **mLUKE** (Ri et al., 2022) model. mLUKE is a multilingual version of the LUKE (Yamada et al., 2020) model based on XLM-RoBERTa that introduces improvement to the original model by using cross-lingual alignment information from Wikipedia entities.

In each case, the text was tokenized before being fed to the language model producing sub-word tokens. The resulting contextualized token embedding produced by the language model was then fed to a fully connected layer, mapping the token embeddings to entities in the "BIO" format. Only the first token of each word was used for predicting the entity, for the other tokens of a given word we assigned a special "-100" label that served as a mask in order not to count them in the loss function. This architecture is shown in Figure 2.

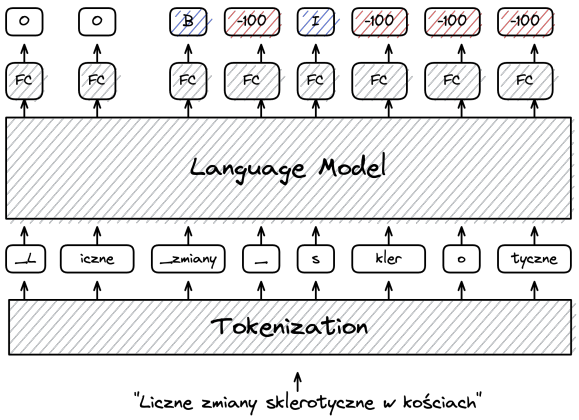

Figure 2: Visualization of deep language model-based approach

We also tested a baseline in form of forward and backward **Flair** (Akbik et al., 2019) embeddings for the Polish language trained on the Polish part of the Common Crawl dataset together with static word GloVe embeddings as suggested by the authors. The embedding layer was then followed by a single BiLSTM layer with a hidden size of 256. This layer was succeeded by a fully-connected layer mapping the hidden states of the BiLSTM layer to the named entities. The model also used Conditional Random Fields (CRF) for prediction, with Viterbi decoding as the loss function. The model was trained for 150 epochs with an initial learning rate of 0.1 which was decreased during training with the "anneal on the plateau" approach.

The models used categorical cross-entropy as the loss function and Adam optimizer with a learning rate of 1e-5 and linear warmup for 10% of steps.

4 Experiments and Results

The results for different models are presented in Table 2.

Table 2: Results of different language models

Model	Precision	Recall	F1-score
HerBERT	0.718	0.798	0.745
Flair	0.749	0.759	0.751
distilroberta	0.752	0.807	0.768
longformer	0.767	0.809	0.778
roberta	0.768	0.811	0.780
mLUKE	**0.791**	**0.826**	**0.809**

These results show that solutions based on deep language models perform better than the ones based on shallower Flair embeddings. The best model was mLUKE achieving an F1 score of 0.81. This can possibly be attributed to the fact that LUKE architecture involves entity-aware self-attention mechanism pre-training schema based on masking entities in large entity-annotated corpus retrieved from Wikipedia, therefore, making it suitable for the end task of sequence labeling. Another observation that can be made is that the best model based on mLUKE is trained solely on Wikipedia texts (as opposed to e.g. Common Crawl dataset used in Roberta pre-training) that have the potential to contain more domain-specific medical knowledge than corpora with casual vocabulary.

After performing additional analysis of the best model shown in Table 3, we observed that the accuracy seems to be the highest for tags with a larger number of examples in the training dataset which follows the standard trend associated with machine-learning-based approaches. However, a few classes (such as pulmonary embolism or aortic dissection) scored lower than average despite being largely represented in the training set. This can be attributed to the fact that those classes contain a lot of variations and clinical observations associated with them can be formulated in a number of ambiguous ways. Similarly, a few classes (such as emphysematous lungs and pulmonary fibrosis) scored well despite having only a few annotated examples. This can also be explained by the fact that those classes rarely appear in the reports and therefore contain fewer possible synonyms.

5 Discussion

In this work, we presented a tool for the parametrization of radiological reports for narrative reports written in natural language. In the interest of standardization and to help further research in this area, we introduced a general annotation scheme that was developed together with clinical experts based on common radiological observations. The results show that general domain language models can successfully be used in the radiology domain, although there is still room for improvements that can possibly be filled with domain-specific models. The detailed analysis of the results shows that the model is able to better capture the entities with fewer variations and higher representation in the training set. It can also be seen that the model rarely confuses different entities, but has some trouble with capturing the spans accurately. However, the model still achieved satisfactory results and with proper verification could successfully be used in clinical practice.

Information extraction is especially challenging with medical terminology since there is some interchangeability between the terms and the structure of a phrase may influence the meaning. For instance, "przepuklina przełykowa" or "przepuklina przełyku" ("hiatal hernia" or "hiatus hernia") can also be phrased as "przepuklina wślizgowa przełyku" ("sliding hiatus hernia"). The literal translation of (parenchymal) pulmonary/lung consolidations is: "zgęszczenia (miąższowe) płuc/płucne" but in reports it usually comes in a phrase "zgęszczenia (miąższowe) w płucu prawym" ("consolidations in the left lung"). Extracting information from a report is a difficult task for the model but it is also non-trivial for a referring physician. From a clinical perspective, the automatic generation of structured reports from free texts combines the benefits of both structured reporting and free text, while limiting the drawbacks of a rigidly structured format.

6 Future Work

The results generated by general domain language models are satisfactory, but far from perfect. This is likely motivated by the fact that the word distribution in the general domain and medical corpora is vastly different, which can result in an array of problems in the NLP of clinical texts. In the future, we are planning to train domain-specific language models using a larger corpus of unlabeled reports using methods such as masked language modeling. Such an approach would most definitely improve the model's results.

Table 3: Classification Report for the best model

Class	Precision	Recall	F-score	Support
adenopatia wnęk	0.36	0.44	0.4	9
adenopatia śródpiersia	0.57	0.68	0.62	50
jamy	0.5	0.61	0.55	33
konsolidacje w płucach	0.84	0.87	0.86	62
kostka brukowa	0.57	0.67	0.62	6
liczne guzki	0.77	0.77	0.77	57
matowa szyba	0.96	0.97	0.96	141
niedodma	0.76	0.71	0.74	63
odma opłucnowa	0.91	0.85	0.88	34
odma podskórna	0.84	0.89	0.86	18
odma śródpiersia	0.7	0.78	0.74	9
patologie nadnerczy	0.61	0.73	0.67	15
patologie opłucnej	0.85	0.85	0.85	39
patologie przewodu pokarmowego	0.33	0.57	0.42	7
plaster miodu	1.0	1.0	1.0	26
pogrubienie ścian oskrzeli	0.57	0.68	0.62	19
pojedyncze guzki	0.73	0.78	0.75	95
poszerzenie pnia płucnego lub tt płucnych	0.42	0.55	0.48	20
powiększenie serca	0.88	0.89	0.88	47
przepuklina rozworu przełykowego	0.62	0.63	0.63	49
płyn w jamie opłucnowej	0.82	0.85	0.83	187
płyn w worku osierdziowym	0.95	0.95	0.95	41
radiologiczne podejrzenie covid	0.74	0.82	0.78	17
rozedma	0.88	0.94	0.91	110
rozstrzenia oskrzeli	0.81	0.87	0.84	62
rozwarstwienie aorty	1.0	1.0	1.0	1
rurka intubacyjna/wkłucie	0.85	0.86	0.85	71
tętniak aorty poszerzenie aorty	0.53	0.8	0.64	10
urządzenia kardiologiczne	0.53	0.6	0.56	15
wydzielina w oskrzelach	0.56	0.59	0.57	17
zastój w krążeniu płucnym	0.67	0.67	0.67	9
zatorowość płucna	1.0	1.0	1.0	1
zmiany miażdżycowe aorty	0.77	0.67	0.71	15
zmiany o typie pączkującego drzewa	0.97	0.97	0.97	40
zmiany w kościach	0.83	0.79	0.81	160
zmiany w tarczycy	0.62	0.8	0.7	20
zmiany w tkankach miękkich	0.48	0.63	0.55	19
zmiany włókniste	0.85	0.92	0.88	165
zmiany zapalne/niedodmowo-zapalne	0.82	0.82	0.82	147
zwapnienia w miąższu	0.46	0.35	0.4	17
zwapnienia w naczyniach wieńcowych	0.93	0.96	0.95	28
złamanie żeber	0.96	0.83	0.89	30
micro avg	0.79	0.83	0.81	1981
macro avg	0.73	0.78	0.75	1981
avg	0.8	0.83	0.81	1981

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Sean Bechhofer, M Tamer Özsu, and Ling Liu. 2009. Owl: Web ontology language. In {*Encyclopedia of Database Systems*}. Springer Nature.

David S Channin, Pattanasak Mongkolwat, Vladimir Kleper, Kastubh Sepukar, and Daniel L Rubin. 2010. The caBIG™ annotation and image markup project. *Journal of Digital Imaging*, 23:217–225.

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. 2022. Evaluation of Transfer Learning for Polish with a Text-to-Text Model. *arXiv preprint arXiv:2205.08808*.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020a. Pre-training Polish Transformer-Based Language Models at Scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020b. Pre-training polish transformer-

based language models at scale. In *International Conference on Artificial Intelligence and Soft Computing*, pages 301–314. Springer.

Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts. 2020. RadLex Normalization in Radiology Reports. In *AMIA Annual Symposium Proceedings*, volume 2020, page 338. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. 2000. To err is human: building a safer health system.

European Society of Radiology (ESR). 2018. ESR paper on structured reporting in radiology. *Insights into imaging*, 9(1):1—7.

Lorenzo Faggioni, Francesca Coppola, Riccardo Ferrari, Emanuele Neri, and Daniele Regge. 2017. Usage of structured reporting in radiological practice: results from an Italian online survey. *European Radiology*, 27(5):1934–1943.

Edson Florez, Frédéric Precioso, Michel Riveill, and Romaric Pighetti. 2018. Named Entity Recognition using Neural Networks for Clinical Notes. In *International Workshop on Medication and Adverse Drug Event Detection*, pages 7–15. PMLR.

Dhakshinamoorthy Ganeshan, Phuong-Anh Thi Duong, Linda Probyn, Leon Lenchik, Tatum A McArthur, Michele Retrouvey, Emily H Ghobadi, Stephane L Desouches, David Pastel, and Isaac R Francis. 2018. Structured reporting in radiology. *Academic Radiology*, 25(1):66–73.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Rada Hussein, Uwe Engelmann, Andre Schroeter, and Hans-Peter Meinzer. 2004. DICOM structured reporting: Part 1. Overview and characteristics. *Radiographics*, 24(3):891–896.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv preprint arXiv:2106.14463*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. AWE-CM Vectors: Augmenting Word Embeddings with a Clinical Metathesaurus. *Scientific data*, 3(1):1–9.

Dariusz Kłeczek. 2020. Polbert: Attacking Polish NLP Tasks with Transformers. In *Proceedings of the PolEval 2020 Workshop*, pages 79–88.

Jun Kong, Leixin Zhang, Min Jiang, and Tianshan Liu. 2021. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 116:103737.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Michał Lew, Aleksander Obuchowski, and Monika Kutyła. 2021. Improving Intent Detection Accuracy Through Token Level Labeling. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Laura Liberman and Jennifer H Menell. 2002. Breast imaging reporting and data system (BI-RADS). *Radiologic Clinics*, 40(3):409–430.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of biomedical informatics*, 42(5):923–936.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. Cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready.

Mathias Prokop, Wouter Van Everdingen, Tjalco van Rees Vellinga, Henriëtte Quarles van Ufford, Lauran Stöger, Ludo Beenen, Bram Geurts, Hester Gietema, Jasenko Krdzalic, Cornelia Schaefer-Prokop, et al. 2020. CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19—definition and evaluation. *Radiology*, 296(2):E97–E104.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text Chunking using Transformation-Based Learning. In *Natural Language Processing using Very Large Corpora*, pages 157–176. Springer.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Oswaldo Solarte-Pabón, Orlando Montenegro, Alberto Blazquez-Herranz, Hadi Saputro, Alejandro Rodriguez-González, and Ernestina Menasalvas. 2021. Information extraction from Spanish radiology reports using multilingual BERT. *CLEF eHealth*.

Krzysztof Sopyła and Łukasz Sawaniewski. 2021. Ermlab/politbert: Polish roberta model trained on polish literature, wikipedia, and oscar. the major assumption is that quality text will give a good model.

Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. 2019. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32:554–564.

Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata, Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, et al. 2021. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.

Buzhou Tang, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen. 2019. De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models. In *AMIA Annual Symposium Proceedings*, volume 2019, page 857. American Medical Informatics Association.

Jan van de Kerkhof. 2016. Convolutional Neural Networks for Named Entity Recognition in Images of Documents.

Michał Wojczulis and Dariusz Kłeczek. 2021. papuGaPT2 - Polish GPT2 language model.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Ruoyu Zhang, Pengyu Zhao, Weiyu Guo, Rongyao Wang, and Wenpeng Lu. 2022. Medical Named Entity Recognition Based on Dilated Convolutional Neural Network. *Cognitive Robotics*, 2:13–20.