






BoneBert: A BERT-based Automated Information Extraction System of Radiology Reports for Bone Fracture Detection and Diagnosis

Zhihao Dai¹ , Zhong Li², and Lianghao Han³  

¹ Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

zhihao.dai@warwick.ac.uk

² InterSystem, Eton SL4 6BB, UK

³ Department of Computer Science, Brunel University, Uxbridge UB8 3PH, UK

lianghao.han@brunel.ac.uk

Abstract. Radiologists make the diagnoses of bone fractures through examining X-ray radiographs and document them in radiology reports. Applying information extraction techniques on such radiology reports to retrieve the information of bone fracture diagnosis could yield a source of structured data for medical cohort studies, image labelling and decision support concerning bone fractures. In this study, we proposed an information extraction system of Bone X-ray radiology reports to retrieve the details of bone fracture detection and diagnosis, based on a bio-medically pre-trained Bidirectional Encoder Representations from Transformers (BERT) natural language processing (NLP) model by Google. The model, named as BoneBert, was first trained on annotations automatically generated by a handcrafted rule-based labelling system using a dataset of 6,048 X-ray radiology reports and then fine-tuned on a small set of 4,890 expert annotations. Thus, the model was trained in a “semi-supervised” fashion. We evaluated the performance of the proposed model and compared it with the conventional rule-based labelling system on two typical tasks: Assertion Classification (AC) for bone fracture status detection (positive, negative or uncertainty) and Named Entity Recognition (NER) related to the fracture type, the bone type and location of a fracture occurs. BoneBert outperformed the rule-based system in both tasks, showing great potential for automated information extraction of the detection and diagnosis of bone fracture from radiology reports, such as, the clinical status, type and location of bone fracture, and more related observations.

Keywords: Electronic medical records · Machine learning · Natural language processing · Semisupervised learning

1 Introduction

Radiology reports in the form of Electronic Health Report (EHR) can be a rich source of data for medical cohort studies, image labelling, and decision support after information extraction techniques are applied for automatic annotation and labeling. Despite abundant researches in recent years, existing information extraction systems of radiology reports are developed for either general [2, 19] or Chest X-ray radiology reports [1, 5, 9, 10, 12–14, 20]. Although some of these system can capture the presence or absence of bone fracture, they either do not extract other significant observations [8, 25], such as the uncertainty, type and location of bone fracture, or rely on regular expressions [22–24], which hamper their performance.

In this study, we have developed a BERT-based information extraction system to retrieve the details related to the detection and diagnosis of bone fracture from bone X-ray radiology reports. It is informed by a large-scale biomedical corpus and syntactic structures of sentences. More specifically, we implement the automated information extraction on the clinical status, type and location of bone fracture. The aim of this study is to convert the information for each fracture recorded from the free-text format into a structured one, such as the following template used by radiologists in the UK:

There is (not/possibly) [STATUS] a [TYPE] fracture of the [LOCATION-BONE-PART] part of the [LOCATION-BONE] bone.

The proposed system employed a BERT-based model which was pre-trained on the PubMed abstracts and MIMIC-III. It was first trained with the annotations automatically generated by a conventional handcrafted rule-based labelling system from a dataset of radiology reports, and then fine-tuned with a small set of expert annotations. The conventional rule-based labelling system was developed to alleviate the burden of manual labelling for generating training datasets. Therefore the proposed system adopted a semi-supervised training approach. The whole dataset comprised of 13,712 Bone X-ray radiology reports in which 4,530 sentences and 4,890 mentions of “fracture” were annotated by clinical experts.

We evaluated the performance of the proposed approach and compared it with the conventional rule-based labelling system on two typical tasks: Assertion Classification (AC) for bone fracture status detection (positive, negative and uncertainty) and Named Entity Recognition (NER) related to the type and location of bone fracture.

2 Related Works

2.1 Rule-Based Approaches

NegBio [14] made uncertainty cases explicit and applied two sets of rules, one for negation and the other for uncertainty, on Universal Dependencies (UD) [6] trees of sentences. CheXpert-labeller [10] used a set of rules to automatically

detect the presence of 14 common observations in Chest X-ray radiology reports, and addressed the problem of misclassification of double negation in NegBio by splitting uncertainty detection into pre-negation and post-negation stages.

2.2 Machine Learning Approaches

Earlier machine learning systems employed traditional methods such as Logistic Regression [1, 8], Decision Trees [25], Support Vector Machine (SVM) [8], Random Forests [8], Conditional Markov Model (CMM) as well as Conditional Random Field (CRF) [9]. Recent advancements in Deep Neural Networks (DNNs) lead to the exploration of Recurrent Neural Network (RNN) [5, 21], Convolutional Neural Network (CNN) [17], as well as Transformers [5] such as BERT [7] and XLNet [26] for information extraction.

2.3 Hybrid Approaches

Training machine learning models on labels automatically generated by rule-based systems, to which some works refer as “weak supervision,” reduces the annotation costs and alleviates the burden of manual annotation. CheXbert [20] and Transfer [12], coupling supervised learning to weakly supervised learning, was a pre-trained BERT/LSTM network fine-tuned successively on labels generated by CheXpert labeller/NegBio and manually produced ground-truths.

3 Methodology

3.1 Dataset

Our dataset was a privately-owned collection of 9,435 radiology studies and 13,712 reports in the UK. All the reports have been anonymised. The dataset also contained a total of 4,890 manual annotations performed by two clinical experts. For each report used for manual annotation, the sentences, where mentions of “fracture” exist, were annotated with four categories of information per mention, and they were **assertion**, **type**, **bone** and **bone part**. Assertion was either **positive (P)**, **negative (N)**, **uncertain (U)** or **ignored (I)**. When there were two mentions and both referred to the same fracture in a sentence, the latter was assigned with an “ignored” label. Also, in the same sentence, words other than the keyword itself and describing type, bone and bone part of the fracture were paired with the mention and assigned to the corresponding categories. Each word was only paired with one mention at most.

Our task is, in effect, a combination of two subtasks, AC for the mention and NER for each of other words. Figure 1 presents an example of manual annotations for a sample report in our dataset.

We split the dataset into the training, validation and test sets consecutively, as shown in Table 1. We ensured that there was no overlap of studies across different sets.

ID: 4e402a05-4cfe-43b4-8e88-e1b7752c3a31 CLINICAL INFORMATION: Fracture review ORDER ITEM: XLOLR, XR Tibia and fibula Rt REPORT: <ol style="list-style-type: none"> 1. There is a healing fracture at the neck of the right fibula. 2. Position as shown.
<ul style="list-style-type: none"> – ASSERTION: positive TYPE: healing – BONE: right, fibula BONE PART: neck

Fig. 1. A sample report from the annotated dataset. The annotation is in color green. (Color figure online)

Table 1. Distributions of Reports, Sentences and Mentions (Either Positive (P), Negative (N), Uncertain (U) or Ignored (I)) in Datasets

Set	Reports	Sentences	Mentions				
			Positive	Negative	Uncertain	Ignored	Sum
Training	9,389	28,000	1,155	1,888	262	27	3,332
Validation	1,903	4,000	203	501	36	4	744
Test	2,420	8,509	270	498	43	3	814
Total	13,712	40,509	1,630	2,896	341	32	4,890
Extra	6,048	51,460	4,733	5,192	365	0	10,290

Additionally, we retrieved 6,048 reports related to bone fracture. The extra set contains 10,290 mentions and is annotated by our rule-based model. Notice that the model does not assign the “ignored” assertion label.

3.2 Information Extraction

Figure 2 presents the BERT-based information extraction pipeline proposed in this study. It consists of a pre-trained BERT model as its core component and a conventional rule-based labeller for automatic annotation. The rule-based labeller applies handcrafted rules on syntactic structures of sentences and words matching to generate “weak labels” or so-called “imperfect labels” for the large extra set. As shown in Fig. 2, the proposed BERT-based model is trained in a so-called “semi-supervised” fashion, that is, first on the aforementioned weak labels and later on a small set of ground-truths annotated by clinical experts for fine-tuning. We named our BERT-based system as BoneBert.

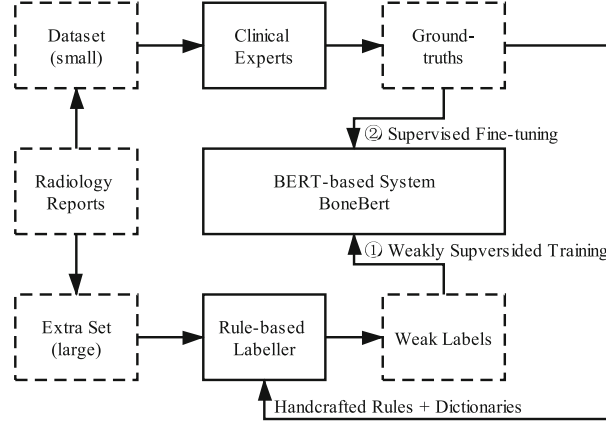


Fig. 2. BERT-based information extraction pipeline.

BERT-based Model. BoneBert was built upon the language model BlueBERT [15], a BERT model pre-trained on the PubMed abstract database and clinical notes MIMIC-III [11] and developed for clinical NLP tasks. We further trained BlueBERT on our dataset of clinical reports, in order to learn Bone X-ray radiology reports containing the information of fracture detection and diagnosis. In this BERT-based model, we converted each mention along with its context into a input-output sequence pair. In the input sequence, the keyword (e.g. fracture) was masked. In the output sequence, we assigned either “POSITIVE”, “NEGATIVE”, “UNCERTAIN”, or “IGNORED” label to the masked token depending on its assertion. For the rest of tokens, we assigned labels in the “type”, “bone”, and “bone part” categories accordingly and empty “O” labels to others.

Rule-Based Labeller. The rule-based labeller was developed to alleviate the burden of large scale manual labelling. The proposed rule-based labeller was based on CheXpert-labeller [10], a rule-based labeller of Chest X-ray radiology reports, and simple words matching for NER.

Figure 3 illustrates the flowchart of the rule-based labeller on bone fracture labelling with a sample sentence extracted from a bone X-ray radiology report. Given a sentence in a report, we first tokenized and extracted mentions of interest. In this study, we looked for any token in the form of “fracture” or “fractures.” We then parsed the sentence to produce a Universal Dependencies (UD) Tree. The tree described the syntactic relations among tokens in the sentence, whereas each arc represented a relation. The head pointed to the dependent of the relation, and the tail the governor.

For the AC task, to detect the presence of bone fracture, we fed the UD Tree into the labeller. Like the CheXpert-labeller, the proposed labeller deployed three sets of rules defined in Semgrep [3], a language for searching dependency relations, of pre-negation, negation, and post-negation consecutively on mentions in the UD Tree to determine the assertion. Mentions matching the pre-negation or post-negation rules were assigned with “uncertain” label, and those matching

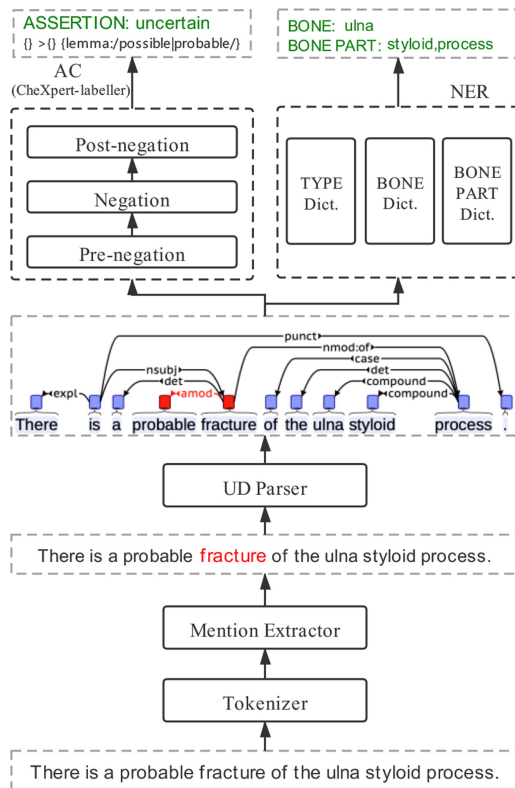


Fig. 3. Rule-based labeller.

the negation rules were assigned with “negative.” Mentions that did not match any rule are considered “positive.”

For the NER task, in order to label fracture type, bone type and bone part where a fracture occurs by the rule-based labeller, we first learned three dictionaries, corresponding respectively to “fracture type,” “bone type,” and “bone part” from the ground-truth labels of the training set. During the training stage, each token with a single NER label was assigned to the dictionary of its corresponding category. Tokens with more than one label were assigned to the dictionary of the most frequent ones only. In addition, we extracted phrases whose roots are “fracture” (type), “bone organ” (bone), and “zone of bone organ” (bone part) from RadLex¹, split them into words, and assigned each word to its corresponding category. During the prediction stage, the system, by default, looks for matches in the sentence where a mention exists. For sentences that have two or more mentions, however, the system limits the matching scope to their respective dependents.

¹ <http://radlex.org>.

3.3 Training and Evaluation

For the rule-based model, we used the “GUM” model² of the Stanford’s Stanza toolkit [16] for tokenisation and the “GENIA+PubMed” model³ of the BLLIP parser [4] for parsing. We converted the resulting trees into Universal Dependencies using the Stanford Dependencies Converter [18].

For the proposed model, BoneBert, we used the base version of BlueBERT as a starting point, which had 12 layers, 12 heads, and 110 million parameters in total. We adapted the original NER code in BlueBERT⁴ for training. Hyperparameters are tuned on the validation set.

4 Experiments

4.1 Assertion Classification

Extra New Rules for Bone X-Ray Radiology Reports. The CheXpert-labeller came with 153 assertion rules across three stages. In our labeller, we expanded the rules bases through manually designing rules for any incorrect prediction by the labeller to account for the style and syntactic differences between our dataset and Chest X-ray reports used for developing the CheXpert-labeller. In the end, the process yielded 19 new rules, of which four were negation, 15 post-negation and no pre-negation. We found that none of the new rules resulted in an incorrect prediction on the validation set.

Table 2. Assertion classification results of the rule-based labellers and the BERT-based systems on the test set. Precision, recall, and F1 score per class as well as overall weighted F1 score are reported in percentile.

Model	Positive			Negative			Uncertain			Overall
	P	R	F1	P	R	F1	P	R	F1	F1
BonePert	80.12	100	88.96	99.58	94.58	97.01	0	0	0	89.19
BonePert+	97.12	100	98.54	99.60	99.80	99.70	100	86.05	92.50	98.93
BlueBERT	99.63	99.26	99.44	100	99.80	99.90	95.56	100	97.73	99.63
BoneBert	99.63	100	99.82	100	99.80	99.90	100	100	100	99.88

Table 2 compares the performance on the test set between the rule-based labeller with the CheXpert’s rule base (we name it **BonePert**) and our expanded rule base (**BonePert+**). Overall, BonePert+ has a weighted average F1 score of 98.93%, a considerable leap from 89.19% by BonePert.

² https://stanfordnlp.github.io/stanza/available_models.html.

³ <https://github.com/BLLIP/bllip-parser/blob/master/MODELS.rst>.

⁴ <https://github.com/ncbi-nlp/bluebert>.

Weak Supervision for BERT. To investigate the effects of weak supervision on BERT’s performance, we fine-tuned a BlueBERT model directly on the training set. We compared it against our BERT-based system (**BoneBert**) trained in the semi-supervised fashion.

Table 2 shows that BlueBERT achieves near-perfect performance on both the “positive” and “negative” classes, with 99.44% and 99.90% F1 scores respectively. The system, however, has a slightly lower F1 score of 97.73% on the “uncertain” class due to the presence of false positives. BoneBert, benefiting from weak supervision by BonePert+, improves its recall in the “positive” class and precision in the “uncertain.” In comparison to BonePert+, BoneBert purges six misclassifications, all of which are uncertain, while making no new ones. Here, we provide an example of an error made by BonePert+ and later corrected by BoneBert.

A repeat lateral projection is advised to confirm or exclude **fracture**.

The “fracture” in the sentence is incorrectly matched to a negation rule and thus classified as “negative” by BonePert+. Both BlueBERT and BoneBert correctly produce an “uncertain” label. As a result, BoneBert has the highest weighted average F1 score of 99.88% thus far.

4.2 Named Entity Recognition

Dictionary Compilation. In the NER subtask, BonePert(+) learned 128 words for “type,” 201 for “bone,” 335 for “bone part” from the training set and RadLex.

Table 3. Named Entity Recognition Results of the rule-based labeller and the BERT-based Systems on the Test Set. Precision, Recall, and F1 score per category as well as average F1 score are reported in percentile.

Model	TYPE			BONE			BONE PART			Average
	P	R	F1	P	R	F1	P	R	F1	F1
BonePert(+)	88.44	95.65	91.91	79.26	91.36	84.88	69.98	97.61	81.52	86.10
BlueBERT	98.29	93.75	95.97	86.44	91.86	89.07	85.56	92.10	88.71	91.24
BoneBert	98.57	93.75	96.10	89.18	92.20	90.67	89.70	93.78	91.70	92.82

Table 3 presents the performance of the BonePert(+) model per category on average. Overall, the model has an average F1 score of 86.10%.

On the test set, 19.71% predicted as “bone” and 25.09% “bone part” in fact do not belong to any category, compared to 11.56% for “type.” Further analysis reveals that most such errors arise from spatial information linked to observations other than bone fractures being constantly mislabelled as “bone” or “bone part” for bone fractures, which explains the relatively lower precision scores in the former two categories.

Weak Supervision for BERT. BlueBERT, informed by large-scale biomedical knowledge, produces 95.97% F1 score for “type”, 89.07% for “bone” and 88.71% for the “bone part.” In contrast, BoneBert, with additional weak supervision from BonePert+, has higher F1 scores of 96.10% for “type”, 91.06% for “bone” and 92.08% for “bone part.” Overall, BlueBERT has an average F1 score of 91.24%, whereas BoneBert has a higher score of 92.95%.

Weak supervision enables BoneBert to capture concepts present in BonePert(+)’s dictionary with greater confidence. Here, we present an example where BlueBERT misses “base” and “fifth” in “base of the fifth metatarsal.”

Left foot- there is a **fracture** through the base of the fifth metatarsal.

Both BonePert(+) and BoneBert correctly identify the words as “bone part” and “bone” respectively since they are included in the corresponding dictionaries.

Since BoneBert is better in separating relevant information from information tied to other observations than BonePert(+), it achieves higher precision scores across all categories. Let’s look at the following sentence in a radiology report as an example,

Left total knee replacement in situ, no **fracture**, dislocation, periprosthetic fracture or evidence of loosening identified.

BonePert(+) mistakes “Left”, “knee,” and “replacement” for the observation of “Left total knee replacement in situ” as the “bone” and “bone part” related to the “fracture,” whereas BoneBert ignores the irrelevant spatial concepts.

The improvements in precision, in comparison to losses in recall, is so significant that the unweighted average F1 score of BoneBert is +6.72% higher than BonePert(+)’s.

5 Discussion

Admittedly, several limitations have arisen in this research. First, our information model covered only the assertion, type and location of bone fractures whereas more information, such as plurality, measurements, and imaging characteristics, could be considered. Second, we developed the expanded rule base for BonePert+ manually. The process was inefficient and hard to scale up if the training set was large. There was no guarantee that the new rules were general enough. Algorithms for learning rules from the training set remains an open task. Last, we considered “fracture” to be the only keyword since it was the most significant indicator of itself. In some rare cases, other words or phrases, say, “bone injury,” might be used to indicate a bone fracture.

6 Conclusion

In this study, we have designed and implemented a BERT-based automatic information extraction system on bone fracture detection and diagnosis from radiology reports, BoneBert. We have also proposed a rule-based labeller for automated annotation to be used for training the proposed model in a weakly supervised fashion.

We evaluated the performance of the model (BoneBert) against the rule-based approach with and without rules specific for fracture diagnosis (named as BonePert+ and BonePert, respectively) on two tasks, namely, AC for detecting the presence of bone fracture and NER for extracting the information about the fracture type, bone type and bone part that a fracture occurs. For the AC task, BoneBert achieved a weighted average F1 score of 99.88%. For the NER task, BoneBert had a higher average F1 score of 92.82%. The proposed approach has shown great potential for automated information extraction on bone fracture detection and diagnosis from bone X-ray radiology reports.

Acknowledgements. This work was supported in part by the SBRI competition: AI supporting early detection and diagnosis in heart failure management.

References

1. Banerjee, I., Chen, M.C., Lungren, M.P., Rubin, D.L.: Radiology report annotation using intelligent word embeddings: applied to multi-institutional chest CT cohort. *J. Biomed. Inform.* **77**, 11–20 (2018). <https://doi.org/10.1016/j.jbi.2017.11.012>, <https://linkinghub.elsevier.com/retrieve/pii/S1532046417302575>
2. Bozkurt, S., Alkim, E., Banerjee, I., Rubin, D.L.: Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *Journal of Digital Imaging* **32**(4), 544–553 (2019). <https://doi.org/10.1007/s10278-019-00237-9>
3. Chambers, N., et al.: Learning alignments and leveraging natural logic. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing - RTE 2007. Association for Computational Linguistics, Morristown, NJ, USA, pp. 165–170 (2007). <https://doi.org/10.3115/1654536.1654570>
4. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL 2005. Association for Computational Linguistics, Morristown, NJ, USA, vol. 1, pp. 173–180 (2005). <https://doi.org/10.3115/1219840.1219862>
5. Datta, S., Si, Y., Rodriguez, L., Shooshan, S.E., Demner-Fushman, D., Roberts, K.: Understanding spatial language in radiology: representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning. *J. Biomed. Inform.* **108**, 103473 (2019). <https://doi.org/10.1016/j.jbi.2020.103473>, <http://arxiv.org/abs/1908.04485>
6. De Marneffe, M.C., et al.: Universal stanford dependencies: a cross-linguistic typology. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pp. 4585–4592 (2014)

7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
8. Grundmeier, R., et al.: Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. *Appl. Clin. Inform.* **7**(4), 1051–1068 (2016). <https://doi.org/10.4338/ACI-2016-08-RA-0129>
9. Hassanpour, S., Langlotz, C.P.: Information extraction from multi-institutional radiology reports. *Artif. Intell. Med.* **66**, 29–39 (2016). <https://doi.org/10.1016/j.artmed.2015.09.007>, <https://linkinghub.elsevier.com/retrieve/pii/S0933365715001244>
10. Irvin, J., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 590–597 (2019). <https://doi.org/10.1609/aaai.v33i01.3301590>, <https://aaai.org/ojs/index.php/AAAI/article/view/3834>
11. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**(1), 160035 (2016). <https://doi.org/10.1038/sdata.2016.35>
12. Liventsev, V., Fedulova, I., Dylov, D.: Deep text prior: weakly supervised learning for assertion classification. In: Tetko, I.V., Kůrková, V., Karpov, P., Theis, F. (eds.) *ICANN 2019. LNCS*, vol. 11731, pp. 243–257. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30493-5_26
13. McDermott, M.B.A., Hsu, T.M.H., Weng, W.H., Ghassemi, M., Szolovits, P.: CheXpert++: approximating the cheXpert labeler for speed, differentiability, and probabilistic output. arXiv preprint [arXiv:2006.15229](https://arxiv.org/abs/2006.15229) (2020)
14. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits Trans. Sci. Proc.* **2017**, 188–196 (2018). <http://www.ncbi.nlm.nih.gov/pubmed/29888070>
15. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 58–65 (2019). <https://doi.org/10.18653/v1/W19-5006>
16. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: a Python natural language processing toolkit for many human languages. arXiv preprint [arXiv:2003.07082](https://arxiv.org/abs/2003.07082) (2020)
17. Santus, E., et al.: Do neural information extraction algorithms generalize across institutions? *JCO Clin. Cancer Inform.* **3**, 1–8 (2019). <https://doi.org/10.1200/CCI.18.00160>
18. Schuster, S., Manning, C.D.: Enhanced English universal dependencies: an improved representation for natural language understanding tasks. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 2371–2378 (2016)
19. Sevenster, M., Buurman, J., Liu, P., Peters, J., Chang, P.: Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. *Appl. Clin. Inform.* **6**(3), 600–610 (2015). <https://doi.org/10.4338/ACI-2014-11-RA-0110>
20. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint [arXiv:2004.09167](https://arxiv.org/abs/2004.09167) (2020)

21. Steinkamp, J.M., Chambers, C., Lalevic, D., Zafar, H.M., Cook, T.S.: Toward complete structured information extraction from radiology reports using machine learning. *J. Digit. Imaging* **32**(4), 554–564 (2019). <https://doi.org/10.1007/s10278-019-00234-y>
22. Tibbo, M.E., et al.: Use of natural language processing tools to identify and classify periprosthetic femur fractures. *J. Arthroplasty* **34**(10), 2216–2219 (2019). <https://doi.org/10.1016/j.arth.2019.07.025>, <https://linkinghub.elsevier.com/retrieve/pii/S0883540319307090>
23. Wang, Y., Mehrabi, S., Sohn, S., Atkinson, E.J., Amin, S., Liu, H.: Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Med. Inform. Decis. Making* **19**(S3), 73 (2019). <https://doi.org/10.1186/s12911-019-0780-5>
24. Wang, Y., et al.: A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Making* **19**(1), 1 (2019). <https://doi.org/10.1186/s12911-018-0723-6>
25. Yadav, K., Sarioglu, E., Smith, M., Choi, H.A.: Automated outcome classification of emergency department computed tomography imaging reports. *Acad. Emerg. Med.* **20**(8), 848–854 (2013). <https://doi.org/10.1111/acem.12174>
26. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* pp. 1–11 (2019)