Application of Deep Learning in Generating Structured Radiology Reports: A Transformer-Based Technique

Seyed Ali Reza Moezzi¹ · Abdulrahman Ghaedi¹ · Mojdeh Rahmanian¹ · Seyyedeh Zahra Mousavi² · Ashkan Sami¹

Received: 16 November 2021 / Revised: 20 June 2022 / Accepted: 27 July 2022 / Published online: 24 August 2022

© The Author(s) under exclusive licence to Society for Imaging Informatics in Medicine 2022

Abstract

Since radiology reports needed for clinical practice and research are written and stored in free-text narrations, extraction of relative information for further analysis is difficult. In these circumstances, natural language processing (NLP) techniques can facilitate automatic information extraction and transformation of free-text formats to structured data. In recent years, deep learning (DL)-based models have been adapted for NLP experiments with promising results. Despite the significant potential of DL models based on artificial neural networks (ANN) and convolutional neural networks (CNN), the models face some limitations to implement in clinical practice. Transformers, another new DL architecture, have been increasingly applied to improve the process. Therefore, in this study, we propose a transformer-based fine-grained named entity recognition (NER) architecture for clinical information extraction. We collected 88 abdominopelvic sonography reports in free-text formats and annotated them based on our developed information schema. The text-to-text transfer transformer model (T5) and Scifive, a pre-trained domain-specific adaptation of the T5 model, were applied for fine-tuning to extract entities and relations and transform the input into a structured format. Our transformer-based model in this study outperformed previously applied approaches such as ANN and CNN models based on ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores of 0.816, 0.668, 0.528, and 0.743, respectively, while providing an interpretable structured report.

Keywords
Structured reporting · Named entity recognition · Relation extraction · Natural language processing · Deep learning · Transformers

Introduction

A radiology report narrative is a text-based interpretation of the acquired images, containing essential information about the patient’s history, imaging procedure, and radiologic findings. However, these reports are mainly written in unstructured, free-text narrations that are hard for computers to interpret and extract clinical information. Besides the complexity and ambiguity of free-format narrations, the variety of medical lexicons used by different radiologists makes the automated analysis more difficult.

Despite these challenges, radiology reports created by different institutions mostly follow the same semantic framework consisting of observations, observation modifiers such as anatomy, quantity, severity, and negation providing substantial opportunity for natural language processing (NLP) techniques to convert these free-text narrations to structured format [1].

Medical NLP experiments mainly consist of two critical steps, named entity recognition and relation extraction. “Named entity recognition (NER)” or “entity extraction” is defined as scanning the unstructured texts to find, label, and classify specific terms into categories such as observation, anatomy, and location. For instance, in the sentence “a simple cortical cyst is seen in the left kidney,” NER can extract the terms “simple,” “cortical cyst,” and “left kidney” and categorize them in the form of “simple: modifier,” “cortical cyst: clinical finding,” and “left kidney: anatomy” without establishing any relations between them. This step can be followed by “relation extraction” which aims to find meaningful relations among the entities identified through NER tasks. Considering the mentioned example, relation extraction methods can establish a relationship among “simple,”
This paper introduces an end-to-end information extraction system that directly utilizes manually generated schemas and requires little to no custom configuration and eliminates the need for additional manual annotation or a heuristic alignment step. We jointly learn to extract, select, and standardize values of interest, enabling the direct generation of a machine-readable data format.

Unlike previous research that commonly requires a large number of labeled instances, our proposed model is trained with a small number of data. We take advantage of transformer-based pre-trained language models which are pre-trained over extensive unlabeled text data using self-supervised learning. Transfer learning allows the reuse of knowledge gained by the model in the source task to perform well in the downstream task. One of the advantages of pre-trained language is that there is no need to train the downstream models from scratch. Once the model is pre-trained on large volumes of unlabeled text data to learn universal language representations of the training data, it can be fine-tuned after adding task-specific layers to be used in various downstream tasks [12].

The main contributions of this paper are as follows:

- Using text-to-text transformers like T5 despite most existing transformer-based clinical NER and relation extraction studies have focused on the Bidirectional Encoder Representations from Transformers (BERT) architecture.
- Fine-tuning a pre-trained model with a small amount of annotated dataset from radiology reports.
- Dealing with a larger number of entity types compared to most NER tasks which usually extract limited entities.
- Generating entirely structured sonography reports interpretable by both physicians and computer systems. For this purpose, we propose ReportQL, a simplified version of GraphQL syntax, which is easy enough for our model to generalize well on its syntaxes while maintaining parsability.

The rest of this paper is organized as follows. The “Related Works” section provides a brief review of related works. The “Structured Reporting Using Transfer Learning” section describes an information model, the corpus, and the annotation scheme. In the “Results and Experience” and “Discussion” sections, we report and discuss the results of our experiments and their limitations. Finally, the “Conclusion” section summarizes and concludes our work.

Related Works

In this section, we have a brief review of literature on structured reporting, strategies used for NER, and relation extraction as the main steps of NLP, and the application of DL for this purpose.
Structured Reporting

Structured reporting was primarily introduced by DICOM in 1999 with generic models of various reports [19]. However, the models did not receive global adoption. In 2007–2008, the Radiological Society of North America (RSNA) created a structured reporting initiative [20]. For this purpose, data mining approaches based on images and clinical texts known as clinical text mining were used [21].

To develop image-based structured reports, Keek et al. and Lambin et al. used “Radiomics” to convert clinical images to quantitative representation for data mining (22, 23). In another approach, Litjens et al. and Lundervold et al. applied deep learning to interpret images and extract relative features [24, 25].

For clinical text mining, many institutions have focused on developing reporting templates such as the disease-specific reporting templates collected in the RadReport or other cloud-based solutions with proprietary template formats such as the Smart Reporting application [26]. In recent years, the conversion of free-text formats to structured data has been advocated. Taira et al. have introduced a natural language processor to automatically structure medical data. For this purpose, NER and relation extraction are also needed [27].

Named Entity Recognition and Relation Extraction

Early applications used for bio-NER such as cTAKES and Metamap matched text phrases with dictionaries and rules. The clinical Text Analysis and Knowledge Extraction System (cTAKES) was created by Mayo Clinic integrating rule-based and ML-based approaches [28]. Metamap, a widely available application created by the National Library of Medicine (NLM), used specific dictionaries to find concepts in the biomedical texts and mapped them to the UMLS metathesaurus [29]. However, rule-based and dictionary-based applications need radiologist expertise to define the rules and templates and create dictionaries manually. Also, due to the large-scale medical terminologies that exist, one dictionary cannot cover all available medical terms leading to missing entities [30, 31]. Recently, these models have been integrated or replaced by machine learning (ML)-based models [32, 33]. In any case, these models also require annotated datasets for the training phase.

As mentioned, the next step of the NLP process after the NER is relation extraction. Three approaches have been introduced for this purpose: co-occurrence approaches [34–37], pattern-based models [38], and ML-aided algorithms [39]. However, to handle complex text mining and relation extraction processes, systems based on a combination of two approaches or ML and DL-based algorithms were created.

Application of DL-Based Approaches

DL-based NLP models including convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), and transformer-based models have outperformed traditional NLP techniques in recent years. Chen et al. created a deep learning convolutional neural network (DL-CNN) model for medical text processing [40]. An automatic text classification method was proposed by Wang et al. in [41]. The results revealed that the CNN model outperformed the rule-based NLP and other ML models in two of three clinical datasets implying that CNN can reveal hidden patterns that are not captured in the rule-based NLP models. RNNs and in particular LSTM have also been used for medical text processing. In a study conducted by Lee et al., LSTM-RNN was applied for the automatic classification of orthopedic images based on the presence of a fracture [42]. Carrodeguas et al. created an LSTM-RNN model to extract follow-up recommendations from radiology reports [43].

Recently, transformer-based models such as the Bidirectional Encoder Representations from Transformers (BERT) models have been applied for text processing. In [44], a BERT-based model named “CheXbert” was created to perform automatic labeling and identify the presence of specific radiologic abnormalities in the chest X-rays. Wood et al. also introduced an attention-based model for automatic labeling of head MRI images (ALARM) regarding the presence and absence of five major neurologic abnormalities [45]. Additionally, transformers and pre-trained language models are used in some other studies for various purposes such as machine translation [13], document generation [46], and text-to-SQL approaches [47].

Overall, although these studies have paved the way for the global adaptation of DL-based models and in particular the transformers in the medical practice, their application in the development of structured radiology reports is less examined.

Structured Reporting Using Transfer Learning

In this section, similar to what has been done in [47], we also formulate the structured reporting problem as a multi-task learning problem by adapting a pre-trained transformer model and translating free-format reports to structured representations.
The development pipeline of the proposed method is depicted in Fig. 1. We first annotated each report using the corresponding schema, to develop our training dataset. After that, a machine-readable representation of the reports is prepared in the form of ReportQL syntax. In addition to the T5 base model, we use Scifive which is pre-trained on a large biomedical corpus. The model was then fine-tuned and tested on the annotated dataset. The following sections go into the details of the suggested technique. We describe how to develop schemas, prepare data, and train tasks that produce structured output.

**Developing Schemas**

During an abdominopelvic ultrasound examination, the ultrasound practitioner should generally examine the anatomic structures. For each organ that is examined in the abdominopelvic ultrasound report, we consider a schema and demonstrate all clinical information that should be evaluated by examining a set of available reports. For feasibility studies, we used only abdominopelvic sonography reports to limit possible information domains. However, information schemas can also be prepared for other parts of the body.

For this purpose, two experts in the field of radiology were asked to extract this information. A separate schema was developed for each organ with all the detailed items for evaluation based on prepared guidelines. Typical reports itemize and describe normal and abnormal findings as well as provide pertinent interpretive comments. Any abnormality is qualified based on its anatomical location, imaging properties, and measurements.

For example, in all reports for the gallbladder, the radiologic findings are expressed in linguistic words such as “stone,” “sludge,” and “distention.” The schema of extracted properties for gallbladder evaluation is shown in Fig. 2. Schemas for other anatomic organs in an abdominopelvic ultrasound examination were created similarly.
Datasets

For our dataset, we used a practical sample of 88 abdominopelvic sonography reports from people who had made a consensus that their medical information could be used for medical research. We eliminated all the personal and gender-specific information. The reports included adult patients of all ages and genders written by various attending radiologists and resident radiologists using their personal reporting templates without a unified format and headings and were written in prose. Our corpus contains 5619 words and 292 unique sentences describing anatomic organs.

The reports were divided into two categories: training and testing sets (80–20%). The train set was manually annotated using the schemas outlined above and was utilized to train our base NER model. Data annotation was performed with two annotators in two steps. In the first step, the same reports were assigned to each annotator. Each section of the report relating to an anatomic structure corresponds to the appropriate schema of that organ. For each report, the annotators read the reports and assigned a value to each corresponding property in the schema (see Fig. 3). The annotation follows the format of ReportQL syntax. ReportQL consists of a hierarchy of entities that represent anatomic organs and their properties, enclosed in curly brackets as depicted in Fig. 3. This syntax uses minimal control characters to facilitate the generalization of this model while remaining highly readable by the physicians. The second step is cross-validation, in which one annotator cross-checks the annotated data from another annotator. In the case of disagreements, the third domain-specific expert has resolved the disagreements.

Transformer Models

One of the main challenges of our research was the limitation of training data. To address this problem, we came up with the idea of using transfer learning. A transfer learning model typically consists of two stages: first, unsupervised training on a general-purpose corpus and then second, supervised training for a specific task known as the downstream task. In the first stage, the transformer-based model is optimized with a large amount of unannotated text data

---

**Fig. 2** An example of the information schema designed as a hierarchical diagram summarizing the information that can be reported for the gallbladder organ in sonography reports including clinical statements or facts and modifiers such as location or quantity.
through language modeling techniques that are independent of specific downstream NLP tasks. In the second stage, the pre-trained transformer-based model is fine-tuned for a particular NLP task using a supervised approach.

We leverage pre-trained language models such as BERT in NLP tasks. BERT is not a unified transfer learning method because BERT-style models can only produce a single prediction for a given input. These models are simply not designed for text generation tasks such as question-answering or summarization. The text-to-text transfer transformer (T5) model overcomes this limitation by outputting a string of text for each input, allowing for question-answering, summarization, and other tasks where a single output is generally insufficient. In our work, we use the T5 model and Scifive, a pre-trained domain-specific adaptation of the T5 model for tasks related to biomedical literature, which is very important for the learning process.

- **T5**: The text-to-text transfer transformer (T5) trained on the “Colossal Clean Crawled Corpus” (C4) uses a basic encoder-decoder transformer architecture as originally proposed by Vaswani et al. [13]. Each encoder block consists of a self-attention layer and a feedforward neural network. Each decoder block consists of a self-attention layer, an encoder-decoder attention layer, and a feedforward neural network. T5 is pre-trained on masked language modeling with a learning objective called “span masking,” where spans of text are replaced with a mask token randomly and the model is trained to predict the real masked spans.

- The T5 has five different model sizes—small (60 million parameters), base (220 million), large (770 million), 3B, and 11B.

- **Scifive**: Like T5, Scifive follows a text-to-text encoder-decoder architecture that transforms an input sequence into an output sequence. Scifive is a domain-specific adaptation of the T5 model pre-trained on a large biomedical corpus. Two different corpora of biomedical language are utilized to generalize the model within the biomedical domain [14]:
  - PubMed Abstract: It contains over 32 million citations and abstracts of biomedical literature.
  - PubMed Central: This database is a corpus of free full-text articles in the domain of biomedical and life sciences.

**Input Representation**

As mentioned, we formulate our task as a text-to-SQL problem. Since the fine-grained structure is a complicated task and hard to reach a consensus on specific terms, to remove the burden of this, we provided a separate schema to our model which acts as the context for our model as explained in the “Developing Schemas” section. This is fairly similar to a recent text-to-SQL architecture where the schema of the tables is concatenated along with input. From our formulation, each input includes the schemas and a report document.

We formulate our schema as:

\[
< t > \quad o_1 \{ c_{11}, \ldots, c_{1j} \} \; < \text{hi} >, \; < t > \; o_2 \\
\{ c_{21}, \ldots, c_{2j} \} \; < \text{hi} >, \ldots, \; < t > \; o_3 \{ c_{i1}, \ldots, c_{ij} \} \; < \text{hi} >
\]

where \( o_i \) represents the name of the organ and \( c_{ij} \) is \( j \)th property for the \( i \)th anatomic organ. We added special tokens \( < t > \) and \( < \text{hi} > \) to separate definitions of organs. We denote schema by \( s \) from now on. Finally, our inputs are formed as:

\[
[PAD] r : r_s : s_i [EOS]
\]

where \( [PAD] \) and \( [EOS] \) are the T5 start and end-of-sentence token, respectively, and \( r_s \) and \( s_i \) are the report and schema pair. Our inputs are then tokenized by the T5 tokenizer and passed to the model for pre-training.
The scarcity of datasets in medical areas, unlike general tasks like question-answering and summarization, may lead to poor performance of the DL models. To mitigate this issue, we used an in-domain T5 model pre-trained on biomedical literature, Scifive [14], to enhance its generalization power.

**Pre-training and Fine-tuning of Transformer Models**

The training process of our model was done in two stages: (i) pre-training and (ii) fine-tuning. For the first stage, the pre-trained model on domain-specific data was reused. Hyperparameters were the same in our experiments. For fine-tuning and evaluating the proposed model, we randomly separated 80% of reports as the training set and 20% as the test set. Models were trained until validation loss increased.

In order to expand the training dataset, we divided the reports into segments based on the anatomic organs described. Then, we randomly sampled combinations of these segments and combined them with the corresponding schemas as input. The total number of synthetic reports produced was 5760. As a result, with a relatively modest number of radiological reports, we were able to generate an extensive training set.

The performance of our models was evaluated using ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L) in addition to SacreBLEU and BLEU scores. Another metric utilized was exact-match, considering the number of key value pairs matching their reciprocal targets. Here, the model should learn not only the granularity of input but also our ReportQL syntax that the output should be structured with.

**Results and Experience**

**Annotation Performance**

Examples of the input free-text reports with the annotated structured results are shown in Table 1. Each input report consists of the description of approximately 17 abdominopelvic organs with an average of 22 fine-grained properties (min: 8, max: 35) in the free-text format which is annotated to extract the mentioned entities and relations and designed in the form of structured data containing the entities related to each organ in the brackets. As illustrated in Table 1, some rare findings such as the presence of horseshoe kidney or some rarely described features of a lesion are missed in the final report that can be solved by including more abnormal reports.

**Evaluation Metrics**

We evaluated the performance of our models using ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L) besides the SacreBLEU and BLEU scores. The ROUGE score (Recall-Oriented Understudy for Gisting Evaluation) measures the recall. It is commonly used in machine translation tasks and for evaluating the quality of the generated text. The number of matching “n-grams” between our model-generated text and a reference text is measured by ROUGE-N. The longest common subsequence (LCS) between our model output and the reference is measured by ROUGE-L [48]. The BLEU score (bilingual evaluation understudy) evaluates the quality of machine-translated text from one natural language to another. BLEU uses modified n-gram precision to compare the candidate translation against ground truth [49].

The fine-tuned T5-base model has achieved a ROUGE-1 score of 0.8151 with subsequent scores of 0.6723 and 0.6957 for ROUGE-2 and ROUGE-L metrics, respectively. The fine-tuned Scifive model also achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.8164, 0.6688, and 0.5280, respectively.

Regarding BLEU scores, the fine-tuned T5-base model has achieved the same SacreBLEU and BLEU scores of 0.7416, respectively. The Scifive-base model has shown the same performance with the same SacreBLEU and BLEU scores of 0.7438, respectively, with a brevity penalty and length ratio of 1.5 and 1.037. Tables 2 and 3 have summarized the mentioned metrics.

Furthermore, we have tested the model with an exact-match score using 10 unstructured abdominopelvic sonography reports obtained from another clinical center apart from the used dataset. Reports were also annotated manually by the annotators to compare. The resulting structured reports were evaluated to measure “precision” and “recall” scores. The mean precision score was 0.892 with a recall score of 0.818, which revealed the acceptable performance of the model despite using another set of unstructured reports with a different format.

**Discussion**

In this study, we proposed a transfer learning model to annotate free-text radiology reports and provide structured texts interpretable by both physicians and data storage software. The model not only achieved BLEU and ROUGE-L scores of 0.7438 and 0.528, but also it can extract entities and relations, identify abbreviations and negation, provide a clinically interpretable structured report, and give the operator freedom on which organs are in the structured report.

Our proposed model has shown significant strengths compared with the previous models. First, it can identify relations of the used entities in the report besides the entity extraction. Considering the sentence “few stones up to 7 mm are seen in the upper pole of the left kidney,” our model can extract medical entities of “few,” “stones,” “up to 7 mm,”
# Table 1 Examples of the free-text abdominopelvic sonography reports with the final structured reports received from our model.

| Free text abdominopelvic sonography report                                                                 | Equivalent structured abdominopelvic sonography report received from our model                                                                 |
|----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| 1 Liver is normal in size and with normal parenchymal echogenicity with no sign of SOL, or bile duct dilatation. | 1 liver (size [normal] echogenicity [normal] lesion [no] bile duct dilatation [no]) GB [seen [no]] | spleen (size [enlarged] echogenicity [normal] lesion [no]) pancreas [seen [no]] | left kidney (size [normal] cortical parenchymal echogenicity [normal] stone [no] stasis [no] perinephric collection [no]) | right kidney (size [normal] cortical parenchymal echogenicity [normal] stone [no] stasis [no] perinephric collection [no]) | bladder (stone [no] wall thickening [no]) prostate (seen [yes] size [normal] echogenicity [normal] lesion [no]) |
| GB is not seen due to previous resection. Spleen is enlarged with normal parenchymal echogenicity with no SOL. | GB (seen [no]) | pancreas (seen [no]) | left kidney (size [normal] cortical parenchymal echogenicity [normal] stasis [no] perinephric collection [no]) | right kidney (size [normal] cortical parenchymal echogenicity [normal] stasis [no] perinephric collection [no]) | bladder (stone [no] wall thickening [no]) prostate (seen [yes] size [normal] echogenicity [normal] lesion [no]) |
| Pancreas could not be evaluated due to severe gas shadow. Both kidneys are normal in size with normal cortical parenchymal echogenicity with no sign of stone, stasis or perinephric collection. | pancreas (seen [no]) | left kidney (size [normal] cortical parenchymal echogenicity [increased] stone [no] stasis severity [no] perinephric collection [no]) | cortical cyst (yes) cortical cyst quantity [two] cortical cyst position [upper pole]) | right kidney (size [normal] cortical parenchymal echogenicity [increased] stone [no] stasis [no] perinephric collection [no]) | bladder (distension [mild] stone [no] wall thickening [no]) abdominopelvic cavity (free fluid [moderate]) |
| Urinary bladder is normal in size and echogenicity with no sign of SOL. Moderate free fluid is seen in the abdominopelvic cavity at present time. | urinary bladder (seen [no]) | urinary bladder (seen [no]) | urinary bladder (seen [no]) | urinary bladder (seen [no]) | urinary bladder (seen [no]) |
| 2 Liver is normal in size, with normal echogenicity with no sign of space occupying lesion. GB is semi distended with two stones up to 8mm in size with a rim of pericholecystic fluid and positive murphy sign. CBD is normal. Spleen is moderately enlarged with normal parenchymal echo with no S.O.L. Pancreas cannot be evaluated due to severe gas shadow. RT. and L.T. kidneys are normal in size with increase parenchymal echogenicity with no sign of stone, stasis or perinephric collection with two cortical cysts in the upper pole of left kidney. Urinary bladder is mildly distended. Moderate free fluid is seen in the abdominopelvic cavity at present time. | liver (size [normal] echogenicity [normal] lesion [no] bile duct dilatation [no]) GB [distension [semi] stone [yes] stone quantity [two] stone size [up to 8 mm] pericholecystic fluid [yes] murphy sign [yes]] | CBD (seen [yes]) | spleen (size [enlarged] echogenicity [normal] lesion [no]) pancreas (seen [no]) | left kidney (size [normal] cortical parenchymal echogenicity [increased] stone [no] stasis severity [no] perinephric collection [no]) | cortical cyst (yes) cortical cyst quantity [two] cortical cyst position [upper pole]) | right kidney (size [normal] cortical parenchymal echogenicity [increased] stone [no] stasis [no] perinephric collection [no]) | bladder (distension [mild] stone [no] wall thickening [no]) abdominopelvic cavity (free fluid [moderate]) |
| 3 Liver is small in size, with increased echogenicity. Multiple hyperechoic cystic lesions with smaller daughter lesions that some of them are partially calcified are seen. CBD is mildly dilated with maximum diameter of 7 mm in middle segment but CBD stone is not seen. GB is contracted and evaluation is impossible. Spleen is enlarged in size with normal parenchymal echogenicity with no sign of solid or cystic lesion. Visualized parts of pancreas are unremarkable. Right kidney is small with increased cortical parenchymal echogenicity with no sign of stone, stasis or perinephric collection. Left kidney is not seen due to previous nephrectomy. Urinary bladder is normal, with no stone or wall thickening. No free fluid is seen in abdominopelvic cavity. | liver (size [small] echogenicity [increased] lesion [yes] lesion quantity [multiple] lesion echogenicity [hyperechoic] lesion component [cystic]) | CBD (dilatation [yes] size [7 mm] position [middle segment] stone [no]) GB (distension [contracted]) | spleen (size [enlarged] echogenicity [normal] lesion [no]) pancreas (seen [yes]) | right kidney (size [small] cortical parenchymal echogenicity [increased] stone [no] stasis severity [no] perinephric collection [no]) | left kidney (size [normal] cortical parenchymal echogenicity [increased] stone [no] stasis [no] perinephric collection [no]) | left kidney (seen [no]) | bladder (stone [no] wall thickening [no]) abdominopelvic cavity (free fluid [no]) |
| 4 Liver is normal in size and with normal parenchymal echogenicity with no sign of SOL, or bile duct dilatation. GB is semi distended with no stone. Spleen has normal size with normal parenchymal echogenicity with no SOL. Both kidneys are small in size with increased cortical parenchymal echogenicity connecting to each other from the lower poles in favor of horseshoe kidney. No sign of stone or perinephric collection is seen. Moderate stasis is seen bilaterally. Urinary bladder is normal, with no stone or wall thickening. Uterus and both ovaries show atrophic change. | liver (size [normal] echogenicity [normal] lesion [no] bile duct dilatation [no]) GB (distension [semi] stone [no]) spleen (size [normal] echogenicity [normal] lesion [no]) | left kidney (size [small] cortical parenchymal echogenicity [increased] stone [no] stasis severity [yes] stasis severity [moderate] perinephric collection [no]) | right kidney (size [small] cortical parenchymal echogenicity [increased] stone [no] stasis [yes] stasis severity [moderate] perinephric collection [no]) | bladder (stone [no] wall thickening [no]) uterus (atrophic change [yes]) | right ovary (atrophic change [yes]) left ovary (atrophic change [yes])
and "upper pole of the left kidney" and then define relations of these entities, and provide the result of "stones {quantity {few}, size {up to 7 mm}, and location {upper pole of the left kidney}}" with identified entities and relations. In contrast, most previous studies have focused only on entity extraction (NER) to answer the presence of a specific finding.

Second, after extracting entities and their relations, the model can use these entities to provide a clinically interpretable structured report that can be used as an alternative to free-text narratives in clinical practice (Table 1). As mentioned, most of the available NLP experiments have focused on entity extraction that aims to answer a specific question, such as the presence of a finding or follow-up indication; however, scant studies have provided an information model capable of transforming reports to structured formats. For instance, a previously DL-based model published by Huang et al. can identify the normality or abnormality of a report; however, the model cannot design the abnormal found entities in the form of a structured text to present, limiting its clinical applicability [50].

Third, heterogeneity of the grammatical formats and terminologies used by different radiologists has made the development of encoded structured reports cumbersome. For example, reports may provide some normal or insignificant findings or normal variations of observations instead of a clear normality statement that should not be interpreted as an abnormal feature. Also, negation detection, heterogeneous sentences used to report abnormalities, synonyms, abbreviations, not using common standardized medical terminologies, etc. challenge medical text processing. However, our model has shown promising results in these aspects, facilitating the development of structured reports. The model has demonstrated significant performance in synonym detection. As an example, all the phrases “dilated bile ducts,” “biliary tree dilatation,” and “intrahepatic biliary duct dilatation” were correctly detected by the model as synonyms. Commonly used medical abbreviations such as “CBD” (common bile duct), “GB” (gallbladder), and “SOL” (space-occupying lesion) were also correctly diagnosed. The model has also shown accurate performance in negation detection and, in particular, in grammatically complex sentences. Besides, in sentences where the same findings are reported for two organs, such as “both kidneys” or “both ovaries,” the model can differentiate and detect both right and left organs separately and mention all the common findings for both of them.

Next, flexibility and lack of word count limitation can also make this model more applicable. The operator can determine which organs will be included in the structured format. For instance, if we aim to collect and analyze characteristics of the liver organ in all the population of an area or patients referring to a specific center, the model can be modified to extract only the information related to the liver organ in all the reports and present them in a structured format to use in the analysis. On the other hand, our model has no limitation regarding the number of words and sentences used in the input reports compared with previously mentioned models. For example, the previously mentioned BERT-based model (cheXbert) can receive a maximum of 512 tokens as the input, limiting the analysis of larger medical texts [44].

From the modeling perspective, it seems that transformer-based models outperformed RNN and CNN while requiring less time to train the model and can be generalized to a wide range of tasks [13]. Regarding the time complexity, our model has been trained for 9 h and 42 m while having an inference time of 5 s for conversion of the input free-text medical reports to a structured output. Besides, the

| Dataset       | Pre-trained models | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU   | SacreBLEU |
|---------------|--------------------|---------|---------|---------|--------|------------|
| Original      | Scifive-base       | 0.732   | 0.633   | 0.635   | 0.727  | 0.738      |
| Synthetic     | Scifive-base       | 0.816   | 0.668   | 0.528   | 0.743  | 0.743      |
| Synthetic     | T5-base            | 0.795   | 0.672   | 0.519   | 0.741  | 0.741      |

| Dataset       | Pre-trained models | Score  | Brevity penalty | Length ratio | Precision          |
|---------------|--------------------|--------|-----------------|--------------|-------------------|
| Original      | Scifive-base       | 0.7382 | 1.0             | 1.010        | [0.84, 0.76, 0.69, 0.62] |
| Synthetic     | Scifive-base       | 0.7438 | 1.0             | 1.037        | [0.85, 0.78, 0.71, 0.63] |
| Synthetic     | T5-base            | 0.7416 | 1.0             | 1.054        | [0.83, 0.77, 0.71, 0.65] |
acceptable performance of our model despite using a small corpus of abdominopelvic sonography reports as the train set compared with the other recent studies would highlight the strengths of the model. This feature is markedly important since requiring large samples of the reports as the input data to train the DL-based models can limit their application. In addition, the model performance was also evaluated using BLEU and ROUGE-L scores. Regarding the BLEU score, the model has achieved a score of 0.7438. The ROUGE-1 and ROUGE-L scores were also measured to be 0.8164 and 0.5280 using the fine-tuned Scifive model. Considering these metrics, using in-domain pre-trained models such as the Sci-five model has outperformed previous similar models [51]. The model also has some limitations:

1. It was only trained based on the abdominopelvic sonography reports. Testing the model performance with reports of other imaging modalities such as CT scans and MRI images related to the chest, brain, and other organs should be performed in the following steps.
2. Although the model has shown perfect performance despite the relatively small corpus of reports used as the train set, including more reports can potentially improve the performance of the model.
3. Evaluation of the model as presented in Table 1 reveals that some rare phrases such as the description of a rare finding (e.g., the horseshoe kidney) or some suggestion statements are missed in the structured report. Including more reports and, in particular, reports containing rare findings can solve this problem.

**Conclusion**

Considering the increasing application of transformer-based NLP models in recent years, using a text-to-text transfer transformer model (T5) and a pre-trained domain-specific adaptation of T5 (Scifive) in this study has shown accepted performance for NER of radiology reports and translation of free-text records to the structured data. The model would also handle annotation of other imaging modalities and also pathological and biological reports with minimal changes in the model schema.

**Acknowledgements** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank Dr. Sadra Valiee (Shiraz University of Medical Sciences) and Dr. Mohammadjavad Gholamzadeh (Shiraz University of Medical Sciences) for their contribution in providing the information schemas.

**Author Contribution** All authors contributed to the study conception and design, and read and approved the final manuscript.

**Data availability** All data generated or analyzed during this study are available from the corresponding author on reasonable request.

**Declarations**

**Ethics Approval** All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study did not contain any experiments performed on humans or animals. To respect the confidentiality of the patients, all the personal data reported in the sonography reports such as the name, age, and clinical history were deleted from the acquired texts before the processing step.

**Consent to Participate** Non-applicable; the study did not contain any experiments performed on humans or animals.

**Consent for Publication** Non-applicable; the study did not contain any experiments performed on humans or animals.

**Competing Interests** The authors declare no competing interests.

**References**

1. Hassanpour S, Langlotz CP: Information extraction from multi-institutional radiology reports. Artif Intell Med 66:29–39; 2016
2. Perera N, Dehmer M, Emmert-Streib F: Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Front Cell Dev Biol 8: 2020
3. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS: Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. J Digit Imaging 32(4):554–64; 2019
4. Sorin V, Barash Y, Konen E, Klang E: Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review. J Am Coll Radiol 17(5):639–48; 2020
5. Monshi MMA, Poon J, Chung V: Deep learning in generating radiology reports: A survey. Artif Intell Med 106; 2020
6. Pons E, Braun LM, Hunink MM, Kors JA: Natural language processing in radiology: a systematic review. Radiology 279(2):329-43; 2016
7. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al.: Natural language processing technologies in radiology research and clinical applications. Radiographics 36(1):176-91; 2016
8. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al.: Huggingface’s transformers: State-of-the-art natural language processing. arXiv Prepr arXiv191003771; 2019
9. Devlin J, Chang MW, Lee K, Toutanova K: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805; 2018
10. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv Prepr arXiv190911942; 2019
11. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv Prepr arXiv190711692; 2019
12. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21:1–67; 2020
13. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al.: Attention Is All You Need. Adv Neural Inf Process Syst 5998–6008; 2017
14. Phan LN, Anibal JT, Tran H, Chanana S, Bahadrogulu E, Peltokian A, Altan-Bonnet G. SciFiive: a text-to-text transformer model for biomedical literature. arXiv preprint arXiv: 2106.03598. 2021 May 28.

15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–40; 2020

16. Si Y, Wang J, Xu H, Roberts K: Enhancing clinical concept extraction with contextual embeddings. J Am Med Informatics Assoc 26(11):1297–304; 2019

17. Peng Y, Yan S, La Z: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv Preprint arXiv:1906.05474; 2019

18. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323; 2019

19. European Society of Radiology (ESR) communications@ myesr.org: ESR paper on structured reporting in radiology. Insights into imaging 9:1-7; 2018

20. Langlotz CP: RadLex: A new method for indexing online educational materials. Radiographics 26(6):1595–7; 2006

21. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al.: Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdiscip Rev Comput Stat 13(6); 2021

22. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al.: Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology 14(12):749-62; 2017

23. Keek SA, Leijenaar RT, Jochems A, Woodruff HC: A review on radiomics and the future of theranostics for patient selection in precision medicine. Br J Radiol 91(1091); 2018

24. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al.: A survey on deep learning in medical image analysis. Med Image Anal 42:60–88; 2017

25. Lundervold AS, Lundervold A: An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 29(2):102–27; 2019

26. Pinto dos Santos D, Baeßler B: Big data, artificial intelligence, and structured reporting. Eur Radiol Exp 2(1); 2018

27. Taira RK, Soderland SG, Jakobovits RM: Automatic structuring of radiology free-text reports. Radiographics 21(1):237–45; 2001

28. Savova GK, Masanz JJ, Ogren P V., Zheng J, Sohn S, Kiper-Schuler KC, et al.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. J Am Med Informatics Assoc 17(5):507–13; 2010

29. Aronson AR, Lang FM: An overview of MetaMap: Historical perspective and recent advances. J Am Med Informatics Assoc 17(3):229–36; 2010

30. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G: Data processing and text mining technologies on electronic medical records: A review. J Healthc Eng; 2018

31. Rebholz-Schuhmann D, Iossifidis V, Arregui M, Gaudan S, Rietveld M, Stoeber P. EBiMed - Text crunching to gather facts for proteins from Medline. Bioinformatics 23(2); 2007

32. Pennington J, Socher R, Manning CD: GloVe: Global vectors for word representation. EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf 1532–43; 2014

33. Hoffmann R, Valencia A: Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21(suppl_2):ii252–8; 2005

34. Garten Y, Altman RB: PharmDxpresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics 10 Suppl 2; 2009

35. Hakenberg J: Mining Relations from the Biomedical Literature 179; 2009

36. Muzaffar AW, Azam F, Qamar U: A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set. Comput Math Methods Med; 2015

37. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al.: Deep learning to classify radiology free-text reports. Radiology 286(3):845–52; 2018

38. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EL, et al.: A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 19(1); 2019

39. Lee C, Kim Y, Kim YS, Jang J: Automatic disease annotation from radiology reports using artificial intelligence implemented by a recurrent neural network. Am J Roentgenol 212(4):734–40; 2019

40. Carrodgeugas E, Laczon R, Swanson W, Khorasani R: Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports. J Am Coll Radiol 16(3):336–43; 2019

41. Smit A, Jain S, Rajpurkar P, Pakeek A, Ng AY, Lungren MP: CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167; 2020

42. Wood DA, Lynch J, Kafiabadi S, Guilehm E, Al Busaidi A, Montvila A, et al.: Townend M, Kiik M. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). InMedical Imaging with Deep Learning 811–826; 2020

43. Liu PJ, Saleh M, Pot E, Goodrich B, Sepassi R, Kaisler L, Shazeer N: Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198; 2018

44. Lyu Q, Chakrabarti K, Hatli S, Kundu S, Zhang J, Chen Z: Hybrid ranking network for text-to-sql. arXiv preprint arXiv:2008.04759; 2020

45. Lin C. Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74–81).

46. Papineni K., Roukos S., Ward T., Zhu W. J.: Bleu: a method for evaluating machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics: 311–318; 2002

47. Huang X, Fang Y, Lu M, Yao Y, Li M: An Annotation Model on former model for biomedical literature. arXiv preprint arXiv:2005.03598. 2021 May 28.

48. Pennington J, Socher R, Manning CD: GloVe: Global vectors for word representation. EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf 1532–43; 2014

49. Hoffmann R, Valencia A: Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21(suppl_2):ii252–8; 2005

50. Garten Y, Altman RB: PharmDxpresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics 10 Suppl 2; 2009

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.