Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations

Francesco Dalla Serra[1,2] **William Clackett**[1] **Chaoyang Wang**[1]
Hamish MacKinnon[1] **Fani Deligianni**[2] **Jeffrey Dalton**[2] **Alison Q O'Neil**[1,3]

[1]Canon Medical Research Europe, Edinburgh, United Kingdom
[2]University of Glasgow, Glasgow, United Kingdom
[3]University of Edinburgh, Edinburgh, United Kingdom
francesco.dallaserra@mre.medical.canon

Abstract

Automated reporting has the potential to assist radiologists with the time-consuming procedure of generating text radiology reports. Most existing approaches generate the report directly from the radiology image, however we observe that the resulting reports exhibit realistic style but lack clinical accuracy. Therefore, we propose a two-step pipeline that subdivides the problem into *factual triple extraction* followed by *free-text report generation*. The first step comprises supervised extraction of clinically relevant structured information from the image, expressed as triples of the form (*entity1, relation, entity2*). In the second step, these triples are input to condition the generation of the radiology report. In particular, we focus our work on Chest X-Ray (CXR) radiology report generation. The proposed framework shows state-of-the-art results on the MIMIC-CXR dataset according to most of the standard text generation metrics that we employ (BLEU, METEOR, ROUGE) and to clinical accuracy metrics (recall, precision and F1 assessed using the CheXpert labeler), also giving a 23% reduction in the total number of errors and a 29% reduction in critical clinical errors as assessed by expert human evaluation. In future, this solution can easily integrate more advanced model architectures – to both improve the triple extraction and the report generation – and can be applied to other complex image captioning tasks, such as those found in the medical domain.

1 Introduction

Chest X-Ray (CXR) studies are among the most frequent radiology studies undertaken in healthcare (NHS England and NHS improvement, 2022). Each CXR is accompanied by a text report written by a radiologist or trained radiographer which describes the findings within the study. Unfortunately, CXR reports are subject to delays, often due to institutional factors, which can result in adverse patient outcomes (Care Quality Commission, 2018). A possible solution to improve the radiology workflow, and to facilitate timely delivery of accurate reports, is to automate the generation of text reports. However, generating clinically accurate radiology reports is a challenging task.

The task of generating a textual description for an image is referred to as image captioning, and recent methods have often adopted encoder-decoder architectures, in which the image embeddings are computed using Convolutional Neural Networks (CNNs) (*e.g.,* He et al., 2016) and the text is generated using Recurrent Neural Networks (RNNs) (*e.g.,* Hochreiter and Schmidhuber, 1997 and Cho et al., 2014), or, more recently, using Transformer-based architectures (Vaswani et al., 2017). Such architectures have been proposed to perform automated report generation in the medical domain, with some custom modules introduced for this specific task. For instance, some recent works in CXR report generation have introduced relational memory modules (Chen et al., 2020) to allow the model to memorise information from previous generation, and cross-modal memory modules (Chen et al., 2021; Qin and Song, 2022) to encourage alignment between visual and textual information. Another line of work has explored ways to inject external knowledge into the model (Liu et al., 2021b; Yang et al., 2022), based on pre-constructed knowledge graphs or by retrieving other similar reports within the dataset. The above methods all attempt to generate the radiology report directly from the image, using only supervision with a standard cross-entropy loss of the generated text compared to the target text, which will reward verbatim replication of the target text (style), whilst not emphasising accurate reporting of the clinically important findings (content). This concern was partially treated by intro-

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 615–624
November 20–23, 2022.

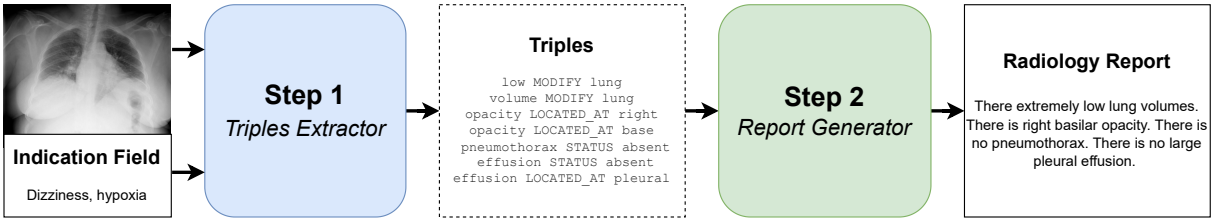

Figure 1: Illustration of the proposed two-step pipeline. **Step 1** – a triples extractor is implemented to extract a set of triples associated with each CXR scan. **Step 2** – a report generator is implemented to generate a radiology report, based on the extracted triples. The CXR image and report shown in this example are both taken from the IU-Xray dataset (Demner-Fushman et al., 2016), while the triples are extracted as described in Section 2.1.

ducing classification of the the findings and pathologies that are present in the image (Alfarghaly et al., 2021), as an auxiliary task. However, in this approach there is no direct link between the classification and reporting outputs, and the transfer of information relies on multi-tasking functioning effectively. Further, this approach does not consider the relations between different classes. Overall, there is a limited effect on the generation process.

We focus our work on improving the clinical utility of the generated reports, by introducing an intermediate step to the generation process. It consists of extracting, from a CXR image, factual information in a structured format, expressed in the form of triples (entity1, relation, entity2). We further categorise the entities and relations according to a clinical schema, in order to remove heterogeneity of expression. This is particularly relevant in the field of radiology, where radiologists can express similar clinical concepts using different phrases i.e. the following phrases all relate to the same clinical concept of edema: "pulmonary oedema", "cardiac decompensation", "fluid overload" and "evidence of acute heart failure". We adopt RadGraph (Jain et al., 2021) to extract four predefined clinically relevant relations (*Suggestive of*, *Located at*, *Modify* and *Status*), and we map medical entities to medical concepts (*e.g.,* "fluid overload" to *«edema»*) according to a scheme devised by a junior physician. Our two-step pipeline is shown in Figure 1, where the first step consists of the triples extraction process which aims at extracting factual information from a CXR image, and the second step corresponds to report generation which uses the image as input alongside (i.e. conditioned by) the extracted triples.

To the best of our knowledge, only Li et al. (2022) have very recently proposed a similar approach for automatic generation of ophthalmic reports. In their work, they show an improvement by extracting, from an ophthalmic image, entities and relations (they consider the extracted triples to represent a latent clinical graph), and injecting them to the text generation process. This varies from our work in three aspects: the definition and generation of triples, the model architecture, and the medical domain application (Ophthalmology *vs.* CXR). In terms of triples annotation, their approach is granular, using the original linguistic terms and relations, without further categorisation and processing: the entities are represented by single words as written in the source text, and they consider the verbs extracted with a dependency parser as the relations. Thus, our annotation pipeline generates a much lower number of entities, relation and triples, standardising and simplifying the triples. Moreover, in terms of model architecture, whilst they train the model end-to-end using a triples restoration loss, we keep the two steps independent from one other, and frame each step as a sequence-to-sequence task.

In summary, our contributions are to:

1. propose using a clinically informed schema to express the information in CXR radiology reports in structured form, using triples (entity1, relation, entity2);

2. propose a two-step pipeline for CXR radiology report generation: *Triples Extractor* followed by *Report Generator*;

3. conduct extensive experiments on the MIMIC-CXR dataset (Johnson et al., 2019a,b; Goldberger et al., 2000), showing state-of-the-art results for NLG and clinical accuracy metrics.

2 Methods

In this section we describe how the ground truth triples were extracted from the Finding section of each original report. Further, we introduce the

two-step pipeline, describing in detail the model architectures. In Figure 1, we show a high level design of the two-step pipeline.

2.1 Ground Truth Triples

We hereby present the steps we adopted to extract the ground truth triples from the Finding sections of the radiology reports; these triples are used to supervise the first step of the proposed two-step pipeline. The triples are represented as (e_1, r, e_2), where e_1 and e_2 are two entities linked together by a relationship r.

The overall annotation pipeline is shown in Figure 3. We use two publicly available tools to annotate the ground truth triples, which are then refined with the help of a junior physician with 2 years of clinical experience. We consider only sentences that can be extracted from a single CXR image, therefore we filter out mentions of comparisons with previous scans, since they are not always available in the MIMIC-CXR dataset.

RadGraph Entity & Relation Extraction We first apply RadGraph (Jain et al., 2021), which extracts entities and relations from a radiology report. RadGraph classifies the extracted entities as *Anatomy* corresponding to anatomical concepts (*e.g., heart* or *lung*), or *Observation* referring to words associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. The *Observation* entities are further categorised as *Definitely Present*, *Uncertain*, and *Definitely Absent*. The schema proposed by RadGraph includes three different relations: *Suggestive Of* – which links two *Observation* entities, where the second entity is implied based on the first entity (*e.g.,* «*opacity* → *SUGGESTIVE_OF* → *pneumonia*»); *Located At* – which indicates where an *Observation* entity is located (*e.g.,* «*fracture* → *LOCATED_AT* → *rib*»); and *Modify* – indicating that the first entity modifies the scope of, or quantifies the degree of, the second entity (*e.g.,* «*dense* → *MODIFY* → *consolidation*»). We use the pre-trained model[1] to extract the entities and relations from the Finding section of MIMIC-CXR radiology reports. Given that we aim to represent each report as a set of triples, we introduce another relation named *Status*, to include the three categorisations that RadGraph associates to each *Observation* entity: *Definitely Present* becomes *STATUS present*, *Uncertain* becomes *STATUS uncertain*, and *Definitely Absent* becomes *STATUS absent*.

ScispaCy Entity Extraction The RadGraph schema was designed to prefer granular entities (mostly represented by single words), linked to one other with many relations, in order to have dense annotations associated with each report. However, to simplify the task, we want to merge triples which could be sensibly represented as a single entity (*e.g.,* «*enteric* → *MODIFY* → *tube*» can be merged into a single medical entity called «*enteric tube*»). Therefore, we additionally use a named-entity recognition model which extracts less granular medical entities, namely ScispaCy's (Neumann et al., 2019) `en_core_sci_scibert` model[2].

Merge Radgraph and SciscpaCy entities The third step consists of merging together the two sets of entities associated with the same report, while keeping the relations extracted with RadGraph. This is performed by prioritising entities extracted using ScispaCy E_{sc} over these extracted using RadGraph E_{rg}. Formally, if there exists $e_{sc} \in E_{sc}$ and $e_{rg} \in E_{rg}$ such that $e_{rg} \subset e_{sc}$ (i.e. e_{rg} is a substring of e_{sc}), then we substitute e_{rg} with e_{sc} and assign to it all the relations originally associated with e_{rg}. Moreover, if $e_{rg,1}$ and $e_{rg,2}$ are linked together with a relation – $(e_{rg,1}, r, e_{rg,2})$ – and $e_{rg,1}, e_{rg,2} \subset e_{sc}$, then we remove the relation r and only keep e_{sc} as a single entity. Otherwise, if $e_{rg} \not\subset e_{sc}\ \forall e_{sc} \in E_{sc}$, then we keep e_{rg} and its associated relations.

Normalise entities and categorise relations according to clinical schema The final step of our annotation process comprises the refinement of the merged entities. With the help of a junior physician, we defined five entity categories: *Anatomy* (*e.g.,* «*heart*»), *Finding/Pathology* (*e.g.,* «*pneumothorax*», «*effusion*»), *Location* (*e.g.,* «*left*», «*top*»), *Modifiers* (*e.g.,* «*large*», «*left*») and *Status* (*e.g.,* «*present*», «*normal*»). For each entity term, we defined a set of synonyms. We then associate the term when one of the synonyms is detected in an entity span. Further, we constrain the triples to a fixed schema, based on the entity labels, as shown in Figure 2, and filter out the triples whose entity types and relations do not appear in that schema. If more than one of the manually selected terms is found inside an entity name, we split the entity and assign the relation based on the same schema. This occurs when

[1] https://physionet.org/content/radgraph/1.0.0/

[2] https://github.com/allenai/scispacy

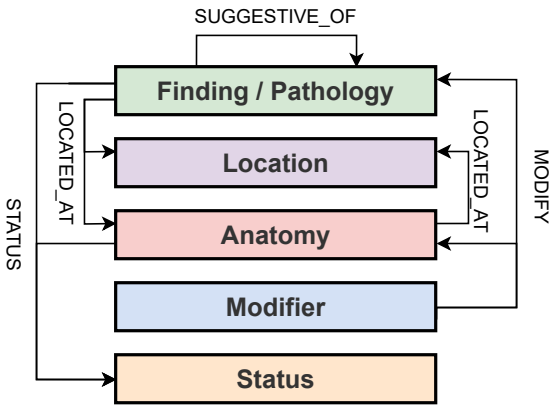

Figure 2: Triples schema. The relations correspond to the edges of the graph, and the type of relation is indicated in capital letters. The entity labels are represented by the nodes of the graph. These represent the triples to which our annotation pipeline is constrained.

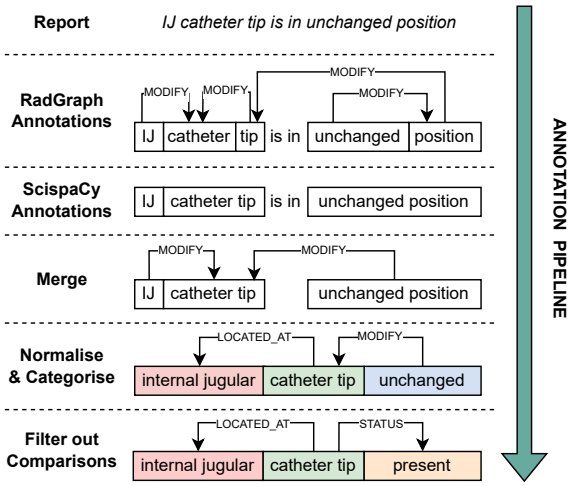

Figure 3: Example of the annotation pipeline to extract the ground truth triples from the radiology report. In the last two steps, we adopt the same color scheme as indicated in Figure 2, to categorise the entities.

ScispaCy detects entities that can be expressed as the combination of two or more separate entities (*e.g., «pulmonary vascular engorgement»* can be expressed as «*engorgement* → *LOCATED_AT* → *pulmonary vascular*»).

Filter out comparisons to previous reports Finally, we substitute the triples that express a change from previous studies of the same patient, since we are aiming to generate the report from a single CXR image, without having access to previous images. We identify the triples expressed as «e_1 → *MODIFY* → e_2», where e_1 corresponds to *«unchanged»*, *«new»*, *«increase»* or *«decrease»*; we then substitute the triple with «e_2 → *STATUS* → *present*», based on the assumption that if the radiologist mentions a change of a pathology or a finding, this is still present and visible in the image.

2.2 Model

We propose a novel framework to perform automated reporting in two steps: *Triples Extraction* and *Report Generation*. Similarly to Chen et al. (2020), we design and train Transformer models with custom architectures from scratch. Figure 4 shows a detailed diagram of the two-step pipeline.

Triples Extractor (TE) The first step consists of extracting the triples associated with each CXR image, whose semi-automated annotation process is described in Section 2.1. We treat this problem as a sequence-to-sequence task, using a multimodal encoder-decoder Transformer as the backbone, with both the CXR image and the indication field (*i.e.*, scan request text) as inputs. The benefit of using the indication field as context for CXR classification in an encoder Transformer model was previously shown by Jacenków et al. (2022).

The multimodal input sequence is the concatenation of the CXR image embedding and the indication field text embedding. The image embedding, denoted $I = \{I_1 \dots I_N\}$, corresponds to the feature map extracted from the last convolutional layer of ResNet-101 and flattened into a 49×2048 image embedding. The text input is tokenised into a $M \times 2048$ token embedding, indicated as $W = \{W_1 \dots W_M\}$. Further, we sum to the input sequence a segment embedding – to allow the model to discriminate between visual and textual inputs – and position embedding – needed by the Transformer to access the order of the input embedding. A `[SEP]` token is used to separate the two input modalities. The target sequence $Trp = \{Trp_1 \dots Trp_K\}$ corresponds to the concatenation of the ground truth triples, each separated by a `[SEP]` token.

We compare two different setups of the triples extractor model *TE-Transformer* to generate the triples (T):

- **CXR → Trp**: a visual Transformer, which only takes a single CXR image as input.
- **CXR + Ind → Trp**: a multimodal Transformer which takes as input the *Indication Field* (Ind), along with the CXR image, to provide additional context to the model.

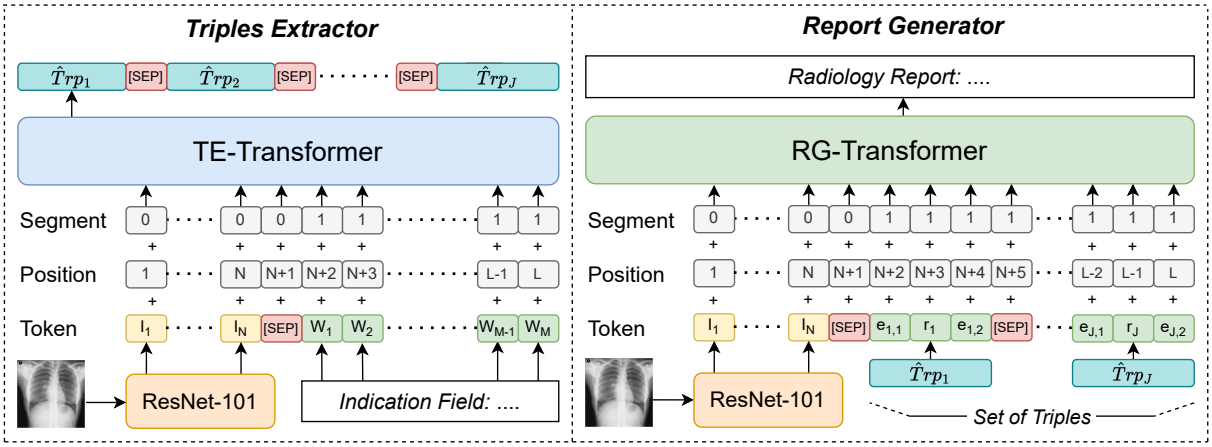

Figure 4: Architecture design of the two models: *Triples Extractor* and *Report Generator*.

Report Generator (RG) The second step of the pipeline corresponds to the generation of the radiology report. The problem is again framed as a sequence-to-sequence task, using a multimodal encoder-decoder Transformer as the model backbone. The multimodal input sequence comprises the CXR image embedding $I = \{I_1 \dots I_N\}$, computed as in step 1; and the text embedding $\hat{T}rp = \{\hat{T}rp_1 \dots \hat{T}rp_J\}$ represents the extracted triples from step 1, which correspond to a single string of text, where the triples are separated by a `[SEP]` token.

During the training phase, we use the concatenation of the ground truth triples $Trp = \{Trp_1 \dots Trp_K\}$, to train our model. To prevent the model focussing only on the triples – which already contain a comprehensive set of information, sufficient to generate a clinically accurate report – and ignoring the CXR image, we also consider randomly masking out 40% of the triples (this percentage was selected empirically). This way, we expect the model to also learn representative features from the image to compensate the missing information. We adopt such a training strategy because step 1 is not expected to be performed perfectly, thus we force the model to still consult the image when generating the final report.

During this step, we compare three different setups of the report generator model *RG-Transformer*, to generate the radiology report (RR):

- **Trp** $\rightarrow$ **RR**: a Transformer which takes only triples as input to generate radiology report.
- **Trp + CXR** $\rightarrow$ **RR**: a multimodal Transformer taking both triples and CXR as inputs.
- **Trp + CXR** $\rightarrow$ **RR (w/ Mask)**: a multimodal Transformer, similar to the above, trained on a random subset of the input triples.

3 Experimental Setup

3.1 Dataset

We conduct our experiments on the MIMIC-CXR dataset, which comprises 377,110 CXR images from 65,379 patients and the associated radiology reports. In this work we adopted the same training/validation/test split as used by Chen et al. (2020)[3] and Chen et al. (2021)[4], for a fair comparison with their methods. This results in 270,790 training images, 2,130 validation images and 3,858 test images, alongside the associated radiology reports. All the images are resized by matching the smaller edge to 256 pixels and maintaining the original aspect ratio.

Following previous methods, we consider only the *Finding Section* of each report as the target text output of our pipeline; this is the section in the report which contains a free-text description of the radiographic findings and/or pathologies which are visualised within the image. Further, we extract the *Indication Field* (sometimes termed *Clinical History*) from the radiology reports, when this is present, as it contains relevant medical history. We use this as additional context for the Triples Extraction step, since this is the part of the report that would be available at imaging time.

3.2 Baselines

We compare our two-step pipeline with:

[3] https://github.com/cuhksz-nlp/R2Gen
[4] https://github.com/cuhksz-nlp/R2GenCMN

- **Lower Bound (CXR → RR)**: an encoder-decoder Transformer architecture which generates the reports from the CXR in one step, without extracting the triples first. This defines the Lower Bound, and we expect our two-step pipeline to outperform this.

- **Upper Bound (GT-Trp → RR)**: we train an encoder-decoder Transformer to generate the radiology report from the ground truth triplets (GT-Trp). This sets an Upper Bound to our problem, as it mimics the scenario where all the triples are perfectly extracted in step 1. This allows us to understand the feasibility of generating a report from the set of triplets.

3.3 Implementation Details

We consider the same model architecture for both steps of the proposed pipeline. A vanilla encoder-decoder Transformer is used as the backbone of our models. Both its encoder and decoder are composed by three Attention Layers, as described by Vaswani et al. (2017), each composed by 8 heads and 512 hidden units, and we initialise them randomly. For both steps, the vocabulary of the tokeniser is defined independently, where each token corresponds to a single word appearing either in the input or output text of the training set; with an additional `[SEP]` token used in the input to separate the image vs text (first step), or image vs triples (second step).

We adopt ResNet-101 as the visual extractor, initialised using ImageNet pre-trained weights (Deng et al., 2009), with the scope of encoding a single CXR image and feeding the embedding to the Transformer as the visual input. During training, we adopt standard data augmentation of the image: random 224×224 crop; random horizontal flip; and random rotation within the range $(-10°, +10°)$. During inference, we take a 224×224 central crop of the image.

For each step, the whole model is trained end-to-end using a cross-entropy loss with Adam optimiser (Kingma and Ba, 2014). The learning rate for the visual extractor is set to 5×10^{-5} and 1×10^{-4} for the remaining parameters, and we decay them by a factor of 0.8 every three epochs.

3.4 Metrics

To evaluate the goodness of step 1, we compute the F1 score between the set of extracted triples $\hat{Trp}$ and the set of ground truth triples Trp.

Model	val F1	test F1
CXR → Trp	0.348	0.275
CXR + Ind → Trp	**0.411**	**0.307**

Table 1: F1 scores for triples (Trp) extracted in step 1 on the validation and test set of MIMIC-CXR. We compare two different versions of the Triples Extractor, as defined in Section 2.2.

Step 2 is evaluated using common Natural Language Generation (NLG) metrics: BLEU score (Papineni et al., 2002), ROUGE score (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). Given that these often fail to capture the semantic meaning of the text, we also consider Clinical Efficiency (CE) metrics. These are computed by applying the CheXpert labeler (Irvin et al., 2019) to the generated reports, which extracts 14 labels: *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax*, and *Support Devices*. Generated labels are then compared with the ground truth labels, provided in the MIMIC-CXR dataset, by computing precision, recall and F1 scores. We note that the CheXpert labeler provides only a partial assessment of clinical accuracy, since attributes are ignored, as well as entities outside of the 14 defined labels. Therefore we also perform a qualitative human evaluation of a subset of the generated reports.

4 Results

Here we evaluate our proposed method on the MIMIC-CXR dataset at each step: *Triples Extraction* and *Report Generation*. Every experiment is repeated 3 times using different random seeds to initialise the model weights and randomise batch shuffling; we report the average scores between the 3 different runs. We also conduct some human evaluation on the generated reports, to further assess their clinical accuracy.

4.1 Results on Triples Extraction

In Table 1, we compare the two models – CXR TE-Transformer and MM TE-Transformer – by computing the F1 score on both the MIMIC-CXR validation and test set. This shows that introducing the *Indication Field* as additional context to the model helps to restore the triples more accurately. This result confirms what has previously

Model		NLG Metrics						CE Metrics		
Step 1	**Step 2**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**P**	**R**	**F1**
Lower Bound: CXR → RR		0.341	0.212	0.145	0.106	0.136	0.280	0.373	0.33	0.334
CXR + Ind → Trp	Trp → RR	0.322	0.219	0.159	0.122	0.150	0.311	**0.454**	0.431	0.442
CXR + Ind → Trp	CXR + Trp → RR	0.336	0.226	0.164	0.125	0.149	0.307	0.439	0.398	0.417
CXR + Ind → Trp	CXR + Trp → RR (w/ Mask)	**0.363**	**0.245**	**0.178**	**0.136**	**0.161**	**0.313**	0.428	**0.459**	**0.443**
Upper Bound: GT-Trp → RR		0.523	0.408	0.332	0.276	0.251	0.466	0.523	0.581	0.551

Table 2: NLG and CE results on the MIMIC-CXR test set, where BL=BLEU, MTR=METEOR, RG=ROUGE, P=Precision and R=Recall. We adopt the two-step pipeline, considering a multimodal TE-Transformer to extract the triples in the 1st step, and comparing different implementation of the 2nd step, defined in Section 2.2. These results are also compared with the Lower Bound and the Upper Bound models, described in Section 3.2.

Model	NLG Metrics						CE Metrics		
	BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**P**	**R**	**F1**
ST (Vinyals et al., 2015)	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
Att2In (Rennie et al., 2017)	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
AdaAtt (Lu et al., 2017)	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
TopDown (Anderson et al., 2018)	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
CA (Liu et al., 2021c)	0.350	0.219	0.152	0.109	0.151	0.283	-	-	-
CMCL (Liu et al., 2021a)	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
PPKED (Liu et al., 2021b)	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
R2Gen CMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
R2Gen CMM+RL (Qin and Song, 2022)	**0.381**	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
Ours	0.363	**0.245**	**0.178**	**0.136**	**0.161**	**0.313**	**0.428**	**0.459**	**0.443**

Table 3: NLG and CE results on the MIMIC-CXR test set. All the results of the comparison methods are taken from Qin and Song (2022).

been found by (Jacenków et al., 2022), and extends their results on a more difficult task.

4.2 Results on Report Generation

In Table 2, we show a comparison of three variants of the Report Generator, which are described in Section 2.2. We also compare the results with a Lower Bound and a Upper Bound model, defined in Section 3.2. During inference, for all three models we input the triples extracted by the MM TE-Transformer, as it yields the highest F1 scores.

It can be seen that the models trained without masking do not consistently outperform the Lower Bound metrics. The reason could be attributed to the fact that, during training, we input to the model the ground truth triples, which contain the necessary information to generate a good quality report. Therefore, the model tends to focus solely on the triples, and always expects to see a set of triples perfectly matching the final report. However, this is not true, as seen from the results in Table 1. We overcome this by masking out some of the ground truth triples during training, which encourages the model to leverage also the CXR image when generating the radiology report. Moreover, it can be noticed that all three models show significantly lower performance compared to the UB. This suggests that there is still a considerable margin of improvement.

In Table 3, we benchmark our pipeline against existing state-of-the-art automated radiology reporting methods. Our two-step approach outperforms other methods for most of the NLG metrics and all the CE metrics, suggesting a good compromise between clinical accuracy and text fluency of the generated radiology reports.

4.3 Human Evaluation

We additionally evaluated the quality of reports using two human evaluators, who compared the reports generated by the baseline model and our two-step model to the original report. The evaluators were junior physicians with 2 and 3 years of clinical experience respectively, including experience of reading CXR reports. Evaluators were blinded to the model type used to generate reports during the exercise. For each example, evaluators were shown the radiologist's report and treated this as the gold standard (they were not shown the underlying CXR image). In line with human evaluation

Original Report	Generated Report
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.	As compared to the previous radiograph there is no relevant change. The extent of the right **pleural effusion** is constant. Constant size of the cardiac silhouette. No newly appeared parenchymal opacities. Omission errors = **Biapical scarring, hilar retraction, pleural thickening, XXXX deviation**
Large left lower lobe opacity is present. There does not appear to be significant mediastinal shift. There is no pneumothorax. the cardiac silhouette is not definitively identified and not fully evaluated. The mediastinal contours are unremarkable.	PA and lateral views of the chest were reviewed and compared to the prior studies. A right **pleural effusion** has increased in size since the prior study. **The left lung is clear**. **There is no pneumothorax.** Omission errors = **Left lower lobe opacity, mediastinal shift, mediastinal contours**

Figure 5: Example of human evaluation undertaken on generated reports. Errors: **Hallucination**, **Omission**, **Attribute error**, **Impression error**. In this data, taken from the IU-Xray dataset (Demner-Fushman et al., 2016), ages (and other patient-identifiable information) is replaced by a placeholder, here indicated by XXXX.

Error Type	Baseline	Two-Step	RC
Hallucinations	101	66	-0.35
Omissions	103	86	-0.17
Attribute Errors	29	25	-0.14
Impression Errors	4	6	+0.50
Grammatical Errors	3	1	-0.67
Total Errors	240	184	-0.23
Critical Errors	31	22	-0.29

Table 4: Number of errors found by the clinical evaluators in 60 reports generated with the baseline and the two-step model. We indicate with RC the relative change between the two models' errors.

methods used to assess voice recognition software (Rana et al., 2005; Quint et al., 2008; Ringler et al., 2017), evaluators counted types of errors which occurred in generated reports. The types of errors available were *1. Hallucination, 2. Omission, 3. Attribute error, 4. Impression error and 5. Grammatical error*. Examples of the use of these errors is shown in Figure 5. There was also the option for evaluators to assign a *critical error* to the first four errors if this was felt to significantly alter the clinical course of action. For example, if a generated report erroneously described a region as being suggestive of pneumonia, this might result in a patient unnecessarily receiving antibiotics. Alternatively, if a report failed to describe a mass, this might result in possible cancer being missed.

The evaluators discussed and agreed the evaluation protocol prior to the exercise. Evaluators received a combined total of 60 ground truth reports alongside the reports generated with the baseline and the two-step approach, including 10 reports shown to both evaluators to compute the inter-annotator agreement. We found a moderate agreement between the two annotators with a Gwet's AC1 score (Gwet, 2014) equal to 0.53.

The number of detected errors are displayed in Table 4. Most of the errors are reduced when using our two-step approach, which is consistent with the results in Section 4.2. This shows that the two-step approach generates more clinically accurate radiology report compared to the single-step baseline. However, the number of clinical error are still significant, which makes this method still unsuitable for real-life diagnostic applications.

5 Conclusion

In this work, we present a two-step framework for CXR automated radiology reporting, which splits the task into *Triples Extraction* and *Report Generation*. We propose a semi-automated annotation schema, which extracts structured information from a radiology report in the form of triples, and serves to supervise the first step of our approach. Further, our method shows state-of-the-art performances on the MIMIC-CXR dataset for most of the NLG metrics and all the CE metrics. Moreover, we conduct human evaluation to assess errors in the generated text, showing how our proposed two-step approach generates 23% fewer errors and 29% fewer critical errors compared to the baseline. Nevertheless, end-to-end supervised report generation from images requires further research on improving clinical accuracy in order to have utility as a diagnostic tool.

In future, this solution can easily integrate more advanced model architectures – to both improve the triple extraction and the report generation – and can be applied to other complex image captioning tasks, such as those found in the medical domain.

References

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Care Quality Commission. 2018. A national review of radiology reporting within the nhs in england. pages 1–26.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, S. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Grzegorz Jacenków, Alison Q O'Neil, and Sotirios A Tsaftaris. 2022. Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

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. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1).

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. 2022. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021c. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

NHS England and NHS improvement. 2022. Diagnostic imaging dataset statistical release. pages 1–17.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland. Association for Computational Linguistics.

Leslie E Quint, Douglas J Quint, and James D Myles. 2008. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. *Journal of the American College of Radiology*, 5(12):1196–1199.

DS Rana, G Hurst, L Shepstone, J Pilling, J Cockburn, and M Crawford. 2005. Voice recognition for radiology reporting: is it good enough? *Clinical radiology*, 60(11):1205–1212.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Michael D Ringler, Brian C Goss, and Brian J Bartholmai. 2017. Syntactic and semantic errors in radiology reports associated with speech recognition software. *Health informatics journal*, 23(1):3–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, page 102510.