Artificial Intelligence In Medicine 116 (2021) 102075

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Evaluating diagnostic content of AI-generated radiology reports of chest X-rays

Zaheer Babar[a,*], Twan van Laarhoven[a], Fabio Massimo Zanzotto[b], Elena Marchiori[a]

[a] *Insitute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands*
[b] *Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy*

ARTICLE INFO

Keywords:
Automated radiology report generation
Assessment of results' quality

ABSTRACT

Radiology reports are of core importance for the communication between the radiologist and clinician. A computer-aided radiology report system can assist radiologists in this task and reduce variation between reports thus facilitating communication with the medical doctor or clinician. Producing a well structured, clear, and clinically well-focused radiology report is essential for high-quality patient diagnosis and care. Despite recent advances in deep learning for image caption generation, this task remains highly challenging in a medical setting. Research has mainly focused on the design of tailored machine learning methods for this task, while little attention has been devoted to the development of evaluation metrics to assess the quality of AI-generated documents. Conventional quality metrics for natural language processing methods like the popular BLEU score, provide little information about the quality of the diagnostic content of AI-generated radiology reports. In particular, because radiology reports often use standardized sentences, BLEU scores of generated reports can be high while they lack diagnostically important information. We investigate this problem and propose a new measure that quantifies the diagnostic content of AI-generated radiology reports. In addition, we exploit the standardization of reports by generating a sequence of sentences. That is, instead of using a dictionary of words, as current image captioning methods do, we use a dictionary of sentences. The assumption underlying this choice is that radiologists use a well-focused vocabulary of 'standard' sentences, which should suffice for composing most reports. As a by-product, a significant training speed-up is achieved compared to models trained on a dictionary of words. Overall, results of our investigation indicate that standard validation metrics for AI-generated documents are weakly correlated with the diagnostic content of the reports. Therefore, these measures should be not used as only validation metrics, and measures evaluating diagnostic content should be preferred in such a medical context.

1. Introduction

The written radiology report is the most important means of communication between a radiologist and the referring clinician and is essential to high-quality patient care [1]. Radiologists wish to produce reports having an appropriate construction, clarity, and clinical focus. However, in daily radiology practice, the report generation task falls towards the end of the radiology workflow, and is a time-consuming and error-prone task. Fig. 1 shows the impression and findings sections of a radiology report. The *impression* section is a single sentence summary, while *findings* describes technical observations in detail of both normal and abnormal characteristics in the image. It consists of sentences covering various aspects such as heart size and lung opacity; any abnormality appearing at lungs, aortic and hilum; and potential diseases such as pneumothorax and consolidation [2].

A computer-aided radiology report system can assist radiologists to generate good reports and decrease their workload. Despite recent advances in deep learning for image caption generation, automated generation of radiology reports remains a challenging task. In particular, assessment of the quality of an AI-generated radiology report is highly domain-dependent. The performance of report generation methods is mainly assessed using conventional quality metrics for natural language processing methods [3–5]. However, the goodness of a radiology report is intrinsically linked to its diagnostic content, and it is not clear whether there is a correlation between text-based metrics and the quality of the diagnostic content of a generated report.

[*] Corresponding author.
E-mail addresses: zbabar@cs.ru.nl (Z. Babar), tvanlaarhoven@cs.ru.nl (T. van Laarhoven), elenam@cs.ru.nl (E. Marchiori).

https://doi.org/10.1016/j.artmed.2021.102075
Received 25 February 2020; Received in revised form 5 February 2021; Accepted 7 April 2021
Available online 15 April 2021
0933-3657/

Fig. 1. Example of impression and findings sections of a radiology report (Indiana U. Chest X-ray dataset). xxxx's are wrongly removed keywords due to de-identification.

Therefore, in order to comparatively assess the diagnostic quality of generated radiology reports, we introduce a new validation measure. We assume a set of diagnostic tags is given (provided by an external source), which are associated to each image in the training and test set. These diagnostic tags are not used for training the report generator but to assess diagnostic quality of generated reports. To this aim, tags are used as class labels of reports, and used to train a probabilistic model. The resulting model estimates the diagnostic score of a report by its capability to correctly predict the diagnostic tags of the corresponding image. The application of this model to reports generated from test images provides a diagnostic measure, that we call the diagnostic content score. Interestingly, this measure can be also used to quantify the diagnostic content of manually generated reports.

Most methods for report generation are based on deep learning and use an encoder-decoder architecture stemming from machine translation [6]: an image is transformed into visual features by the encoder and the decoder transforms visual features into a textual description of that image. The textual description consists of a sequence of elements from a given dictionary. The dictionary used in all conventional methods consists of the set of words occurring in the reports of the training data. However, radiologists tend to use a well-focused vocabulary of 'standard' sentences to describe their findings [7], with ongoing standardization initiatives favoring this trend [8,7]. Therefore, it is interesting to investigate the impact on the diagnostic quality of a dictionary of sentences instead of a dictionary of words.

We perform a comparative analysis of the use of word- and sentence-based dictionary on two recent attention-based models that automatically learn to describe the content of images [9,2]. We use the publicly available Indiana U. Chest X-rays dataset from the Open-i image collection [10].

We show experimentally that using a dictionary of sentences instead of a dictionary of words, as all current image caption generation methods do, does not significantly change the quality of the generated reports, both in terms of text-based metrics and in terms of diagnostic content. As a by-product, a significant training speed-up is achieved compared to models trained on a dictionary of words.

Results of our experimental investigation also show that validation metrics for natural language processing methods and our diagnostic content validation metric are only weakly correlated. Therefore, these measures should be not used as the only validation metrics, and metrics evaluating diagnostic content should be used as well in medical contexts.

Our contributions toward a computer-aided radiology report system can be summarized as follows: (a) an external quality measure to assess the diagnostic content of AI-generated radiology reports; (b) an in-depth comparative analysis of the proposed quality measure and standard quality metrics for natural language processing methods; (c) extensive experiments on a publicly available dataset, using two machine learning methods for text generation from images (a popular image caption method and a recent method specifically developed for radiology report generation), and two types of dictionaries, the standard word-based dictionary and a sentence-based one.

2. Related work

In recent years many techniques have been proposed to address the task of automatic image captioning. Initial models were mainly based on feed-forward neural networks [11] and consisted of multimodal architectures which could be conditioned on other modalities. Later on, feed-forward neural networkswere replaced by recurrent neural networks, see e.g. [12].

The release of the Microsoft COCO dataset [13] stimulated research on caption generation from images. Significant progress was achieved especially through the introduction of models that combined Convolutional Neural Networks (CNN's) with Recurrent Neural Networks (CNN-RNN's) [14,9,15]. CNN-RNN's were inspired by the encoder-decoder approach used in machine translation. Rather than translating text to text, a CNN-RNN translates an image to a text. The CNN is used as an encoder and an RNN as decoder, as described in the next section.

2.1. Conventional techniques

The first use of the encoder-decoder architecture for image captioning was in the *Show and Tell* method [14]. This method uses a deep CNN to extract features from an image, and then uses these features as an input for the first time-step of a Long Short Term Memory (LSTM) network. In [9] a more involved model called *Show, Attend and Tell* was proposed, which introduces two attention-based mechanisms called soft attention and hard attention. These mechanisms mimic the human eye's characteristic to switch focus between different parts of an image. Donahue et al. [16] also proposed an encoder-decoder model which uses a recurrent CNN. At each time-step, the model takes a different variation of the same image as an input. The resulting model is also applicable to generate descriptions of videos.

2.2. Specialized techniques

Compared to automatic image captioning, generating a textual description for radiology images is a more challenging task. Therefore, a number of specialized computational methods have been introduced.

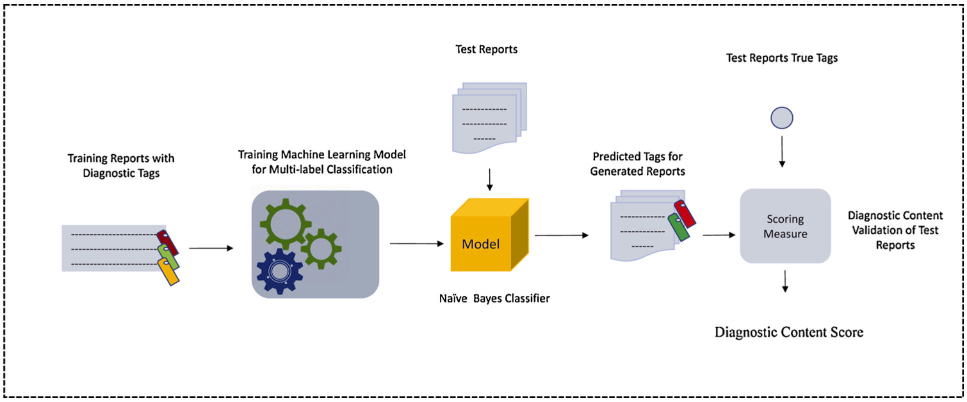

Test Reports
Test Reports True Tags
Training Reports with Diagnostic Tags
Training Machine Learning Model for Multi-label Classification
Predicted Tags for Generated Reports
Model
Scoring Measure
Diagnostic Content Validation of Test Reports
Naïve Bayes Classifier
Diagnostic Content Score

Fig. 2. Methodology: diagnostic content score.

In [17] a machine learning method was introduced for annotating chest X-rays with Medical Subject Headings annotations. This method involves a CNN for classifying X-ray images into given disease-based classes, RNN models for learning representations of the context of disease-based classes, and a cascade model which combines image and context representations to generate annotations. In [18] an unsupervised procedure, called Looped Deep Pseudo-Task Optimization, was introduced to discover disease-based classes of radiology images. The procedure starts from pseudo class labels derived from text reports, which ate fine-tuned by means of an iterative procedure.

A hierarchical report generation approach was proposed in [19]: a CNN is used to learn visual features from images as well as for predicting their diagnostic tags, followed by a LSTM in combination with a co-attention mechanism to generate a textual report. This method relies on the availability of diagnostic tags at training time, and may generate reports containing repetitions.

In [2] a recurrent generative model was introduced to generate both the impression sentence and the finding paragraph of a radiology report. This method makes use of global image features to generate the impression sentence of a report. Next, it uses the generated impression sentence together with local image features as input to generate the findings paragraph, in a sentence-by-sentence fashion. Although the method is reported to achieve good results on the IU chest X-ray dataset, it is not capable to generate sentences that did not already occur in the training set. This limitation is shared by the sentence-based dictionary approach we propose. In [20] a CNN-RNN-RNN based model was proposed: a CNN is used to learn visual feature representations, which are fed to a RNN to generate a topic for each sentence. Next, an RNN uses topic and visual features to generate sequence of words.

In [21] an enhanced encoder-decoder approach is used. In particular, the encoder is pre-trained with a large number of chest X-ray images to classify 14 common radiographic observations, fine tuned to extract the most frequent medical concepts from the X-ray images.

Specialized (hybrid) Information Retrieval (IR) methods for radiology report generation have been introduced. Although in our analysis we will use only machine learning baseline methods, we briefly summarize two recent methods based on IR.

In [22] a Knowledge-driven Encode, Retrieve, Paraphrase (KERP) method was introduced which reconciles traditional knowledge- and retrieval-based methods with modern learning-based methods. KERP employs an encoder module that transforms visual features into a structured abnormality graph by incorporating prior medical knowledge; then a Retrieve module retrieves text templates based on the detected abnormalities; finally, a Paraphrase module rewrites the templates according to specific cases.

In [23] a method called Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) was proposed. This method considers sentences along with words to generate diagnostic reports. It uses a retrieval policy module to decide whether at a certain point it should retrieve a template sentence from a off-the-shelf database or invoke a generation module to generate a new sentence.

2.3. Datasets

Various publicly available datasets from the radiology domain have been introduced. In [24] a multilabel ChestX-ray8 dataset was released. This dataset was mainly used to classify and localize commonly occurring thoracic diseases. Two other recent publicly available chest X-ray

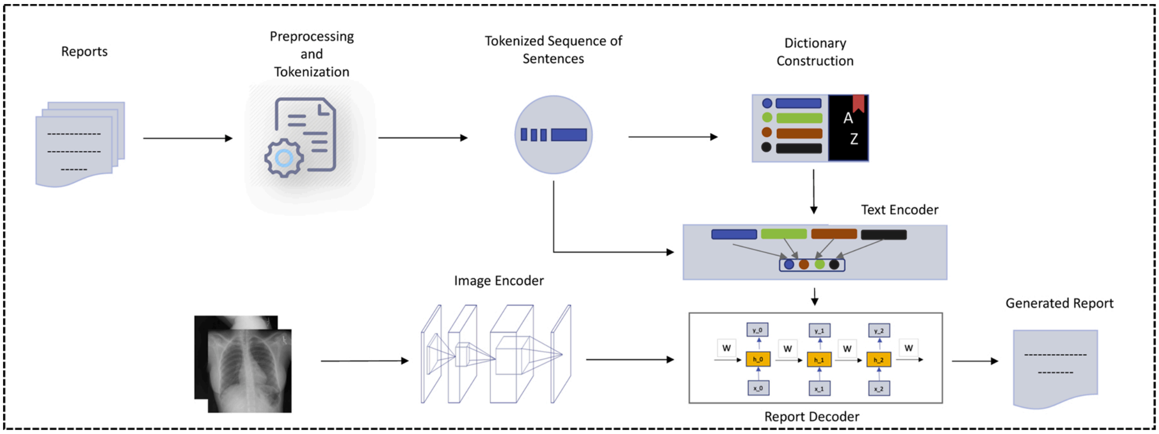
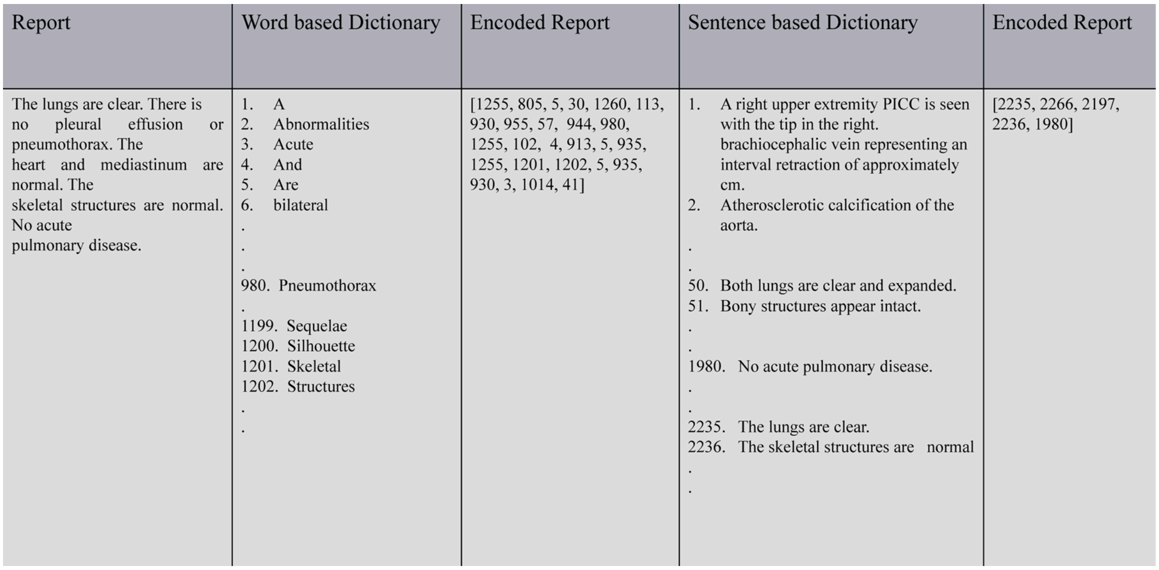


Fig. 3. Methodology: training using sequence of sentences.



Report	Word based Dictionary	Encoded Report	Sentence based Dictionary	Encoded Report
The lungs are clear. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal. No acute pulmonary disease.	1. A 2. Abnormalities 3. Acute 4. And 5. Are 6. bilateral . . . 980. Pneumothorax . 1199. Sequelae 1200. Silhouette 1201. Skeletal 1202. Structures . .	[1255, 805, 5, 30, 1260, 113, 930, 955, 57, 944, 980, 1255, 102, 4, 913, 5, 935, 1255, 1201, 1202, 5, 935, 930, 3, 1014, 41]	1. A right upper extremity PICC is seen with the tip in the right. brachiocephalic vein representing an interval retraction of approximately cm. 2. Atherosclerotic calcification of the aorta. . . 50. Both lungs are clear and expanded. 51. Bony structures appear intact. . . 1980. No acute pulmonary disease. . . 2235. The lungs are clear. 2236. The skeletal structures are normal . .	[2235, 2266, 2197, 2236, 1980]

Fig. 4. Word-based encoding vs. sentence-based encoding.

datasets are MIMIC-CXR [25] and ChestXpert [26]. Unfortunately, radiology reports from these datasets are not directly accessible. The only publicly available radiology dataset we are aware of, containing images and radiology reports, is the Indiana U. Chest X-rays dataset from the Open-i image collection [10]. This dataset has been extensively used to test methods for radiology report generation, and will be used in our experimental analysis.

2.4. *Diagnostic quality measures*

Few diagnostic-based quality measures have been introduced. In [2] the so-called Keywords Accuracy (KA) was proposed which considers the ratio of the number of diagnostic keywords in the generated report to the number of all diagnostic keywords in ground truth report. Diagnostic keywords are diagnostically relevant terms extracted using a Medical Text Indexer (MTI). While the KA measure quantifies diagnostic content of a generated report by the fraction of relevant keywords it contains, the diagnostic score we propose uses the prediction of a probabilistic classifier built on training data consisting of reports with medical tags as their class labels.

In [20] a clinical accuracy score based on the CheXpert labeler [26] is used. The CheXpert labeler involves extracting mentions from a list of observations from the impression part of radiology reports; a rule-base classifier is built to classify such mentions. Rules for mention classification are designed on the universal dependency parse of the report. CheXpert is used to compute annotations of generated and ground truth reports on 14 different categories related to thoracic diseases and support devices, which are then used to compute accuracy, precision, and recall values. The rule-based classifier is constructed using a dataset whose characteristics are possibly different than those of the reports it is applied to when computing clinical accuracy, hence the resulting assessment could be biased.

3. Methodology

In order to assess the diagnostic quality of generated reports, we propose a new scoring measure called Diagnostic Content Score (DCS) (see Fig. 2). Also, we investigate the use of sentence-based dictionary instead of word-based dictionary to train a report generator model (see Fig. 3). Dictionary construction and text encoding are the prime differences between a sentence-based and word-based setting. As illustrated in the Fig. 4, a sentence-based text encoding is simpler than a word-based one.

3.1. *Diagnostic content score (DCS)*

Conventional evaluation metrics for text generation from images include BLEU, ROUGE, and METEOR [4,27,28] (see Section 4.3). To date, these are the most widely used tools in machine translation and summarization for evaluating how good a generated text is as compared to the ground truth one. Such metrics measure goodness in a generic context, independent of the application domain. They score a generated report only in terms of how well it matches the corresponding ground truth report. As such, they can miss important clinical information contained in the generated report. For example, “Heart is normal” and “Heart is not normal” are syntactically close but semantically very different sentences. Therefore, we propose to assess the clinical quality of radiology reports using an external source of information (see Fig. 2). We assume an external source of ground truth knowledge in the form of a set of diagnostic tags for each report, see an example in Fig. 15. These diagnostic tags are external knowledge because they are not used for training the report generator model.

By considering tags as class labels associated to a report, we build a probabilistic model (on the training data) that predicts the class labels of a report. The probabilistic model is applied to the generated reports of the test set, and the model performance is used as quantitative estimate of the diagnostic quality of the generated reports.

Specifically, for predicting tags of a report generated from images (from the test set), we train a multinomial Naive Bayes classifier on the ground truth reports and associated external tags (from the training set), using n-grams of words as features. Let $w_1, w_2, \cdots, w_N$ be the words or n-grams that make up a report R. Then for each tag t, we compute $\widehat{\mathbb{P}}(t = 1|R)$, the probability that tag t is present in R, using Bayes rule as

$$\mathbb{P}(t = 1|R) \propto \widehat{\mathbb{P}}(t = 1) \prod_{i=1}^{N} \widehat{\mathbb{P}}(w_i|t = 1), \tag{1}$$

where $\widehat{\mathbb{P}}(t=1)$ is the empirical probability of a tag t, and $\widehat{\mathbb{P}}(w_i|t=1)$ is the empirical probability of reports that contain the word or n-gram w_i among reports labeled with tag t. These empirical probabilities are estimated with Laplace smoothing, that is, as the count of reports with the given tag and of the given word in the training set, plus 1, normalized to give a multinomial distribution. Similarly, we estimate $\widehat{\mathbb{P}}(t=0|R)$, the probability of a tag being absent from report R, based on reports in the training set that do not have that tag.

Then, for each tag t we predict that tag to be associated to the report if $\widehat{\mathbb{P}}(t=1|R) > \widehat{\mathbb{P}}(t=0|R)$. We use the F_1 score to compare the set of predicted tags to the true diagnostic tags associated with that report. We call this the diagnostic content score (DCS) of a report R:

$$\mathrm{DCS}(R) = 2\frac{|T\cap\widehat{T}|}{|T|+|\widehat{T}|},$$

where T is the set of true tags of R, and $\widehat{T}$ is the set of predicted tags. This score is a number between 0 and 1, where 1 means that the predicted tags are identical to the true tags, and 0 means that the sets are completely disjoint. We define the DCS over a set $\mathcal{R}$ of reports as the average DCS over the reports in $\mathcal{R}$: $\sum_{R\in\mathcal{R}}\mathrm{DCS}(R)$.

3.2. Sentence-based automated image report generation

In order to exploit the standardization of reports, instead of using a dictionary of words, as current image captioning methods do, we investigate the use of a dictionary of sentences. The assumption underlying this choice is that radiologists use a well-focused vocabulary of 'standard' sentences, which should suffice for composing most reports.

Fig. 3 shows the training procedure for radiology report generation models using a dictionary of sentences. Here there is no need for preprocessing such as stop words removal, punctuation removal, etc. Dictionary construction and text encoding are the prime differences between this setting and the word-based setting. To construct a dictionary of sentences, the training reports are tokenized into sentences of the dictionary (Fig. 3). Fig. 4 shows the dictionary construction and text encoding. As illustrated in the figure, a sentence-based text encoding is simpler than a word-based one.

We use two state-of-the-art machine learning baseline methods to investigate the impact of a dictionary of sentences. The first one is the popular attention-based CNN-LSTM model (Show, Attend, and Tell) [9]. The second baseline is the Multimodal Recurrent Attention model [2], a specialized method to automatically generate radiology reports.

We modify these baselines and consider their sentence-based dictionary variants. This approach illustrated schematically in Fig. 3.

3.2.1. Sentence-based show, attend, and tell model (SAT):

In a medical context, the goal of SAT is to generate a radiology report R given an input image X. To generate the report, an encoder-decoder architecture is used. The encoder generates visual features from the image. As in [9], we use the pre-trained deep CNN VGG-19 as encoder.

For the decoder, we modify and use the attention mechanism proposed in [9], which uses the encoded image as input for an LSTM. Here, we consider a dictionary of sentences instead of words, so at time t a sentence S_t instead of a word is output. At each time-step, this network focuses its attention on a specific part of the input image.

Formally, in our setting a report of length L consists of sentences S_1, S_2, $\cdots$, S_L. Each of these sentences is drawn from a dictionary of K different possible sentences that can occur in radiology reports. The likelihood of the whole report is

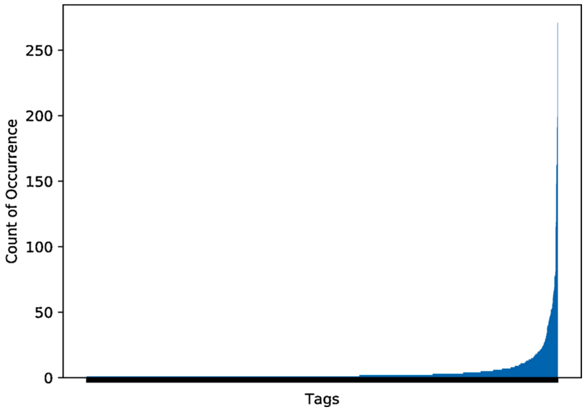


Fig. 5. Number of occurrences of diagnostic tags (excluding tag 'Normal') in the training data.

$$\mathbb{P}(R|X) = \prod_{i=1}^{L}\mathbb{P}(S_i|S_1,\cdots,S_{i-1},X). \tag{2}$$

3.2.2. Sentence-based multimodal recurrent attention model (MRA):

In MRA, we extract local and global features of an input image as described in [2], using the pre-trained deep CNN ResNet152.

For the decoder part, MRA uses a hierarchical approach to generate each sentence which comprises a sequence of words. MRA recurrently generates each sentence using the encoded image and the previous sentence as inputs.

Formally, each sentence is generated separately, depending only on the previous sentence and on the image X. This means that the likelihood of a sentence $S_i=[w_1,w_2,\cdots,w_{n_i}]$ of length n_i is

$$\mathbb{P}(S_i|X) = \prod_{j=1}^{n_i}\mathbb{P}(w_j|X,S_{i-1},w_1,\cdots,w_{j-1}). \tag{3}$$

The sentences together form the report of length L, which has a likelihood

$$\mathbb{P}(R|X) = \prod_{i=1}^{L}\mathbb{P}(S_i|X,S_{i-1}). \tag{4}$$

In a sentence-based model, the likelihood of a sentence in terms of its constituent words is replaced by a direct estimation of the likelihood based on a dictionary of sentences. This leaves Eq. (4) to describe the likelihood of a particular model.

4. Experiments

4.1. Data

We conduct experiments on the Indiana University (IU) Chest X-Ray collection [10], consisting of radiology diagnostic records of 3999 patients. Each record contains one or more chest X-ray images along with a corresponding textual report. Each report consists of four sections: impression, findings, comparison, and indication. As in [19], we use the concatenation of impression and findings as the target report to be generated.

Each report in the collection is associated with a subset of diagnostic tags. In total 1600+ unique tags are given (frequence of tags given in Fig. 5). On average a report is associated with 3.5 tags.

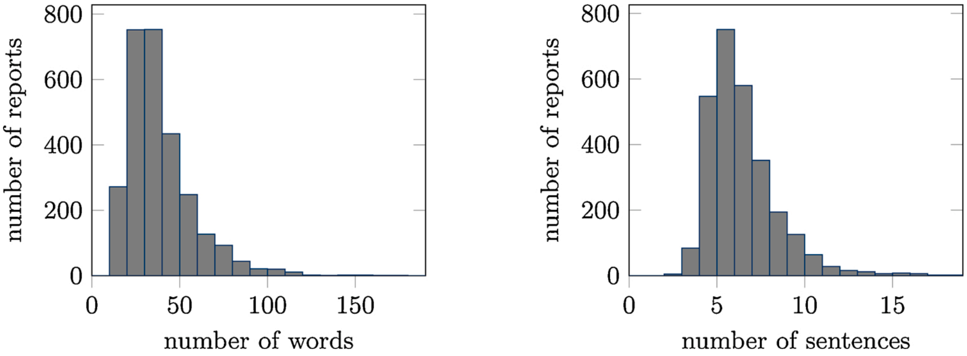


Fig. 6. Histogram of the length of reports in the IU Chest X-ray dataset.

Table 1
Evaluation of word- and sentence-based methods based on conventional metrics. (Standard deviation values are shown with '±').

Algo	Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
SAT Two-Images	Word-based	0.35 ±0.01	0.22 ±0.01	0.15 ±0.00	0.10 ±0.00	0.16 ±0.00	0.29 ±0.01
	Sentence-based	0.35 ±0.01	0.23 ±0.02	0.16 ±0.02	0.12 ±0.02	0.16 ±0.01	0.27 ±0.01
MRA Two-Images	Word-based	0.33 ±0.02	0.21 ±0.02	0.14 ±0.02	0.09 ±0.01	0.16 ±0.01	0.28 ±0.01
	Sentence-based	0.35 ±0.01	0.21 ±0.01	0.14 ±0.01	0.09 ±0.01	0.15 ±0.01	0.28 ±0.01

Almost each report in the collection is associated with two images: a frontal and a lateral view.

These images can be used independently, as done in [19], or jointly, as done in [2]. We consider the latter approach, which directly preserves the relatedness of the two views.

Originally, SAT processes one image at a time. To be able to process a pair of images together, we use a modified encoder which extracts features for each of the image views and concatenate them.

We consider only records containing two image views and complete textual report. After this filtering, we get a total of 2775 records (each consisting of a pair of images and associated medical report).

For each report, we convert all tokens to lowercase and remove all non-alphabetic tokens.

We maintain two separate dictionaries, one of words and one of sentences. We apply word-based and sentence-based tokenization to create these dictionaries.

The dataset has a total of 1933 unique words and 5100 unique sentences.

4.2. Experimental setup

We follow the same experimental setup as in [2]: out of 2775 records we randomly select 250 samples to form the test set and use the remaining data for the training. We repeat the split of the data into training and test set 5 times, and average the performance on the test sets of the models trained on the corresponding training sets. All models are trained for 60 epochs using a batch size of 16.

We have used existing implementations of SAT[1] and MRA[2], which have been adapted for the sentence-based setting.

For the word-based SAT model, we use 110 time-steps, since nearly all (over 99%) of the reports in the training data have fewer than 110 words (see Fig. 6). For the sentence-based SAT model we use a limit of 14 time-steps, which covers over 99% of the reports in the training data.

For the word-based MRA model, we use 40 time-steps to generate a single sentence. This is the maximum length for any sentence. For the sentence-based variant of MRA, it takes only one time step to generate a single sentence, excluding starting and end token.

Table 2
Evaluation of word- and sentence-based methods based on the CheXpert labeler.

Algo	Approach	Accuracy	Precision	Recall	F1-score
SAT Two-Images	Word-based	0.86	0.22	0.21	0.21
	Sentence-based	0.88	0.23	0.23	0.23
MRA Two-Images	Word-based	0.87	0.25	0.26	0.25
	Sentence-based	0.86	0.20	0.21	0.20

4.3. Validation

We assess the performance of the models using conventional validation measures for text analysis including BLEU, ROUGE and METEOR as well as diagnostic-based measures.

BLEU [4] is a precision-based metric that counts how many *n*-grams of the generated report are in the ground-truth reference(s). BLEU has a correction to penalize reports that repeat *n*-grams. In fact, *n*-grams in system outputs cancel *n*-grams references when counted. The BLEU metric favors short outputs, and all *n*-grams are equally weighted.

ROUGE [28] is a recall-based metric that counts how many *n*-grams in references are covered by the system outputs. The most interesting version of this metric is the one that takes into account the longest sub-sequence in common between references and system outputs. The ROUGE metrics favor long system outputs, which may contain a lot of useless information. As with BLEU, all *n*-grams are weighted equally.

METEOR [27] is the harmonic mean between precision and recall over unigrams with a penalizing factor. METEOR is computed after the application of an alignment between references and system outputs. Hence, the penalizing factor is used for fragmentation, that is, a count of the unigrams which are close in the reference and far in the system output.

As diagnostic assessment measures, we use the DCS score introduced in Section 3.1, to quantify the diagnostic relevance of sentence-based and word-based generated reports to the given diagnostic tags.

[1] https://github.com/yunjey/show-attend-and-tell
[2] https://github.com/wangleihitcs/MedicalReportGeneration

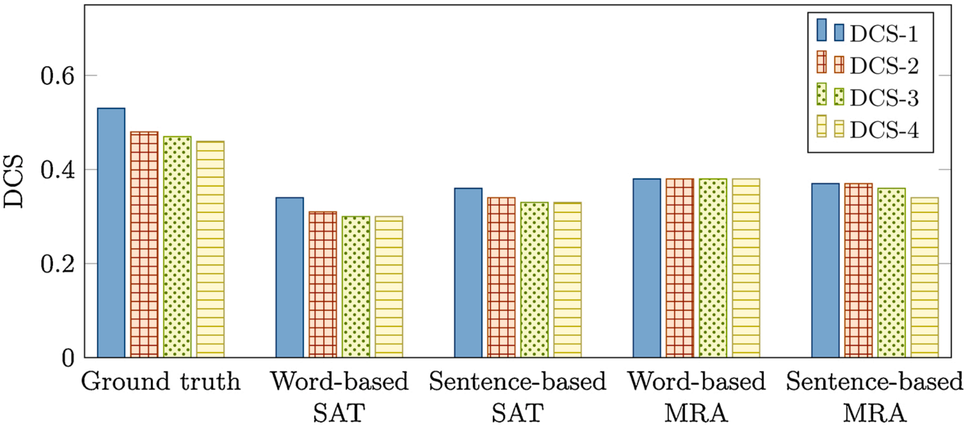


Fig. 7. Average DCS of ground truth and generated reports. Where DCS-1,2,3, and 4 are computed using 1,2,3, and 4 grams respectively.

As in the other validation metrics, we consider a range of *n*-grams (n=1,2,3, and 4). We refer to these scores as DCS-1 up to DCS-4. Note that since this score is based on tags and not on text comparison, the ground truth reports will not automatically get a score of 1. We also looked at the DCS of the ground truth reports, which represents an upper bound on the diagnostic content of generated reports. Also, we consider diagnostic metrics based on the CheXpert labeler [26,20]: accuracy and other standard classifier performance metrics are computed by comparing the predicted annotations of generated reports with the predicted annotations of the corresponding ground truth reports, where prediction is performed with the CheXpert labeler.

4.4. Results

Table 1 contains performance scores based on conventional metrics, for SAT and MRA with word- and with sentence-based dictionary. Overall results indicate similar performance of sentence-based and word-based methods. Sentence-based SAT performs better than word-based SAT when BLEU 2-4 are considered. In terms of ROUGE word-based SAT is slightly ahead but not significantly different from its sentence-based variant (p-value from Wilcoxon test >0.05). For MRA, results show similar performance of word- and sentence-based methods, except for BLEU-1, where sentence-based MRA performs better than the word-based variant. However, these differences are not significant (p-value from Wilcoxon test >0.05).

Table 2 contains performance results of word- and sentence-based SAT and MRA, computed using CheXpert labeler evaluation metrics. According to these diagnostic metrics, sentence-based SAT performs better than its word based variant, while word-based MRA performs slightly better than its sentence based variant.

Fig. 7 shows bar plots with DCS scores (for different n-grams) of ground truth test set reports, word-based generated reports, and sentence-based generated reports for SAT and MRA. Results indicate that ground truth test reports have much higher scores than generated reports for both sentence- and word-based SAT and MRA. Also, sentence-based SAT performs better than its word-based variant. Differences between MRA word- and sentence-based variants are marginal, with the word-based variant being slightly better.

We applied the Wilcoxon test to assess the significance of different DCS performance between the considered models. We find that for SAT the difference between the word-based and sentence-based models is not significant (p-value >0.2 for DCS n-gram variants).

Differences of DCS's results between word- and sentence-based MRA are not significant for DCS-1 and DCS-2 (p-values >0.1), but become significant when longer n-grams are used (p-value 1.5×10^{-2} for DCS-3 and 0.12×10^{-3} for DCS-4).

5. Discussion

We perform a qualitative and quantitative comparison of DCS and other metrics, and analyze the role of the (specific) classifier used in DCS. Next, we discuss advantages of sentence-based report generation models.

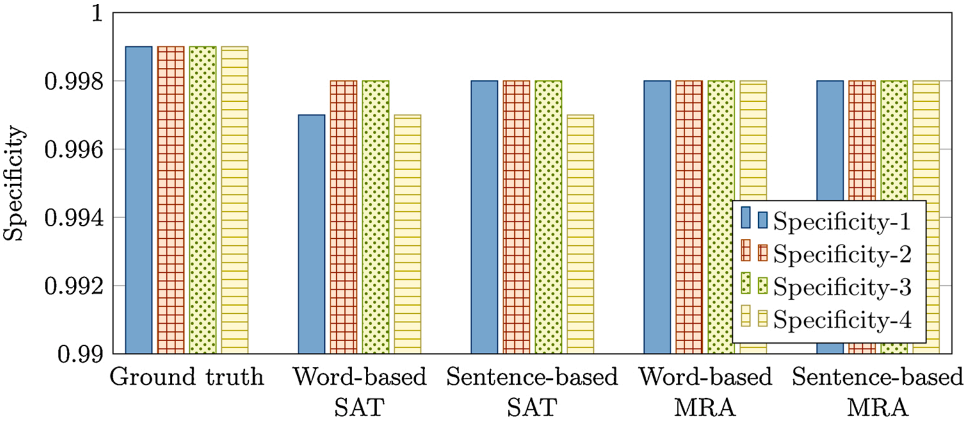


Fig. 8. Average specificity of the diagnostic tag classifier on of ground truth and generated reports. Where Specificity-1,2,3, and 4 are computed using 1,2,3, and 4 grams respectively.

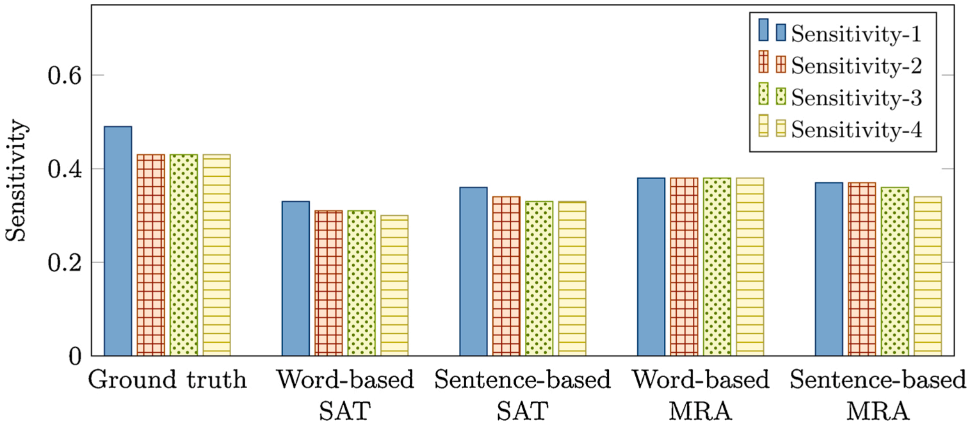


Fig. 9. Average sensitivity of the diagnostic tag classifier on ground truth and generated reports. Where Sensitivity-1,2,3, and 4 are computed using 1,2,3, and 4 grams respectively.

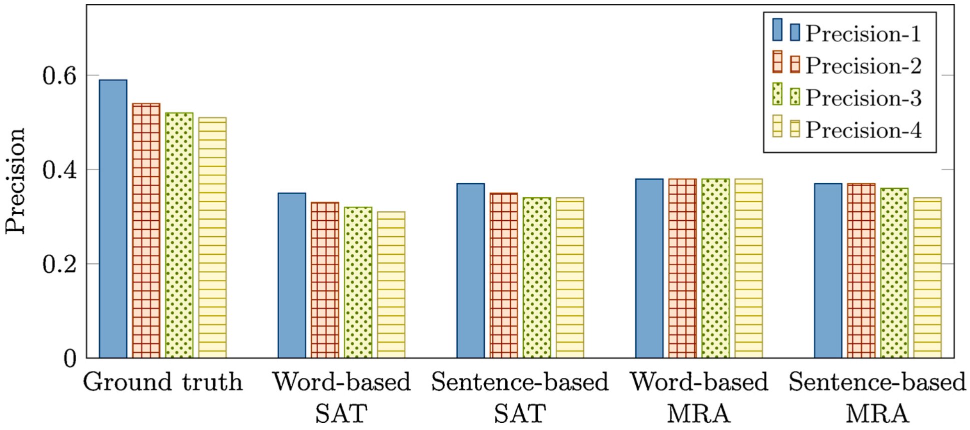


Fig. 10. Average precision of the diagnostic tag classifier on ground truth and generated reports. Where Precision-1,2,3, and 4 are computed using 1,2,3, and 4 grams respectively.

5.1. Comparison between DCS and other metrics

In Figs. 11 and 12, we see examples of word-based and sentence-based generated reports where the DCS value is 0 while BLEU values are relatively high. This phenomenon occurs because conventional metrics give the same weight to each word, even though adding or removing even a single word can completely change the overall diagnostic content of a report. On the other hand, DCS indirectly places a larger weight on the incorrect words (highlighted in italic in the Figures) since they change the meaning of the report.

In Fig. 11 we see that overall the text is tilted towards “normal” except for the words “thoracic spondylosis” which point to a specific non-normal diagnosis. This is the only difference between the generated text and the ground truth. DCS detects this difference, classifies the report as non-normal, resulting in a DCS score of 0.

A similar behavior occurs in the absence of a key sentence or phrase. In Fig. 12, we see that the sentence-based generated report is mostly in line with the ground truth reference text, which has tag “Cardiomegaly/mild, Cardiomegaly”. But since the key phrase “mild Cardiomegaly” is missing, the generated report is classified as “normal” and gets a DCS score of 0. Instead, the BLEU score cannot capture the diagnostic relevance of single parts of the report.

In the examples we also see a disadvantage of DCS: it assumes very coarse-grained values, which are often equal to 0 or 1. DCS also focuses exclusively on text that is relevant for the diagnostic tags. That means that DCS can miss many other important aspects of the generated report, such as correct grammar, sentence structure and the detailed motivation for a report’s findings.

To investigate more in depth the relation between DCS and other metrics, we have looked at their correlation. We computed Pearson’s correlation between each pair of metric values of generated test reports for word-based SAT. In Fig. 14, we see that DCS is only weakly correlated with the conventional metrics, while these metrics are highly correlated with each other. DCS is weakly correlated also with CheXpert accuracy metric, and this metric has an even lower correlation with conventional metrics. Results for sentence-based SAT and for MRA are given in Section A.

Overall, our experimental analysis indicates that neither conventional metrics nor diagnostic quality metrics like DCS or CheXpert accuracy can fully characterize the quality of a medical report. However, conventional and diagnostic-based metrics complement each other and cover different syntactic and semantic aspects of a radiology report.

Diagnostic metrics like DCS have intrinsic biases induced by the specific method employed to extract relevant keywords/tags, and by the type of classifier (Naive Bayes for DCS, rule-based for CheXpert) and how it is trained (using the original dataset for DCS, using an external

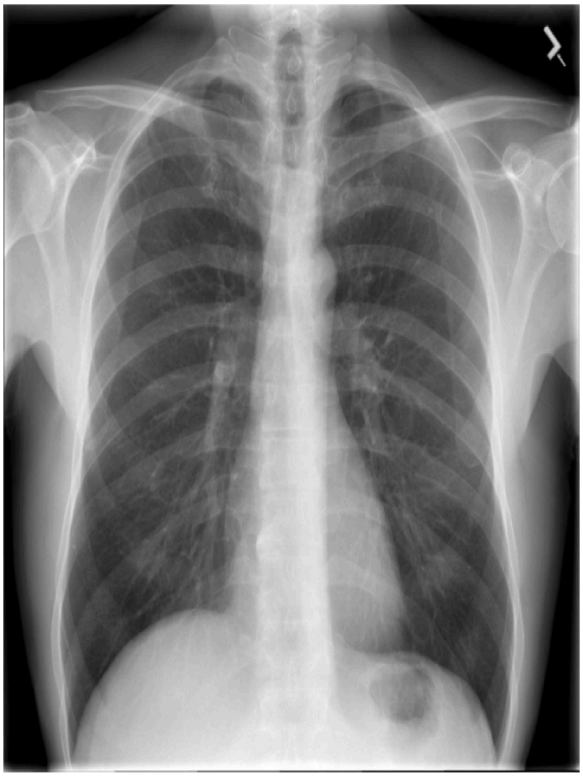


Fig. 11. Example of a generated report with high BLEU scores. The generated report has an incorrect diagnosis, marked in italic, and it gets a low DCS.

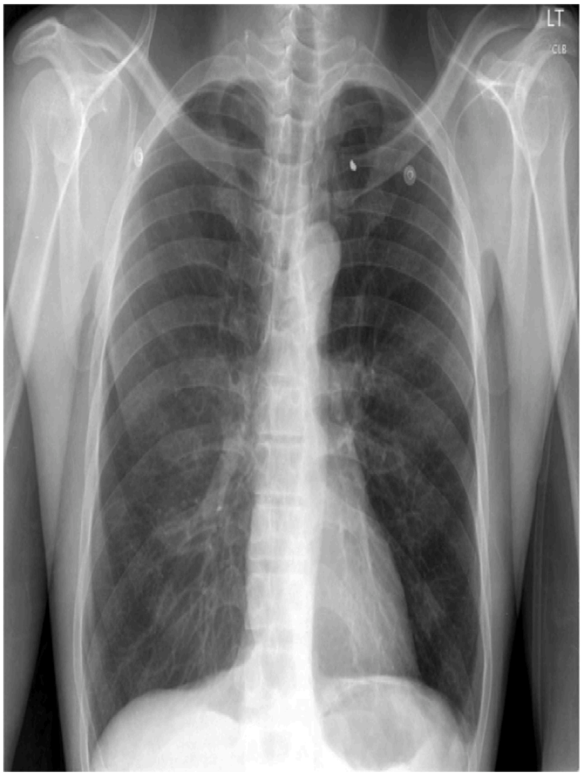


Fig. 12. Example of a sentence-based generated report with incorrect diagnostic content and low DCS, but with high BLEU scores.

dataset and class labels for CheXpert). In particular, for DCS, a low score could be due to the bad performance of the (Naive Bayes) classifier used to compute DCS values, and not necessarily to the bad performance of the report generation model. To investigate this phenomenon, we show sensitivity, specificity and precision, of the Naive Bayes classifier over the ground truth reports in the test set. Sensitivity is computed as follows: first, locally for each ground truth report in the test set, as the fraction of actual tags present in the report that are correctly predicted as such; next, as the average over all reports' sensitivities. Specificity is computed similarly, where for each ground truth report of the test set, the fraction of actual tags not present in the report that are correctly predicted as such is considered. To compute precision, for each report, the number of actual tags present in the report divided by the total number of tags labeled as belonging to that report is considered. Results are shown in Figs. 8–10. Although sensitivity and precision are overall much lower than specificity, due to the high majority of 'normal' reports, the relatively higher performance of the classifier on ground truth reports, indicate that the report generator model is mainly responsible for the classifier performance as assessed by DCS, not the classifier employed to compute the DCS score. This is also substantiated by the relative agreement of evaluation results across different score metrics.

Tag prevalence and DCS score values are related: tags with a very high prevalence get better DCS score. Tag "Normal" occurs in about 42% of the test data and yields an average DCS score of 0.88. While the rest of the tags has average DCS of 0.27. In Table A.3 we break down the DCS score for the most frequent tags: for all tags the performance of the classifier is much better on the ground truth reports compared to generated reports. In addition, the performance of the classifier on the ground truth reports is still reasonable for less frequent tags, but the classifier never produces these tags on generated reports. In a concrete case this could mean that the classifier is able to predict rare tags from the ground truth reports, while these true rare tags are not correctly predicted for the generated report. This indicates that the differences in DCS scores are mainly due to the quality of the generator, not to the classifier used to compute DCS. Fig. 13 shows one example where this happens.

In order to assess whether the type of classifier used to define DCS affects the results, we computed DCS using other classifiers. Specifically, we considered Decision Trees, Random Forest, and KNN instead of Naïve Bayes.

Results indicate similar performance trend across these classifiers. In particular, the relatively higher performance of these classifiers on ground truth reports (see Table A.4), further substantiates the central role of the report generator model, not of the classifier used to compute the DCS score.

5.2. Sentence-based and word-based generated reports

In order to analyze the impact of sentence-based dictionary on the generated reports, we compare sample reports generated by both sentence-based and word-based approaches.

Fig. 15 shows the best and worst reports generated by the sentence-based SAT model, the corresponding ground truth reports and word-based generated reports. In the best case, the report generated by the sentence-based approach exactly matches the ground truth. For this example, the word-based generated report also looks good, but is less detailed than the ground truth. The worst sentence-based SAT generated report does not include any abnormal finding mentioned in the ground truth report. In this case the word-based approach generated the same report as the sentence-based approach. When looking at the best and worst reports according to the scores for the SAT word-based model, the situation is very similar, with the best word-based generated reports exactly matching the ground truth, and the worst reports missing important parts of the text contained in the ground truth. An example is

Least Frequent Tag Example: no pulmonary consolidation fracture deformity proximal right humerus hyperinflation lungs no pulmonary consolidation xxxx opacity left base compatible xxxx atelectasis or xxxx scarring the cardiomediastinal silhouette appears unremarkable mild atherosclerotic calcification aorta prior chest surgery costophrenic xxxx clear visualized spine vertebrae appear normal in xxxx and alignment fracture deformity proximal right humerus

Diagnostic Tags: "Fractures, Bone/humerus/right", **"Deformity/humerus/right"**, "Lung/hyperdistention", "Opacity/lung/base/left", "Pulmonary Atelectasis/base/left", "Cicatrix/lung/base/left", "Atherosclerosis/aorta/mild", "Humerus", "atelectases", "deformity", "fracture", "hyperinflation lungs", "scarring", "chest surgery", "Humeral Fractures"
DCS-1: 0.22, **DCS-2:** 0.17, **DCS-3:** 0.16, **DCS-4:** 0.29

Most Frequent Tag Example: no acute findings cardiac and mediastinal contours are within normal limits the lungs are clear bony structures are intact

Diagnostic Tag: Normal
DCS: 1.0

Fig. 13. Ground truth test reports associated with a least frequent (top part) and most frequent (bottom part) tag and relative DCS's.

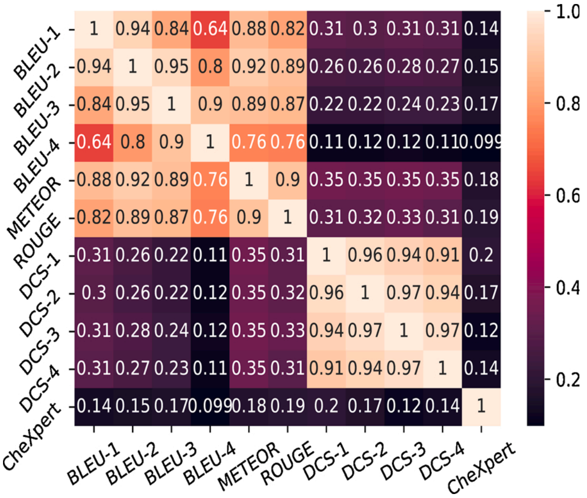


Fig. 14. Pearson correlation between DCS and other metrics for the word-based SAT method. Here, CheXpert represents F1-score computed on CheXpert based labels of generated report and ground truth report.

included in the Appendix.

An advantage of using a sentence-based dictionary is a drastic reduction in training time. We compared the running time of SAT and MRA word-based and sentence-based variants, using a GPU implementation.

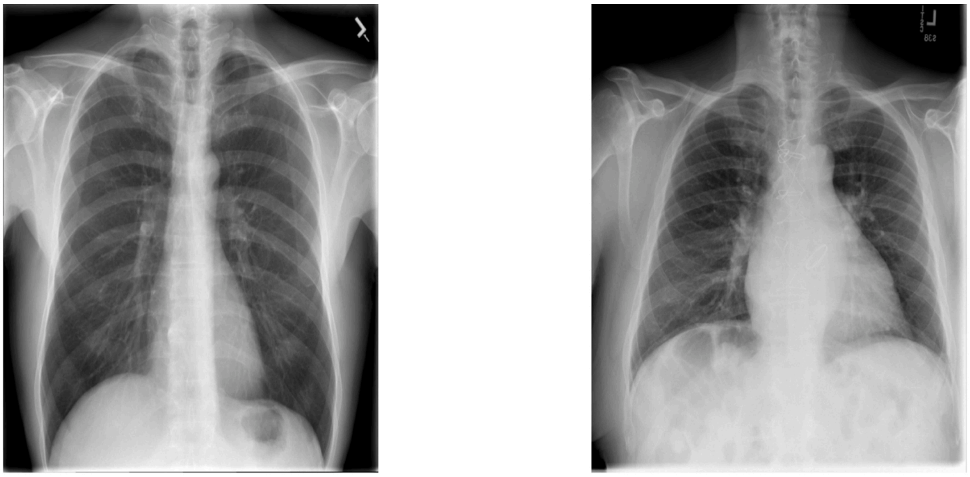
Fig. 16 (a) shows that for the sentence-based SAT model, training takes less than 0.05 seconds for a single batch (16 images) versus 0.30 seconds for the word-based model. As a result, the sentence-based model takes around 10 minutes to train, while the word-based model takes almost 1.16 hours. For MRA results in Fig. 16 (b) show that the sentence-based variant is about three times faster than the word-based one.

6. Conclusion

In summary, in this paper we investigated the downsides of using conventional text-based validation metrics as the only measure to validate automatically generated radiology reports. Results of our extensive experimental analysis indicate that conventional metrics do not capture the quality of the diagnostic content of generated reports. We have shown that DCS has a low correlation with conventional text-based metrics such as the BLEU score. These results indicate that DCS capture other properties of a report, namely diagnostic content. On the other hand, low correlation of DCS and another recent diagnostic-based validation metric, indicates that results of these type of metrics depend on the way they are computed, for instance using the original dataset with given diagnostic tags (as in DCS) or an external dataset and different tags (as in CheXpert).

Overall our investigation indicates the usefulness of both conventional and diagnostic-based metrics.

We have investigated the use of a sentence-based dictionary for radiology report generation methods and showed that a sentence-based dictionary yields results of similar quality as those obtained by using a



Diagnostic Tags: Normal

Reference: the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings. no acute cardiopulmonary findings.

Sentence-based: the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings. no acute cardiopulmonary findings.

Bleu-1: 1.0,**Bleu-2:** 1.0, **Bleu-3:** 1.0, **Bleu-4:** 1.0, **DCS:** 1.0

Word-based: heart size is normal the lungs are clear there are no focal air space consolidations no pleural effusions or pneumothoraces the hilar and mediastinal contours are normal normal pulmonary vascularity no acute abnormality

Bleu-1: 0.51, **Bleu-2:** 0.30, **Bleu-3:** 0.20, **Bleu-4:** 0.13, **DCS:** 1.0

Diagnostic Tags: Cardiomegalymild, Implanted Medical Deviceaortic valve, Spinedegenerative

Reference: the heart size is mildly enlarged. the patient is post aortic valve replacement. the xxxx sternotomy xxxx are intact. the pulmonary xxxx and mediastinum are within normal limits. there is no pleural effusion or pneumothorax. there is no focal air space opacity to suggest a pneumonia. there mild degenerative changes of the spine. no acute cardiopulmonary disease. critical result notification documented through primordial.

Sentence-based: the cardiac contours are normal. the lungs are clear. thoracic spondylosis. no acute process.

Bleu-1: 0.014,**Bleu-2:** 0.006, **Bleu-3:** 4.15×10^{-08}, **Bleu-4:** 1.16×10^{-10}, **DCS:** 0.0

Word-based: the cardiac contours are normal the lungs are clear thoracic spondylosis no acute process

Bleu-1: 0.014,**Bleu-2:** 0.006, **Bleu-3:** 4.15×10^{-08}, **Bleu-4:** 1.16×10^{-10}, **DCS:** 0.0

Fig. 15. The best (left) and worst (right) generated reports, according to BLEU-1 for the sentence-based SAT model.

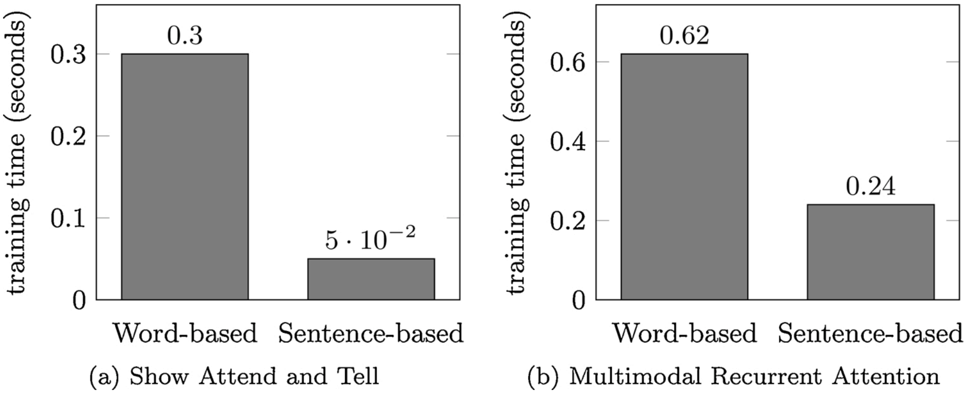


Fig. 16. Comparison of training time (in seconds) of a single batch (16 samples) for the word- and sentence-based approaches.

word-based dictionary, and improves efficiency of the training process.

A limitation of sentence-based dictionary is that it suffers from a similar drawback as [2], that is, it does not generate new sentences that have never appeared in the training set. Therefore, it relies on the assumption that the training set is sufficiently informative.

Another limitation stems from the variability between sentences. Two similar sentences will be seen as completely different in a sentence-based dictionary, and a model might thereby miss related information. Because of the standardized terminology used by radiologists, this is not a big problem in practice, as also substantiated by our experiments. However, it might be possible to further improve the sentence-based approach by using pre-processing techniques to transform sentences of the training set into an informative and diverse dictionary, or even to use generative models to create a better dictionary of sentences.

Conflict of interest

The authors declare no conflict of interest.

Appendix A. Additional results

Fig. A.17
Fig. A.18
Fig. A.19
Fig. A.20
Table A.3
Table A.4

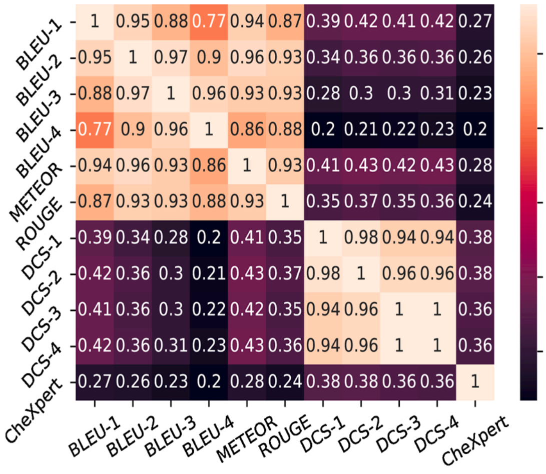


Fig. A.17. Correlation between DCS and other metrics for sentence-based SAT. Here, CheXpert represents F1-score computed on CheXpert based labels of generated report and ground truth report.

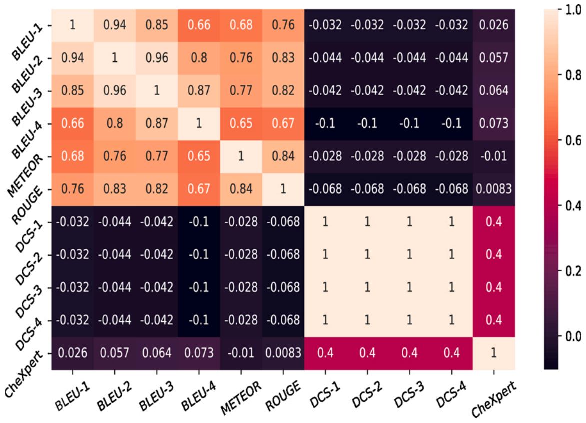


Fig. A.18. Correlation between DCS and other metrics for word-based MRA. Here, CheXpert represents F1-score computed on CheXpert based labels of generated report and ground truth report.

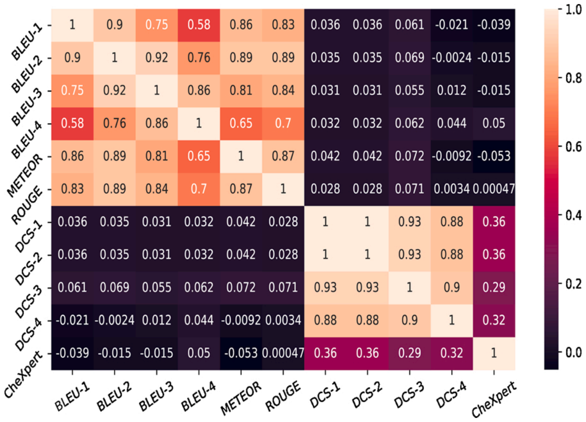
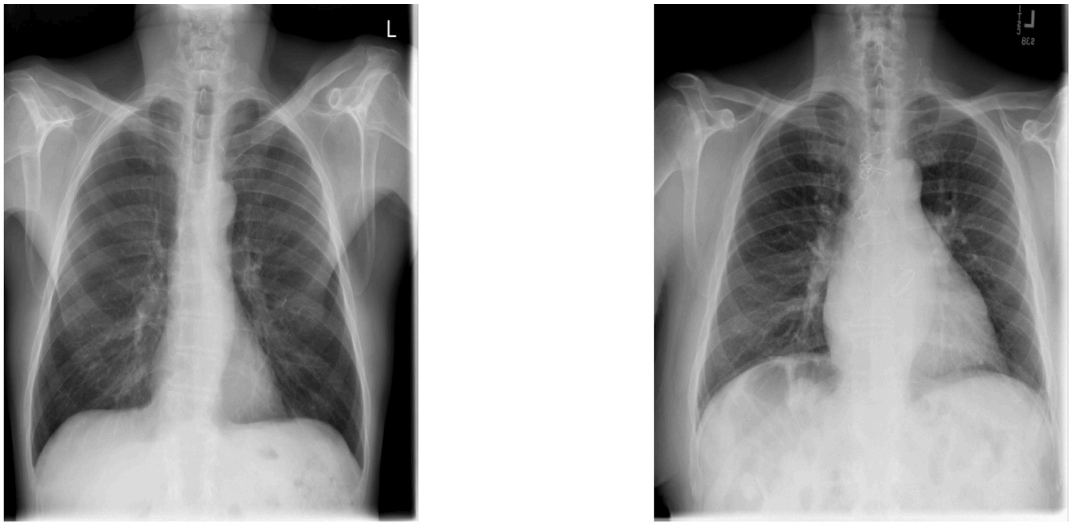


Fig. A.19. Correlation between DCS and other metrics for sentence-based MRA. Here, CheXpert represents F1-score computed on CheXpert based labels of generated report and ground truth report.



Diagnostic Tags: Normal

Reference: the heart size and pulmonary vascularity appear within normal limits. the lungs are free of focal airspace disease. no pleural effusion or pneumothorax is seen. no evidence of active disease.

Sentence-based: the heart is normal in size. the mediastinum is unremarkable. the lungs are clear. no acute disease. **Bleu-1:** 0.30,**Bleu-2:** 0.19, **Bleu-3:** 0.13, **Bleu-4:** 0.09, **DCS:** 1.0

Word-based: the heart size and pulmonary vascularity appear within normal limits the lungs are free of focal airspace disease no pleural effusion or pneumothorax is seen no evidence of active disease

Bleu-1: 1.0,**Bleu-2:** 1.0, **Bleu-3:** 1.0, **Bleu-4:** 1.0, **DCS:** 1.0

Diagnostic Tags: Cardiomegalymild, Implanted Medical Deviceaortic valve, Spinedegenerative

Reference: the heart size is mildly enlarged. the patient is post aortic valve replacement. the xxxx sternotomy xxxx are intact. the pulmonary xxxx and mediastinum are within normal limits. there is no pleural effusion or pneumothorax. there is no focal air space opacity to suggest a pneumonia. there mild degenerative changes of the spine. no acute cardiopulmonary disease. critical result notification documented through primordial.

Sentence-based: the cardiac contours are normal. the lungs are clear. thoracic spondylosis. no acute process.

Bleu-1: 0.014,**Bleu-2:** 0.006, **Bleu-3:** 4.15×10^{-08}, **Bleu-4:** 1.16×10^{-10}, **DCS:** 0.0

Word-based: the cardiac contours are normal the lungs are clear thoracic spondylosis no acute process

Bleu-1: 0.014,**Bleu-2:** 0.006, **Bleu-3:** 4.15×10^{-08}, **Bleu-4:** 1.16×10^{-10}, **DCS:** 0.0

Fig. A.20. The best (left) and worst (right) generated reports, according to the conventional metrics for the word-based model.

Table A.3
Top 10 most frequent tags and breakdown of DCS.

Tag	% of training reports	F_1 on ground truth	F_1 on word based SAT	F_1 on sentence based SAT
normal	39%	0.91	0.57	0.57
degenerative change	11.3%	0.74	0.12	0.13
opacity	10.7%	0.53	0.24	0.09
atelectases	7.9%	0.61	0.28	0.11
atelectasis	7.5%	0.58	0.15	0.11
cardiomegaly	6.4%	0.51	0.21	0.13
lung/ hypoinflation	6.4%	0.61	0.20	0.00
calcified granuloma	5.8%	0.30	0.00	0.00
lung/ hyperdistention	4.7%	0.35	0.00	0.00
scarring	4.5%	0.69	0.00	0.00

Table A.4
DCS of test reports generated using word-based SAT, where DCS is computed using random forest, decision tree, and KNN instead of Naive Bayes. These models are used with default parameters from python based Scikit-learn library.

Algo	Approach	1-gram	2-gram	3-gram	4-gram
Random Forest	Ground Truth	0.55	0.5	0.47	0.42
	Word-based	0.33	0.3	0.25	0.16
	Sentence-based	0.35	0.34	0.29	0.25
Decision Tree	Ground Truth	0.71	0.7	0.64	0.54
	Word-based	0.34	0.27	0.10	0.10
	Sentence-based	0.36	0.26	0.16	0.13
KNN	Ground Truth	0.46	0.42	0.40	0.41
	Word-based	0.36	0.37	0.36	0.37
	Sentence-based	0.35	0.36	0.34	0.34

Appendix B. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.artmed.2021.102075.

References

[1] Lukaszewicz A, Uricchio J, Gerasymchuk G. The art of the radiology report: practical and stylistic guidelines for perfecting the conveyance of imaging findings. Can Assoc Radiol J 2016;67(4):318–21.

[2] Xue Y, Xu T, Long LR, Xue Z, Antani S, Thoma GR, Huang X. Multimodal recurrent model with attention for automated radiology report generation. Proceedings of the international conference on medical image computing and computer-assisted intervention (MICCAI). Granada, Spain: Springer; 2018. p. 457–66.

[3] Kilickaya M, Erdem A, Ikizler-Cinbis N, Erdem E. Re-evaluating automatic metrics for image captioning. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics: volume 1, long papers, association for computational linguistics; 2017. p. 199–209.

[4] Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, Association for Computational Linguistics; 2002. p. 311–8.

[5] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. In: Proceedings of the ieee conference on computer vision and pattern recognition (CVPR); 2015. p. 4566–75.

[6] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), association for computational linguistics; 2014. p. 1724–34. arXiv:1406.1078.

[7] Hong Y, Kahn CE. Content analysis of reporting templates and free-text radiology reports. J Digit Imaging 2013;26(5):843–9.

[8] Ganeshan D, Duong P-AT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, Ghobadi EH, Desouches SL, Pastel D, Francis IR. Structured reporting in radiology. Acad Radiol 2018;25(1):66–73.

[9] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the international conference on machine learning (ICML); 2015. p. 2048–57.

[10] Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2015;23(2): 304–10.

[11] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In: Proceedings of the international conference on machine learning; 2014. p. 595–603.

[12] Mao J, Xu W, Yang Y, Wang J, Yuille A. Deep captioning with multimodal recurrent neural networks (m-rnn). In: Proceedings of the international conference on learning representations; 2015.

[13] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll ár P, Zitnick CL. Microsoft COCO: common objects in context. European conference on computer vision 2014:740–55.

[14] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 3156–64.

[15] Wang C, Yang H, Bartz C, Meinel C. Image captioning with deep bidirectional lstms. In: Proceedings of the ACM international conference on multimedia; 2016. p. 988–97.

[16] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 2625–34.

[17] Shin H-C, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2497–506.

[18] Wang X, Lu L, Shin H-c, Kim L, Nogues I, Yao J, Summers R. Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database. 2016. CoRR abs/1603.07965.

[19] Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers); 2018. p. 2577–86.

[20] Liu G, Hsu T-MH, McDermott M, Boag W, Weng W-H, Szolovits P, Ghassemi M. Clinically accurate chest x-ray report generation. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, Wiens J, editors. Proceedings of the 4th machine learning for healthcare conference, vol. 106 of proceedings of machine learning research. Ann Arbor, Michigan: PMLR; 2019. p. 249–69.

[21] Yuan J, Liao H, Luo R, Luo J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Proceedings of the international conference on medical image computing and computer-assisted intervention (MICCAI); 2019. p. 721–9.

[22] Li CY, Liang X, Hu Z, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33; 2019. p. 6666–73.

[23] Li CY, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. In: Proceedings of the 32nd international conference on neural information processing systems, NIPS'18; 2018. p. 1537–47.

[24] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chest x-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 2097–106.

[25] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-y, Mark RG, Horng S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6.

[26] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball RL, Shpanskaya KS, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence; 2019.

[27] Banerjee S, Lavie A. 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; 2005. p. 65–72.

[28] Lin C-Y. Rouge: a package for automatic evaluation of summaries. Text summarization branches out. 2004. p. 74–81.