Journal of Digital Imaging (2021) 34:374–384
https://doi.org/10.1007/s10278-021-00423-8

Automatic Fully-Contextualized Recommendation Extraction from Radiology Reports

Jackson Steinkamp[1] · **Charles Chambers**[1] · **Darco Lalevic**[1] · **Tessa Cook**[1]

Received: 15 April 2020 / Revised: 7 December 2020 / Accepted: 11 January 2021 / Published online: 10 February 2021

Abstract
Recommendations are a key component of radiology reports. Automatic extraction of recommendations would facilitate tasks such as recommendation tracking, quality improvement, and large-scale descriptive studies. Existing report-parsing systems are frequently limited to recommendations for follow-up imaging studies, operate at the sentence or document level rather than the individual recommendation level, and do not extract important contextualizing information. We present a neural network architecture capable of extracting fully contextualized recommendations from any type of radiology report. We identified six major "questions" necessary to capture the majority of context associated with a recommendation: recommendation, time period, reason, conditionality, strength, and negation. We developed a unified task representation by allowing questions to *refer* to answers to other questions. Our representation allows for a single system to perform named entity recognition (NER) and classification tasks. We annotated 2272 radiology reports from all specialties, imaging modalities, and multiple hospitals across our institution. We evaluated the performance of a long short-term memory (LSTM) architecture on the six-question task. The single-task LSTM model achieves a token-level performance of 89.2% at recommendation extraction, and token-level performances between 85 and 95% F1 on extracting modifying features. Our model extracts all types of recommendations, including follow-up imaging, tissue biopsies, and clinical correlation, and can operate in real time. It is feasible to extract complete contextualized recommendations of all types from arbitrary radiology reports. The approach is likely generalizable to other clinical entities referenced in radiology reports, such as radiologic findings or diagnoses.

Keywords Natural language processing · Radiology reports · Information extraction · Machine learning

Introduction

One of the most important functions of a radiology report is to convey recommendations to the ordering clinician. These recommendations include a wide variety of follow-up tasks (imaging, tissue sampling, physical exam, lab tests, subspecialty consultation) and are prevalent in reports; however, previous work has shown low adherence to follow-up recommendations [1]. Centralized systems for identifying and tracking follow-up recommendations aim to increase adherence [2] and are often implemented using structured codes or macros within reports. However, simple structured buckets and ontologies do not have the expressive power of natural language and may fail to capture the linguistic or logical complexities of follow-up recommendations (e.g., recommendation strength, chains of conditional recommendations, negations); for this reason, recommendation free-text is almost always included even alongside structured coding systems.

Another approach to improving radiologists' documentation efficiency is to build automated information extraction systems which extract and structure recommendations and their properties from free text. Previous work has shown the feasibility of complete information extraction from radiology reports, including extraction of relations and "modifier" entity properties (e.g., size of a finding, time for a recommendation) with small corpora of annotated reports[3]. There is a large body of previous work focusing on recommendation extraction in particular, with many methodologies and task definitions [1, 4–8]. Almost all work performs binary classification at the document or sentence level (e.g., does the document—or sentence—contain *at least one* recommendation?);

✉ Jackson Steinkamp
jacksonsteinkamp@gmail.com

[1] Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, USA

however, many reports and even sentences contain multiple recommendations, and an ideal tracking system would identify and track each recommendation individually. Furthermore, most extant systems focus only on *follow-up imaging* recommendations, neglecting other types of clinically useful recommendations. The majority rely on multi-step pipelines involving hand-coded rules or heuristics; although some incorporate machine learning techniques, it is usually as a part of a longer pipeline [8].

Ideal tracking systems will extract and display all of the contextual information needed to understand the recommendation, including *timing*, recommendation *strength*, *conditions* under which the recommendation should be performed, *reasons* for the recommendation, and associated radiologic *findings* which can be linked to the recommendation and tracked over time. While many systems extract a subset of these pieces of information, to the authors' knowledge, none extract all of it. In this study, we present an end-to-end deep learning architecture for complete, fully contextualized recommendation extraction of *all types* of clinical recommendations found in radiology reports.

Materials and Methods

Dataset

Our study was retrospective, performed with data collected for routine clinical care, and was approved by our institution's IRB. Our dataset was composed of a random sample of 2272 reports from our institution's entire report database, including adult patients of all ages and genders. Our reports were drawn from all subspecialties of radiology, including interventional radiology, from multiple hospitals within a single academic institution, and included a variety of report templates, including completely free-text.

Task Definition

Similar to prior work on generalized information extraction, we opted to conceive of the recommendation task as a set of hierarchical "questions" to be answered [3] which are sufficient to provide a fully contextualized recommendation to the clinician end user, or serve as an engine for automated descriptive analyses. The top-level question is "What are all of the recommendations in this document?" For each answer to this question (i.e., for each recommendation), there exist downstream questions, e.g., "when should *this recommendation* be done?" or "what is the strength of *this recommendation*?," which refer back to the recommendation in question.

Each question can also be framed as one of two types. The first question type represents a named entity recognition (NER) task, in which each token in the document must be labeled as part of an answer to the question or not, and the output is a set of contiguous text spans from the document corresponding to the answers to the question. The second question type represents a classification task, in which the system's output is a single categorical prediction (e.g., is this recommendation "strong" or "weak"). In the latter question type, evidence from the text is still helpful in interpreting the machine's prediction (e.g., the system can point to the word "consider" in "consider follow-up imaging" as evidence for a weak recommendation rather than a strong one); precedent for this exists in the literature [9, 10]. Note that the named entity recognition task is strictly harder than the document or sentence classification tasks used for recommendation extraction in previous work, as it extracts the exact text span corresponding to the recommendation or modifying property, rather than merely the sentence or document containing it. Furthermore, there may be multiple acceptable boundaries for the beginning or end of an answer (e.g., a recommendation or reason); the exact boundaries are inherently somewhat subjective.

A list of six questions was developed through iterative examination of recommendations within reports. After viewing approximately 250 reports with recommendations, we were satisfied that we had captured the majority of relevant contextual factors and modifiers. These properties are listed in Table 1.

For this study, we opted not to perform coreference resolution (CR) [11], i.e., linking multiple mentions of the same recommendation within a document. Coreference is

Table 1 List of questions answered by our system

Information subtask	Referents	Type
1. What are all of the recommendations in this report?	None	Named entity recognition
2. What is the desired time period(s) for this recommendation?	Recommendation	Named entity recognition
3. What are the stated reasons for this recommendation?	Recommendation	Named entity recognition
4. Under what conditions should this recommendation be performed?	Recommendation	Named entity recognition
5. What is the strength of this recommendation?	Recommendation	Classification ("strong" or "weak")
6. Is the recommendation explicitly negated?	Recommendation	Classification ("yes" or "no")

a common phenomenon in radiology reports; frequently, a radiologist will list a recommendation after stating an identified finding within the main body of the report, and state it again (often with different wording) in the report impression or conclusion. Furthermore, when a trainee's report is read by an attending physician supervisor, the same recommendation may be stated multiple times. Although automated resolution of coreferences is useful for a downstream clinical user (to reduce duplicated information and aggregate information between mentions of the same entity), it represents a difficult NLP problem which often requires large datasets to address appropriately. For further details on specific questions and their annotation, see Appendix 1.

Annotation

Documents were tokenized using the spaCy Python package. We annotated the 2272 reports at the token level using a custom-built labeling application. For each question, annotations consisted of a set of contiguous text spans from the document (starting token and ending token). For classification questions, the annotation also included the target output class (e.g., strong recommendation vs. weak recommendation). Each text-span answer to the top-level question (i.e., each recommendation) required annotations to be made for each of the downstream modifier questions. Further details and screenshots of the labeling application are given in Appendix 2.

Models

We aimed to build a unified task representation allowing for a single model architecture to answer all of the above questions. To do this, we designed the neural network models to take three distinct inputs: (1) the radiology report itself, in the form of continuous word vectors; (2) a single integer representing the question being asked; and (3) a vector representing which tokens in the document are being asked *about* (i.e., a vector of referents). For instance, in the question "What is the time period for this recommendation," which refers back to the recommendation "follow-up chest x-ray," the referent vector would consist of all zeros, except for the tokens "follow-up chest x-ray," which would be represented by *ones*. The question index would be "2" as shown in Table 1. The network contains additional embedding layers for both the question index and the referent vector, enabling it to learn how to use this additional contextual information to provide the answer to the question. To extend this architecture to additional questions, or questions with multiple referents, one would simply add additional question and referent indices.

The model produces two outputs—a NER output corresponding to the label of each word token, and a classification output corresponding to the categorical answer to the question (e.g., "weak" vs. "strong"). For questions which do not have categorical answers and only have NER components, the network's classification output is disregarded. For questions with categorical answers, the total network loss is obtained by summing the loss from the token-level prediction (averaged over all tokens) and the classification prediction.

We evaluated a simple long short-term memory (LSTM) architecture [12] on the task. We used a combination of custom-trained fastText vectors, trained on our institution's entire repository of radiology reports, with Global Vectors [13] trained on the Common Crawl dataset.

We compared an LSTM model with *a single* set of weights for all questions (i.e., a multi-task learning paradigm) vs. an LSTM model with different weights for each question (where each sub-network independently learns to solve a single question). Details of the full model are described in Appendix 3.

Training and Validation

For each question, the dataset was split into training, validation, and test sets with an 80%/10%/10% split of question instances. Models were trained using an Adam optimizer with a learning rate of 1e−4 until validation loss stopped decreasing, with a patience of two epochs. All models were trained on a machine with a single GPU, in less than 2 hours.

To provide a more complete picture of model performance, the recall, precision, and F1 score (the harmonic mean of precision and recall) are calculated using three different criteria, following the convention of 2013 SemEval Task 9 [14]. These are as follows: (1) token-level metrics (e.g., at the level of the prediction for each word token); (2) entity-level exact match, a strict condition where only perfect matches between predicted and gold standard text spans are counted as true positives; and (3) partial match, a more generous condition which counts partial text span matches (e.g., a prediction of "MRI" when the gold-standard annotation is "MRI examination") as true positives. For each metric, confidence intervals for precision and recall are calculated using the Wilson score for binomial proportions.

Results

Dataset

Table 2 shows a breakdown of the named entity recognition annotations. Question 1 is asked once of each report document, whereas questions 2–6 are asked of each recommendation (i.e., each answer to question 1). The

Table 2 Summary statistics

Question	Number of question instances	Number of answers	No. of unique answers
1. What are all of the recommendations in this report?	2,272	1820	899
2. What is the desired time period(s) for this recommendation?	1,820	796	224
3. What are the stated reasons for this recommendation?	1,820	2948	1467
4. Under what conditions should this recommendation be performed?	1,820	428	196
5. What is the strength of this recommendation?	1,820	1612	207
6. Is the recommendation explicitly negated?	1,820	108	10

"number of text span answers" column represents the number of contiguous text spans which serve as an answer to that question (for instance, if a report has two recommendations in it, then that adds 2 to the overall total for question 1). The "number of unique text span answers" is designed to capture the variability of the ways in which recommendations are talked about; for instance, there are very few unique ways in our corpus to negate a recommendation, but a huge variety of unique text spans corresponding to "reasons" for recommendations, with multiple answers per recommendation.

Our dataset contained 1820 recommendations in total, with 899 *unique* recommendations. The five most common recommendations were "screening mammogram" (89), "further evaluation" (57), "clinical correlation" (55), "normal interval follow-up" (55), and "MRI" (32). There were a large variety of unique recommendations in our dataset, suggesting that simple rule- or ontology-based algorithms for extracting them would be insufficient for capturing the full variety. Many reports which contain at least one recommendation contain *more than one* recommendation (e.g., a mammogram which gives a separate recommendation for each breast, or a restatement of the same recommendation by an attending physician and a resident).

Some modifier questions, such as negation, had very stereotyped patterns of answers ("no," "do not," "does not meet criteria"), while others, such as "reason for recommendation," were significantly more varied in representation.

For questions 5 and 6, which have categorical answer sets, the breakdown was as follows. A total of 108 recommendations were negated, while 1712 were positive recommendation statements. Exactly 1295 recommendations were denoted as "strong," while 565 were denoted as "weak" recommendations.

We also examined the types of recommendations made across the corpus. The majority of recommendations fell into one of the following seven categories: correlation with clinical history or physical exam, subsequent imaging study, comparison to existing imaging study, laboratory measurement, other diagnostic study (including tissue sampling, colonoscopy), consultation of a specific clinical service, and treatment/therapeutic action (e.g., "incision and drainage," "diuresis"). We added an eighth category for vague recommendations which did not specify a particular action (e.g. "Continued follow-up," "appropriate management"). Total counts of each recommendation type are given in Table 3.

Differences Across Modalities and Specialties

Our corpus consisted of 1,124 x-ray studies, 449 CT studies, 245 MR studies, 357 ultrasound studies, and 88

Table 3 Breakdown of recommendations and reports by study type

	X-ray	CT	MR	Ultrasound	Other
Total studies	1124	449	245	357	97
Total recommendations	621	563	197	381	58
# Recommendations: Correlate with History/Exam	71	70	29	31	4
# Recommendations: Additional Imaging Study	485	391	115	257	35
# Recommendations: Compare with Existing Imaging	21	12	6	5	5
# Recommendations: Laboratory Test	1	9	1	20	0
# Recommendations: Other Diagnostic Study	13	34	10	44	5
# Recommendations: Consultation	4	5	4	4	1
# Recommendations: Therapeutic Action	13	4	0	7	4
# Recommendations: Vague	13	38	32	13	4

"other" studies (including nuclear medicine and other miscellaneous studies). In total, the X-ray reports contained 621 recommendations (of any type, not just follow-up imaging), the CT reports contained 563 recommendations, the MR studies contained 197, the ultrasounds contained 381, and the "other" studies contained 58 recommendations. Table 3 shows a detailed breakdown of recommendation types by study type.

Model Performance

Named Entity Recognition

Performance of the single-task and multi-task paradigm LSTM models on named entity recognition is given below, in Table 4. The single-task model is able to achieve good performance on the majority of tasks, including relatively complex tasks such as "reasons for recommendation." Perhaps unsurprisingly, tasks with longer text spans and more subjective "beginning" and "end" boundaries (e.g., question 3) are more difficult for the network within the "exact match" metrics. In particular, answers to question 3 often take the form of multiple sentences describing a single finding which motivated a recommendation. The multi-task model tends not to perform as well as the single-task model on the majority of tasks.

Classification

Table 5 demonstrates the LSTM model's performance on questions which include categorical classification of a recommendation. On our test dataset, the models were able to achieve near-perfect performance at these tasks.

Real-Time Application

Our system can be used in real-time to extract contextualized recommendations from free text. First, we "ask" the model to answer the top-level recommendation extraction question by passing as arguments the index of the top-level question, and a referent vector consisting of all zeros, along with the report text. For each "answer" text span predicted by the model, we create a new set

Table 4 Model performance on each question

Question	Performance (single-task model, exact match)	Performance (single-task model, partial match)	Performance (single-task model, token-level)	Performance (multi-task model, token-level)
1. What are all of the recommendations in this report?	Recall: 70.7% (63.7–77.6)% Precision: 69.0% (62.0–76.0)% F1: 69.9%	Recall: 91.0% (86.4–95.6)% Precision: 88.9% (84.0–93.8)% F1: 89.9%	Recall: 87.1% (84.5–89.7)% Precision: 91.3% (89.0–93.6)% F1: 89.2%	Recall: 85.8% (83.1–88.3%) Precision: 73.2% (70.2–76.2%) F1: 79.0%
2. What is the desired time period(s) for this recommendation?	Recall: 83.3% (75.4–91.2)% Precision: 77.3% (68.9–85.8)% F1: 80.2%	Recall: 91.1% (84.7–97.5)% Precision: 84.5% (77.0–92.0)% F1: 87.7%	Recall: 84.5% (80.3–88.6)% Precision: 88.8% (85.0–92.5)% F1: 86.6%	Recall: 82.6% (77.7–87.4)% Precision: 82.6% (77.7–87.4)% F1: 82.6%
3. What are the stated reasons for this recommendation?	Recall: 43.8% (38.1–49.5)% Precision: 44.0 (38.3–49.7)% F1: 43.9%	Recall: 88.4% (84.5–92.1)% Precision: 88.7% (84.9–92.4)% F1: 88.5%	Recall: 90.0% (88.9–90.9)% Precision: 82.9% (81.6–84.1)% F1: 86.3%	Recall: 70.2% (68.8–71.6)% Precision: 83.4% (82.2–84.7)% F1: 76.3%
4. Under what conditions should this recommendation be performed?	Recall: 82.2% (70.4–94.0)% Precision: 77.1% (64–8-89.3)% F1: 79.6%	Recall: 88.9% (78.7–99.1%) Precision: 83.3% (72.2–94.5)% F1: 86.0%	Recall: 90.4% (86.5–94.2)% Precision: 90.8% (87.0–94.6)% F1: 90.5%	Recall: 93.3% (90.5–96.2)% Precision: 91.0% (87.8–94.2)% F1: 92.2%
5. What is the strength of this recommendation? (textual evidence)	Recall: 87.5% (57.5–100)% Precision: 77.8% (47.6–100)% F1: 82.3%	Recall: 100% (74.2–100)% Precision: 88.9% (61.3–100)% F1: 94.1%	Recall: 90.9% (67.3–100%) Precision: 90.9% (67.3–100)% F1: 90.9%	Recall: 87.9% (84.4–91.4)% Precision: 94.1% (91.5–96.8)% F1: 90.9%
6. Is the recommendation explicitly negated? (textual evidence)	Recall: 82.2% (76.0–88.2)% Precision: 82.7% (76.6–88.7)% F1: 82.4%	Recall: 93.6% (89.5–97.7)% Precision: 94.2% (90.2–98.2)% F1: 93.9%	Recall: 93.6% (90.7–96.6)% Precision: 86.3% (82.4–90.2)% F1: 89.8%	Recall: 75.0% (64.8–85.2)% Precision: 100% (95.1–100%) F1: 85.7%

Table 5 Model performance on categorical classification

Question	F1 score (classification, single-task model)	F1 score (classification, multi-task model)
5. What is the strength of this recommendation? (strong/weak)	Recall (weak): 100% (94.2–100)% Precision (weak): 93.8% (85.4–100)% F1 (weak): 96.8%	Recall (weak): 99.2% (96.7–100)% Precision (weak): 97.7% (94.6–100)% F1 (weak): 98.5%
6. Is the recommendation explicitly negated? (yes/no)	Recall (negated): 100% (74.2–100)% Precision (negated): 100% (74.2–100)% F1 (negated): 100%	Recall (negated): 100% (74.2–100)% Precision (negated): 100% (74.2–100)% F1 (negated): 100%

of downstream modifier questions to be answered. The system is capable of operating in real time (as a user is typing) or in batch mode (for large scale analysis of millions of radiology reports).

Discussion

In summary, we define a set of tasks which, together, provide complete contextualized free-text recommendation extraction from arbitrary radiology reports. We build a dataset consisting of 2272 randomly sampled reports including a wide variety of patients, report templates, indications, study types, and radiologic subdisciplines. Our task extracts the exact text corresponding to recommendations and recommendation modifiers, which represents an improvement over document-level or sentence-level classification and extracts all relevant associated contextual factors (reason, negation, conditional, time period). The extracted information can be used in clinical tracking systems, population health applications, and automated descriptive studies of large volumes of radiology reports.

Our unifying question-answering framework enables a single representation of each of the questions to be passed to the neural network, rather than requiring multiple types of networks arranged in a pipeline. For downstream "modifier" questions, explicit representations of the referent entities (in this case, the recommendations) in the form of referent vectors were sufficient to teach the network to focus on the context surrounding the specific recommendation, even though it had access to the full document if necessary. Our framework is extensible to arbitrary sets of questions and referents, allowing for simple addition of new questions. In particular, we plan to build a similar system for extraction of radiologic findings and link findings to recommendations with similar referent-based questions, enabling a finding-level summary of a patient's radiologic history.

The model does not perform as well on the "exact" match metric for some questions, notably questions 1 and 3, the answers to which often consist of multiple sentences or sentence parts and may even have multiple reasonable correct answers. The exact match metric imposes an "all or nothing" standard on the model, where if it predicts the phrase "head MRI" as the answer instead of "repeat head MRI," or "Recommend correlation with physical exam findings" instead of "correlation with physical exam findings," it is counted as a 0% for that example. In these cases, the token-level match metric provides a better indicator of the model's performance, because it is capable of measuring partial or near-perfect matches. In the case of questions 1 and 3, the token-level metric indicates that the model is identifying the majority of tokens involved in the question answer, even for these more complex or subjective questions.

We compared the single-task paradigm LSTM to the multi-task LSTM to evaluate whether the model would benefit from learning each question separately vs. jointly learning all questions simultaneously. From this study, although the multi-task models were able to learn features associated with each question and answer differently depending on which question passed to the network, the multi-task network does not provide significant performance benefits over the single-task paradigm on the majority of questions. However, there are a wide variety of paradigms and approaches for multi-task deep learning; another approach might prove more fruitful [15].

Machine learning algorithms are capable of combing through large volumes of data and identifying cases which merit close human scrutiny much faster than a human user. These strengths complement the strengths of the human user, who has fluid general knowledge and the ability to handle situations outside of the machine learning system's training distribution. For the foreseeable future, such applications will still require human expertise to handle out-of-distribution or unexpected cases; however, clinicians' expertise can be put to its best use in combination with automated screening systems. Particularly, if the alternative to a machine-assisted system is no standardized tracking system whatsoever, as is often the case, an imperfect system is certainly preferable.

Tools such as this have potential not only for point-of-care or population-based abnormality tracking systems but also for large-scale automated descriptive studies [16, 17], which can provide deep analysis of existing large-scale practice patterns and motivate quality and safety interventions. Many of these analyses are currently done manually, by having human clinicians spend hours reading or skimming thousands of documents, but machine learning-based information

extraction systems have great potential to increase the ease of such analyses. A system such as ours, which is capable of analyses on the individual recommendation level, rather than the sentence or document level, is required for the deepest understanding of clinical entities such as recommendations. In particular, the inclusion of negations and conditional statements is *critical* for any such study to accurately describe the statistics of the document corpus. In future work, we plan on conducting large-scale descriptive analyses to examine how recommendations vary with imaging modality and radiology specialty.

One limitation of our work is that it has only been validated on our institution's reports. Although our dataset includes reports from many different radiologists, other institutions may have different documentation practices, so our models may not generalize perfectly. Another limitation is the significant up-front human cost to label the data. Finding ways to embed annotation naturally into existing radiologist workflow, to create datasets which can ultimately be used to produce tools for radiologists themselves, remains a useful endeavor. Given the limited amount of common "questions" one might ask of a radiology report, creating large collaborative cross-institutional annotated datasets could reasonably be created for the majority of common use cases. Alternatively, federated ML approaches can be used; in the federated paradigm, slices of a complete dataset are stored in different locations, training is performed locally, and model weight updates are pushed from the local sites to a central copy of the model [18]. Either approach would enable all institutions, including those without significant research departments or specific vendor products, to benefit from such tools.

Conclusion

We have demonstrated the feasibility of complete, fully contextualized recommendation extraction from all types of radiology reports, including all types of recommendations (e.g., clinical correlation, follow-up imaging, tissue biopsy, and others). Such tools may ultimately prove useful in a wide variety of clinical, research, and quality improvement applications.

Appendix 1. Annotation Protocol for Specific Questions

Question 1: What are all of the Recommendations in this Report?

To fully capture the diversity of recommendations in reports, as well as to enable the maximal variety of future descriptive analyses or clinical tracking systems, we defined recommendations as broadly as possible. Recommendations were defined as any span of text indicating that something should or should not be done or considered by the ordering clinician. This includes recommendations for imaging studies, lab tests, tissue samples, specialty referrals, correlation with physical exam findings, comparison with prior imaging, and nonspecific statements of "clinical correlation." Statements that "No further follow-up is required" are annotated as negated recommendations; i.e., "follow-up" is tagged as the text span answer to the top-level question, and "no" is the classification answer to the downstream question "Is this recommendation negated?" Statements that "XXX may be considered for better visualization" were included as recommendations.

Question 2: What is the Desired Time Period for this Recommendation?

Any statement indicating a time period, whether exact, relative, or vague, was annotated as a text-span answer to this question (e.g., "Screening mammogram in 12 months," "short interval follow-up," "nonemergent CT." "MRI when patient's clinical condition permits").

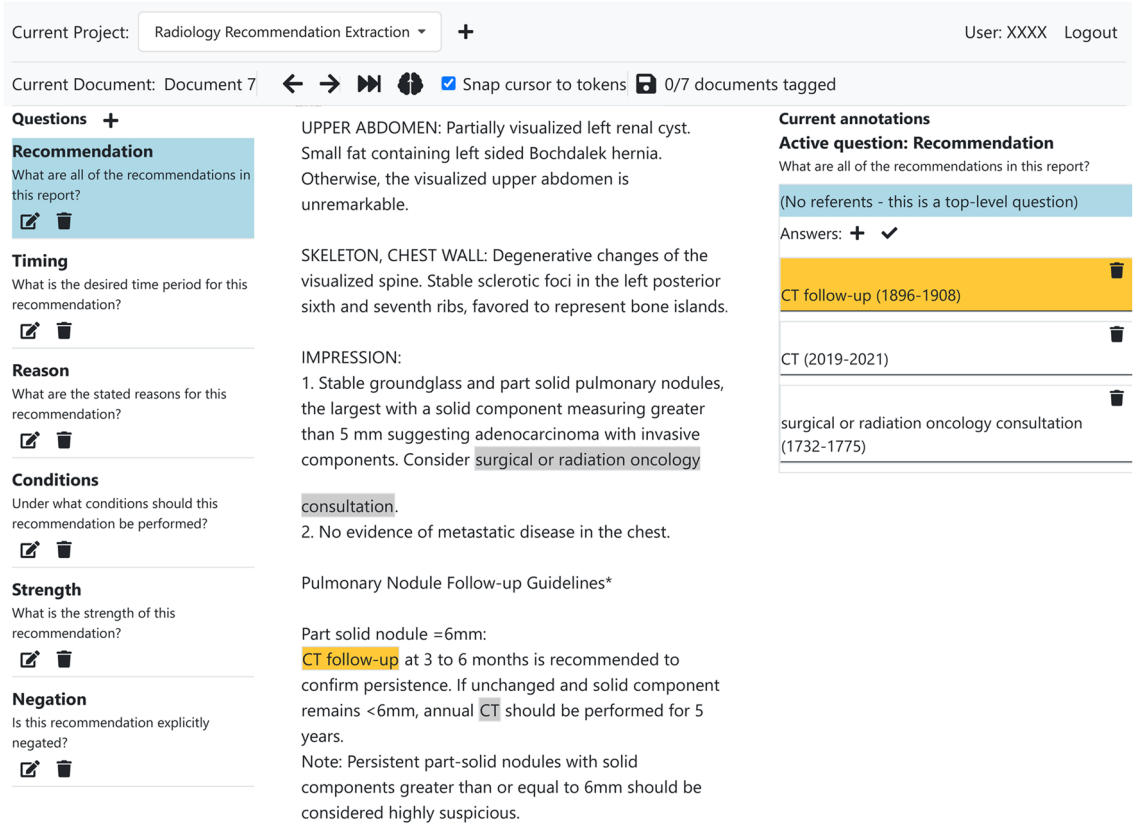
Question 3: What are the Stated Reasons for this Recommendation?

Text spans were annotated as answers to Question 3 if they provided context for the recommendation regarding its purpose, reason, or goal. This includes broad indication categories ("screening mammogram," "follow-up imaging," "surveillance"), specific stated goals for follow-up ("to rule out malignancy," "to ensure resolution of consolidation"), rationale or justification statements ("given stability over the past 12 months"), and findings which clearly motivated the recommendation ("'Indeterminate left adrenal mass may represent atypical adenoma and should be characterized. Recommend…"). For negated recommendations, reasons for *not* performing the follow-up were annotated. This was designed to be used in downstream systems which link recommendations to findings over time, with the understanding that it is somewhat more difficult than the other questions to define exactly.

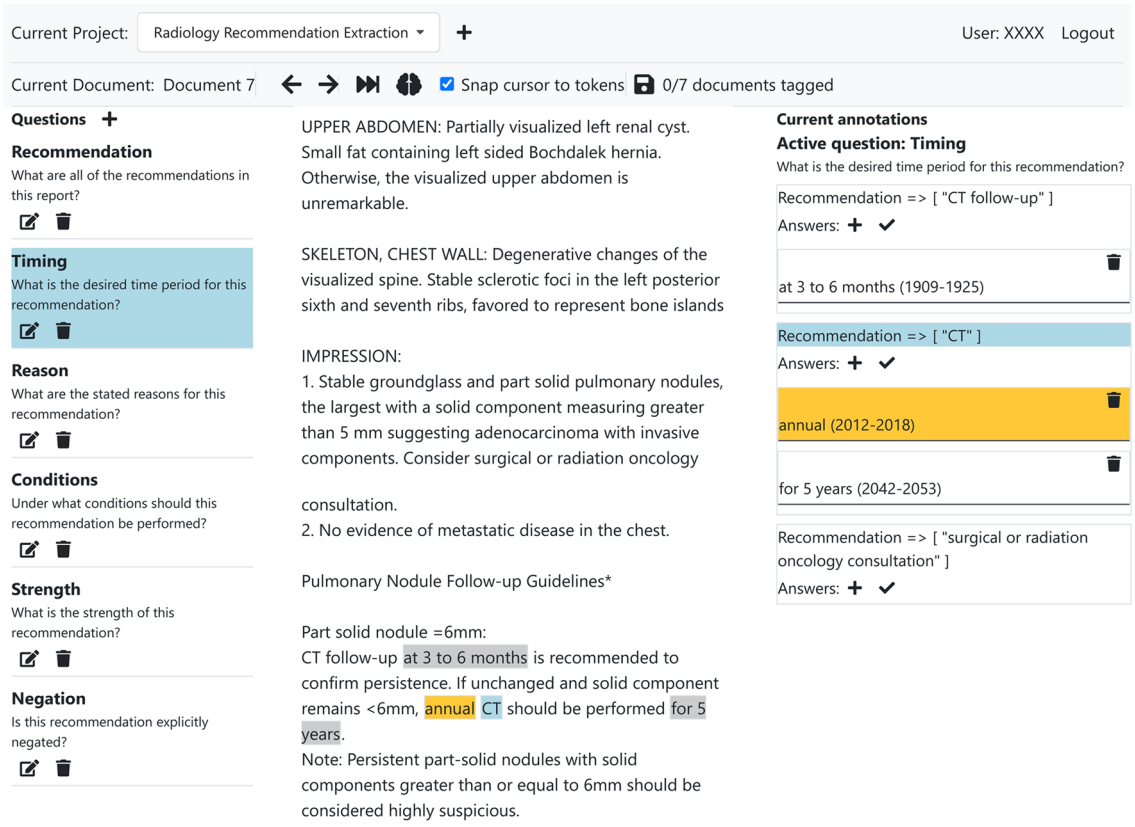
Question 4: Under What Conditions Should this Recommendation be Performed?

Many recommendations include conditions which should be met before the recommendation is done (e.g., "if the

(a)


Current Project: Radiology Recommendation Extraction
User: XXXX Logout
Current Document: Document 7
Snap cursor to tokens
0/7 documents tagged
Questions
Recommendation
What are all of the recommendations in this report?
Timing
What is the desired time period for this recommendation?
Reason
What are the stated reasons for this recommendation?
Conditions
Under what conditions should this recommendation be performed?
Strength
What is the strength of this recommendation?
Negation
Is this recommendation explicitly negated?
UPPER ABDOMEN: Partially visualized left renal cyst. Small fat containing left sided Bochdalek hernia. Otherwise, the visualized upper abdomen is unremarkable.
SKELETON, CHEST WALL: Degenerative changes of the visualized spine. Stable sclerotic foci in the left posterior sixth and seventh ribs, favored to represent bone islands.
IMPRESSION:
1. Stable groundglass and part solid pulmonary nodules, the largest with a solid component measuring greater than 5 mm suggesting adenocarcinoma with invasive components. Consider surgical or radiation oncology consultation.
2. No evidence of metastatic disease in the chest.
Pulmonary Nodule Follow-up Guidelines*
Part solid nodule =6mm:
CT follow-up at 3 to 6 months is recommended to confirm persistence. If unchanged and solid component remains <6mm, annual CT should be performed for 5 years.
Note: Persistent part-solid nodules with solid components greater than or equal to 6mm should be considered highly suspicious.
Current annotations
Active question: Recommendation
What are all of the recommendations in this report?
(No referents - this is a top-level question)
Answers:
CT follow-up (1896-1908)
CT (2019-2021)
surgical or radiation oncology consultation (1732-1775)

(b)


Current Project: Radiology Recommendation Extraction
User: XXXX Logout
Current Document: Document 7
Snap cursor to tokens
0/7 documents tagged
Questions
Recommendation
What are all of the recommendations in this report?
Timing
What is the desired time period for this recommendation?
Reason
What are the stated reasons for this recommendation?
Conditions
Under what conditions should this recommendation be performed?
Strength
What is the strength of this recommendation?
Negation
Is this recommendation explicitly negated?
UPPER ABDOMEN: Partially visualized left renal cyst. Small fat containing left sided Bochdalek hernia. Otherwise, the visualized upper abdomen is unremarkable.
SKELETON, CHEST WALL: Degenerative changes of the visualized spine. Stable sclerotic foci in the left posterior sixth and seventh ribs, favored to represent bone islands
IMPRESSION:
1. Stable groundglass and part solid pulmonary nodules, the largest with a solid component measuring greater than 5 mm suggesting adenocarcinoma with invasive components. Consider surgical or radiation oncology consultation.
2. No evidence of metastatic disease in the chest.
Pulmonary Nodule Follow-up Guidelines*
Part solid nodule =6mm:
CT follow-up at 3 to 6 months is recommended to confirm persistence. If unchanged and solid component remains <6mm, annual CT should be performed for 5 years.
Note: Persistent part-solid nodules with solid components greater than or equal to 6mm should be considered highly suspicious.
Current annotations
Active question: Timing
What is the desired time period for this recommendation?
Recommendation => ["CT follow-up"]
Answers:
at 3 to 6 months (1909-1925)
Recommendation => ["CT"]
Answers:
annual (2012-2018)
for 5 years (2042-2053)
Recommendation => ["surgical or radiation oncology consultation"]
Answers:

◂**Fig. 1** **(a)** The custom web interface used to annotate reports. This shows a user annotating a top-level question (question 1) for an anonymized radiology report. The user uses the left-hand column to select the question to annotate answers for and uses the center column to highlight and tag spans of text as answers to that particular question. In this document, there are three separate recommendations, corresponding to three answers to the "recommendation" question. (**b)** In this screenshot, the user is annotating question 2 ("What is the desired time period for this recommendation?") for each of the three previously identified recommendations in this report. The right-hand column is used to select which follow-up recommendation (i.e. which answer to question 1) to annotate. Note that downstream questions such as question 2 may also have multiple answers, as in this case –"CT" should be performed "annually" and "for 5 years." Both of these text spans provide temporal context for the recommendation

patient is at increased risk for pulmonary metastatic disease," "if further evaluation of this finding is desired," "if clinically indicated," "otherwise"). Any such condition was tagged as a free-text answer to this question.

Question 5: What is the Strength of this Recommendation?

While this can be a category of dispute even among radiologists, we opted to use a few simple heuristics for annotation. Statements such as "is recommended," "is advised," or imperative directives ("Follow up in 6 months") were annotated as strong recommendations, whereas those which "suggested" asked ordering clinicians to "consider" a recommendation or posited that a follow-up study "may be useful" were annotated as weak recommendations. Similarly, recommendations which merely stated that a follow-up study has the potential to better distinguish differential diagnoses were annotated as weak recommendations.

Question 6: Is this Recommendation Explicitly Negated?

Many recommendations are actually statements *not* to do something, or that *nothing* is necessary to be done. If a recommendation was of this type (e.g., "No further follow-up is required," "Does not meet criteria for follow-up"), the recommendation was annotated as *negated*, and the negation phrase (e.g., "no," "nothing," "does not") was tagged as a free-text rationale for the classification.

Appendix 2. Custom Labeling Application

Our custom web application for report annotation is demonstrated below. During annotation time, it was hosted behind our institution's firewall, accessible only on the institution's internal network. The application was hosted using Python's Flask web server software, with a MongoDB database for storing documents and annotations. The front-end client-side interface was written using Vue.js. The web application allowed for the creation of annotation projects, each of which has their own set of "questions." Documents were loaded into the application and tokenized using SpaCy. The interface enables users to navigate between different questions and annotate documents at the token or character-level. In the case of this project, annotations were performed at the token level.

Figure 1 shows the web interface being used to annotate a sample anonymized radiology report.

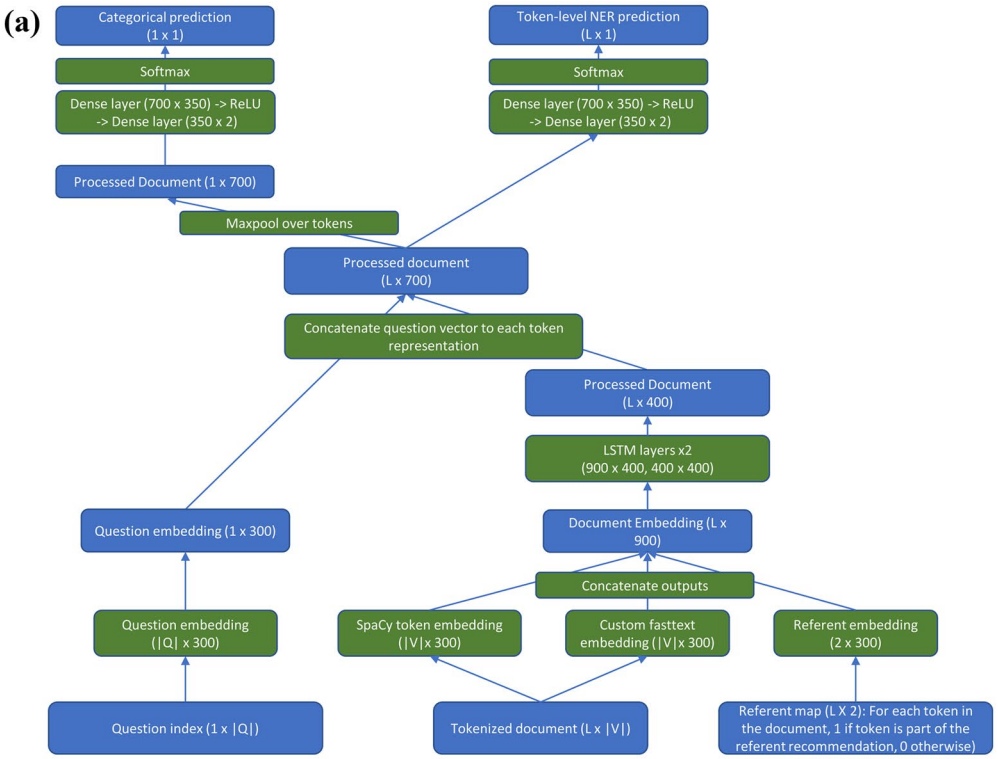
Appendix 3. Neural Network Architectures

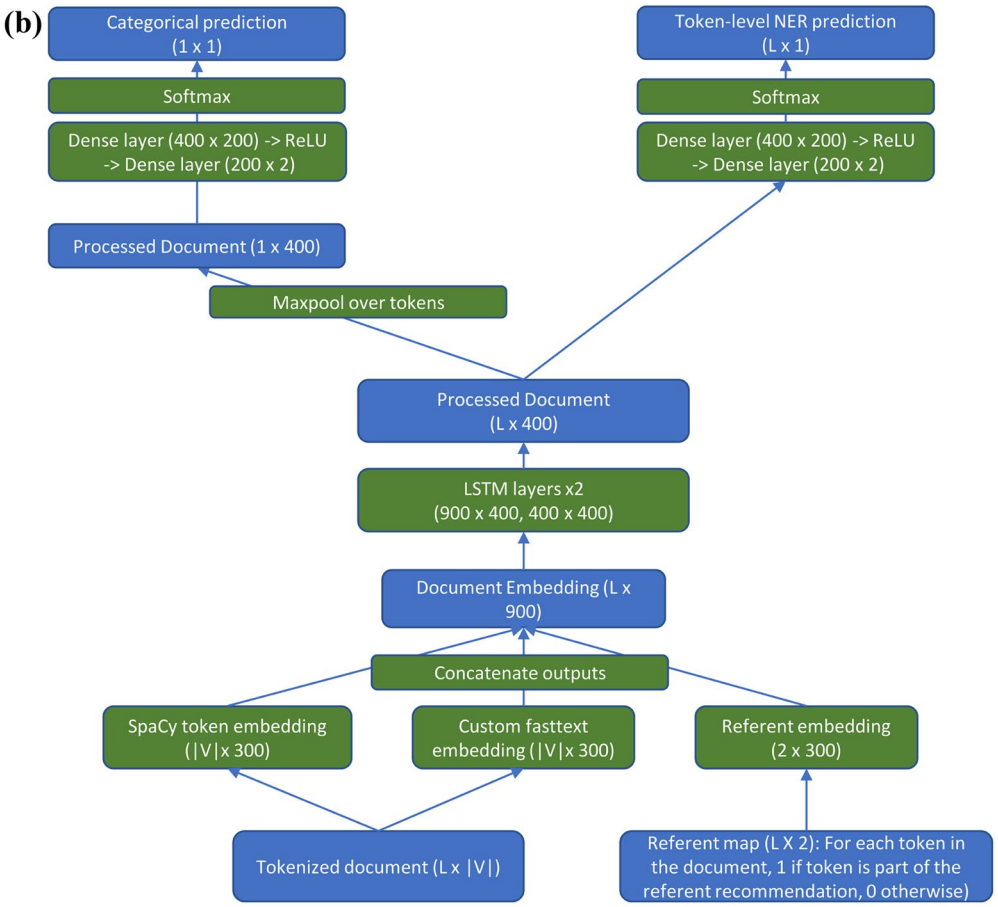
Word Vectors

In order to capture the specifics of the relatively limited domain of radiologic text while also maintaining the benefits of word embeddings trained on a large and diverse English corpus, we opted to combine two word embeddings: (1) a set of general-purpose English text embeddings from spaCy's "en_core_web_lg" model (trained on the Common Crawl dataset as Global Vectors (GloVe)) and (2) custom-trained fastText vectors on our institution's entire corpus of radiology reports. We used the fastText implementation from the gensim python package with the skip-gram training procedure, 300-dimensional embeddings, and all other parameters set to gensim's default values.

LSTM Multi-Task Model

The multi-task LSTM consisted of a document embedding layer with an output consisting of 600 dimensions; 300 of these came from the default SpaCy vectors ('en_core_web_lg' model), and 300 from our custom-trained fastText vectors. In addition, there was a separate referent embedding layer which embedded the referent vector in a 300-dimensional space. For this study, there was only one type of referent (recommendations), so the embedding layer processed vectors consisting entirely of zeros (indicating that the token was not part of a referent) and ones (indicating the token *was* part of a recommendation referent), although it is generalizable to an arbitrary number of referent types. These two word embeddings were concatenated to produce the final token embedding of size 900. The concatenated embeddings were then processed by 2 bidirectional LSTM layers of dimension 400. The processed text was concatenated, token-wise, with the question embedding. The question embedding was a third, separate embedding layer which embedded the question type,


(a)
Categorical prediction (1 x 1)
Softmax
Dense layer (700 x 350) -> ReLU -> Dense layer (350 x 2)
Processed Document (1 x 700)
Maxpool over tokens
Token-level NER prediction (L x 1)
Softmax
Dense layer (700 x 350) -> ReLU -> Dense layer (350 x 2)
Processed document (L x 700)
Concatenate question vector to each token representation
Processed Document (L x 400)
LSTM layers x2 (900 x 400, 400 x 400)
Document Embedding (L x 900)
Concatenate outputs
Question embedding (1 x 300)
Question embedding (|Q| x 300)
Question index (1 x |Q|)
SpaCy token embedding (|V|x 300)
Custom fasttext embedding (|V|x 300)
Tokenized document (L x |V|)
Referent embedding (2 x 300)
Referent map (L X 2): For each token in the document, 1 if token is part of the referent recommendation, 0 otherwise)


(b)
Categorical prediction (1 x 1)
Softmax
Dense layer (400 x 200) -> ReLU -> Dense layer (200 x 2)
Processed Document (1 x 400)
Maxpool over tokens
Token-level NER prediction (L x 1)
Softmax
Dense layer (400 x 200) -> ReLU -> Dense layer (200 x 2)
Processed Document (L x 400)
LSTM layers x2 (900 x 400, 400 x 400)
Document Embedding (L x 900)
Concatenate outputs
SpaCy token embedding (|V|x 300)
Custom fasttext embedding (|V|x 300)
Referent embedding (2 x 300)
Tokenized document (L x |V|)
Referent map (L X 2): For each token in the document, 1 if token is part of the referent recommendation, 0 otherwise)

◂**Fig. 2** (**a**) Diagram of the multi-task neural network model. Green boxes represent model actions; blue boxes represent the state of data as it passes through the network. $|V|$: number of tokens in vocabulary. $|Q|$: number of unique questions to be answered by the network. L: length of document. Batch sizes are ignored for ease of comprehension. (**b**) Diagram of the single-task neural network model. Green boxes represent model actions; blue boxes represent the state of data as it passes through the network. In this model, the question type is used to select *which* of the sub-networks handles the question, and therefore requires no question embedding required. Each sub-network has its own unique weights and is trained only on a single question. $|V|$: number of tokens in vocabulary. $|Q|$: number of unique questions to be answered by the network. L: length of document. Batch sizes are ignored for ease of comprehension

enabling the same network to answer all six different questions. A two-layer dense neural network (dimensions: 700×350 followed by 350×2), with an intermediate ReLU layer to allow for nonlinearity, was then used to process each token to produce the final two-dimensional output (part of an answer vs. not). For questions with categorical inputs, a maxpool over all tokens was applied before a separate 2-layer dense network (700×350, 350×2). The network is shown in Fig. 2a.

LSTM Single-Task Model

This model is very similar to the multi-task model; the only difference is that there is a separate network (with unique weights) for each question, which only answers that question. The type of question now merely determines *which sub-network* each question is passed to. The network is shown in Fig. 2b.

Declarations

Disclosure JS is part-owner of River Records LLC., a healthcare company focused on developing tools to reduce clinician documentation burden and improve healthcare quality. The company had no involvement in this study.

References

1. Mabotuwana T, Hall CS, Hombal V, Pai P, Raghavan UN, Regis S, et al: Automated tracking of follow-up imaging recommendations. AJR Am J Roentgenol. 2019;1–8.
2. Zafar HM, Chadalavada SC, Kahn CE Jr, Cook TS, Sloan CE, Lalevic D, et al: Code Abdomen: An Assessment Coding Scheme for Abdominal Imaging Findings Possibly Representing Cancer. J Am Coll Radiol. 2015;12: 947–950.
3. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS: Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. J Digit Imaging. 2019;32: 554–564.
4. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ: Extraction of recommendation features in radiology with natural language processing: exploratory study. AJR Am J Roentgenol. 2008;191: 313–320.
5. Dang PA, Kalra MK, Blake MA, Schultz TJ, Stout M, Lemay PR, et al: Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol. 2008;5: 197–204.
6. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, et al: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology. 2005;234: 323–329.
7. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH: A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform. 2013;46: 354–362.
8. Mabotuwana T, Hall CS, Dalal S, Tieder J, Gunn ML: Extracting Follow-Up Recommendations and Associated Anatomy from Radiology Reports. Stud Health Technol Inform. 2017;245: 1090–1094.
9. Reddy S, Chen D, Manning CD: CoQA: A Conversational Question Answering Challenge. Transactions of the Association for Computational Linguistics. 2019;7: 249–266.
10. Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, et al: Multimodal explanations: justifying decisions and pointing to the evidence. arXiv [cs.AI]. 2018. Available: http://arxiv.org/abs/1802.08129
11. Sukthanker R, Poria S, Cambria E, Thirunavukarasu R: Anaphora and coreference resolution: a review. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1805.11824
12. 12. Hochreiter S, Schmidhuber J: Long short-term memory. Neural Comput. 1997;9: 1735–1780.
13. [PDF]GloVe: Global vectors for word representation—Stanford NLP. Available: https://nlp.stanford.edu/pubs/glove.pdf
14. Segura Bedmar I, Martínez P, Herrero Zazo M: Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics; 2013. Available: https://e-archivo.uc3m.es/bitstream/handle/10016/20455/semeval_SEMEVAL_2013.pdf?sequence=3
15. Zhang Y, Yang Q: A survey on multi-task learning. arXiv [cs.LG]. 2017. Available: http://arxiv.org/abs/1707.08114
16. Sistrom CL, Dreyer KJ, Dang PP, Weilburg JB, Boland GW, Rosenthal DI, et al: Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology. 2009;253: 453–461.
17. Cochon LR, Kapoor N, Carrodeguas E, Ip IK, Lacson R, Boland G, et al: Variation in Follow-up Imaging Recommendations in Radiology Reports: Patient, Modality, and Radiologist Predictors. Radiology. 2019;291: 700–707.
18. Kaissis GA, Makowski MR, Rückert D, Braren RF: Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging. Nature Machine Intelligence 2020;6: 305–11.