An end-to-end neural model based on cliques and scopes for frame extraction in long breast radiology reports

Perceval Wajsbürt, Xavier Tannier

An end-to-end neural model based on cliques and scopes for frame extraction in long breast radiology reports

Perceval Wajsbürt[1,2] and **Xavier Tannier**[1]
[1]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, Paris, France
[2]Innovation and Data unit, IT department,
Assistance Publique-Hôpitaux de Paris, Paris, France

Abstract

We consider the task of automatically extracting various overlapping frames, i.e, structured entities composed of multiple labels and mentions, from long clinical breast radiology documents. While many methods exist for related topics such as event extraction, slot filling, or discontinuous entity recognition, a challenge in our study resides in the fact that clinical reports typically contain overlapping frames that span multiple sentences or paragraphs. We propose a new method that addresses these difficulties and evaluate it on a new annotated corpus. Despite the small number of documents, we show that the hybridization between knowledge injection and a learning-based system allows us to quickly obtain proper results. We will also introduce the concept of scope relations and show that it both improves the performance of our system, and provides a visual explanation of the predictions.

1 Introduction

In this study, we will address the task of structuring breast radiology reports.[1] This end to end task consists in extracting different structured entities (frames) of various types, each one composed of multiple labels and multiple mentions organized as a list of fields. The creation of a frame is triggered by a "trigger" mention, and some of its fields may be justified by "attribute" mentions. An example of structured lesion frames extracted from a fictitious document is illustrated in Figure 1 and Table 1, using the scheme described in Appendix A. A first difficulty is related to the size of the documents: frames can group mentions that span several sentences or paragraphs. The second difficulty is related to overlapping frames that may share multiple mentions and even trigger mentions. Therefore, modelling the relations between different mentions is necessary to distinguish between multiple overlapping frames, such as Frames 1 and 3 in the example. Our contributions [2] are the following:

- an end-to-end model to extract frames from texts
- a clique-based method for dealing with overlaps
- the concept of scope-relations to group mentions

Breast ultrasound:
Left:
Two cysts located on the 8 o'clock radius at 3 cm, and at 2cm on the 6 o'clock radius.
These nodules are millimetric.

Right:
No abnormal masses to report.

CONCLUSION :
Multiple cysts on the left.

Figure 1: Fictitious excerpt, annotated with lesion related-mentions colored according to their frames

2 Related works

The extraction of structured information from clinical reports has been the subject of many studies. Most of these works are not specific to breast imaging reports and the objectives vary greatly, in terms of their scope, granularity and form. Interested readers can refer to existing surveys on the state of NLP in radiology reports (Bitterman et al., 2021; Miwa et al., 2014). Several works are only concerned with the extraction of a few report-level attributes, and therefore view the task as a classification or term extraction task for items such as ACR scores, histological grade or primary site of

[1]This study was approved by the institutional review board at APHP (CSE 190022) as part of the EZMammo project. Only previously pseudonimized documents were used in this study (Paris et al., 2019).

[2]An implementation of the model presented in this article can be found at https://github.com/percevalw/breast-imaging-frame-extraction

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 156–170

	Frame 1		Frame 2		Frame 3	
field	**value**	**mention(s)**	**value**	**mention(s)**	**value**	**mention(s)**
trigger		[cysts], [nodules]		Multiple [cysts]		[cysts], [nodules]
organ	breast	[Breast]	breast		breast	[Breast]
laterality	left	[Left]:	left	on the [left]	left	[Left]:
time	current		current		current	
quadrant						
size		[millimetric]				[millimetric]
distance	30mm	[3 cm]			20mm	[2 cm]
angle	8	[8 o'clock radius]			6	[6 o'clock radius]

Table 1: Lesion frames of the example in Figure 1.

lesions (He et al., 2017; Qiu et al., 2018; Alawad et al., 2018; Moore et al., 2017; Castro et al., 2017). Other features have also been the subject of specialized systems such as locations (Datta et al., 2020). An extensive survey of the different systems proposed for different features was conducted by Datta et al. (2019). Other works have sought to produce a more detailed and global extraction, and to detect several types of entities at the same time. The earliest work was the one of Taira et al. (2001), who proposed a frame based representation and method for annotating abnormal findings, anatomy, and medical procedures frames in radiology reports. Lacson et al. (2015) used a rule-based system and terminologies to extract abnormal findings and ACR scores. The DeepPhe system was proposed by Savova et al. (2017) as a fully integrated software built on cTakes (Savova et al., 2010) to extracts document and patient level cancer summaries (akin to frames) in clinical reports. Steinkamp et al. (2019) proposed a fact-based scheme, in which each fact is structured around an anchor and may contain modifiers. However, their model makes the assumption that all the mentions that characterize an entity are adjacent inside the fact span. Several methods decompose the problem into a first NER step followed by a relation detection step that allows arguments to be non adjacent. Roberts et al. (2019) proposed a frame based scheme for annotating cancer information in clinical reports and a method to perform the prediction (Si and Roberts, 2018). Their method first extracts triggers and modifiers with a NER system, and predicts their relations to form frames, but makes the assumption that there is no overlap between the different entities. Recently, a more complex scheme has been proposed by Jain et al. (2021) to annotate nested relationships between different entities. However, this work does not specifically address the case of complex or distant relations between entities.

The closest task to ours is the one of Event Extraction, in which models extract one event per trigger mention, and look for related mentions that might be part of the same event. However, trigger mentions (e.g., the first [cysts] in Figure 1) may belong to different overlapping frames (the *3cm 8 o'clock lesion* frame and the *2cm 6 o'clock lesion* frame in Figure 1) that can only be distinguished by considering relations between their attribute mentions ([2cm], [3cm], [8 o'clock radius], and [6 o'clock radius]). To address this issue, an approach consists in listing all the possible combinations of mentions, then filtering them with a classifier (Miwa et al., 2010; Heimonen et al., 2010; Björne and Salakoski, 2011, 2013, 2015; Liu et al., 2015; Trieu et al., 2020). However, this solution becomes computationally unsatisfactory when the number of mentions that compose a frame grows.

3 Method

We now detail a neural network based end-to-end method to automatically extract frames from clinical reports. We encode each document as word embeddings and share these with the downstream decoding components. Like most relation and event extraction models, our model operates as a pipeline. As illustrated by Figure 2, the first two mention-level decoders extract the named entities, or mentions (step ①), that are likely to be used in the composition of structured entities, and normalize them (step ②) to obtain the value of the field they apply to. The next two decoders focus on frame-level extractions. The frame extraction decoder (step ③) extracts groups of mentions to form frames. For each frame, the last frame classification decoder (step ④) predicts the values of the fields for which no explicit mention was found, such as a past temporality, which could be indicated by a verb tense.

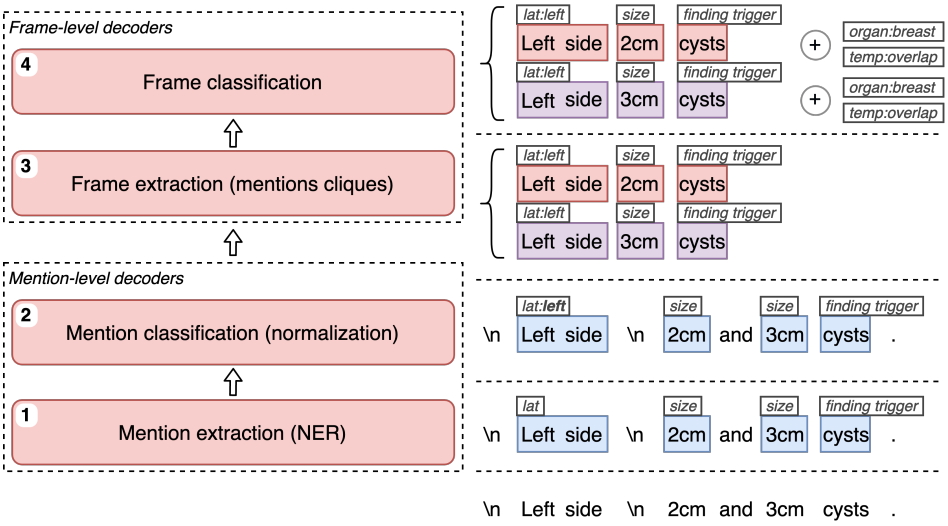

Figure 2: Overview of the decoding steps: we ① extract mentions, ② normalize them, ③ group them into frames and ④ predict the values of the fields for which no explicit mention was found

3.1 Text encoder

Our documents are written in French, therefore we use a pretrained CamemBERT model (Martin et al., 2020). To reduce the sequence size and ensure that the NER step does not predict boundaries inside words, we average the wordpieces embeddings of a word to obtain one embedding per word. Moreover, we split the document into sentences using a regular expression. We add the left and right contexts ("document context") of each sentence before running it through the Transformer, up to a maximum total number of wordpieces, as it proved useful in other studies (Devlin et al., 2019; Kantor and Globerson, 2020; Yu et al., 2020; Schweter and Akbik, 2020; Luoma and Pyysalo, 2021). Next, we apply multi-layer BiLSTM on the concatenation of the BERT embeddings generated for each sentence of the document. BERT models usually focus on sentences and replace the "line break" character by a single space. To keep this informative token in our long clinical documents, we replace all line breaks with the rarely used "_" character.

3.2 Mention recognition and normalization

We use a sequence labeling NER model based on the BIOUL tag scheme with multiple parallel CRF layers. Each independent layer is responsible for the extraction of entities of a given type, such that predicted entities of different types may overlap.

Each mention is then classified, or normalized, to obtain the values of the fields to which it applies. A subset of the available values for each field/mention is given in Table 7.

For example, "bilateral" is normalized as both "left" and "right". The allowed multi-label combinations is defined manually. We compute a max-pooled representation for each mention m and project it to obtain one score per label:

$$\text{score}^{\text{label}}(m) = V^{\text{label}} \cdot \underset{w \in \text{words}(m)}{\text{maxpool}} E(w)$$

Finally the score of each possible legal label combination L_{mention} is computed as the score of the labels present in the combination. The probability is computed by normalizing over all legal combinations.

To be processed in the next layers, each mention is represented by the average embedding of its words.

3.3 Frame extraction

We now seek to extract the frames, that is, group the extracted named entities together. We will now describe a method to overcome the previously discussed issue of overlapping frames. The overall frame extraction component and its training procedure are described in Figure 3.

3.3.1 Clique extraction

Our approach consists in answering the following question for each pair of mentions: "are these two mentions part of the same frames ?" We can then extract maximal cliques of entities, i.e., groups in which mentions agree with each other on belonging to the same frame. For two mentions u and v, we compute the score $r(u, v)$ computed by u of

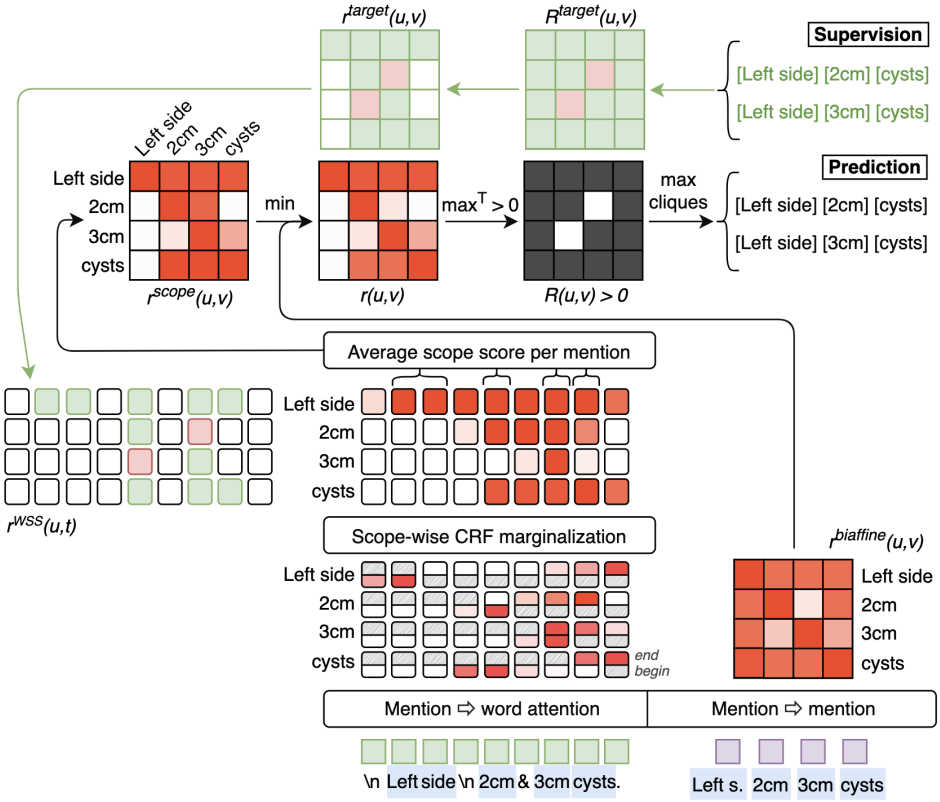

Figure 3: Overview of the frame extraction process and its supervision. Forbidden scope begin and end locations (because they are located after or before the mention) are grayed out. Green matrices and arrows at the left and top of the Figure show the possible supervision signals: red means no relation, green means that a relation should be predicted, and white means no supervision.

v belonging to its frames, and the score $r(v, u)$ computed by v of u belonging to its frames: the final agreement score between the two mentions is the maximum

$$R(u, v) = \max{}^T r = \max(r(u, v), r(v, u))$$

meaning that one of the two mentions can be uncertain about the relationship. At this point, we could have assumed a symmetric function for r to avoid the max computation, but as we will see in the next sections, both biaffine and scopes scores are asymmetric.

3.3.2 Biaffine relation scores

A simple baseline to compute $r(u, v)$ consists in a biaffine model. In our case, we compute this score as an attention score between the mentions representations. Additionally, we inject the relative distances between mentions inside the attention mechanism using a similar mechanism to He et al. (2020). This attention is the sum of a content-content attention (the original dot product attention of Vaswani et al. (2017)), a content-position attention and a position-content attention.

3.3.3 Scope relation scores

We propose another approach for the same relation extraction task, based on the concept of scopes. Scopes are annotations of contiguous text zones on which a named entity referred to as a "cue" applies its meaning. Scopes have been mostly studied in the context of negation and uncertainty detection (Vincze et al., 2008; Li and Lu, 2018; Dalloux et al., 2020; Khandelwal and Sawant, 2020). We extend this concept to all types of named entities and make it the primary mode of relation extraction in our task. Indeed, it may be simpler for the model to detect where the scope of a mention starts and stops, and to retrieve all entities between these boundaries, than inferring the value of the relation for each pair of mentions. In the example of Figure 1, the scope of laterality [Left] covers all the section and therefore applies its effect to all frames composed of these mentions.

For the mathematical details of our formulation, we will call u and v two mentions, and t a word. Each scope is represented with the BIOUL format. We compute two attention matrices $S_B(u, t)$

and $S_L(u,t)$ between the mentions and words, using the relative attention mechanism previously described to obtain start (B) and end (L) scope scores for each word. We constrain B to be before the mention and L after. The score S_U of the tag U (scope that only contains one word) can be computed as the sum of the start and end scores, and the scores S_I and S_O of I and O tags are set to zero and will be inferred by a CRF layer (Lafferty et al., 2001).

To predict if a word is in the scope of a mention, i.e., is labeled I, B, L or U, we compute the marginalized probabilities $S^m_{\cdots}$ of a CRF with the forward-backward algorithm on the scope of each mention. The Scope CRF is parameter-less but illicit transitions (such as I $\longrightarrow$ B or L $\longrightarrow$ I) between tags are prevented, i.e., all CRF weights are 0 or $-\infty$. The score $r^{\text{scope}}(u,t)$ of each word t being in the scope of u is therefore:

$$r^{\text{scope}}(u,t) = \ln\left[e^{S^m_{\text{B}}} + e^{S^m_{\text{L}}} + e^{S^m_{\text{U}}} + e^{S^m_{\text{I}}} - e^{S^m_{\text{O}}}\right]$$
$$\text{with } S^m_{\text{BIOUL}}(u) = \text{ForwardBackward}(S_{\text{BIOUL}}(u))$$

and, the score $r^{\text{scope}}(u,v)$ of v being in the scope of u, i.e., the average of the scores of each word of v of being in the scope of u: $\frac{1}{|v|}\sum_{t\in v} r^{\text{scope}}(u,t)$

Using a CRF allows us to never explicitly compute the score that a word is in the scope of a given mention. Instead, we let the network predict the start and end of the scope for each mention and use the CRF to "paint" the inside of the scopes in a differentiable way.

3.3.4 Score combination

The scope relation and biaffine relation scores are combined together. Because we defined scopes as being continuous spans of text, it is possible that a mention falls in the scope of another mention and yet does not belong to its frame. In the example *"Mammography: we find the left mass biopsied in 2010. Nothing else in the right breast."*, the scope of [Mammography] contains the temporality [2010] but the two mentions are not part of a same frame. Therefore, a relation between two mentions is only predicted if both components (biaffine-based and scope-based) predict this relation, which we formulate as $r(u,v) = \min(r^{\text{scope}}(u,v), r^{\text{biaffine}}(u,v))$.

3.3.5 Frame relation supervision

Training the frame extraction module raises several difficulties. For two compatible mentions u and v, we supervise R with the supervision matrix R^{target} via a binary cross-entropy loss.

$$R^{\text{target}}(u,v) = \begin{cases} 1 \text{ if } u \text{ and } v \text{ are linked} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

The score $R(u,v)$ is the result of the maximum of a matrix $r(u,v)$ and its transpose, which, from a scope perspective, means that one mention can be within the scope of another without the reverse being true. This non-differentiable maximum can be hard to learn for the model.

For this reason, we propose to supervise one of the two relation directions scores (i.e., $r(u,v)$ or $r(v,u)$) specifically, instead of the maximum, with the asymmetric target matrix $r^{\text{target}}(u,v)$. The difference between these two supervision modes is illustrated at the top of the Figure 3. If the two mentions u and v are not part of the same frames, both directions scores should be negative, since $max(r(u,v), r(v,u)) = R(u,v) < 0$. However, if the two mentions share the same frames, we "explore" the two different supervision directions by performing stochastic sampling of r^{target}, according to a categorical distribution parameterized by the relation probability computed by the model:

$$[r^{\text{tgt}}(u,v), r^{\text{tgt}}(v,u)] \sim \text{softmax}(r(u,v), r(v,u))$$

The model should explore a few ways of arranging the scopes at the beginning of the training when the probabilities are close to 0.5, and stick to a strategy that leads to low entropy of the above distribution as the training progresses and its confidence increases in either direction.

3.3.6 Supervision heuristics

We also experiment with heuristics in the supervision matrix $r^{\text{target}}(u,v)$. If u belongs to strictly more frames than v, we maximize $r(u,v)$. If both belong to the same number of frames, we choose the direction that leads to the smallest number of wrong erroneous memberships due to the contiguity of scopes. Finally, if no heuristic can be applied, we sample a direction as previously described.

3.3.7 Word-level scope supervision (WSS)

Finally, we also propose to supervise the scopes scores $r^{\text{scope}}(u,t)$ directly using partial word-level annotation r^{WSS} generated from the r^{target} matrix, as illustrated on the left side of Figure 3. Indeed, using the r^{target} matrix for a given mention u, we can determine which words t of other mentions

should be contained in its scope, which words of other mentions should not, and which words are not supervised. Because scopes are contiguous, if a mention v that is not part of the frame of u is contained within its partially supervised scope, i.e., if it is between two mentions that belong to the scope of u, we do not supervise its words and leave the biaffine component handle the non-relation detection.

3.4 Frame classification

Some labels of a frame such as its temporality or laterality may not be explicitly supported by a mention. Each frame is therefore fed through a constrained multi-label classifier. We represent each frame by an embedding computed as a projection of the max-pooling output of the embeddings of its mentions, and then project it to give a score per label. The score of a label combination is computed as the score of the labels in the combination. The probability is computed by normalizing over all legal combinations. During prediction, the label combinations are filtered to keep only those that contain the normalized labels of the mentions in the frame.

3.5 Optimization

The different components are trained jointly. We use the CRF Forward algorithm to compute the NER loss, cross-entropy to compute the mention normalization and the frame classification losses. The frame extraction decoder relation loss $\mathcal{L}_{\text{relation}}$ is the sum of binary cross entropy for every valid supervised mention-mention pair and the partial word-level supervision Scope CRF loss $\mathcal{L}_{\text{scope}}$ is the CRF Forward algorithm. The losses are combined into a weighted average, the specifics of which are detailed in Appendix B.

3.6 Knowledge injection

Data augmentation We augment the training data in two ways. First, we randomly extract parts of documents such that no frame is cut, and add them as new documents to the dataset. This is somewhat akin to sentence splitting, but for multi-sentence entities. Second, we build synthetic sentences from a manually pre-defined lexicon of mentions, and add these sentences as NER samples to the dataset. The sentence creation process is the following: we randomly pick a synonym from the lexicon such as [ACR 6] and insert it in a randomly picked context from a predefined list such as *"There is {} ."* to generate *"There is [ACR 6]."* The documents generated from these augmentations are mixed with the original documents such that every batch approximately contains $\frac{1}{3}$ of each (original, doc parts and lexicon sentences).

Output constraints Some background knowledge can be injected by constructing rules such as the fact that "left" and "right" are exclusive, or the fact that a mammogram is always performed on the breasts. During the frame extraction step, relations between mentions that cannot be part of the same frame are filtered out during learning and prediction. Similarly, as mentioned in Section 3.2, illegal label combinations are filtered out during training and prediction. This filtering reduces the number of possibilities that the model must evaluate, and alleviate the need for the model to "learn" the annotation scheme.

4 Experiments

We evaluate our proposed approach on the test set of a new annotated dataset described in Appendix A, and perform several ablation experiments to investigate the design choices of our model. The dataset is composed of 120 French breast imaging clinical reports annotated with frames. There are five types of objects: ACR (cancer risk) scores, breast density scores, diagnostic procedures, therapeutic procedures and lesions. The document-level statistics are detailed in Table 2. The model is evaluated with 3 retrieval metrics: the mention metrics evaluate the mention and normalization prediction with approximate boundaries, the Frame Support evaluates the frames through their mentions, and the Frame Label evaluates them through their labels. These metrics are further described in the following section.

	train	**test**
count	80	40
average words	361.0750	362.175
average lines	45.7375	45.475
average frames	19.4750	18.425

Table 2: Document level statistics for the corpus

4.1 Relaxed retrieval metrics

We use three metrics to evaluate the predictions at the mention and frame level, and provide algorithms to compute them in Appendix C.

Unlike the exact match NER metric for which a true positive is unambiguously counted when two elements of the predicted and gold entities match, defining and computing relevant metrics between more complex sets of objects becomes more difficult as the number of element attributes increases. One option is to lower the minimum similarity threshold required between predicted and gold features to account for small errors such as mismatch between mention boundaries. However, this leads to ambiguities in the metric computation, since several predicted elements may match a single gold element, and vice versa. We explicitly formulate a greedy matching procedure to compute a maximum bipartite greedy match between the elements of two sets, in the algorithm 1 to avoid double counting true positives.

The NER metric (Algorithm 2) uses a score function that returns 1 if the Dice overlap of words in two mentions is higher than 0.5. The procedure is described in the Algorithm 2.

The Frame Support metric (Algorithm 3) scores a pair of two frames with a non-zero match score if some of their mentions overlap, and a perfect score if all their mentions overlap, and 0 otherwise. This score between 0 and 1 is the Dice/F1 overlap between the mentions of the two frames. It is used as a "relaxed" true positive when computing the retrieval metrics.

Finally, the Frame Label metric (Algorithm 4) scores a pair of two frames with a matching score of 1 if their labels match and their trigger mentions overlap, and 0 otherwise. This score is used as a true positive when computing retrieval metrics.

4.2 Experimental setup

Hyperparameters were manually selected by trial and error on 20 documents from the training dataset, and the model were trained for 2000 steps with a batch size of 16. Hyperparameters are further described in Appendix B. All experiments were averaged on 3 runs.

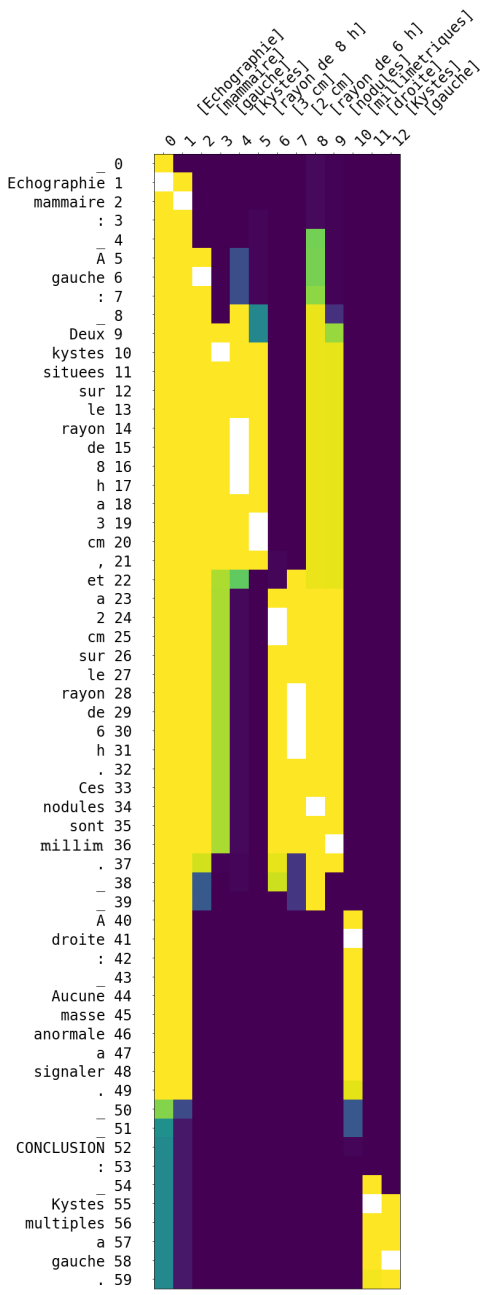

Figure 4: Visualization of the predicted mentions and scopes on the example of Figure 1 (original French version). The vertical axis represents the words, and the horizontal axis represents the mentions. The predicted mention is marked in white, the scope in yellow.

4.3 Main results

Table 3 shows the performance on the different types of frames. The model performs better for frames with fewer fields such as Cancer risks or Breast densities. We visualize the predicted scopes of the proposed model on the right side of Figure 4. We observe that the scopes coarsely follow the structure of the document, i.e., that the predicted boundaries are located at the beginning or the end of the different sections. This observation suggests that our approach may effectively leverage the structure commonly found within clinical documents. It is worth keeping in mind that these scopes have only been supervised with the requirement that they contain or exclude certain mentions, and that no information regarding the precise location of their boundaries has been given. Moreover we note that the reading of these scopes gives a par-

	Frame support			Frame label		
Type	P	R	F1	P	R	F1
ACR score	89.6	95.7	92.5	80.6	86.1	83.3
Density	84.9	96.9	90.5	82.6	94.3	88.1
Diag proc	82.1	91.7	86.6	74.0	82.7	78.1
Ther proc	86.2	87.1	86.6	68.3	69.0	68.6
Finding	74.0	82.4	78.0	59.6	66.5	62.9
Overall	81.1	90.0	85.3	68.7	76.2	72.2

Table 3: Performance of the model at the frame level

tial explanation of the predicted relations, whereas the outputs of relation prediction models are usually hardly explainable.

4.4 Impact of scopes

Table 4 shows the effect of ablating the model scopes. In this configuration, the model can only predict the relations through the biaffine model. We can observe that ablating scopes results in an overall loss of 5.3 pt for the Frame Label metric and 4.9 pt for the Frame Support metric. We believe that this is due to the inability of standard neural components to reason with intervals, i.e., to answer queries such as "what word is between these two words". Given that scopes improve the quality of predictions, the question arises as to what kind of supervision is needed for learning them. As shown in Table 4, when the scopes are learned directly using word-level partial annotations, the model performs better than with distant supervision on the $r(u, v)$ matrix. If we directly supervise the symmetric matrix $R(u, v)$ instead of the asymmetric matrix $r(u, v)$, the performance collapses and we lose between 10 and 15 pt for the Frame metrics. The learning of scopes must be hindered by the uncertainty related to the supervision of this matrix alone and the small amount of data. Interestingly, if we remove the relation supervision heuristic described in Section 3.3.6 and let the model explore different configurations on its own, the performance shown in Table 4 remains on par with the proposed approach. Since these heuristics aim at injecting information about the hierarchy of mentions and the structure of the text, this suggests that the model is able to infer this information itself.

4.5 Impact of the relative attention

We evaluated the effect of the added information on the relative position of the word-mention and mention-mention attention mechanisms. From the Table 4, we can observe that this added information leads to a performance gain of 1.3 pt of F1 frame support and 1.8 pt of F1 frame label. Without it, a mention is "positionally blind" and must rely on the inductive bias of the LSTM to find its neighboring words or mentions. Therefore, we expected a larger drop in performance, especially in the context of long documents. Nevertheless, relative attention proves to be an effective way to improve performance.

4.6 Impact of the size of the training data

Figure 5 shows the overall performance of the model when trained with different numbers of annotated samples. On one hand, we can note that our system requires only a small amount of documents to achieve "correct" accuracy, i.e., it can be used to pre-annotate more documents. This "data efficiency" is important when tackling new domains in order to allow quick feedback and possible changes regarding the annotation scheme. However, given the complexity of the task and the evolution of performance with the training set size, we also note that a large number of annotated documents might be needed to approach a perfect score.

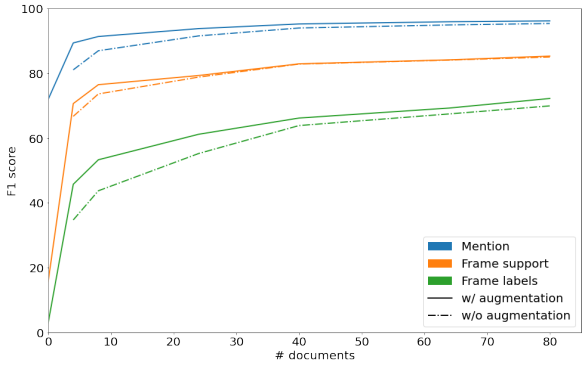

Figure 5: Plotted evolution of the F1 scores with the number of annotated documents. The plain lines show the performance with data augmentation (synthetic sentences and document splitting), while the dashed lines show the performance without augmentation.

4.6.1 Impact of the augmentations

We remove the augmented samples from the training data and show the effect on performance in Table 5 and Figure 5. We observe that adding synthetic lexicon sentences only slightly helps improving the model mention detection performance (+0.3 pt). However, this improved performance has a larger effect of 1.5 pt on the Frame Label metric. This is typical of the phenomenon of error propa-

	Mention	Frame support	Frame label
Full model	96.2	85.3	72.2
− relative attention	95.6 (−0.5)	84.0 (−1.3)	70.5 (−1.8)
− relation supervision heuristics	96.1 (−0.1)	85.4 (+0.1)	71.8 (−0.4)
− WSS	96.1 (−0.1)	82.1 (−3.2)	69.5 (−2.7)
− WSS − asymmetric supervision	95.9 (−0.3)	74.4 (−10.9)	57.5 (−14.8)
− scopes (only biaffine scores)	96.2 (+0.0)	80.4 (−4.9)	66.9 (−5.3)

Table 4: Architecture ablation experiments . WSS stands for Word-level Scope Supervision. All reported metrics are F1-scores.

	Mention	Frame support	Frame label
All	96.2	85.3	72.2
− doc splitting (1)	96.1 (−0.0)	85.3 (+0.1)	71.5 (−0.7)
− synthetic lexicon sentences (2)	95.4 (−0.8)	85.0 (−0.3)	70.8 (−1.5)
− data augmentations (1+2)	95.4 (−0.8)	85.0 (−0.3)	69.9 (−2.3)
− constraints	96.2 (−0.0)	84.0 (−1.3)	69.4 (−2.8)

Table 5: Knowledge and data ablation experiments. All reported metrics are F1-scores.

gation, since a missing or mislabelled mention can have an effect on multiple frames.

As we reduce the number of annotated documents in the training set, the effect of augmentation becomes more important, and with only 4 annotated documents we obtain an average performance of 89.4 F1 in mention extraction versus 81.1 F1 without, and an average performance of 45.7 F1 in Frame Label F1 versus 34.7 without. Finally, we can see that a model trained only with synthetic sentences, i.e., 0 training document in Figure 5) already obtains decent retrieval performances, which is valuable when tackling a new domain with unlabeled data only. The non-zero Frame metrics can be explained by the presence of frames containing only one mention, and the constraints preventing the system from predicting illicit label combinations.

4.6.2 Impact of constraints

We train the model without the constraints described in section 3.6 (but we still apply these constraints during the evaluation phase to avoid illicit predictions). In this configuration, the model learns that each pair of mentions is legal. We observe in Table 5 this leads to a loss of 2.3 pt in the Frame label F1-score and 1.3 pt in the Frame support F1-score. This can be explained by the fact that the model has to "learn" the annotation scheme and its inevitable imperfect representations of the reports. These constraints can also help the model focus on the actual uncertainties of the task, and leave what is already known to the modeled constraints.

5 Acknowledgment

We thank the clinical data warehouse (Entrepôt de Données de Santé, EDS) of the Greater Paris University Hospitals for its support and the realization of data management and data curation tasks.

6 Conclusion

In this work, we presented a system for extracting structured entities from clinical breast radiology reports. We have shown that the addition of synthetic sentences can improve the performance in the context of a small amount of data. This information is valuable for the annotation and development of new information retrieval systems in other domains, where key words or phrases are known in advance. The method we described introduces the notion of frame extraction in the form of mention cliques, and we have shown that a formulation of the relation extraction task via scopes improves the performance of our system. Future work will evaluate this approach on other structured entity extraction tasks such as event extraction.

References

Mohammed Alawad, Hong Jun Yoon, and Georgia D. Tourassi. 2018. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. *2018 IEEE EMBS International Conference*

on Biomedical and Health Informatics, BHI 2018, 2018-Janua:218–221.

Danielle S. Bitterman, Timothy A. Miller, Raymond H. Mak, and Guergana K. Savova. 2021. Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. *International Journal of Radiation Oncology Biology Physics*, 110(3):641–655.

Jari Björne and Tapio Salakoski. 2011. Generalizing Biomedical Event Extraction. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task Portland Oregon June Association for Computational Linguistics*, page 183–191.

Jari Björne and Tapio Salakoski. 2013. TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. *BioNLP Shared Task 2013 Workshop*, pages 16–25.

Jari Björne and Tapio Salakoski. 2015. TEES 2.2: Biomedical Event Extraction for Diverse Corpora. *BMC Bioinformatics*, 16(16):1–20.

Sergio M. Castro, Eugene Tseytlin, Olga Medvedeva, Kevin Mitchell, Shyam Visweswaran, Tanja Bekhuis, and Rebecca S. Jacobson. 2017. Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics*, 69:177–187.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2020. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Surabhi Datta, Elmer V. Bernstam, and Kirk Roberts. 2019. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E. Shooshan, Dina Demner-Fushman, and Kirk Roberts. 2020. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning. *Journal of Biomedical Informatics*, 108(February):103473.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

Tiancheng He, Mamta Puppala, Richard Ogunti, James J. Mancuso, Xiaohui Yu, Shenyi Chen, Jenny C. Chang, Tejal A. Patel, and Stephen T.C. Wong. 2017. Deep learning analytics for diagnostic support of breast cancer disease management. *2017 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2017*, pages 365–368.

Juho Heimonen, Jari Björne, and Tapio Salakoski. 2010. Reconstruction of semantic relationships from their projections in biomolecular domain.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. (NeurIPS).

Ben Kantor and Amir Globerson. 2020. Coreference resolution with entity equalization. Technical report.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 5739–5748.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Ronilda Lacson, Kimberly Harris, Phyllis Brawarsky, Tor D. Tosteson, Tracy Onega, Anna N. A. Tosteson, Abby Kaye, Irina Gonzalez, Robyn Birdwell, and Jennifer S. Haas. 2015. Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry. *Journal of Digital Imaging*, 28(5):567.

John Lafferty, Andrew McCallum, and Fernando C N Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 8(June):282–289.

Hao Li and Wei Lu. 2018. Learning with structured representations for negation scope extraction. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2:533–539.

Xiao Liu, Antoine Bordes, and Yves Grandvalet. 2015. Extracting biomedical events from pairs of text entities. *BMC Bioinformatics*, 16(S10):S8.

Jouni Luoma and Sampo Pyysalo. 2021. Exploring Cross-sentence Contexts for Named Entity Recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguis-*

tics, pages 904–914, Stroudsburg, PA, USA. International Committee on Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric de la Clergerie, Djame Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Stroudsburg, PA, USA. Association for Computational Linguistics.

Makoto Miwa, Rune Sætre, Jin Dong Kim, and Jun'Ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology*, 8(1):131–146.

Makoto Miwa, Paul Thompson, Ioannis Korkontzelos, and Sophia Ananiadou. 2014. Comparable study of event extraction in newswire and biomedical domains. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, pages 2270–2279.

Carlton Moore, Ashraf Farrag, and Evan Ashkin. 2017. Using Natural Language Processing to Extract Abnormal Results from Cancer Screening Reports. *Journal of patient safety*, 13(3):138.

Nicolas Paris, Matthieu Doutreligne, Adrien Parrot, Xavier Tannier, N Paris, M Doutreligne, A Parrot, and X Tannier. 2019. Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. In *TALMED 2019 : Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical*.

John X. Qiu, Hong Jun Yoon, Paul A. Fearn, and Georgia D. Tourassi. 2018. Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports. *IEEE Journal of Biomedical and Health Informatics*, 22(1):244–251.

Kirk Roberts, Yuqi Si, Anshul Gandhi, and Elmer V Bernstam. 2019. A framenet for cancer information in clinical narratives: Schema and annotation. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 272–279.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Guergana K Savova, Eugene Tseytlin, Sean Finan, Melissa Castine, Timothy Miller, Olga Medvedeva, David Harris, Harry Hochheiser, Chen Lin, Girish Chavan, and Rebecca S Jacobson. 2017. DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Research*, 77(21):e115–e118.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition.

Yuqi Si and Kirk Roberts. 2018. A Frame-Based NLP System for Cancer-Related Information Extraction.

Jackson M. Steinkamp, Charles Chambers, Darco Lalevic, Hanna M. Zafar, and Tessa S. Cook. 2019. Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. *Journal of Digital Imaging*, 32(4):554–564.

Pontus Stenetorp, Sampo Pyysalo, Goran Topíc, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based tool for NLP-Assisted text annotation.

Ricky K. Taira, Stephen G. Soderland, and Rex M. Jakobovits. 2001. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245.

Hai Long Trieu, Thy Thy Tran, Khoa N.A. Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. DeepEventMine: End-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem:5999–6009.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(SUPPL. 11):1–9.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Annotation scheme appendix

We detail here the annotation scheme and the resulting dataset. We focus on entities related to therapeutic (e.g. surgery) or diagnostic (e.g. mammography) procedures, radiological observations (e.g. cysts or masses), and breast density or ACR (or BI-RADS) cancer risk scores. The relevant entities to extract were the result of discussions with a physician expert in the field. The annotation scheme itself was the result of many iterations between annotations and scheme revision. The corpus consists of 120 annotated clinical documents, 80 for the training set and 40 for the evaluation set. The document-level statistics are detailed in Table 2.

A.1 Mention annotation

First, we annotate several types of mentions, each justifying the value of a field in a frame. In our scheme, each mention has an *effect* that can be combined with other effects to describe an entity. Some mentions have the effect of justifying the existence of a frame: we will refer to these mentions as "triggers". Other mentions have the effect of specifying an attribute of an object: we will refer to them as "attribute" mentions. No frame is created if there is no trigger, even if several attributes are present. In the example 1, the trigger [Ultrasound] mention has the effect of creating at least one "Diagnostic procedure" frame, whereas the [millimetric] attribute has the effect of giving a size to the frames that it is part of.

The trigger mention types are *ACR score*, *Breast density*, *Diagnostic procedure*, *Therapeutic procedure* and *Radiological lesion*. The additional attribute mention types are *Diagnostic procedure type*, *Therapeutic procedure type*, *Breast density type*, *ACR score type*, *Organ*, *Laterality*, *Temporality*, *Size*, *Distance*, *Angle* and *Breast quadrant*.

We have chosen to annotate mentions describing attributes (such as laterality or size) even if they are not part of any frame. On the other hand, trigger mentions are not annotated if they do not justify the presence of an object. In the sentence *"No suspicious mass on the right"*, only [right] is annotated as potentially justifying the laterality of an object, but not [mass] since it is preceded by a negation, and therefore does not justify the creation of any radiological lesion object.

Finally, each mention is classified, or normalized, according to a predetermined set of values. For example, a trigger mention "Breast density" may be labeled exclusively "type 1", "type 2", "type 3", "type 4". A laterality can take the values "left", "right", or "left + right".

A.2 Frame annotation

Frames describe conjunction of triggers and attributes that share their effect (or concept) on a given entity. In the above example, [8 o'clock radius] (applying an angle), [3cm] (applying a distance), [Left] (applying a laterality), [Breast] (applying an organ) and the trigger [cysts] (applying the effect of existing) share their respective effect on a same slice of an object. These mentions may be located in different sentences or paragraphs, and a field in a given frame may be justified by several mentions. On the other hand, if an object is described in several places in the text, we annotate it with several distinct frames. The notion of "several places" and the choice to split a same object into multiple frames is sometimes ambiguous. We choose to annotate a single frame for an object if it is described on several juxtaposed sentences, and split it into multiple frames otherwise. For instance, the [cysts] trigger is combined with the [nodules] trigger because they are found in juxtaposed sentences, and [nodules] is clearly referring to the previously mentioned [cysts].

All frames follow a specific scheme that constraints the set of labels and mentions (or effects) combinations. A summary of the frame schemes is shown in Figure 7. In practice, these constraints take the form of a list of 2502 label tuples that enumerates every possible mention / label combination. For example, a ACR Cancer Risk type 0 on the right breast at the time of the exam is described by the following tuple:

(acr_trigger, acr_type_0, temp_overlap,
organ_breast, lat_right)

As shown in the structured output 1 of example 1, five frames are annotated:

- the ultrasound "Diagnostic procedure" frame for its left location, composed of the [Breast], [ultrasound] and [left] mentions on lines 1 and 2

- the ultrasound "Diagnostic procedure" frame for its right location, composed of the [Breast], [ultrasound] and [right] mentions on lines 1 and 7

Diag. procedure 1	Frame 1		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[Ultrasound]		[Ultrasound]
type	ultrasound	[Ultrasound]	ultrasound	[Ultrasound]
organ	breast	[Breast]	breast	
laterality	left	[Left]:	right	[Right]:
temporality	overlap		overlap	

Lesion 1	Frame 3		Frame 5	
field	**value**	**justification**	**value**	**justification**
trigger		[cysts], [nodules]		Multiple [cysts]
organ	breast	[Breast]	breast	
laterality	left	[Left]:	left	on the [left]
temporality	overlap		overlap	
quadrant				
size		[millimetric]		
distance	30mm	[3 cm]		
angle	8	[8 o'clock radius]		

Lesion 2	Frame 4		Frame 5	
field	**value**	**justification**	**value**	**justification**
trigger		[cysts], [nodules]		Multiple [cysts]
organ	breast	[Breast]	breast	
laterality	left	[Left]:	left	on the [left]
temporality	overlap		overlap	
quadrant				
size		[millimetric]		
distance	20mm	[2 cm]		
angle	6	[6 o'clock radius]		

Table 6: All mentions, frames and objects extracted from the example 1

- the first "Finding" frame of the first nodule, with two trigger mentions: [cysts] and [nodules] and attribute mentions [8 o'clock position], [3cm] and [millimetric] on lines 1, 2, 3, 4 and 5
- the first "Finding" frame of the second nodule, with two trigger mentions: [cysts] and [nodules] and attribute mentions [6 o'clock position], [2cm] and [millimetric] on lines 1, 2, 4 and 5
- the second "Finding" frame of both nodules in the conclusion: composed of the trigger [cysts] and the laterality [left] on line 11

Since the mass negation on line 8 is not an indication of the presence of an object, we do not annotate it. The temporality of each frame overlaps the exam, although no explicit mention can support this fact, so we fill the temporality field of the frames with the value "overlap" and leave the justification empty.

A.3 Object annotation

Finally, the different frames are grouped into objects, although we do not extract them in the model presented in this study. Objects are union of frames. For a given set of concepts, multiple frames might be required to describe a same object. In the context of growing lesions, a union of multiple (temporality, size) conjunctions can represent the evolution. In an other setting with moving objects, a union of (temporality, localisation) labels could be used. In our case, as we represent lateralities with two exclusive "left" and "right" concepts, bilateral objects are described with two co-referent frames.

In the previous example, three objects are an-

Type	Field	Field value
ACR risk score	score trigger	
	score type	type 0/... type 6
	laterality	left/right
	temp	overlap/before
Breast density	dens. trigger	
	dens. type	type 1/... type 4
	laterality	left/right
	temp	overlap/before
Diag. proc.	diag. trigger	
	diag. type	mammo/mri/...
	organ	breast/other
	laterality	left/right
	temp	overlap/before/after
Ther. proc.	ther. trigger	
	ther. type	surgery/other
	organ	breast/other
	laterality	left/right
	temp	overlap/before/after
Lesion	lesion trigger	
	organ	breast/other
	laterality	left/right
	temp	overlap/before
	quadrant	lower inner/...
	size	
	distance	
	angle	

Table 7: Schemes of the extracted frames. Each frame is composed of multiple fields that can take a value.

notated, grouping two frames for the ultrasound procedure and two frames for each cyst. The last nodule frame in the conclusion is a case of plural coreference, since it its attributes apply to both objects. In this case, the frame describing several objects is added to each one. The statistics of objects in the annotated documents are described in Table 8. This step amounts to annotating coreferences between frames. We did not address this task of frame-coreference prediction in this study.

A.4 Annotation process

Clinical documents were de-identified automatically beforehand and the manual annotation was performed with BRAT (Stenetorp et al., 2012) by two annotators. 120 clinical reports were sampled from a from of query the APHP clinical data warehouse that combined the substrings "mamm" (to obtain breast related reports), "ACR" and "BI-?RADS" (to obtain ACR scores). Some sampled

	train		**val**	
	obj	frame	obj	frame
lesion	279	449	122	210
diagnostic proc.	285	795	141	379
therapeutic proc.	51	83	22	29
ACR score	152	152	82	82
breast density	98	98	52	52

Table 8: Frame and object statistics in the annotated corpus

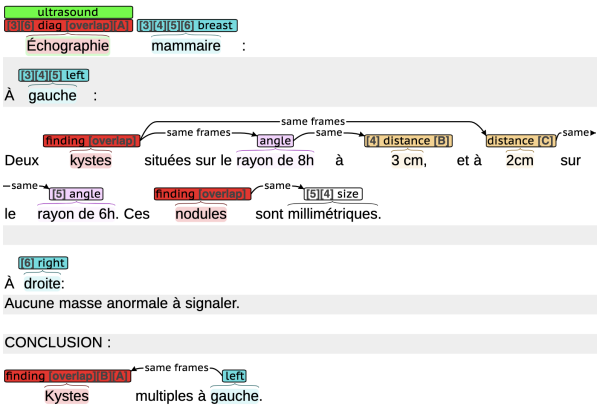

Figure 6: BRAT annotation of Example 1

reports were not breast radiology reports, yet we annotated and kept them as negative samples. Using the "Event" or "Relation" annotations in BRAT turned out to be impractical. We choose instead to annotate frames using a mix of identifier attributes (frame1, frame2, ...) on mentions, and relations on close-by mentions. Co-references, i.e., object annotation, were annotated using identifier attributes (objectA, objectB, ...) for the same reason. The BRAT annotations of Example 1 are shown in Figure 6. The direction of the annotated relations is only used to extract the paths along which the frames are clustered, but is not used as directed relation in our model, since it is not consistent.

B Hyperparameters appendix

We optimize the model weights with the Adam optimizer (Kingma and Ba, 2015) without weight decay and use a first learning rate lr_{BERT}, linearly decayed from 5×10^{-5} with a 10% warmup, for the pretrained CamemBERT (base) weights, and a second lr_{main}, linearly decayed from 5×10^{-4}, for the other parameters. The models were trained with a batch size of 16 for 2000 steps. The maximum wordpiece sequence size is 192, a dropout of 0.5 is applied on the output of BERT, and a dropout of 0.2 in the attention matrices computation. There

are 3 BiLSTM layers of hidden size 200. The loss weights are set to $\alpha_{\text{NER}} = 2$, $\alpha_{\text{normalization}} = 1$, $\alpha_{\text{relation}} = 1$, $\alpha_{\text{WSS}} = 1$, $\alpha_{\text{frame\_classification}} = 0.5$

C Relaxed retrieval metrics algorithms

Algorithm 1 Procedure to compute the maximum sum of greedily matched items between two sets of predicted and gold items P and G according to the MATCH_SCORE function

function MATCH_SUM(P, G, MATCH_SCORE)
  scores ← empty matrix ▷ *match scores*
  matched ← {} ▷ *matched (p,g) items*
  result ← 0 ▷ *aggregated score*
  for each predicted item p ∈ P **do**
    for each gold item g ∈ G **do**
      scores[p, g] ← MATCH_SCORE(p, g)
  while $|P\backslash\text{matched} \times G\backslash\text{matched}| > 0$ **do**
    Take the 1st p ∈ P\matched
    g ← $\text{argmax}_{g \in G\backslash\text{matched}}$(scores[p])
    if scores[p, g] > 0 **then**
      result ← result + scores[p, g]
      matched ← matched ⋃ {p, g}
  return result

Algorithm 2 Procedure for the approximate mention retrieval metric

function SCORE_NER(p, g)
  ▷ *return 1 if p and g have a word dice overlap ≥ 0.5 and the same label, 0 otherwise*
  return 2·|p.words ⋂ g.words| / (|p.words| + |g.words|) > 0.5 and p.label = g.label

function HALF_NER(P, G)
  tp ← MATCH_SUM(P, G, SCORE_NER)
  f1 ← 2 · tp/(|G|+|P|)
  return f1

Algorithm 3 Procedure to compute the Frame Support retrieval metrics

function OVERLAP(a, b)
  ▷ *return 1 if a and b share a word and have the same label, 0 otherwise*

function SCORE(p, g)
  ▷ *return the Dice score between spans (=mentions) of p and g, between 0 if there is no overlap and 1 if all mentions match)*
  tp ← MATCH_SUM(p.spans,g.spans,OVERLAP)
  return 2 · tp/(|g.spans| + |p.spans|)

function FRAME_SUPPORT(P, G)
  ▷ *return the retrieval metrics, where relaxed true positives between P and G are computed with* SCORE
  relaxed_tp ← MATCH_SUM(P, G, SCORE)
  f1 ← 2·relaxed_tp/(|G|+|P|)
  return f1

Algorithm 4 Procedure to compute the Frame Label retrieval metrics

function SCORE(p, g)
  ▷ *return 1 if all labels of g are in p, all labels of p are in g or a non conflicting frame of the same object and triggers overlap, 0 otherwise*

function FRAME_LABEL(P, G)
  ▷ *return the retrieval metrics, where true positives between P and G are computed with* SCORE
  tp ← MATCH_SUM(P, G, SCORE)
  f1 ← 2 · tp/(|G|+|P|)
  return f1