Event-independent temporal positioning: application to French clinical text

Nesrine Bannour, Bastien Rance, Xavier Tannier, Aurélie Névéol

Event-independent temporal positioning: application to French clinical text

Nesrine Bannour[1], Bastien Rance[2,3,4], Xavier Tannier[5], and Aurélie Névéol[1]

[1]Université Paris-Saclay, CNRS, LISN
[2]Inserm, CRC, UMRS 1138, Université de Paris, Université Sorbonne Paris Cité
[3]HeKA, Inria Paris, France
[4]Assistance Publique - Hôpitaux de Paris, Hôpital Européen Georges Pompidou
[5]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS

Abstract

Extracting temporal relations usually entails identifying and classifying the relation between two mentions. However, the definition of temporal mentions strongly depends on the text type and the application domain. Clinical text in particular is complex. It may describe events that occurred at different times, contain redundant information and a variety of domain-specific temporal expressions. In this paper, we propose a novel event-independent representation of temporal relations that is task-independent and, therefore, domain-independent. We are interested in identifying homogeneous text portions from a temporal standpoint and classifying the relation between each text portion and the document creation time. Temporal relation extraction is cast as a sequence labeling task and evaluated on oncology notes. We further evaluate our temporal representation by the temporal positioning of toxicity events of chemotherapy administrated to colon and lung cancer patients described in French clinical reports. An overall macro F-measure of 0.86 is obtained for temporal relation extraction by a neural token classification model trained on clinical texts written in French. Our results suggest that the toxicity event extraction task can be performed successfully by automatically identifying toxicity events and placing them within the patient timeline (F-measure .62). The proposed system has the potential to assist clinicians in the preparation of tumor board meetings.

1 Introduction

Temporal information extraction from text is a critical task in natural language processing (NLP) research, and it has been employed in a wide range of NLP applications, including narrative construction (Do et al., 2012; Ning et al., 2017; Han et al., 2019), temporal question answering (Llorens et al., 2015), and clinical text processing (Tourille et al., 2017c; Moharasan and Ho, 2019; Lin et al., 2020). Temporal information extraction was first addressed in the news article domain (Bögel et al., 2014; Chambers et al., 2014; Ning et al., 2017; Vashishtha et al., 2019). At the same time, there has been a significant interest in temporal information extraction from clinical narratives through the i2b2-2012 challenge (Sun et al., 2013) and the Clinical TempEval shared tasks (Bethard et al., 2015, 2016, 2017).

Narrative texts embedded in Electronic Health Records (EHR) contain essential temporal information, which can help better understand the clinical healthcare pathway. Temporal information extraction involves detecting events (EVENT), identifying temporal expressions (TIMEX), and extracting temporal relations between them. Several challenges arise when representing clinical temporal information (Najafabadipour et al., 2020). Temporal expressions vary widely, including domain-specific, non-standard, and abbreviated date expressions. Moreover, clinical narrative text is often ungrammatical and goes back and forth through time, making it difficult to link the events to temporal expressions. Sometimes, the time related to the clinical event is not explicitly specified. Redundant information in clinical text is another major problem when determining the chronology of events.

Temporal relation extraction denotes temporal ordering between text mentions, indicating events or temporal expressions. The TimeML (Pustejovsky et al., 2003) annotation scheme was initially developed to model general-domain events, temporal expressions, and their temporal links (TLINKs). THYME-TimeML (Pustejovsky and Stubbs, 2011; Styler IV et al., 2014), a similar annotation scheme adapted to the clinical domain, has been proposed, introducing another form of temporal relations linking clinical events to the Document Creation Time (DCT), namely DocTimeRel. Temporal relation extraction models evolved from rule-based models (Chang et al., 2013; Wang et al., 2016;

Najafabadipour et al., 2020) to machine learning-based (Tourille et al., 2016b; Chikka, 2016; Tourille et al., 2017c; Viani et al., 2019b) and deep learning-based models (Tourille et al., 2017a; Liu et al., 2019; Lin et al., 2020; Alfattni et al., 2021; Chokwijitkul et al., 2018). However, the performance of the proposed models is still insufficient for practical applications (Gumiel et al., 2021). Indeed, event modeling is task-specific and heavily dependent on the nature of the text, and annotating temporal relations in clinical texts is more challenging since it requires medical expertise, which is costly and time-consuming. As a result, inter-annotator agreement (IAA) is poor in available datasets (Ning et al., 2018), and annotated corpora are limited for non-English languages. Only a few studies, in particular, use French corpora (Tourille et al., 2016a, 2017c).

In this work, we introduce a novel event-independent representation of temporal relations in clinical texts that is task-independent and easily adaptable to different domains. Each narrative portion is assigned to a temporal category, describing its relation to the DCT. Unlike the DocTimeRel extraction task, we do not begin by identifying the clinical events and then extracting their relation to the DCT. Instead, we extract the temporal positioning of each text portion according to the DCT, regardless of events. The clinical events could then be identified depending on the task, and each event will have the same temporal positioning as the text portion that includes it. Our main contributions are the following:

- We propose a novel representation of temporal relations that allows us to identify homogeneous text portions from a temporal standpoint and to classify their temporal positioning, regardless of the domain and the extraction task. This leads to a task much faster and easier for human annotators, as well as more reproducible through different event types.
- To evaluate our temporal representation, we annotate a corpus of clinical reports written in French using the THYME-TimeML annotation scheme, and we define the temporal positioning extraction task as a sequence classification task. The classification model is compared to a rule-based baseline model.
- To validate the effectiveness of our temporal representation, we apply our classification model to another clinical corpus, identify the chemotherapy toxicity events in this corpus, and then infer the temporal positioning of these events according to the DCT.

2 Related Work

Prior works on temporal relation extraction in clinical text are based on hand-crafted rules. To define these rules, (Gaizauskas et al., 2006; Hernández et al., 2016; Viani et al., 2019a; Najafabadipour et al., 2020) used lexical and grammatical features such as part-of-speech (POS) tags, tense and aspect of events. Wang et al. (2016) proposed a model to extract relations between events and time expressions that relies only on the properties of relation entities without the need for extra grammatical information to ensure a minimal dependency on the text quality. Rule-based models need human expertise to create domain-specific rules, and such models are difficult to adapt to other domains.

Later, a variety of supervised Machine Learning approaches has been used, such as Support Vector Machine (SVM) classifiers and Conditional Random Fields (CRFs), with different sets of features, including syntactic, lexical, and semantic features (Lin et al., 2016; Tourille et al., 2016b, 2017c; Viani et al., 2019b; Barros et al., 2016; Chikka, 2016). Tourille et al. (2016a) presented a model based on the Random Forest (RF) algorithm, and Cohan et al. (2016) used a Logistic Regression (LR) classifier to extract the temporal relations between clinical events and the DCT. Velupillai et al. (2015) addresses the DocTimeRel task by using a CRF model and the narrative container relation sub-task by building a hybrid approach based on CRF and a rule-based technique. (Chang et al., 2013) also proposed a hybrid model combining a rule-based method with a maximum entropy model.

Deep neural networks have been used for temporal relation extraction in recent years. Li and Huang (2016) utilized Convolution Neural Networks (CNNs) to identify the relation between events and DCT. Dligach et al. (2017) proposed models based on CNNs and Long Short-Term Memory (LSTM) for extracting the event-event and event-time contains relations from the THYME corpus. Tourille et al. (2017a,b) used LSTMs to create inter-sentence and intra-sentence relation classifiers. Galvan et al. (2018) obtained state-of-art performance on the 2016 Clinical TempEval challenge for temporal relation extraction using a

tree-based LSTM model relying on dependency information. Liu et al. (2019) proposed an attention mechanism to enhance the overall performance of LSTM and GRU neural models for containment relations. Alfattni et al. (2021) investigated the attention mechanism built into a Bi-LSTM model on a broader set of temporal relations in clinical discharge summaries, including intra-sentence, cross-sentence, and DocTimeRel temporal relations. Lin et al. (2019, 2020) introduced BERT (Kenton and Toutanova, 2019) based models using the combination of global embeddings and multi-task learning to extract TLINKs and DocTimeRel relations jointly.

Supervised Machine Learning and Deep Learning models require large amounts of annotated data. However, few annotated corpora in the clinical domain are available in languages other than English. As a result, few research efforts addressed French corpora (Tourille et al., 2016a, 2017c).

3 Material and Methods

3.1 Overview of the temporal relation representation

As illustrated in Figure 1a, temporal relations in the text are often represented by DocTimeRel and TLINKs relations. The extraction of DocTimeRel refers to identifying events and classifying their temporal relations with the Document Creation Time (DCT). According to the THYME-TimeML scheme, each event will be assigned to one of the following categories: *Before (orange)*, *Before_Overlap (green)*, *Overlap (yellow)*, and *After (blue)*. However, since the events vary according to the task they are devised for, the DocTimeRel extraction task varies from domain to domain, and no generalization is possible. Additional challenges are also encountered in the definition of clinical events due to the complexity and the variety of medical terminologies prevalent in clinical narratives. The extraction of TLINKs relations starts with extracting possible pairs of events and temporal expressions. The most common strategy is to select the pairs in the same sentence and extract the intra-sentence temporal relations. Nevertheless, the characteristic of clinical text, such as the use of punctuation marks and the omission of sentence start and finish marks, make identifying sentence boundaries challenging. Moreover, other strategies must be adopted to resolve long-distance dependencies if the event and the temporal expression are in different sentences.

Therefore, we introduce a novel event-independent representation of temporal relations. As described in Figure 1b, homogeneous text portions from a temporal standpoint are identified and assigned to a category of the THYME-TimeML annotation scheme that reflects the relation with the DCT. Events will subsequently have the same temporal category as the text portion that includes them. Thus, we do not have to deal with sentence boundaries or long dependency issues. Although this representation is coarser than the traditional representation of temporal information, it is totally independent of the type of mentions to be extracted and, therefore, of the application domain. Figure 2 illustrates an example of our event-independent temporal positioning representation.

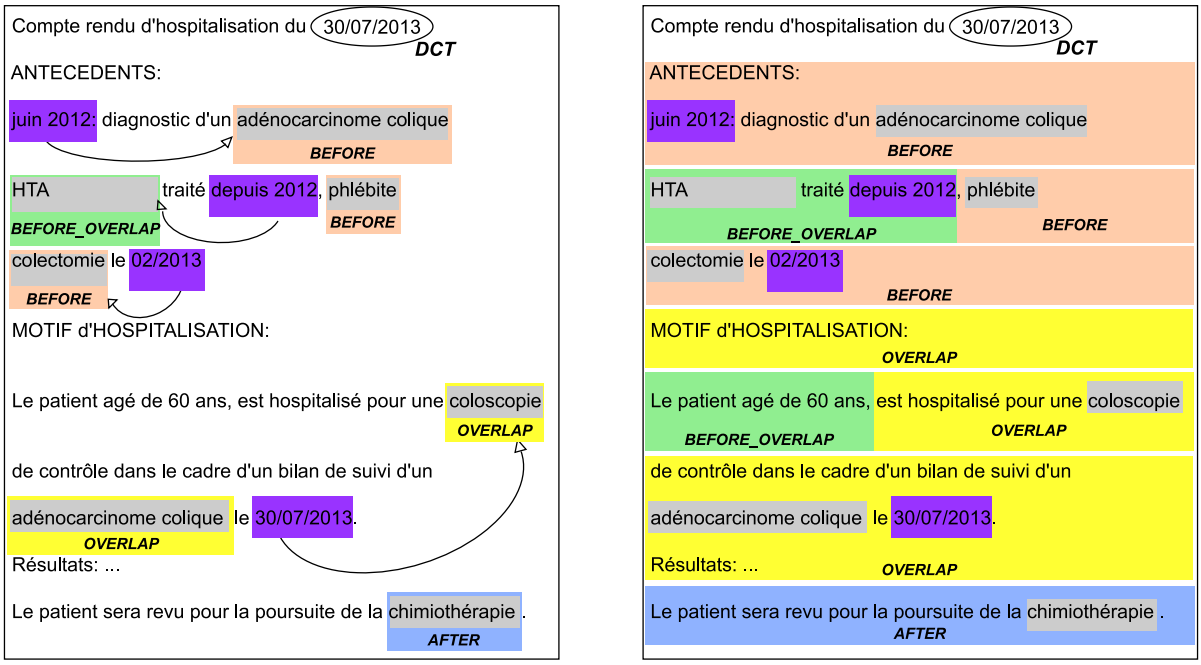
3.2 Corpora description

To develop and evaluate our temporal relation representation and our temporal positioning models, we use the following two clinical corpora[1]:

3.2.1 Temporal extraction corpus

This corpus is restricted and is built with randomly selected de-identified hospital, operative, and consultation reports of colon cancer patients from a French clinical data warehouse of the Georges Pompidou European Hospital (Jannot et al., 2017). We annotate 220 documents to train and validate our model and 57 documents for evaluation. We use the temporal categories of the THYME-TimeML annotation scheme and two more categories, namely *TemporalReference* and *End_Scope*. The *TemporalReference* category is used to identify the beginning of a clinical report associated with a new Document Creation Time (DCT), which is useful when multiple clinical reports are concatenated in the same document. The *End_Scope* category marks the end of a text portion if the following portion is a heading or signature. This only allows us to exclude these sections in the preprocessing step. Three annotators annotated a selection of 9 clinical documents. The inter-annotator agreements between annotator pairs in terms of macro F-measure are: 0.62, 0.73, and 0.69, which is higher than the agreement previously observed for temporal relations in clinical corpora in French and English (Tourille et al., 2017c). For our temporal positioning classification task, we will thus have these five categories:

[1] The scientific and ethical committee of AP-HP approved access to the clinical data (CSE21-15_TALONCO).



(a) Traditional representation of temporal information

(b) Event-independent temporal positioning

Figure 1: Temporal information representation. The DCT is surrounded, temporal expressions are represented in purple, events are represented in gray and encased by their DocTimeRel relations, and TLINKs are represented by arrows. Figure 1a illustrates the traditional representation of DocTimeRel between the DCT and the events and TLINKs between the events and the temporal expressions. Figure 1b depicts our representation of the temporal positioning of text portions according to the DCT, regardless of events. Translation of the mock narrative into English: "Discharge summary of 07/30/2013. PAST MEDICAL HISTORY: Adenocarcinoma of the colon was diagnosed in June 2012. Hypertension treatment was initiated in 2012. Phlebitis. Patient had large bowel resection on 02/2013. HISTORY OF PRESENT ILLNESS: This is a 60 y.o. male admitted on 07/30/2013 for a routine colonoscopy planned in the course of follow-up for known colon adenocarcinoma. RESULTS: ... The patient is scheduled for a new round of chemotherapy. "

TemporalReference, *Before*, *Before_Overlap*, *Overlap*, and *After*. The default temporal category for *TemporalReference* is *Overlap*. The detailed annotation guideline is provided in Appendix A.

3.2.2 Toxicity corpus

This corpus is restricted and is built with randomly selected de-identified hospital clinical reports containing toxicity information of chemotherapy administrated to colon and lung cancer patients from the same French clinical data warehouse as the temporal extraction corpus (Jannot et al., 2017). An expert manually validated the toxicity events annotations on 43 clinical documents. The annotation process is done using the BRAT annotation tool (Stenetorp et al., 2012). This corpus includes 16 documents regarding colon cancer and 27 about lung cancer and is used to validate the efficacy of our temporal positioning approach.

Table 1 presents descriptive statistics for each category in the temporal extraction training and test corpora.

3.3 Temporal relation extraction

With this representation, we cast temporal relation extraction as a supervised sequence labeling task. The main goal is to identify homogeneous text portions from a temporal standpoint and to classify each text portion into a pre-defined temporal category, describing its relationship with the Document Creation Time (DCT). We train a token classification model using the French model CamemBERT (Martin et al., 2020) from the HuggingFace transformers library (Wolf et al., 2020). We classify each token based on the BIO (Beginning-Inside-Outside) tagging scheme. Hence, the model can identify tokens that indicate a temporal shift in the clinical text. The model weights were optimized with Adam (Kingma and Ba, 2014) without weight decay for 20 epochs. The batch size was set to 32. All the models were trained using a GPU NVIDIA Quadro P5000.

	# text portions (test)	# text portions (train)
TemporalReference	57 (12.2%)	253 (10.3%)
Before	106 (22.7%)	562 (22.9%)
Before_Overlap	92 (19.70%)	476 (19.4%)
Overlap	165 (35.3%)	861 (35.1%)
After	47 (10.1%)	302 (12.3%)
Total	467	2454

Table 1: The number of text portions for each category in the temporal extraction training and test corpora.

3.4 Chemotherapy toxicity event extraction

For the first pre-annotation and extraction of chemotherapy toxicity events, we use a dictionary-based model consisting of a simple matching between the clinical corpus and a chemotherapy toxicity dictionary (Rogier et al., 2021), containing French toxicity terms from different terminologies. This model is built using the QuickUMLS (Soldaini, 2016) algorithm. The obtained pre-annotations, as previously stated, are manually verified and corrected by a domain expert.

3.5 Baseline model

We compared our model with a defined rule-based baseline model. We map entire sections to a temporal positioning, based on terms that are often used to denote medical sections, in particular in hospital and operative clinical reports such as "Antécédents" (*Case history*), "Indication" (*Indication*), "Gestes réalisés" (*Operative actions*), "Plan de traitement" (*Treatment plan*), etc. These keywords are typically useful for the temporal annotation process, even though they do not cover all types of clinical reports. This baseline model will be evaluated on the temporal extraction test corpus.

3.6 Evaluation metrics

In our work, we are interested in identifying temporal shifts between large text portions. In this case, segmentation into sentences and tokens is no longer needed. We evaluate the performance of our models at the character level by measuring the macro Precision, Recall, and F-measure.

True positives, false positives, and false negatives are denoted as TP, FP, and FN, respectively.

The three used evaluation metrics per category are defined below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 \times (Recall \times Precision)}{Recall + Precision}$$

Furthermore, using the *empirical bootstrap* method (Dekking et al., 2005, p.275), we compute the 95% confidence intervals of our classification results. For this, we sample our test corpus with replacement 1000 times. Evaluation metrics will be calculated for each sample.

To measure[2] the carbon footprint of training and testing our temporal positioning models, we use the Carbon tracker tool (Anthony et al., 2020) .

4 Results

Table 2 summarizes the overall results of the baseline model and our temporal positioning model. The best results are obtained with our model with an F-measure of 0.86, which is higher than the inter-annotator agreements. The results are much lower with the baseline model, with an F-measure of 0.35. The CO_2 emissions from training and testing our temporal positioning model are estimated to be 199 g.

Table 3 presents the detailed performance of our temporal positioning model over all categories on the temporal extraction test corpus.

Figure 2 shows a clinical text sample with predicted results of temporal positioning of homogeneous text portions.

Table 4 illustrates the toxicity events extraction performance, the results of event-independent temporal positioning of text portions, and the temporal positioning of toxicity events on the toxicity corpus. An F-measure of 0.59 is obtained for extracting toxicity events using the QuickUMLS algorithm with chemotherapy toxicity events. Our model achieves 0.8 of F-measure on extracting and temporal positioning the text narrative portions of the toxicity corpus. Table 4 also provides further performance

[2]Note that these estimates are approximative and are computed by using the World-wide average carbon intensity of electricity production in 2019.

	Precision	Recall	F-Measure	CO_2 eq (g.)
Baseline model	0.39 [0.33-0.46]	0.55 [0.48-0.61]	0.35 [0.29-0.41]	-
Temporal positioning model	**0.87 [0.84-0.90]**	**0.86 [0.83-0.90]**	**0.86 [0.84-0.89]**	199

Table 2: Overall results on the temporal extraction test corpus.

	P	R	F
TemporalReference	0.94	0.88	0.91
Before	0.82	0.90	0.86
Before_Overlap	0.79	0.76	0.77
Overlap	0.93	0.87	0.90
After	0.85	0.90	0.88
Overall	**0.87**	**0.86**	**0.86**

Table 3: Results per category for the temporal positioning model on the temporal extraction test corpus.

details based on the type of cancer described in the toxicity corpus documents. Our model yields better results on colon narrative portions than lung narrative portions (an F-measure of 0.81 vs. an F-measure of 0.79). For temporal positioning of the toxicity events, inferior results are obtained with an F-measure of 0.62.

5 Discussion

5.1 Performance of temporal positioning models

As reported in Table 2, our temporal positioning model outperforms the baseline model on the temporal extraction test corpus, with an exact macro F-measure of 0.86 vs. 0.35. Table 3 presents the results per category of our model. The most prevalent categories are the best predicted (see Table 3). Thus, an F-measure of 0.9 is obtained for the *Overlap* category, representing 35.1% of the training corpus, and 0.86 for the *Before* category, representing 22.9% of the training corpus. However, high F-measures are also reported for less represented categories such as *TemporalReference* (10.3% and an F-measure of 0.91), *After* (12.3% and an F-measure of 0.88). This may be due to the well-specified boundaries of these categories. The text portions with the *Before_Overlap* category are often sentences included in text portions with the *Before* category with a temporal indication that shows consistency in time, such as "depuis le" (*since the*) (see Figure 2). This temporal shift is not always predicted, and despite the coverage of the *Before_Overlap* category (19.4% in the training corpus) training corpus) the performance is lower (0.77 of F-measure). Except for the second 'Follow-up' text span in Figure 2, most homogeneous text portions are adequately retrieved and classified. In particular, the temporal shift between the *Before* and the *Before_Overlap* categories is well predicted. The text portion "patient de 56 ans dans le contexte de" has been correctly assigned to the *Before_Overlap* category. The two text portions beginning with "Suivi et évolution dans le service:" and "Suivi:" respectively, are on follow-up care. The first one depicts the follow-up during the hospital stay and is well classified into the *Overlap* category. However, the second text portion starting with 'Suivi:' is wrongly assigned to the *Overlap* category when, in fact, it should be assigned to the *After* category since we are discussing future follow-up after discharge, including future treatments and medications. Other mistakes may occur when predicting temporal categories. For instance, text portions starting with 'Soins post-opératoires' (*Post-operative care*) and 'Soins de support' (*Support care*) are about patient care. The first span, usually described in operative reports, discusses post-operative care and should be assigned to the *After* category. In contrast, the second statement, usually in discharge summaries, examines whether or not there is supportive care and should be classified as *Overlap*.

As previously stated, an F-measure of 0.35 is observed for the baseline model. Note that we do not use the *End_Scope* category to avoid the heading and signature sections in this baseline model since there is no defined term to identify such sections. Therefore, the precision of this model remains low. The *TemporalReference* category has poor precision because it specifies the start of a clinical report and is usually in the heading section. Moreover, we use the terms "Paris" and "Compte-rendu" (*report*). The first keyword usually indicates the start of consultation reports, as healthcare professionals begin by writing the location and date of the report. However, such terms may occur in various parts of the clinical text. The second keyword denotes the start of hospital and operative reports, which begin with a title such as "Compte rendu opératoire" (*Op-*

	Precision	Recall	F-Measure
Extraction of toxicity events	0.47 [0.39-0.55]	0.79 [0.72-0.85]	0.59 [0.52-0.66]
Temporal positioning of narrative portions	0.84 [0.82-0.86]	0.77 [0.74-0.81]	0.80 [0.76-0.83]
Colon narrative portions	0.88 [0.84-0.91]	0.77 [0.71-0.83]	0.81 [0.76-0.86]
Lung narrative portions	0.82 [0.79-0.85]	0.78 [0.75-0.82]	0.79 [0.75-0.82]
Temporal positioning of toxicity events	0.62 [0.55-0.89]	0.62 [0.55-0.88]	0.62 [0.54-0.88]

Table 4: Performance of extraction of toxicity events, event-independent temporal positioning of narrative portions, and temporal positioning toxicity events on the toxicity corpus.

erative report) or "Compte rendu d'hospitalisation" (*Hospitalization report*). Similar observations are obtained for the *After* category, which tends to be at the end of the clinical report and just before the signature part. The keywords used in the rule-based model do not cover the consultation reports, which contain narrative text describing the patient visit summary. As a result, the baseline model also suffers from a low recall rate.

We also test the performance of our model on the toxicity corpus as shown in Table 4. An overall F-measure of 0.8 is obtained, which is slightly lower than the performance on the temporal extraction corpus (an F-measure of 0.86). This might be due to differences in the cancer types described in the texts in each corpus. Indeed, our temporal positioning model was trained on the temporal extraction corpus, which only includes clinical reports of colon cancer patients, but the toxicity corpus contains clinical reports of both colon and lung cancer patients. As a result, the performance of temporal positioning clinical reports of colon cancer patients in the toxicity corpus is better than that of lung cancer patients in the same corpus as reported in Table 4 (an F-measure of 0.81 vs. an F-measure of 0.79). This good performance show that our model can adapt to other corpora, including other types of cancer.

5.2 Environmental impact

The carbon footprint of our event-independent temporal positioning model is reported in Table 2 in terms of CO_2 equivalent measure in grams. A total of 199 g of CO_2 emissions is estimated from training and testing our model, which is roughly equivalent to 1.85 km traveled by car based on CO_2 performance of new passenger cars in Europe[3]. Note that Carbon tracker fails to fetch the IP address and, therefore, to determine the geographic location dynamically. As a result, it uses the World-wide average carbon intensity of electricity production in 2019 (475 gCO_2/kWh) instead of the used value for France (around 58 gCO_2/kWh in 2021), which yields to overestimated CO_2 equivalent measures. Moreover, Carbon tracker does not take into consideration the execution environment or the technique of energy production. Thus, the obtained carbon footprint measures remain very approximative.

5.3 Performance of toxicity events extraction

As reported in Table 4, an F-measure of 0.59 is obtained for the toxicity event extraction using the quickUMLS algorithm. The toxicity events extraction model extracts all toxicity events in clinical text. However, we are solely interested in toxicity events related to chemotherapy treatments. As a result, the precision of this model remains low. For instance, if "HTA" (*hypertension*, high blood pressure) is included in the comorbidity medical section, we do not consider it as a toxicity event. However, if such event is mentioned while describing the toxicities of previous chemotherapy cures, it will be retained as a toxicity event.

It is also worth noting that we extract even the negated toxicity events. In fact, "anémie de grade 0" (*anemia of grade 0*) and "pas d'anémie" (*no anemia*) are synonyms for the absence of such toxicity event. However, extracting the toxicity event *anemia* is still important to propose better treatment strategies.

5.4 Temporal positioning of chemotherapy toxicity events

This experiment aims to determine how effectively we can recognize and characterize the temporal relation between toxicity events and the DCT. To address this question independently of how well event recognition can be achieved, we have used the gold standard toxicity event annotations, which

[3]https://www.eea.europa.eu/ims/co2-performance-of-new-passenger

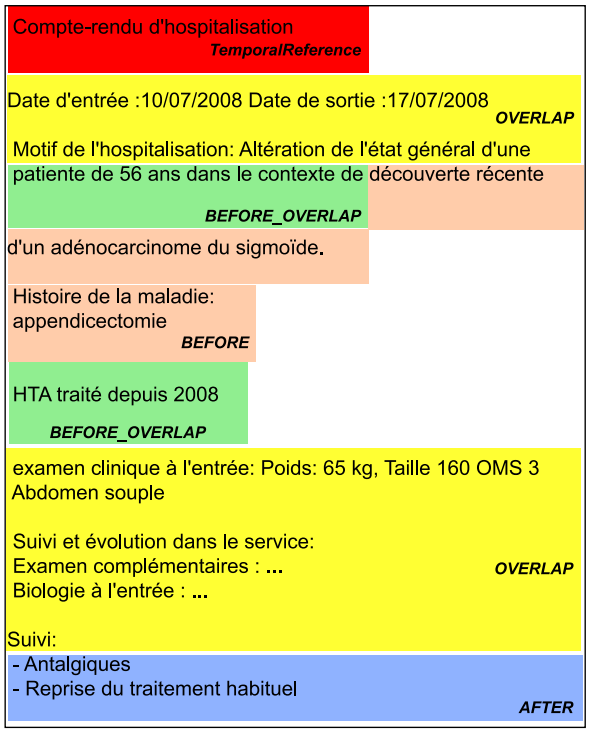


Figure 2: An example of predicted temporal positioning of text portions. Translation of text into English: *"Discharge summary. Admission date: 07/10/2008 Discharge date: 07/17/2008. Reason for admission: 56 y.o female presented with asthenia, weight loss and lack of appetite following the recent discovery of sigmoid adenocarcinoma. Past medical history: appendectomy Hypertension treatment was initiated in 2008. Physical examination on admission: Weight: 65 kg, Size 160 OMS 3 Abdomen was soft. Hospital course: Further medical exams: Tests on admission: ... Discharge instructions/Follow-up: - analgesics - patient should continue her usual care."*

are, therefore, 'perfectly' recognized. As reported in Table 4, an F-measure of 0.62 is obtained. Looking at the outcomes by category, the majority of toxicity events are temporarily well-positioned into the three categories *Before*, *Before_Overlap*, and *Overlap*. Nevertheless, in our toxicity corpus, just one toxicity event matches the *After* category. This event, mentioned in a hypothesis statement, is incorrectly positioned as a *Overlap* category. As a result, the performance in terms of macro F-measure is a bit low (vs. a micro F-measure of 0.82). The good performance of temporal positioning of chemotherapy toxicity events validates the efficacy of our event-independent temporal representation of temporal information.

6 Conclusion

In this paper, we introduced a novel event-independent representation of temporal relations that is task-independent and, hence, domain-independent. The temporal relation classification problem is cast as a sequence token classification task using our representation. The main goal of this classification task is to identify homogeneous text portions and to classify them into temporal categories reflecting their relations with the document creation time. To develop and evaluate our model, we annotate a corpus of clinical reports written in French using the THYME-TimeML annotation scheme. Our temporal positioning model yields good results when recognizing and categorizing text portions. Moreover, experiments on the temporal positioning of chemotherapy toxicity events for patients with colon and lung cancers have also shown that good results could be achieved using our representation of temporal relations. This problem modeling might be the initial step toward constructing a patient timeline to order all its medical events.

Limitations

In our work, we manually annotated small portions of corpora. Such limited size is justified by the time-consuming task of temporal annotations and the requirement of expertise for toxicity event annotations. Although our temporal representation seems to perform well with other clinical reports containing information about a different type of cancer from that on which it was trained (e.g. lung cancer vs. colon cancer), such results must be validated on clinical reports containing information about additional cancer types. Additional experiments are also needed to validate the generalizability of our event-independent representation, such as evaluating it on other hospital or data warehouse clinical reports with various structures and evaluating it on other extraction tasks with different event definitions.

Ethics Statement

This study uses de-identified clinical data with the approval of the partner French hospital scientific and ethical committee (IRB equivalent). This work might be recommended to clinicians as a useful tool for assisting in the systematic analysis of large patient records. Unfortunately, the annotated cor-

pora developed in this work cannot be shared with the community due to confidentiality restrictions.

Acknowledgements

We would like to thank the scientific and ethical council of the AP-HP health data warehouse and the Georges Pompidou European Hospital, who gave us access to the corpora used in this work. Nesrine Bannour received funding from the ITMO Cancer Aviesan. Bastien Rance is supported by the SIRIC CARPEM program. The authors also acknowledge the support of ANR under grant CODEINE ANR-20-CE23-0026-01.

References

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2021. Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *Journal of Biomedical Informatics*, 123:103915.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. ArXiv:2007.03051.

Marcia Barros, Andre Lamurias, Gonçalo Figueiro, Marta Antunes, Joana Teixeira, Alexandre Pinheiro, and Francisco M. Couto. 2016. ULISBOA at SemEval-2016 task 12: Extraction of temporal expressions, clinical events and relations using IBEnt. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1263–1267, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Thomas Bögel, Jannik Strötgen, and Michael Gertz. 2014. Computational narratology: Extracting tense clusters from narrative texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 950–955, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2013. Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics*, 46:S54–S62.

Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1237–1240, San Diego, California. Association for Computational Linguistics.

Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. 2018. Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In *Proceedings of the BioNLP 2018 workshop*, pages 18–27, Melbourne, Australia. Association for Computational Linguistics.

Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. Guir at semeval-2016 task 12: Temporal information processing for clinical narratives. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1248–1255.

Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. 2005. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687, Jeju Island, Korea. Association for Computational Linguistics.

Rob Gaizauskas, Henk Harkema, Mark Hepple, and Andrea Setzer. 2006. Task-oriented extraction of temporal information: The case of clinical narratives.

In *Thirteenth International Symposium On Temporal Representation And Reasoning (time'06)*, pages 188–195. IEEE.

Diana Galvan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2018. Investigating the challenges of temporal relation extraction from clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Yohan Bonescki Gumiel, Lucas Emanuel Silva e Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, and Deborah Ribeiro Carvalho. 2021. Temporal relation extraction in clinical texts: A systematic review. *ACM Computing Surveys (CSUR)*, 54(7):1–36.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Eddie Paul Hernández, Alexandra Pomares Quimbaya, and Oscar Mauricio Muñoz. 2016. Htl model: A model for extracting and visualizing medical events from narrative text in electronic health records. In *ICT4AgeingWell*.

Anne-Sophie Jannot, Eric Zapletal, Paul Avillach, Marie-France Mamzer, Anita Burgun, and Patrice Degoulet. 2017. The georges pompidou university hospital clinical data warehouse: A 8-years follow-up experience. *International Journal of Medical Informatics*, 102:21–28.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Peng Li and Heng Huang. 2016. UTA DLNLP at SemEval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1268–1273, San Diego, California. Association for Computational Linguistics.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2016. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, 23(2):387–395.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 70–75, Online. Association for Computational Linguistics.

Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. 2019. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 134–139, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé 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, Online. Association for Computational Linguistics.

Gandhimathi Moharasan and Tu-Bao Ho. 2019. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, 3(2):220–244.

Marjan Najafabadipour, Massimiliano Zanin, Alejandro Rodríguez González, María Torrente, Beatriz Nuñez García, Juan Luis Cruz Bermudez, Mariano Provencio, and Ernestina Menasalvas Ruiz. 2020. Reconstructing the patient's natural history from electronic health records. *Artificial intelligence in medicine*, 105:101860.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Portland, Oregon, USA. Association for Computational Linguistics.

Alice Rogier, Adrien Coulet, and Bastien Rance. 2021. Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from EHRs. In *Medinfo 2021 - 18th World Congress on Medical and Health Informatics*, Virtual conference, Australia.

Luca Soldaini. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20 5:806–13.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016a. Extraction de relations temporelles dans des dossiers électroniques patient (extracting temporal relations from electronic health records). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, pages 459–466, Paris, France. AFCP - ATALA.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016b. Limsi-cot at semeval-2016 task 12: Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1136–1142.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2017a. Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230, Vancouver, Canada. Association for Computational Linguistics.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017b. Limsi-cot at semeval-2017 task 12: Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017c. Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 739–745, Valencia, Spain. Association for Computational Linguistics.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy Chapman. 2015. BluLab: Temporal information extraction for the 2015 clinical TempEval challenge. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 815–819, Denver, Colorado. Association for Computational Linguistics.

Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Robert Stewart, Rashmi Patel, and Sumithra Velupillai. 2019a. Annotating temporal relations to determine the onset of psychosis symptoms. In *MedInfo*, pages 418–422.

Natalia Viani, Timothy A Miller, Carlo Napolitano, Silvia G Priori, Guergana K Savova, Riccardo Bellazzi, and Lucia Sacchi. 2019b. Supervised methods to extract clinical events from cardiology reports in italian. *Journal of biomedical informatics*, 95:103219.

Wei Wang, Kory Kreimeyer, Emily Jane Woo, Robert Ball, Matthew Foster, Abhishek Pandey, John Scott, and Taxiarchis Botsis. 2016. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. *Journal of Biomedical Informatics*, 62.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Temporal Annotation scheme for our clinical corpus

A.1 Definitions of temporal categories

To annotate the temporal information in a clinical report, we define a temporal annotation scheme based on the Document Creation Time (DCT) and the possible categories of the Document creation Time Relation (DocTimeRel). The DCT might be the current medical visit date, usually stated in the document heading. It might also be the length of time spent in the hospital. The DCT does not need to be annotated.

A.1.1 Document creation Time Relation

Document creation Time Relation is the relation between events and Document Creation Time. We consider these four possible categories for this time relation: Before, Before_Overlap, Overlap, and After. We annotate only the first word of each temporal portion. We consider that the start of a temporal portion denotes the end of the previous one.

A.1.2 Before

The Before category is used to annotate narrative portions referring to what occurred before the Document Creation Time.

Examples

- Antécédents, antécédents médicaux, antécédents chirurgicaux, Antécédents familiaux, Histoire de la maladie, Rappel clinique, Rappel sur la pathologie → All terms referring to the medical history section.
- **Except:** Maladie traitée depuis le → Before_Overlap since we have a temporal indication that the procedure/disease is still ongoing for the patient (cf. Figure 3).

A.1.3 Before_Overlap

The Before_Overlap category is used to annotate narrative portions that started before the document creation time and are still ongoing at that time.

Examples

- Comorbidités, Mode de vie, Autonomie, traitement habituel, traitement à l'entrée, Allergies, Traitements concomitants, Facteurs de risque, Indication, Indication opératoire, décision d'une intervention, Tolérance inter-cure
- Patient de 70 ans
- HTA traitée depuis, dans le cadre d'un suivi d'un cancer → The patient is still suffering from the disease.
- METASTASES HEPATIQUES D'UN ADENOCARCINOME → The disease's name as a title in operative reports, which is generally capitalized (cf. Figure 4).

A.1.4 Overlap

The Overlap category is used to annotate narrative portions that happen at the same time as the document creation time.

Examples

- Examen pratique, Au total, Conclusion, Gestes opératoires, Gestes réalisés, Motif d'hospitalisation, Biologie, Biologie de sortie, INTERVENTION, constantes à l'arrivée, Date d'hospitalisation, Date d'entrée, Date de l'intervention, Motif
- Examens complémentaires, Examens paracliniques → Sometimes, some complementary exams are conducted before the document creation time but because they are done for the purpose of the hospital stay, we annotate them as Overlap (cf. Figure 3).
- Je vois ce jour, Je revois en consultation

A.1.5 After

The After category is used to annotate narrative portions referring to what occurs after the document creation time.

Examples

- Traitement de sortie, Prochains rendez-vous, Rendez-vous à venir, Prescription de médicaments, Date de la prochaine cure, Ordonnance de sortie, Prochains examens
- Je reverrai ce patient, je prévois une coloscopie
- La pièce est envoyée pour un examen histologique

A.2 Other categories

A.2.1 TemporalReference

Because several medical reports might be written in the same document, the TemporalReference category specifies the beginning of a new clinical report. Because several medical reports might be written in the same document, the TemporalReference category specifies the beginning of a new clinical report. Each clinical report will then have its own Document Creation Time, and the annotations will be based on this DCT. The TemporalReference category's default Document Time Relation is assumed to be Overlap and does not need to be annotated.

Examples

- Compte-rendu opératoire, Compte-rendu d'hospitalisation, Paris, le 14 octobre 2018

A.2.2 End_Scope

We do not consider heading and signature information in our annotation. Therefore, we use the category End_Scope to mark the ending of a narrative portion if the next narrative portion is a heading or a signature. This way, we avoid annotating the contact information for the health care unit, which may be repeated in several clinical reports. Despite the fact that the clinical documents are de-identified, we avoid annotating specific patient information. In cases other than headings or signatures, the end of a temporal portion is implicitly considered the start of a new temporal portion.

A.3 Examples of annotations made in accordance with the above scheme and guidelines

Annotations of the first example (cf. Figure 3)

- From *Compte* to *d'hospitalisation* as TemporalReference
- From *Hospitalisé* to 30/07/2013 as Overlap
- From *Motif* to *d'HOSPITALISATION :* as Overlap, note that we don't annotate the temporal portion after the End_Scope containing contact information of doctors
- From *HISTOIRE* to *ANTECEDENTS* as Before
- From *HTA* to *2012,* as Before_Overlap since we have a temporal indication that the disease is still ongoing for the patient
- From *phlébite* to *07/2012* as Before since it's part of the medical patient history
- From *ALLERGIES* to *Autonome* as Before_Overlap
- From *Examens* to *et* as Overlap despite the fact that the medical exams are conducted before the date of hospital admission
- From *sera* to *10/09/2013* as After. The signature of the document after the End_Scope category is not annotated

Figure 3: A first example of hospital report annotations

Annotations of the second example (cf. Figure 4)

- From *COMPTE* to *OPERATOIRE* as Temporal Reference
- *ADENOCARCINOME* as Before_Overlap
- *COLECTOMIE* as Overlap
- From *Rappel* to *clinique:* as Before
- From *Indication* to *opératoire.* as Before_Overlap
- From *Gestes* to *réalisés:* as Overlap

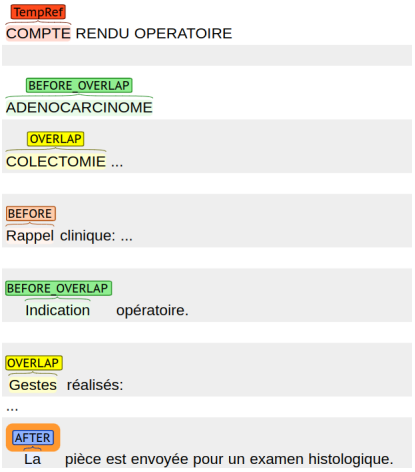


Figure 4: A second example of annotating an operative report

- From *La* to *histologique.* as After

Annotations of the third example (cf. Figure 5)

- From *Paris* to *2014,* as TemporalReference
- From *Je* to *jour* as Overlap
- From *Monsieur* to *comme* as Before_Overlap for the patient's age and since it is stated that the purpose of the medical visit is a disease follow-up
- From *antécédent* to *Rappel:* as Before
- From *Examen* to *pratique:* as Overlap
- From *A* to *mois* as After
- From *Dossier* to *staff* as TemporalReference, it's a new clinical report
- From *Dernières* to *2014:* as Before, based on the document creation time of the second clinical report.
- From *Décisions* to *staff:* as Overlap
- From *Le* to *consultation.* as After

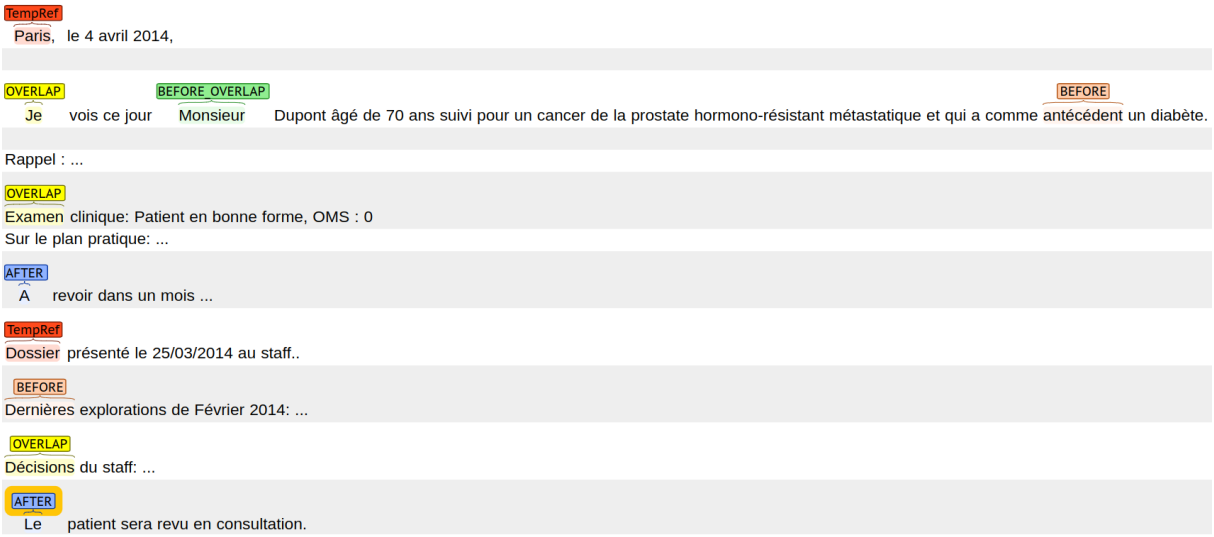


Figure 5: A third example of annotating a clinical document containing two clinical reports