Information Extraction from Electronic Health Records : Studies on temporal ordering, privacy and environmental impact

Nesrine Bannour

université PARIS-SACLAY

Information Extraction from Electronic Health Records: Studies on temporal ordering, privacy and environmental impact

Extraction d'Informations à partir des Dossiers Patients Informatisés : Etudes en temporalité, confidentialité et impact environnemental

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire Interdisciplinaire des Sciences du Numérique (Université Paris-Saclay, CNRS)**, sous la direction de **Aurélie NÉVÉOL**, Directrice de Recherche, et le co-encadrement de **Bastien RANCE**, Maître de conférences des Universités - Praticien Hospitalier, **Xavier TANNIER**, Professeur des Universités

Thèse soutenue à Paris-Saclay, le 30 Novembre 2023, par

Nesrine BANNOUR

Composition du jury

Membres du jury avec voix délibérative

Fatiha SAIS Professeure des Universités, Université Paris Saclay	Présidente
Maxime AMBLARD Professeur des Universités, Université de Lorraine	Rapporteur & Examinateur
Timothy MILLER Assistant Professor, Harvard University, Boston Children's Hospital	Rapporteur & Examinateur
Fleur MOUGIN Professeure des Universités, Université de Bordeaux	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASG082

Acknowledgments

I want to take some time to express my gratitude to all of the people who supported and helped me during my three years of PhD and without whom the journey would not have been possible.

First of all, I would like to thank my wonderful mom, Najiba, who has always been there for me and never stopped encouraging me in whatever I do. Without her, I would not be able to succeed in life. I would also like to thank my three sisters, Sondes, Ines, and Fetia, who were good examples to follow in their studies and everyday life. You constantly inspire me and help me overcome obstacles; thus, I consider myself extremely lucky to have you in my life. I also want to express my gratitude to my brothers-in-law Sami, Hassene, and Lotfi, especially Sami, who always had insightful advice to share and who motivated me to keep moving forward despite many encountered hurdles. A special thanks to my niece Sarah and my nephews Kaysser, Sinan, and Asser for their encouragement and understanding of why their auntie couldn't always play with them.

Besides my family, I had the support of my friends. More specifically, and in no particular order, thank you Oumaima, Mouhanned, Ala, Amen, Ahmed, and Amine for all the laughter and moments we had together, which helped me to forget about the stress I was under. I also want to thank my two best friends of a lifetime, Asma and Sabrine who have never let me down and have been an indispensable presence during this PhD and for more than ten years now. Words can never express my gratitude for having you in my life.

During this PhD, I was fortunate to make new friends, who made my daily life more pleasant. I would like to thank, in no particular order, Camille, Nicolas, Marco, Juan, Atila, Sofiya, Juliette, Paul, Mathilde, Lisa, Emmanuella, Mathieu, Armand, Iskandar, Nadège, Sophie, Katherine, Paritosh, Claire, François, Hugo, as well as all the people that I have met and with whom I shared brief or long discussions. A special thanks to those with whom I enjoyed an amazing trip to Chamonix, full of good vibes, laughing, and unforgettable memories. Thank you for helping me through my tough times and for making this experience joyful after a challenging first year due to the Covid lockdown.

I would like to express my deepest gratitude to my Master's supervisor, Gaël Dias, for his generosity and outstanding supervision, which enabled me to conduct research in the NLP field. I am also grateful for his guidance, support, and attempts to offer me several PhD opportunities. Sadly, we haven't had the chance to collaborate during this PhD, but I'm confident we'll do so in the near future. And, of course, thanks to my PhD supervisors, Aurélie, Bastien, and Xavier, for their constant guidance, professional expertise, and support throughout these years. Without you, I wouldn't have been able to complete this PhD successfully.

I am also grateful to my jury members, Maxime Amblard, Tim Miller, Fleur Mougin, and Fatiha Sais, for agreeing to spend valuable time reading and evaluating my thesis. I also want to thank everyone I have had the opportunity to cooperate and work with, especially Perceval, Sahar, Anne-Laure, and the members of the HeKa research team.

Finally, I would like to acknowledge the ITMO Cancer Aviesan for funding my PhD, the Georges Pompidou European Hospital, and everyone who made available and granted access to the resources we used in our experiments.

Résumé

Extended French summary

Les dossiers patients informatisés (DPIs) contiennent des informations importantes sur les patients. L'extraction automatique de ces informations est cruciale car elle permet d'améliorer la prise de décision pour les soins médicaux et la recherche clinique. Cependant, la plupart de ces informations résident dans des données textuelles non structurées. La tâche d'extraction d'informations inclut l'extraction des entités cliniques telles ques les maladies, les symptôtmes, les traitements, etc, via la reconnaissance d'entités nommées (REN) ainsi que l'extraction des relations temporelles entre les entités dans le but de construire une chronologie patient.

Cette thèse présente un travail pluridisciplinaire qui se situe au carrefour des domaines du Traitement Automatique des Langues (TAL) et de l'informatique médicale. Notre travail s'est appuyé sur des méthodes d'apprentissage pour la reconnaissance de séquence et a été guidé par le cadre applicatif de l'oncologie. Dans cette thèse, nous nous concentrons sur la REN et l'extraction de relations temporelles pour le domaine clinique, avec des questions annexes sur la confidentialité des données, l'apprentissage automatique avec peu de ressources et l'impact environnemental des approches de TAL. Un des objectifs principaux de notre travail est de proposer des méthodes et des ressources pour la recherche clinique en français.

De bons résultats ont été obtenus en utilisant des modèles neuronaux pour plusieurs tâches de TAL, notamment l'extraction d'informations. Néanmoins, ces modèles nécéssitent de quantités importantes de données annotées. Le processus d'annotation est long, coûteux et nécessite une expertise dans le domaine, ce qui limite la disponibilité des corpus annotés, en particulier pour les langues autres que l'anglais. De plus, le texte clinique est complexe, peu formel, comportent une variété de terminologie médicale, des informations temporelles implicites, ambigues et spécifiques au domaine, ainsi que plusieurs entités imbriquées. Cela rend le processus d'annotation et d'extraction plus difficile. Le traitement des textes cliniques présente des défis supplémentaires dû à leur caractère confidentiel. Par conséquent, le partage de données est difficile et strictement encadré par des réglementations telles que RGPD. Les performances ne sont donc pas encore suffisantes pour la pratique.

Dans ce contexte, nous étudions l'apprentissage par mimétisme pour la REN dans les rapports cliniques écrits en français, en utilisant des corpus publics et privés. Le principe de l'apprentissage par mimétisme consiste à annoter des données publiques non étiquetées à l'aide d'un *modèle enseignant* privé qui a été entraîné sur les données sensibles originales. Les données publiques nouvellement étiquetées sont ensuite utilisées pour entraîner des *modèles élèves*. Ces modèles peuvent être partagés sans révéler les données privées d'origine ou exposer le modèle privé construit avec ces données. Notre architecture de modèles préservant la confidentialité des données permet aux institutions hospitalières de générer

des modèles partageables, en particulier lorsqu'aucun corpus annoté n'est disponible publiquement. Nous montrons que notre stragéie offre un bon compromis entre la performance et la préservation de la confidentialité.

Notre seconde contribution concerne l'extraction des relations temporelles, reliant des évènements à un ancrage temporel représenté par des expressions temporelles. Cette tâche est difficile, très spécifique au domaine d'application et nécessite des corpus bien annotés. La tâche se révèle encore plus complexe pour le domaine clinique, car le texte clinique va et vient dans le temps, décrivant plusieurs évènements survenus à des moments différents. Dans certains cas, le moment associé à un évènement n'est même pas explicitement mentionné. De plus, la performance des systèmes d'extraction de relations temporelles dépend largement de la performance d'extraction des évènements. Or, la définition des évènements est très spécifique au domaine et aucune généralization n'est possible. Pour cela, nous nous intéressons à une simplification de l'extraction des relations temporelles en proposant une nouvelle représentation des relations temporelles, qui est indépendente des évènements et donc du domaine d'application. Le but de cette représentation est d'identifier des portions de textes homogènes du point de vue temporel et de classifier la relation temporelle de chaque portion de texte avec la date de création du document. L'annotation et l'extraction des relations temporelles sont ainsi plus faciles et plus reproductibles à travers différents types d'événements, vu qu'il n'est pas nécessaire de définir et d'extraire les événements au préalable. Nous évaluons notre représentation par le positionnement temporel des événements de toxicité des chimiothérapies décrits dans des rapports cliniques d'oncologie rédigés en français. Nos résultats suggèrent que ces événements pourraient être placés avec succès dans la chronologie du patient par la suite.

En plus de ces contributions, nous proposons une étude préliminaire des principaux outils existants pour calculer l'empreinte carbone des modèles de TAL, en particulier les modèles à base d'apprentissage profond. Notre objectif est de fournir une analyse comparative de l'utilisation de ces outils en comparant les mesures qu'ils produisent. Nous utilisons un des outils étudiés pour calculer l'empreinte carbone de tous les modèles proposés au cours de la thèse, en considérant qu'il s'agit d'un premier pas vers la sensibilisation et la production de modèles plus efficaces avec de faibles émissions de carbone.

Dans l'ensemble, nous pensons que notre travail apporte des contributions à la recherche clinique en français grâce aux méthodes et aux ressources que nous mettons à la disposition des cliniciens. En particulier, nous proposons de nouvelles architectures et représentations qui facilitent l'extraction d'informations cliniques pour une application pratique plus efficace.

Contents

List of Figures

List of Tables

Glossary

AI Artifical Intelligence. 68–70, 74, 84

APHP Assistance Publique des Hopitaux de Paris (Public Paris hospitals). 113

BERT Bidirectional Encoder Representations from Transformers, a language model based on the transformer architecture. 35, 43, 45, 50, 62, 63, 90, 92

Bi-LSTM Bidirectional Long Short Term Memory network, a recurrent neural network which is composed of two LSTMs, allowing to have both forward and backward input information. 43, 60–62, 90, 103

BoW Bag of Words, a model that converts text into fixed-length vectors without any information about word order. 33

BPE Byte Pair Encoding, a compression algorithm method that iteratively merges the most common subwords. 35

CBOW Continuous Bag-Of-Words, a neural network that predicts a center word based on its given context words. 34

CNN Convolutional Neural Network, a feed-forward neural network that enables greater extraction of features from input (text or image). 43, 44, 51, 52, 61, 90, 103

CoNLL Conference on Computational Natural Language Learning. 37, 40, 43, 92

CPU Central Processing Unit. 68, 77, 78, 80

CRF Conditional Random Fields, a discriminative model that models the dependency between variables by considering the neighboring contextual information. 40–43, 46, 50–52, 59, 60, 63, 90, 91, 103

DCT Document Creation Time. 28, 48, 53, 55–58, 61, 64, 110, 111, 113, 114, 123–125, 128, 135–137

Deep Learning Deep Learning, a class of machine learning methods that seeks to develop multi-layered neural networks. 24–28, 33, 35, 43, 50, 64, 68–71, 73, 78, 86, 127, 131

DocTimeRel Document Time Relation, a temporal relation between an event and the document creation time. 53, 55–62, 64, 111, 113, 124, 135, 137

DPI Dossier Patient Informatisé (Electronic Health Record). 5, 199

DRAM Dynamic Random Access Memory, a temporary memory for your computer that stores data for quick, short-term access. 78

DT Decision Tree, a supervised Machine Learning approach whose purpose is to predict the value of a target variable based on multiple input variables. 40, 41, 60

EHR Electronic Health Record, a digitized version of a patient's health information over time. 23, 39, 62, 63, 65, 66, 86, 110, 124, 130, 200

ELMo Embeddings from Language Models, contextualized word embeddings. 35, 45

GDPR General Data Protection Regulation, a European Union regulation on Information privacy in the European Union (EU) and the European Economic Area (EEA). 46, 86

GLOVE Global Vector for Word Representation, an unsupervised learning algorithm that generates distributed word embeddings by encoding how often two words appear within a given window. 34

GPT Generative Pre-trained Transformer, a type of Large Language Models relying on deep learning to produce human-like texts from a given text input. 35, 44, 67, 69, 89

GPU Graphical Processing Unit. 68, 77, 78, 80, 83, 94, 117

HEGP Hôpital européen Georges-Pompidou (Georges Pompidou European Hospital). 25, 105, 106, 115

HMM Hidden Markov Model, a probabilistic generative model that describes the probabilistic relationship between a set of observations and a set of hidden states. 40–42

ICD10 10th revision of the International Statistical Classification of Diseases and Related Health Problems, which is a medical classification list by the World Health Organization (WHO). 38, 39, 88

IDF Inverse-Document-Frequency, a measurement of the proportion of documents in the corpus that include a given word. 33

IE Information Extraction, an NLP task for automatically extracting structured information from unstructured documents. 23–29, 32, 36, 62, 86, 127, 130

LISN Laboratoire Interdisciplinaire des Sciences du Numérique (LISN lab). 25, 84

LLM Large Language Model, a large-scale generative deep learning model aimed to interpret and produce natural language. 63, 67, 130, 131

LSTM Long Short Term Memory network, a recurrent neural network that can capture long-term dependencies. 43, 50, 52, 60, 61

ME Maximum Entropy, a discriminative model that maximizes the entropy of the data to generalize as much as possible for the training data. 40–42, 59

MEMM Maximum Entropy Markov Model, a discriminative model combining features of Hidden Markov and Maximum Entropy models. 42

ML Machine Learning, a subfield of Artificial Intelligence, seeking to develop systems that could learn and make predictions. 23–26, 48, 50, 59, 60, 62, 63, 69, 70, 78, 86, 106

MLM Masked Language Model, a trained model that predicts a missing token in a sequence using the context given by the words around it. 65

MLP MultiLayer Perceptron, a feed-forward neural network with multiple connected layers. 52, 61

MRC Machine Reading Comprehension, a branch of NLP in which machines are trained to understand and respond to queries about unstructured text. 44

NB Naive Bayes, a supervised learning algorithm based on Bayes' Theorem. 59

NE Named Entity. 36, 40, 41

NER Named Entity recognition, an NLP task for identifying and classifying key information from text. 23–28, 32–34, 36–48, 50, 52, 60, 66, 67, 70, 71, 74, 78, 80, 83, 84, 86, 87, 89–92, 94, 96, 97, 101–103, 106, 127–131, 200

NLP Natural Language Processing, a field that combines linguistics and computer science which allows computers to understand human language. 23–25, 27–29, 32, 33, 35, 37, 39, 42, 43, 48, 64–71, 73–76, 78, 80, 83–86, 115, 127, 129–131, 200

OHE One-Hot-Encoding, an encoding technique that converts categorical variables to a numerical format. 33

PLM Pre-trained Language Model, a trained model on large corpora and that could be fine-tuned for a specific task. 43, 44, 47

POS Part Of Speech. 34, 39, 59

PUE Power Usage Effectiveness, a ratio that describes how efficiently a computer data center uses energy. 77

RE Relation Extraction, a subtask of Information Extraction aiming to identify relations between entities in text. 23, 24, 32, 130

REN Reconnaissance d'entités nommées (Named Entity Recognition). 5, 199

RGPD Règlement Général sur la Protection des Données (General Data Protection Regulation, GDPR). 5

RNN Recurrent Neural Network, a bi-directional neural network that can handle sequential input by sending information over time steps. 50, 61, 63

SG Skip-Gram, a neural network that predicts the likelihood of a word being a context word for a given word. 34

SVM Support Vector Machines, a supervised Machine Learning approach whose goal is to find the optimum hyperplane for classifying samples in a dataset. 40–42, 46, 50–52, 59–61

TAL Traitement Automatique des Langues (Natural Language Processing). 5, 6, 199

TF Term-Frequency, a measurement of how often a given word occurs within a document compared to the total number of words in the document. 33

THYME-TimeML THYME-TimeML, a specialization of TimeML for the clinical domain. 53, 55, 57, 111, 113, 125

TIE Temporal Information Extraction, an NLP task for extracting temporal information that could be used to order events. 24, 32, 48, 49, 51, 54, 56–58, 62, 125, 127, 128

TimeML Time Markup Language, a specification language for events, time expressions, and temporal relations in text. 50, 51, 53, 54, 56

TLINK Temporal link, a temporal relation between event and/or temporal expressions. 53–59, 64, 111, 124

TRE Temporal Relation Extraction, an NLP task aiming to identify temporal relations between mentions. 7, 24–26, 28, 31–33, 48, 49, 53, 54, 56, 58–64, 71, 110, 112, 117, 124, 125, 127, 128, 130, 200

UMLF Unified Medical Lexicon for French, a medical lexicon built using several French resources and terminologies. 88

UMLS The Unified Medical Language System, a compendium of many controlled vocabularies in the biomedical sciences. 46, 52, 59, 88, 96

Word2Vec Word2Vec, an algorithm that generates distributed word embeddings from large corpora. 34

WordPiece WordPiece, a subword-based tokenization algorithm. 35

Chapter 1

Introduction

1.1 Context

Although free text is the most convenient and easiest way to communicate, it is hard to process automatically due to its unstructured nature. Therefore, Natural Language Processing and Machine Learning methods have been used to understand and gain access to the useful information contained in the text. Information Extraction (IE) is the process of identifying the key elements of information that are relevant to a specific domain, including extracting entities via the Named Entity Recognition task and relations between entity mentions via the Relation Extraction task. However, addressing information extraction in specialized domains increases the task difficulties since domain expertise and greater effort are required to adapt information extraction systems using in-domain data. In the clinical domain, up to 80% crucial information contained in Electronic Health Records are in the form of unstructured text (Escudié et al., 2017). For many years, clinicians have been collecting and analyzing clinical narratives to identify important patient information, resulting in a waste of valuable expert time. IE allows the automatic identification and extraction of relevant information, which minimizes human labor and speeds up healthcare decision-making. Nevertheless, sharing resources and information extraction methods is difficult owing to clinical data privacy. As a result, creating and sharing resources in the clinical domain while preserving the privacy of sensitive data is needed, especially for non-English languages with lower resources like French. This thesis covers multidisciplinary research at the crossroads of

the fields of natural language processing and medical informatics. This work is based on machine learning methods for sequence recognition and is motivated by oncology applications.

Named Entity Recognition (NER) consists of identifying the target entities and classifying them into pre-defined categories. According to Ehrmann (2008), "Given an application model and a corpus, a named entity is any linguistic expression that refers to a unique entity of the model autonomously in the corpus." A named entity could be a word or a group of words with a beginning, an ending, and a type. The NER task is crucial for extracting general and domain-specific concepts. For instance, building clinical IE systems requires developing an accurate NER system for extracting medical concepts such as diseases, anatomical locations, drugs, symptoms, etc. Named entities can be nested, meaning they can be embedded in other entities, making their identification more challenging. Several efforts have addressed the NER task. However, until recent years, most of these research efforts have been only interested in extracting simple entities and neglected embedded overlapping entities. Extracting nested entities is, therefore, still under active research. The NER task is also a crucial step for other NLP tasks, such as Relation Extraction (RE). Indeed, the performance of extracting relations between entities relies on how effectively the entities are extracted.

Temporal ordering between entity mentions is also crucial in understanding language. Extracting temporal relations between mentions is essential, in particular, to building clinical patient timelines, which offer a better understanding of the patient's prior medical history, disease progression, treatment effects, etc. This also allows better decision-making about future treatment plans. Temporal information extraction implies, in the first place, the extraction of clinical event mentions and temporal expressions and then the extraction of temporal relations. Although the extraction of mentions might be solved using NER systems, the definition of event mentions largely depends on the text type, the application task, and the domain, making generalization across domains difficult. Temporal Relation Extraction also depends on the quality of extraction of clinical events and temporal expressions, which raises more difficulties in designing end-to-end systems.

Access to data is essential to create efficient information extraction systems. However, using highly sensitive data, such as personal patient health information, is problematic. For this, several studies have been conducted to address the de-identification of clinical narratives. However, with the rapid expansion of machine learning, particularly deep learning methods, and their outstanding performance in many NLP tasks, several privacy risks have arisen. Indeed, while training and deploying models on sensitive data, there is a risk of accidental memorization, which might result in the leakage of personal data (Bender

et al., 2021). As a result, simply de-identifying the clinical narratives is no longer sufficient to ensure the privacy of sensitive data. Moreover, sensitive data privacy limits the ability to share data and the models trained on this data, limiting collaborations and research. So, there is a need to propose ways to be able to create shareable models.

Neural-based methods have been proven to yield the best results for many NLP tasks, including NER and Temporal Relation Extraction (TRE), outperforming rule-based and traditional machine learning methods. However, such methods require a sufficient amount of annotated data. The annotation process is known to be time-consuming and costly as it requires domain expertise. Furthermore, clinical training data is often limited, in particular for non-English languages, which makes the clinical French NER task more challenging, as it is considered a low-resource problem. Designing annotation schemes to annotate clinical temporal relations also remains difficult, as shown by moderate inter-annotator agreement (Tourille et al., 2017b). This is due to the domain-dependent nature of the task, which requires extensive domain knowledge. Furthermore, clinical narrative text is often ungrammatical and goes back and forth through time, making it difficult to link events to temporal expressions. The time related to the clinical event is not always explicitly specified, and redundant information could be a major problem when determining the chronology of events.

Aside from annotation requirements and privacy concerns, using machine and deep learning models has a significant environmental impact. Recently, research efforts have been made to measure the carbon footprint of these models, encouraging the research community to evaluate their model carbon emissions and attempt to construct efficient green models with lower carbon emissions (Strubell et al., 2019; Bender et al., 2021; Wu et al., 2022).

Having funding from the ITMO Cancer Aviesan allowed us to conduct our research in the LISN lab as part of a multidisciplinary project with the Hôpital Européen Georges Pompidou (HEGP) and Sorbonne University, combining expertise in Natural Language Processing and the biomedical domain. By working on real oncology data, we developed effective NLP approaches, with the goal of assisting clinicians in extracting relevant information for tumor board meetings.

1.2 Research questions

Information extraction from narrative text is important for several domains, especially the clinical domain. Unstructured reports do contain crucial information that is essential for understanding the clinical patient history and

proposing better treatment strategies. However, the majority of available resources and information extraction methods are in English. This leads us to these broad research questions: **How can we provide relevant resources and information extraction tools for languages with lower resources?** and **how accurate can these systems be in a specialized domain, particularly the clinical domain?**

Over the years, several NER systems have been proposed, ranging from rule-based to traditional machine learning and deep learning-based models. Clinical NER is more challenging, as the clinical text is complex, containing a variety of medical terminologies, ambiguous entities, and multiple nested entities. Neural-based models have emerged as the most effective method for building high-performance NER systems, particularly when dealing with nested entities. However, due to the personal and sensitive nature of the clinical narratives, sharing these NER models trained on clinical data is restricted, limiting research and collaborations between institutions. As a result, **how can we develop shareable efficient NER models that, aside from handling nested entities, could preserve patient information privacy?** Moreover, **can we create these shareable models when few resources are available, as in the French clinical NER task?**

To get an accurate temporal relation extraction system, high performance in extracting the mentions, i.e., events and temporal expressions, is required. However, the definition of events is extremely dependent on the target task and the application domain. **Can we thus address the temporal relation extraction task independently from the domain in order to improve cross-domain generalization?** Furthermore, as indicated by moderate inter-annotator agreement in several shared tasks, the annotation of temporal relations remains a difficult task. So, **how can we simplify the representation of temporal relations to reduce the annotation efforts and simplify the task while still allowing for useful clinical text practical applications?**

Finally, with the growing need for annotated data in widely used deep learning methods, **can we create resources, particularly in French, that may be used for future clinical research and be useful to clinicians?** Also, **is it possible to raise awareness about the significant environmental impact of the intensive use of these deep learning methods?**

1.3 Contributions

This section highlights the main contributions of our work to answering the previous research questions. We propose novel architectures, approaches, and

resources for information extraction in clinical narrative text:

1. A novel *Privacy-Preserving Mimic Models architecture* that enables the creation of *shareable privacy-preserving models* for clinical French Named Entity Recognition in a low resource setting by leveraging both public and private corpora and using neural-based NER models. (cf. Chapter 4)

2. A novel *event-independent representation of temporal relations* in *complex narrative text* that facilitates annotating and extracting temporal relations, independently from the target task and the domain application, with an application on French clinical text. (cf. Chapter 5)

3. *Resources for further clinical NLP research* by providing annotations of two publicly available French clinical corpora generated in our NER experiments, making our shareable privacy-preserving NER model available to hospital institutions and sharing annotation guidelines to enable easier modeling of the temporal relation extraction task. Discussions are currently underway to integrate our NER model into the medkit tool[1], which is a Python library built by the HeKA team[2] to facilitate the development of applications for learning health systems. (cf. Chapter 4 and Chapter 5)

It is worth highlighting that all of our approaches were applied to the French language because few resources are available for this language, notably in the clinical domain, emphasizing the importance of efficient contributions to clinical research in French.

Additionally, although we take advantage of the rise of neural networks to address the information extraction task, we are conscious of their significant environmental impact. As a first step to raise awareness, we conduct a review of the existing tools for calculating the carbon footprint of NLP models, essentially deep learning models. We evaluate the tools by measuring the carbon emissions of NER experiments, and we choose one of these tools to calculate the carbon footprint of all our thesis experiments. (cf. Chapter 3)

Overall, our contributions tackle clinical information extraction in French, with the goal of assisting clinicians by simplifying and accelerating the collection of relevant patient information while preserving patient privacy. For instance, the identification and extraction of information from unstructured clinical narratives might benefit the decision-making process in tumor board meetings. Furthermore, our contributions allow further clinical research, in

[1] https://github.com/TeamHeka/medkit
[2] https://team.inria.fr/heka/fr/

particular, for French. Although we focused on clinical text in our methods, it is worth noting that our proposed representations and architectures could be adapted to other types of text, particularly those with similar privacy concerns.

1.4 Dissertation Outline

The remainder of this dissertation is structured as follows:

Chapter 2 - first describes the input text representations, progressing from hand-crafted to neural learned representations, and then introduces the background of textual information extraction, with a focus on Named Entity Recognition and Temporal Relation Extraction tasks, as well as the data privacy and Natural Language Processing environmental impact, both of which are relevant to most NLP tasks.

Chapter 3 - provides a preliminary study of the main existing tools for calculating the carbon footprint of NLP models, in particular, the computationally expensive deep learning models. This study offers a comparative analysis based on estimated environmental impact measurements and usability. The evaluation of selected tools was made by measuring the impact of NER experiments in two computational set-ups.

Chapter 4 - addresses the task of generating shareable Named Entity Recognition models in clinical narratives written in French. It puts forward a novel Privacy-Preserving Mimic Models architecture that leverages both public and private corpora and enables the sharing of neural models without disclosing patient data privacy. This architecture is evaluated through a neural-based NER model, which covers flat and nested clinical entities, providing a good compromise between performance and privacy preservation.

Chapter 5 - addresses the task of temporal relation extraction by proposing a novel event- and task-independent representation of temporal relations. This representation allows the identification of homogeneous text portions from a temporal standpoint and the classification of their temporal positioning according to the Document Creation Time without needing the prior definition of events and temporal expressions. The main goal of the proposed temporal representation is to simplify the temporal relation annotation efforts, which remain challenging, enhance the performance of extraction models for better practical use, and make the task more reproducible through different event types. Our event-independent representation of temporal relations is

evaluated on clinical tumor narratives, with a use case on temporal positioning of chemotherapy toxicity events.

Chapter 6 - concludes this dissertation by summarizing our main contributions and giving insight into research directions and perspectives for information extraction, particularly in the clinical domain.

1.5 Published Work

The material presented in Chapter 3 is based on a SustaiNLP EMNLP workshop paper (Bannour et al., 2021). The material presented in Chapter 4 is based on two publications, one at the Journal of Biomedical Informatics (JBI) (Bannour et al., 2022b) and one at the ATALA Day dedicated to Robustness of NLP systems (Bannour et al., 2022a). The material presented in Chapter 5 is based on two publications, one at the 2023 TALN conference (Bannour et al., 2023b) and one at the BioNLP workshop associated with the ACL conference (Bannour et al., 2023a).

List of Publications

- **Bannour, N.**, Ghannay, S., Névéol, A., & Ligozat, A. L. (2021, November). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing (pp. 11-21).

- **Bannour, N.**, Wajsbürt, P., Rance, B., Tannier, X., & Névéol, A. (2022). Privacy-preserving mimic models for clinical named entity recognition in French. Journal of Biomedical Informatics, 130, 104073.

- **Bannour, N.**, Wajsbürt, P., Rance, B., Tannier, X., & Névéol, A. (2022, November). Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français. In Journée d'étude sur la robustesse des systemes de TAL.

- **Bannour, N.**, Tannier, X., Rance, B., & Névéol, A. (2023). Positionnement temporel indépendant des évènements: application à des textes cliniques en français. In 18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 1-14). ATALA.

- **Bannour, N.**, Rance, B., Tannier, X., & Neveol, A. (2023, July). Event-independent temporal positioning: application to French clinical text. In The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks (pp. 191-205).

Chapter 2

Background and Related Work

2.1 Introduction

Information Extraction (IE) emerged as a method for mining the growing amount of publicly available information contained in unstructured and semi-structured data. Information Extraction begins with a collection of texts and converts them into structured information that can be easily used and evaluated (Cowie and Lehnert, 1996). It could be divided into five major tasks: Segmentation, Named Entity Recognition, Relation Extraction, Normalization, and Coreference resolution (Simoes et al., 2009). In this thesis, we are interested in building clinical French named entity recognition models in a low-resource setting while preserving the privacy of patient health information. Such models could be shared and used by clinicians and hospital institutions to propose better patient treatment strategies. To assist with the systematic analysis of large patient records, we are also interested in proposing temporal information extraction approaches to temporally ordering clinical events. Therefore, in this chapter, we dive deeper into the related work of Named Entity Recognition (NER), temporal information extraction, in particular Temporal Relation Extraction (TRE), data privacy concerns, as well the environmental impact of NLP methods.

The chapter is structured as follows. In Section 2.2, we start by reviewing the input text representations, going from hand-crafted to neural learned representations. In Section 2.3, we describe the publicly available corpora for the Named Entity Recognition task. Then, we delve into the several proposed

NER methods ranging from rule-based to feature-based and neural methods, as well as the proposed methods for French clinical NER, some low-resource NER strategies, and the evaluation metrics used for this task. Next, in Section 2.4, we start by defining the mentions involved in temporal relation extraction: time expressions, events, and temporal relations. We also review resources for temporal relation extraction, including annotation schemes, proposed shared tasks, and corpora in Section 2.4.4, and we go over the several proposed approaches for TRE, including rule-based, traditional learning-based, and neural methods. In Section 2.5, we explore the privacy risks raised by NLP methods, in particular, the deep learning approaches while dealing with sensitive data, and we enumerate some privacy-preserving strategies. Finally, we study the NLP environmental impact, particularly for modern NLP methods in Section 2.6, before concluding the chapter in Section 2.7.

2.2 Input text representations

Text represents a rich source of information, but because it is unstructured, it is difficult to extract and leverage this information. Indeed, computers cannot process the raw text data derived from natural language. Therefore, converting text into suitable numerical representations is a critical step in every Natural Language Processing work. However, it is important to create simple and easy-to-use representations while maintaining text semantics and meanings. The text is segmented into textual units, and a numerical vector represents each of these units. A sentence can be divided into words with vectorial representations for each word or into characters or n-grams for a finer-grained representation. In this section, we review the main earlier and modern text representations.

2.2.1 Earlier representations

Earlier approaches for word representation were either based on mappings of words into a list of terms, such as gazetteers, lexicons, and dictionaries or on statistical approaches based on word frequencies, such as One-Hot-Encoding (OHE), Bag of Words (BoW), Term-Frequency (TF), Inverse-Document-Frequency (IDF). The One-Hot-Encoding, for instance, consists in creating a vocabulary-size vector by inserting one in the index corresponding to the word in the sentence. Manning and Schutze (1999) go through further word representations and their use in earlier statistical NLP methods. Such earlier approaches for textual representations are simple to use but yield sparse high-dimensional vector representations that need a lot of memory and include several features that may not always be essential. Therefore, they are limited to small-scale corpora. Techniques such as feature selection and fea-

ture transformation were employed to overcome these issues, as explained in Patil et al. (2023). Feature selection tempts to keep only the main terms or features and dismiss the remaining ones, whereas feature transformation aims to map the vector representations to a smaller space with fewer dimensions. As a result, each textual unit is represented by a set of features. Some examples of extracted features used in the NER task include punctuation, morphological properties, or Part Of Speech (POS) tags that represent the grammatical categories of words. However, such methods do not include word meanings in representations, which are required to understand semantic concepts such as polysemy.

2.2.2 Modern representations

To better integrate the semantics of words into representations, two main categories of distributed word representations or word embeddings were proposed. The first category is static representations that are low-dimensional dense and fixed-length vectors, built assuming that words with similar contexts have the same meaning (Harris, 1954). Contextualized embeddings are the second type of embeddings, which incorporate context information into vector representations based on the premise that a word can have several meanings depending on context. These two types of embeddings will be discussed further in the following sections. Such feature embeddings are learned automatically, removing the need for the laborious feature engineering process.

Static embeddings. Prediction-based or count-based models were used to generate static embeddings. Mikolov et al. (2013b,a) introduced *prediction-based models* by proposing two models for learning embeddings, namely the continuous bag-of-words (CBOW) and skip-gram (SG) models. Both models are based on feed-forward neural networks. The CBOW model predicts a center word based on its given context words, while the SG model predicts the likelihood of a word being a context word for a particular target word. These two architectures are implemented into the Word2Vec toolkit[1]. An improvement of SG models was proposed in the fastText toolkit[2] (Bojanowski et al., 2017; Joulin et al., 2017) by using character-level representations to tackle the out-of-vocabulary problem. The main contribution for *count-based models* is the Global Vector for Word Representation (GLOVE)[3] model proposed by Pennington et al. (2014). This model learns word embeddings by encoding how frequently two words appear within a given window. For instance, if two words co-occur several times, they are semantically close.

[1] https://code.google.com/archive/p/word2vec/
[2] https://research.facebook.com/downloads/fasttext/
[3] https://nlp.stanford.edu/projects/glove/

Contextualized embeddings. In 2018, context-dependent models were presented to provide contextualized embeddings that go beyond traditional static embeddings. They are based on the assumption that a good model should be able to understand the various meanings of words given the context. Peters et al. (2018) proposed the Embeddings from Language Models (ELMo) model, which uses deep learning techniques to create contextualized representations. A word may have distinct embeddings depending on its context and position in a sentence. This model was followed by the Bidirectional Encoder Representations from Transformers (BERT) model proposed by Devlin et al. (2019) and based on transformers (Vaswani et al., 2017), an attention mechanism that learns contextual relation between words. BERT model uses a WordPiece tokenization algorithm (Wu et al., 2016). This algorithm starts by initializing the vocabulary with individual characters in the training corpus. During training, merging rules are learned, producing iteratively subwords known as wordpieces. Given this subword vocabulary, each out-of-vocabulary token will be segmented into a sequence of frequent subwords. This tokenization is similar to the Byte Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016) compression algorithm, in which the most frequent subwords are recursively merged. Unlike BPE, the WordPiece tokenization selects the pairs that increase the likelihood of the training data once added to the vocabulary rather than the most common pairs. The main goal of these two tokenization algorithms is to split rare words into smaller meaningful subwords rather than splitting frequently used words, which addresses concerns with word-based and character-based representations, such as high vocabulary size, out-of-vocabulary tokens, and the presence of less significant individual tokens. BPE has been used in GPT models (Radford and Narasimhan, 2018). Several transformer-based models have since been proposed such XLNET (Yang et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Zhuang et al., 2021), BART (Lewis et al., 2020) and many other models (Qiu et al., 2020; Han et al., 2021). El Boukkouri et al. (2020) proposed the CharacterBERT model, a variant of BERT that does not rely on wordpieces but instead consults the characters of each token to build word-level representations by using ELMo's Character-CNN module instead of the BERT's wordpiece embedding layer. French versions were also proposed such as CamemBERT (Martin et al., 2020), FlauBERT (Le et al., 2020) and recently CamemBERT-bio (Touchent et al., 2023), DrBERT (Labrak et al., 2023) and ALiBERT (Berhe et al., 2023) which are designed for the biomedical domain.

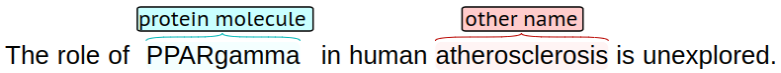
All the previously described neural embeddings are feature-based. They could be used as pre-trained embeddings, but we could also use these previous models as a model backbone of various NLP tasks and learn the input embedding from scratch during training. This type of embedding may be called fine-tuning-based embedding. Readers can find more details regarding feature

representations in Patil et al. (2023).

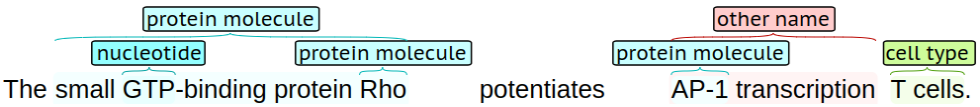
2.3 Overview of Named Entity Recognition

Named Entity Recognition (NER) is one of the five significant tasks of Information Extraction. It refers to identifying named entities in text and classifying them into pre-defined categories. The term "Named Entity" (NE) was initially used in the Message Understanding Conference (MUC) in the 1990s (Grishman and Sundheim, 1996), where the purpose was primarily to identify persons, organizations, localization, and numerical expressions such as time. In addition to these generic named entities, several domain-specific entities have been introduced. A named entity could be a word or phrase with a beginning, an ending, and a type.

Named entities can be nested, meaning they can include mentions of other entities. Recognizing such entities is known as a "nested NER". Nested entities might be of the same or of a distinct entity type, making the extraction task more challenging. On the other hand, the NER task with no nested entities is referred to as a "flat NER" task or simply a "NER" task. Figure 2.1 illustrates two GENIA dataset (Kim et al., 2003) samples, the first representing the traditional flat NER task and the second showing a nested NER task. In the example of Figure 2.1b, the five nested entities are *"Small GTP-binding protein Rho"*, *"GTP"*, *"Rho"*, *"AP-1 transcription"* and *"AP-1"*. The traditional flat NER does not consider the recognition of these nested entities.



(a) Flat NER



(b) Nested NER

Figure 2.1: Two GENIA samples showing the difference between flat and nested NER.

In this section, we describe available corpora for NER, the several proposed methods of NER ranging from rule-based to feature-based and neural methods. We also review the main proposed approaches for French clinical NER, some low-resource NER strategies, and the evaluation metrics.

2.3.1 Corpora for Named Entity Recognition

Several annotated corpora have been proposed for the NER task. These corpora differ in language, domain, entity types, and whether or not they contain nested entities. Most annotated available corpora have been proposed in major NLP evaluation campaigns. The MUC-6 (Grishman and Sundheim, 1995) and the MUC-7 (Chinchor and Robinson, 1997) English corpora were provided in the shared tasks at the 6th and 7th Message Understanding Conference (MUC). The CoNLL-2002 (Tjong Kim Sang, 2002) and CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) corpora were introduced in the Conference of Natural Language Learning (CoNLL) (Tjong Kim Sang and Buchholz, 2000) shared tasks and covered four languages (Spanish, Dutch, English, and German). OntoNotes is a NER corpus that has been developed in the OntoNotes project[4] and covered three languages: English, Chinese, and Arabic. The QUAERO Broadcast News Extended Named Entity Corpus[5] has been introduced with French named entities. These datasets described above are for the general domain and do not contain nested entities. However, the ACE[6] corpus is a major benchmark general domain corpus in which around 35% of the sentences include nested entities for the English language (Wang et al., 2022b).

Other NER datasets have been proposed for specific domains. For biomedical and clinical domains, various datasets have been proposed, such as GENIA (Kim et al., 2003), i2b2-2010 (Sun et al., 2013), BC5CDR (Li et al., 2015), and NCBI-disease (Doğan et al., 2014) corpora. About 17% of the GENIA corpus entities are embedded within other entities (Wang et al., 2022b). Most annotated available corpora in the clinical domain are devoted to English. Few publicly available corpora have been proposed for French, like the QUAERO French Medical (Névéol et al., 2014), CAS (Grabar et al., 2018) and, DEFT-2020 (Cardon et al., 2020) corpora. We describe the publicly available French corpora used in the thesis experiments below. Table 2.1 presents descriptive statistics about these datasets, which are publicly available for research purposes through a data use agreement.

QUAERO Broadcast News Extended Named Entity. This corpus (Galibert et al., 2010) comprises manually fully annotated radio broadcast news and broadcast conversation data. This corpus is freely available for non-commercial use and does not contain nested entities.

[4] https://catalog.ldc.upenn.edu/LDC2013T19

[5] http://catalog.elra.info/en-us/repository/browse/ELRA-S0349/

[6] https://catalog.ldc.upenn.edu/LDC2006T06

	QUAERO French News	QUAERO French Med		CAS	DEFT	CépiDc
		EMEA	MEDLINE			
Language	French	French	French	French	French	French
Domain	News	Biomedical	Biomedical	Clinical	Clinical	Clinical
Documents	167	38	2,498	717	167	23,750
Tokens	1,347,368	40,257	31,926	231,662	57,188	237,777
Entities	79,632	7,159	9,074	-	12,867	-
Unique entities	19,876	1,880	5,895	-	8,831	-
Nested entities	-	1,009	2,280	-	5,352	-
% Nested entities	-	14,27%	25,31%	-	41,60%	-
Max Depth	1	4	4	-	4	-

Table 2.1: Descriptive statistics for the French NER corpora used in this thesis.

QUAERO French Medical. This corpus is made of manually annotated biomedical EMEA documents and MEDLINE titles, written in French, and was used in the CLEF eHealth Lab in 2015 and 2016 (Névéol et al., 2014; Névéol et al., 2016) for the clinical Named Entity Recognition task and is freely available for non-commercial use. 10 types of clinical entities that are annotated, namely *anatomy*, *Chemicals & Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology*, and *Procedures*. Named entities can nest up to 3 levels, reaching a maximum depth of four.

CAS. This corpus is available for research purposes through a data use agreement. It consists of clinical cases described in French scientific literature. It is annotated with two types of demographic entities (age, gender) and two types of clinical entities (origin of the visit, outcome). The CAS corpus contains 717 clinical documents with a total of 231,662 tokens.

DEFT. This corpus represents a subset of 167 clinical cases from the CAS corpus, introduced in the 2020 DEFT challenge[7]. This corpus is available for research purposes through a data use agreement. It is annotated with 13 different types of clinical entities, namely *anatomy*, *substance*, *dose*, *administration method*, *treatment*, *pathology*, *sign or symptom*, *medical procedures*, *value*, *date*, *frequency*, *moment* and, *duration*. Named entities can nest up to 3 levels, reaching a maximum depth of four.

CépiDC. This corpus is available under a data use agreement[8]. It was used in the CLEF eHealth ICD10 coding challenge (Névéol et al., 2018b) and contains free-text descriptions of causes of death extracted from death certificates

[7] https://deft.limsi.fr/2020/index-en.html
[8] http://www.cepidc.inserm.fr/

submitted electronically over the period 2006-2015. The certificates are annotated at the document level with codes from the International Classification of Diseases (ICD10). The CépiDC corpus contains 23,750 death certificates with a total of 237,777 tokens.

2.3.2 Rule-based and terminology-based approaches

Early methods for Named Entity Recognition were rule-based and terminology-based methods. These methods are based on human handcrafted sets of rules and domain-specific lexicons. Several rule-based NER systems were proposed, such as NetOwl (Krupka and Hausman, 1998), Facile (Black et al., 1998), LaSIE-II (Humphreys et al., 1998) and LTG (Mikheev et al., 1999). Rau (1991) proposed one of the first systems to extract company names from financial text using rules, heuristics, and patterns. Farmakiotou et al. (2000) proposed a rule-based Greek NER system using gazetteers and grammars, which was evaluated on a financial news corpus. Gattani et al. (2013) created a Wikipedia-based method for Named Entity Recognition in social media, in which pertinent words were linked to Wikipedia pages. Wikipedia is used as a lexicon containing a large number of entities. Rocha et al. (2016) build a named entity recognition system using regular expressions to identify common candidate entities and a new collection of regular expressions on the Part-Of-Speech (POS) tags to filter specific candidates. To extract nested entities, early work on nested NER used rule-based post-processing. For instance, Shen et al. (2003) proposed four main patterns relating to different cascaded or nested entities. Like much early work on nested entities, their technique is combined with a learning-based strategy. These approaches will be explored in the following sections.

Clinical domain. Many rule-based and terminology-based NLP systems have been developed for clinical concepts extraction such as MedLee (Friedman et al., 1994), MedEx (Xu et al., 2010), MetaMap (Aronson and Lang, 2010), cTakes (Savova et al., 2010) and MedXN (Sohn et al., 2014). Hanisch et al. (2005) introduced a synonym dictionary-based NER system for extracting protein and gene entities. Deléger et al. (2010) introduced a rule-based system for extracting medication information. Quimbaya et al. (2016) proposed a combined dictionary-based approach for NER in Electronic Health Records (EHRs). This approach combined a direct match technique with fuzzy matching and stemmed matching. Eftimov et al. (2017) introduced a rule-based NER method to extract dietary concepts in clinical texts. Rule-based techniques cannot be generalized because they rely heavily on the quality of manually defined language and domain-specific rules. Moreover, developing such rules is time-consuming and expensive. Exhaustive lexicons are required for terminology-based approaches. As a result of domain-specific rules and incomplete

dictionaries, rule-based and terminology-based techniques have high accuracy but low recall.

2.3.3 Feature-engineering-based supervised methods

Feature-engineering-based supervised methods aim to generate an inferred function that maps incoming input to a pre-defined category by learning from a labeled corpus. Named entity recognition may be represented as a classification task for each token, independently from each other. Traditional supervised methods formalize the NER task as a sequence labeling task using sequence tag schemes, where the goal is to assign a label to each element in a sequence and then combine the elements to identify named entities. Each element or token will be given a tag with an entity type and an indication of the token's position in a named entity.

Tagging Schemes. Several tagging schemes for encoding the named entities were proposed. The first proposed tag schemes were IO, IOB, BIO, and IOE. Each token in the IO scheme is classified as an (I)nside tag or (O)utside tag. The (O)utside tag denotes the absence of any type of entity at a given position. To represent named entities with multiple tokens, additional tags are given in the IOB, BIO and IOE tag schemes to indicate whether the token is at the (B)eginning or (E)nd of an entity. In the IOB tagging method, the (B)eginning tag is solely used to differentiate successive items of the same type, which is not allowed with the IO tagging scheme. However, in the BIO scheme, the (B)eginning tag is added to all entities. The BIO tagging scheme gained popularity when it was adopted by the Conference of Natural Language Learning (CoNLL). The IOBES labeling scheme is an extension of the IOB scheme, where the (E)nd tag is used to identify the last token of the entity, and the (S)ingle tag is used for single-token entities. This encoding scheme is known by several names, such as BMEWO scheme and BIOUL scheme, using the (L)ast tag instead of the (E)nd tag for ending tokens and the (U)nit tag instead of the (S)ingle tag for single-token entities. This tagging format obtained the best performance of the CoNLL dataset (Ratinov and Roth, 2009). Figure 2.2 illustrates an example of all these tagging schemes, derived from the French DEFT 2020 challenge[9], with an example of nested entities represented on two levels with the BIOUL tagging scheme.

Methods. The common earlier used supervised techniques for the NER task are Hidden Markov Model (HMM) (Baum and Petrie, 1966), Maximum Entropy (ME) model (Berger et al., 1996), Decision Tree (DT) (Wu et al., 2008), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and Conditional

[9] https://deft.lisn.upsaclay.fr/2020/

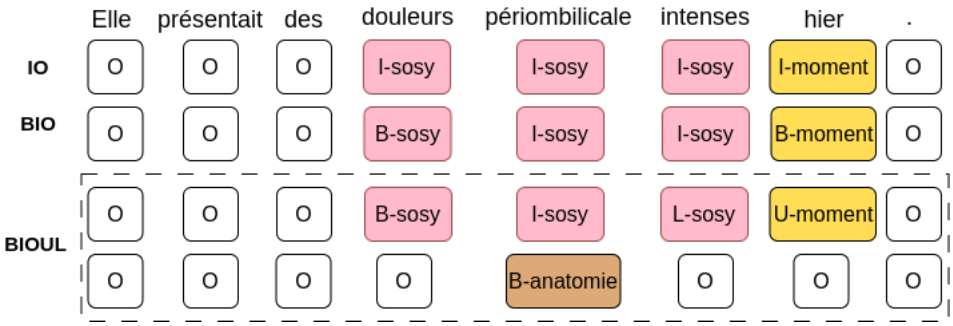


Figure 2.2: Example with different tagging schemes. Translation of text into English: *"She had severe periumbilical pain yesterday."*.

Random Fields (CRF) (Lafferty et al., 2001). Bikel et al. (1999) introduced the first HMM model for NER task evaluated on English and Spanish texts. The HMM model learns the probability of the current token's label given the previous token's label and the probability of generating a token given its label. Bikel et al. (1999) use the Viterbi algorithm (Viterbi, 1967), a dynamic programming algorithm, to find the most probable sequence of labels or hidden states based on a sequence of observations. Morwal et al. (2012) proposed a similar language-independent NER system trained and tested on Indian languages. Dahan et al. (2015) presented an Arabic NER system using a HMM model, outperforming rule-based techniques. However, HMM implies that all tokens are independent of each other, which limits the contextual information available to the NER model. Chieu and Ng (2003) proposed two ME based NER systems that did not only use the local context inside a phrase but also exploited word occurrences throughout the same document and incorporated additional features from external name lists. Cowie (1995) proposed the AutoLearn system, which used the ID3 algorithm (Quinlan, 1986) to build a DT that could identify the start and the end of certain named entities. Bennett and Aone (1997) introduced proposed the RoboTag NER system, which uses an improved version of the ID3 algorithm, namely the decision-tree induction algorithm C4.5 (Salzberg, 1994), and outperforms AutoLearn on the MUC-6 data due to the use of gazetteers and other lexical resources. Sekine et al. (1998) addresses the problem of wrongly identifying person names included in organization NEs by searching the most probable sequence of output tags that provide a valid combined solution using a human rule set. To allow the decision tree to categorize the NEs directly, Paliouras et al. (2000) proposed a pre-processing step that consists in extracting noun phrases using a separate parser. This is done under the premise that NEs are noun phrases. Li et al. (2005) developed an SVM-based system and evaluated it on the CoNLL-2003 dataset and CMU seminars. This system comprises two SVM classifiers for each entity type, one for recognizing the beginnings of the named entity and

another for the ends. They experimented with various window sizes and features, and they used a variant of the SVM, the SVM with uneven margins (Li and Shawe-Taylor, 2003), which outperforms the original SVM in terms of generalization performance. SVMs can learn various combinations of features but do not take neighboring words into account when predicting an entity label. McCallum and Li (2003) introduced a feature induction for CRFs in NER and evaluated their method on the CoNLL-2003 dataset for English and German. Torisawa et al. (2007) proposed a CRF-based NER model using features from Wikipedia as external knowledge. Krishnan and Manning (2006) build a two-stage NER model using two CRFs. To capture non-local dependencies, the second CRF uses features obtained from the output of the first CRF. CRFs may capture both local and global contexts and represent deep domain knowledge using features. They are, however, unidirectional and can only represent connections between labels in the forward direction. To reduce the need for manually created expert rules and for annotated data, some works adopt a hybrid NER model combining rules and statistical learning methods. For instance, Shaalan and Oudah (2014) presented a hybrid NER approach to enhance the performance of the Arabic NER task. To solve the nested entities problem, Lu and Roth (2015) developed a directed hypergraph-based approach that allows the representation of many possible combinations of overlapping mentions of different types. Muis and Lu (2017) proposed an improvement to this approach by modeling mention edges along with the features.

Clinical domain. Takeuchi and Collier (2003) introduced an SVM-based biomedical NER approach using a collection of MEDLINE abstracts. Wang and Patrick (2009) presented a cascading clinical NER system that reclassifies the extracted entities using a CRF model, an SVM model, and a Maximum Entropy model with a voting strategy. Wang et al. (2014) demonstrated that the CRF approach outperforms HMM and MEMM models for recognizing symptoms in Chinese clinical text. Xu et al. (2014) proposed a joint model based on CRF for segmentation and NER on Chinese discharge summaries. Chieu and Ng (2003) conducted an evaluation of active learning methods for named entity recognition in clinical text using the NER corpus from the 2010 i2b2/VA NLP challenge. Cheng et al. (2019) proposed a hybrid model incorporating expert rules with a BiLSTM-CRF approach to extract Chinese clinical named entities. To address the nested biomedical NER task, Zhang et al. (2004) introduced a layered HMM-based approach on the GENIA corpus. For this, two HMMs are trained, one to identify short nested entities and the other to extend short entities. Alex et al. (2007) structured the nested NER problem as cascaded flat NER tasks. Each NER task consists in a CRF model that is trained by using the previous CRF's output as a feature for the current one. The main drawback of this technique is that it does not handle overlapping entities of

the same type. Finkel and Manning (2009) used a tree-based parsing model for the nested NER task. In fact, entities were represented as subtrees, and a CRF approach was used to detect the nested entities.

Overall, fully-supervised methods need a large amount of annotated data, and their performance heavily depends on the annotation quality. Moreover, more features usually result in better performance. However, annotating large corpora is time-consuming and highly expensive.

2.3.4 Neural-based approaches

Collobert and Weston (2008) introduced the first neural-based model for the NER task with manually constructed feature vectors. Deep learning feature representations, i.e., word embeddings, were used in the later proposed NER models (Collobert et al., 2011). Collobert et al. (2011) proposed a one-layer Convolutional Neural Network (CNN) (Waibel et al., 1989) based on word embeddings, followed by a CRF output layer. Huang et al. (2015) proposed a similar architecture using Long-Short-Term-Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Bidirectional LSTM (Bi-LSTM) models to take into account the context and demonstrated that adding a CRF layer on top of the Bi-LSTM enhanced performance on the English CoNLL dataset. Lample et al. (2016) presented a similar NER model but using character-based word features rather than hand-crafted features. Chiu and Nichols (2016) proposed a Bi-LSTM-CNN hybrid model that captures both character-based and word-based features. Ma and Hovy (2016) introduced a hybrid NER architecture based on Bi-LSTM, CNN, and CRF and obtained better performance on the English CoNLL dataset. This architecture is an end-to-end system that does not require feature engineering or data pre-processing. Panchendrarajan and Amaresan (2018) introduced a NER architecture that combines a Bi-LSTM model and a bidirectional CRF (Bi-CRF) layer, which describes the dependency between labels in both directions. It is demonstrated that the backward CRF can extract complex entities. Later, contextual representations produced by Pre-trained Language Models (PLMs) considerably increased NER system performance (Peters et al., 2018; Devlin et al., 2019). In fact, several transformer-based PLMs such as BERT (Devlin et al., 2019), RoBERTA (Liu et al., 2019b), ALBERT (Lan et al., 2019) and T5 (Raffel et al., 2020) achieved high performance in many NLP tasks. Therefore, pretraining models on a vast amount of text and fine-tuning it on task-specific corpora is now the common approach in modern NLP tasks, including the NER task. Multiple works use these models to enhance the performance of NER models (Liu et al., 2019a; Luo et al., 2020). However, all these approaches address only flat NER. Katiyar and Cardie (2018) proposed a hypergraph representation for nested entities and used an LSTM-based sequence labeling model to learn the structure. Ju et al.

(2018) proposed a stacked neural layered model built with flat NER layers. Inspired by the state-of-the-art model proposed in (Lample et al., 2016), each layer is based on a Bi-LSTM-CRF model. Other than the error propagation from layer to layer, a limitation of this model is that an inner entity cannot be identified when an outer entity is extracted first. Aside from token-based NER formulation, span-based NER approaches have recently gained popularity, where the goal is to identify and classify all possible continuous sequences of tokens independently and then deal with the overlap conflict as a post-processing step. Wang et al. (2020) introduced a span-based neural layered model, namely Pyramid, which consists of a stack of linked layers and recognizes entities in a bottom-up manner. Li et al. (2020b) cast the nested NER task as a Machine Reading Comprehension (MRC) task by prompting a pre-trained language model with queries containing the entity categories and asking the model to identify the spans corresponding to these categories. Straková et al. (2019) formulated the nested NER task as a sequence-to-sequence generation problem with an input sequence of tokens and a target sequence of labels. Multiple combinations of context-based embeddings were also studied. Yu et al. (2020) used a biaffine model (Dozat and Manning, 2017) to score all candidate spans in a sentence and predict both flat and nested entities using contextual embeddings. Some works treated the NER task in a generative way. Indeed, Yan et al. (2021) proposed a sequence-to-sequence unified generative model with pointer network (Vinyals et al., 2015) and based on BART (Lewis et al., 2020). Shen et al. (2021) addresses the nested NER task by a two-stage approach, which is commonly used in the computer vision field. Wang et al. (2023) proposed a NER method based on GPT (Brown et al., 2020) to explore the use of large generative language models for both flat and nested NER tasks. Shen et al. (2023) explored a novel generative method for both flat and nested NER tasks that cast the NER task as a boundary denoising diffusion process and generate named entities from noisy spans using diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020).

Clinical domain. Yao et al. (2015) proposed one of the first neural models for the biomedical NER task based on CNN and with a multi-layered structure. Zhao et al. (2017) formulated the NER task as a classification task and presented a multi-label CNN model to extract disease and chemical entities. Habibi et al. (2017) developed the model LSTM-CRF for several biomedical NER tasks and performed better than the other tested NER systems. Wu et al. (2018) proposed a method that combines medical knowledge embeddings with word embeddings in the LSTM-CRFs model to deal with the medical terminologies that are not often used in general domain corpora. Xu et al. (2018) introduced an attention-based neural clinical NER model to incorporate the document-level global information with the local context information using

representations from pre-trained bidirectional language models with attention. Zhang et al. (2019) obtained significant performance improvement when using BERT embedding as input features to a Bi-LSTM-CRF model for extracting clinical concepts from Chinese clinical breast cancer notes. Wei et al. (2019) introduced an attention-based BiLSTM-CRF model to enhance the ability to extract significant context information in the biomedical NER task. Yang et al. (2020) explored transformer-based models such as BERT in the clinical concept extraction task on different shared tasks corpora and highlighted the benefits of contextual embeddings. To improve the performance on biomedical and clinical NER tasks, several domain-specific language models were also introduced, such as BioBERT (Lee et al., 2020), clinicalBERT (Huang et al., 2019), BioALBERT (Naseem et al., 2021) and PubMedBERT (Gu et al., 2021). For the nested NER task, the stacked neural model proposed by Ju et al. (2018) outperforms state-of-the-art feature-based models on the GENIA dataset. Wang et al. (2020) showed that their method obtained state-of-the-art results on different nested NER corpora, including the GENIA corpus. Sohrab and Miwa (2018) proposed a neural model that enumerates all possible spans as potential entity mentions and classifies each span into a specific category or a non-entity. Each span is represented with its word embeddings. Zheng et al. (2019) introduced a boundary-aware NER model that combines a sequence labeling model to identify boundaries with a span classification model to predict nested entities based on the detected boundaries. Straková et al. (2019) proved that by using contextual embeddings such as ELMo, BERT, and Flair, their sequence-to-sequence model performs better on many datasets, including the GENIA dataset, and for both flat and nested NER tasks. Yu et al. (2020) evaluated their biaffine model on the biomedical GENIA dataset, and a significant gain was obtained compared to earlier systems.

The main strength of neural approaches is their ability to learn complex input representations, which reduces the effort of hand-crafting features. However, to achieve high performance, such neural models require sufficient human-annotated data, which can be costly and time-consuming. Moreover, the biases involved with training large language models raise many ethical and legal issues, such as patient and data privacy (Bender et al., 2021). We will review the clinical data privacy concerns in Section 2.5.

2.3.5 Proposed approaches for French clinical NER

Several studies have addressed the NER task in French using different domain corpora such as news corpora (Galibert et al., 2012; Dupont, 2017; Dekhili and Sadat, 2020; Labusch and Neudecker, 2020) and Twitter texts (Sileo et al., 2017; Peres et al., 2017). In the clinical domain, most studies focus on texts written in English or Chinese. Few studies were proposed on French cor-

pora (Névéol et al., 2014, 2018a). As part of the CLEF eHealth 2015 workshop (Névéol et al., 2015), Soualmia et al. (2015); Jiang et al. (2015) proposed a combination of CRF-based models with hand-crafted features and lexicons for geographical entities and d'Hondt et al. (2015) presented a three classifiers NER system to deal with nested entities using CRFs and SVM models. Van Mulligen et al. (2016) proposed a dictionary-based NER approach using French Unified Medical Language System (UMLS) terms with translated English UMLS terms, and Ho-Dac et al. (2016) used a CRF-based model with diverse linguistic features for the CLEF eHealth 2016 clinical NER task (Névéol et al., 2016). The CLEF eHealth 2015 and 2016 shared NER tasks were based on the annotated QUAERO French Medical corpus (Névéol et al., 2014). Lerner et al. (2020) proposed a hybrid system that combines expert rules with a Bidirectional Gated Recurrent Unit with a CRF (BiGRU-CRF) architecture to extract five types of entities on a proposed French corpus of 147 clinical documents. Jouffroy et al. (2021) created a hybrid approach that uses a BiLSTM-CRF model, contextual word embeddings trained on clinical text, and a combination of knowledge base and expert rules. Their model is evaluated using a private French clinical data warehouse. As part of the DEFT 2020 challenge (Cardon et al., 2020), Minard et al. (2020); Copara et al. (2020) proposed CRF-based French clinical NER approaches, Lemaitre et al. (2020) used a rule-based system and Wajsbürt et al. (2020) proposed two NER models: a layered BiLSTM-CRF model and a greedy NER model, using CamemBERT embeddings. Their models take into account the extraction of nested entities. Le Clercq de Lannoy et al. (2022) introduced a hybrid approach that combined specialized knowledge with language model (CamemBERT) adaptation on several biomedical corpora.

The fact that few research works have been conducted on French corpora may be related to the extra challenges encountered when dealing with French clinical data. In fact, as mentioned in Section 2.3.1, few publicly available French annotated datasets are available. On the one hand, the annotation process is time-consuming and extremely expensive due to the need for rich domain knowledge, representing a big challenge for low-resource languages. On the other hand, due to the sensitive nature of clinical data, sharing such data is restricted. Indeed, sharing data is difficult in practice and is governed by laws and regulations such as General Data Protection Regulation (GDPR)[10]. As a result, limited collaborations could be done across hospital institutions.

2.3.6 A word about low-resource NER strategies

To address the lack of annotated corpora and address the low-resource NER task, prior works used either semi-supervised learning, which aims to learn from

[10] https://gdpr-info.eu/

both labeled and unlabeled data (Liao and Veeramachaneni, 2009; Liu et al., 2011; Gao et al., 2021), or data augmentation methods that expand the training set by applying transformations without changing their labels (Dai and Adel, 2020; Phan and Nguyen, 2022) or active learning methods which assume the presence of a human annotator, who may be queried to get ground-truth labels for the most relevant unlabeled instances to be added to the training set. As a result, only data that can increase performance are annotated (Tomanek and Hahn, 2009; Shen et al., 2017; Liu et al., 2022; Naguib et al., 2023; Le et al., 2023). Other methods seek distant supervision, which uses external knowledge rather than propagating the knowledge to either label more data (Cao et al., 2019; Lison et al., 2020; Liang et al., 2020; Wang et al., 2021) or incorporate meta information such as context and prompts to facilitate training (Lee et al., 2022a). With the rise of pre-trained language models, few-shot and zero-shot learning methods were proposed to learn better using only a few labeled instances (Košprdić et al., 2023; Zhang et al., 2023; Agrawal et al., 2022; Yohannes and Amagasa, 2022).

2.3.7 Evaluation metrics

Precision, Recall, and F-measure are commonly used evaluation metrics for information extraction systems. These measures are calculated based on the number of true positives (TP), false positives (FP), and false negatives (FN) as defined in the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F - measure = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{2.3}$$

In an **exact-match or strict evaluation** setting, an extracted entity is considered a true positive if both entity type and boundaries are correctly extracted, a false positive if it was wrongly labeled, and a false negative if it was not annotated.

However, we can also evaluate our NER models in a **partial-match or relaxed-match evaluation** setting, allowing entities to match if their boundaries overlap. Indeed, an extracted entity could be counted as a true positive if it shares half of the tokens with the gold entity. This evaluation method may be sufficient for some tasks since it is regarded as a more intuitive metric that could consider the annotation mistakes. In our experiments, we use the

BRATEval tool[11] to calculate our evaluation metrics.

2.3.8 Summary

The first approaches to Named Entity Recognition were based on hand-crafted techniques such as expert rules or dictionary-based matching. However, statistical approaches have gained popularity recently, ranging from traditional machine learning to modern neural approaches. Traditional supervised machine learning approaches depend heavily on the quality of hand-crafted input feature representations. Neural approaches discard the feature engineering process but still require a lot of annotation effort. Indeed, to obtain high-performing NER models, large amounts of annotated corpora are needed. Some recent strategies, including semi-supervised approaches, have evolved to leverage partially and few labeled datasets. Nevertheless, there are several challenges to overcome in the clinical domain. Indeed, clinical text is complicated, containing a variety of medical terminologies, ambiguity, and nested entities. Although most NER models are devoted to flat entities, many methods seek to deal with nested entities. Due to the personal and sensitive nature of clinical text, particularly in French, annotated clinical corpora are often limited. As a result, only a few studies addressed the task of French clinical NER. In our thesis, we are interested in proposing shareable French clinical NER models while preserving patient privacy. We are also interested in temporality between mentions with the objective of creating patient timelines. In the following section, we will go through the main methods that have been proposed for Temporal Relation Extraction.

2.4 Overview of Temporal Relation Extraction

Temporal Information Extraction (TIE) can be defined as extracting meaningful information that could enable ordering in unstructured text. Temporal Information Extraction may be divided into two subtasks: (1) identification of events and time expressions and (2) extraction of temporal relations. The first subtask consists in detecting both event and time expressions. Events and time expressions can be considered entities, and this first subtask might be tackled as a NER task. The second subtask of TIE is to extract temporal relations that could be between events and/or time expressions, as well as relations between events and the Document Creation Time (DCT). Temporal Relation Extraction (TRE) is important for many NLP tasks, such as Question-answering systems, Machine Translation, and document summariza-

[11] https://bitbucket.org/nicta_biomed/brateval/src/master/

tion. TRE is also a fundamental task for the biomedical and clinical domains since clinicians need to identify and order relevant clinical events to create patient timelines to understand, for instance, the disease progression.

In this section, we briefly introduce the notion of events and time expressions, as well as some methods of their extraction for both general and clinical domains, with a particular focus on the temporal relation extraction subtask. We review some annotated corpora for TRE and the proposed approaches for this task, ranging from rule-based methods to traditional machine learning and modern neural-based methods. We also go over some research efforts that attempt to structure clinical narrative text by developing section segmentation methods.

Figure 2.3 illustrates an example of the Temporal Information Extraction task, including two events (EVENT), a time expression (TIMEX3), and two types of temporal relations, namely relations between events and the document creation time and relations between an event and a temporal expression. More information regarding temporal information is provided in the following sections.

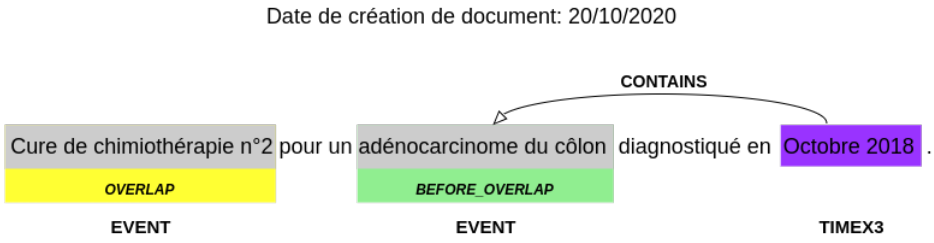


Figure 2.3: A temporal information example that includes events, time expressions, and temporal relations. Translation of text into English: *"Document Creation Time: 10/20/2020. Chemotherapy round n°2 for a colon adenocarcinoma which was diagnosed in October 2018."*

2.4.1 Time expressions

Time or temporal expressions are used to give information about *when*, *how long*, or *how often* something occurs (Derczynski, 2017). There are four types of time expressions: dates, times, durations, and sets. Several temporal annotation schemes have been proposed to standardize temporal information, including time expressions. The earliest modelization of time expressions was a simple temporal value attribute TIMEX (Sundheim, 1993) that could take the value *date* or *time*. Then, in the TILDES TIMEX2 annotation scheme (Ferro et al., 2001), an expanding set of attributes has been added, such as a value tag for the normalized value of the temporal (VAL), a modifier tag (MOD), a set tag (SET) that is marked as yes if the time expression is representing a

set, etc. The Time Markup Language (TimeML) (Pustejovsky et al., 2003a) and ISO-TimeML (Pustejovsky et al., 2010) defined a final version of TIMEX annotation (TIMEX3). The TIMEX3 tag is based on the previous TIMEX tags and includes, among other tags, the type of time expression (TYPE) and BEGINPOINT and ENDPOINT tags when the time expression is a duration, among other tags. Another annotation scheme, namely the SCATE scheme, has also been developed by Bethard and Parker (2016) to take into account the fine-grained aspect of time expressions.

Time expressions have been widely studied as part of the TempEval challenges (Verhagen et al., 2007, 2010; UzZaman et al., 2013), which focused on English news articles documents. Indeed, various rule-based approaches have been developed for time expression recognition, such as TempEx (Mani and Wilson, 2000), SUTime (Chang and Manning, 2012) and, HeidelTime (Strötgen and Gertz, 2013). These systems achieve good performance in the TempEval challenges. Strötgen and Gertz (2015) described a method to extend the HeidelTime system to all languages and created a new baseline of 200 languages, including French (Moriceau and Tannier, 2014). Machine learning systems have also been introduced for the task of time expressions recognition and normalization, based on CRF (UzZaman and Allen, 2010), SVM (Bethard, 2013), and other 'machine learning algorithms (Ding et al., 2019; Ning et al., 2018). Lee et al. (2014) proposed a hybrid system using Combinatory Categorial Grammar (Steedman and Baldridge, 2011), combining hand-crafted and trained rules and outperformed the state-of-the-art (SOTA) temporal tagging systems. Later on, deep learning methods were developed, using RNNs (Laparra et al., 2018), BERT embeddings (Chen et al., 2019) and LSTMs (Lange et al., 2020). Cao et al. (2022) presented the XLTime framework for multilingual time expression extraction, which is cast as a sequence labeling task, similar to NER. Note that deep learning-based techniques for time expression recognition are less frequent and produce results that are comparable to or worse than rule-based SOTAs (Cao et al., 2022).

Clinical domain. In the clinical domain, modifications have been made to take into account domain particularities. Indeed, Styler IV et al. (2014) proposed the Time Markup Language guidelines in 2014 for annotating temporal information from clinical texts. In particular, a new tag is added to the TIMEX3 tags, namely PREPOSTEXP, which refers to clinically relevant and temporally complex terms such as *preoperative*, *postoperative*, and *intraoperative* (Olex and Mcinnes, 2021).

Time expression extraction has received interest in the clinical domain through the i2b2-2012 challenge (Sun et al., 2013) and the Clinical TempEval shared tasks (Bethard et al., 2015, 2016, 2017). Jindal and Roth (2013) used

the HeidelTime system and developed several rules to extract complex clinical time expressions. Sohn et al. (2013) presented the rule-based system MayoTime that adapts the HeidelTime Framework to the clinical domain. Most of the other proposed methods in the clinical challenges are hybrid. Indeed, Lin et al. (2013) introduced the MedTime system, which used the initial tagging from HeidelTime, a specific FREQUENCY tagger, and a CRF-based model that identifies the domain-specific time expressions. Velupillai et al. (2015) created a time expression recognizer based on ClearTK (Bethard, 2013) and SVM classifiers. Tapi-Nzali et al. (2015) studied time expression extraction across three domains (news, historical, and medical) in French narratives, using the Heideltime outputs as features of a CRF-based system. Lin et al. (2017) proposed a CNN-based time expression recognition system that outperformed previous methods on the THYME corpus. Tourille et al. (2017a) proposed a hybrid LSTM-CRF model to extract the TIMEX3 entities. To sum up, extracting time expressions in clinical narratives remains a challenging task. Indeed, there is a variety of time expressions that could be ambiguous, relative, or even implicit, referring, for instance, to other medical events (Olex and Mcinnes, 2021).

2.4.2 Events

Aside from temporal information extraction, there are other event-related tasks, such as *event extraction*, *Slot filling* and *Topic Detection and Tracking* (Tourille, 2018). For instance, *Event extraction* aims to extract event triggers and classify event types for a given event mention, which is usually a sentence in which the event is described, as formulated in the ACE 2005 program (Doddington et al., 2004). In our work, we are interested in the definition of events according to temporal information extraction, where the purpose is, however, to locate an event in time rather than extract its arguments.

According to TimeML (Pustejovsky et al., 2003a) and ISO-TimeML (Pustejovsky et al., 2010), "an event is a cover term for situations that *happen* or *occur*, including predicates describing *states* or *circumstances* in which something obtains or holds true". An event may also be defined as something that occurs, and that can be associated with a timestamp. Events are generally conveyed using tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases. However, the definition of events is highly domain and application-dependent. In the context of temporal information extraction, the event extraction task is defined as in the TempEval campaigns, with the purpose of identifying the extent of the events in a text as described by the TimeML EVENT tag and their associated CLASS.

There were few rule-based methods to tackle the temporal event extraction task (Zavarella and Tanev, 2013), but most strategies were learning-based,

including methods based on SVM (Chambers et al., 2007; Bethard and Martin, 2006; Bethard, 2013) and on CRF (Llorens et al., 2010; Kumar Kolya et al., 2010; MacAvaney et al., 2017) models. Few deep learning-based techniques have been also presented, such as LSTMs (Meng et al., 2017). Han et al. (2019) proposed a neural structural SVM model to extract simultaneously events and their temporal relations.

Clinical domain. In the clinical domain, the definition of an event differs from the general domain. An event is defined as a clinically relevant situation (Galescu and Blaylock, 2012). According to Styler IV et al. (2014), any entities that come under these Unified Medical Language System (UMLS) (Bodenreider, 2004) categories can be defined as events: *Disorder*, *Chemical/Drug*, *Procedure* and *Sign/Symptom*.

Most approaches in the clinical domain cast the event extraction task as a sequence labeling temporal NER task. For the i2b2-2012 and the Clinical TempEval challenges, several event extraction approaches were proposed, in particular learning-based approaches. Roberts et al. (2013) identified the clinical events using a CRF model and detected event attributes using SVM classifiers. Lee et al. (2016) presented an HMM-SVM model to identify the spans of event mentions and time expressions along with their types. Barros et al. (2016) used CRFs classifiers to extract event mentions and by considering an event as a single word mention. MacAvaney et al. (2017); Chikka (2016) also applied CRFs and SVMs approaches to extract clinical events and their attributes. Neural methods have also been suggested for the event extraction task. Li and Huang (2016) used a CNN network to learn hidden feature representations and a MultiLayer Perceptron (MLP) to identify event spans and attribute values. Tourille et al. (2017a) introduced an LSTM-based approach for identifying time expressions and events. As already mentioned earlier, there is no universally agreed definition for events since definitions may vary depending on the application task and domain. However, extracting relevant clinical events in clinical narratives is crucial to understanding the patient's longitudinal medical history.

2.4.3 Temporal relations

Temporal relations consist of relations between pairs of text mentions, such as between time expressions (TIMEX-TIMEX) or events (EVENT-EVENT) or between time expressions and events (TIMEX-EVENT). Temporal relations were first illustrated using Allen's representation (Allen, 1983), consisting of seven relations between points in time, such as BEFORE, MEET, OVERLAP, DURING, and others. Based on these representations, several annotation schemes have been introduced, with a simplification or domain adaptation

of relations. For instance, the TimeML scheme represents the temporal relations in a TLINK tag but does not address OVERLAP relations. Gumiel et al. (2021) presented an overview of these different temporal relation representations, with a comparison of the proposed representations for the clinical domain.

Clinical domain. Based on Allen's representations and the TimeML annotation scheme, THYME-TimeML was developed as an adaptation to the clinical domain under the THYME project[12]. This scheme created a new category of temporal relation, namely the DocTimeRel relations, which consist of relations between the events and the Document Creation Time (DCT) and are considered as an event attribute. Indeed, the DCT is useful for examining the patient's clinical history and future plans, as described by doctors in clinical notes. Styler IV et al. (2014) also introduced in the THYME-ML scheme the concept of narrative containers (Pustejovsky and Stubbs, 2011), which can be thought of as a cluster of EVENTs that could be represented or *anchored* by a time expression, an abstract concept or durative EVENTs, which may involve multiple events. Instead of annotating each TLINK between each event, each event will be linked to its narrative container, and links will be established between those containers. As a result, the contained events will be linked by inference. Styler IV et al. (2014) claims that using the narrative container concept improves annotation quality by increasing the inter-annotator agreement, having the necessary annotations, and removing the confusing ones. They also state that using containers better illustrates the story-telling structure of both general and clinical domains, as doctors tend to cluster discussions around a certain date. Using the THYME-TimeML annotation scheme, each event is assigned to one of four containers: BEFORE, OVERLAP, BEFORE/OVERLAP, or AFTER the DCT. Once these DocTimeRel relations are assigned, TLINKs must be annotated with one of the following five temporal relations: BEFORE, OVERLAP, BEGINS-ON, ENDS-ON, and CONTAINS.

Temporal relation extraction models evolved from rule-based models to machine learning-based and deep learning-based models. We review these models in detail in Sections 2.4.5, 2.4.6, and 2.4.7 for both general and clinical domains. Note that for the TLINK extraction using learning-based models, a relevant step when developing approaches is first to select a strategy for generating candidate pairs. Available annotations often include only positive relation samples. Therefore, negative samples should be generated for TRE models. A widely adopted strategy was to restrict intra-sentence relations by considering all pairs within the same sentence. However, additional effort has been made to address cross-sentence relations, such as by restricting candidate pairs based

[12] http://thyme.healthnlp.org

on token windows, for instance. More details about these strategies may be found in Gumiel et al. (2021). Before delving into the TRE methods, we go over the main resources for the TRE task, including details about the proposed annotation schemes, the shared tasks, and their respective corpora.

2.4.4 Resources for temporal relation extraction

In this section, we review the main temporal annotation schemes for both general and clinical domains. We also present some known shared tasks and their associated corpora that have been proposed to the community to tackle the Temporal Information Extraction task, including the TRE subtask.

Annotation schemes. Two main annotation schemes were used to annotate corpora in the literature. In this thesis, we are more interested in the TRE task. Therefore, we will briefly review the guidelines for annotating the events and time expressions, and we will focus more on annotation details of temporal relations.

ISO-TimeML - The ISO-TimeML (Pustejovsky et al., 2010) specification is a standardization of TimeML (Pustejovsky et al., 2003a). Here, we review this specification for annotating events, time expressions, and temporal relations. Time expressions are mentions of dates, times, durations, and sets and are represented by TIMEX3 tags, as discussed in Section 2.4.1. Several other attributes could be included in the TIMEX3 tag, such as modifier tag (MOD), the function of the TIMEX3 tag within the document (functionInDocument) that could take these values: *creation_time, modification_time, publication_time, release_time, reception_time, expiration_time or None*, and other attributes. As described in Section 2.4.2, events are defined as "a cover term for situations that happen or occur". Both EVENT and MAKEINSTANCE tags are used to represent events. Indeed, the MAKEINSTANCE tag enables modeling difficult examples that require two event instances (Tourille, 2018). The main attributes that represent events are CLASS, TENSE, ASPECT, POS, POLARITY, etc.

Temporal relations can occur between two events, between two time expressions, or between an event and a time expression. The TLINK tag represents these relations using the following obligatory attributes:

- **eventInstanceID** or **timeId:** the ID of the first involved entity in the temporal link.
- **relatedToEventInstance** or **relatedToTime:** the ID of the second involved entity in the temporal link, which is associated with the event instance with ID=eventInstanceID or time expression with ID=timeID.

- **RELTYPE:** the type of relation holding between the entities with the following possible values: *BEFORE*, *AFTER*, *INCLUDES*, *IS_INCLUDED*, *DURING*, *DURING_INV*, *SIMULTANEOUS*, *IAFTER*, *IBEFORE*, *IDENTITY*, *BEGINS*, *ENDS*, *BEGUN_BY*, *ENDED_BY*.

Apart from TLINKs, there are two other links, SLINK and ALINK. The SLINK is used to annotate subordination links between two events, and the ALINK is used to represent the relation between an aspectual event and its argument event. The ISO-TimeML annotation scheme also includes a SIGNAL tag, which is a textual element that makes explicit the relation between two entities. Signals can be temporal prepositions (e.g., *on, at, to*), temporal conjunctions (e.g., *before, to when*), prepositions signaling modality (e.g., *to*) or special characters (e.g., -, / in time expressions denoting ranges).

THYME-TimeML - The THYME-TimeML (Styler IV et al., 2014) is a temporal annotation scheme developed to annotate the temporal information in clinical documents and is based on the ISO-TimeML standard. This annotation scheme has been mainly established to annotate the THYME corpus. As already discussed in Section 2.4.2, the definition of events is extended to include clinically relevant events such as diagnosis, diseases, or procedures. Several modifications have been made to the ISO-TimeML annotation scheme. For instance, event modality is no longer represented by the SLINK tag anymore but with three event attributes: contextual modality, contextual aspect, and permanence. To represent the various hypothetical statements in clinical notes, the contextual modality attribute, for example, may take the value *HYPOTHETICAL*, among other possible values. The American Joint Committee on Cancer Staging Codes (AJCC) tumor type codes are also annotated as events since they provide useful information for clinicians. The major change in representing the time expressions within the TIMEX3 tag is the addition of the new tag PREPOSTEXP, as mentioned in Section 2.4.1. Styler IV et al. (2014) also point out that SETs are common in the medical domain, particularly regarding medications and treatments. They also note that addressing time expressions in the clinical domain is more difficult since many time expressions are anchored to events rather than dates.

For temporal relations, the THYME-TimeML annotation scheme reduces the number of annotated relations to decrease the annotators' conflicts. Indeed, as discussed in Section 2.4.3, the use of the narrative container concept will result in just necessary relations being annotated, which overcomes both under- and over-annotation issues. Within this context, another category of temporal relations has been annotated, the DocTimeRel relations, which model the relation between each event and the Document Creation Time (DCT). Note

that even though the document was created after the medical examination, the DCT is considered the same as the time the clinician saw the patient. The DocTimeRel relation is annotated as an event attribute and takes the following potential values:

- **BEFORE:** when the event occurred and ended before the DCT.
- **OVERLAP:** when the event occurs during the DCT.
- **BEFORE-OVERLAP:** when the event started before and continues to be true at the DCT.
- **AFTER:** when the event is planned in the future.

As mentioned in Section 2.4.3, TLINKs between events and/or time expressions have five different types: BEFORE, OVERLAP, BEGINS-ON, ENDS-ON, and CONTAINS. Styler IV et al. (2014) recommend only annotating these links if they provide more information than the information in the DocTimeRel attribute.

Shared tasks and corpora. Many shared tasks have been proposed to solve the temporal information extraction task, particularly the temporal relation extraction task. Here, we cover these shared tasks and describe the several corpora that were introduced in the literature for both general and clinical domains.

TimeBank and AQUAINT TimeML corpora - The TimeBank corpus (Pustejovsky et al., 2003b) contains 183 English news articles and was annotated using the ISO-TimeML (Pustejovsky et al., 2010) specification. The AQUAINT TimeML corpus is quite similar to TimeBank in content, and it has also been annotated using the ISO-TimeML scheme. There are 73 news reports in this corpus. Other annotated corpora have been created in other languages based on the same specifications, such as the French TimeBank corpus (Bittar et al., 2011) and the Portuguese TimeBankPT corpus (Costa and Branco, 2012). Cassidy et al. (2014) introduced the TimeBank-Dense corpus, which contains a subset of 36 documents of the TimeBank corpus and addresses the sparsity problem in the TimeBank corpora.

TempEval shared tasks corpora - The corpora proposed in the three editions of TempEval shared tasks ((Verhagen et al., 2007, 2010; UzZaman et al., 2013)) are based on the TimeBank corpus. The corpora were annotated using a simplified version of ISO-TimeML that includes a set of six temporal relations: *before, after, overlap, before-or-overlap, overlap-or-after*, and

vague. The first shared task TempEval-1, focused on extracting three types of temporal relations: those between EVENTs-TIMEX3s in the same sentence, those between EVENTs and document creation time, and those between two EVENTS in adjacent sentences. Note that Document Creation Time is represented by a TIMEX3 tag. Three tasks were added to the second and third editions of TempEval to extract time expressions, events, and temporal relations between EVENTs which are in a syntactic dependency relation. While the TempEval-1 proposed the corpus for English, the TempEval-2 provided this corpus for six languages, and the TempEval-3 challenge presented the corpus for English and Spanish languages.

I2b2 Corpus - The Informatics for Integrating Biology & the Bedside (i2b2) corpus (Sun et al., 2013) is an annotated English corpus of 310 discharge summaries that was used during the i2b2-2012 challenge on clinical temporal information extraction. This corpus was annotated based on ISO-TimeML and an earlier version of THYME-TimeML with an extended set of relations. However, since a low inter-annotator agreement was noticed for multiple relation types, this set was restricted to three temporal relations: *before*, *after*, and *overlap*. The i2b2-2012 challenge comprises three tasks: events and time expression extraction, TLINK extraction using gold entities, and an end-to-end extraction task that combines the first two tasks.

THYME corpus - The Temporal Histories of Your Medical Event (THYME) corpus (Styler IV et al., 2014) contains clinical notes and pathology reports from cancer patients at the Mayo Clinic. This corpus was annotated using the THYME-TimeML annotation scheme and was used in the Clinical TempEval shared tasks (Bethard et al., 2015, 2016, 2017). Several subtasks have been proposed for these challenges: time expression extraction, event extraction, CONTAINS relation extraction between events and/or time expressions, and DocTimeRel relations between events and document creation time. Only the CONTAINS TLINK was used in these challenges because of the limited annotations for the other relation types. The methods in the first two editions of Clinical TempEval were developed and tested on colon cancer notes. However, in the third edition, systems were trained on colon cancer reports and tested on brain cancer reports to conduct domain adaptation. Indeed, two phases are proposed: unsupervised domain adaptation, which aims to train on colon cancer annotations and test on brain cancer annotations, and supervised domain adaptation, in which few annotations for brain cancer patients are available and could be integrated with the colon cancer annotations in training.

MERLOT - The Medical Entity and Relation LIMSI annOtated Text (MERLOT) corpus (Campillos et al., 2018) is a restricted clinical corpus built with 500 de-identified clinical notes written in French related to the Hepato-gastro-enterology and Nutrition specialties. The temporal annotation scheme in the MERLOT corpus is based on the ISO-TimeML standard. Temporal expressions are represented as TIMEX3 entities with a type attribute taking the following values: *date, time, duration* or *frequency*. Two types of temporal relations are annotated: those between events and/or time expressions and those between events and document creation time. There are six types of relations between events and/or time expressions: *before, begins on, during, ends on, overlap* or *simultaneous*. The relations between events and document creation time are represented by an event attribute with a value of *before, before-overlap, overlap* or *after*.

Gumiel et al. (2021) review other clinical corpora that were proposed for the temporal information extraction task in the clinical domain, mostly in English.

2.4.5 Rule-based methods

Few rule-based methods were proposed for temporal relation extraction. For instance, to address the three tasks of the first edition of TempEval, Hagège and Tannier (2007) presented a rule-based system based on a customized XIP parser (Aït-Mokhtar et al., 2002). Most of the other proposed approaches for TempEval shared tasks were hybrid, using both heuristic rules and statistical methods. We review these approaches in the following sections.

Clinical domain. Prior works on clinical temporal relation extraction are based on manually created rules. Gaizauskas et al. (2006) presented an algorithm for extracting a specific set of TLINKs between EVENTs and TIMEXs in the same sentence and DocTimeRel relations between EVENTs and DCT in clinical narratives. Their algorithm is based on the tense and aspect of relation entities. Wang et al. (2016) developed a method for extracting relations between EVENTs and TIMEXs using domain-specific rules on the i2b2-2012 corpus. Najafabadipour et al. (2020) identified temporal relations from Spanish clinical texts by building dependency trees using the Universal Dependency Pipe (UDPipe) tool (Straka et al., 2016). Based on the provided dependency trees, rules are created to identify the time expression related to each event.

Rules-based models require human expertise to create domain-specific rules, and such models are difficult to adapt to other domains and ensure generalization.

2.4.6 Feature-engineering-based supervised methods

A variety of data-driven approaches has been used for the TRE task, starting with traditional machine learning-based methods. Mani et al. (2006) developed a ME classifier for extracting temporal relations using extracted features from raw text. Chambers et al. (2007) addressed the identification of the same temporal relations by introducing a two-stage NB approach, adding event-specific features. Bethard and Martin (2007) proposed an SVM-based system to extract DocTimeRel relations in the challenge TempEval-1 using a set of lexical, syntactic, and semantic features. Puşcaşu (2007) proposed the best system for the three temporal relation extraction tasks of TempEval-1 by combining knowledge-based and statistical methods. In TempEval-2, participants focused on extracting temporal relations from English and Spanish texts. Llorens et al. (2010) introduced the TIPSem algorithm, which is based on CRFs using general and semantic features, achieving good results for all tasks and both languages Cheng et al. (2007) used an HMM_SVM sequence labeling model with features generated by dependency parsing. Bethard (2013) achieved the best performance on the TLINK identification and classification task in the TempEval-3 challenge by developing an SVM-based system and by using additional verb-clause relations from Bethard et al. (2007). Laokulrat et al. (2013) proposed a hybrid system, UTTime, which identifies temporal links based on a rule-based approach and then filters out some links by a classifier. Chambers et al. (2014) created the CAVEO system, which is a pipeline with ordered sieves. Each sieve is either a rule-based model or a machine learning-based model. This system was considered the best-performing feature-based system for the TimeBank-Dense corpus.

Clinical domain. There was a wide use of feature-engineering-based methods in the TLINKs classification task in the i2b2-2012 challenge. Cherry et al. (2013) divided the task into four sub-tasks: anchoring EVENTs to section time, intra-sentence EVENTs-TIMEX3s relations, inter-sentence OVERLAP relations between EVENTs, and extracting causal relations induced TLINKs, using both SVM and ME classifiers. Grouin et al. (2013), Xu et al. (2013) also divided the TLINKs into more specific subtasks. Roberts et al. (2013) presented an SVM-ranker to identify EVENTs-TIMEX3s relations and a multiclass SVM classifier for the TLINK category, using a large selection of features, including POS tags from the GENIA tagger and UMLS features. Tang et al. (2013) proposed a hybrid system that achieved the best performance for the TRE task in the i2b2-2012 challenge, using event positional information, POS tags, n-grams, dependency-related, time-related, and event-related features. Miller et al. (2013) conducted a preliminary study on the THYME corpus and explored the extraction of within-sentence CONTAINS relations using an SVM classifier with Tree Kernels.

In the 2015 Clinical TempEval challenge, Velupillai et al. (2015) presented a CRF classification approach to tackle DocTimeRel relation extraction using token-level features and used certain rules to extract Container relations. Lin et al. (2016) addressed the TRE task on three levels: a coarse level by extracting DocTimeRel relations using a multiclass SVM, a medium-grained one by extracting the CONTAINS relations, and a fine-grained level to extract Allen-style event-event and time-event relations using a combination of SVM classifiers and rules. Tourille et al. (2016) addressed the 2016 Clinical TempEval DocTimeRel extraction with a model based on the Random Forest (RF) algorithm and the container relation extraction with a Linear SVM model. For both models, they compared the use of lexical, contextual, and structural features to the use of word embeddings, which are computed on the MIMIC II clinical corpus (Saeed et al., 2011). They concluded that using features provides a more balanced system than word embeddings. Tourille et al. (2017b) adapted and evaluated their feature-based approach (Tourille et al., 2016) for DocTimeRel and intra-sentence narrative container relation extraction on the French MERLOT corpus and achieved comparable results when compared to the English THYME corpus, by replacing language sensitive resources in the preprocessing step. The UTHealth SVM-based system proposed by Lee et al. (2016) was the top system for the TRE tasks in the 2016 Clinical TempEval challenge, and it used a variety of lexical, morphological, syntactic, discourse, and word representation features. P R et al. (2017) used a stacked ensemble of gradient-boosted decision trees, random forest, and extra trees classifiers to extract the narrative container relations. By using the ClearTK CRF-based NER classifier (Bethard, 2013) for the DocTimeRel task, they achieved good results. MacAvaney et al. (2017) proposed a hybrid system combining CRFs, rules, and decision trees with a large set of features. This system outperformed other participating systems for the CONTAINS relation extraction task and unsupervised domain adaptation.

To sum up, several traditional machine learning methods have been proposed for the temporal relation extraction task. However, the performance of these methods relies heavily on human-engineered features that allow a better understanding of contextual information. Moreover, most of these approaches are restricted to extracting within-sentence relations.

2.4.7 Neural-based methods

Neural-based methods attracted interest in temporal relation extraction. Indeed, Cheng and Miyao (2017) presented a dependency path-based Bi-LSTM model to extract event-event, event-time, and DocTimeRel relations and showed good results on the TimeBank-Dense corpus without the use of any explicit features or external resources. Meng et al. (2017) proposed LSTM-based

models to extract intra-sentence, cross-sentence, TIMEX-TIMEX, and DocTimeRel relations, using shortest dependency paths as input. Their method outperformed state-of-the-art systems. Han et al. (2019) proposed a neural structural SVM model to extract events and their temporal relations jointly. Cheng et al. (2020) an event-centric model that allows learning dynamic event representations across event-event, event-time, and DocTimeRel relations using multi-task transfer learning and RNNs models. Good performance has been reached for TRE on English TimeBank-Dense and Japanese BCCWJ-TimeBank (Asahara et al., 2014) corpora. Wang et al. (2022a) proposed a DCT-centered Temporal Relation Extraction model to identify the temporal relations among events, TIMEXs, and DCT jointly using multi-task learning. Input representations are obtained using pre-trained models, and a DCT-indicator sentence is added at the beginning of the document to provide a representation for the DCT as well. Recently, Yuan et al. (2023) evaluated the ChatGPT's ability on zero-shot TRE task, and they claim that ChatGPT performs better for small classes than SOTA methods, but the performance is still very low on the TRE task, in particular for long-distance dependencies.

Clinical domain. Li and Huang (2016) used a CNN network to learn hidden feature representations, domain-specific features from the cTAKES toolkit and a MultiLayer Perceptron (MLP) to extract DocTimeRel relations in the 2016 Clinical TempEval challenge. Tourille et al. (2017a) introduced a neural Bi-LSTM architecture for the CONTAINS relation task, in which input vectors are constructed by concatenating a word2vec embedding, a Bi-LSTM character-based embedding, one embedding per Gold Standard attribute and one embedding for the type of DocTimeRel relations. To address domain adaptation, their strategy to block further training of the pre-trained word embeddings during training gave the best results in the 2017 Clinical TempEval challenge. For the supervised domain adaptation phase, combining brain cancer samples with colon cancer samples during training outperformed the results of other proposed systems. Leeuwenberg and Moens (2017a) presented another top-ranking system that is based on a document-level structured perceptron proposed by Leeuwenberg and Moens (2017b) for extracting both DocTimeRel and narrative container relations. For domain adaptation, they tried assigning a higher weight to the brain cancer training samples and representing unknown words in the input vocabulary. Dligach et al. (2017) evaluated the use of CNNs and LSTMs in extracting event-event and event-time CONTAINS relations on the THYME corpus. Galvan et al. (2018) outperformed the best 2016 Clinical TempEval system for the TRE task using a tree-based LSTM model relying on dependency information. Zhao et al. (2019) suggest carving each instance into three segments depending on the entity pair position and using associative attention networks to emphasize the related information of each

segment and reconstruct the semantic structure between the segments. Their method obtained state-of-the-art performance on the THYME corpus. Lin et al. (2019) introduced a window-based BERT-fine-tuned model for within- and cross-sentence CONTAINS relations. They evaluated their models for in- and cross-domain tasks on the THYME corpus. The best-performing model was the fine-tuned BioBERT using non-XML tags and adding generated "silver instances" using the self-training technique proposed by Lin et al. (2018). Lin et al. (2020) adapted the one-pass encoding mechanism initially proposed by Wang et al. (2019) by incorporating global embeddings for long-distance relations and jointly extracting the CONTAINS and DocTimeRel relations. Alfattni et al. (2021) studied an attention mechanism built into a Bi-LSTM model on a large set of temporal relations in clinical discharge summaries, including intra-sentence, cross-sentence, and DocTimeRel temporal relations. Dligach et al. (2022) explored the use of sequence-to-sequence generative models for the 2016 Clinical TempEval TRE task by designing a variety of input/output representations. Prompting one entity at a time was the most successful representation, and using a T5 model produced competitive results with the state-of-the-art. Recently, Miller et al. (2023) proposed a multi-task end-to-end system for temporal information extraction using a multi-headed attention mechanism over a pre-trained transformer encoder. High performance has been obtained for in-domain and cross-domain settings, compared to the best systems in the 2016 and 2017 Clinical TempEval challenges.

With the emergence of deep learning methods, more works have been proposed to tackle inter and intra-sentence relations. Such methods outperformed machine learning methods, particularly attention-based models. However, more annotated corpora are required to evaluate and compare developed approaches.

2.4.8 A word about clinical section segmentation

Unstructured narrative text in EHRs contains crucial information about each patient. Clinical section segmentation seeks to automatically structure clinical text as a pre-processing step for multiple clinical information extraction tasks. Indeed, this is useful to help clinicians identify the probable location where certain information should be. For instance, if a doctor is interested in finding the drug codes, they are likely in the Medication section (Rosenthal et al., 2019). However, there is no obligation for doctors to follow a certain format and indicate sections, and even if they do, this structure is not uniform across EHRs from various hospital institutions. Therefore, the clinical section segmentation task is challenging. Section segmentation includes detecting the boundaries of sections and assigning a pre-defined label to a section. Prior works focused more on section classification, which consists in mapping sec-

tions into standard section types, either using heuristic rules or using machine learning models (Denny et al., 2008, 2009; Li et al., 2010; Haug et al., 2014). Other works focused on both section identification and classification (Dai et al., 2015; Apostolova et al., 2009; Ganesan and Subotin, 2014; Tepper et al., 2012). Deléger and Névéol (2014) introduced an automatic system to separate the core medical content from other document sections, such as headers and footers, using a CRF-based model and applied it to French clinical text. Tepper et al. (2012); Ganesan and Subotin (2014) evaluated domain adaptation by considering several corpora and came to the conclusion that there is a significant drop in performance across domains. To tackle this problem, Rosenthal et al. (2019) used sections from medical literature similar to those in EHRs to train two models: an RNN and a BERT based models. These models will then be used to predict sections in EHRs via transfer learning. Their results demonstrated that the use of medical literature data improved the performance on EHRs data. Kuling et al. (2022) built a contextualized embedding BERT model using breast radiology reports and discovered that using the contextual embedding in conjunction with auxiliary data helps to better understand the global report context in the section segmentation task. To facilitate domain adaptation, Zhou et al. (2023b) cast the section classification task as a SOAP ("Subjective", "Object", "Assessment" and "Plan") classification task and used continued pre-training to improve the transferability of BERT-based models, showing that continued pre-training only improves transferability when target domain samples are included. Zhou et al. (2023a) evaluated the ability of large language models (LLMs) to perform SOAP classification and showed that an ensemble method combining BERT and LLMs produced the best results and that LLMs performed better on the rare category while BERT performed better on the most prevalent categories.

2.4.9 Summary

While good results can be obtained for extracting entities, including temporal expressions and events, temporal relation annotation and extraction remain challenging. Indeed, temporal relations have poor inter-annotator agreement scores, which are much lower than other clinical tasks, such as event and temporal expressions annotation tasks (Verhagen et al., 2007). Although attempts have been made to increase inter-annotator agreement scores by reducing the set of temporal relations, such efforts are insufficient to annotate temporal relations in clinical texts. For the clinical domain, the annotation process involves specific domain expertise, which is costly and time-consuming. Therefore, most of the proposed works on TRE are related to datasets provided by shared tasks. The several shared tasks proposed for both general and clinical domains helped the research community in developing and comparing their extraction methods. However, this limits the evaluation of methods for other texts or languages. For

instance, only a few works were proposed for French due to the lack of publicly available annotated resources (Tourille et al., 2017b). Temporal relations can be DocTimeRel relations between an event and the DCT or TLINKs, which are relations between event or/and temporal expressions. DocTimeRel relations can be used to generate a coarse-level temporal ordering, although this ordering is considered too generic for some tasks. Adding TLINKs, however, results in a more precise and fine-grained temporal representation at the expense of increasing task difficulty. DocTimeRel relation extraction depends on how well task-specific events are defined but remains less complicated than TLINK extraction (Olex and Mcinnes, 2021), which performance is still relatively low. As reviewed in this section, TRE approaches have evolved throughout time, from rule-based methods to traditional learning-based methods using different features and to neural-based methods. Attention-based techniques appear to produce superior results, particularly for TLINK extraction, where the adopted strategy for pair selection is to use token windows to cover inter-sentence relations. Nevertheless, generalization across domains is still difficult, and the performance of most proposed systems remains far from adequate for practical applications (Najafabadipour et al., 2020). Indeed, representing temporal relations between events largely depends on event definition, as well as the quality of event extraction, which makes the practical application more challenging. Even though the construction of a complete medical patient timeline is important, a coarse-level timeline would still be useful in the clinical domain to extract past, current, and future events, in particular for decision support systems that struggle with processing temporal relations. A task simplification might result in more efficient practical results. As a result, in Chapter 5, we propose a novel event-independent representation of temporal relations, making the task easier and more reproducible through different event types. In the next section, we go through the major challenges of data privacy and the main methods that have been proposed to preserve patient privacy, particularly with the wide use of modern deep learning methods in addressing NLP tasks.

2.5 Data privacy

Deep learning approaches have been the key to the technological progress of NLP methods in recent years, yielding outstanding results for many NLP tasks. However, training or deploying models on sensitive data may raise privacy concerns. Among these is the issue of accidentally memorizing sensitive data from training data, as stated in Bender et al. (2021). For instance, memorization in the biomedical domain might result in public leakage and disclosure of sensitive private patient health information. Therefore, preserving privacy is crucial for developing NLP models, particularly when dealing with clinical

personal and private data. In this section, we will review the main data privacy threats when developing NLP applications. Then, we will go over the main privacy preservation methods that have been proposed for NLP tasks, with a focus on proposed approaches for the clinical domain.

2.5.1 A categorization of privacy attacks for NLP

Sousa and Kern (2022) yield a precise classification of privacy-preserving NLP techniques in the literature by first considering the following types of threats: threats emerging from datasets, threats related to model development, and threats associated with computation scenarios.

Data threats. The most common text data privacy attacks target personal and private information contained in the text, such as author identity, demographic information, or even patient health information for clinical text. To preserve sensitive information, methods such as anonymisation (Meystre et al., 2010; Larbi et al., 2022) or de-identification (Grouin, 2013; Grouin and Névéol, 2014) have been proposed. Indeed, while diffusing corpora for research purposes, documents such as Electronic Health Records (EHRs) and legal reports are required to remove or obfuscate any identifying information. However, such methods may omit some indirect information that may lead to the identification of persons. Moreover, there have been attempts of re-identification attacks (El Emam et al., 2011; Carrell et al., 2019).

Threats targeting NLP models. Model privacy concerns might arise from attacks that could leak private data used for training, presenting a risk of revealing personal information, e.g., patient health information (Carlini et al., 2021; Pan et al., 2020). In this same context, another type of attack could target linking pieces of information that the model could unintentionally memorize, such as unique or rare training instances, which could lead to identifying individuals. Identifying a patient with a rare disease is an example of this scenario. Another known attack is the membership inference attack (Hu et al., 2022), which seeks to recover information about whether or not a certain person was in the training data samples. Unlike prior works on data leakage in Masked Language Models (MLMs) (Lehman et al., 2021; Vakili and Dalianis, 2021), Mireshghallah et al. (2022) demonstrated that MLMs are susceptible to memorization using a principled ratio-based membership inference attack. However, even though membership inference attacks are usually suggested to quantify memorization, Vakili and Dalianis (2023) demonstrated that such attacks fail to distinguish between a model trained using real or pseudonymized data. Xie et al. (2023) recently presented a novel privacy attack targeting prompt-tuning methods. Their experiments revealed that memorization also exists in these methods.

Threats targeting the computation scenario. Privacy could be put at risk while working with centralized cloud servers or distributed processing architectures. Indeed, several potential attacks could be made on servers and client devices (Sousa and Kern, 2022), leading to the leak of locally stored private data.

In the next section, we will cover the main privacy-preservation methods according to this classification of threats.

2.5.2 Privacy-preservation approaches

To address the several privacy issues that could be encountered while developing NLP models, multiple research works have been proposed according to the different types of attacks we mentioned in the previous section. In this section, we will focus on the main data privacy preservation methods that have been proposed in the literature, with a focus on the clinical domain and the NER task.

The process of data anonymization consists in removing all pieces of personal information that could lead to the identification of a person, according to the National Commission on Informatics and Liberty (CNIL)[13]. Several anonymization methods were first introduced to avoid data attacks (Sousa and Kern, 2022; Raj and D'Souza, 2021) However, true anonymization is hard to achieve (Kushida et al., 2012) and may result in losing valuable information for research purposes (Langarizadeh et al., 2018). In particular, for the NER task, we can lose information that is required for the comprehension of EHRs. Therefore, multiple strategies, such as de-identification, have been developed to reach a compromise between data privacy preservation and data value. De-identification methods have been widely used in the clinical domain (Meystre et al., 2010; Norgeot et al., 2020; Grouin and Névéol, 2014), and they consist in deleting or replacing personal health identifiers (PHI) in clinical documents, making it difficult to rebuild a link between a person and their information. De-identification approaches seem to preserve data privacy without reducing data quality and without harming the performance of the NER task (Berg et al., 2020). However, re-identification attacks are always possible, which is problematic in the clinical field (Grouin et al., 2015; El Emam et al., 2011), and sharing de-identified data remains challenging. Other strategies (Basu et al., 2021; Klymenko et al., 2022; Feyisetan et al., 2019) have also been adopted to preserve textual data, such as differential privacy (DP) (Dwork, 2008), representing a mathematical guarantee for privacy using a noise-adding mechanism.

Recently, with the success of statistical and deep learning models, many attacks have targeted shared statistical NLP models to recover sensitive train-

[13] https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

ing data through model parameters or predictions (Boulemtafes et al., 2020). Carlini et al. (2021) revealed that diverse data extraction attacks could be performed on large language models such as GPT-2 to recover training sensitive data. Membership inference attacks can also lead to privacy leakage (Shokri et al., 2017; Mireshghallah et al., 2022). To prevent data sharing, a widely used strategy is transfer learning (Ruder, 2019), which involves applying neural networks to tasks that differ from those targeted by earlier training. However, the vocabularies of models may contain specific private information. Alawad et al. (2020) introduced a cancer information extraction system, and in order to address privacy concerns and securely share their model, they limited the word embedding vocabulary to filter outpatient health information. Another potential solution has been adopted in coping with data privacy issues, namely federated learning (McMahan et al., 2017; Brauneck et al., 2023; Sheller et al., 2020). Federated learning allows distributed training of models through a central server while keeping private training data on local devices. Ge et al. (2020) proposed a NER method based on federated learning. Thus, several studies have demonstrated that it is still possible to reconstruct training data from model updates sent to servers (Truong et al., 2021; Lyu et al., 2022). Baza et al. (2020) use the mimic learning paradigm to overcome privacy problems in smart health applications. Mimic learning entails using a trained model on sensitive data to annotate large amounts of unlabeled data and using these annotations to train a new model. The main idea is to perform a knowledge transfer and be able to share the newly trained model without sharing the sensitive data. Another category of methods, particularly for cloud environments, is encryption methods that aim to encrypt data to perform training on encrypted datasets (Lee et al., 2022b; Pulido-Gaytan et al., 2021). However, such methods may lead to memory costs that are hard to manage (Sousa and Kern, 2022).

2.5.3 Summary

To sum up, we examined the main types of attacks that could target sensitive data, and we focus on proposed methods to avoid such attacks, particularly in the clinical domain, while maintaining good performance in handling NLP tasks. To truly prevent sensitive data attacks, it is best to avoid sharing the data even if it is de-identified. The same goes for statistical models that are trained on sensitive personal data in order to avoid leakage through model parameters and learning weights. As a result, sharing research findings in the clinical community remains problematic, and privacy-preservation sharing strategies should be provided. We believe that another major concern when constructing privacy-preserving neural models is the lack of metrics or methods for detecting whether or not the trained model leaks personal information or how efficiently the model preserves data privacy. Membership

inference attacks have been proposed as a privacy-preservation evaluation approach (Shokri et al., 2017; Mireshghallah et al., 2022). However, Vakili and Dalianis (2023) shows that such methods fail to detect the privacy benefits of models using pseudonymized data. To deal with all these privacy issues, there is a growing interest in creating synthetic corpora, particularly for the biomedical domain (Hiebel et al., 2023; Venugopal et al., 2022). Aside from privacy concerns, a further major problem with statistical models is their energy consumption and carbon emissions, which considerably impact the environment. In the following section, we review the carbon emissions of NLP methods.

2.6 NLP environmental impact

Recent advances in computational resources, such as Graphical Processing Units (GPUs), have enabled the intensive use of deep learning models, particularly given their impressive performance across NLP tasks. However, due to their high demand for computer resources, energy, and materials, such models have a significant environmental impact regarding Greenhouse Gas (GHG) emissions, CO_2 equivalent emissions, or carbon footprint. Other indicators include abiotic resource depletion, blue water shortage, human toxicity, etc. In this section, we first describe the main sources of CO_2 equivalent emissions that should be considered to evaluate the environmental impact of NLP computational experiments. Then, we review studies that evaluated the carbon footprint of machine learning and NLP methods, the main tools available for computing the carbon emissions, as well as the recent efforts that encourage the conduct of green AI research.

2.6.1 Sources of carbon footprint

The environmental impact in terms of carbon footprint needs to account for the entire lifecycle of Information and Communication Technology (ICT) equipment from production through use and, finally, end of life. Life Cycle Analysis usually allocates part of the GHG emitted during equipment production to the use. This phase is challenging to examine for ICT equipment since statistics on GHG emissions during manufacture are not always easily available. It should be noted that production can account for a significant portion of total GHG emissions. A French study on a data center (with Central Processing Unit (CPU) servers only) in Grenoble discovered that around 40% of the total emissions released during one hour of CPU use were due to the production phase (including emissions due to the equipment alone) (Berthoud et al., 2020). Similarly, according to another recent study, most of the environmental impacts of mobile and data center computing equipment are attributed to hardware manufacturing and infrastructure, while the impact of operating

energy consumption is decreasing (Gupta et al., 2021). Due to the lack of data, assessing the end-of-life phase of ICT is also extremely difficult. To conclude, at least four sources of CO_2 equivalent emissions should be included when assessing the environmental impact of computational experiments: 1/ production of hardware equipment: router, PC, server; 2/ idle use of the hardware; 3/ dynamic use of the hardware; and 4/ end of life of the equipment.

2.6.2 Empirical studies

Recent efforts have been undertaken to evaluate the environmental impact of NLP methods, particularly those using deep learning models. Strubell et al. (2019) were among the first to study this, examining the carbon impact of training various state-of-the-art NLP models and concluding that we need to lower the carbon footprint of training and using models. Schwartz et al. (2020) introduce a new study topic, namely Green AI, which refers to AI research that considers the environmental cost and impact. Bender et al. (2021) underlined the environmental consequences of constantly growing the scale of AI models in general. Researchers have been more concerned with enhancing state-of-the-art task performance by intensively executing multiple experiments to create models without evaluating the environmental risks of their trials. A first step towards creating Green AI models is to measure the impact of our developed methods. Cao et al. (2020a) carried out experiments on the energy measurements of NLP models. Further research has been done to compute the energy use and the carbon footprint of deep-learning NLP architectures and large language models, such as T5 (Patterson et al., 2021), GPT-3 (Patterson et al., 2021), and Bloom (Luccioni et al., 2022). Other studies discussed the impacts of privacy-preserving machine learning methods such as Federated Learning (Qiu et al., 2023) and differential privacy (Naidu et al., 2021). For the evaluated NLP task, Naidu et al. (2021) showed that increasing the degree of privacy can entail a significant computing cost, which will inevitably raise the carbon footprint of the model's training. Therefore, more efforts are required to strike a compromise between privacy and reduced carbon emissions. Luccioni and Hernandez-Garcia (2023) recently presented a survey of carbon emissions of 95 ML models over time and across different tasks in NLP and computer vision.

2.6.3 Tools for measuring carbon footprint

To assess the carbon footprint of AI and NLP models, several tools have been proposed. Some tools run in parallel to model training and compute the energy use and the CO_2 equivalent measures, such as Carbontracker (Anthony et al., 2020), Experiment Impact Tracker (Henderson et al., 2020), Energy Usage (Lottick et al., 2019) and Cumulator (Tristan Trebaol and Ghadikolaei, 2020), while others are online tools and provide emission estimations based

on user-supplied information, such as Green Algorithms (Lannelongue et al., 2021) and ML CO2 Impact (Lacoste et al., 2019). In Chapter 3, we review the availability and the use of these six tools to measure the carbon footprint of NLP methods. Following our study, Bouza Heguerte et al. (2023) present a study that gives additional details, including measurements process, infrastructure, default values, and sources of information used by some of these tools with the addition of the newest version of ML CO2, namely CodeCarbon[14] and Eco2AI (Budennyy et al., 2023) tool. As it will be covered in Chapter 3, the carbon emissions estimations produced by the studied tools vary significantly, making it difficult to determine which tool is best for carbon emissions measures. Moreover, all the tools evaluate the carbon footprint by only considering one source of emissions: dynamic use of the hardware equipment.

2.6.4 Towards the development of efficient models

Lately, efforts have been made to encourage sustainable AI by building models efficiently. This includes efficiency in data use, designing and training, experiments and infrastructure, and hardware (Wu et al., 2022). Developing efficient models results in a considerable reduction in carbon footprint. This research is gaining attention in the research community through workshops such as SustaiNLP[15] and EMC2[16]. Xu et al. (2021) highlighted the progress achieved so far in developing Green deep learning methods by examining the most efficient Green approaches. Luccioni and Hernandez-Garcia (2023) showed that good performance could be achieved with low carbon emissions using the recent advances in training machine learning methods efficiently.

2.6.5 Summary

To summarize, we went through the main sources of CO_2 emissions that should be considered to evaluate the carbon footprint of deep learning methods. We then review surveys and research works that have been working on the evaluation of carbon footprint, in particular for the NLP domain. These studies helped raise awareness about the huge carbon footprint of deep learning algorithms, which are becoming increasingly popular due to their high performance on several NLP tasks. To measure the carbon footprint, several tools have been proposed, giving measures during the training process, as well as measures post-training based on user-provided information. In the next chapter, we review six tools and evaluate their measurements on the NER task. Although recent studies aim to provide standards for reducing carbon footprints when creating and training deep learning systems, more awareness is required for the community, which is currently focusing on enhancing the per-

[14] https://codecarbon.io/
[15] https://sites.google.com/view/sustainlp2023
[16] https://www.emc2-ai.org/

formance of models while neglecting environmental impacts. As a first step in developing efficient models, it is important to create carbon tracking methodologies that are easy to adopt and understand. Moreover, energy and carbon measurements should be reported while studying the performance of novel proposed models. More efforts should also be made to examine and compute the carbon footprint of models during their whole life cycle, not only the training phase.

2.7 Conclusion

Throughout this chapter, we have introduced the main concepts and tasks that interest us in this thesis, namely named entity recognition, privacy preservation, and temporal relation extraction. In the remainder of this manuscript, we will tackle these tasks by proposing methods for French clinical NER, where the goal is constructing privacy-preserving shareable models and a novel event-independent temporal information representation that could be applied to several domains. Besides the privacy concerns in deep learning models, such models can also have a high environmental impact regarding CO_2 equivalent emissions. As a result, in the next Chapter 3, we examine available carbon footprint measuring tools and evaluate their application on NLP methods, in particular, the NER task. A tool will be selected for measuring the carbon footprint of all our thesis experiments.

Chapter 3

Towards a better understanding of NLP environmental impact: A review of existing carbon footprint measurement tools

The material of this chapter is based on the publication in the SustaiNLP EMNLP workshop (Bannour et al., 2021).

3.1 Introduction

Modern NLP makes intensive use of deep learning approaches because of the great performance they provide for a variety of tasks. However, as discussed in Section 2.6, such methods can have a significant environmental impact in terms of carbon footprint due to the consumption of computational facilities used to run them. This impact has been increasing over the years and is affecting populations that can be different from those generating the impact (Schwartz et al., 2020). To tackle the environmental impact of these methods, the first

key step is to have the appropriate tools to measure and compare the carbon footprints. Some tools have been developed to assess the carbon emissions of statistical models, in particular, the training phase. However, there is yet no standard measurement tool for calculating carbon footprint. Therefore, in our work, we aim to conduct a systematic review of tools available for measuring the impact of NLP tools and to offer a comparative analysis from the perspective of calculated impact measures and usability. We seek to understand the methods implemented by the tools and the criteria used to assess the impact. For that, we identify the list of available measurement tools, characterize them with respect to the scope of impact information provided and usability, and apply the selected tools to assess the impact of named entity recognition experiments and compare the obtained carbon measurements in two computational set-ups (local server vs. computing facility).

In this chapter, we describe our study on existing carbon footprint measurement tools and their use to assess the impact of NER experiments. In Section 3.2, we explain our tool selection approach, our defined evaluation criteria, and the list of final selected tools. Then, we report, discuss, and compare the application of these tools to evaluate NER experiments, which are performed on different computational set-ups, in Section 3.3. Finally, in Section 3.4, we summarize our study and briefly go over some of the recent studies.

3.2 Selection of tools

In this section, we will present the selection procedure, the defined evaluation criteria as well as the final list of selected tools for our study. We essentially aim to evaluate these tools by understanding their methods of implementation and their criteria to assess impact.

3.2.1 Selection process

We started with a preliminary set of tools that were identified by a Working group on the environmental impact of AI in the French group EcoInfo[1] (Experiment Impact Tracker, Pyjoules, and Carbontracker). We then extended this list by using snowballing to collect publications that cited these tools (according to Google Scholar). We also assessed "related papers" for papers published on ArXiv when available. This process was repeated for each newly identified tool. Note that the selected tools should be freely available, usable in our programming environment (Mac/Linux terminal), documented in a scientific publication, suitable for NLP experiments, and providing a CO_2 equivalent

[1] https://ecoinfo.cnrs.fr/

measure. As a result, tools like pyJoules[2] were excluded since they do not provide a CO_2 equivalent measure. The same goes for under-development tools that do not include code or platform (Zhang et al., 2020; Shaikh et al., 2021). Figure 3.1 shows detailed results of our literature search for identifying carbon footprint measurement tools. Google Scholar yielded 94 publications, with an additional 20 from ArXiv core related works. 85 articles were examined after de-duplication. We found that many (N=43) offered opinions or discussions of carbon impact measurement in machine learning, NLP, and other fields. Another 27 (shown by the orange flow) reported studies that measured the environmental impact of experiments using one of the selected tools. Strubell et al. (2019) presented research examining the impact of NLP experiments using approaches (Nvidia and Intel RAPL system management interface) that are currently implemented in several of the selected tools.

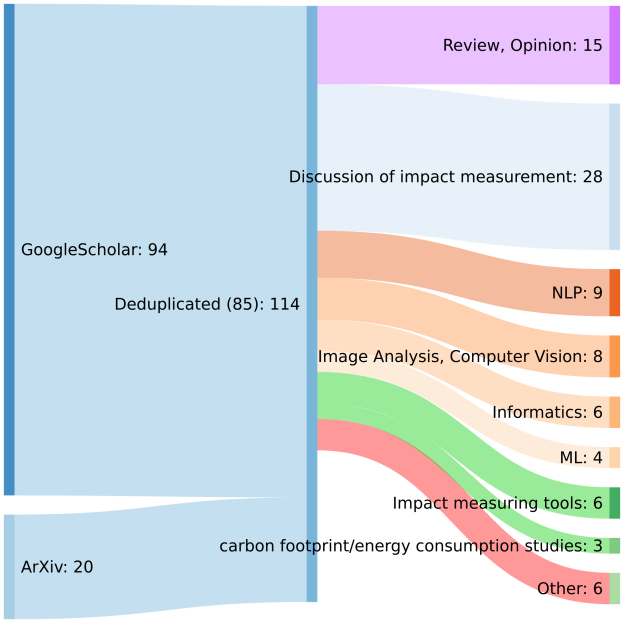


Figure 3.1: Sankey[3] diagram showing the publications reviewed in our literature search for selecting carbon impact measurement tools.

3.2.2 Evaluation criteria

In order to evaluate and compare the selected tools, we defined several criteria to characterize the availability and documentation of tools as well as the technical aspects. This comprises the type of hardware covered, the type of measure provided, the details used to determine data center electricity, and the carbon intensity for electricity production based on location. We split our

[2] https://pypi.org/project/pyJoules/

evaluation criteria into 4 categories: (a) publication, (b) technical, (c) configuration, and (d) functional criteria. Publication criteria concern scientific publications and citation details. Technical criteria are information regarding the tool's availability, installation, documentation, and output formats. Configuration criteria refer to the ability to do manual configurations of measurement computation and which configuration aspects are considered. Functional criteria cover the type of emission sources and hardware taken into account. These criteria are presented in detail below.

Publication criteria

P1 - Year of the last publication;

P2 - Citations in Google Scholar (as of 11 May 2021);

P3 - Citations for measuring NLP experiments (as of 11 May 2021).

Technical criteria

T1 - Date of the last version (as of 11 May 2021);

T2 - Availability of the source code;

T3 - Online availability for use;

T4 - Easiness of installation; We evaluated it as "Poor" if we did not manage to install it, "Fair" if we managed to install it but needed system administration access, "Good" if we managed to install it as an ordinary user.

T5 - Quality of the documentation (companion publication or code documentation); We evaluated it as "Poor" if we did not find documentation on the tool, "Fair" if documentation is available but lacks practical usage details, "Good" if the available documentation addresses usage questions such as parameter settings and country localization.

T6 - Type of license

T7 - Output formats

[3] http://www.sankeymatic.com/build/

Configuration criteria

C1 - Local values for carbon intensity: Are local values automatically taken into account, or is a global energy mix used?

C2 - Possible (manual) configuration of carbon intensity; Yes if it is possible to configure the carbon intensity without changing the code; No otherwise; We also note whether instructions are provided to the users as to where adequate values can be found.

C3 - Possible (manual) configuration of PUE; Yes if it is possible to configure the PUE without changing the code; No otherwise; We also note whether instructions are provided to the users as to where adequate values can be found.

C4 - Platforms taken into account; which type of equipment is covered by the measurements: PC, server, cloud?

C5 - Other configuration features

Functional criteria

F1 - CO_2 equivalent emission sources taken into account; We consider the following sources, described in section 2: *production*, *idle use*, *dynamic use* and *end of life*.

F2 - Hardware taken into account: does the calculation model account for emissions from data transmission between equipment types as well as from the hardware executing the experiments?

All the tools are supposed to take both CPU and GPU consumption into account, so we did not include this criterion in our analysis.

3.2.3 List of selected tools

The following six tools were finally selected for our study:

- Green Algorithms[4] (Lannelongue et al., 2021): an online tool that calculates the energy usage and carbon footprint of computer use based on information provided by the user in a web interface: runtime, number of cores, memory requested, type of platform used (PC, local server, cloud computing), type of cores, location.

[4] http://www.green-algorithms.org/

- ML CO2 Impact[5] (Lacoste et al., 2019): an online tool that determines the energy consumption and carbon footprint of computer use based on the user-provided information including hardware, runtime, cloud provider and location of the computing facilities operated. A new version of the tool is being developed with the *Code Carbon*[6] initiative. However, by the publication of our work, it was not yet described in a scientific publication, so we have decided to assess ML CO2, which the NLP research community has used.

- Energy Usage[7] (Lottick et al., 2019): a python package developed to calculate and report the energy usage of machine learning methods.

- Experiment impact tracker[8] (Henderson et al., 2020): a python package introduced to assist researchers in measuring and reporting the impact of their machine learning experiments.

- Carbontracker[9] (Anthony et al., 2020): a python package proposed for tracking and predicting the energy consumption and carbon footprint of training deep learning models.

- Cumulator[10] (Tristan Trebaol and Ghadikolaei, 2020): a python package that estimates the energy consumption of computation based on runtime, GPU load, and carbon intensity, with a fixed value for consumption of a typical GPU. It also estimates the energy consumption of communication based on the file sizes and the 1-byte model from The Shift Project (The Shift Project, 2018). The three preceding Python programs obtain information about a machine learning program's energy usage from its GPU, CPU, and DRAM.

Table 3.1 shows the evaluation of these tools according to the previously defined criteria.

3.3 Measuring the impact of NER methods

To evaluate the use of the studied tools, we present experiments on the NER task using two computational set-ups: the use of a server within the laboratory and the use of an external shared computer facility. Two NER methods

[5] https://mlco2.github.io/impact/#compute
[6] https://codecarbon.io/
[7] https://github.com/responsibleproblemsolving/energy-usage
[8] https://github.com/Breakend/experiment-impact-tracker
[9] https://github.com/lfwa/carbontracker
[10] https://github.com/epfl-iglobalhealth/cumulator

	Carbon Tracker (Anthony et al., 2020)	**Green Algorithms** (Lannelongue et al., 2021)	**Experiment Impact Tracker** (Henderson et al., 2020)	**ML CO2 Impact** (Lacoste et al., 2019)	**energy usage** (Lottick et al., 2019)	**Cumulator** (Tristan Trebaol and Ghadikolaei, 2020)
P1	2020	2021	2020	2019	2019	2020
P2	18	4	33	35	4	0
P3	1 (Parcollet and Ravanelli, 2021)	1 (Liu et al., 2021)	3 (Cao et al., 2020b; Prasanna et al., 2020; Peng et al., 2021)	4 (Sarti, 2020; Selby et al., 2021; Chaudhary et al., 2020; Gencoglu, 2020)	0	0
T1	Dec 8, 2020	Dec 17, 2020	April 29, 2021	May 4, 2021	July 10, 2020	April 29, 2021
T2	Yes	Yes	Yes	Yes	Yes	Yes
T3	No	Yes (online)	No	Yes (online)	No	No
T4	Good	No install needed	Fair	No install needed	Poor	Good
T5	Good	Fair	Fair	Fair	Fair	Good
T6	MIT	CC-BY-4.0	MIT	MIT	Apache	MIT
T7	Generates a statement to report results	Generates a statement to report the results	Generates a statement and graphs to report results	Generates text and LaTeXcode to report the results	Generates text and pdf reports	Generates a text report
C1	Yes: carbon intensity from Energi data service for Denmark, Carbon intensity API for the UK, CO2 signal API otherwise, and European Environment Agency for the default value	Yes, carbon intensity from carbonfootprint	Yes, carbon intensity from electricitymap	Yes, pointers supplied to user, including electricitymap	Yes, carbon intensity from U.S. Energy Information Administration data and U.S. Environmental Protection Agency eGRID data	No
C2	No	No	No	Yes	No	No, but indications are given in the documentation about how to change the default value
C3	No, default value of PUE = 1.67 (2019)	No, default value of PUE = 1.67 (2019)	Yes. Default PUE (1.58) can be adjusted	Partly, for Google, Amazon, Azure cloud providers.	No PUE used but PSU loss can be set	No PUE used
C4	Install dependent	PC, local server, cloud	Install dependent	3 specific providers, private infrastructure	Install dependent	Install dependent
C5	Consumption prediction based on a number of epochs, monitoring of chosen components, conversion to interpretable numbers...	Pragmatic Scaling Factor to take into account the number of experiments, conversion to interpretable numbers, comparison with other locations	Asserting certain hardware	No	Year for the data, comparison with other locations	No
F1	Dynamic use	Dynamic use	Dynamic use	Dynamic use	Dynamic use	Dynamic use
F2	Hardware only	Hardware only	Hardware only	Hardware only	Hardware only	Hardware and communication

Table 3.1: Evaluation of the tools according to the publication (P), technical (T), configuration (C), and functional (F) criteria.

were used, one that addresses flat entity recognition (Ma and Hovy, 2016) and one that addresses both flat and nested entity recognition (Wajsbürt, 2021). These methods were evaluated on the QUAERO Broadcast News Extended Named Entity dataset and the QUAERO French Med dataset described in Section 2.3.1. In this section, we start by describing the experimental settings of our experiments. Then, we present, discuss, and compare the measurement given by our six reviewed tools when evaluated on the NER task.

3.3.1 Experimental settings

The configuration is as follows: 1 core is used; GTX 1080 Ti Nvidia GeForce GPUs are used on the lab server, while Tesla V100 Nvidia GPUs are used on the external shared computing facility; memory is 11GB on the server and 32GB on the facility; 20 CPUs are used on the facility; and the experiments were carried out in France. We documented France as the location for experiments in the impact measurement tools as directed by the tools' documentation. We hypothesize that this type of set-up can be available to NLP researchers and that it is relevant to document the implications of choosing one or the other for a set of experiments. To reproduce the NER model proposed by Wajsbürt (2021), we encode the words using the pre-trained Language Model CamemBERT$_{BASE}$ provided by the HuggingFace library (Wolf et al., 2020)[11].

3.3.2 Results and discussion

Table 3.2 presents the results of the experiments, with the CO_2 equivalent measures for training each model. To understand potential differences in CO_2 measures, we also report the corresponding energy consumption in Table 3.3.

NER performance. In Table 3.2, we report the performance of systems in terms of precision, recall, and F-measure for the NER task. As expected, the performance is higher with the nested entity recognition tool than the flat entity recognition tool, and the performance of both systems is above the median and average of participants in the 2016 CLEF eHealth task where the QUAERO French Med was used as a benchmark. However, while the results of the nested entity recognition tool are directly comparable to the shared task results, those of the flat entity extraction tool are not since 14-25% of nested entities are not taken into account. We can also note that the state-of-the-art on the Quaero French Med dataset remains the low-carbon cost dictionary-based method submitted to the CLEF eHealth shared task by the Erasmus team (Van Mulligen et al., 2016).

[11] https://huggingface.co/camembert-base

		CO_2 equivalent (g.)						Runtime (mins.)	NER metrics		
		CT	EIT	EU	Cu	MLCI	GA		P	R	F
NER (Wajsbürt, 2021)	**French News**										
	Server	237.96	78	0.496	302	290	350.15	163:39	87.49	74.85	80.68
	Facility	161.16	48	0.979	222	250	260.26	118:04	88.05	74.71	80.83
	EMEA										
	Server	9.70	30	0,00131	19	20	16.67	9:31	73.78	59.74	66.02
	Facility	8.07	1	0,002	13.7	10	14.31	6:51	77.58	58.71	66.84
	MEDLINE										
	Server	13.44	30	0,00128	26.1	20	20.68	11:55	66.62	62.11	64.28
	Facility	10.50	1	0,00259	19.4	20	20.03	9:11	79.73	78.35	78.98
NER (Ma and Hovy, 2016)	**French News**										
	Server	87.62	12	5.1	100.04	125	104.40	58:30	78.49	69.77	73.87
	Facility	46.43	6	2.87	79.05	99	102.08	46:44	80.75	70.67	75.38
	EMEA										
	Server	2.23	0.004	0.117	4.31	0	3.83	02:14	61.77	50.27	55.43
	Facility	2.28	0	0.151	3.23	0	4.99	02:27	57.46	51.98	54.58
	MEDLINE										
	Server	2.99	0	0.137	5.20	0	5.57	03:11	43.97	41.08	42.47
	Facility	2.74	0	0	0.176	0	5.67	02:58	52.39	36.68	43.15

Table 3.2: Results of NER experiments. The upper part of the table presents the results obtained with an implementation of the method by Wajsbürt (2021) while the bottom part presents the results obtained with an implementation of the method by Ma and Hovy (2016). The CO_2 equivalent measures are reported according to the six selected tools in this study, Carbontracker (CT), Green Algorithms (GA), Experiment Impact Tracker (EIT), ML CO2 Impact (MLCI), Energy Usage (EU), and Cumulator (Cu).

Differences in carbon footprint measurements. There are major differences in measures obtained by our six tools, as shown in Table 3.2. This could be due to the different values used for the average carbon intensity. For instance, in our trials, Carbontracker uses the average carbon intensity for EU-28 in 2017 (294.21 gCO_2eq/kWh) rather than the French value (around 30 to 40 gCO_2eq/kWh according to electricityMap), which overestimates the CO_2 equivalent cost. Green Algorithms uses the 2020 values from electricityMap, giving 39 gCO_2eq/kWh. Experiment impact tracker uses the 2018 electricityMap value, which gives a 47.60 gCO_2eq/kWh for France. Energy Usage is based on worldwide energy mix data from the U.S. Energy Information Administration data for 2016 and assumed carbon equivalencies by energy type. Thus, the value for France seems to be 424 gCO_2eq/kWh. ML CO2 impact uses a default value regardless of the location, which is 432 gCO_2/kWh. We looked into the data sources offered by the tools to find a more exact estimate for France, but the results for Carbon intensity varied significantly. Indeed, it

		Energy consumption (kWh)					
		CT	EIT	EU	Cu	MLCI	GA
NER (Wajsbürt, 2021)	**French News**						
	Server	0.809	1.399	0,00117	n/a	0.68	1.38
	Facility	0.548	0.865	0,00231	n/a	0.59	1.03
	EMEA						
	Server	0.033	0.053	0,0000034	n/a	0.04	0.07
	Facility	0.027	0.017	0,0000047	n/a	0.03	0.06
	MEDLINE						
	Server	0.046	0.045	0,0000030	n/a	0.05	0.08
	Facility	0.036	0.021	0,0000061	n/a	0.05	0.08
NER (Ma and Hovy, 2016)	**French News**						
	Server	0.298	0.209	0.012	n/a	0.29	0.41
	Facility	0.158	0.102	0.0068	n/a	0.23	0.40
	EMEA						
	Server	0.0072	0.007	0.00028	n/a	0.01	0.02
	Facility	0.0078	0.004	0.00036	n/a	0.01	0.02
	MEDLINE						
	Server	0.010	0.007	0.00032	n/a	0.015	0.02
	Facility	0.0094	0.005	0.0004	n/a	0.015	0.02

Table 3.3: Energy consumption in kWh for each method and experimental condition. The upper part of the table presents the results obtained with an implementation of the method by Wajsbürt (2021) while the bottom part presents the results obtained with an implementation of the method by Ma and Hovy (2016). The measures are reported according to the six selected tools in this study, Carbontracker (CT), Green Algorithms (GA), Experiment Impact Tracker (EIT), ML CO2 Impact (MLCI), Energy Usage (EU) and Cumulator (Cu).

is 53 gCO_2/kWh on Carbon footprint[12], leading to the 2018 emissions; with the most recent data available, from 2020, the carbon intensity for France is 38,95. At the time of our experiments, the value on electricityMap was 31 gCO_2/kWh. The European Commission link again gives varying values depending on the kind of electricity considered and based on 2013 values. To summarize, the carbon intensity values are different, even when considering the same country.

As illustrated in Table 3.2, the CO_2 equivalent values obtained by Green algorithms and ML CO2 Impact are higher than those returned by other tools.

[12] https://www.carbonfootprint.com

This could be explained by the fact that these two tools do not perform direct measurements of the energy consumption but estimate it based on user provided information and due to hardware options offered by these online algorithms that do not exactly correspond to our equipment. In fact, we used Tesla V100-PCIE-32GB GPUs on the computing facility. However, ML CO2 Impact only provides V100-PCIE-16GB or V100-SXM2-32GB, and Tesla V100 is the only GPU option available for Green Algorithms. Consequently, we presented the results for the V100-SXM2-32GB, which may result in a lack of precision in the results. Green algorithms yield higher results than ML CO2 Impact as well, which could be explained by the differences in measuring the carbon intensity values as illustrated in Table 3.3. Similar CO_2 equivalent measures are obtained with Carbontracker and Experiment Impact Tracker, which use the same calculation methods. The differences in measures by comparing the computer facility and local server could be explained by a difference in equipment, such as the type of GPUs. Energy usage returned lower results compared to other tools, and this seems to be because it does not consider GPU consumption.

Which tool is more efficient for measuring the carbon footprint of NLP experiments? Although these tools seem to be a good start to measuring the carbon footprint, the carbon footprint is still underestimated. Indeed, it is only evaluated based on energy consumption during the dynamic use phase of equipment, which counts only for a quarter of the emission sources. Emissions resulting from the production and end-of-life phases should be taken into account. The online tools (Green Algorithms and ML CO2 impact) are very convenient to use as no installation is necessary. Since they are used separately from running the experiments, an estimate of the experiment's impact can be obtained after conducting the experiment. However, some of the required information, such as "memory requirement" (GA) or "carbon intensity" (MLCI), is difficult to figure out. In our experience, even with direct powercap access, using the Python packages tracking real-time energy consumption (Carbon Tracker, Experiment Impact Tracker, and Energy consumption) required specific permission to read RAPL results. Therefore, admin help was necessary to use the tools. Note also that the short training times for the NER experiments, due to the use of modest sized datasets, yield impact measures of 0, as illustrated in Table 3.2, which suggests that the reviewed tools are not sensitive enough to measure small impacts. As shown in Table 3.1 , the availability of tools is quite recent and moderately used in the NLP field. As a result, further research is required to better understand the differences between the tools and to account for all sources of carbon emissions.

3.4 Conclusion

In this chapter, we discussed our conducted review of six tools measuring the carbon footprint of NLP methods by explaining our selection process and our defined evaluation criteria. We evaluated these tools according to these criteria, then evaluated and compared their use to assess the impact of NER experiments. Based on our findings, we note that the differences in measures and the used parameters of the carbon footprint measuring tools could not lead to a recommendation of tool for NLP methods. However, we chose to use the Carbontracker tool for measuring the CO_2 equivalent for all our thesis experiments. This tool is easy to use and incorporate with our equipment. Note that several versions of this tool were used due to various settings, which we will cover in the next chapters.

As mentioned in Section 2.6.3, following our work, Bouza Heguerte et al. (2023) presented a more detailed study of available carbon emissions measurement tools and evaluated two more tools, namely CodeCarbon and Eco2AI. Several studies have also promoted the creation of green AI models (Hershcovich et al., 2022; Verdecchia et al., 2023; Ligozat et al., 2020), and more research papers are measuring the impact of their proposed methods (Lakim et al., 2022; Luccioni et al., 2022). In France, the labos1point5[13] collective proposes support to research labs interested in evaluating their carbon footprint. Recently, Morand (2023) conducted a review of the environmental impacts of Natural Language Processing methods during his master's degree in the LISN lab.

[13] https://labos1point5.org/

Chapter 4

Privacy-Preserving Mimic Models for Named Entity Recognition: Application to French clinical corpus

The material of this chapter is based on these two publications: one at the Journal of Biomedical Informatics (JBI) (Bannour et al., 2022b) and one at the ATALA Day about Robustness of NLP systems (Bannour et al., 2022a).

4.1 Introduction

Electronic health records (EHRs) are typically regarded as having enormous potential to enhance clinical research. However, the majority of data contained in EHRs is in free-text form (Fu et al., 2020). Free text is the easiest and most natural way for clinicians to communicate. Moreover, up to 80% of important clinical information is only available in the form of unstructured text (Escudié et al., 2017; Jouffroy et al., 2021). To gain easier access to this information, several Natural Language Processing (NLP) techniques - information extraction methods in particular - have been proposed over the past years (Wang et al., 2018; Névéol et al., 2018a). Named Entity Recognition (NER) is the process of identifying named entities in text and classifying them into predefined categories. Having an accurate NER model for the extraction of medical concepts, such as Disease, Anatomy, Drug, Sign Or Symptom, etc., is essential for building clinical Information Extraction (IE) systems. As reviewed in Chapter 2, the NER models progressed from traditional rule-based and terminology-based models to machine learning-based and complex deep learning-based models. Supervised neural models have become the go-to approach for solving this NLP task, achieving higher performance than rule-based and terminology-based systems (Li et al., 2020a). However, to obtain high-performing supervised NER systems, large amounts of manually annotated corpora are required. The annotation process is known to be time-consuming and highly expensive. Moreover, despite the technological progress in NLP models, there are still several challenges to address in the clinical domain. The clinical narrative text is complex, incorporating many medical terminologies, abbreviations, ambiguity, poor grammar, and nested entities (Bose et al., 2021). Annotated clinical training data is often limited, in particular for non-English languages. Furthermore, the personal and sensitive nature of clinical text restricts the possibility of sharing data across institutions. Indeed, sharing data is difficult in practice and is managed by law and regulation, such as General Data Protection Regulation (GDPR)[1]. As a result, researchers can only build and test their models on the datasets owned by their institutions, and limited collaborations could be done with other institutions. Transferring NLP algorithms from one institution to another can also lead to reduced performances, as shown in Wagholikar et al. (2012). Therefore, a research challenge arises about how we can construct shareable models that maintain the right balance between performance and data privacy, particularly in a low-resource setting. In our work, we address the task of shareable named entity recognition in clinical narratives written in French. Few studies have been proposed for the French clinical NER task, which is regarded as a low-resource problem

[1] https://gdpr-info.eu/

because of the lack of publicly available annotated clinical corpora due to privacy concerns, as stated in Section 2.3 of Chapter 2. To this end, we propose a Privacy-Preserving Mimic Models architecture that enables the generation of shareable models using the *mimic learning* approach. Indeed, following the work of Baza et al. (2020), we investigate the possibility of using the *mimic learning* approach to leverage both public and private data sets. The idea of *mimic learning* is to annotate unlabeled public data through a private *teacher model* trained on the original sensitive data. The newly labeled public dataset is then used to train the *student models*. These generated *student models* could be shared without sharing the data itself or exposing the *private model* that was directly built on this data. These shareable models aim to improve knowledge transfer among clinicians and other medical institutions without revealing the personal health information of patients.

The remainder of this chapter is structured as follows. In Section 4.2, we describe the corpora we are using in our experiments. We present our Privacy-Preserving Mimic Models architecture and our used NER model, which tackles both flat and nested entities, in Section 4.3. We then describe our experiments in Section 4.4, presenting our generated shareable mimic models, the experimental settings, and the baseline models. We discuss the obtained results in Section 4.5. Finally, we report a real-world use of our contribution in Section 4.6 before concluding the chapter in Section 4.7 with our final remarks. For further research, we make available the silver annotations for two publicly available clinical corpora produced in our experiments, the source code of a NER system that addresses both flat and nested entities, as well as our best Privacy-Preserving Mimic model.

4.2 Corpora description

To conduct our experiments and evaluate our proposed models, we use these three publicly available clinical French corpora: CAS (Grabar et al., 2018), DEFT (Cardon et al., 2020), and CépiDC. These datasets are described in detail in Section 2.3.1, including their descriptive statistics in Table 2.1. The CAS and CépiDC corpora will both be used as unlabeled corpora with an equivalent number of tokens. The DEFT corpus will be used as an annotated corpus and split into a training set of 85 documents, a validation set of 20 documents, and a test set of 62 documents.

We also use a private clinical French corpus, namely MERLOT (Campillos et al., 2018). This is a restricted corpus built with de-identified patient records related to the Hepato-gastro-enterology and Nutrition specialties obtained through a use agreement with a French hospital. This corpus is not pub-

	MERLOT
Language	French
Domain	clinical
Documents	500
Tokens	148,476
Entities	39,616
Unique entities	13,830
Nested entities	3,772
% Nested entities	9.60%
Max Depth	4

Table 4.1: Descriptive statistics for the private MERLOT corpus used in our study.

licly available. However, the annotation scheme and guidelines are available to the community. The annotation scheme covers 21 entities, 11 attributes, and 37 relations. For our use, we split this corpus into 320 documents for training, 80 documents for validation, and 100 documents for testing. Table 4.1 presents descriptive statistics for the MERLOT corpus, including details about nested entities.

We also use two medical dictionaries that were available in-house:

- **UMLS-derived dictionary** – a dictionary containing French terms from the 2012AA and 2020AA versions of the Unified Medical Language System (UMLS) (Lindberg et al., 1993), terms from the Unified Medical Lexicon for French (UMLF) (Zweigenbaum et al., 2003), some terms from the International SNOMED and ICD10 terminologies, translated terms from the English version of UMLS 2012AA and validated on French corpus as well as additional synonyms (Van Mulligen et al., 2016).
- **Jeux de Mots** – a dictionary drawn from the knowledge base JeuxDeMots, in particular its specialized clinical terms component (Lafourcade and Nathalie, 2020; Lemaître et al., 2020).

Scheme annotation alignment. To compare the performance of our models, we perform an alignment step between the entity types of our used annotated corpora: MERLOT and DEFT. Table A.1 (Appendix A) describes the details of this alignment step. Note that six entities from MERLOT (i.e., Hospital, Localization, Concept_Idea, Genes_Proteins, Devices, BiologicalProcessOrFunction) have no equivalent.

There is a major ambiguity issue between diseases and signs or symptoms

since diseases can be considered symptoms in some cases (Hassan et al., 2015). Therefore, we merged these two types of entities by including the Sign Or Symptom category into the Disorder category.

4.3 Privacy-Preserving Mimic Models

In this section, we go over our proposed Privacy-Preserving Mimic Models architecture, which is based on *Mimic learning* and describe the NER model that we are using to address the task of clinical NER and which tackles both flat and nested entities.

4.3.1 Privacy-Preserving Mimic Models architecture

The main goal of our approach is to enable data providers to generate shareable models that end users could use without sharing sensitive data. Data providers could be hospital institutions with medical data warehouses having large medical patient reports. End users could be other hospital institutions, clinicians, or physicians who aim to use these models to propose better treatment strategies. Figure 4.1 depicts an overview of our proposed architecture.

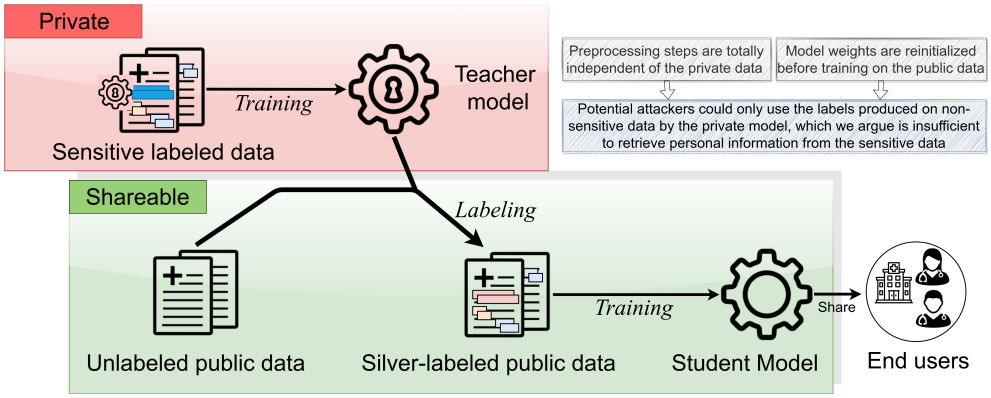


Figure 4.1: Architecture of the Privacy-Preserving Mimic Models.

Teacher model. As described in Figure 4.1, the sensitive clinical narrative reports are used to train an accurate *teacher model*. Several studies have revealed that it is possible to approximately rebuild a portion of training data by just observing the predictions (Abadi et al., 2016; Chang and Li, 2018; Boulemtafes et al., 2020). Carlini et al. (2021) revealed that diverse data extraction attacks could be performed on large language models such as GPT-2 (Radford et al., 2019) to recover training sensitive data. As a result, this private *teacher model* will only be used to produce silver annotations for public data, which will be used to train the shareable *student models*. Indeed,

the *teacher model* will be kept private, and similarly to sensitive data, it could not be shared for public use.

Student model To generate a *student model*, we use the *teacher model* to annotate the unlabeled publicly available corpus. This way, we could create a new annotated corpus. The latter is used to train the *student model*. Although we follow the same training process as the *teacher model*, this *student model* training might be viewed as a knowledge transfer process between the *teacher* and the *student model* in a privacy-preserving manner. We assess the performance of the *student model* on the original sensitive data.

As illustrated in Figure 4.1, the preprocessing steps are totally independent of the private sensitive data, and the model weights are reinitialized before training these *student models* on the silver-labeled public data. Thus, potential attackers could only use the silver labels generated by the private model on non-sensitive public data, which we argue is insufficient to retrieve personal health information from the sensitive data.

4.3.2 The NER model

The named entity recognition model, illustrated in Figure 4.2, addresses both flat and nested entity recognition and comprises three elements: the text encoder, the word tagger, and the bounds matcher.

The text encoder computes features for each word in the sequence. We first concatenate the embeddings produced by a CamemBERT (Martin et al., 2020) model, a character level features using a char CNN encoder and static French FastText embeddings (Bojanowski et al., 2017). We compute the BERT embedding of a word by averaging the embeddings of each of its subwords since BERT uses a sub-word tokenization scheme. These embeddings are then fed to a multilayer Bi-LSTM with sigmoid residual connections.

The word tagger component consists of a set of Conditional Random Field layers (Lafferty et al., 2001) that predicts entities by labeling each word of the input sequence with the BIOUL tagging scheme. Since multiple entity types in a corpus may overlap, we run multiple CRFs in parallel, one for each entity type, with five possible tags each. Each word in the sequence is classified as a (B)egin word, (I)nner word, (O)uter word, (U)nary word (a word that is both a begin word and an end word), or (L)ast word for each label. The scores obtained for each word are then run through a CRF, and the most likely BIOUL tag sequence for each label is then extracted by running the Viterbi algorithm (Viterbi, 1967). These tags are then decoded to produce candidate triplets (Begin, End, Label).

Finally, the bound matcher is used for nested entities of the same type.

Because the previous labeling scheme may generate false positives, these are filtered by checking that the begin and end words of a candidate do indeed bound an entity of the given label. This component projects each word into n_{label} begin and n_{label} end embeddings. Each candidate (B, E, L) is then scored by computing the dot product between the L^{th} begin embedding of its begin word B, and the L^{th} end embedding of its end word E. During prediction, if the tagger predicted a begin or end bound but is not associated with any other bound by the matcher, we match it with the begin or end bound that gives the highest score, even if this score is negative. This ensures that each terminal predicted by the tagger is part of at least one entity. This bound matcher is a biaffine decoder similar to the one of Yu et al. (2020).

This NER model is trained jointly by minimizing the sum of the losses of each component. The loss of the tagger is computed by summing the loss of each CRF (one per label), while the loss of the matcher is computed by adding the binary cross-entropy loss for each (begin, end, label) valid triplet. A triplet is valid when begin $\leq$ end, and the length is below the maximum entity size. In our experiments, we set the maximum entity size to 40 words. Further details about this NER model are presented in Wajsbürt (2021).

4.4 Experiments

In this section, we describe our *teacher model* trained on the clinical private corpus, the three generated Privacy-Preserving Mimic *student models* using the three publicly available corpora, and the experimental settings we use to assess the effectiveness of our approach. We also discuss the defined baseline models against which our models are compared.

4.4.1 Generated Privacy-Preserving Mimic models

As illustrated in Figure 4.3a, based on a *teacher model* trained on the MERLOT corpus, we build three Privacy-Preserving Mimic student Models trained on the three corpora: DEFT, CAS and CépiDc. The training corpus is the only difference between these three Privacy-Preserving Mimic student Models. To train these models, we incrementally augment the small portions of gold standard annotations in our disposal with silver annotations generated by the *teacher model*. The gold standard annotations are created by manually correcting the silver annotations of 20 documents (7,433 tokens) for the DEFT/CAS corpora and the silver annotations of 206 documents (2,456 tokens) for the CépiDc corpus using the MERLOT annotation scheme guideline. The agreement between the gold and the silver annotations in terms of exact F-measure

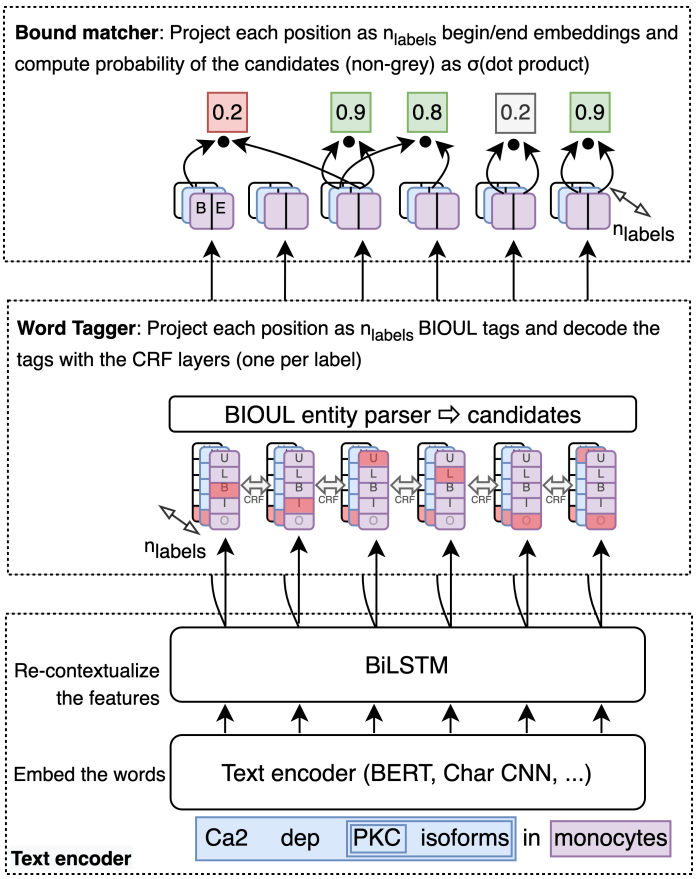


Figure 4.2: Architecture of the NER model.

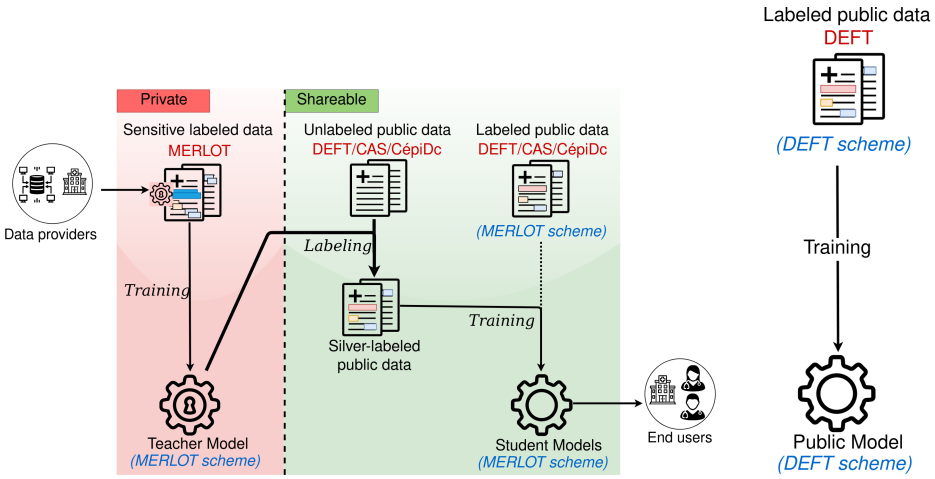
is equal to 0.758 for the DEFT/CAS corpora and 0.487 for the CépiDc corpus[2]. Figure 4.4 shows a text sample with silver annotations automatically produced by the *teacher model*.

4.4.2 Experimental settings

For our experiments, the NER model weights (including CamemBERT) were optimized with Adam without weight decay for 20 epochs. Note that this neural NER model[3] achieves 0.931 of exact F-measure, using large BERT (Devlin et al., 2019) embeddings, on the coNLL English dataset (Tjong Kim Sang and De Meulder, 2003), containing only flat entities and 0.784 of exact F-measure, using large BioBERT (Lee et al., 2020) embeddings, on GENIA (Kim et al., 2003), a widely used biomedical English dataset containing both flat and

[2]The gold and silver annotations used to create the DEFT/CAS student models are available at https://zenodo.org/record/6451361.

[3]The source code for the NER system is available at https://github.com/percevalw/nlstruct



(a) Generated Privacy-Preserving Mimic Models (b) Public Model

Figure 4.3: Figure 4.3a describes the generation process of our three Privacy-Preserving mimic student models, which are trained using three corpora: DEFT, CAS and CépiDC. Figure 4.3b illustrates a public baseline model trained on the original publicly available annotations of the DEFT corpus.

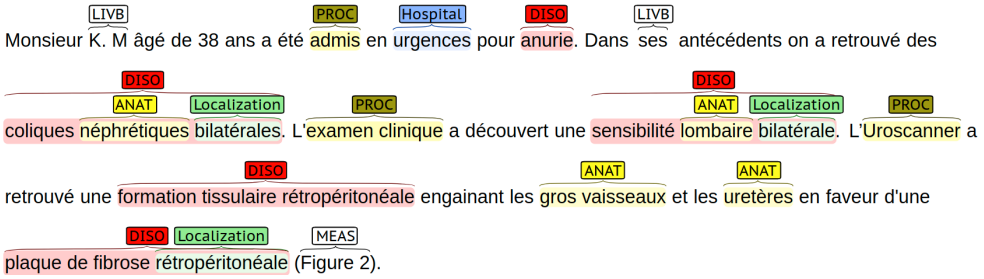


Figure 4.4: **Excerpt of the CAS corpus with silver annotations.** Translation of text into English: *"Mr K. M is a 38 yo male who was admitted to the ER for anuria. His antecedents are notable for bilateral renal colic. Upon evaluation, he was noted to have tenderness in the lower back area bilaterally. CT scan of the urinary tract showed a retroperitoneal growth encasing arteries and ureters consistent with retroperitoneal fibrosis (Figure 2)."* The annotations are correctly produced for the three first sentences, including nested entities. However, in the last sentence, the word "rétropéritonéale" ("retroperitoneal") is an anatomy entity type that was not annotated in the first occurrence and was incorrectly annotated as a Localization entity type in the second. We can also note that the annotation of "Figure 2" as a measure entity is incorrect.

nested entities. Most of our developed models in this work are trained using an early stopping with a patience of 3 epochs, except two models trained on small portions of documents (less than 8,000 tokens) where an early stopping with a patience of 10 epochs is used. We repeat each experiment 5 times. All models were trained using a GPU NVIDIA GeForce GTX 1080 Ti.

4.4.3 Baseline models

The performance of our Privacy-Preserving Mimic Models was compared to three defined baseline models: a Private Model, a Public Model, and a Dictionary-based method evaluated on two medical dictionaries. The following section details the defined baseline models and their implementation.

1. **Private Model:** This model is the *teacher model* illustrated in Figure 4.3a. The *teacher model* is trained on the original sensitive corpus.
2. **Public Model:** This model, as shown in Figure 4.3b, is trained on publicly available clinical corpora, assuming that the annotation scheme is relatively similar to the original sensitive corpus.
3. **Dictionary-based Models:** These models consist of a simple matching between the original sensitive corpus and the dictionary terms. To build these models, we use the QuickUMLS algorithm (Soldaini and Goharian, 2016).

These models are evaluated on the test set of the original sensitive corpus MERLOT.

4.5 Results & discussion

In this section, we report and discuss the results of our experiments by analyzing the privacy-preservation strategy and the NER performance. Our models are evaluated using the evaluation metrics provided in Section 2.3.7. We make a brief comparison of our models to the related work. As mentioned in the previous chapter, we also report the carbon footprint of our experiments.

4.5.1 Privacy-preservation analysis

According to the European Working Party on the protection of individuals concerning the processing of personal data[4], privacy-preserving techniques should be evaluated based on three criteria: (i) is it possible to identify an

[4] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

individual directly (ii) correlate multiple pieces of information that could lead to the identification of an individual and (iii) is it possible to infer information related to an individual. We provide below an evaluation of each of these risks related to the data and models we are releasing.

Risks related to (i) have been evidenced in solidly de-identified corpus (Carrell et al., 2019). However, we are not sharing the sensitive data itself or the private model built on this data. Therefore, we believe that retrieving personal information from sensitive data is not directly possible. Indeed, we share only the silver labels produced on public non-sensitive data by the private *teacher model*, which we argue is insufficient to retrieve personal information directly.

Risks related to (ii) involve identifying a person by linking numerous pieces of information about the same individual in the same corpus or in two distinct corpora. A worse-case scenario situation would be that the transfer of annotations from the private corpus to the public corpus consists in marking in the public corpus only entities that are present in the private corpus. In this worse-case scenario, the "silver annotations" would consist of excerpts of the private corpus. We have established that no direct identifiers can be leaked that way because the private corpus was unidentified, and the public corpus does not contain identifying information. Furthermore, the risk of recovering phenotypes (e.g., a combination of disorders or symptoms experienced by one patient) is also void because the set of annotations in the public corpus is globally aggregated. The analysis of the public annotations produced by the private model reveals that we are not dealing with the worst-case situation because many entities not present in the private corpus are, in fact, annotated.

An example of a potential attack concerning the third criterion mentioned above (iii) is the membership inference attack, which seeks to recover information about whether a specific person was in the training data samples or not. The membership inference attack model is a binary classifier whose inputs are a target data sample, a target model, and some auxiliary knowledge (Zou et al., 2020). We can consider three possible scenarios: an attack could be made on (1) the *teacher model* to infer the membership status of the private dataset, (2) the *student model* to infer the membership status of the student dataset and (3) the *student model* to infer the membership status of the private dataset. Given that we do not share the private *teacher model*, revealing information about the private corpus is not possible. As a result, the first scenario is ruled out. In the second scenario, we believe that having access to the *student model* may result in the disclosure of student dataset information. However, the student dataset consists of publicly available clinical narratives with produced silver annotations, which we make available for future research. Therefore, there is no risk of disclosure of sensitive data in this case. Concerning the third scenario, we think that access to the *student model* would not leak

information about the private corpus. Indeed, only the student dataset stated in the preceding scenario would be released, and we argue that no potential attack could reveal information about sensitive private data using the silver annotations generated by the *teacher model* on publicly available non-sensitive data. Zou et al. (2020) explored comparable attacks in the context of transfer learning and reached similar conclusions.

However, we acknowledge that the evolution of technology and the definition of privacy risks may evolve over time; the annotations and *student model* that we release may contribute to future exploration of privacy attacks.

4.5.2 Performance of NER models

We report and discuss the overall performance of our NER models on the test set of the private MERLOT corpus, as well as the results per entity type for our best Privacy-Preserving student model[5]. We investigate the impact of training data size and corpus genre, and we report additional NER experiments using two newly released biomedical language models for French.

Overall results. Table 4.2 summarizes the overall results based on an exact match of our baseline models and our three Privacy-Preserving Mimic Models trained on a combination of gold and silver standard annotations. The best results are obtained with the private *teacher model* with an F1 score of 0.857. The dictionary-based models have the worst results, with an F-measure of 0.089 for the model using the JDM dictionary and an F-measure of 0.2 for the model using the UMLS dictionary. The best performance obtained with the CAS privacy-preserving model is inferior to that of the teacher private model (0.706 vs. 0.857 of F-measure) but well above the performance of the other baseline models (0.465 of F1 score for the public NER model trained on DEFT corpus using the original gold standard annotations according to the DEFT annotation scheme). The CépiDc privacy-preserving model has the higher CO_2 equivalent measure (169 g), and the public DEFT model has the lowest carbon footprint with 22 g of CO_2 equivalent measure.

[5]Our best CAS Privacy-Preserving Mimic model is available at https://huggingface.co/NesrineBannour/CAS-privacy-preserving-model

	P	R	F	CO_2 eq. (g.)
Private Model (*MERLOT, teacher model*)	0.852	0.862	0.857	123
Public Model (*DEFT*)	0.592	0.383	0.465	22
Dictionary-based Model (*JDM*)	0.153	0.062	0.089	-
Dictionary-based Model (*UMLS*)	0.246	0.168	0.200	-
Privacy-Preserving Mimic Model (*DEFT, student model*)	0.604	0.743	0.666	30
Privacy-Preserving Mimic Model (*CAS, student model*)	**0.628**	**0.806**	**0.706**	169
Privacy-Preserving Mimic Model (*CépiDc, student model*)	0.580	0.710	0.638	394

Table 4.2: Overall results on test corpus.

Although the best results are obtained with the private *teacher model* as reported in Table 4.2, the use of this private model to create silver standard annotations on the public corpus DEFT/CAS seems to be a successful strategy to increase the performance of clinical NER with a model trained on the public corpus. In fact, a gain of 20 pts is obtained when comparing the DEFT public model trained using the DEFT original annotation scheme (0.465 of F-measure) and the DEFT privacy-preserving model (0.666 of F-measure). Good performance is also noticed for the CépiDc privacy-preserving model with an F-measure of 0.638. This solution offers a good trade-off between performance and privacy preservation. As mentioned earlier, the lowest results are obtained with the dictionary-based models. Note that no pre-processing has been performed on the dictionaries utilized in the study, and not all entity types are present in these dictionaries. In fact, only these five entity types are present: ANAT, CHEM, DISO, LIVB, and PROC. Moreover, there is a lot of ambiguity in short names and abbreviations. For instance, the word "être" can denote the infinitive form of the verb *to be* or the generic noun for *living being*. It is listed in our dictionaries as a LIVB entity, whereas the verb form is more frequent in the corpus than the noun. Due to these issues, the precision of these models remains low. Dictionary-based methods suffer as well from a low recall rate due to large variations in medical terminology and due to possible differences in the definition of entity types boundaries with the annotation guideline of our corpus.

Table 4.3 compares the performance of *student models* trained on gold annotations augmented by silver annotations produced by the *teacher model* to that of *student models* trained solely on silver standard annotations for CAS and CépiDc corpora. The performance of models trained on only silver standard annotations is very close to the performance of models trained on the combination of a small set of gold standard annotations and silver annotations

(an F1 score of 0.707 vs. 0.706 for CAS and an F1 score of 0.634 vs. 0.638 for CépiDc). These findings further demonstrate the good quality of the produced silver annotations for both CAS and CépiDc corpora. Indeed, we can observe similar results to our augmentation strategy without the need for any manual or corrected annotations for the two public corpora.

	P	**R**	**F**	**CO_2 eq. (g.)**
Privacy-Preserving Mimic Model (*CAS, student model*)	0.628	0.806	0.706	169
Privacy-Preserving Mimic Model (*CAS, only Silver annotations*)	0.631	0.804	0.707	200
Privacy-Preserving Mimic Model (*CépiDc, student model*)	0.580	0.710	0.638	394
Privacy-Preserving Mimic Model (*CépiDc, only Silver annotations*)	0.575	0.707	0.634	412

Table 4.3: Comparison of models trained on only silver annotations versus models trained on a combination of both gold and silver annotations.

Results per entity type of our best model. Table 4.4 presents the results per entity type of the CAS privacy-preserving mimic model that delivers the best results based on exact and partial matches.

	Exact match			**Partial match**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
ANAT	0.823	0.858	0.840	0.903	0.930	0.924
DISO	0.728	0.763	0.745	0.867	0.900	0.882
CHEM	0.866	0.903	0.884	0.902	0.940	0.921
MEAS	0.660	0.850	0.737	0.722	0.924	0.804
LIVB	0.336	0.875	0.486	0.377	0.952	0.540
TEMP	0.859	0.886	0.872	0.940	0.958	0.949
PROC	0.680	0.784	0.728	0.768	0.882	0.821
MODE	0.747	0.705	0.725	0.747	0.705	0.725
DOSE	0.791	0.741	0.762	0.958	0.858	0.905
Localization	0.589	0.665	0.624	0.683	0.772	0.724
BiologicalProcessOrFunction	0.625	0.535	0.570	0.672	0.571	0.610
Devices	0.654	0.716	0.679	0.864	0.902	0.885
Concept_Idea	0.668	0.775	0.717	0.699	0.812	0.751
Genes_Proteins	0	0	0	0	0	0
Hospital	0.319	0.602	0.415	0.381	0.722	0.497
Overall	**0.628**	**0.806**	**0.706**	**0.704**	**0.893**	**0.787**

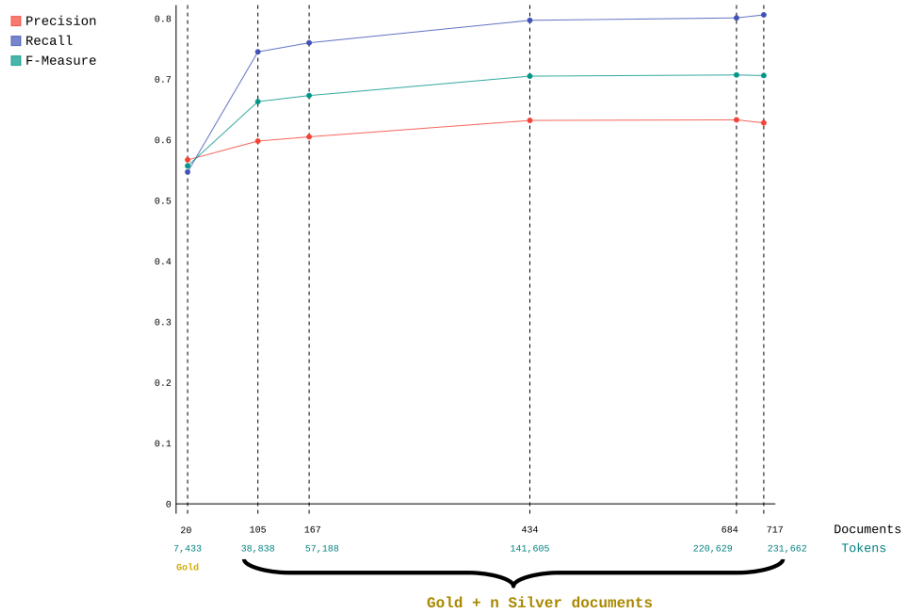
Table 4.4: Results per type entity for the CAS Privacy-Preserving Mimic Model on test corpus.

The largely covered entity types in the MERLOT distribution (see Figure 4.6) obtain the best results based on exact match. For instance, an exact

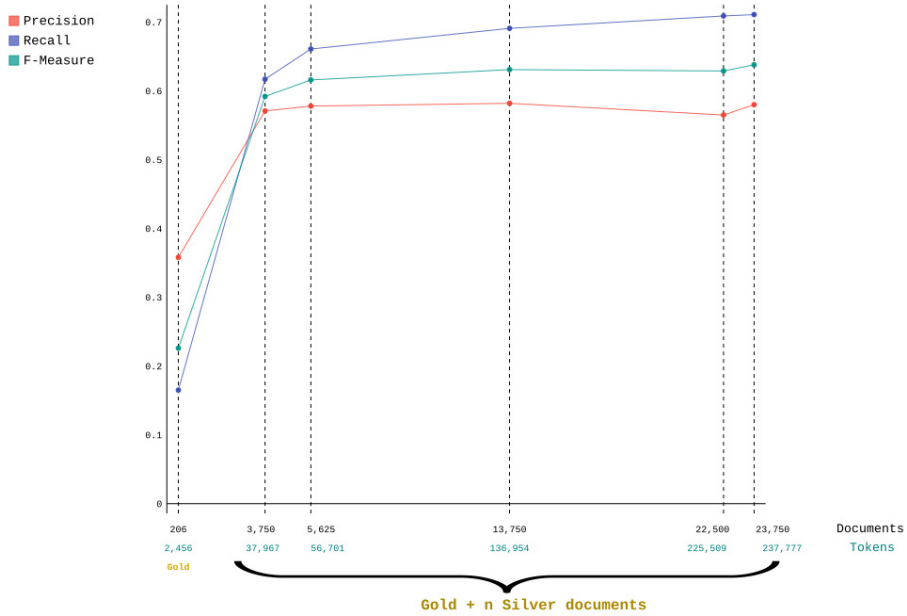
F-measure of 0.84 is obtained for the anatomy entities (ANAT) representing 12.43% of MERLOT annotations. Similar results are observed for disorders (DISO), measurement (MEAS), temporal expressions (TEMP), and medical procedures (PROC). Since these entity types are well represented in the MERLOT distribution, the *teacher model* can produce accurate silver CAS annotations, and therefore good performance is achieved by the CAS *student model* for these relevant entities. For poorly represented entities such as Genes and proteins (Genes_Proteins) (0.014% of MERLOT annotations), Living beings and persons (LIVB) (0.16% of MERLOT annotations), healthcare institutions (Hospital) (2.25% of MERLOT annotations) and Biological process or Function (2.53% of MERLOT annotations), low F-measures are observed (less than 0.6 of exact F-measure for LIVB, Hospital and Biological process or Function and 0 for Genes_Proteins). However, high F-measures are also reported for some poorly represented entities in MERLOT, such as chemical drugs (CHEM) (3.84% and an exact F-measure of 0.884), drug forms and administration routes (MODE) (0.7% and an exact F-measure of 0.725), dosage and strength (DOSE) (2.59% and an exact F-measure of 0.762) and concepts and ideas (Concept_Idea) (8.28% and an F-measure of 0.717). This may be due to the well-defined nature of these entities. As for the Localization and the diagnosis or treatment devices (Devices), which account respectively for 2.35% and 2.97% of MERLOT distribution, an exact F-measure of 0.624 and 0.602 are respectively observed. Localization entities are often embedded in anatomy entities. As a result, it is difficult to distinguish the boundaries of the two entities. For example, in the MERLOT annotation guideline, "membres inférieurs" (*"lower limbs"*) is annotated as an anatomy entity type, whereas the CAS privacy-preserving model also predicts "inférieurs" (*"lower"*) as Localization. We can also have Localization entities such as "au niveau antérieur" (*"at the anterior level"*) in MERLOT while the CAS predicted entity is rather "antérieur" (*"anterior"*). That is why we can notice a difference of 10% between the exact match F-measure and the partial match F-measure for the Localization entity type. Issues with boundary definition are common for the device's entity type, particularly for extended device names. For instance, "Coloscope CFQ 145I (194315) BIO 194315 Et Vidéo PCF 160 AL (194315)" is predicted by our CAS model as two devices entities "Coloscope CFQ 145I" and "Vidéo PCF 160 AL (194315)". This explains the observed difference of 20.6% between exact match F-measure and partial match F-measure for this entity type.

Influence of training data size. Figures 4.5a and 4.5b present the impact of increasing the training corpus size on the performance of the DEFT/CAS and CépiDc privacy-preserving models. Each experiment is realized using an equivalent number of tokens for both DEFT/CAS and CépiDc

corpora. Better performance in terms of F-measure is noticed while augmenting the training corpus size with Silver annotated documents.



(a) DEFT/CAS



(b) CépiDc

Figure 4.5: Performance as the training data size increases.

As shown in 4.5a and 4.5b, exact F-measures of 0.226 and 0.557 are obtained respectively for the CépiDc and DEFT/CAS corpora, when using solely gold standard annotations (206 documents of CépiDc corresponding to 2,456 tokens and 20 documents of DEFT corresponding to 7,433 tokens) in the training corpora. However, by incrementally adding produced silver annotations, we reach maximum performance with respective F-measures of 0.706 and 0.638 for the DEFT/CAS and CépiDc corpora, respectively. This performance is achieved using an equivalent number of tokens for both corpora: a total of 717 documents corresponding to 231,662 tokens for DEFT/CAS and a total of 23,750 documents corresponding to 237,777 tokens for CépiDc. Building such a number of manually annotated documents is difficult and time-consuming. Therefore, we believe that generating silver standard annotations is a good way to increase performance and generate accurate privacy-preserving models.

Influence of the annotation scheme. Figure 4.6 illustrates the frequency distribution of gold annotations of entity types for MERLOT and DEFT corpora as well as the frequency distribution of the generated silver annotations of entity types for CAS and CépiDc corpora.

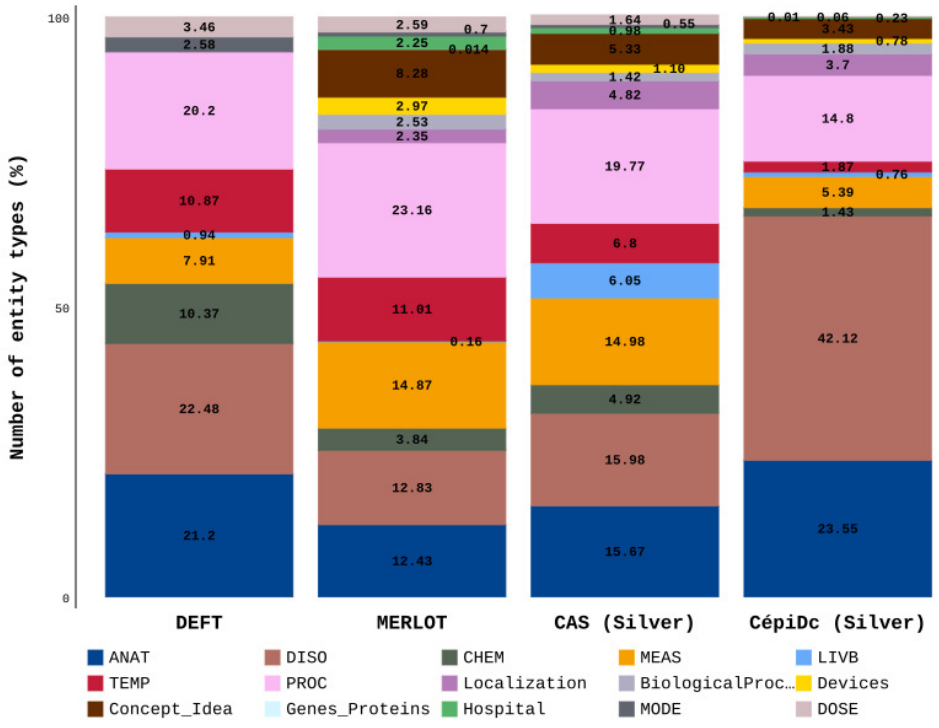


Figure 4.6: Frequency distribution of annotations of entity types.

The best results for NER are obtained with the privacy-preserving model

that shows the closest distribution to the private data, namely, CAS silver standard annotations compared to MERLOT. We can also notice that for the DEFT corpus, the best results are also obtained when the annotation scheme used in training data is the same as that of the target private data (MERLOT). Despite the equivalence drawn between the DEFT public annotation scheme and the MERLOT annotation scheme, the lower performance of NER for the public model implies that the definition of equivalent entities differs significantly. An analysis of the annotated data shows that the entities in the DEFT scheme tend to have larger spans than in the MERLOT scheme, and in some cases, the two schemes diverge on entity types to be assigned to specific text snippets. For example, the phrase "tension artérielle de la patiente demeure acceptable (91–106/53–59 mm Hg)" *(patient blood pressure remained adequate (91–106/53–59 mm Hg)* was annotated as a sign and symptom entity in DEFT while it would be annotated partly as a Biological Process Or Function ("tension artérielle" / *blood pressure*), person ("patiente" / *patient*) and measure ("acceptable" / *adequate* as qualitative measure and "91–106/53–59 mm Hg" as quantitative measure). This type of divergence in schemes impacts both precision and recall when comparing the two options. The good performance of the Public Model on the DEFT test data supports this hypothesis (Precision: 0.778, Recall: 0.798, F-measure: 0.788).

Influence of corpus genre. Death certificates are short documents (on average, 10 tokens/document vs. 323 tokens/documents for CAS and 297 for MERLOT) with a specific structure, where each line contains information on the cause of death, starting with the most immediate cause and going back to the general health status of the patient. We also computed a measure of similarity between the language distributions in the study corpora (Seddah et al., 2012) and found that CAS was closer to MERLOT (noisiness score of 0.27) than CepiDc (noisiness score of 1.02). The entities found in death certificates are mainly disorders and anatomy: Figure 4.6 shows that these two entity types account for 2/3 of all entities in the corpus. This is due to the nature of the documents, which relate to the medical problems experienced by the patient leading to their death. The focus is, therefore, on problem description rather than treatment, diagnosis or procedures, which are also found in clinical notes - and case reports contained in CAS.

NER results using recent biomedical language models. With the recent release of two biomedical language models for French, CamemBERT-bio (Touchent et al., 2023) and DrBERT (Labrak et al., 2023), we have done additional experiments by generating teacher and student models using these language models and using the MERLOT and the CAS corpora. We use the DrBERT-4GB and the camembert-bio-base models. Table 4.5 compares

the results of these experiments with our private teacher model and our best CAS privacy-preserving mimic model based on the CamemBERT model. The best results are obtained using the CamemBERT-bio embeddings, with an F-measure of 0.869 for the teacher private model and an F-measure of 0.723 for the CAS privacy-preserving mimic model. Lower results are obtained using the DrBERT model, with an F-measure of 0.848 for the teacher model and an F-measure of 0.707 for the CAS privacy-preserving student model. Despite being trained from scratch using domain-specific biomedical corpora, the DrBERT model performs worse than the CamemBERT-bio model, developed via continual pretraining from a French model. Moreover, we can notice that slight improvements could be obtained using domain-specialized language models, particularly the CamemBERT-bio embeddings. Such observations make us wonder if building and using these specialized models is worthwhile, especially given their considerable computing and environmental impacts and privacy leakage risks.

	Precision	**Recall**	**F-Measure**	**CO_2 eq. (g.)**
Private Model, CamemBERT (*MERLOT, teacher model*)	0.852	0.862	0.857	123 (*)
Private Model, CamemBERT-bio (*MERLOT, teacher model*)	**0.863**	**0.876**	**0.869**	**30**
Private Model, DrBERT (*MERLOT, teacher model*)	0.845	0.852	0.848	46
Privacy-Preserving Mimic Model, CamemBERT (*CAS, student model*)	0.628	0.806	0.706	169 (*)
Privacy-Preserving Mimic Model, CamemBERT-bio (*CAS, student model*)	**0.650**	**0.814**	**0.723**	**38**
Privacy-Preserving Mimic Model, DrBERT (*CAS, student model*)	0.671	0.748	0.707	41

Table 4.5: Comparison of our models versus models trained using French biomedical language models. (*) denotes that these measures are calculated by a previous version of the Carbontracker tool

4.5.3 Comparison to related work

Ge et al. (2020) introduced a privacy-preserving medical NER method based on federated learning. A private module, composed of Bi-LSTM and CRF layers, is used to capture the characteristics of the locally stored medical data, and a shared module, composed of word-level CNN and embeddings layers, is used to capture the shared knowledge among different medical platforms. Baza et al. (2020) used the *mimic learning* approach to address the privacy

issues. This approach implies using a model trained on the original sensitive training data in order to annotate a large set of unlabeled data and using these annotations to train a new model. This way, a knowledge transfer from the original model to the newly trained one is initiated without sharing the sensitive data.

Compared to these related works (Ge et al., 2020; Baza et al., 2020), our strategy seems to better preserve the privacy of personal patient information since neither the original sensitive data nor the private model weights are shared. Despite that Federated Learning (McMahan et al., 2017) used in Ge et al. (2020) have been originally proposed to better preserve privacy by only exchanging model parameters between local nodes through a centralized server, personal information could still be extracted from local training parameters (Truong et al., 2021; Melis et al., 2019; Hitaj et al., 2017).

A direct comparison with Baza et al. (2020) is difficult due to differences in the used datasets. In fact, we encounter extra challenges while dealing with narrative clinical text due to the complexity and the variety of medical terminologies presented in the clinical text. However, our results are in agreement with the results presented in Baza et al. (2020) since *student models* are proved to be able to mimic the *teacher model* performance without access to the original private data.

4.5.4 Carbon footprint

As stated in Chapter 3, we use the Carbontracker tool to measure the carbon footprint of training our models. Note that the used version at the time of our experiments computes its estimates by using the average carbon intensity in the European Union in 2017 instead of the France value, even if it successfully detects France as the location of the experiments. Carbon footprint is reported in Table 4.2 in terms of CO_2 equivalent measure in grams. The highest CO_2 emissions are observed when training the CépiDc privacy-preserving mimic student model (394 g). Our best CAS privacy-preserving model has lower CO_2 emissions: 169 g. However, to obtain this model, we first train the *private model* to produce the silver annotations. Therefore, a total of 292 g of CO_2 emissions is estimated. Although CAS and CépiDc corpora are equivalent in the number of tokens, the CO_2 emissions value is higher for the CépiDc corpus (a total of 517 g). This could be due to the high number of documents used for training the CépiDc corpus (23,750 documents). As mentioned in Strubell et al. (2019), deep learning models can have a significant environmental impact due to the high energy consumption of the computing equipment necessary to execute them. The estimated CO_2 emissions from training both the *teacher model* and the CAS *student model* is roughly equivalent to 2.52 km traveled by car. The estimated CO_2 emissions from training both the *teacher model* and

the CépiDc *student model* is equivalent to 4.37 km traveled by car.

The carbon footprint of the newly generated models, based on CamemBERT-bio and DrBERT, is reported in Table 4.5. The DrBERT-based models have the highest CO_2 emissions: 46g for the teacher model and 41g for the privacy-preserving model. The CamemBERT-bio-based models have lower CO_2 emissions: 30g for the private model and 38g for the student model. These measures are obtained with a newer version of Carbontracker, which explains the major differences in CO_2 emissions compared to our models, based on CamemBERT. Indeed, in this version, the average carbon intensity of the detected country[6] is used.

For France, the used value is now 85 gCO_2eq/kWh instead of the old average carbon intensity in the European Union in 2017, i.e., 294.21 gCO_2eq/kWh. Moreover, the value for estimating the CO2 equivalent emission for km traveled by car is also modified. To sum up, we believe that the reported carbon footprint of our initial models is overestimated, and similar measures to those of the newly generated models could be obtained using these new parameters of Carbontracker. It is worth noting that the online tool Green Algorithms[7], which is based on information provided by the user, has created a feature that allows users to store their experiment settings and roll back to older versions of the tool when needed, allowing for better impact measurement traceability.

4.6 Practical use

For a more practical assessment of our best *student model*, our shareable CAS privacy-preserving mimic model was used on semi-structured radiology documents by computer scientists working in the medical informatics team at the Georges Pompidou European Hospital (Hôpital Européen Georges Pompidou, HEGP). These clinical documents describe the tumor progression of patients who were followed at the hospital. The analysis of radiology reports is guided by the Response Evaluation Criteria In Solid Tumours (RECIST 1.1) to define and monitor target lesions, non-target lesions, and the appearance of new tumor lesions. Indeed, the radiologist classifies response to treatment into four categories: Stable Disease (SD), Progressive Disease (PD), Partial Response (PR), or Complete Response (CR). An overall conclusion is also provided for the response to treatment of all the patient's tumor lesions, as shown in Figure 4.7. Such medical information is not stored in databases and can only be obtained in the plain text of the imaging reports. Some information from these reports was extracted using an in-house tool based on regular expressions,

[6] https://ourworldindata.org/grapher/carbon-intensity-electricity
[7] http://www.green-algorithms.org/

namely Py-Rex. However, certain information, such as anatomical locations, required the use of a more modular tool. Our privacy-preserving mimic model was therefore used to automatically extract the lesions, the anatomical entities as well as the location of target lesions (i.e., the Anatomy, Disorder, and Localization entities as defined in the MERLOT annotation scheme (Campillos et al., 2018)). Figure 4.7 illustrates an example of a radiology report with the three extracted types of entities. There has been no formal evaluation at scale, but an empirical evaluation indicates that our approach performs well in extracting both flat and nested entities. Indeed, 1864 entities describing the anatomical lesion sites are extracted from 859 radiology reports of HEGP patients from 2010 to 2020. This suggests that our method might be helpful to clinicians in future research. As a result, we make available our Privacy-Preserving NER model[8] and is now being discussed for integration into the medkit tool[9], a Python library built by the HeKA team[10] to facilitate the development of applications for learning health systems.

4.7 Conclusion

Throughout this chapter, we tackled the task of shareable Named Entity Recognition in clinical narratives in French, which may be defined as a low-resource problem from the machine learning perspective since no annotated clinical corpus is publicly available. Indeed, we studied the use of the *mimic learning* approach to leverage both public and private corpora by proposing a Privacy-Preserving Mimic Models architecture. This architecture enables a knowledge transfer to a *student model* through a *teacher model* trained on private sensitive data. In fact, the *teacher model* is used to annotate unlabeled public data. The newly labeled public corpus is then used to train the *student model*. As a result, the generated *student models* could be shared without revealing the private data itself or exposing the *private model* that was directly built on this data. Experiments on different medical corpora have shown that our strategy offers a good compromise between performance and data privacy preservation. We also provide a use case of our best shareable privacy-preserving mimic model carried out by the medical informatics team at the Georges Pompidou European Hospital as an example of a real-world use of our models. We make available the generated silver annotations for the two publicly available corpora (i.e., DEFT and CAS), the source code of the NER system that tackles both flat and nested entities, as well as our best

[8] https://huggingface.co/NesrineBannour/CAS-privacy-preserving-model

[9] https://github.com/TeamHeka/medkit

[10] https://team.inria.fr/heka/fr/

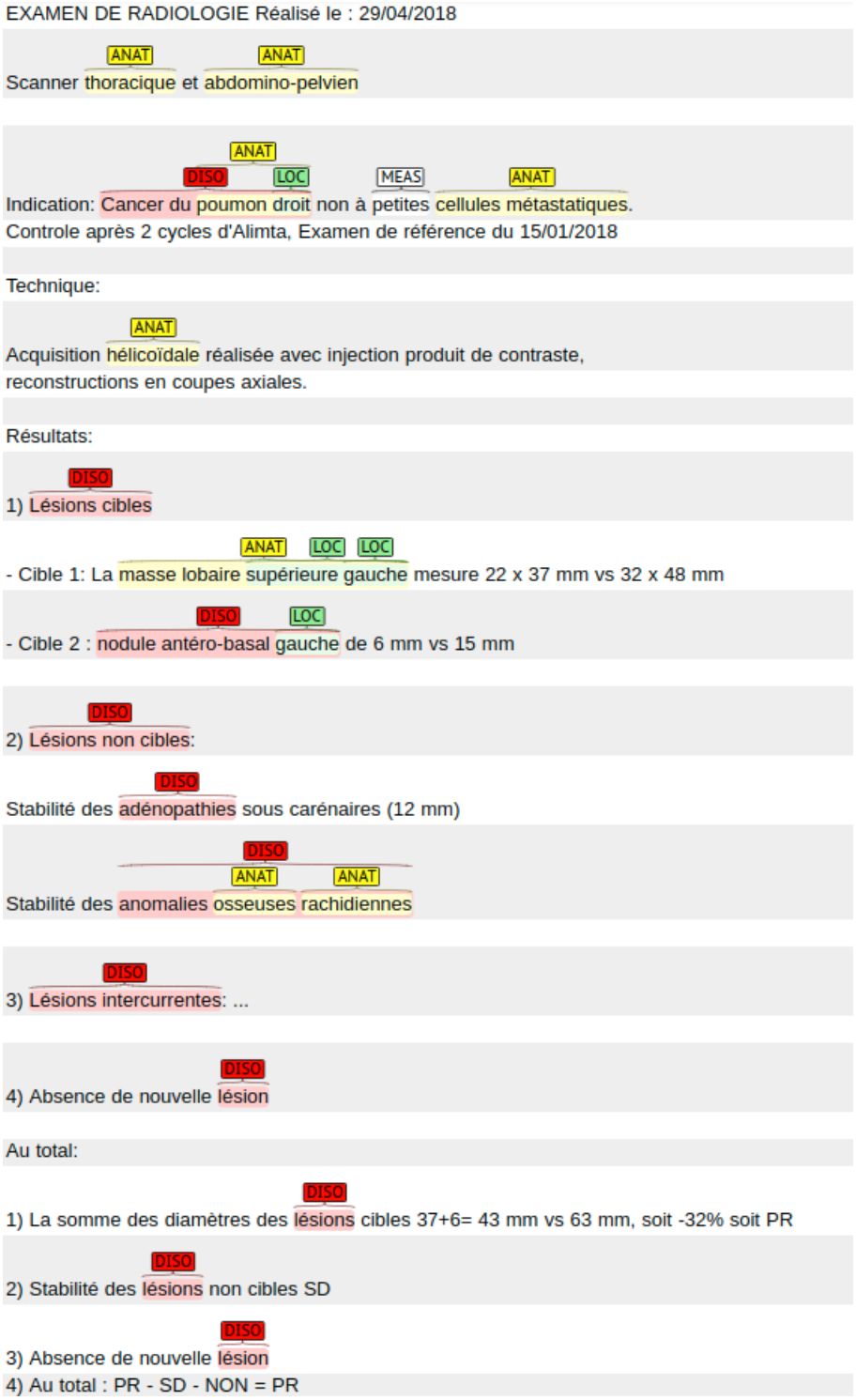


Figure 4.7: An example of the extracted types of entities from a radiology report, using our shared privacy-preserving CAS mimic model.

Privacy-Preserving Mimic Model. Note that the data privacy preservation was assessed empirically by analyzing the various attacks that could be performed, but as mentioned in Section 2.5 (Chapter 2), it would be better to have metrics or methods that could identify whether or not the trained models leak personal information or how well the models respect data privacy. The Silver annotations and the student models we offer could also be useful to future investigations of privacy attacks.

Chapter 5

Event-independent temporal positioning: application to French clinical text

The material of this chapter is based on two publications: one at the TALN conference (Bannour et al., 2023b) and one at the BioNLP workshop associated with the ACL conference (Bannour et al., 2023a).

5.1 Introduction

Constructing patient timelines entails extracting the key elements from unstructured clinical free-text notes in Electronic Health Records (EHRs), such as major clinical events, temporal expressions, and temporal relations. Temporal relations enable the ordering of temporal information about a patient's past treatments, disease evolution, treatment responses, and toxicity rates. The performance of temporal relation extraction relies heavily on the quality of extraction of events and temporal expressions, which increases the challenges in developing end-to-end systems for timeline construction, particularly when working with real-world data. Moreover, the definition of events strongly depends on the text type, the application task, and the domain, making cross-domain generalization challenging. As reviewed in Section 2.4 of Chapter 2, most proposed research efforts for temporal relation extraction on the clinical text were through shared tasks and their related datasets. This is due to the costly and time-consuming annotation process, which requires domain expertise and remains difficult even for humans, as evidenced by moderate inter-annotator agreement (Verhagen et al., 2007; Tourille et al., 2017b). In our work, we propose a novel event- and task-independent representation of temporal relations that allows the identification of homogeneous text portions from a temporal standpoint and classify their temporal positioning according to the Document Creation Time (DCT). This results in a much faster and easier task for human annotators through a simpler annotation scheme, as well as more reproducible through different event types. We argue that the loss of expressiveness of this scheme does not preclude useful applications on clinical reports. Such problem modeling does not require the prior definition of events and temporal expressions. The temporal relation extraction is cast as a sequence token classification problem. To evaluate our temporal positioning models, we have defined and extracted a posteriori the clinical events that interest us, i.e., the chemotherapy toxicity events, and infer the temporal positioning of these events using our models. Each event will have the same temporal positioning as the text portion that includes it.

In this chapter, we describe our created corpora of clinical text written in French in Section 5.3, including our annotation process and guidelines and the main challenges we experienced. Then, we present the traditional temporal relation representation and our novel event-independent representation

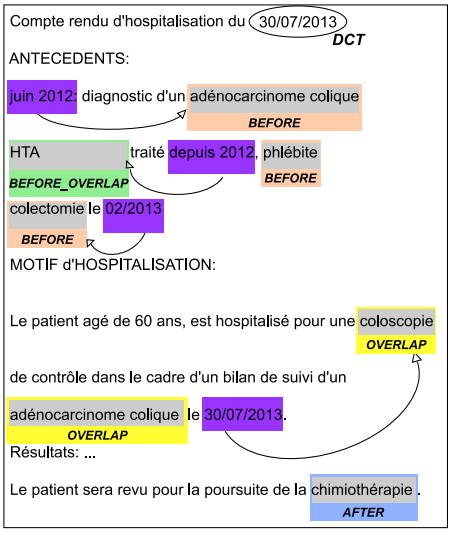
of temporal relations in Section 5.2. Section 5.4 describes our temporal positioning models, chemotherapy toxicity event extraction model, the developed rule-based baseline model, and the used evaluation metrics. We discuss the performance of our models in Section 5.5, including their performance in identifying and classifying the temporal relations between chemotherapy toxicity events and the DCT. Before concluding the chapter Section 5.7, we study the overall challenges encountered when attempting to construct patient timelines in Section 5.6.

5.2 Overview of the temporal relation representation

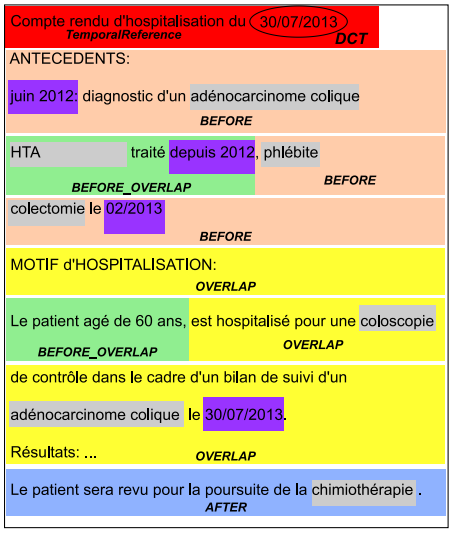
As discussed in Section 2.4 and as illustrated in Figure 5.1a, temporal relations in the text are often represented by DocTimeRel and TLINKs relations. DocTimeRel relations refer to the temporal relation of each event with the Document Creation Time (DCT). According to the THYME-TimeML scheme, each event will be assigned to one of the following categories: *Before (orange)*, *Before_ Overlap (green)*, *Overlap (yellow)*, and *After (blue)*. However, since events vary depending on the task for which they are defined, the DocTimeRel extraction task differs from domain to domain, and no generalization is possible. Moreover, the definition of clinical events presents further challenges due to the complexity and variety of medical terminology used in clinical narratives. Extracting 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 characteristics 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, as mentioned in Section 2.4.3, other strategies must be adopted to resolve long-distance dependencies if the event and the temporal expression are in different sentences. Overall, the traditional representation of temporal relations is task-dependent and requires accurate results in extracting events and temporal expressions, making the annotation and extraction tasks difficult.

Therefore, we introduce a novel event-independent representation of temporal relations. As shown in Figure 5.1b, 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. The *TemporalReference* label is assigned to the narrative portion that marks the beginning of the clinical reports, and that could include the DCT. Events will subsequently have the same temporal category as the text portion that in-

cludes them. Thus, we do not have to deal with sentence boundaries or long dependency issues. Moreover, although this representation is coarser than the traditional representation of temporal information, it is totally independent of the type of mentions to be defined and extracted and, therefore, of the application domain. Figure 5.4 illustrates an example of our event-independent temporal positioning representation.



(a) Traditional representation of temporal information



(b) Event-independent temporal positioning

Figure 5.1: 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 5.1a illustrates the traditional representation of DocTimeRel between the DCT and the events and TLINKs between the events and the temporal expressions. Figure 5.1b 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."*

5.3 Corpora description

There are no publicly temporally annotated resources that are available for French, as mentioned in Section 2.4.4. Therefore, to address the TRE task in

French and develop and evaluate our event-independent representation of temporal relations and our temporal positioning models, we create and annotate two French clinical corpora, namely the **temporal extraction corpus** and the **toxicity corpus**[1]. In this section, we detail the annotation process and guidelines, including the encountered challenges, before moving on to the description of the constructed corpora. For the creation and annotation of these corpora, we followed the methodology and annotation steps indicated by Fort (2012).

5.3.1 Annotation process

To annotate the temporal relations in clinical documents, we define an annotation scheme based on the Document Creation Time (DCT) using the possible temporal categories that DocTimeRel relations can take in the THYME-TimeML annotation scheme (*Before, Before_Overlap, Overlap, After*), as described in Section 2.4.4 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 default temporal category for *TemporalReference* is *Overlap*. 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 portions in the preprocessing step.

The DCT might be the current medical visit date or the period of time spent in the hospital, which is usually indicated in the document heading. The DCT does not need to be annotated. For each identified homogeneous text portion from a temporal standpoint, we assign a temporal category. For instance, the *Before* category could be assigned to narrative portions describing past medical events. For simplicity, we only annotate the first word of each temporal narrative portion, and we consider that the start of a temporal portion denotes the end of the previous one.

Figure 5.2 illustrates an example of annotating a clinical document containing two clinical reports. As stated earlier, the *TemporalReference* category, also denoted as *TempRef*, indicates the beginning of the clinical report. According to our annotation scheme, the narrative portion going from *Paris* to *2014* will then be assigned to the *TemporalReference* category. The text portion from *Monsieur* to *comme* is annotated 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. The second *TemporalReference* category assigned to the portion from *Dossier* to *staff* marks the start of a new clinical report, and annotations will be adapted

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

to the new DCT ('25/03/2014' in this example).

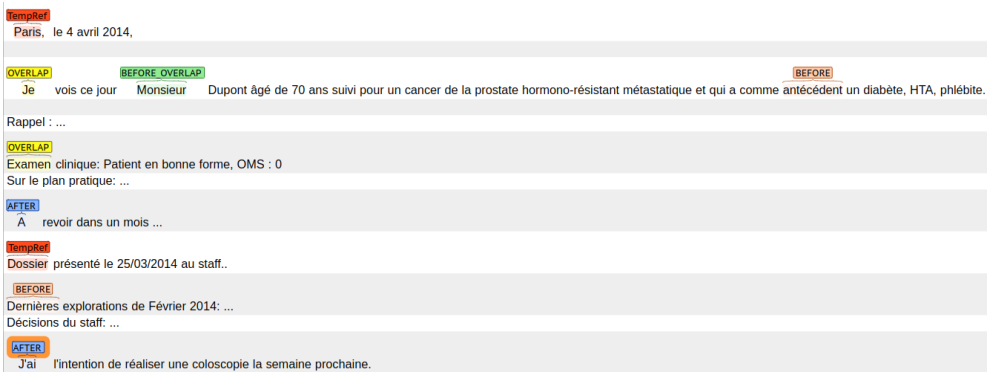


Figure 5.2: An example of annotating a clinical document containing two clinical reports. Translation of the text portion into English: *"Paris, April 4th, 2014. Mr. Dupont is a 70 y.o male with hormone-resistant metastatic prostate cancer and a medical history of diabetes, hypertension, phlebitis. Clinical history: ... Physical examination: Patient in good state of health, OMS: 0. Practical course of action: ... Follow-up in one month... Record presented on 03/25/2014 to staff... Prior workup in February 2014: ... Staff decisions: ...Colonoscopy scheduled next week."*

Annotation challenges and inter-annotator agreement. Establishing a solid annotation guideline is essential to get high-quality annotations. Therefore, despite the simplicity of our annotation scheme, several discussions were held with the annotators using draft annotation guidelines before reaching the final version.

A notable example of discussion was annotating narrative portions about medical and case history. Indeed, when describing the medical history, we can have temporal expressions that could indicate the exact time of past surgeries or diagnosis, but we can also have medical events such as *'phlébite'* (phlebitis) or *'HTA'* (*hypertension*, high blood pressure) without any related temporal expressions, as illustrated in Figure 5.2. Phlebitis can be understood as a past medical event and easily marked with a *Before* label. However, HTA is a chronic condition and could be understood as an ongoing event or a past event. Such decisions need medical expertise, particularly for complex medical events. To handle these issues, an agreement has been established by annotating medical history events as *Before* unless there is a clear temporal indication that the event is still ongoing (*depuis 2014*, since 2014) in Figure 5.3) and which needs to be annotated as *Before_ Overlap*. Another agreement has been made when annotating sentences that the doctor writes at the end of a narrative clinical report, such as *J'ai l'intention de réaliser une coloscopie la semaine prochaine* (*I intend to perform a colonoscopy next week*), in Figure 5.2. In this

sentence, a strict choice of language might be to annotate the first part from *J'ai l'intention de réaliser* (*I* to *perform a*) as *Overlap* and the second part as *After*. Such annotation, however, is both unhelpful and not practical since the annotation boundaries are unclear, and we are more interested in annotating the main medical events. As a result, the whole sentence will be annotated as *After*. Three annotators with NLP backgrounds applied to health data annotated a sample 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., 2017b).

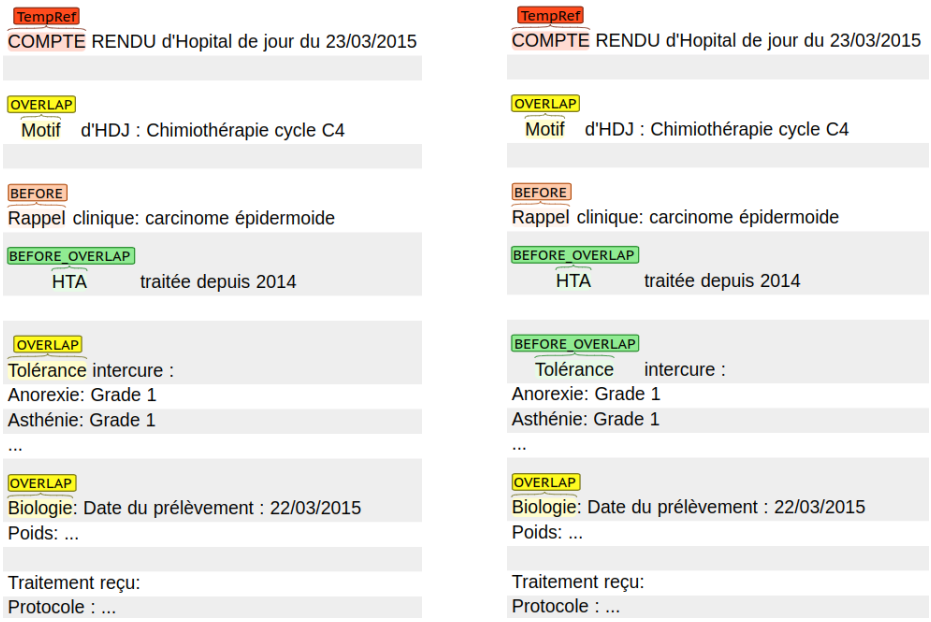
Some changes were made to the annotation guidelines when annotating our second created corpus, namely the toxicity corpus, built with documents containing toxicity treatment information. Figure 5.3 illustrates an example of an annotation modification. According to our first version of the annotation guideline, the prior chemotherapy response (*Tolérance intercure*), which is reported in each chemotherapy administration report, is annotated as *Overlap* (cf. Figure 5.3a), with the assumption that it is useful in understanding the current report. However, after discussing with a domain expert, we concluded that it is more convenient to assign a *Before_ Overlap* label to such information since it started in the past but is still true, and it is crucial to have the chemotherapy toxicity information for the actual chemotherapy administration (cf. Figure 5.3b). Note that biology tests (*Biologie* in Figure 5.3) are annotated as *Overlap*, even if they were done before the hospitalization because these tests are only meaningful and interpreted for hospitalization. The detailed final version of our annotation guidelines is provided in Appendix B, with more illustrated examples. Despite the few difficulties we encountered in defining the annotation scheme, our annotation process is easier than the standard method of annotating temporal relations and yields better inter-annotator agreement.

It is worth noting that although our representation of temporal relations is intended to be event-independent, we are aware that we may have considered medical events while developing annotation guidelines.

5.3.2 Corpora

According to our annotation guidelines, we created and annotated the following corpora:

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 annotated



(a) Prior annotation. (b) Final annotation.

Figure 5.3: An illustration of annotation modifications. Translation of the text portion into English: *"Discharge summary of 03/23/2015. Reason for admission: chemotherapy cycle C4. Clinical history: squamous cell carcinoma Hypertension treatment was initiated in 2014. Toxicity since last cycle: Anorexia: Grade 1 Asthenia: Grade 1 ... Lab workup: date of sampling: 03/22/2015 Weight: ... Regimen: Protocol: ..."*

222 documents to train and validate our models and 57 documents for evaluation. It is worth noting that our corpora always contain an equal amount of clinical documents from every category.

Table 5.1 presents descriptive statistics for each temporal category in the two temporal extraction training and test corpora.

	# text portions (test)	**# text portions (train)**
TempRef	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 5.1: The number of text portions for each category in the temporal extraction training and test corpora.

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. 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.

5.4 Experiments

In this section, we describe our temporal relations and the chemotherapy extraction models. To evaluate our event-independent temporal positioning model, we also describe the developed rule-based baseline model and the used evaluation metrics.

5.4.1 Temporal relation extraction

Using our temporal representation, we cast the temporal relation extraction task 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 from these five categories: *TemporalReference*, *Before*, *Before_Overlap*, *Overlap*, and *After*. 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 as belonging to a narrative portion using 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.

5.4.2 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). This dictionary is created using French toxicity terms from two reference terminologies: the 5th version of Common Terminology Criteria for Adverse Events (CTCAE) and the World Health Organisation Terminology (WHOART). To extract the chemotherapy toxicity events, we use the QuickUMLS (Soldaini and Goharian, 2016) algorithm. The obtained pre-annotations, as previously stated, are manually verified and corrected by a domain expert. Note that we are only interested in toxicity events related to chemotherapy.

5.4.3 Baseline model

We compared our temporal positioning 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 surgery clinical reports such as 'Antécédents' (*Case history*), 'Indication' (*Indication*), 'Gestes réalisés' (*Operative actions*), 'Plan de traitement' (*Treatment plan*), etc. For instance, if we have the keyword 'Antécédents' (*Case history*), the assigned label for the text portion *Before* until another keyword is encountered. These keywords are typically useful for the temporal annotation process, even though they do not cover all types of clinical reports. This baseline model is evaluated on the temporal extraction test corpus.

5.4.4 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. 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 are calculated for each sample. We use the BRATEval tool[2] to assess the entity-level performance of toxicity event extraction. To measure the carbon footprint of training and testing our temporal positioning models, we use, as usual, the Carbontracker tool. Note that we are using the newer version of Carbontracker, which uses the World-wide average carbon intensity of electricity production in 2019 if it fails to detect the location, and its measures are based on CO_2 performance of new passenger cars in Europe[3].

5.5 Results & Discussion

In this section, we present the results of our temporal positioning models using the temporal extraction corpus and the toxicity corpus. We then discuss the overall performance of our temporal positioning models and their carbon footprint, the performance of toxicity events extraction, and the temporal positioning of chemotherapy toxicity events.

[2] https://bitbucket.org/nicta_biomed/brateval/src/master/
[3] https://www.eea.europa.eu/ims/co2-performance-of-new-passenger

5.5.1 Results

Table 5.2 summarizes the overall results of the baseline model and our temporal positioning model on the temporal extraction test corpus. Our model provides the best results, with an F-measure of 0.86, which is also greater than the inter-annotator agreements. The baseline model gives lower results, 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.

	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 5.2: Overall results on the temporal extraction test corpus.

Table 5.3 presents the detailed performance of our temporal positioning model over all categories on the temporal extraction test corpus.

	Precision	**Recall**	**F-Measure**
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 5.3: Results per category for the temporal positioning model on the temporal extraction test corpus.

Figure 5.4 shows a clinical text sample with predicted results of temporal positioning of homogeneous text portions.

Table 5.4 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 0.55 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 5.4 also provides further performance 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.

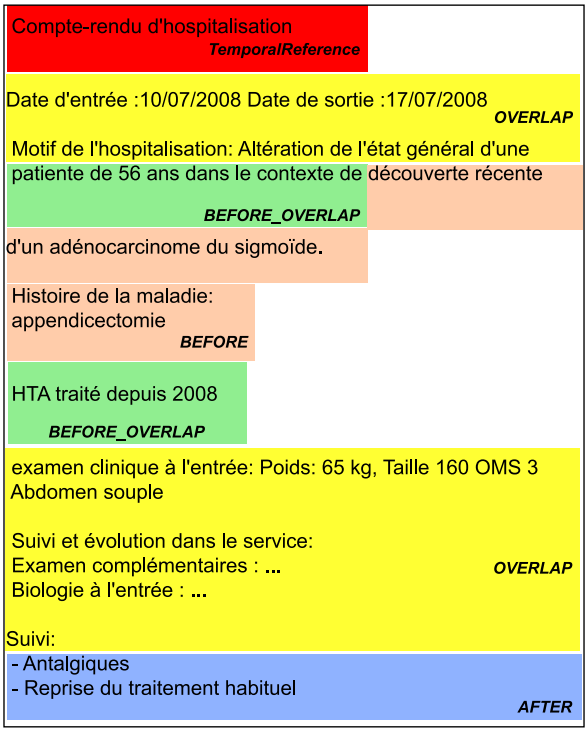


Figure 5.4: An example of predicted temporal positioning of text portions. Translation of text into English: *"Discharge summary. Admission date: 07/10/2008 Discharge date: 07/17/2008. Reason for admission: 56 y.o female presented with asthenia, weight loss and lack of appetite following the recent discovery 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."*

	Precision	**Recall**	**F-Measure**
Extraction of toxicity events	0.43	0.77	0.55
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 5.4: Performance of extraction of toxicity events, event-independent temporal positioning of narrative portions, and temporal positioning toxicity events on the toxicity corpus.

5.5.2 Performance of temporal positioning models

As reported in Table 5.2, 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 5.3 presents the results per category of our model. The most prevalent categories are the best predicted. 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 around 23% of the training corpus. However, high F-measures are also reported for less represented categories such as *TemporalReference* (10.3% of 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 an assigned *Before* category including a temporal indication that shows consistency in time, such as "depuis le" (*since the*) (cf. Figure 5.4). This temporal shift is not always predicted, and despite the coverage of the *Before_ Overlap* category (19.4% in the training corpus), the performance is lower (.77 of F-measure). Except for the second 'Follow-up' text span in Figure 5.4, 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" (*Operative 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 without any pre-defined structured sections. 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 5.4. 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 5.4 (an F-measure of 0.81 vs. an F-measure of 0.79). This good performance shows that our model can adapt to other corpora, including other types of cancer.

5.5.3 Carbon footprint

The carbon footprint of our event-independent temporal positioning model is reported in Table 5.2 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. Note that Carbontracker 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, as already discussed in Chapter 3, Carbontracker does not take into consideration the execution environment or the technique of energy production or other factors such as life cycle impact of hardware. Thus, the obtained carbon footprint measures remain very approximation.

5.5.4 Performance of toxicity events extraction

As reported in Table 5.4, an F-measure of 0.55 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 an 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, checking for toxicity types is part of patient assessment protocols, and extracting such negated events allows for a more thorough clinical evaluation and is vital for enhancing the quality of the care process.

5.5.5 Temporal positioning of chemotherapy toxicity events

This experiment aims to determine how effectively our temporal positioning model 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 are, therefore, 'perfectly' recognized. In our first trials, we evaluated a small selection of 5 documents belonging to both temporal and toxicity corpora (Bannour et al., 2023b). A good performance with an F-measure of 0.7 is observed, and an average of 10 toxicity events per document were temporally positioned. Following these findings and as reported in Table 5.4, an F-measure of 0.62 is obtained on the toxicity test corpus. 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.

5.6 Challenges in building patient timelines

Unstructured text in Electronic Health Records contains significant temporal information. There has been a major interest in constructing patient timelines by the temporal analysis of clinical narratives in order to understand better the clinical history, encompassing disease progress and the quality of healthcare provided, such as the long-term effects of medications. This entails not just analyzing individual clinical notes but also integrating collected information from multiple sources. Indeed, patient information can be dispersed among many clinical notes and structured data, which may not be completely up to date with the most recent patient information, generating further challenges. In our work, we studied single document temporal analysis, which is still difficult. Even in English, the integration of temporal information has seldom been attempted (Raghavan et al., 2014).

Creating patient timelines requires extracting temporal relations between mentions. While extracting DocTimeRel relations between events and DCT offers a coarse-grained temporal ordering, the additional extraction of TLINKs provides more fine-grained timelines. In this section, we study the difficulties encountered while building patient timelines.

The unstructured and complex nature of the clinical text is the first hurdle. Indeed, clinical text is often ungrammatical and includes a wide range of temporal expressions, such as task-specific, non-standard, and abbreviated expressions. Moreover, the clinical text shifts back and forth in time, describing, in some cases, clinical events occurring at different times in the same sentence and including redundant information due to copy-pasted text portions from previous clinical documents. Clinical text also includes a variety of narrative portions that do not have to be pre-defined structured sections with a temporal anchor point.

As discussed in the previous sections, major challenges arise when tackling the temporal relation extraction task, such as the difficulties in defining clinical events, which is task-dependent and makes it difficult to generalize across domains, and the difficulties of extracting TLINKs inter- and intra-sentences. In Section 5.2, we introduced our novel event-independent representation of temporal relations, which makes the TRE task faster and easier since it does not rely on the task-dependent and challenging definition of events. Using the traditional temporal representation, extracting temporal relations requires the prior accurate extraction of events and temporal expressions. However, our temporal representation could provide a coarse level of temporal ordering without requiring prior extraction of mentions.

To get a fine-grained patient timeline, a potential idea is to leverage the

structured information stored in clinical databases. For instance, while annotating some of our clinical documents, we discovered incoherence between the mentioned DCT at the beginning of the document and the time expressions in the narrative portions. As humans and for annotation purposes, we assumed that some of these time expressions were incorrectly represented and attempted to create the most consistent storyline for annotation. However, in some clinical texts, copy-paste portions are written instead of the actual real DCT. To gain a better understanding of temporal ordering, we looked for chemotherapy hospitalization dates in structured admission data. This strategy could help address the incoherence that might occur in narrative text. However, mistakes and incorrectly indicated information are still possible. As a result, merging extracted unstructured and structured information is also a hard task.

Furthermore, in order to generate a patient timeline, it is necessary to incorporate extracted information from several clinical documents concerning the same patient, which increases the risk of having inconsistent information and brings more challenges, such as how to represent the cross-document temporal information. In summary, constructing patient timelines for practical applications remains very difficult and challenging.

5.7 Conclusion

In this chapter, we introduced a novel event-independent representation of temporal relations, which is task- and domain-independent. Using this representation, the temporal relation classification is cast as a sequence token classification task. The main goal of our work is to identify homogeneous narrative portions from a temporal standpoint and classify them into temporal categories reflecting their relations with the DCT. Our representation makes the temporal relation extraction task easier and faster for human annotators. Indeed, based on the THYME-TimeML annotation scheme, we propose a simpler annotation scheme that provides better inter-annotator agreement than the previously reported for the TRE task. Using this scheme, we annotated two corpora of clinical reports written in French. 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. Developing end-to-end systems for temporal information extraction usually yields low results, particularly when evaluated with real data. Indeed, the performance of events and temporal expressions extraction has a high impact on the quality of temporal relation extraction. The TRE task is addressed separately from mentions extraction using our representation, making it more reproducible through dif-

ferent event types. This problem modeling might be the initial step toward constructing a patient timeline to order all its medical events.

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 hospitals or data warehouse clinical reports with various structures and other extraction tasks with different event definitions. Furthermore, since our representation allows the extraction of only coarse-level temporal information, additional approaches should be proposed for extracting fine-grained temporal relations, such as using structured information in clinical data warehouses.

Chapter 6

Conclusions and perspectives

This chapter summarizes our several contributions to information extraction that were mainly applied to clinical narratives in Section 6.1 and discusses further research directions and perspectives in Section 6.2.

6.1 Summary of contributions

In this dissertation, we have proposed novel representations and architectures to address clinical information extraction from unstructured clinical narratives. In Chapter 2, we reviewed relevant background on Named Entity Recognition and Temporal Information Extraction and, in particular, Temporal Relation Extraction. We also covered the main privacy concerns and attacks on personal sensitive text and possible privacy-preservation methods. Moreover, we introduced several aspects of NLP environmental impact, including the main sources of carbon emissions, the main tools for measuring the carbon footprint of statistical models, as well as some research efforts encouraging the conduct of efficient and green NLP experiments.

Following that, in Chapter 3, we examined existing tools for measuring the carbon footprint of statistical models, essentially deep learning models, in order to gain a better understanding of the environmental impact of these models, which are widely used in modern NLP. These tools were selected and evaluated based on specific defined criteria, such as availability, documentation, and technical aspects. We then assessed their use for evaluating the impact of NER experiments, using two different NER models and two computational set-ups. According to our findings, several tools underestimate carbon footprint, and there are a lot of differences in measurements due to different setups, making it difficult to assess and propose an effective tool for NLP approaches. However, we selected one of the discussed tools to report the carbon footprint of all our thesis experiments because it is simple to use and integrate with our equip-

ment. We believe that calculating the carbon emissions of developed models is the first step toward raising awareness and moving toward the development of more efficient models with fewer carbon emissions.

In Chapter 4, we focused on creating shareable NER models that could preserve privacy. We investigated using *mimic learning* to generate shareable student models. The main goal of *mimic learning* is to first train a private *teacher model* on the sensitive data, then use this model to create silver annotations on publicly available corpora, and finally train a *student model* using these produced silver annotations. The generated *student models* could be shared without exposing the *private model* or the sensitive data it was trained on. Our privacy-preserving mimic models architecture allowed us to leverage both public and private corpora in a low-resource setting. Indeed, there are only a few publicly available annotated clinical corpora in French. Using clinical narratives in French, we generated several shareable models and achieved a good compromise between performance and privacy preservation. It should be noted that the used NER model addresses flat and nested entities. For a more practical evaluation, our best privacy-preserving model was shared with a French hospital institution and performed well in extracting lesions, anatomical entities, and the location of target lesions from semi-structured radiology documents.

In Chapter 5, we addressed the temporal relation extraction task from a different perspective. Temporal relation extraction involves identifying and extracting relations between events and/or temporal expressions. The definition of events is highly task- and domain-dependent, making cross-domain generalization challenging. Furthermore, since the performance of TRE systems is closely tied to the performance of mentions extraction, developing end-to-end temporal information extraction systems that could be used in practical applications is difficult. To this end, we propose a novel event-independent representation of temporal relations, providing a coarse level of temporal ordering without requiring prior mentions extraction. Using our novel representation, homogeneous text portions from a temporal standpoint are identified and temporarily positioned according to the Document Creation Time. This results in a much faster and easier task for human annotators through a simpler annotation scheme, as well as more reproducible through different event types. The TRE task was cast as a sequence token classification problem. Although we may lose expressiveness in our proposed annotation scheme, we argue that this does not preclude useful applications on clinical reports. To evaluate our temporal positioning models, we created and annotated two French clinical corpora and demonstrated that good results could be obtained for temporal positioning text portions, as well as chemotherapy toxicity events. Finally, we enumerated the difficulties we have encountered when working with clinical

text and the remaining challenges that arise for constructing patient timelines.

Throughout this thesis, we also provide useful resources for further research. For instance, the silver annotations we provide in Chapter 4 would benefit clinical research as well as research on privacy attacks. We also presented and published the source code of the used NER tool, addressing both flat and nested entities (Wajsbürt et al., 2020). Moreover, we made available our best privacy-preserving model for practical use, and integration into the medkit[1] Python library is under consideration. Finally, we made available our annotation guidelines to annotate corpora with temporal relations using our novel representation.

6.2 Perspectives

Following the promising results of our thesis work, several research areas arise. Working with real-world data made us realize how complex the clinical text is. In fact, the clinical text is often ungrammatical, with spelling errors and redundant information owing to copy-pasted text segments. As a result, digitizing clinical notes may generate errors that are difficult for automated systems to handle. To guarantee high-performance extraction systems, more effective tools are required to manage the maximum amount of errors in preprocessing steps. In addition, clinical narratives shift back and forth in time, describing, in some cases, clinical events occurring at different times in the same sentence. It also contains a lot of domain-specific vocabulary, including task-specific, non-standard, and abbreviated temporal expressions, requiring medical knowledge to understand. Hence, there is a need for more collaborations between hospital institutions and the NLP community. Dealing with clinical narratives written in French added further difficulties. In contrast to English, few resources and processing tools are available for the French language, particularly in the biomedical domain, which limits the clinical NLP research, in particular, due to sensitive data sharing restrictions.

Although efficient NER models could be obtained that efficiently address nested entities, there are still some challenges. To achieve good performance, these NER models require sufficient annotated data. However, the annotation process is time-consuming and involves several phases, beginning with drafting annotation guidelines and then working on annotating documents. These difficulties are heightened when working with specialized domains. For instance, getting domain experts to annotate clinical documents is challenging owing to their professional commitments, yet it is necessary to obtain high-quality annotations. Moreover, as previously stated, disclosing sensitive data is re-

[1] https://team.inria.fr/heka/fr/

stricted. In Chapter 4, we presented a privacy-preservation architecture that could enable sharing models. But, this requires a minimum amount of publicly available data. Therefore, possible perspectives could be to propose methods that reduce the annotation efforts, such as active learning, which aims to annotate only relevant instances for training (Naguib et al., 2023; Le et al., 2023). To deal with all these privacy issues, there is also a growing interest in creating synthetic corpora, particularly for the biomedical domain (Hiebel et al., 2023; Venugopal et al., 2022). Another avenue of improvement is to develop metrics that can detect whether or not trained models leak personal information or how well models protect data privacy.

Developing end-to-end systems for timeline construction, in particular clinical timelines, is far from being a solved task. It requires extracting temporal relations between mentions. However, as discussed in Chapter 5, major challenges are encountered when addressing the temporal relation extraction task. Indeed, TRE entails accurately defining and extracting event and temporal mentions, which is highly dependent on the task and the application domain. Additional difficulties occur when attempting to address inter- and intra-sentence temporal relations. Since extracting fine-grained temporal relations is not a trivial task, getting a coarse-level timeline may be more beneficial for practical applications. Hence, we tried simplifying the TRE task in Chapter 5 by proposing a novel event-independent representation of temporal relations. Further experiments are required to evaluate our representation's generalizability, particularly on other domains and other extraction tasks with different event definitions. Most research efforts focus on within-document timeline extraction. Nevertheless, to offer a temporal analysis of a specific EHR and incorporate information from many clinical notes, cross-document temporal information must be included. More efforts are needed to create annotated corpora with cross-document temporal information. Furthermore, information obtained via structured data should be incorporated with information retrieved from unstructured clinical narratives, whether to resolve ambiguities or to introduce new knowledge that could benefit clinical timelines.

Recently, large language models have emerged, promising good results on various downstream NLP tasks. ChatGPT[2] is now one of the most popular LLMs due to its excellent capacity for interpreting and producing human-like answers. Some studies evaluated the performance of ChatGPT on Information Extraction tasks (Gao et al., 2023; Han et al., 2023). These studies conclude that ChatGPT performs well with simple tasks but struggles with more complicated tasks such as NER, Event extraction, and RE, as evidenced by a major performance gap compared to SOTA approaches. Similar studies have also been conducted for the biomedical domain and have reached the same conclu-

[2] https://openai.com/blog/chatgpt

sion concerning this model (Hu et al., 2023; Chen et al., 2023). This may be due to their lack of domain-specific knowledge. However, a possible research direction could be to use LLMs to solve target tasks without using a pipeline with several sub-tasks. For example, LLMs may be able to handle the relation extraction problem without first solving the NER task. Nevertheless, ChatGPT may run into privacy issues owing to the fact that this model involves transferring patient data to external hosting platforms (Liu et al., 2023).

Aside from biases and ethical concerns, another major drawback of adopting LLMs and deep learning-based models, in general, is their high computational cost. As reviewed in Chapter 3, it is important to quantify the carbon footprint of trained models in order to design more efficient models with low carbon emissions. However, further studies are required to better understand the environmental impact of NLP models by presenting efficient measurement tools and standards to conduct Green NLP research.

Finally, while we focused on clinical texts in our thesis work, we think all of our proposed representations and methods could be adapted for other domains.

Appendix A

Privacy-Preserving Mimic Models for Named Entity Recognition

To be able to compare our privacy-preservation mimic models with our baseline models, described in 4.4.3, we perform an alignment step between entity types. Table A.1 represents this alignment step across the two used French clinical corpora MERLOT and DEFT.

Common category	DEFT entity type	MERLOT entity type
ANAT	anatomie	Anatomy
CHEM	substance	Chemicals_Drugs
DISO	pathologie signe ou symptôme	Disorder SignOrSymptom
LIVB	Living Beings genre	LivingBeings Persons
PROC	traitement examen	MedicalProcedure
TEMP	date durée fréquence moment age	Temporal
DOSE	dose -	Dosage Strength
MODE	mode -	AdministrationRoute DrugForm
MEAS	valeur	Measurement

Table A.1: Alignment between entity types across French clinical corpora; alignments are not always one-to-one.

Appendix B

Temporal Annotation scheme for our clinical corpus

B.0.1 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.

B.0.1.1 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.

B.0.1.2 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 B.1).

B.0.1.3 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 intercure
- 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 B.2).

B.0.1.4 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 B.1).
- Je vois ce jour, Je revois en consultation

B.0.1.5 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

B.0.2 Other categories

B.0.2.1 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

B.0.2.2 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.

B.0.3 Examples of annotations made in accordance with the above scheme and guidelines

Annotations of the first example (cf. Figure B.1)

- 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

Annotations of the second example (cf. Figure B.2)

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

Annotations of the third example (cf. Figure B.3)

- From *Paris* to *2014,* as TemporalReference

Figure B.1: A first example of hospital report annotations

- 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

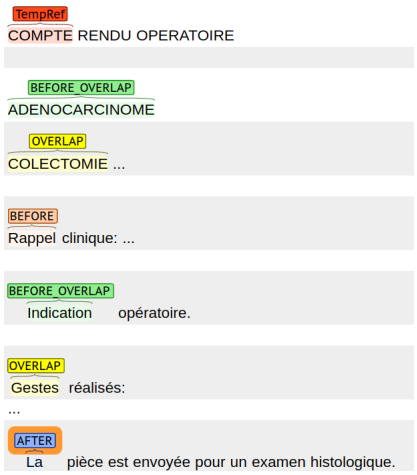


Figure B.2: A second example of annotating an operative report

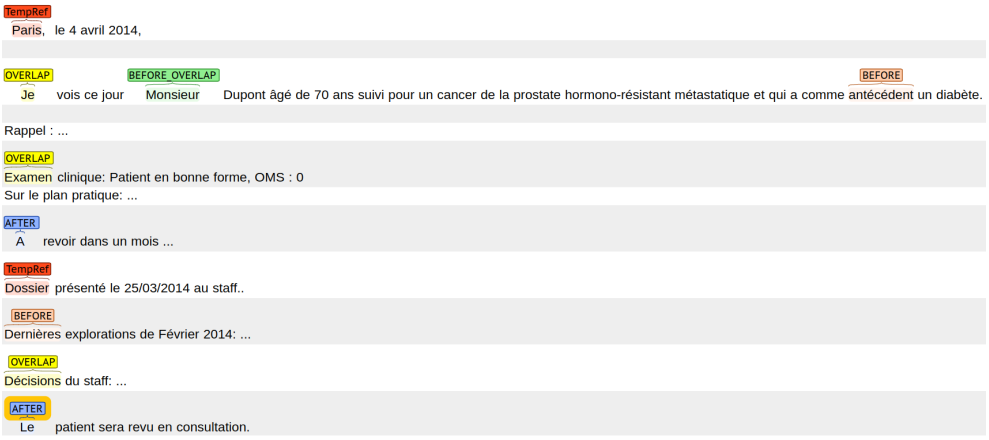


Figure B.3: A third example of annotating a clinical document containing two clinical reports

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. (**cited on** p. 89)

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. (**cited on** p. 47)

Salah Aït-Mokhtar, J-P Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144. (**cited on** p. 58)

Mohammed Alawad, Hong-Jun Yoon, Shang Gao, Brent Mumphrey, Xiao-Cheng Wu, Eric B Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Linda Coyle, et al. 2020. Privacy-preserving deep learning nlp models for cancer registries. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1219–1230. (**cited on** p. 67)

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 42)

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2021. Attention-based bidirectional long short-term memory networks for extracting temporal relation-

ships from clinical discharge summaries. *Journal of Biomedical Informatics*, 123:103915. (**cited on** p. 62)

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843. (**cited on** p. 52)

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. In *ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems"*. (**cited on** p. 69, 78, 79)

Emilia Apostolova, David S Channin, Dina Demner-Fushman, Jacob Furst, Steven Lytinen, and Daniela Raicu. 2009. Automatic segmentation of clinical texts. In *2009 annual international conference of the IEEE engineering in medicine and biology society*, pages 5905–5908. IEEE. (**cited on** p. 63)

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236. (**cited on** p. 39)

Masayuki Asahara, Sachi Kato, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. 2014. BCCWJ-TimeBank: Temporal and event information annotation on Japanese text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 3, September 2014*. (**cited on** p. 61)

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics. (**cited on** p. 29, 73)

Nesrine Bannour, Bastien Rance, Xavier Tannier, and Aurelie Neveol. 2023a. Event-independent temporal positioning: application to French clinical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 191–205, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 29, 110)

Nesrine Bannour, Xavier Tannier, Bastien Rance, and Aurélie Névéol. 2023b. Positionnement temporel indépendant des évènements : application à des textes cliniques en français. In *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 1–14, Paris, France. ATALA. (**cited on** p. 29, 110, 123)

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. 2022a. Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français. In *Journée d'étude sur la robustesse des systemes de TAL*, Paris, France. ATALA. (**cited on** p. 29, 85)

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. 2022b. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130:104073. (**cited on** p. 29, 85)

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. (**cited on** p. 52)

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh, and Fatemehsadat Mireshghallah. 2021. Benchmarking differential privacy and federated learning for bert models. *arXiv preprint arXiv:2106.13973*. (**cited on** p. 66)

Leonard E Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563. (**cited on** p. 40)

Mohamed Baza, Andrew Salazar, Mohamed Mahmoud, Mohamed Abdallah, and Kemal Akkaya. 2020. On sharing models instead of data using mimic learning

for smart health applications. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 231–236. IEEE. (**cited on** p. 67, 87, 103, 104)

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623. (**cited on** p. 24, 25, 45, 64, 69)

Scott W. Bennett and Chinatsu Aone. 1997. Learning to tag multilingual texts through observation. In *Second Conference on Empirical Methods in Natural Language Processing*. (**cited on** p. 41)

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics. (**cited on** p. 66)

Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71. (**cited on** p. 40)

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 35)

Francoise Berthoud, Bruno Bzeznik, Nicolas Gibelin, Myriam Laurens, Cyrille Bonamy, Maxence Morel, and Xavier Schwindenhammer. 2020. Estimation de l'empreinte carbone d'une heure.coeur de calcul. Research report, UGA - Université Grenoble Alpes ; CNRS ; INP Grenoble ; INRIA. (**cited on** p. 68)

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics*

*(*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA. Association for Computational Linguistics. (**cited on** p. 50, 51, 52, 59, 60)

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. (**cited on** p. 50, 57)

Steven Bethard and James H. Martin. 2006. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154, Sydney, Australia. Association for Computational Linguistics. (**cited on** p. 52)

Steven Bethard and James H. Martin. 2007. CU-TMP: Temporal relation classification using syntactic and semantic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 129–132, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 59)

Steven Bethard, James H Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *International Conference on Semantic Computing (ICSC 2007)*, pages 11–18. IEEE. (**cited on** p. 59)

Steven Bethard and Jonathan Parker. 2016. A semantically compositional annotation scheme for time normalization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3779–3786, Portorož, Slovenia. European Language Resources Association (ELRA). (**cited on** p. 50)

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. (**cited on** p. 50, 57)

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. (**cited on** p. 50, 57)

Daniel M. Bikel, Richard M. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231. (**cited on** p. 41)

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French timebank: an iso-timeml annotated reference corpus. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 130–134. (**cited on** p. 56)

William J Black, Fabio Rinaldi, and David Mowatt. 1998. FACILE: Description of the NE system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. (**cited on** p. 39)

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270. (**cited on** p. 52)

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146. (**cited on** p. 34, 90)

Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319. (**cited on** p. 86)

Amine Boulemtafes, Abdelouahid Derhab, and Yacine Challal. 2020. A review of privacy-preserving techniques for deep learning. *Neurocomputing*, 384:21–45. (**cited on** p. 67, 89)

Lucia Bouza Heguerte, Aurélie Bugeau, and Loïc Lannelongue. 2023. HOW TO ESTIMATE CARBON FOOTPRINT WHEN TRAINING DEEP LEARNING MODELS? A GUIDE AND REVIEW. Working paper or preprint. (**cited on** p. 70, 84)

Alissa Brauneck, Louisa Schmalhorst, Mohammad Mahdi Kazemi Majdabadi, Mohammad Bakhtiari, Uwe Völker, Jan Baumbach, Linda Baumbach, and Gabriele Buchholtz. 2023. Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: Scoping review. *Journal of Medical Internet Research*, 25:e41588. (**cited on** p. 67)

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901. (**cited on** p. 44)

Semen Andreevich Budennyy, Vladimir Dmitrievich Lazarev, Nikita Nikolaevich Zakharenko, Aleksei N Korovin, OA Plosskaya, Denis Valer'evich Dimitrov, VS Akhripkin, IV Pavlov, Ivan Valer'evich Oseledets, Ivan Segundovich Barsola, et al. 2023. Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. In *Doklady Mathematics*, volume 106, pages 1–11. Springer. (**cited on** p. 70)

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601. (**cited on** p. 58, 87, 106)

Qingqing Cao, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020a. Towards Accurate and Reliable Energy Measurement of NLP Models. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 141–148, Online. Association for Computational Linguistics. (**cited on** p. 69)

Qingqing Cao, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020b. Towards accurate and reliable energy measurement of NLP models. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 141–148, Online. Association for Computational Linguistics. (**cited on** p. 79)

Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 261–270, Hong Kong, China. Association for Computational Linguistics. (**cited on** p. 47)

Yuwei Cao, William Groves, Tanay Kumar Saha, Joel Tetreault, Alejandro Jaimes, Hao Peng, and Philip Yu. 2022. XLTime: A cross-lingual knowledge transfer framework for temporal expression extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1931–1942, Seattle, United States. Association for Computational Linguistics. (**cited on** p. 50)

Remi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. Association pour le Traitement Automatique des Langues. (**cited on** p. 37, 46, 87)

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6. (**cited on** p. 65, 67, 89)

David S Carrell, David J Cronkite, Muqun Li, Steve Nyemba, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2019. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hid-

ing in plain sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544. (**cited on** p. 65, 95)

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506. (**cited on** p. 56)

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. (**cited on** p. 59)

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 173–176, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 52, 59)

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA). (**cited on** p. 50)

Shan Chang and Chao Li. 2018. Privacy in neural network learning: Threats and countermeasures. *IEEE Network*, 32:61–67. (**cited on** p. 89)

Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. 2020. TopicBERT for energy efficient document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1682–1690, Online. Association for Computational Linguistics. (**cited on** p. 79)

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*. (**cited on** p. 131)

Sanxing Chen, Guoxin Wang, and Börje F. Karlsson. 2019. Exploring word representations on time expression recognition. Technical Report MSR-TR-2019-46, Microsoft Research. (**cited on** p. 50)

Fei Cheng, Masayuki Asahara, Ichiro Kobayashi, and Sadao Kurohashi. 2020. Dynamically updating event representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357, Online. Association for Computational Linguistics. (**cited on** p. 61)

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 60)

Ming Cheng, Liming Li, Yafeng Ren, Yinxia Lou, and Jianbo Gao. 2019. A hybrid method to extract clinical information from chinese electronic medical records. *IEEE Access*, 7:70624–70633. (**cited on** p. 42)

Yuchang Cheng, Masayuki Asahara, and Yuji Matsumoto. 2007. NAIST.Japan: Temporal relation identification using dependency parsed tree. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 245–248, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 59)

Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013. A la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 nlp challenge. *Journal of the American Medical Informatics Association*, 20(5):843–848. (**cited on** p. 59)

Hai Leong Chieu and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 160–163. (**cited on** p. 41, 42)

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. (**cited on** p. 52)

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21. (**cited on** p. 37)

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370. (**cited on** p. 43)

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. (**cited on** p. 43)

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537. (**cited on** p. 43)

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized French language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France. ATALA et AFCP. (**cited on** p. 46)

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297. (**cited on** p. 40)

Francisco Costa and António Branco. 2012. TimeBankPT: A TimeML annotated corpus of Portuguese. In *Proceedings of the Eighth International Conference*

on Language Resources and Evaluation (LREC'12), pages 3727–3734, Istanbul, Turkey. European Language Resources Association (ELRA). (**cited on** p. 56)

Jim Cowie. 1995. CRL/NMSUDescription of the CRL/NMSU systems used for MUC-6. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. (**cited on** p. 41)

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91. (**cited on** p. 32)

Fadl Dahan, Ameur Touir, and Hassan Mathkour. 2015. First order hidden markov model for automatic arabic name entity recognition. *International Journal of Computer Applications*, 123(7). (**cited on** p. 41)

Hong-Jie Dai, Shabbir Syed-Abdul, Chih-Wei Chen, Chieh-Chen Wu, et al. 2015. Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. *BioMed research international*, 2015. (**cited on** p. 63)

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867. (**cited on** p. 47)

Ghaith Dekhili and Fatiha Sadat. 2020. Hybrid statistical and attentive deep neural approach for named entity recognition in historical newspapers. In *CLEF (Working Notes)*. (**cited on** p. 45)

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. (**cited on** p. 118)

Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. 2010. Extracting medication information from french clinical texts. In *MEDINFO 2010*, pages 949–953. IOS Press. (**cited on** p. 39)

Louise Deléger and Aurélie Névéol. 2014. Automatic identification of document sections for designing a French clinical corpus (identification automatique de zones dans des documents pour la constitution d'un corpus médical en français) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 568–573, Marseille, France. Association pour le Traitement Automatique des Langues. (**cited on** p. 63)

Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard III. 2008. Development and evaluation of a clinical note section header terminology. In *AMIA annual symposium proceedings*, volume 2008, page 156. American Medical Informatics Association. (**cited on** p. 63)

Joshua C Denny, Anderson Spickard III, Kevin B Johnson, Neeraja B Peterson, Josh F Peterson, and Randolph A Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815. (**cited on** p. 63)

Leon RA Derczynski. 2017. *Automatically ordering events and times in text*. Springer. (**cited on** p. 49)

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. (**cited on** p. 35, 43, 92)

Eva d'Hondt, Brigitte Grau, and Pierre Zweigenbaum. 2015. Limsi@ clef ehealth 2015-task 2. In *CLEF 2015*. (**cited on** p. 46)

Wentao Ding, Guanji Gao, Linfeng Shi, and Yuzhong Qu. 2019. A pattern-based approach to recognizing time expressions. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. (**cited on** p. 50)

Dmitriy Dligach, Steven Bethard, Timothy Miller, and Guergana Savova. 2022. Exploring text representations for generative temporal relation extraction. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 109–113, Seattle, WA. Association for Computational Linguistics. (**cited on** p. 62)

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, Valencia, Spain. Association for Computational Linguistics. (**cited on** p. 61)

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). (**cited on** p. 51)

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10. (**cited on** p. 37)

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734. (**cited on** p. 44)

Yoann Dupont. 2017. Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique (feature exploration for French named entity recognition with machine learning). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 42–55, Orléans, France. ATALA. (**cited on** p. 45)

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer. (**cited on** p. 66)

Tome Eftimov, Barbara Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE*, 12. (**cited on** p. 39)

Maud Ehrmann. 2008. *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Theses, Paris Diderot University. (**cited on** p. 24)

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915. (**cited on** p. 35)

Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071. (**cited on** p. 65, 66)

Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. 2017. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, 17(1):1–10. (**cited on** p. 23, 86)

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78. (**cited on** p. 39)

Lisa Ferro, Inderjeet Mani, Beth Sundheim, and George Wilson. 2001. Tides temporal annotation guidelines version 1.0. 2. *The MITRE Corporation, McLean-VG-USA*. (**cited on** p. 49)

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE*

International Conference on Data Mining (ICDM), pages 210–219. IEEE. (**cited on** p. 66)

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics. (**cited on** p. 43)

Karën Fort. 2012. *Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus*. Ph.D. thesis, Université Paris-Nord-Paris XIII. (**cited on** p. 113)

Carol Friedman, Philip Alderson, John Austin, James Cimino, and Stephen Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association : JAMIA*, 1:161–74. (**cited on** p. 39)

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. 2020. Clinical concept extraction: A methodology review. *Journal of Biomedical Informatics*, 109:103526. (**cited on** p. 86)

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38. (**cited on** p. 35)

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. (**cited on** p. 58)

Lucian Galescu and Nate Blaylock. 2012. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 715–720. (**cited on** p. 52)

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. 2010. Named and

specific entity detection in varied data: The quæro named entity baseline evaluation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). (**cited on** p. 37)

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2012. Extended named entities annotation on OCRed documents: From corpus constitution to evaluation campaign. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3126–3131, Istanbul, Turkey. European Language Resources Association (ELRA). (**cited on** p. 45)

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. (**cited on** p. 61)

Kavita Ganesan and Michael Subotin. 2014. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40. IEEE. (**cited on** p. 63)

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*. (**cited on** p. 130)

Shang Gao, Olivera Kotevska, Alexandre Sorokine, and J Blair Christian. 2021. A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, 16(2):e0246310. (**cited on** p. 47)

Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. 2013. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137. (**cited on** p. 39)

Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. Fedner: Privacy-preserving medical named entity recognition with federated learning. *arXiv preprint arXiv:2003.09288*. (**cited on** p. 67, 103, 104)

Oguzhan Gencoglu. 2020. Large-scale, language-agnostic discourse classification of tweets during covid-19. *Machine Learning and Knowledge Extraction*, 2(4):603–616. (**cited on** p. 79)

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics. (**cited on** p. 37, 87)

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. (**cited on** p. 37)

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. (**cited on** p. 36)

Cyril Grouin. 2013. *Anonymisation de documents cliniques: performances et limites des méthodes symboliques et par apprentissage statistique*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI. (**cited on** p. 65)

Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. 2013. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association*, 20(5):820–827. (**cited on** p. 59)

Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. 2015. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39, Lisbon, Portugal. Association for Computational Linguistics. (**cited on** p. 66)

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus development. *Journal of Biomedical Informatics*, 50:151–161. Special Issue on Informatics Methods in Medical Privacy. (**cited on** p. 65, 66)

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23. (**cited on** p. 45)

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. (**cited on** p. 53, 54, 58)

Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2021. Chasing carbon: The elusive environmental footprint of computing. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA 2021)*. IEEE. (**cited on** p. 69)

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48. (**cited on** p. 44)

Caroline Hagège and Xavier Tannier. 2007. XRCE-T: XIP temporal module for TempEval campaign. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 492–495, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 58)

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*. (**cited on** p. 130)

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics. (**cited on** p. 52, 61)

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250. (**cited on** p. 35)

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):1–9. (**cited on** p. 39)

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162. (**cited on** p. 34)

Mohsen Hassan, Olfa Makkaoui, Adrien Coulet, and Yannick Toussaint. 2015. Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs. In *Proceedings of BioNLP 15*, pages 71–80, Beijing, China. Association for Computational Linguistics. (**cited on** p. 89)

Peter J Haug, Xinzi Wu, Jeffery P Ferraro, Guergana K Savova, Stanley M Huff, and Christopher G Chute. 2014. Developing a section labeler for clinical documents. In *AMIA Annual Symposium Proceedings*, volume 2014, page 636. American Medical Informatics Association. (**cited on** p. 63)

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43. (**cited on** p. 69, 78, 79)

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. (**cited on** p. 84)

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics. (**cited on** p. 68, 130)

Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. 2017. Deep models under the gan: Information leakage from collaborative deep learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. (**cited on** p. 104)

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851. (**cited on** p. 44)

Lydia-Mai Ho-Dac, Ludovic Tanguy, Céline Grauby, Aurore Heu Mby, Justine Malosse, Laura Rivière, Amélie Veltz-Mauclair, and Marine Wauquier. 2016. Litl at clef ehealth2016: recognizing entities in french biomedical documents. In *CLEF eHealth 2016*. (**cited on** p. 46)

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780. (**cited on** p. 43)

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37. (**cited on** p. 65)

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*. (**cited on** p. 131)

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*. (**cited on** p. 45)

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*. (**cited on** p. 43)

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. (**cited on** p. 39)

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. (**cited on** p. 115, 117)

Jingchi Jiang, Yi Guan, and Chao Zhao. 2015. Wi-enre in clef ehealth evaluation lab 2015: Clinical named entity recognition based on crf. In *CLEF (Working Notes)*. (**cited on** p. 46)

Prateek Jindal and Dan Roth. 2013. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics*, 46:S13–S19. (**cited on** p. 50)

Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz, et al. 2021. Hybrid deep learning for medication-related information extraction from clinical texts in french: Medext algorithm development study. *JMIR medical informatics*, 9(3):e17934. (**cited on** p. 46, 86)

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics. (**cited on** p. 34)

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers), pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics. (**cited on** p. 43, 45)

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana. Association for Computational Linguistics. (**cited on** p. 43)

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_-1):i180–i182. (**cited on** p. 36, 37, 92)

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980. (**cited on** p. 117)

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics. (**cited on** p. 66)

Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. 2023. A transformer-based method for zero and few-shot biomedical named entity recognition. *arXiv e-prints*, pages arXiv–2305. (**cited on** p. 47)

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1121–1128, Sydney, Australia. Association for Computational Linguistics. (**cited on** p. 42)

George R. Krupka and Kevin Hausman. 1998. IsoQuest inc.: Description of the NetOwl^{TM} extractor system as used for MUC-7. In *Seventh Message Under-*

standing Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998. (**cited on** p. 39)

Grey Kuling, Belinda Curpen, and Anne L Martel. 2022. Bi-rads bert and using section segmentation to understand radiology reports. *Journal of Imaging*, 8(5):131. (**cited on** p. 63)

Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. 2010. JU_CSE_TEMP: A first step towards evaluating events, time expressions and temporal relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 345–350, Uppsala, Sweden. Association for Computational Linguistics. (**cited on** p. 52)

Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. 2012. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82. (**cited on** p. 66)

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pretrained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 35, 102)

Kai Labusch and Clemens Neudecker. 2020. Named entity disambiguation and linking historic newspaper ocr with bert. In *CLEF (Working Notes)*. (**cited on** p. 45)

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. (**cited on** p. 70, 78, 79)

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*. (**cited on** p. 41, 90)

Mathieu Lafourcade and Le Brun Nathalie. 2020. Game design evaluation of GWAPs for collecting word associations. In *Workshop on Games and Natural Language Processing*, pages 26–33, Marseille, France. European Language Resources Association. (**cited on** p. 88)

Imad Lakim, Ebtesam Almazrouei, Ibrahim Abualhaol, Merouane Debbah, and Julien Launay. 2022. A holistic assessment of the carbon footprint of noor, a very large Arabic language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 84–94, virtual+Dublin. Association for Computational Linguistics. (**cited on** p. 84)

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics. (**cited on** p. 43, 44)

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. (**cited on** p. 35, 43)

Mostafa Langarizadeh, Azam Orooji, Abbas Sheikhtaheri, and D Hayn. 2018. Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. *eHealth*, 248:80–87. (**cited on** p. 66)

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 103–109, Online. Association for Computational Linguistics. (**cited on** p. 50)

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms:

Quantifying the carbon footprint of computation. *Advanced Science*, page 2100707. (**cited on** p. 70, 77, 79)

Natsuda Laokulrat, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. UTTime: Temporal relation classification using deep syntactic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 88–92, Atlanta, Georgia, USA. Association for Computational Linguistics. (**cited on** p. 59)

Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics*, 6:343–356. (**cited on** p. 50)

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2022. Which anonymization technique is best for which nlp task?–it depends. a systematic study on clinical text processing. *arXiv preprint arXiv:2209.00262*. (**cited on** p. 65)

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association. (**cited on** p. 35)

Linh Le, Gianluca Demartini, Guido Zuccon, Genghong Zhao, and Xia Zhang. 2023. Active learning with feature matching for clinical named entity recognition. *Natural Language Processing Journal*, page 100015. (**cited on** p. 47, 130)

Tiphaine Le Clercq de Lannoy, Romaric Besançon, Olivier Ferret, Julien Tourille, Frédérique Brin-Henry, and Bianca Vieru. 2022. Stratégies d'adaptation pour la reconnaissance d'entités médicales en français (adaptation strategies for biomedical named entity recognition in French). In *Actes de la 29e Conférence sur le*

Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, pages 215–225, Avignon, France. ATALA. (**cited on** p. 46)

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022a. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2700, Dublin, Ireland. Association for Computational Linguistics. (**cited on** p. 47)

Garam Lee, Minsoo Kim, Jai Hyun Park, Seung-won Hwang, and Jung Hee Cheon. 2022b. Privacy-preserving text classification on BERT embeddings with homomorphic encryption. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3169–3175, Seattle, United States. Association for Computational Linguistics. (**cited on** p. 67)

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, California. Association for Computational Linguistics. (**cited on** p. 52, 60)

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. (**cited on** p. 45, 92)

Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Baltimore, Maryland. Association for Computational Linguistics. (**cited on** p. 50)

Artuur Leeuwenberg and Marie-Francine Moens. 2017a. KULeuven-LIIR at SemEval-2017 task 12: Cross-domain temporal information extraction from clinical records. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1030–1034, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 61)

Artuur Leeuwenberg and Marie-Francine Moens. 2017b. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1150–1158, Valencia, Spain. Association for Computational Linguistics. (**cited on** p. 61)

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics. (**cited on** p. 65)

Thomas Lemaitre, Camille Gosset, Mathieu Lafourcade, Namrata Patel, and Guilhem Mayoral. 2020. DEFT 2020 - extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance (fine-grained information extraction in clinical data : Dedicated terminologies and knowledge graphs). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 55–65, Nancy, France. ATALA et AFCP. (**cited on** p. 46)

Thomas Lemaître, Camille Gosset, Mathieu Lafourcade, Namrata Patel, and Guilhem Mayoral. 2020. Deft 2020 - extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance (fine-grained information extraction in clinical data : Dedicated terminologies and knowledge graphs). In *JEPTALNRECITAL*. (**cited on** p. 88)

Ivan Lerner, Nicolas Paris, and Xavier Tannier. 2020. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of biomedical informatics*, 102:103356. (**cited on** p. 46)

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. (**cited on** p. 35, 44)

Jiao Li, Yueping Sun, R Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2015. Annotating chemicals, diseases, and their interactions in biomedical literature. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 173–182. The Fifth BioCreative Organizing Committee. (**cited on** p. 37)

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70. (**cited on** p. 86)

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. (**cited on** p. 52, 61)

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics. (**cited on** p. 44)

Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2005. Svm based learning system for information extraction. In *Deterministic and Statistical Methods*

in Machine Learning: First International Workshop, Sheffield, UK, September 7-10, 2004. Revised Lectures, pages 319–339. Springer. (**cited on** p. 41)

Yaoyong Li and John Shawe-Taylor. 2003. The SVM with uneven margins and Chinese document categorization. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 216–227, Sentosa, Singapore. COLIPS PUBLICATIONS. (**cited on** p. 42)

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM international health informatics symposium*, pages 744–750. (**cited on** p. 63)

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1054–1064. (**cited on** p. 47)

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65. (**cited on** p. 47)

Anne-Laure Ligozat, Aurélie Névéol, Bénédicte Daly, and Emmanuelle Frenoux. 2020. Ten simple rules to make your research more sustainable. *PLoS Computational Biology*, 16(9):e1008148. (**cited on** p. 84)

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. (**cited on** p. 60)

Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the Ninth Interna-*

tional Workshop on Health Text Mining and Information Analysis, pages 165–176, Brussels, Belgium. Association for Computational Linguistics. (**cited on** p. 62)

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP 2017*, pages 322–327, Vancouver, Canada,. Association for Computational Linguistics. (**cited on** p. 51)

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. (**cited on** p. 62)

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. (**cited on** p. 62)

Yu-Kai Lin, Hsinchun Chen, and Randall A. Brown. 2013. Medtime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46 Suppl:S20–8. (**cited on** p. 51)

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods of information in medicine*, 32(4):281–291. (**cited on** p. 88)

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics. (**cited on** p. 47)

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does roberta know and when? *CoRR*, abs/2104.07885. (**cited on** p. 79)

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. Ltp: a new active learning strategy for crf-based named entity recognition. *Neural Processing Letters*, 54(3):2433–2454. (**cited on** p. 47)

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA. Association for Computational Linguistics. (**cited on** p. 47)

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017. (**cited on** p. 131)

Yijin Liu, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019a. GCDT: A global context enhanced deep transition architecture for sequence labeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2431–2441, Florence, Italy. Association for Computational Linguistics. (**cited on** p. 43)

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. (**cited on** p. 43)

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden. Association for Computational Linguistics. (**cited on** p. 52, 59)

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. In *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*. (**cited on** p. 69, 78, 79)

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867. (**cited on** p. 42)

Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476*. (**cited on** p. 69, 70)

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*. (**cited on** p. 69, 84)

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020. Hierarchical contextualized representation for named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*. (**cited on** p. 43)

Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*. (**cited on** p. 67)

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics. (**cited on** p. 15, 43, 80, 81, 82)

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 52, 60)

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics. (**cited on** p. 59)

Inderjeet Mani and D. George Wilson. 2000. Robust temporal processing of news. In *Annual Meeting of the Association for Computational Linguistics*. (**cited on** p. 50)

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT Press. (**cited on** p. 33)

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. (**cited on** p. 35, 90, 117)

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191. (**cited on** p. 42)

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR. (**cited on** p. 67, 104)

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. (**cited on** p. 104)

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pages 887–896, Copenhagen, Denmark. Association for Computational Linguistics. (**cited on** p. 52, 60)

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16. (**cited on** p. 65, 66)

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway. Association for Computational Linguistics. (**cited on** p. 39)

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. (**cited on** p. 34)

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26. (**cited on** p. 34)

Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 18–26, Sofia, Bulgaria. Association for Computational Linguistics. (**cited on** p. 59)

Timothy Miller, Steven Bethard, Dmitriy Dligach, and Guergana Savova. 2023. End-to-end clinical temporal information extraction with multi-head attention. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 313–319, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 62)

Anne-Lyse Minard, Andréane Roques, Nicolas Hiot, Mirian Halfeld Ferrari Alves, and Agata Savary. 2020. DOING@DEFT : cascade de CRF pour l'annotation

d'entités cliniques imbriquées (DOING@DEFT : cascade of CRF for the annotation of nested clinical entities). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 66–78, Nancy, France. ATALA et AFCP. (**cited on** p. 46)

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. (**cited on** p. 65, 67, 68)

Clément Morand. 2023. Evaluation of the environmental impacts of natural language processing methods. (**cited on** p. 84)

Véronique Moriceau and Xavier Tannier. 2014. French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3239–3243, Reykjavik, Iceland. European Language Resources Association (ELRA). (**cited on** p. 50)

Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1. (**cited on** p. 41)

Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2608–2618, Copenhagen, Denmark. Association for Computational Linguistics. (**cited on** p. 42)

Marco Naguib, Aurélie Névéol, and Xavier Tannier. 2023. Stratégies d'apprentissage actif pour la reconnaissance d'entités nommées en français. In

18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 232–247. ATALA. (**cited on** p. 47, 130)

Rakshit Naidu, Harshita Diddee, Ajinkya Mulay, Aleti Vardhan, Krithika Ramesh, and Ahmed Zamzam. 2021. Towards quantifying the carbon emissions of differentially private machine learning. *arXiv preprint arXiv:2107.06946*. (**cited on** p. 69)

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. (**cited on** p. 58, 64)

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. (**cited on** p. 45)

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018a. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13. (**cited on** p. 46, 86)

Aurélie Névéol, Julien Grosjean, Stéfan Darmoni, and Pierre Zweigenbaum. 2014. Language resources for French in the biomedical domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2146–2151, Reykjavik, Iceland. European Language Resources Association (ELRA). (**cited on** p. 46)

Aurélie Névéol, Cyril Grouin, Kevin B Cohen, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier,

and Pierre Zweigenbaum. 2016. Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *Proc of CLEF eHealth Evaluation lab*, Evora, Portugal. (**cited on** p. 38, 46)

Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In *Proc of ShARe/CLEF Evaluation Lab*, Toulouse, France. (**cited on** p. 46)

Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. 2018b. Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF (Working Notes)*. (**cited on** p. 38)

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics. (**cited on** p. 50)

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. 2020. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57. (**cited on** p. 66)

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30. (**cited on** p. 37, 38, 46)

Amy L. Olex and Bridget Mcinnes. 2021. Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *Journal of biomedical informatics*, page 103784. (**cited on** p. 50, 51, 64)

Sarath P R, Manikandan R, and Yoshiki Niwa. 2017. Hitachi at SemEval-2017 task 12: System for temporal information extraction from clinical notes. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1005–1009, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 60)

Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis, and Constantine D Spyropoulos. 2000. Learning decision trees for named-entity recognition and classification. In *ECAI Workshop on Machine Learning for Information Extraction*. (**cited on** p. 41)

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE. (**cited on** p. 65)

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional lstm-crf for named entity recognition. In *Proceedings of the 32nd Pacific Asia conference on language, information and computation*. (**cited on** p. 43)

Titouan Parcollet and Mirco Ravanelli. 2021. The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. Working paper or preprint. (**cited on** p. 79)

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*. (**cited on** p. 34, 36)

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv e-prints*, pages arXiv–2104. (**cited on** p. 69)

Xutan Peng, Guanyi Chen, Chenghua Lin, and Mark Stevenson. 2021. Highly efficient knowledge graph embedding learning with Orthogonal Procrustes Analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2364–2375, Online. Association for Computational Linguistics. (**cited on** p. 79)

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. (**cited on** p. 34)

Rafael Peres, Diego Esteves, and Gaurav Maheshwari. 2017. Bidirectional lstm with a context input window for named entity recognition in tweets. In *Proceedings of the Knowledge Capture Conference*, pages 1–4. (**cited on** p. 45)

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. (**cited on** p. 35, 43)

Uyen Phan and Nhung Nguyen. 2022. Simple semantic-based data augmentation for named entity recognition in biomedical texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 123–129, Dublin, Ireland. Association for Computational Linguistics. (**cited on** p. 47)

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics. (**cited on** p. 79)

Bernardo Pulido-Gaytan, Andrei Tchernykh, Jorge M Cortés-Mendoza, Mikhail Babenko, Gleb Radchenko, Arutyun Avetisyan, and Alexander Yu Drozdov. 2021. Privacy-preserving neural networks with homomorphic encryption: C hallenges and opportunities. *Peer-to-Peer Networking and Applications*, 14(3):1666–1691. (**cited on** p. 67)

Georgiana Puşcaşu. 2007. WVALI: Temporal relation identification by syntactico-semantic analysis. In *Proceedings of the Fourth International Workshop on Se-*

mantic Evaluations (SemEval-2007), pages 484–487, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 59)

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34. (**cited on** p. 50, 51, 54)

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK. (**cited on** p. 56)

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). (**cited on** p. 50, 51, 54, 56)

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. (**cited on** p. 53)

Xinchi Qiu, Titouan Parcollet, Javier Fernandez-Marques, Pedro PB de Gusmao, Yan Gao, Daniel J Beutel, Taner Topal, Akhil Mathur, and Nicholas D Lane. 2023. A first look into the carbon footprint of federated learning. *J. Mach. Learn. Res.*, 24:129–1. (**cited on** p. 69)

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897. (**cited on** p. 35)

Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel

Alberto Garcia Peña, and Cyril Labbé. 2016. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61. (**cited on** p. 39)

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1:81–106. (**cited on** p. 41)

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. (**cited on** p. 35)

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. (**cited on** p. 89)

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551. (**cited on** p. 43)

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 998–1008, Baltimore, Maryland. Association for Computational Linguistics. (**cited on** p. 124)

Anushree Raj and Rio D'Souza. 2021. Anonymization of sensitive data in unstructured documents using nlp. *International Journal of Mechanical Engineering and Technology (IJMET)*, 12(4):25–35. (**cited on** p. 66)

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics. (**cited on** p. 40)

Lisa F Rau. 1991. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society. (**cited on** p. 39)

Kirk Roberts, Bryan Rink, and Sanda M Harabagiu. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, 20(5):867–875. (**cited on** p. 52, 59)

Conceição Rocha, Alípio Mário Jorge, Roberta Akemi Sinoara, Paula Brito, Carlos Pimenta, and Solange Oliveira Rezende. 2016. Pampo: using pattern matching and pos-tagging for effective named entities recognition in portuguese. *ArXiv*, abs/1612.09535. (**cited on** p. 39)

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. (**cited on** p. 117)

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4864–4873, Hong Kong, China. Association for Computational Linguistics. (**cited on** p. 62, 63)

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway. (**cited on** p. 67)

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952. (**cited on** p. 60)

Steven L. Salzberg. 1994. Book review: C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16:235–240. (**cited on** p. 41)

Gabriele Sarti. 2020. Interpreting neural language models for linguistic complexity assessment. Master's thesis, University of Trieste. (**cited on** p. 79)

G. Savova, J. Fan, Zi Ye, Sean P. Murphy, Jiaping Zheng, C. Chute, and I. Kullo. 2010. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2010:722–6. (**cited on** p. 39)

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63. (**cited on** p. 69, 73)

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee. (**cited on** p. 102)

Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. 1998. A decision tree method for finding and classifying names in Japanese texts. In *Sixth Workshop on Very Large Corpora*. (**cited on** p. 41)

Kira A. Selby, Yinong Wang, Ruizhe Wang, Peyman Passban, Ahmad Rashid, Mehdi Rezagholizadeh, and Pascal Poupart. 2021. Robust embeddings via distributions. *CoRR*, abs/2104.08420. (**cited on** p. 79)

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. (**cited on** p. 35)

Khaled Shaalan and Mai Oudah. 2014. A hybrid approach to arabic named entity recognition. *Journal of Information Science*, 40(1):67–87. (**cited on** p. 42)

Omar Shaikh, Jon Saad-Falcon, Austin P Wright, Nilaksh Das, Scott Freitas, Omar Asensio, and Duen Horng Chau. 2021. Energyvis: interactively tracking and exploring energy consumption for ml models. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. (**cited on** p. 75)

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598. (**cited on** p. 67)

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics. (**cited on** p. 39)

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 47)

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics. (**cited on** p. 44)

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. *arXiv preprint arXiv:2305.13298*. (**cited on** p. 44)

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE. (**cited on** p. 67, 68)

Damien Sileo, Camille Pradel, Philippe Muller, and Tim Van de Cruys. 2017. Synapse at cap 2017 ner challenge: Fasttext crf. *arXiv e-prints*, pages arXiv–1709. (**cited on** p. 45)

Gonçalo Simoes, Helena Galhardas, and Luisa Coheur. 2009. Information extraction tasks: a survey. *Simpósio de Informática*, 540:1–550. (**cited on** p. 32)

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR. (**cited on** p. 44)

Sunghwan Sohn, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute, and Hongfang Liu. 2014. MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5):858–865. (**cited on** p. 39)

Sunghwan Sohn, Kavishwar B Wagholikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association*, 20(5):836–842. (**cited on** p. 51)

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics. (**cited on** p. 45)

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4. (**cited on** p. 94, 117)

Lina Fatima Soualmia, Chloé Cabot, Badisse Dahamna, and Stéfan Jacques Darmoni. 2015. Sibm at clef e-health evaluation lab 2015. In *CLEF (Working Notes)*. (**cited on** p. 46)

Samuel Sousa and Roman Kern. 2022. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*. (**cited on** p. 65, 66, 67)

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, pages 181–224. (**cited on** p. 50)

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA). (**cited on** p. 58)

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics. (**cited on** p. 44, 45)

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47:269–298. (**cited on** p. 50)

Jannik Strötgen and Michael Gertz. 2015. A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 541–547, Lisbon, Portugal. Association for Computational Linguistics. (**cited on** p. 50)

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics. (**cited on** p. 25, 69, 75, 104)

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. (**cited on** p. 50, 52, 53, 55, 56, 57)

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813. (**cited on** p. 37, 50, 57)

Beth M. Sundheim. 1993. TIPSTER/MUC-5 information extraction system evaluation. In *TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993*, pages 147–163, Fredericksburg, Virginia, USA. Association for Computational Linguistics. (**cited on** p. 49)

Koichi Takeuchi and Nigel Collier. 2003. Bio-medical entity extraction using support vector machines. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 57–64, Sapporo, Japan. Association for Computational Linguistics. (**cited on** p. 42)

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835. (**cited on** p. 59)

Mike Donald Tapi-Nzali, Xavier Tannier, and Aurélie Névéol. 2015. Automatic extraction of time expressions accross domains in french narratives. In *Conference on Empirical Methods in Natural Language Processing*. (**cited on** p. 51)

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008, Istanbul, Turkey. European Language Resources Association (ELRA). (**cited on** p. 63)

The Shift Project. 2018. Lean ICT: Towards Digital Sobriety. Technical report, The Shift Project. Directed by Hugues Ferreboeuf. (**cited on** p. 78)

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. (**cited on** p. 37)

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*. (**cited on** p. 37)

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. (**cited on** p. 37, 92)

Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, Suntec, Singapore. Association for Computational Linguistics. (**cited on** p. 47)

Kentaro Torisawa et al. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 698–707. (**cited on** p. 42)

Rian Touchent, Laurent Romary, and Eric Villemonte de La Clergerie. 2023. Camembert-bio: Un modèle de langue français savoureux et meilleur pour la santé. (**cited on** p. 35, 102)

Julien Tourille. 2018. *Extracting clinical event timelines: temporal information extraction and coreference resolution in electronic health records*. Ph.D. thesis, Université Paris Saclay (COmUE). (**cited on** p. 51, 54)

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. 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, San Diego, California. Association for Computational Linguistics. (**cited on** p. 60)

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017a. 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, Vancouver, Canada. Association for Computational Linguistics. (**cited on** p. 51, 52, 61)

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017b. 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. (**cited on** p. 25, 60, 64, 110, 115)

Martin Jaggi Tristan Trebaol, Mary-Anne Hartley and Hossein Shokri Ghadikolaei. 2020. A tool to quantify and report the carbon footprint of machine learning computations and communication in academia and healthcare. *Infoscience EPFL: record 278189*. (**cited on** p. 69, 78, 79)

Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. 2021. Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Computers & Security*, 110:102402. (**cited on** p. 67, 104)

Naushad UzZaman and James F Allen. 2010. Event and temporal expression extraction from raw text: First step towards a temporally aware system. *International Journal of Semantic Computing*, 4(04):487–508. (**cited on** p. 50)

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics. (**cited on** p. 50, 56)

Thomas Vakili and Hercules Dalianis. 2021. Are clinical bert models privacy preserving? the difficulty of extracting patient-condition associations. In *HUMAN@ AAAI Fall Symposium*. (**cited on** p. 65)

Thomas Vakili and Hercules Dalianis. 2023. Using membership inference attacks to evaluate privacy-preserving language modeling fails for pseudonymizing data.

In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 318–323, Tórshavn, Faroe Islands. University of Tartu Library. (**cited on** p. 65, 68)

Erik M Van Mulligen, Zubair Afzal, Saber Akhondi, Dang Vo, and Jan Kors. 2016. Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. In *Conference and Labs of the Evaluation Forum*. (**cited on** p. 46, 80, 88)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30. (**cited on** p. 35)

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. (**cited on** p. 51, 60)

Rohit Venugopal, Noman Shafqat, Ishwar Venugopal, Benjamin Mark John Tillbury, Harry Demetrios Stafford, and Aikaterini Bourazeri. 2022. Privacy preserving generative adversarial networks to model electronic health records. *Neural Networks*, 153:339–348. (**cited on** p. 68, 130)

Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. A systematic review of green ai. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1507. (**cited on** p. 84)

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics. (**cited on** p. 50, 56, 63, 110)

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International*

Workshop on Semantic Evaluation, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics. (**cited on** p. 50, 56)

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28. (**cited on** p. 44)

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269. (**cited on** p. 41, 90)

Kavishwar B. Wagholikar, Manabu Torii, Siddhartha R. Jonnalagadda, and Hongfang Liu. 2012. Feasibility of pooling annotated corpora for clinical concept extraction. *AMIA Summits on Translational Science Proceedings*, 2012:38 – 38. (**cited on** p. 86)

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339. (**cited on** p. 43)

Perceval Wajsbürt. 2021. *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université. (**cited on** p. 15, 80, 81, 82, 91)

Perceval Wajsbürt, Yoann Taillé, Guillaume Lainé, and Xavier Tannier. 2020. Participation de l'équipe du LIMICS à DEFT 2020 (participation of team LIMICS in the DEFT 2020 challenge). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 108–117, Nancy, France. ATALA et AFCP. (**cited on** p. 46, 129)

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, pages 1371–1377, Florence, Italy. Association for Computational Linguistics. (**cited on** p. 62)

Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online. Association for Computational Linguistics. (**cited on** p. 44, 45)

Liang Wang, Peifeng Li, and Sheng Xu. 2022a. DCT-centered temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. (**cited on** p. 61)

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*. (**cited on** p. 44)

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*, 62:78–89. (**cited on** p. 58)

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. (**cited on** p. 47)

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49. (**cited on** p. 86)

Yaqiang Wang, Zhonghua Yu, Li Chen, Yunhui Chen, Yiguang Liu, Xiaoguang Hu, and Yongguang Jiang. 2014. Supervised methods for symptom name recognition

in free-text clinical records of traditional chinese medicine: an empirical study. *Journal of biomedical informatics*, 47:91–104. (**cited on** p. 42)

Yefeng Wang and Jon Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 42–49, Borovets, Bulgaria. Association for Computational Linguistics. (**cited on** p. 42)

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022b. Nested named entity recognition: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29. (**cited on** p. 37)

Hao Wei, Mingyuan Gao, Ai Zhou, Fei Chen, Wen Qu, Chunli Wang, and Mingyu Lu. 2019. Named entity recognition from biomedical texts using a fusion attention-based bilstm-crf. *IEEE Access*, 7:73627–73636. (**cited on** p. 45)

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. (**cited on** p. 80, 117)

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813. (**cited on** p. 25, 70)

Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. 2008. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37. (**cited on** p. 40)

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al.

2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*. (**cited on** p. 35)

Yonghui Wu, Xi Yang, Jiang Bian, Yi Guo, Hua Xu, and William Hogan. 2018. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1110. American Medical Informatics Association. (**cited on** p. 44)

Shangyu Xie, Wei Dai, Esha Ghosh, Sambuddha Roy, Dan Schwartz, and Kim Laine. 2023. Does prompt-tuning language model ensure privacy? *arXiv preprint arXiv:2304.03472*. (**cited on** p. 65)

Guohai Xu, Chengyu Wang, and Xiaofeng He. 2018. Improving clinical named entity recognition with global neural attention. In *Web and Big Data: Second International Joint Conference, APWeb-WAIM 2018, Macau, China, July 23-25, 2018, Proceedings, Part II 2*, pages 264–279. Springer. (**cited on** p. 44)

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24. (**cited on** p. 39)

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A survey on green deep learning. *arXiv preprint arXiv:2111.05193*. (**cited on** p. 70)

Yan Xu, Yining Wang, Tianren Liu, Jiahua Liu, Yubo Fan, Yi Qian, Junichi Tsujii, and Eric I Chang. 2014. Joint segmentation and named entity recognition using dual decomposition in chinese discharge summaries. *Journal of the American Medical Informatics Association*, 21(e1):e84–e92. (**cited on** p. 42)

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):849–858. (**cited on** p. 59)

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822. (**cited on** p. 44)

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942. (**cited on** p. 45)

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32. (**cited on** p. 35)

Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Waqas Anwar. 2015. Biomedical named entity recognition based on deep neutral network. *Int. J. Hybrid Inf. Technol*, 8(8):279–288. (**cited on** p. 44)

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 837–844. (**cited on** p. 47)

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, Online. Association for Computational Linguistics. (**cited on** p. 44, 45, 91)

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 61)

Vanni Zavarella and Hristo Tanev. 2013. FSS-TimEx for TempEval-3: Extracting temporal information from text. In *Second Joint Conference on Lexical and*

*Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 58–63, Atlanta, Georgia, USA. Association for Computational Linguistics. (**cited on** p. 51)

Huaizheng Zhang, Yizheng Huang, Yonggang Wen, Jianxiong Yin, and Kyle Guan. 2020. Inferbench: Understanding deep learning inference serving with an automatic benchmarking system. *arXiv preprint arXiv:2011.02327*. (**cited on** p. 75)

Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of biomedical informatics*, 37(6):411–422. (**cited on** p. 42)

Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. 2023. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv e-prints*, pages arXiv–2305. (**cited on** p. 47)

Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International journal of medical informatics*, 132:103985. (**cited on** p. 45)

Shiyi Zhao, Lishuang Li, Hongbin Lu, Anqiao Zhou, and Shuang Qian. 2019. Associative attention networks for temporal relation extraction from electronic health records. *Journal of biomedical informatics*, 99:103309. (**cited on** p. 61)

Zhehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. *BMC medical genomics*, 10:75–83. (**cited on** p. 44)

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 357–366, Hong Kong, China. Association for Computational Linguistics. (**cited on** p. 45)

Weipeng Zhou, Majid Afshar, Dmitriy Dligach, Yanjun Gao, and Timothy Miller. 2023a. Improving the transferability of clinical note section classification models with BERT and large language model ensembles. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 125–130, Toronto, Canada. Association for Computational Linguistics. (**cited on** p. 63)

Weipeng Zhou, Meliha Yetisgen, Majid Afshar, Yanjun Gao, Guergana Savova, and Timothy Miller. 2023b. Improving model transferability for clinical note section classification models using continued pretraining. *medRxiv*, pages 2023–04. (**cited on** p. 63)

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China. (**cited on** p. 35)

Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. 2020. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. *ArXiv*, abs/2009.04872. (**cited on** p. 95, 96)

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Eric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Benoît Thirion, and Stefan Darmoni. 2003. UMLF: a Unified Medical Lexicon for French. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2003:1062. (**cited on** p. 88)

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Extraction d'Informations à partir des Dossiers Patients Informatisés : Etudes en temporalité, confidentialité et impact environnemental

Mots clés: Extraction d'information, Représentation temporelle, Traitement Automatique des Langues cliniques, Confidentialité, Réseaux de neurones, Empreinte carbone

Résumé: L'extraction automatique des informations contenues dans les Dossiers Patients Informatisés (DPIs) est cruciale pour améliorer la recherche clinique. Or, la plupart des informations sont sous forme de texte non structuré. La complexité et le caractère confidentiel du texte clinique présente des défis supplémentaires. Par conséquent, le partage de données est difficile dans la pratique et est strictement encadré par des réglementations. Les modèles neuronaux offrent de bons résultats pour l'extraction d'informations. Mais ils nécessitent de grandes quantités de données annotées, qui sont souvent limitées, en particulier pour les langues autres que l'anglais. Ainsi, la performance n'est pas encore adaptée à des applications pratiques. Outre les enjeux de confidentialité, les modèles d'apprentissage profond ont un important impact environnemental. Dans cette thèse, nous proposons des méthodes et des ressources pour la Reconnaissance d'entités nommées (REN) et l'extraction de relations temporelles dans des textes cliniques en français.

Plus précisément, nous proposons une architecture de modèles préservant la confidentialité des données par mimétisme permettant un transfert de connaissances d'un modèle enseignant entraîné sur un corpus privé à un modèle élève. Ce modèle élève pourrait être partagé sans révéler les données sensibles ou le modèle privé construit avec ces données. Notre stratégie offre un bon compromis entre la performance et la préservation de la confidentialité.

Ensuite, nous introduisons une nouvelle représentation des relations temporelles, indépendante des événements et de la tâche d'extraction, qui permet d'identifier des portions de textes homogènes du point de vue temporel et de caractériser la relation entre chaque portion du texte et la date de création du document. Cela rend l'annotation et l'extraction des relations temporelles plus facile et reproductible à travers différents types d'événements, vu qu'aucune définition et extraction préalable des événements n'est requise.

Enfin, nous effectuons une analyse comparative des outils existants de mesure d'empreinte carbone des modèles de TAL. Nous adoptons un des outils étudiés pour calculer l'empreinte carbone de nos modèles, en considérant que c'est une première étape vers une prise de conscience et un contrôle de leur impact environnemental.

En résumé, nous générons des modèles de REN partageables préservant la confidentialité que les cliniciens peuvent utiliser efficacement. Nous démontrons également que l'extraction de relations temporelles peut être abordée indépendamment du domaine d'application et que de bons résultats peuvent être obtenus en utilisant des données d'oncologie du monde réel.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Information Extraction from Electronic Health Records: Studies on temporal ordering, privacy and environmental impact

Keywords: Information Extraction, Temporal Representation, Clinical Natural Language Processing, Confidentiality, Neural Networks, Carbon Footprint

Abstract: Automatically extracting rich information contained in Electronic Health Records (EHRs) is crucial to improve clinical research. However, most of this information is in the form of unstructured text. The complexity and the sensitive nature of clinical text involve further challenges. As a result, sharing data is difficult in practice and is governed by regulations. Neural-based models showed impressive results for Information Extraction, but they need significant amounts of manually annotated data, which is often limited, particularly for non-English languages. Thus, the performance is still not ideal for practical use. In addition to privacy issues, using deep learning models has a significant environmental impact. In this thesis, we develop methods and resources for clinical Named Entity Recognition (NER) and Temporal Relation Extraction (TRE) in French clinical narratives.

Specifically, we propose a privacy-preserving mimic models architecture by exploring the mimic learning approach to enable knowledge transfer through a teacher model trained on a private corpus to a student model. This student model could be publicly shared without disclosing the original sensitive data or the private teacher model on which it was trained. Our strategy offers a good compromise between performance and data privacy preservation.

Then, we introduce a novel event- and task-independent representation of temporal relations. Our representation enables identifying homogeneous text portions from a temporal standpoint and classifying the relation between each text portion and the document creation time. This makes the annotation and extraction of temporal relations easier and reproducible through different event types, as no prior definition and extraction of events is required.

Finally, we conduct a comparative analysis of existing tools for measuring the carbon emissions of NLP models. We adopt one of the studied tools to calculate the carbon footprint of all our created models during the thesis, as we consider it a first step toward increasing awareness and control of their environmental impact.

To summarize, we generate shareable privacy-preserving NER models that clinicians can efficiently use. We also demonstrate that the TRE task may be tackled independently of the application domain and that good results can be obtained using real-world oncology clinical notes.