Université Sorbonne Université

Ecole Doctorale Informatique, Télécommunications et Electronique - ED130

Inria de Paris / Équipe ALMAnaCH

Thèse de doctorat

Discipline : Informatique

Présentée par

Rian TOUCHENT-SAAD

Pour obtenir le grade universitaire de

Docteur de l'Université Sorbonne Université

Improving Clinical Information Extraction with Public-Data Pretraining

---

Présentée et soutenue publiquement le DATE devant le jury composé de :

Mathieu Constant	ATILF (CNRS / Université de Lorraine)	Rapporteur
Aurélie Névéol	CNRS, LISN, Université Paris-Saclay	Rapportrice
Natalia Grabar	CNRS, Université de Lille	Rapportrice
Pierre Zweigenbaum	CNRS, LISN, Université Paris-Saclay	Examinateur
Thierry Charnois	LIPN, Université Sorbonne Paris Nord	Examinateur
Laurent Romary	Inria - ALMAnaCH	Directeur
Éric de la Clergerie	Inria - ALMAnaCH	Co-encadrant

Abstract

Contents

1 Introduction

> "The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. [...] We should build in only the meta-methods that can find and capture this arbitrary complexity. [...] The eventual success is tinged with bitterness, and often incompletely digested."
>
> - Richard S. Sutton, *The Bitter Lesson*

Healthcare has been one of the earliest and most prominent application domains for artificial intelligence. Use cases range from diagnostic support to medical record management, cohort selection, report generation, and clinical information extraction.

As early as the 1970s, MYCIN, the iconic expert system developed at Stanford, diagnosed bacterial infections and recommended antibiotics. It already achieved performance comparable to human specialists in its narrow domain. INTERNIST-1, later known as CADUCEUS, tackled internal medicine diagnosis across hundreds of diseases. However, these systems relied on explicit rules and static knowledge bases, which limited their ability to handle the complexity and variability of real-world medical data. They were rigid and required constant maintenance to keep up with medical advances. Building such systems demanded expensive medical experts to annotate clinical data and write rules by hand.

With the rise of machine learning in the 1990s and 2000s, data-driven approaches began to dominate healthcare AI. Supervised and unsupervised learning algorithms trained models directly from data. The interpretability of rule-based systems was traded for the performance of deep neural networks capable of capturing complex patterns. R2 ImageChecker (1998) became the first FDA-approved computer-aided detection system for mammography, using neural networks to spot suspicious microcalcifications. PAPNET applied neural networks to cytological screening for cervical cancer.

Then came the scaling era. The recipe became simple: pretrain a language model on billions of words of raw text, then fine-tune on the target task. This approach routinely outperformed systems built with years of domain expertise. The need for expensive annotation, hand-crafted rules, and curated knowledge bases faded, beaten by data and compute.

Despite healthcare offering many structured and curated knowledge bases such as UMLS, SNOMED-CT, and RxNorm, experience has shown that general-purpose language models pre-trained on massive unstructured data outperform approaches based on domain-specific rules and knowledge bases. It seems our models cannot effectively integrate these structured resources, or perhaps these knowledge bases do not cover the full complexity and variability of medical knowledge as it is used in practice.

However, healthcare faces a limitation that other domains do not: the confidentiality and scarcity of annotated clinical data. Medical data is sensitive and difficult to obtain in large quantities for

model training. Works adapting language models to the medical domain often rely on public data such as PubMed or patient forums, but these do not always reflect the real clinical language used in electronic health records. The clinical domain has unique jargon, abbreviations, and writing styles that differ from general medical texts. Moreover, while MIMIC exists in the United States, public clinical datasets remain scarce in other countries such as France.

This leads to a fundamental problem: the paradigm that has won is pretraining at scale, but in healthcare we lack massive public clinical data to pretrain language models. Can we find new ways to exploit public data for pretraining language models adapted to the clinical domain, without using sensitive clinical data?

Part I

Related Works

2 Language Models

2.1 Introduction

Un language model c'est un modele probabiliste sur des sequences de tokens:

$$P(\mathbf{w}) = \prod_{t=1}^{L} P(w_t \mid w_{<t})$$

Factorisation autoregressive (chain rule). Fondation de tout le NLP moderne.
Pourquoi c'est important pour la these:

- Clinical IE repose sur des LMs pretrained
- Domain adaptation = defi central
- French + biomedical + clinical = triple rarete de donnees

-> plan du chapitre: on remonte des n-grams jusqu'aux archis modernes, puis on zoome sur le continual pretraining qui est au coeur de la these

2.2 From Statistical to Neural Language Models

2.2.1 N-gram Models

Hypothese de Markov: $P(w_t \mid w_{<t}) \approx P(w_t \mid w_{t-n+1}, \ldots, w_{t-1})$.

- Kneser-Ney smoothing
- Probleme: curse of dimensionality, $|\mathcal{V}|^n$ explose

2.2.2 Neural LMs

- Bengio et al. (2003): neural probabilistic LM
  - Embeddings denses, resout la sparsity
  - Mots similaires -> vecteurs proches
- Word2Vec (Mikolov et al., 2013): skip-gram / CBOW
  - $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$
- ELMo (Peters et al., 2018): embeddings contextuels via biLSTM

  - Meme mot, contextes differents -> vecteurs differents
  - Precurseur du pretrain-then-finetune

2.2.3 RNNs

- LSTM (Hochreiter & Schmidhuber, 1997)
- En theorie: long-range dependencies
- En pratique: vanishing gradients, pas parallelisable

-> bottleneck pour scaler -> besoin d'une nouvelle archi

2.3 Transformer

2.3.1 Self-Attention

Vaswani et al. (2017).

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

- Interactions directes cross-position (vs RNN indirect via hidden state)
- $O(L^2)$ temps et memoire
- Parallelisable
- Multi-head: h tetes, chacune specialise (syntaxe, semantique, position)

2.3.2 Positional Encodings

- APE sinusoidal (Vaswani 2017) -> extrapolation limitee
- APE learned (Devlin 2019) -> pas d'extrapolation
- ALiBi (Press et al. 2022) -> bias lineaire, mieux
- RoPE (Su et al. 2021) -> matrices de rotation, le meilleur

RoPE = standard maintenant: LLaMA, Mistral, ModernBERT.

2.3.3 Encoder vs Decoder

- Encoder: attention bidirectionnelle, MLM, classification/NER (BERT, RoBERTa)
- Decoder: attention causale, CLM, generation (GPT, LLaMA)
- Encoder-decoder: T5 (Raffel et al., 2020), seq2seq, plus trop utilise

-> maintenant qu'on a l'archi, question: quel objectif de pretraining?

2.4 Pretraining Objectives

2.4.1 CLM (Causal Language Modeling)

GPT (Radford et al., 2018).

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{L} \log P(w_t \mid w_{<t})$$

- Tous les tokens contribuent au gradient
- Supervision dense
- Naturel pour la generation

2.4.2 MLM (Masked Language Modeling)

BERT (Devlin et al., 2019).

$$\mathcal{L}_{\text{MLM}} = -\sum_{t \in \mathcal{M}} \log P(w_t \mid \mathbf{w}_{\setminus t})$$

- 15% tokens masked (80% [MASK], 10% random, 10% unchanged)
- Contexte bidirectionnel -> representations plus riches
- Mais: seulement 15% contribuent -> gradient sparse
- Contraste avec CLM ou chaque token compte

2.4.3 Objectifs hybrides (2024-2025)

Important pour le Ch.5 de la these.

Approche CLM puis MLM:

- CLM d'abord puis decay vers MLM = meilleur que MLM seul
- CLM: gradients denses (tous les tokens), MLM: sparse (15%)
- Hypothese: CLM "compresse" la connaissance, MLM restaure la bidirectionnalite

Refs:

- "Should We Still Pretrain Encoders with MLM?" (arXiv:2507.00994, 2025) -> biphasic CLM vers MLM bat pure MLM
- GPT-BERT (arXiv:2410.24159, 2024) -> unifie CLM+MLM
- AntLM (CoNLL 2024) -> alternance CLM/MLM

Note: ELECTRA (Clark et al., 2020) avait propose RTD pour signal plus dense, mais ModernBERT montre que bonne archi + MLM suffit (bat DeBERTaV3 qui utilise RTD).

-> on a les archis et les objectifs, question maintenant: est-ce que ca scale?

2.5 Scaling Laws & LLMs

2.5.1 Scaling

- GPT-3 (Brown et al., 2020): 175B params, few-shot emergent, prompting remplace le finetuning
- Chinchilla (Hoffmann et al., 2022):

$$L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E$$

  -> compute-optimal: smaller model + more data bat larger model + less data
- LLaMA (Touvron et al., 2023): 7B rivalise avec bcp plus gros via 1T tokens -> open weights, democratisation

2.5.2 Post-ChatGPT

- RLHF (Ouyang et al., 2022) -> preference learning
- Instruction tuning
- ChatGPT, Claude, Gemini
- Emergent abilities (Wei et al., 2022) -> chain-of-thought etc.

-> ok ca scale pour le general, mais nous on a un domaine specifique -> comment adapter?

2.6 Continual Pretraining

Section centrale pour la these. C'est le coeur de ce qu'on fait (CamemBERT-bio, ModernCamemBERT-bio).

2.6.1 Fondements et terminologie

"Don't Stop Pretraining" (Gururangan et al., ACL 2020):

- DAPT = Domain-Adaptive Pre-Training: 2eme phase sur corpus domaine
- TAPT = Task-Adaptive Pre-Training: sur donnees non-annotees de la tache
- DAPT + TAPT combine -> meilleurs resultats
- Jusqu'a +3 F1 meme avec petit corpus

Taxonomie moderne (Wang et al., CSUR 2025):

- CPT = nouvelles donnees generales au fil du temps
- DAP = specialisation domaine
- CFT = adaptation sequentielle de taches

2.6.2 Catastrophic Forgetting

Le probleme (Kirkpatrick et al., 2017):

- Les reseaux ecrasent les poids necessaires aux taches precedentes
- EWC (Elastic Weight Consolidation): penaliser les changements sur les poids importants

Pourquoi ca compte pour les LLMs:

- Continual PT peut degrader les capacites generales
- Capacite a suivre des instructions particulierement fragile
- Petits modeles plus sensibles (Yildiz et al., 2024)

"Spurious forgetting" (2024): les baisses de perf refletent peut-etre une perte d'alignement, pas une vraie perte de connaissance.
-> ok c'est un vrai probleme, quelles solutions?

2.6.3 Solutions recentes (2024-2025)

Ibrahim et al. (2024), "Simple and Scalable Strategies":

- LR re-warming: relancer le learning rate quand on ajoute des donnees
- LR re-decaying: cosine decay apres warmup
- Replay: mixer 1% des donnees precedentes -> ca suffit
- Valide a 405M et 10B
- -> egalise from-scratch pour une fraction du compute

Stability gap (arXiv:2406.14833, 2024):

- Chute temporaire au debut du continual PT, puis recovery
- Solutions: subset propre, sous-corpus de qualite, data mixing
- 36.2% -> 40.7% avec seulement 40% du budget training

Data selection:

- Donnees non pertinentes = degradation
- 10% du corpus bien selectionne fait aussi bien que 100% vanilla continual PT

2.6.4 From scratch vs continual PT

- From scratch (PubMedBERT, DrBERT): vocab custom, pas de forgetting, mais tres cher
- Continual (BioBERT, Gururangan): efficient, preserve le general, mais tokenizer sous-optimal

Cout compute:

- DrBERT from scratch: 128x V100, 20h
- AliBERT from scratch: 48x A100, 20h
- BioBERT continual: 8x V100, ˜10 jours
- -> from-scratch coute ˜30x plus de compute

2.6.5 Encoders vs decoders: dynamiques differentes

Decoders (LLMs), ca marche pas bien:

- BioMistral: -0.9 points apres continual PT
- Dorfner et al. (2024): biomedical LLMs sous-performent les generalistes
- OpenBioLLM-8B: 30% vs Llama-3-8B: 64.3% sur cas NEJM
- -> hypothese: perte de connaissances generales pendant la specialisation

Encoders, c'est different:

- Contexte bidirectionnel peut-etre plus robuste au forgetting
- MLM moins sujet au catastrophic forgetting?
- -> voir Ch.5: approche CLM vers MLM sur les encoders

-> tension narrative: continual PT marche pas pour les decoders, et pour les encoders? -> c'est la question de la these

2.6.6 Tentatives anterieures en francais biomedical

- Copara et al. (2020): 31K articles -> aucune amelioration (corpus trop petit)
- Le Clercq de Lannoy et al. (2022): 136M mots -> +2 EMEA seulement
- Dura et al. (2022): 21M docs APHP -> +3% mais donnees privees

Question ouverte: pourquoi des resultats contradictoires? DrBERT dit que continual PT marche pas. CamemBERT-bio (cette these) montre que si, avec le bon corpus.
-> ok on sait ce qu'est le continual PT, maintenant quels modeles existent?

2.7 Modeles pretrained biomed

2.7.1 Encoders biomed anglais

- BioBERT (Lee et al., 2019): PubMed+PMC 18B mots, continual -> +0.62% NER, +2.80% RE
- SciBERT (Beltagy et al., 2019): Semantic Scholar 1.14M papers, from scratch -> 42% vocab overlap avec BERT
- PubMedBERT (Gu et al., 2022): PubMed only, from scratch -> introduit BLURB benchmark
- ClinicalBERT (Alsentzer et al., 2019): MIMIC-III, continual -> acces restreint
- BioLinkBERT (Yasunaga et al., 2022): PubMed + citation links, from scratch -> +7% BioASQ

SOTA 2025: BioClinical ModernBERT, cf section architectures modernes.

2.7.2 Modeles francais

General:

- CamemBERT (Martin et al., 2020): RoBERTa sur OSCAR 138GB -> resultat interessant: 4GB donne a peu pres la meme perf que 138GB
- FlauBERT (Le et al., 2020)
- CamemBERT 2.0 / CamemBERTa (2024): architecture DeBERTaV3, nouveau tokenizer

Biomedical francais:

- DrBERT (Labrak et al., 2023): from scratch, public + prive -> pretend que continual PT marche pas pour le francais biomed
- AliBERT (Berhe et al., 2023): from scratch, Unigram regularise -> pas disponible publiquement

-> tension: DrBERT dit from-scratch necessaire, CamemBERT-bio (cette these) montre le contraire

2.7.3 Decoders biomed

- BioMistral (Labrak et al., 2024): 3B tokens PMC, 32x A100 20h, base Mistral
- Meditron (Chen et al., 2023): 46B tokens, 128x A100 332h, base Llama-2
- PMC-LLaMA (Wu et al., 2024): 75B tokens, massif

Continual PT pour decoders = resultats mitiges:

- BioMistral: -0.9 points, recupere seulement via model merging
- Dorfner et al. (2024): “Biomedical LLMs Seem not to be Superior to Generalist Models”
- -> contraste avec les encoders, voir Ch.5

-> les modeles existent mais les archis ont evolue, quoi de neuf?

2.8 Architectures modernes

2.8.1 Modernisation des encoders

DeBERTa (He et al., ICLR 2021):

- Disentangled attention: separe contenu et position
- Premier a battre l’humain sur SuperGLUE
- -> influence sur ModernBERT, CamemBERT 2.0

MosaicBERT (Portes et al., NeurIPS 2023):

- 30% masking au lieu de 15% BERT
- FlashAttention + ALiBi + GLU + unpadding
- BERT-base en 1.13h sur 8x A100 (˜$20)
- -> fondation directe de ModernBERT

2.8.2 Efficient Attention

FlashAttention (Dao et al., 2022):

- Reorganise le calcul pour la hierarchie memoire GPU (tiling, kernel fusion)
- Meme $O(L^2)$ mais 2-4x plus rapide en pratique
- Standard maintenant

Autres:

- GQA (Ainslie et al., 2023): partage de KV-heads -> Llama-2, Mistral
- RMSNorm (Zhang & Sennrich, 2019): LayerNorm simplifie

2.8.3 Long context

Le probleme: BERT = 512 tokens, comptes-rendus cliniques = souvent 2000+. Tronquer = perdre de l'info diagnostique.

- Longformer (Beltagy et al., 2020): 4096 tokens, sliding window + global attention
- BigBird (Zaheer et al., 2020): 4096, sparse attention
- Clinical Longformer (Li et al., 2022): 4096, Longformer continue pretrained sur MIMIC-III
- ModernBERT (Warner et al., 2024): 8192 tokens
  - FlashAttention v2/v3
  - RoPE
  - Alternance local/global attention
  - 30% masking (de MosaicBERT)
  - -> 16x le contexte de BERT

Adaptations cliniques 2025:

- Clinical ModernBERT (Lee et al., 2025): PubMed + MIMIC-IV + ontologies
- BioClinical ModernBERT (2025): 53.5B tokens, 20 datasets cliniques -> SOTA actuel

2.9 Tokenization

2.9.1 Methodes subword

- BPE (Sennrich et al., 2016): merge les paires les plus frequentes
- SentencePiece (Kudo & Richardson, 2018): language-independent
- Unigram (Kudo, 2018): selection probabiliste

2.9.2 Domain mismatch

SciBERT: vocab general vs scientifique = 42% intersection seulement.
Exemple biomed francais:

- Tokenizer general: "echocardiographie" -> ["echo", "##cardi", "##ographie"]
- Tokenizer domaine: "echocardiographie" -> ["echocardiographi", "e"]

Tradeoff:

- From-scratch = vocab custom mais cher
- Continual PT = tokenizer sous-optimal mais efficient

2.10 Limites et transition

- Rarete des donnees: texte clinique prive (CNIL), modeles hospitaliers pas partageables -> besoin d'alternatives publiques -> Ch.3 Corpus Annotation
- Knowledge gap: LMs captent des patterns, pas la structure des ontologies. Codage medical = connaissances hierarchiques -> approches knowledge-enhanced -> Ch.4 Clinical IE
- Documents longs: comptes-rendus hospitaliers depassent les limites de BERT. Longformer et ModernBERT aident mais faut adapter au domaine

-> ce chapitre a couvert les fondations techniques du language modeling. Le chapitre suivant traite de comment construire et annoter des corpus pour entrainer ces modeles.

3 Pretraining Data: Corpora and Knowledge

3.1 Introduction

Le chapitre precedent couvrait les modeles. Maintenant faut les nourrir.

Les LMs modernes sont data-hungry:

- GPT-3 (Brown et al., 2020): 300B tokens
- LLaMA (Touvron et al., 2023): 1-2T tokens
- Chinchilla (Hoffmann et al., 2022): compute-optimal = plus de data, modele plus petit

La qualite et la composition du corpus de pretraining determinent largement les capacites du modele. "Data is the new bottleneck."

Deux types de donnees de pretraining:

- Texte brut: web crawls, articles scientifiques, livres
- Connaissances structurees: ontologies medicales, knowledge graphs

Pourquoi c'est central pour la these:

- Part 1 = construire un corpus biomed francais a partir de sources publiques
- Ch.6 = injecter des connaissances d'ontologies medicales pendant le pretraining
- Triple rarete: francais + biomedical + clinique

-> plan du chapitre: on part des corpus web generalistes, on zoome sur le filtrage qualite, puis sur les sources biomed specifiques, et enfin sur les ontologies medicales et comment les integrer dans le pretraining

3.2 Web-Scale Pretraining Corpora

3.2.1 Common Crawl

Source de base de presque tous les grands corpus de pretraining:

- Web crawl ouvert, 2.5B pages par snapshot mensuel, 250-400 TiB non compresse
- Snapshots mensuels depuis 2008, petaoctets cumules
- Contenu tres heterogene: articles, forums, spam, code, pubs, boilerplate...
- Pas utilisable directement -> necessite des pipelines de nettoyage lourdes

3.2.2 Premiers grands corpus

C4 (Raffel et al., 2020):

- Colossal Clean Crawled Corpus, 750GB d'anglais extrait d'un seul snapshot CC (avril 2019)
- Filtres: langdetect >= 99% anglais, suppression phrases incompletes, dedup
- Dodge et al. (2021) documentent le contenu: brevets, sites militaires US, texte machine-generated, exemples de benchmarks
- Blocklist filtering retire disproportionnellement du texte de/sur les minorites

The Pile (Gao et al., 2020):

- 800GB anglais, 22 sources diversifiees (PubMed, ArXiv, GitHub, StackExchange, Wikipedia, livres, etc.)
- Idee: diversite des sources > volume brut d'une seule source
- Open-source (EleutherAI)

OSCAR (Ortiz Suarez et al., 2019):

- Pipeline asynchrone pour classifier CC par langue
- Optimise pour infra moyenne/basse ressource (I/O-bound)
- Produit le corpus de CamemBERT: 138GB de texte francais
- Shuffled au niveau ligne pour eviter les problemes de copyright (ironiquement)

3.2.3 Corpus modernes (2024-2025)

FineWeb (Penedo et al., 2024):

- 15T tokens, 96 snapshots CC
- Documentation detaillee des strategies de dedup et filtrage
- Pipeline: extraction texte, filtrage langue, heuristiques qualite, dedup URL + MinHash
- FineWeb-Edu: sous-ensemble 1.3T tokens filtre pour contenu educatif (cf section suivante)

RedPajama-V2 (Weber et al., 2024):

- Dataset ouvert pour entrainer des LLMs
- Inclut subset francais (RPv2-Fr)
- Metriques de qualite pre-calculees pour faciliter le filtrage

TxT360 (Tang et al., 2024):

- "Perfect blend" de sources
- Dedup globale sur 99 snapshots CC: 20TB -> 4.83T tokens (reduction 80%)
- Pipeline similaire a FineWeb + near-dedup globale additionnelle

Autres:

- Dolma (Soldaini et al., 2024): 3T tokens, corpus ouvert pour recherche (OLMo)
- ROOTS (Laurencon et al., 2023): 1.6TB, multilingue composite (BigScience/BLOOM)

-> on a des milliards de tokens de texte web, mais la qualite varie enormement -> comment selectionner les bons documents?

3.3 Data Quality and Curation

3.3.1 Filtrage heuristique

Approches classiques, devenues standard depuis Gopher (Rae et al., 2022):

- Regles: longueur min/max du document, ratio mots/symboles, proportion de majuscules, lignes dupliquees
- Filtrage par perplexite: KenLM entraine sur Wikipedia, documents a perplexite trop haute = bruit
- Language ID: fastText pour filtrer par langue
- Ces heuristiques sont reprises par presque tous les pipelines modernes (FineWeb, C4, Dolma)

Limites: les heuristiques sont necessaires mais insuffisantes. Elles retirent le bruit evident mais ne distinguent pas la qualite du contenu.

3.3.2 Classifieurs de qualite

Paradigme recent: LLM-as-annotator -> distillation vers petit classifieur -> filtrage a grande echelle.

FineWeb-Edu (Penedo et al., 2024):

- Llama-3-70B-Instruct annote 460K pages web sur "educational value" (echelle 0-5)
- Echelle additive: le LLM evalue chaque critere et construit le score pas a pas
- Prompt cible le niveau primaire/college pour eviter de favoriser les papiers trop techniques
- Distillation vers classifieur de regression (Snowflake-arctic-embed, F1 = 82% en binaire avec seuil >= 3)
- Classification de 15T tokens: 6K heures H100

- Retire 92% des donnees, bat quand meme le corpus complet sur MMLU, ARC, OpenBookQA

DCLM / DataComp-LM (Li et al., 2024):

- Classifieur FastText entraine pour separer instructions synthetiques (Open Hermes 2.5) du web general
- Favorise les structures Q&A resolues: question courte + reponse directe + raisonnement bref
- Utilise dans le mix de pretraining d'OLMo2

Nemotron-CC (NVIDIA, 2024):

- Classifieur de qualite multi-criteres (accuracy, clarity, coherence, etc.)
- Approche plus granulaire que le score educatif unique de FineWeb-Edu

3.3.3 Organisation par domaine

WebOrganizer (Wettig et al., 2025):

- Organise le web en taxonomies par topic et par format
- Annotations LLM distillees vers classifieurs efficaces
- Montre que FineWeb-Edu a un biais implicite vers Science & Technology et Academic Writing
- L'efficacite de FineWeb-Edu vient en partie de preferences de domaine alignees avec les benchmarks (MMLU, HellaSwag)
- Combiner organisation par domaine + filtrage qualite > qualite seule

3.3.4 Deduplication

- URL-level: retirer les duplicats exacts d'URL
- Document-level: MinHash + LSH pour near-duplicates
- TxT360: dedup globale sur 99 snapshots CC, reduction de 80%
- Crucial: la duplication biaise le training vers les textes les plus copies (boilerplate, templates)

3.3.5 Contamination

Le filtrage agressif pose un probleme de contamination des benchmarks:

- Deng et al. (2024): ChatGPT et GPT-4 devinent 52% et 57% des options manquantes de MMLU -> les QA pairs sont sur le web
- Xu et al. (2024): survey des techniques de detection de contamination dans les LLMs
- InfiniGram (Liu et al., 2024): outil pour identifier des exact matches dans les training sets

Tension fondamentale: filtrer plus agressivement = meilleurs scores benchmarks, mais potentiellement via contamination implicite plutot que vraie qualite.

-> ok le filtrage marche pour les corpus web generalistes, mais en biomedical c'est different: les corpus sont plus petits, le domaine est tres specifique, les criteres de qualite ne sont pas les memes -> quelles sources biomed existent?

3.4 Biomedical and Clinical Corpora

3.4.1 Grandes sources anglophones

PubMed:

- Base de donnees de la NLM (National Library of Medicine)
- 39M citations et abstracts d'articles biomedicaux
- Texte structure, vocabulaire technique, qualite editoriale
- Abstracts seulement (pas le full-text)
- Utilise par BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2022)

PMC Open Access:

- 4.5M articles full-text en acces libre
- 98% anglais (Labrak et al., 2024)
- Contenu heterogene: articles de recherche, reviews, case reports, editoriaux, guidelines
- Utilise par BioMistral (3B tokens), Meditron (46B tokens), PMC-LLaMA (75B tokens)
- Probleme: granularite article. Un article de recherche peut contenir a la fois des cas cliniques pertinents et des sections methodologiques non pertinentes

MIMIC (Johnson et al., 2016/2023):

- MIMIC-III, MIMIC-IV: notes cliniques du Beth Israel Deaconess Medical Center
- Acces restreint (data use agreement, formation ethique)

- Utilise par ClinicalBERT (Alsentzer et al., 2019), Clinical Longformer (Li et al., 2022), BioClinical ModernBERT (Sounack et al., 2025)
- Gold standard pour le texte clinique anglais, mais un seul hopital americain -> pas generalizable

3.4.2 Curation article-level pour le biomed

Meditron (Chen et al., 2023):

- Score 0-1 par article base sur: MeSH tags, type de publication, reputation du journal, recence, nombre de citations
- Upsampling des articles a haut score, filtrage des bas scores
- 46B tokens au total, 128x A100, 332h pour le 70B
- Limite: granularite article. Un article peut melanger contenu pertinent et non-pertinent

BioClinical ModernBERT (Sounack et al., 2025):

- 169B tokens: PubMed + PMC + 20 datasets cliniques (MIMIC-III, MIMIC-IV, CheXpert Plus, etc.)
- SOTA actuel sur les benchmarks biomed anglais
- Mais: la plupart des sources cliniques necessitent des data use agreements -> pas vraiment "public"

3.4.3 Sources francophones

Ressources publiques existantes:

- ISTEX: base de 27M references scientifiques, permet d'extraire des documents francais bio/med
- CLEAR (Grabar & Cardon, 2018): notices medicamenteuses, articles d'encyclopedie medicale en francais technique et simplifie
- E3C (Magnini et al., 2021): corpus multilingue de cas cliniques, notices, resumes de these medicale
- HAL / Halvest (Kulumba et al., 2024): articles scientifiques francais et anglais extraits du depot HAL (Hyper Articles en Ligne)
- Wikipedia biomedical: articles des portails medecine, pharmacie, biologie. Libre et multilingue

Modeles entraines from scratch sur corpus francais:

- DrBERT (Labrak et al., 2023): corpus de 7GB, from scratch sur 128x V100

  - Pretend que le continual PT ne marche pas pour le francais biomed
  - Mais: evaluation potentiellement biaisee (cf Ch.4 de la these)

- AliBERT (Berhe et al., 2023): ScienceDirect + theses Sudoc, tokenizer Unigram regularise, 48x A100
  - Modele non disponible publiquement

Tentatives de continual pretraining en francais biomed:

- Copara et al. (2020): 31K articles seulement -> aucune amelioration sur CamemBERT-large
- Le Clercq de Lannoy et al. (2022): PubMed + Cochrane + ISTEX + Wikipedia = 136M mots -> +2 F1 sur EMEA, rien sur MEDLINE
- Dura et al. (2022): 21M documents cliniques APHP -> +3% sur APMed (prive), mais donnees non partageables

Resultats décevants: DrBERT dit que continual PT marche pas, Copara et Le Clercq confirment. Mais le probleme est peut-etre le corpus (trop petit, pas assez cible), pas la methode.

Triple rarete: francais + biomedical + clinique. Les textes cliniques sont prives (CNIL), les modeles hospitaliers ne sont pas partageables entre etablissements. Les sources publiques francophones sont fragmentees et petites comparees a PubMed/PMC.

-> le texte biomedical aide pour le pretraining, mais les LMs captent des patterns statistiques, pas la structure des ontologies. Pour le codage medical (CIM-10, CCAM) il faut des connaissances hierarchiques explicites -> peut-on utiliser directement les ontologies comme donnees de pretraining?

3.5 Medical Ontologies and Knowledge Graphs

3.5.1 Principales ontologies medicales

UMLS (Lindberg et al., 1993):

- Unified Medical Language System, meta-ontologie biomedical de la NLM
- Integre 189 vocabulaires sources (SNOMED-CT, MeSH, ICD, ATC, etc.)
- 3.4M concepts, 16.7M noms de concepts uniques
- Standard international pour la recherche en NLP biomed
- Contient: synonymes, relations hierarchiques, relations associatives, definitions

SNOMED-CT:

- 371K concepts cliniques structures
- Relations hierarchiques (is-a) + associatives (finding site, causative agent, etc.)

- Standard pour les systemes d'information clinique internationaux
- Couverture: diagnostics, procedures, anatomie, substances, organismes

CIM-10 / ICD-10 (OMS, 2019):

- Classification internationale des maladies
- Hierarchie: chapitres -> blocs -> categories -> sous-categories
- Version francaise enrichie par l'ATIH avec: exclusions, inclusions, notes, libelles courts/longs
- Utilise pour le codage PMSI (Programme de Medicalisation des Systemes d'Information) en France
- 19K codes dans la version ATIH

CCAM (ATIH):

- Classification Commune des Actes Medicaux, 8K codes
- Specifique a la France, structure hierarchique
- Relations: actes associes, incompatibilites, modificateurs

ATC (OMS):

- Anatomical Therapeutic Chemical, 6K codes au 5eme niveau
- 5 niveaux: anatomique -> therapeutique -> pharmacologique -> chimique -> substance
- Standard pour la classification des medicaments

3.5.2 Representations formelles

- RDF (Resource Description Framework): triplets sujet-predicat-objet, standard W3C
- OWL (Web Ontology Language): logique descriptive, raisonnement automatique, classes, proprietes
- SKOS (Simple Knowledge Organization System): prefLabel, altLabel (synonymes), broader/narrower (hierarchie), related, exactMatch

Les ontologies codent explicitement ce que les LMs doivent apprendre implicitement:

- Hierarchie: E11 (diabete type 2) $\subset$ E10-E14 (diabete sucre)
- Exclusions: E11 $\neq$ E10 (diabete type 1) – a ne pas confondre
- Synonymes: un meme concept a plusieurs libelles (prefLabel + altLabel)
- Relations associatives: complications, traitements, liens entre ontologies

-> les ontologies sont riches en connaissances structurees, mais sous forme de graphes: pas directement exploitables par les LMs qui attendent du texte en language natu -> comment les integrer dans le pretraining?

3.6 Knowledge-Enhanced Pretraining

3.6.1 Embeddings statiques de graphes

Approches qui produisent des vecteurs fixes (pas contextuels) a partir de graphes:

- Node2Vec (Grover & Leskovec, 2016): random walks sur graphe -> skip-gram, apprentissage de representations de noeuds
- RDF2Vec (Ristoski & Paulheim, 2016): adapte le paradigme walk + skip-gram aux graphes RDF. Fondation des approches de walks sur ontologies
- Snomed2Vec (Agarwal et al., 2019): Node2Vec applique a SNOMED-CT, 5-6x amelioration sur concept similarity
- OWL2Vec* (Chen et al., 2021): walks sur ontologies OWL (teste sur Gene Ontology)

Limite: embeddings statiques, pas integres dans un transformer. Un concept = un vecteur fixe quel que soit le contexte. Pas de contextualisation.
-> on peut faire mieux en integrant les KG directement dans les transformers

3.6.2 Injection de connaissances dans les transformers

Premiere generation: injecter des entites dans l'architecture:

- ERNIE (Zhang et al., 2019): aligne entity embeddings (TransE) avec representations textuelles pendant le pretraining. Entity linking implicite
- KnowBERT (Peters et al., 2019): integre des entity linkers comme couches d'attention supplementaires dans BERT
- K-BERT (Liu et al., 2020): injecte des triples de KG comme soft-position embeddings dans l'input. Modifie la sequence d'entree directement

Deuxieme generation: multi-task pretraining language + KG:

- KEPLER (Wang et al., 2021): MLM + TransE knowledge embedding sur Wikidata simultanement. Les deux objectifs partagent le meme encoder
- CoLAKE (Sun et al., COLING 2020): construit des word-knowledge graphs qui unifient contexte textuel et contexte KG, puis pretraine avec MLM modifie sur ces graphes
- DRAGON (Yasunaga et al., NeurIPS 2022): joint pretraining bidirectionnel MLM + link prediction sur UMLS
  - PubMed pour le texte, UMLS pour le KG
  - +3% accuracy sur MedQA, +10% sur raisonnement multi-step
  - Reference pour le multi-task MLM + KG

3.6.3 Approches contrastives medicales

- SapBERT (Liu et al., NAACL 2021): self-alignment pretraining sur synonymes UMLS
  - Contrastive learning: synonymes UMLS = positifs, non-synonymes = negatifs
  - Metric learning scalable sur 4M+ concepts UMLS
  - SOTA entity linking biomed sur 6 benchmarks
  - Un seul modele pour tous les types d'entites ("one-model-for-all")
- CODER (Yuan et al., 2022): contrastive pretraining sur 87M triplets de relations UMLS
  - Relations + synonymes dans l'objectif contrastif
  - Medical term normalization

3.6.4 Graph-to-text pour le pretraining

Approche recente: transformer les graphes en texte lisible par les LMs.
"Textbooks Are All You Need" (Gunasekar et al., 2023):

- Donnees synthetiques de qualite textbook generees par GPT -> petits modeles tres performants
- Montre que la qualite et la structure des donnees importent plus que le volume brut
- Paradigme: generer des donnees synthetiques structurees plutot que crawler du texte brut

AntGLM-Med (Zhou et al., 2023):

- Reecrit des sous-graphes de KG medical en prose naturelle
- Utilise pour le continual pretraining de LLMs medicaux
- Plus proche precedent de la generation de texte a partir d'ontologies pour pretraining

EntiGraph (Yang et al., ICLR 2025):

- Algorithme d'augmentation entity-centric: decompose un corpus en entites puis genere du texte sur les relations entre entites
- 1.3M tokens reels -> 600M tokens synthetiques via GPT-4-turbo
- Scaling log-lineaire de l'accuracy avec le nombre de tokens synthetiques
- Cible les decoders (Llama 3 8B), pas encore teste sur encoders

Limites actuelles du knowledge-enhanced pretraining:

- La plupart des travaux sont en anglais et sur UMLS
- Graph-to-text pour le pretraining est recent (2023-2025)

- Pas encore explore pour les encoders bidirectionnels (DRAGON fait du link prediction, pas du graph-to-text)
- Les ontologies nationales (CIM-10 ATIH, CCAM) n'ont jamais ete utilisees pour du pretraining
- SapBERT et CODER exploitent UMLS mais uniquement via contrastive learning, pas via generation de texte

-> methodes prometteuses mais gap important: langues non-anglaises, ontologies nationales, pretraining d'encoders via graph-to-text

3.7 Limites et transition

- Rarete biomed non-anglais: PMC = 98% anglais, les corpus francais biomed sont petits et fragmentes. Les pipelines de filtrage (FineWeb-Edu, DCLM) sont concus pour l'anglais -> besoin de strategies de collection et de filtrage adaptees au domaine et a la langue
- Granularite de filtrage: les approches existantes (Meditron, FineWeb-Edu) filtrent au niveau document ou article. Mais un article biomed peut contenir a la fois des cas cliniques pertinents et des sections non pertinentes -> besoin de filtrage a granularite plus fine (paragraphe)
- Criteres de qualite non adaptes: les criteres "educatifs" de FineWeb-Edu ne correspondent pas au texte biomedical (cas cliniques, notices medicamenteuses, protocoles). Les biais vers Science & Technology identifies par WebOrganizer peuvent etre differents en biomed -> besoin de criteres domaine-specifiques
- Ontologies nationales sous-exploitees: UMLS domine la recherche, mais les systemes hospitaliers francais utilisent CIM-10/ATIH, CCAM, ATC. Aucun travail n'a integre ces ontologies dans le pretraining -> gap pour le codage medical en francais
- Graph-to-text naissant: AntGLM-Med et EntiGraph montrent la voie mais sont limites aux decoders en anglais. Pas encore explore pour les encoders bidirectionnels ni pour les ontologies non-anglophones -> possibilite de generer du texte "textbook" a partir des ontologies ATIH

-> ce chapitre a couvert les donnees de pretraining: sources textuelles web et biomed, methodes de curation et filtrage, ontologies medicales et leurs integrations dans le pretraining. Le chapitre suivant traite de la tache finale: l'extraction d'information clinique, et de comment les modeles pretrained sont utilises pour cette tache.

4 Clinical Information Extraction

Part II

Building a Biomedical Corpus

5 Collecting Biomedical Text

6 Filtering by Quality Signals

7 Detecting Content Types

Part III

Pretraining Language Models

8 Encoder Models for French Biomedicine

9 When Decoder Continue-Pretraining Stops Working

10 Beyond Masked Language Modeling

Part IV

Adapting to Clinical Tasks

11 Limits of Direct Fine-Tuning

12 Architectures for Low-Resource Extraction

13 Synthetic Data for Task Adaptation

14 Conclusion

Acronyms

PCA	Principal component analysis
SNF	Smith normal form
TDA	Topological data analysis

Glossary

LaTeX	A document preparation system
$\mathbb{R}$	The set of real numbers

Appendix