Adaptation des grands modèles de langue au domaine : le point de vue du Traitement Automatique des Langues

Olivier Ferret

Université Paris-Saclay
CEA-List – Laboratoire LASTI

PRÉAMBULE

- **Inspirations**

  - Projet ANR ADDICTE (2018-2022) : Analyse distributionnelle en domaine de spécialité
    - LS2N, LIMSI (→ LISN), CEA List, CLLE-ERSS

  - Thèse d'Hicham El Boukkouri. Domain Adaptation of Word Embeddings Through the Exploitation of In-domain Corpora and Knowledge Bases, Université Paris-Saclay, novembre 2021
    - Directeur : Pierre Zweigenbaum
    - Co-encadrants : Thomas Lavergne et Olivier Ferret

  - Projet Confiance.ai (2022-) : Methodology for Confident NLP Models with Limited Training Data
    - Renault, CEA List (Romaric Besançon et Julien Tourille), IRT SystemX

PLAN DE LA PRÉSENTATION

- **Domaine de spécialité et LLM : vue d'ensemble**
- **Préentraînement d'un LLM pour un domaine de spécialité**
  - Principes généraux et variantes
  - À partir de zéro ou d'un LLM généraliste
  - Modularité et efficacité : hiérarchie thématique de LLM
- **LLM pour un domaine de spécialité : la question du vocabulaire**
  - Ajouter du vocabulaire
- **LLM pour un domaine de spécialité : la question des connaissances**
  - Injecter des connaissances dans les LLM

DOMAINE DE SPÉCIALITÉ ET LLM : VUE D'ENSEMBLE

NOTION DE DOMAINE

- **Définition de (Pan & Yang, 2010 ; Ruder, 2019)**
  - Domaine $\mathcal{D} = \{\chi, P(X)\}$
    - χ : espace de représentation des données (espace des caractéristiques)
    - $P(X)$: distribution de probabilité sur les données
      - $X = \{x_1, \dots, x_n\}$, avec x_i, la i^{ème} caractéristique de l'espace de représentation
  - Transposition aux grands modèles de langue neuronaux (LLM)
    - χ = ensemble des tokens constituant le vocabulaire du modèle
    - données = séquences de tokens
    - $P(X)$: caractérise le fait que
      - pour un modèle de langue « général » (`gpt-2`), la continuation de la séquence « I would like to buy » est par exemple :

I would like to buy **a new car.**

      - alors que pour un modèle de langue du domaine financier (`finance-gpt2`), elle est plutôt :

I would like to buy **some shares of an underpriced industrial REIT**

NOTION DE DOMAINE DE SPÉCIALITÉ

- **Notion très liée au champ de la terminologie**
  - Domaine de spécialité → langue de spécialité
  - Langue de spécialité : « Sous-système linguistique qui comprend l'ensemble des moyens linguistiques propres à un champ d'expérience particulier (discipline, science, technique, profession, etc.) » (Grand Dictionnaire Terminologique, 1985)

- **Quelques caractéristiques notables**
  - Domaine : lié à une discipline, science, technique, profession…
    - Médecine, finance, droit, environnement…
  - Importance de la notion de terme, donc du vocabulaire
    - En particulier, présence forte de termes multi-mots
    - Moindre ambiguïté sémantique des mots / langue générale
  - Présence de ressources terminologiques : forme de connaissances sur le domaine
  - Types de documents spécifiques
    - Ex. : comptes-rendus médicaux
    - Accessibilité parfois limitée : volumétrie réduite ou contraintes de confidentialité

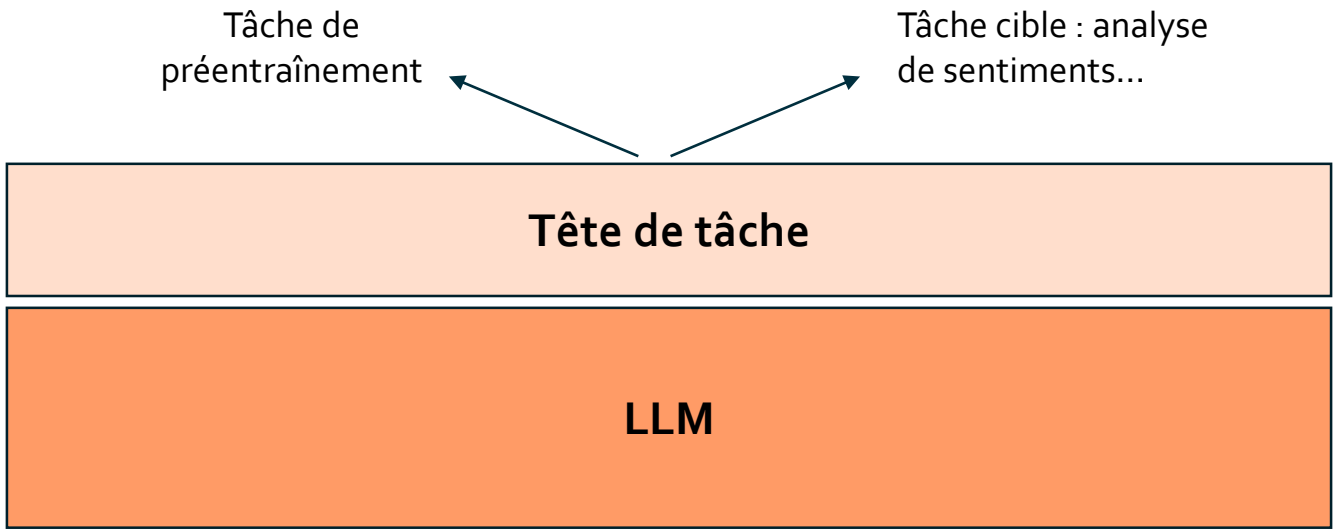
DOMAINE DE SPÉCIALITÉ : DOMAINE VS TÂCHE

- **Domaine**
  - Défini au travers de l'espace de représentation des données et de la distribution de probabilité sur ces données
- **Tâche**
  - Défini selon (Pan & Yang, 2010 ; Ruder, 2019) par
    - Υ : espace des étiquettes
    - $P(Y)$: probabilité a priori sur cet espace
    - $P(Y/X)$: probabilité conditionnelle sur les étiquettes en fonction des données
- **Domaine de spécialité**
  - Défini en premier lieu en tant que domaine au sens de (Pan & Yang, 2010)
  - Mais peut aussi se caractériser par des tâches propres
    - Ex. : extraction de concepts médicaux

  - Dans cette présentation : on fait l'hypothèse de l'existence de données annotées pour les tâches propres au domaine de spécialité
    - Pas d'adaptation au domaine au sens de (Pan & Yang, 2010 ; Ruder, 2019)

LA PHILOSOPHIE LLM

- **LLM : une approche fondée sur l'apprentissage par transfert**
  - 1ère phase : préentraînement du LLM à partir d'un grand corpus
    - Tâche de type « modélisation de langue » : prédire le mot suivant par ex.
  - 2ème phase : entraînement du LLM pour la tâche cible
    - reconnaissance d'entités nommées, classification thématique…
  - ➢ Apprentissage par transfert séquentiel
    - 2 tâches différentes : tâche de préentraînement, puis tâche cible
- **LLM : la philosophie des outils multi-têtes**



LLM ET DOMAINES DE SPÉCIALITÉ

- **Trois grandes approches**
  - Préentraînement à partir de zéro d'un LLM sur un corpus du domaine cible
    - Pré-requis : avoir un corpus du domaine cible de taille suffisante
      - pas nécessairement évident pour tous les domaines de spécialité
  - Adaptation d'un LLM existant avec un corpus du domaine cible
    - Adaptation au domaine du LLM par le biais de la tâche de préentraînement : apprentissage par transfert séquentiel avec la même tâche mais des données différentes
    - Hypothèse : pas d'oubli catastrophique entre les deux applications de la tâche de préentraînement
  - Utilisation d'un LLM « universel »
    - Hypothèse : corpus de préentraînement suffisamment grand, diversifié et couvrant des domaines différents → adaptation non nécessaire pour ce qui est du domaine
    - Philosophie actuelle des modèles d'IA générative à la suite de GPT-3
    - Focalisation sur l'adaptation à la tâche cible (*label shift* plutôt que *domain shift*)
      - Adaptation reposant sur une forme de méta-learning : apprentissage d'instructions

LLM ET DOMAINES DE SPÉCIALITÉ

- **Trois grandes approches**
  - Préentraînement à partir de zéro d'un LLM sur un corpus du domaine cible
    - Pré-requis : avoir un corpus du domaine cible de taille suffisante
      - pas nécessairement évident pour tous les domaines de spécialité
  - Adaptation d'un LLM existant avec un corpus du domaine cible
    - Adaptation au domaine du LLM par le biais de la tâche de préentraînement : apprentissage par transfert séquentiel avec la même tâche mais des données différentes
    - Hypothèse : pas d'oubli catastrophique entre les deux applications de la tâche de préentraînement
  - Utilisation d'un LLM « universel »
    - Hypothèse : corpus de préentraînement suffisamment grand, diversifié et couvrant des domaines différents → adaptation non nécessaire pour ce qui est du domaine
    - Philosophie actuelle des modèles d'IA générative à la suite de GPT-3
    - Focalisation sur l'adaptation à la tâche cible (*label shift* plutôt que *domain shift*)
      - Adaptation reposant sur une forme de méta-learning : apprentissage d'instructions

PRÉENTRAÎNEMENT D'UN LLM POUR UN DOMAINE DE SPÉCIALITÉ

PRINCIPES GÉNÉRAUX ET VARIANTES

ADAPTATION PAR PRÉENTRAÎNEMENT : PRINCIPE GÉNÉRAL

- **En entrée**
  - LLM préentraîné sur un corpus du « domaine général »
    - Ex. : `bert-base` pour l'anglais (Devlin et al., 2019)
    - corpus : Wikipédia en anglais + BookCorpus → ~ 3 milliards de mots
  - Corpus du domaine cible
    - Ex. : corpus biomédical (Le Clercq de Lannoy et al., 2022) → ~ 136 millions de mots
- **En sortie**
  - LLM initial adapté pour le domaine cible
- **Méthode**
  - Application de la tâche de préentraînement initial du modèle sur le corpus du domaine cible
  - BERT
    - tâche dite de Masked Language Modeling (MLM)
    - tâche prédiction de la phrase suivante (NSP) généralement laissée de côté
  - GPT
    - tâche de prédiction du mot suivant

DAPT ET TAPT (GURURANGAN ET AL., 2020)

- **2 variantes de préentraînement d'adaptation au domaine**
  - Variantes se distinguant par le corpus utilisé pour le préentraînement

  - DAPT : Domain Adaptation Pretraining
    - corpus représentatif du domaine cible
    - non annoté
    - de taille significative (a minima une centaine de millions de mots)

  - TAPT : Task Adaptation Pretraining
    - corpus d'entraînement de la tâche cible
    - sans les annotations
    - généralement de taille limitée

- **Possibilité de conjuguer les deux variantes**
  - DAPT puis TAPT

DAPT ET TAPT : CADRE D'EXPÉRIMENTATION

- **4 domaines + taille (en milliards de tokens) des corpus de préentraînement associés**
  - BIOMED : articles dans le domaine biomédical 7,55
  - CS : publications dans le domaine de l'informatique 8,10
  - NEWS : articles journalistiques du média REALNEWS 6,66
  - REVIEWS : avis des utilisateurs du site Amazon 2,11

- **8 tâches considérées**

Domain	Task	Label Type
BIOMED	CHEMPROT	relation classification
	†RCT	abstract sent. roles
CS	ACL-ARC	citation intent
	SCIERC	relation classification
NEWS	HYPERPARTISAN	partisanship
	†AGNEWS	topic
REVIEWS	†HELPFULNESS	review helpfulness
	†IMDB	review sentiment

† : gros jeu de données

DAPT ET TAPT : ÉVALUATIONS

Domain	Task	RoBERTa	DAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$

Gain les plus notables

- **DAPT**
  - Gain de performance systématique mais notable dans un peu moins de la moitié des cas
    - pas notable pour les jeux de données les plus gros

DAPT ET TAPT : ÉVALUATIONS

Domain	Task	RoBERTa	DAPT	TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$

- **TAPT**
  - Gain également systématique et notable dans la moitié des cas
  - TAPT > DAPT dans un peu plus de la moitié des cas mais pas toujours de façon notable

➢ Proximité des données d'adaptation / données de test > volume des données d'adaptation

DAPT ET TAPT : ÉVALUATIONS

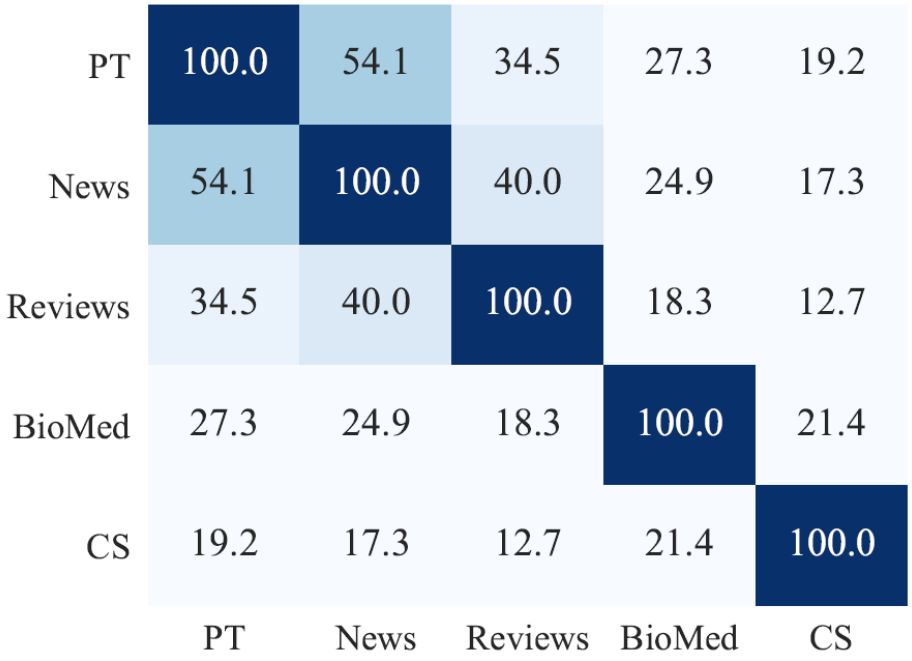
Domain	Task	RoBERTa	DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

- **DAPT + TAPT**
  - Gain presque systématique par rapport aux autres conditions
  - ➢ Meilleures performances dans presque tous les cas

  - ➢ Les deux types d'adaptation sont globalement complémentaires

SIMILARITÉ ENTRE DOMAINE INITIAL ET DOMAINE CIBLE

Recoupement (%) entre les vocabulaires des domaines



PT : corpus d'entraînement de RoBERTa

- **Relation entre corpus de préentraînement initial du LLM et corpus d'adaptation**
  - CS : domaine le plus éloigné du corpus d'entraînement de RoBERTa
  - CS : aussi le domaine pour lequel l'adaptation apporte un gain notable pour toutes les conditions

➢ Intérêt de l'adaptation d'autant plus important pour les domaines de spécialité, généralement peu représentés dans les corpus d'entraînement des LLM

PRÉENTRAÎNEMENT D'UN LLM POUR UN DOMAINE DE SPÉCIALITÉ

À PARTIR DE ZÉRO OU D'UN LLM GÉNÉRALISTE

(EL BOUKKOURI ET AL., 2022)

PROBLÉMATIQUE

- **Adaptation vs entraînement à partir de zéro**
  - Entraînement à partir de zéro beaucoup plus coûteux
  - Mais performance attendue supérieure

- **Un facteur susceptible d'influer sur l'adaptation des LLM : leur vocabulaire**
  - LLM : modèles fondés sur l'architecture Transformeur
  - Gestion d'un vocabulaire ouvert par un mécanisme de découpage en wordpieces
    - Vocabulaire du modèle se répartissant entre
      - un ensemble de mots complets
      - un ensemble de parties de mots – les wordpieces – permettant, par concaténation, de représenter potentiellement tout mot rencontré
    - Vocabulaire construit automatiquement à partir du corpus d'entraînement du modèle
  - Adaptation par préentraînement : utilisation du vocabulaire du modèle originel et non du vocabulaire correspondant au domaine d'adaptation
    - Quel impact sur la performance ?

DOMAINE ET VOCABULAIRE DES LLM

- **Comparaison de 2 domaines**
  - 2 domaines → 2 corpus (en anglais) → entraînement de 2 tokeniseurs BERT → 2 vocabulaires
  - Domaine général : corpus = Wikipédia EN + BookCorpus
  - Domaine médical : corpus = MIMIC-III + PubMed

- **Analyse qualitative**

Terme de référence	Vocabulaire médical	Vocabulaire général
paracetamol	[paracetamol]	[para, ##ce, ##tam, ##ol]
choledocholithiasis	[choledoch, ##olithiasis]	[cho, ##led, ##och, ##oli, ##thi, ##asi, ##s]
borborygmi	[bor, ##bor, ##yg, ##mi]	[bo, ##rb, ##ory, ##gm, ##i]

  - Termes médicaux moins découpés par le vocabulaire du BERT médical que par le vocabulaire du BERT général

CONFIGURATIONS TESTÉES

- **V = vocabulaire, C1 = corpus d'entraînement initial de BERT, C2 = corpus d'adaptation de BERT**

- **V = général, C1 = général, C2 = ∅**
  - entraînement à partir de zéro dans le domaine général
    - corpus : / `bert-base`, remplacement de BookCorpus par OpenWebText
  - comparable à `bert-base`
- **V = général, C1 = général, C2 = médical**
  - DAPT
  - équivalent à BlueBERT (Peng et al., 2019)
- **V = médical, C1 = médical, C2 = ∅**
  - entraînement à partir de zéro dans le domaine médical → modèle natif
  - comparable à PubMedBERT (Gu et al., 2021)
- **V = médical, C1 = médical, C2 = médical**
  - entraînement à partir de zéro dans le domaine médical + réentraînement dans le domaine médical avec le même corpus
  - même taille de corpus que V = général, C1 = général, C2 = médical

TÂCHES D'ÉVALUATION

- **Extraction de concepts médicaux**
  - jeu de données i2b2 médical ; mesure = F1 stricte
  - modèle = BERT + couche linéaire + CRF

The patient had headache that was relieved only with oxycodone. A CT scan of the head showed microvascular ischemic changes. A followup MRI which also showed similar changes. This was most likely due to her multiple myeloma with hyperviscosity.

Clinical Concept Types

Problem Treatment Test

- **Inférences inter-phrastiques**
  - jeu de données MedNLI ; mesure = exactitude
  - modèle = BERT + couche linéaire

Sentence 1 : Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4.
Sentence 2 : Patient has normal Cr.

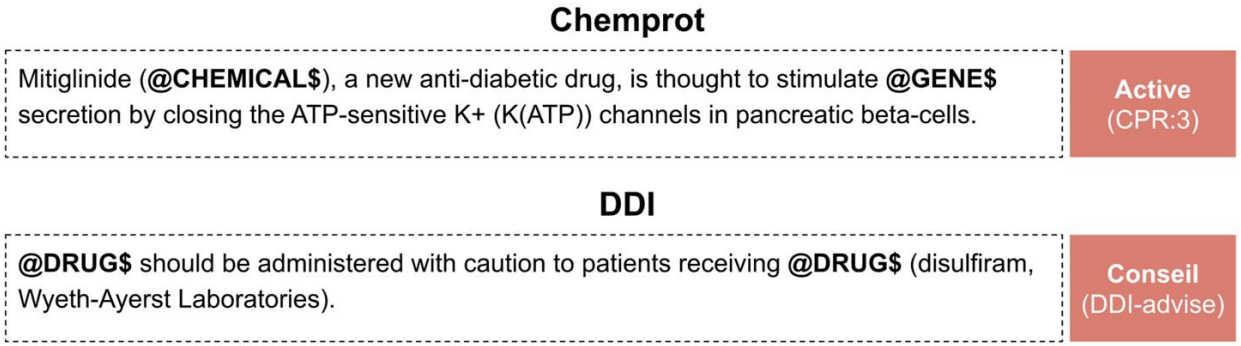
Contradiction

Sentence 1 : Nystagmus and twiching of R arm was noted.
Sentence 2 : The patient had abnormal neuro exam.

Entailment

TÂCHES D'ÉVALUATION

- **Extraction de relations**
  - 2 jeux de données : DDI (Drug Drug Interaction) et ChemProt
    - mesure = micro-F1
  - modèle = BERT + couche linéaire



RÉSULTATS

gras : meilleure performance, souligné : deuxième meilleure

G : général
M : médical

Moyenne ± écart-type sur 10 graines aléatoires

Model			Evaluation Task			
V	C_1	C_2	I2B2	MEDNLI	CHEMPROT	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	89.00 ± 0.17	**84.91 ± 0.46**	72.29 ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	79.40 ± 1.15
M	M	M	**89.20 ± 0.20**	84.32 ± 0.73	**72.97 ± 0.46**	**80.11 ± 0.79**
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BLUEBERT			88.70 ± 0.21	84.53 ± 0.76	68.35 ± 0.61	77.89 ± 0.65

RÉSULTATS

gras : meilleure performance, souligné : deuxième meilleure

G : général
M : médical

Moyenne ± écart-type sur 10 graines aléatoires

Model			Evaluation Task			
V	C_1	C_2	i2b2	MedNLI	ChemProt	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	89.00 ± 0.17	**84.91 ± 0.46**	72.29 ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	79.40 ± 1.15
M	M	M	**89.20 ± 0.20**	84.32 ± 0.73	**72.97 ± 0.46**	**80.11 ± 0.79**
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BlueBERT			88.70 ± 0.21	84.53 ± 0.76	68.35 ± 0.61	77.89 ± 0.65

- **Modèle pour le domaine général (G, G, ∅) vs baseline `bert-base`**
  - Performance un peu moindre mais proche de `bert-base` → bonne base d'expérimentation

RÉSULTATS

gras : meilleure performance, souligné : deuxième meilleure

G : général
M : médical

Moyenne ± écart-type sur 10 graines aléatoires

Model			Evaluation Task			
V	C_1	C_2	i2b2	MedNLI	ChemProt	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	89.00 ± 0.17	**84.91 ± 0.46**	72.29 ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	79.40 ± 1.15
M	M	M	**89.20 ± 0.20**	84.32 ± 0.73	**72.97 ± 0.46**	**80.11 ± 0.79**
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BlueBERT			88.70 ± 0.21	84.53 ± 0.76	68.35 ± 0.61	77.89 ± 0.65

- **Modèle domaine général (G, G, ∅) < tous les modèles domaine médical (*, M, *)**

RÉSULTATS

gras : meilleure performance, <u>souligné</u> : deuxième meilleure

G : général
M : médical

Moyenne ± écart-type sur 10 graines aléatoires

Model			Evaluation Task			
V	C_1	C_2	i2b2	MedNLI	ChemProt	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	<u>89.00</u> ± 0.17	**84.91 ± 0.46**	<u>72.29</u> ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	<u>79.40</u> ± 1.15
M	M	M	**89.20 ± 0.20**	84.32 ± 0.73	**72.97 ± 0.46**	**80.11 ± 0.79**
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BlueBERT			88.70 ± 0.21	<u>84.53</u> ± 0.76	68.35 ± 0.61	77.89 ± 0.65

- **Modèle domaine général (G, G, ∅) < tous les modèles domaine médical (*, M, *)**
- **Modèle domaine général adapté (G, G, M) légèrement > modèle domaine médical entraîné de zéro (M, M, ∅)**
  - Modèle adapté > BlueBERT, correspondant à la même condition

RÉSULTATS

gras : meilleure performance, souligné : deuxième meilleure

G : général
M : médical

Moyenne ± écart-type sur 10 graines aléatoires

Model			Evaluation Task			
V	C_1	C_2	I2B2	MEDNLI	CHEMPROT	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	89.00 ± 0.17	**84.91 ± 0.46**	72.29 ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	79.40 ± 1.15
M	M	M	**89.20 ± 0.20**	84.32 ± 0.73	**72.97 ± 0.46**	**80.11 ± 0.79**
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BLUEBERT			88.70 ± 0.21	84.53 ± 0.76	68.35 ± 0.61	77.89 ± 0.65

- Modèle domaine général < tous les modèles domaine médical
- Modèle domaine général adapté (G, G, M) légèrement > modèle domaine médical entraîné de zéro (M, M, ∅)
- **Modèle domaine général adapté (G, G, M) légèrement < modèle domaine médical entraîné de zéro + réentraînement (M, M, M)**

PRÉENTRAÎNEMENT D'UN LLM POUR UN DOMAINE DE SPÉCIALITÉ

MODULARITÉ ET EFFICACITÉ : HIÉRARCHIE THÉMATIQUE DE LLM

LIMITES DE LA STRATÉGIE DAPT

- **DAPT**
  - 1 domaine → 1 corpus → adaptation du LLM cible par préentraînement sur ce corpus
  - ➢ Stratégie nécessitant une adaptation du LLM par domaine cible

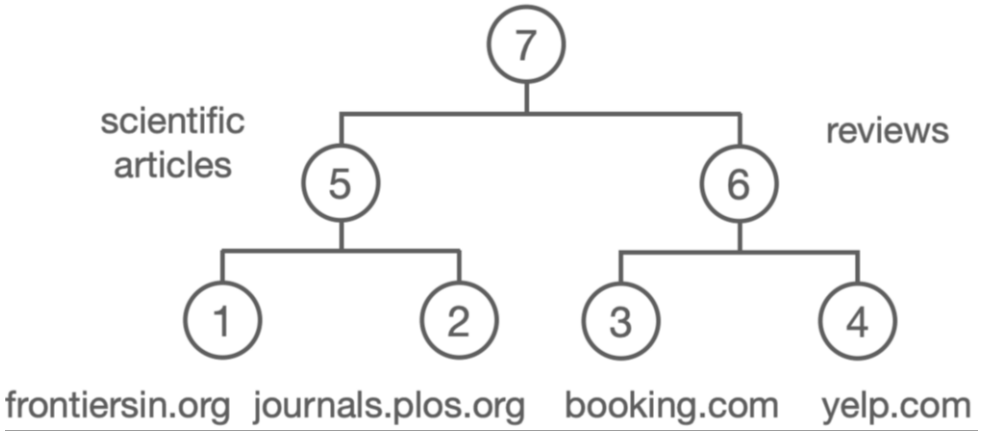
- **Double problème**
  - Coût computationnel
    - Préentraînement d'un LLM : opération coûteuse, même avec des corpus de taille limitée
  - Volumétrie du corpus d'adaptation
    - Nécessité de disposer d'un corpus de taille suffisante pour le domaine cible

- **(Chronopoulou et al., 2022) : stratégie double**
  - Utilisation d'Adapteurs → limitation du nombre de paramètres à entraîner → coût moindre du préentraînement d'adaptation
  - Hiérarchie de domaines : permet d'exploiter les recouvrements entre domaines, en particulier pour limiter le volume des corpus d'adaptation

(CHRONOPOULOU ET AL., 2022) : PRINCIPES

- **Structuration hiérarchique des domaines**
  - Exemple :

Sites Web issus du corpus C4



niveau plus général

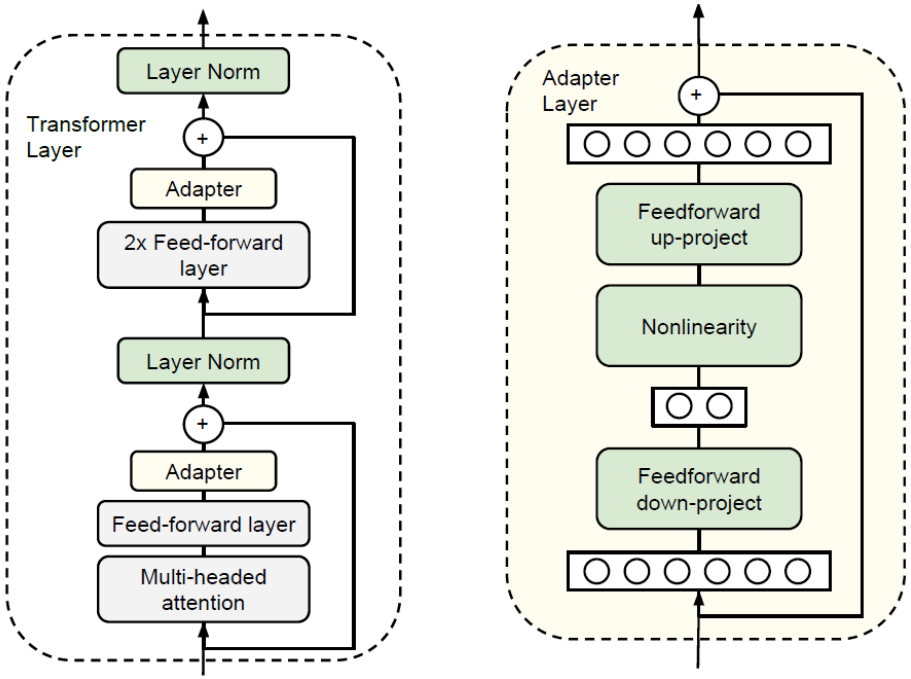
domaines spécifiques

  - Structuration
    - définie manuellement (cf. exemple ci-dessus) ou
    - résultant d'un algorithme de regroupement hiérarchique

- **Structuration du modèle**
  - 1 LLM fixe
  - 1 adapteur pour chaque nœud de la hiérarchie : seule structure mise à jour lors de l'adaptation

UN MOT SUR LES ADAPTEURS

- **Principe (Houlsby et al., 2019)**
  - Au sein de chaque couche d'un modèle Transformeur, insertion de structures de type « bottleneck »
    - Les paramètres du modèle sont gelés
    - Seuls les paramètres des structures insérées sont mises à jour
- **Plus globalement, un exemple de méthode PEFT (Parameter Efficient Fine-Tuning)**

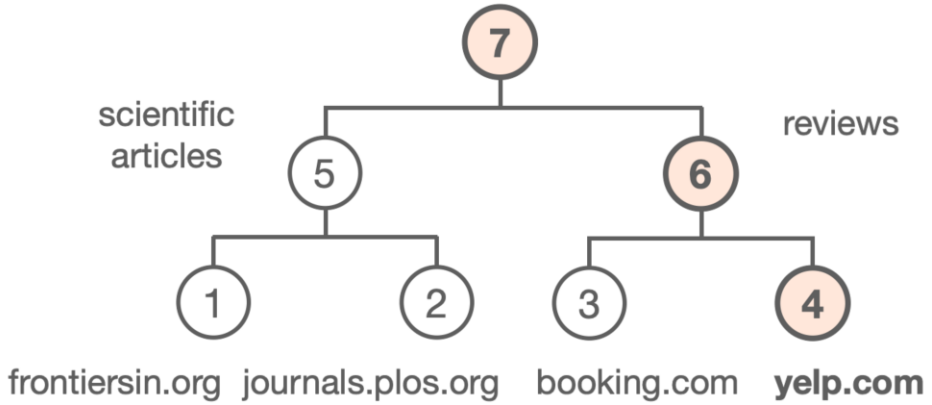


Source : (Houlsby et al., 2019)

FONCTIONNEMENT DU MODÈLE HIÉRARCHIQUE DES DOMAINES

- **Principe**
  - Passe *forward* pour un mini-batch d'exemples appartenant à un domaine spécifique (feuille de la hiérarchie)
    - Pour chaque couche du LLM (figé) → sortie h_i
    - h_i : entrée non seulement de l'adapteur correspondant au domaine cible mais également des adapteurs associés à tous les nœuds entre ce domaine et la racine de la hiérarchie
      - Adapteurs concernés traités en parallèle
    - Finalement, sortie de la couche = moyenne des sorties de tous les adapteurs mobilisés

- **Exemple**
  - Mini-batch contenant des données de **yelp.com**
  - Pour chaque couche du LLM
    - Sortie finale = moyenne des sorties des adapteurs associés aux nœuds 4, 6 et 7



FONCTIONNEMENT DU MODÈLE HIÉRARCHIQUE DES DOMAINES

- **Caractéristiques du modèle**
  - Lors de l'entraînement, les domaines les plus généraux sont mis à jour plus souvent → représentation plus solide
  - En inférence, un domaine spécifique profite des domaines plus généraux qui sont ses ascendants → permet de limiter les données nécessaires pour apprendre un modèle représentation de ce domaine spécifique

- **Si le domaine n'est pas connu a priori**
  - Recherche du domaine connu le plus proche en s'appuyant sur la proximité des nouvelles données avec les données d'entraînement de chaque domaine connu

ÉVALUATION

- **Cadre d'évaluation**
  - LLM : modèle GPT-2 EN
  - Mesure d'évaluation : perplexité
    - à minimiser
  - 2 conditions
    - en domaine : données utilisées pour entraîner le modèle
      - sites booking.com, yelp.com, frontiersin.org et journals.plos.org
    - hors domaine ~ données de test ; comparables aux données d'entraînement mais différentes
      - ex. : pages du site tripadvisor.com, à comparer aux sites booking.com et yelp.com en entraînement
  - 2 baselines
    - multi-adapter : 1 seul adapteur pour tous les domaines spécifiques
    - single adapter : 1 adapteur indépendant des autres / domaine spécifique ≡ DAPT

ÉVALUATION

- **Résultats « en domaine »**

	GPT-2	single adapters	multi adapters	hierarchical adapters
frontiersin.org	22.2	16.1	15.8	15.5
journals.org	24.5	16.6	16.3	15.8
booking.com	29.7	9.7	9.9	9.2
yelp.com	36.2	24.3	25.3	23.8
average	27.7	15.8	15.9	15.2

- Globalement, forte chute de la perplexité avant et après préentraînement → impact notable de l'adaptation
  - Impact globalement fort mais variable selon les domaines : plus fort pour les avis que les articles scientifiques
    - Influence probable de la proximité / corpus de préentraînement du LLM
- Différences assez faibles entre les différentes méthodes d'adaptation
- Avantage tout de même au modèle hiérarchique

ÉVALUATION

- **Résultats hors domaine**

	GPT-2	single adapters	multi adapters	hierarchical adapters
ncbi	20.5	18.2	17.6	16.9
link.springer	27.7	24.5	22.7	22.2
tripadvisor	41.3	36.6	34.1	31.8
techcrunch	27.7	27.1	26.3	25.5
medium	29.1	30.0	27.9	27.1
lonelyplanet	35.5	27.1	24.3	25.3
scholars.duke	22.7	20.1	20.3	19.7
average	29.2	26.2	24.8	24.1

- Logiquement : impact plus faible de l'adaptation
- Persistance de différences parfois notables entre domaines
- Écarts plus importants entre les différentes méthodes d'adaptation
  - hiérarchie d'adapteurs > multi-adapteurs > single adapteurs
  - intérêt plus évident des stratégies intégrant plusieurs domaines / DAPT mono-domaine

LLM POUR UN DOMAINE DE SPÉCIALITÉ : LA QUESTION DU VOCABULAIRE

AJOUT DE VOCABULAIRE

ADAPTATION PAR AJOUT DE VOCABULAIRE : PROBLÉMATIQUE

- **Constat concernant le vocabulaire des LLM**
  - Surdécoupage des mots pour les domaines de spécialité en utilisant des LLM « généralistes »
  - BERTRAM (Schick & Schütze, 2020) : illustration de l'impact négatif du découpage des mots en word pieces
  - Utilisation d'un vocabulaire adapté au domaine
    - Gains modestes par rapport à une adaptation d'un LLM possédant un vocabulaire « généraliste »
    - Stratégie coûteuse du fait de la nécessité d'un entraînement à partir de zéro du LLM

- **Alternative : étendre le vocabulaire du LLM**
  - Conservation du vocabulaire originel
  - Ajout de mots entiers du domaine cible
  - Inconvénient : augmentation de la taille du modèle

PROCESSUS D'EXTENSION DU VOCABULAIRE

- **Sélection du vocabulaire à ajouter**
  - Compromis entre la taille du nouveau vocabulaire (souvent ~10 000 mots) et sa pertinence par rapport au domaine
- **Initialisation des plongements associés au vocabulaire ajouté**
  - Aléatoire, à partir de plongements statiques ou sur la base du vocabulaire existant (moyenne des représentations des sous-mots)
- **Phase de préentraînement des représentations du vocabulaire ajouté**
  - Lors du préentraînement d'adaptation ou lors de l'étape d'affinage propre à une tâche cible
- **Méthode de contextualisation des représentations du vocabulaire ajouté**
  - En gelant le LLM et en y ajoutant une extension dédiée au vocabulaire ajouté
  - En incorporant le vocabulaire ajouté au vocabulaire existant, sans distinction

UN EXEMPLE : (MOSIN ET AL., 2023)

- **Notion de Vocabulary transfer**
  - Processus d'adaptation du vocabulaire d'un LLM (V_{LLM}) à un domaine spécifique

- **Détail du processus pour un modèle BERT**
  - Construction du vocabulaire V_{DOM} du domaine cible à partir d'un corpus, en fixant a priori sa taille
  - Initialisation des représentations associées à ce vocabulaire
    - Si le mot $\in V_{LLM}$, représentation du token dans V_{DOM} = représentation du mot dans V_{LLM}
    - Sinon, application de l'algorithme VIPI (*Vocabulary Initialization with Partial Inheritance*) pour initialiser la représentation du mot dans V_{DOM}
  - Application d'une époque de Masked Language Modeling sur le corpus du domaine cible pour contextualiser les tokens du vocabulaire V_{DOM}
  - Utilisation du modèle résultant pour l'affinage sur une tâche cible

INITIALISATION DES NOUVEAUX MOTS : VIPI

- **Pour chaque nouveau mot du vocabulaire**
  - Production de tous les découpages possibles suivant V_{LLM}
  - Choix du découpage minimisant le nombre de tokens
  - Si découpage non unique
    - Sélection de tous les découpages possédant le plus long token
  - Pour tous les découpages sélectionnés
    - Construction de la représentation du mot par moyennage des représentations des tokens du découpage selon V_{LLM}
  - Représentation du mot dans V_{DOM} : moyennage des représentations de tous les découpages

CADRE D'ÉVALUATION

- **3 jeux de données**
  - Quora Insincere Questions Detection Dataset
    - Taille : 150 Mo
    - Tâche : détection de questions correspondant à des opinions et non des questions
  - Sentiment 140
    - Taille : 300 Mo
    - Tâche : détection de polarité dans des tweets
  - Hyperpartisan
    - Taille : 2,2 Go
    - Tâche : détection de dépêches de presse défendant des points de vue fortement univoques

RÉSULTATS

Dataset	Number of tokens in vocabulary	Original tokenization	New tokenization		
		Pretrained body	Random body	Pretrained body	
		Pretrained embeddings	Random embeddings		VIPI
Quora	8 000	95.64	95.82	96.01	**96.03**
	16 000	95.91	95.92	96.03	**96.11**
	32 000	95.70	95.97	95.83	**96.11**
Sentiment 140	8 000	85.65	85.62	85.71	**85.73**
	16 000	85.64	85.71	85.67	**85.86**
	32 000	84.53	85.23	85.78	**85.80**
Hyperpartisan	8 000	88.66	88.72	88.66	**89.05**
	16 000	86.24	86.51	88.03	**88.58**
	32 000	86.39	86.95	89.17	**89.74**

- **3 tailles de vocabulaire pour chaque tâche**
- **Modèle BERT**
  - Embeddings : plongements associés aux tokens du vocabulaire
  - Body : l'ensemble des couches Transformeur

RÉSULTATS

Dataset	Number of tokens in vocabulary	Original tokenization		New tokenization	
		Pretrained body	Random body	Pretrained body	
		Pretrained embeddings	Random embeddings		VIPI
Quora	8 000	95.64	95.82	96.01	**96.03**
	16 000	95.91	95.92	96.03	**96.11**
	32 000	95.70	95.97	95.83	**96.11**
Sentiment 140	8 000	85.65	85.62	85.71	**85.73**
	16 000	85.64	85.71	85.67	**85.86**
	32 000	84.53	85.23	85.78	**85.80**
Hyperpartisan	8 000	88.66	88.72	88.66	**89.05**
	16 000	86.24	86.51	88.03	**88.58**
	32 000	86.39	86.95	89.17	**89.74**

- **3 tailles de vocabulaire pour chaque tâche**
- **Modèle BERT**
  - Embeddings : plongements associés aux tokens du vocabulaire
  - Body : l'ensemble des couches Transformeur
  - Original tokenization + pretrained embeddings & body ≡ DAPT
  - Random body & embeddings ≡ entraînement à partir de zéro
  - New tokenization + pretrained body = 2 façons d'initialiser les plongements du nouveau vocabulaire

RÉSULTATS

Dataset	Number of tokens in vocabulary	Original tokenization	New tokenization		
		Pretrained body	Random body	Pretrained body	
		Pretrained embeddings	Random embeddings		VIPI
Quora	8 000	95.64	95.82	96.01	**96.03**
	16 000	95.91	95.92	96.03	**96.11**
	32 000	95.70	95.97	95.83	**96.11**
Sentiment 140	8 000	85.65	85.62	85.71	**85.73**
	16 000	85.64	85.71	85.67	**85.86**
	32 000	84.53	85.23	85.78	**85.80**
Hyperpartisan	8 000	88.66	88.72	88.66	**89.05**
	16 000	86.24	86.51	88.03	**88.58**
	32 000	86.39	86.95	89.17	**89.74**

- **Globalement, scores assez proches les uns des autres**
  - Impact limité de la taille du vocabulaire
  - DAPT et entraînement à partir de zéro sont quasi-équivalents
  - Léger gain apporté par l'ajout de vocabulaire, que les plongements soient initialisés aléatoirement ou en fonction du vocabulaire initial
  - Avantage systématique pour la stratégie d'initialisation VIPI

LLM POUR UN DOMAINE DE SPÉCIALITÉ : LA QUESTION DES CONNAISSANCES

INJECTER DES CONNAISSANCES DANS LES LLM

UNE AUTRE APPROCHE POUR L'ADAPTATION DE LLM

- **Injection de connaissances dans les LLM**
  - Problématique générale visant à enrichir les LLM, en particulier pour faciliter les tâches impliquant des inférences
    - Beaucoup de travaux dans le domaine général, avec des bases de connaissances telles que Wikidata

- **Domaines de spécialité**
  - Souvent associés à des ontologies ou des bases de connaissances, a minima des ressources terminologiques
    - Ex. : UMLS dans les domaines médical et biomédical

  ➢ Adapter un LLM à un domaine de spécialité en y injectant les connaissances associées à ce domaine
    - A priori, complémentaire d'une approche de type DAPT
      - En particulier, certains types de relations, comme les relations paradigmatiques, sont peu explicités dans les textes mais très présentes dans les bases de connaissances

UNE VARIANTE ÉLÉMENTAIRE DU DAPT

- **Approche générale**
  - Transformation de la base de connaissances à injecter en un ensemble de phrases
  - Application de l'approche DAPT sur le corpus généré à partir de la base de connaissances

- **Application au domaine médical : (Roy & Pan, 2021)**
  - Base de connaissances : UMLS
  - Génération de 1,6 million de phrases
    - Ex. : (fever, may_be_treated_by, ibuprofen) → fever may be treated by ibuprofen
  - Évaluation sur la tâche d'extraction de relations médicales 2010 i2b2/VA
    - Performances en F1
      - Modèle ClinicalBERT : 0,9127
      - Modèle ClinicalBERT + DAPT UMLS : 0,9078
    - ➢ Injection pas nécessairement source de gains sur une tâche applicative

UN EXEMPLE D'INJECTION PLUS PROFONDE (EL BOUKKOURI ET AL., 2022)

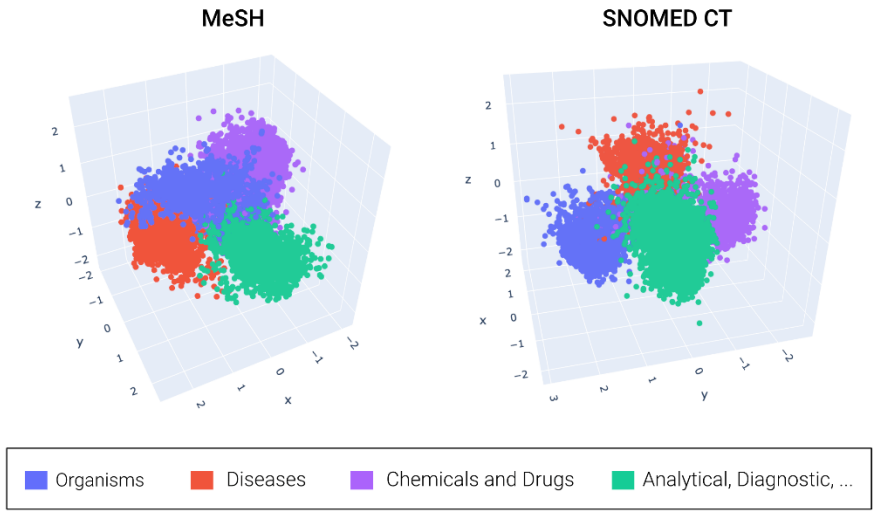
- **Principe**
  - LLM → plongements T_{emb} pour le texte au niveau de chacune de ses couches
  - Base de connaissances → plongements T_{kb}
    - Base de connaissances : ensemble de concepts et de relations binaires entre ces concepts
  - Au niveau de chaque couche du LLM, association de T_{emb} et T_{kb} par une concaténation
  - Entraînement du modèle
    - Point de départ : LLM préentraîné pour le domaine général
    - Préentraînement sur un corpus du domaine cible dans lequel les concepts de la base de connaissances sont présents
      - LLM = CharacterBERT (El Boukkouri et al., 2020)
      - Préentraînement : Masked Language Modeling (MLM) et Next Sentence Prediction (NSP)

BASE DE CONNAISSANCES : REPRÉSENTATION

- **Base de connaissances**
  - Domaine médical
  - Sous-ensemble de l'UMLS : terminologies MeSH et SNOMED_CT
    - MeSH : 29,738 concepts ; SNOMED_CT : 389,872 concepts
  - Relations de type is_a : relations de nature hiérarchique
    - A priori, les plus complémentaires / texte
- **Représentation des connaissances**
  - 1 seul type de relations → pas de nécessité de prendre en compte le type des relations
  - ➢ Utilisation d'un algorithme de plongement de graphe : *node2vec* (Grover & Leskovec, 2016)
    - 1 plongement / terminologie

ACP appliquée aux plongements pour chaque base de connaissances :

recoupement avec 4 grands types de concepts médicaux



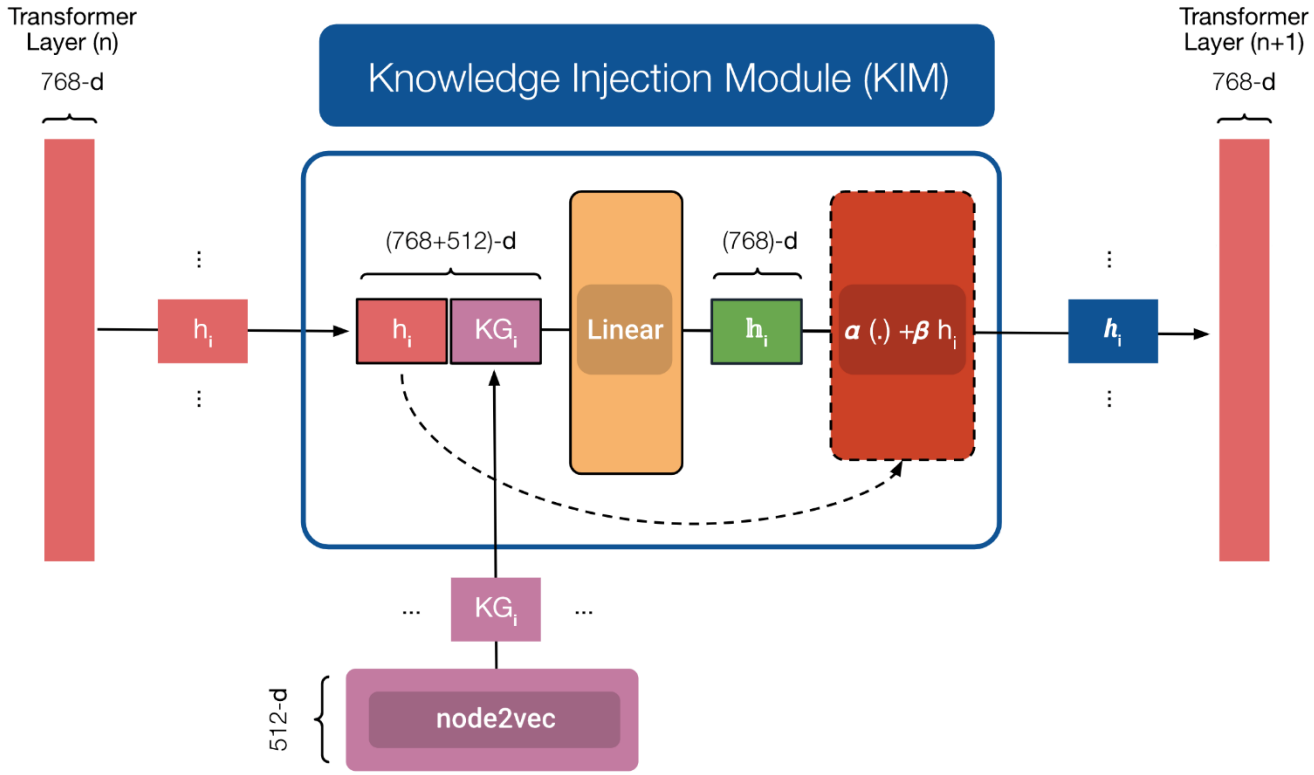
BASE DE CONNAISSANCES : LIEN AVEC LE TEXTE

- **Liage référentiel (*entity linking*)**
  - Identification dans le texte des mentions des concepts
  - En pratique, appariement strict en termes de chaîne de caractères entre
    - les différentes formes linguistiques des concepts dans l'UMLS et
    - les mots du texte
  - Même concept associé à tous les tokens couvrant une mention du concept

- **Représentation interne d'un token**
  - Concaténation du plongement issu du LLM originel (h_i) et du plongement du concept associé (KG_i)
  - Plongement du concept associé
    - Vecteur nul si pas de concept associé
    - Concaténation des plongements MeSH et SNOMED dans le cas contraire
      - Si concept présent dans une seule base de connaissances → vecteur nul pour le plongement de l'autre base

INJECTION DES CONNAISSANCES

- **Modules KIM**
  - Insertion après chaque couche du LLM
    - Concaténation des représentations issues du texte et des représentations issues des bases de connaissances
    - Combinaison linéaire entre cette concaténation et la représentation textuelle



CADRE D'ÉVALUATION

- **Tâches**
  - Extraction de concepts médicaux
    - i2b2
    - *BC5-chemical/disease*
  - Inférences inter-phrastiques
    - MedNLI
    - *BIOSSES, ClinicalSTS* : similarité de phrases
      - Mesure d'évaluation : corrélation de Pearson
  - Extraction de relations
    - DDI
    - ChemProt

CADRE D'ÉVALUATION

- **Modèles testés**
  - Modèle de base : CharacterBERT (CBERT)
  - Modèle adapté au domaine médical par DAPT : $CBERT_{med}$
    - Corpus d'adaptation : MIMIC-III et PubMed
  - Injection de connaissances par concaténation externe de plongements au niveau des tokens
    - [$Emb(CBERT_{med})$, Emb(KB)]
      - Emb(KB) : produit de la même façon que pour les KIM
  - Injection interne de connaissances via les KIM
    - $CBERT(KIM)_{med}$

ÉVALUATION

Modèle	i2b2	BC5 Disease	BC5 Chemical	ChemProt	DDI	BIOSSES	ClinicalSTS	MedNLI
CBERT	88,08	80,89	88,74	70,57	79,39	90,58	84,49	78,86
$CBERT_{med}$	**89,83**	83,60	92,06	**73,61**	80,61	87,52	83,63	**84,64**
[Emb($CBERT_{med}$), Emb(KB)]	89,68	83,90	**92,39**	73,29	**81,54**	87,36	82,84	84,61
$CBERT(KIM)_{med}$	89,77	**85,03**	92,08	73,01	79,35	**92,65**	**84,42**	84,32

- **CharacterBERT vs modèles adaptés**
  - Intérêt de l'adaptation au domaine, quelle que soit la méthode
- **Parmi les modèles adaptés**
  - Tendance générale difficile à dégager
    - Chaque méthode domine sur à peu près le même nombre de jeux de données
  - Intérêt de l'injection interne de connaissances pour les tâches de similarité de phrases
    - Pas de transposition à l'inférence entre phrases
  - Pas de tendance claire pour les tâches d'extraction de concepts et de relations

EN CONCLUSION

SYNTHÈSE

- **3 grandes méthodes d'adaptation des LLM à un domaine cible**
  - Corpus représentatif du domaine + poursuite du préentraînement du LLM sur ce corpus
  - Ajout au LLM du vocabulaire le plus représentatif du domaine cible
  - Injection/association de connaissances caractéristiques du domaine cible dans le/au LLM
- **Caractère central du préentraînement sur un corpus du domaine**
  - Cœur de la première méthode
  - Généralement utilisé par les deux autres méthodes pour entraîner les ajouts faits aux modèles
- **Intérêt de l'adaptation**
  - Gain de performance assez systématique, mais variable selon les cas
    - Ajout de vocabulaire + préentraînement sur corpus du domaine a priori la meilleure option
    - Injection de connaissances : pas systématiquement intéressante ; peut-être à réserver à certaines tâches
  - Option intéressante par rapport à un entraînement à partir de zéro pour le domaine cible

POUR ALLER PLUS LOIN

- **Mieux caractériser les conditions d'une bonne adaptation**
  - Caractérisation en fonction de la « distance » entre le LLM initial, le corpus utilisé pour représenter le domaine cible et les données de test
    - Comment anticiper le gain attendu sur une tâche en fonction de ces trois éléments et de leur degré de proximité ?
    - LLM initial : corpus d'entraînement non nécessairement connu ou accessible (problème de droit, de taille…)

- **Adaptation des gros LLM (BLOOM, Llama…)**
  - Affinage reposant généralement sur des méthodes PEFT (Parameter Efficient Fine-Tuning)
    - Adapters, LoRA (low rank adapter)…
  - Adaptation par préentraînement sur corpus du domaine → opération coûteuse devant s'appuyer sur des méthodes PEFT
    - cf. (Chronopoulou et al., 2022) ou AdapterSoup (Chronopoulou et al., 2023)
  - Comment combiner les adapteurs dédiés au domaine et ceux dédiés à la tâche ?
    - AdapterHub (Pfeiffer et al., 2020) ; AdapterFusion (Pfeiffer et al., 2021)

Merci de votre attention

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT, 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list,cea,fr

Établissement public à caractère industriel | RCS Paris B 775 685 019

Bibliographie

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021.
Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training.
In *2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022.
Efficient hierarchical domain adaptation for pretrained language models.
In *2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, Alexander Fraser, and Jesse Dodge. 2023.
AdapterSoup: Weight averaging to improve generalization of pretrained language models.
In *Findings of the Association for Computational Linguistics : EACL 2023*, pages 2054–2063, Dubrovnik, Croatia. Association for Computational Linguistics.

Bibliographie

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.
In *29ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2022)*, pages 215–225, Avignon, France.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
In *2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bibliographie

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 *28th International Conference on Computational Linguistics (COLING 2020)*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022a.
Re-train or train from scratch? comparing pre-training strategies of bert in the medical domain.
In *13th Language Resources and Evaluation Conference (LREC 2022)*, pages 2626–2633, Marseille, France. European Language Resources Association.

Bibliographie

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022b.
Specializing static and contextual embeddings in the medical domain using knowledge graphs: Let's keep it simple.
In *13th International Workshop on Health Text Mining and Information Analysis (LOUHI 2022)*, pages 69–80, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016.
node2vec : Scalable feature learning for networks.
In *22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

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, 3(1).

Bibliographie

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020.
Don't Stop Pretraining: Adapt Language Models to Domains and Tasks.
In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019.
Parameter-efficient transfer learning for nlp.
In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Vladislav Mosin, Igor Samenko, Borislav Kozlovskii, Alexey Tikhonov, and Ivan P. Yamshchikov. 2023.
Fine-tuning transformers: Vocabulary transfer.
Artificial Intelligence, 317 :103860.

Sinno Jialin Pan and Qiang Yang. 2010.
A Survey on Transfer Learning.
IEEE Transactions on Knowledge and Data Engineering, 22(10).

Bibliographie

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019.
Transfer learning in biomedical natural language processing : An evaluation of bert and elmo on ten benchmarking datasets.
In *2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021.
AdapterFusion: Non-destructive task composition for transfer learning.
In *16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020.
AdapterHub: A framework for adapting transformers.
In *2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Bibliographie

Barbara Plank. 2016.
What to do about non-standard (or non-canonical) language in nlp.
In *13th Conference on Natural Language Processing (KONVENS 2016)*, pages 13–20.

Alan Ramponi and Barbara Plank. 2020.
Neural Unsupervised Domain Adaptation in NLP—A Survey.
In *28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arpita Roy and Shimei Pan. 2021.
Incorporating medical knowledge in BERT for clinical relation extraction.
In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder. 2019.
Neural Transfer Learning for Natural Language Processing.
Ph.D. thesis, National University of Ireland, Galway.