Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Stratégies d'apprentissage actif pour la reconnaissance d'entités nommées en français

Marco Naguib[1] Aurélie Névéol[1] Xavier Tannier[2]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay cedex, France
(2) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, 75006 Paris, France
marco.naguib@lisn.upsaclay.fr, aurelie.neveol@lisn.upsaclay.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

L'annotation manuelle de corpus est un processus coûteux et lent, notamment pour la tâche de reconnaissance d'entités nommées. L'apprentissage actif vise à rendre ce processus plus efficace, en sélectionnant les portions les plus pertinentes à annoter. Certaines stratégies visent à sélectionner les portions les plus représentatives du corpus, d'autres, les plus informatives au modèle de langage. Malgré un intérêt grandissant pour l'apprentissage actif, rares sont les études qui comparent ces différentes stratégies dans un contexte de reconnaissance d'entités nommées médicales. Nous proposons une comparaison de ces stratégies en fonction des performances de chacune sur 3 corpus de documents cliniques en langue française : MERLOT, QuaeroFrenchMed et E3C. Nous comparons les stratégies de sélection mais aussi les différentes façons de les évaluer. Enfin, nous identifions les stratégies qui semblent les plus efficaces et mesurons l'amélioration qu'elles présentent, à différentes phases de l'apprentissage.

ABSTRACT

Sampling strategies in active learning for named entity recognition in French

Manual corpus annotation for NLP can be labor intensive and expensive. Active learning aims to achieve high accuracy with fewer training data by allowing a model to select the data to be annotated and used for learning. Some sampling strategies aim to select the most representative instances, while others aim to capture the instances that are most informative for the language model. Despite a growing interest in active learning in recent years, few studies provide a thorough comparison between those strategies in the context of medical named entity recognition. In this work, we apply these stategies on 3 French corpora in the clinical domain : MERLOT, QuaeroFrenchMed, and E3C. We provide an extensive comparison of those strategies and discuss various ways of evaluating them. Finally we determine those which seem most effective, and measure the estimated improvement they can provide at different stages of the training process.

MOTS-CLÉS : Reconnaissance d'entités nommées ; Documents cliniques ; Apprentissage actif.

KEYWORDS: Named entity recognition ; Clinical narratives ; Active learning.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 232–247.

1 Introduction

L'apprentissage supervisé repose sur l'hypothèse de la disponibilité de données annotées de haute qualité. Or, la collecte et l'annotation de telles données peuvent s'avérer coûteuses en ressources et en temps (Fort *et al.*, 2012; Grouin *et al.*, 2014). Ceci est particulièrement vrai dans le domaine des textes cliniques, où la tâche d'annotation doit être confiée à des personnes faisant preuve d'un haut niveau d'expertise (Campillos *et al.*, 2017). La question d'efficacité en données est donc primordiale. L'*active learning* (Lewis & Gale, 1994) ou apprentissage actif propose d'augmenter l'efficacité d'un algorithme d'apprentissage supervisé, en lui permettant d'interagir directement avec la source de données (souvent l'annotateur). On l'oppose à un algorithme d'apprentissage supervisé, dit « passif », où l'intégralité de la supervision, à savoir, des données annotées servant au processus d'apprentissage, est disponible avant l'exécution de l'algorithme. L'*active learning* propose, quant à lui, de faire intervenir l'algorithme d'apprentissage dans le processus de sélection des données à annoter, avant qu'elles ne le soient. L'objectif étant de parvenir à sélectionner les données les plus pertinentes pour l'apprentissage. Ici, nous nous intéressons au schéma d'*active learning* dit *pool-based* (Cheng *et al.*, 2013), dans lequel les données non annotées sont toutes disponibles avant l'apprentissage, et l'algorithme d'apprentissage cherche à estimer la pertinence de chaque portion de données. Ceci imite le contexte d'usage clinique, où disposer de bases de données non annotées de taille raisonnable peut se montrer bien plus facile que d'annoter ces données (Névéol *et al.*, 2014).

Si l'*active learning* connaît du succès dans le domaine de l'image ou la classification de texte, il n'y a pas de façon standard de l'employer, ni de l'évaluer, dans le contexte des tâches de prédiction structurée, comme la reconnaissance d'entités nommées. En particulier, de nombreuses stratégies de sélection d'exemples existent mais aucune d'entre elles ne s'impose, faute de succès. De plus, malgré l'intérêt grandissant pour l'*active learning*, les études l'appliquant à la reconnaissance d'entités nommées sont rares. Par exemple, la dernière en date s'intéressant au français est Claveau & Kijak (2015). Nous nous intéressons ici à appliquer, combiner et comparer quelques unes de ces stratégies, sur trois corpus de documents cliniques en langue française : MERLOT (Campillos *et al.*, 2017), QuaeroFrenchMed (Névéol *et al.*, 2014) et E3C (Magnini *et al.*, 2021). Les contributions de cet article sont les suivantes :

— Nous fournissons une comparaison approfondie des stratégies classiques de sélection d'exemples dans le domaine médical.
— Nous discutons de différentes façon d'évaluer l'*active learning* dans le cadre de la reconnaissance d'entités nommées.
— Nous montrons que deux stratégies simples basées sur la similarité de vocabulaire se révèlent les plus efficaces, et nous mesurons l'amélioration apportée par celles-ci.
— Nous mettons à disposition l'ensemble des scripts utilisés dans nos expériences [1].

2 État de l'art

L'*active learning* a été étudié depuis plus de vingt ans, (Cohn *et al.*, 1994a,b; Lewis & Catlett, 1994; Lewis & Gale, 1994). Zhang *et al.* (2022b) observent que si le nombre de publications étudiant l'*active learning* a connu un pic entre 2008 et 2010, cet intérêt semble avoir diminué entre 2011 et 2019, années de l'essor du *deep learning*. Plus récemment, on observe un regain d'intérêt

1. https://github.com/marconaguib/active-nlstruct

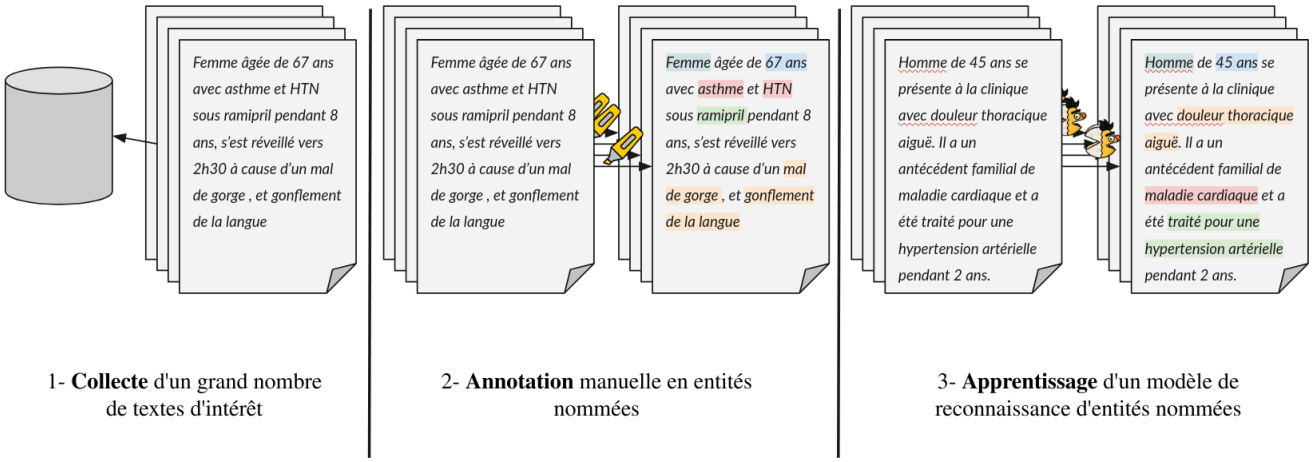

FIGURE 1 – Apprentissage supervisé (« passif »)

pour l'*active learning* qui se traduit dans l'*ACL anthology* par une montée du nombre d'articles l'étudiant en 2020. Ce regain d'intérêt pour l'*active learning* vise essentiellement à le combiner au *deep learning* (Ren *et al.*, 2021; Zhan *et al.*, 2022), et ainsi à rendre l'apprentissage des modèles plus efficace.

Une boucle de sélection, d'annotation et d'apprentissage Pour réaliser une tâche comme la reconnaissance d'entité nommées, la solution que propose généralement le *deep learning* est l'apprentissage supervisé. Elle consiste en 3 grandes étapes (cf. figure 1). A l'étape 1, on collecte un grand nombre de textes du domaine d'intérêt. L'étape 2 consiste à annoter manuellement ces textes en entité nommées. Cette étape peut selon les domaines nécessiter plus ou moins de compétence, et peut ainsi s'avérer plus ou moins coûteuse. Enfin, cette base de données ainsi annotée sert pour l'apprentissage d'un modèle de reconnaissance d'entités nommées à l'étape 3.

L'*active learning* part de l'intuition que tous les textes à annoter n'ont pas la même pertinence. Ainsi, il vise à sélectionner ceux qui sont les plus pertinents à annoter, pour réduire ce coût d'annotation et augmenter l'efficacité de l'entraînement. Il propose ainsi de remplacer les étapes 2 et 3 par une boucle (cf. figure 2) où l'on sélectionne d'abord un petit nombre d'exemples qu'on estime les plus pertinents à annoter (2a). Ceux-ci sont ensuite annotés (2b) et utilisés pour démarrer l'entraînement (2c), avant de sélectionner à nouveau un petit nombre d'exemples etc.

Stratégies de sélection Il existe de nombreuses façon d'estimer la pertinence des exemples (et ainsi sélectionner ceux qui sont les plus pertinents). On peut distinguer deux familles de stratégies de sélection.

1. Les stratégies d'**informativité** visent à repérer les exemples qui semblent les plus informatifs pour le modèle.

   Il est courant d'estimer l'informativité de chaque exemple par l'**incertitude** qu'exprime le modèle face à celui-ci. Ainsi, les exemples les plus pertinents à annoter seraient ceux pour lesquelles le modèle est le plus incertain (Lewis & Gale, 1994). Dans un modèle de classification probabiliste, appliquer cette méthode revient à calculer la distribution de probabilité pour chaque exemple sur les différentes classes possibles, puis de trier les exemples selon l'entropie de cette distribution (Tang *et al.*, 2002; Chen *et al.*, 2006; Zhu & Hovy, 2007),

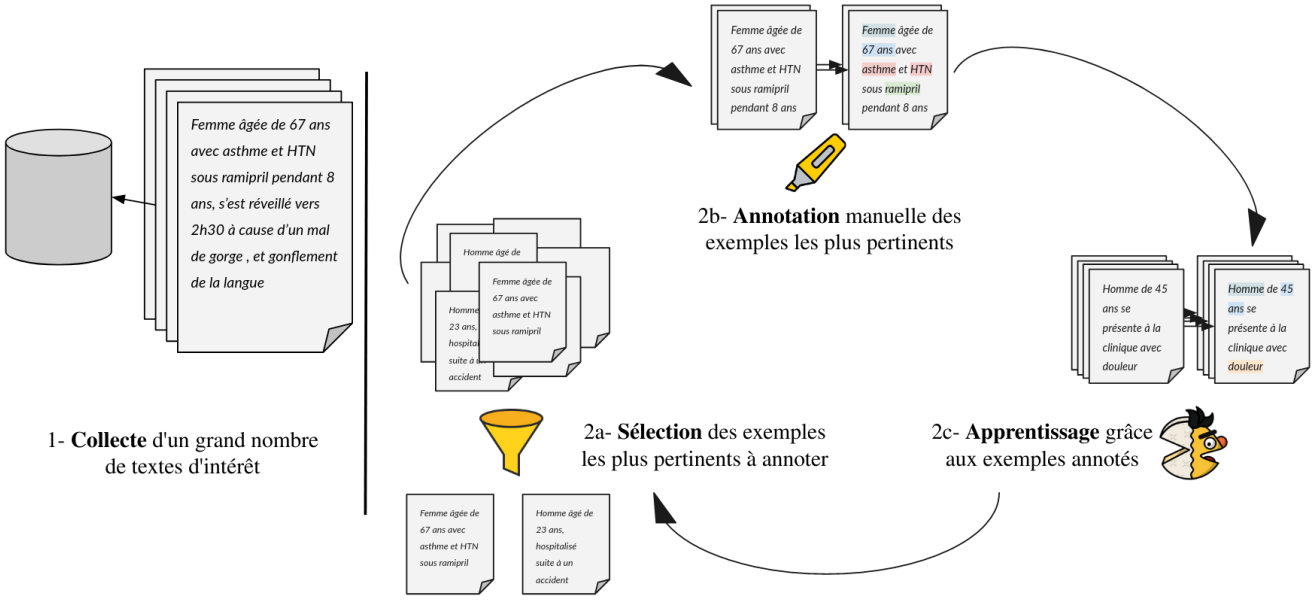

FIGURE 2 – Aprentissage actif

selon la probabilité attribuée à la classe la plus probable (Lewis & Gale, 1994; Culotta & McCallum, 2005) ou selon l'écart de probabilité entre les deux classes les plus probables (Scheffer *et al.*, 2001; Schein & Ungar, 2007).

Il n'existe pas de façon standard de **transposer** cette approche d'incertitude à la reconnaissance d'entités nommées. En effet, cette dernière associe à chaque séquence d'entrée (des tokens) une séquence de prédictions, ayant chacun une incertitude séparée. Ainsi, il faut agréger ces incertitudes en une incertitude globale à l'échelle de l'exemple (phrase ou document), pour pouvoir décider de l'informativité et ainsi la pertinence globale de celle-ci. Culotta & McCallum (2005) proposent de calculer d'abord la « confiance » du modèle en une prédiction $c_i = 1 - u_i$ où u_i est son incertitude pour le $i^{ème}$ élément de la séquence d'entrée, puis de calculer l'incertitude globale $\mathcal{I} = 1 - \Pi_{i=1}^{n} c_i$. Cette méthode a tendance à préférer les séquences longues. Tang *et al.* (2002) et Shen *et al.* (2018) proposent en revanche de moyenner les incertitudes pour les normaliser par la taille de la séquence.

Il existe par ailleurs d'autres approches pour estimer l'informativité. La **divergence locale** consiste à examiner les prédictions du modèle dans la région locale de chaque exemple grâce à la recherche des voisins les plus proches (Margatina *et al.*, 2021), à des perturbations locales (Zhang *et al.*, 2022a) ou à l'augmentation de données (Jiang *et al.*, 2020). Le **désaccord multi-modal** (Shen *et al.*, 2018; McCallum & Nigam, 1998; Houlsby *et al.*, 2011) vise, lui, à entraîner un « comité » de modèles et se servir du désaccord entre eux.

2. Les stratégies de **représentativité** visent à prendre en compte la similarité des exemples entre eux, pour éviter la redondance et la sélection d'intrus, problèmes auxquels les approches d'informativité peuvent être vulnérables (Roy & McCallum, 2001; Karamcheti *et al.*, 2021).

   Par exemple, l'approche de **densité** vise à sélectionner les exemples qui présentent un vocabulaire le plus similaire en moyenne à celui de tous les autres exemples. Cette similarité peut être mesurée par la fréquence de mots ou n-grams (McCallum & Nigam, 1998; Settles & Craven, 2008; Zhao *et al.*, 2020)

   L'approche de **diversité**, elle, vise à éviter la redondance, en sélectionnant des exemples variés qui représentent la variété de l'ensemble des entrées. Ainsi, l'on peut présenter des

exemples qui présentent le moins de similarité entre eux (Brinker, 2003), et/ou le moins de similarité avec les données déjà annotées (Eck *et al.*, 2005; Bloodgood & Callison-Burch, 2010; Erdmann *et al.*, 2019). Cette sélection peut se faire de façon itérative (Shen *et al.*, 2004; Geifman & El-Yaniv, 2017; Sener & Savarese, 2018) ou on employant la classification non supervisée (Zhdanov, 2019; Yu *et al.*, 2022).

3. Ces deux familles d'approches ne sont pas incompatibles et l'on peut les **combiner** de plusieurs façons. Par exemple, Chen *et al.* (2011) proposent de calculer les « scores de pertinence » de chaque exemple, puis de calculer simplement une somme pondérée de ces scores. Mirroshandel & Nasr (2011) et Tang *et al.* (2002) appliquent d'abord une première stratégie pour sélectionner un sous ensemble des données, puis en utilisent un autre pour sélectionner les exemples. On peut également procéder à des combinaisons dynamiques. Ambati *et al.* (2011) et Wu *et al.* (2017) proposent de reposer d'abord sur une approche de représentativité, puis, au fur et à mesure de l'apprentissage, passer à une approche d'informativité.

Dans ce travail, nous examinons des stratégies d'incertitude, de densité et de diversité, ainsi qu'une concaténation simple de ces deux dernières, dans le contexte de la reconnaissance d'entités nommées médicales.

3 Expérimentation

Corpus utilisés Dans ce travail, nous étudions le comportement de l'*active learning* sur trois corpus cliniques annotés en entités nommées. Le tableau 1 indique les différent types d'entités dans chaque corpus, ainsi que leurs nombres.

1. **MERLOT** (Campillos *et al.*, 2017) est un corpus privé de documents cliniques pseudonymisés collectés d'hôpitaux français. Ils couvrent différents types de documents : compte-rendus de prise en charge, compte-rendus post-opératoires, et courriers dans le domaine de la gastro-entérologie. Le corpus est annoté en 21 types d'entités nommées. Cependant, nous choisissons de suivre les remplacements indiqués par Bannour *et al.* (2022), pour n'obtenir ainsi que 15 types d'entités.
2. **QuaeroFrenchMed** (Névéol *et al.*, 2014) est un corpus composé de deux parties que nous traitons séparément. La première partie, **EMEA** est une collection de 13 notices patient concernant des médicaments commercialisés en Europe, fournis par l'Agence Européenne des Médicaments. La seconde partie **MEDLINE**, consiste en 2500 titres d'articles scientifiques indexés dans la base de données MEDLINE [2]. Ces deux parties sont annotées en 10 types d'entités nommées.
3. **E3C** (Magnini *et al.*, 2021) est un corpus européen de cas cliniques. Nous utilisons la partie française, composée de 1615 cas cliniques collectionnés du domaine public : articles scientifiques indexés sur PubMed [2] et articles scientifiques sous license CC-by. Il est annoté en 6 types d'entités nommées cliniques, dont un type : entité clinique (`CLINENTITY`) que nous faisons le choix de désagréger en sous-types pour avoir un schéma d'annotation avec une diversité se rapprochant de celui des autres corpus. Chaque entité de ce type est annotée par un attribut `CUI` qui désigne l'identifiant unique du concept associé dans le métathésaurus UMLS (Unified Medical Language System). Grâce à cet identifiant, nous pouvons ainsi récupérer le type sémantique (McCray *et al.*, 2001) correspondant (pathologie, symptôme...).

2. http://pubmed.ncbi.nlm.nih.gov/

MERLOT	EMEA	MEDLINE	E3C
ANAT (4449)	ANAT (583)	ANAT (1499)	ACTOR (427)
CHEM (1374)	CHEM (2482)	CHEM (1055)	Acquired_Abnormality (15)
Concept_Idea (2964)	DEVI (170)	DEVI (128)	Anatomical_Abnormality (25)
DEVI (1068)	DISO (1510)	DISO (2843)	Bacterium (1)
DISO (4593)	GEOG (65)	GEOG (131)	BODYPART (659)
DOSE (928)	LIVB (817)	LIVB (941)	Cell_or_Molecular_Dysfunction (7)
Genes_Proteins (5)	OBJC (174)	OBJC (100)	Congenital_Abnormality (11)
Hospital (806)	PHEN (62)	PHEN (158)	Disease_or_Syndrome (374)
LIVB (3921)	PHYS (344)	PHYS (469)	EVENT (1100)
Localization (840)	PROC (952)	PROC (1750)	Finding (178)
MEAS (5322)			Injury_or_Poisoning (24)
MODE (252)			Mental_or_Behavioral_Dysfunction (22)
PHEN (905)			Neoplastic_Process (99)
PROC (8291)			Pathologic_Function (149)
TEMP (3940)			RML (508)
			Sign_or_Symptom (179)
			TIMEX3 (333)
			Virus (1)

TABLE 1 – Types d'entités présents dans les différents corpus

En appliquant l'*active learning* au traitement automatique des langues, se pose la question de l'unité documentaire considérée, et donc du découpage du texte. Même si les corpus sont naturellement découpés par document (sauf MEDLINE où chaque titre est un document séparé), il est plus courant dans l'*active learning* de considérer un découpage par phrase (Chen *et al.*, 2015). En effet, toutes les phrases d'un document n'ont pas la même pertinence, et on peut préférer en sélectionner certaines et pas d'autres. De plus, les approches d'informativité se basent souvent sur les représentations et les prédictions faites par le modèle pour chaque mot. Considérer les documents dans leur globalité pourrait donc noyer l'information que le modèle peut émettre à l'échelle des mots. C'est d'ailleurs ce qui encourage Radmard *et al.* (2021) à découper les corpus en n-grammes et évaluer l'informativité de chacun. Nous en restons à un découpage par phrases pour des raisons d'efficacité. Nous procédons donc à ce découpage pour MERLOT, EMEA et QuaeroFrenchMed, grâce à une expression régulière permettant une séparation aux ponctuations fortes et aux multiples sauts à la lignes.

Stratégies de sélection Dans ce travail, nous examinons et comparons plusieurs stratégies de sélection comme suit.

— `random` sélectionne aléatoirement des exemples.

— `common_vocab` identifie les 500 n-grammes les plus fréquents dans le corpus, puis trie les exemples selon le nombre de n-grammes fréquents qu'ils contiennent. Elle est inspirée de Zhao *et al.* (2020), qui l'appliquent à la tâche de traduction.

— `diverse_vocab` sélectionne un exemple de façon aléatoire, puis, itérativement, sélectionne l'exemple qui présente le plus de *n-grams* non déjà vus dans les exemples déjà sélectionnés. Elle est inspirée de Kirsch *et al.* (2019).

— `diverse_pred` sélectionne un exemple de façon aléatoire, puis, itérativement, sélectionne l'exemple qui présente le plus de types d'entités prédites, non déjà prédites dans les exemples précédemment sélectionnés.

— `uncertainty_mean_min3` calcule les confiances du modèle pour chaque entité prédite [3], puis trie les exemples selon la confiance moyenne croissante. Pour éviter les phrases « trop courtes », on restreint le choix aux exemples présentant plus de 3 entités prédites. Elle

3. Pour mieux tenir compte des entités imbriquées, notre modèle classifie chaque partie (chaque *span*) de l'exemple. Ainsi, ce sont les confiances en ces prédictions qu'on moyenne sur l'exemple et qu'on trie.

est inspirée de Shen *et al.* (2018).

Il est important de noter que `common_vocab` et `diverse_vocab` ne dépendent pas des prédictions du modèle, et peuvent donc être appliquées avant l'entraînement. `diverse_pred` et `uncertainty_mean_min3`, en revanche, nécessitent une inférence sur toutes les données non-annotées du corpus, l'annotateur doit donc attendre l'entraînement du modèle. En pratique cependant, on peut imaginer un scénario où l'annotateur a une longueur d'avance par rapport à l'entraînement du modèle : à la première étape, le modèle sélectionne 2 *batchs* d'exemples à annoter, l'annotateur annote d'abord le premier *batch* et le soumet, puis, pendant qu'il annote le deuxième *batch*, le modèle commence l'entraînement sur le premier, etc.

Déroulement de la simulation Pour examiner ces stratégies, nous simulons une boucle d'*active learning*. Pour ce faire, le jeu d'entraînement est initialisé à l'ensemble vide, puis, à chaque itération, on y ajoute seulement les documents sélectionnés par la stratégie en question.

Afin de ne pas multiplier les expériences, nous utilisons un unique modèle pour évaluer toutes ces stratégies, NLStruct [4]. Il s'agit d'un Bi-LSTM-CRF, combiné à un modèle de langue CamemBERT-base (Martin *et al.*, 2020), il est décrit en détail par Wajsbürt (2021). Nous utilisons tous les hyperparamètres proposés par défaut.

Nous fixons le nombre de phrases annotées à chaque itération à 10. Ce qui correspond par exemple à 10-20 minutes en moyenne dans le corpus MEDLINE. À chaque itération, la stratégie en question sélectionne donc 10 phrases du corpus, et l'on simule l'annotation manuelle en dévoilant les annotations *gold standard* concernant ces phrases. On applique d'abord cette stratégie $k = 2$ fois pour choisir 20 phrases comme jeu de validation. Puis pendant 10 itérations, on intègre les phrases annotées dans le jeu d'entraînement et entraîne le modèle sur l'ensemble des données annotées. Nous discutons de ce choix de 10 itérations dans la partie 5. Chaque entraînement consiste en 1000 étapes d'optimisation avec arrêt prématuré si aucune amélioration sur le jeu de validation n'est observée pendant 300 étapes.

Mesures Nous mesurons les performances du modèle en fonction de l'effort à fournir par l'annotateur au fil des itérations. Afin de mesurer précisément l'effort annotateur, nous choisissons de rapporter **le nombre de mots annotés** (Chen *et al.*, 2015). Quant à la mesure de performance, nous rapportons la mesure classique du score f_1 **toutes classes confondues**, que nous appelons f_1^{micro}. De plus, pour tenir compte de la performance sur les classes rares, nous rapportons également la **moyenne simple des scores** f_1 obtenus sur chaque classe séparément, sans pondération. Nous l'appelons f_1^{macro}. La figure 3 rapporte l'évolution de ces 2 scores au fil de l'apprentissage. Chaque courbe représente la moyenne du score sur 3 graines aléatoires utilisées, dans un intervalle de confiance à 95%.

Nous rapportons également dans le tableau 2 pour chaque corpus c et stratégie s une quantité que nous appelons **Performance relative à** 1000 **mots** ou $\mathcal{P}_{1000}(c, s)$. Cette quantité mesure le ratio entre un score de performance obtenu en entraînant un modèle sur au moins 1000 mots du corpus c choisis selon la stratégie s, et celui obtenu en entraînant le modèle sur l'ensemble de c.

4. https://github.com/percevalw/nlstruct

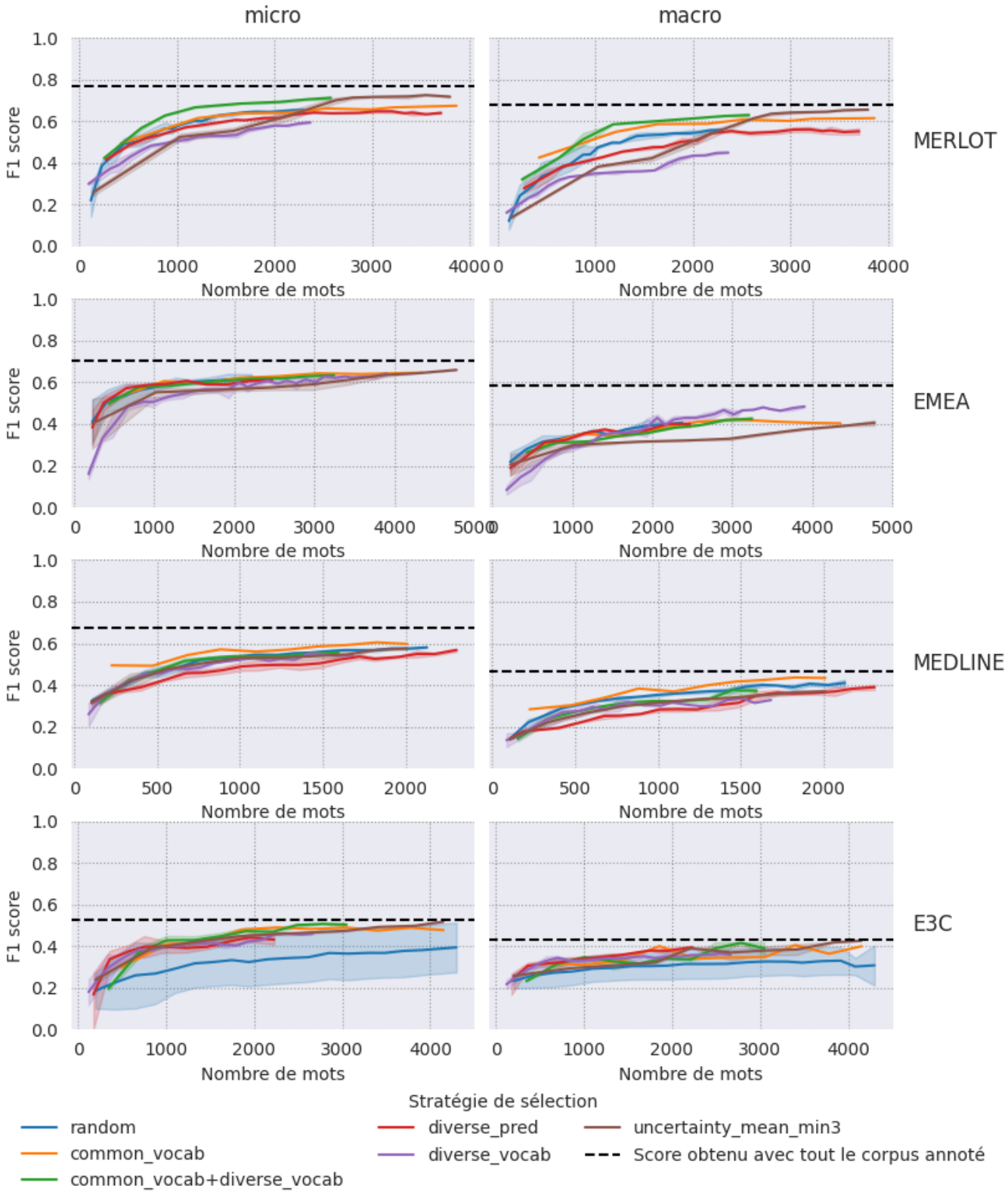

FIGURE 3 – Courbes d'évolution de f_1^{micro} et f_1^{macro} en fonction du nombre de mots annotés. Noter que les courbes ne s'arrêtent pas toutes à la même abscisse. En effet, les stratégies sélectionnent des séquences de tailles différentes. `common_vocab+diverse_pred` désigne la concaténation des deux stratégies : chacune d'entre elles sélectionne 5 exemples.

Stratégie	Corpus							
	MERLOT		EMEA		MEDLINE		E3C	
`random`	0.76	(0.70)	0.83	(0.60)	0.81	(0.77)	0.56	(0.67)
`common_vocab`	0.80	(0.81)	**0.86**	(0.60)	**0.83**	(**0.79**)	0.79	(0.76)
`diverse_vocab`	0.69	(0.61)	0.74	(0.50)	0.78	(0.64)	0.76	(0.76)
`diverse_pred`	0.71	(0.58)	0.85	(**0.61**)	0.73	(0.60)	0.74	(**0.81**)
`common_vocab+diverse_vocab`	**0.87**	(**0.86**)	0.83	(0.54)	0.80	(0.69)	**0.82**	(0.80)
`uncertainty_mean_min3`	0.68	(0.56)	0.79	(0.51)	0.79	(0.68)	0.80	(0.71)

TABLE 2 – Performances relatives à 1000 mots selon f_1^{micro} (f_1^{macro}).

4 Résultats

De prime abord, nous remarquons que des performances raisonnables peuvent être atteintes avec peu de phrases annotées, même quand celles-ci sont tirées aléatoirement. Par exemple, 1000 mots tirés aléatoirement du corpus MERLOT (<1 % du corpus) et annotés sont suffisants pour atteindre 0,61 de score f_1^{micro}, soit 70 % du score atteint en entraînant le même modèle sur l'intégralité du corpus annoté. On observe ce phénomène dans toutes les applications similaires, mais il semble être accentué par la redondance particulière aux documents cliniques (Cohen *et al.*, 2013; Searle *et al.*, 2021). Effectivement, le corpus MEDLINE, composé de titres d'articles scientifiques (on peut donc penser qu'il est particulèrement peu redondant) présente une performance relative à 1000 mots de 77 %. De manière générale, nous observons que dans chaque graphique, une ou plusieurs stratégies ont une meilleure évolution que `random`. Cependant, il n'y a pas une stratégie qui semble la meilleure pour tous les corpus. La stratégie `common_vocab` semble tout de même souvent efficace. Il est intéressant de voir qu'une simple concaténation des méthodes `common_vocab` et `diverse_vocab` peut dans certains cas avoir des meilleurs résultats que chacune d'elles séparément.

Nous observons par ailleurs que la stratégie `uncertainty_mean_min3` qui est la plus adoptée dans la littérature (Chen *et al.*, 2015; Shen *et al.*, 2018) ne semble pas adaptée au contexte de peu de phrases annotées dans lequel nous nous plaçons (cf. partie 5).

Une grande majorité des travaux sur l'*active learning* (Chen *et al.*, 2015; Shen *et al.*, 2018; Liu *et al.*, 2022; Radmard *et al.*, 2021) font le choix de cadrer les graphiques d'évolution de performance sur la partie supérieure, afin de mieux visualiser l'écart entre les performances des différentes méthodes. Ici, nous faisons le choix de les cadrer entre 0 et 1, ce qui permet de voir le caractère marginal des améliorations obtenues grâce aux meilleures stratégies.

5 Discussion et perspectives

Mesure de l'effort Il est courant d'estimer l'effort de l'annotateur par le nombre de phrases ou de mots annotés. Mais certains types d'entités peuvent être plus difficiles à annoter que d'autres. Notamment, les stratégies d'informativité visent précisément à sélectionner les exemples ambigus et difficile à annoter. Aussi, une comparaison juste des stratégies de sélection prendrait-elle cet effort en compte. Fort *et al.* (2012) fournissent une modélisation de cet effort en fonction du schéma d'annotation. C'est donc une perspective d'amélioration que nous considérons.

Une meilleure estimation de l'effort d'annotation peut même guider les stratégies de sélection. En

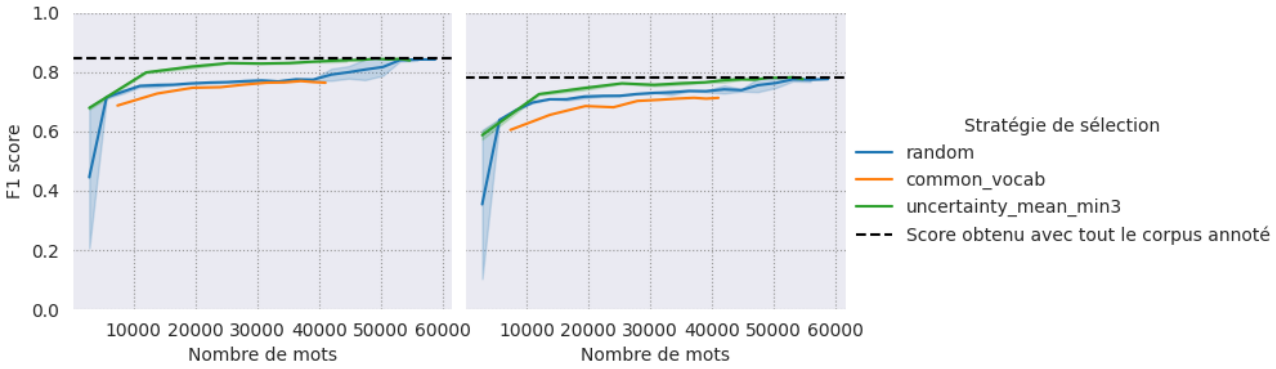

FIGURE 4 – Comparaison à grande échelle des différentes stratégies sur MERLOT

	Corpus			
Stratégie	MERLOT	EMEA	MEDLINE	E3C
`common_vocab`	<1	<1	<1	<1
`diverse_vocab`	3	2	2	<1
`diverse_pred`	105	45	70	55
`uncertainty_mean_min3`	103	42	68	55

TABLE 3 – Comparaison du temps d'exécution moyen en secondes des différentes stratégies

effet, certains travaux (Tomanek & Hahn, 2010; Wei *et al.*, 2019) développent des stratégies qui visent à estimer non seulement la pertinence mais aussi le coût d'annotation de chaque exemple, pour ensuite les trier selon ce que Haertel *et al.* (2008) appellent « retour sur investissement ».

Intérêt et budget d'annotation La plupart des travaux sur l'*active learning* l'étudient dans toutes ses phases (Shen *et al.*, 2018; Radmard *et al.*, 2021). Par exemple, ces derniers l'évaluent pour un nombre de mots annotés allant jusqu'à 1 million, pour l'anglais et le chinois. En revanche, nous nous sommes intéressés au contexte plus naturel d'une centaine de phrases annotées, ce qui vaut entre une et deux heures de travail pour l'annotateur. Pour pouvoir comparer notre implémentation à celles de l'état de l'art, nous procédons à une simulation d'*active learning* sur MERLOT (le plus grand de nos corpus) en fixant le nombre de phrases à annoter à chaque itération à 250 ($\approx$ 5 % du corpus). La figure 4 montre ainsi les courbes d'évolution, et l'on trouve en effet que `common_vocab` n'est plus très intéressante à grande échelle et `uncertainty_mean_min3` devient plus intéressante dans ce cadre. Elle apporte, en effet, des amélioration similaires à celles obtenues par Liu *et al.* (2022) et Zhou *et al.* (2021) Ainsi, nous pouvons émettre l'hypothèse qu'une stratégie de combinaison dynamique qui passe progressivement d'une stratégie de représentativité à une stratégie d'informativité mériterait une attention particulière, dans un prochain travail.

Temps d'attente Le tableau 3 montre la moyenne de temps d'exécution d'une requête de sélection pour chaque stratégie, mesurées sur une carte GeForce GTX 1080 Ti (11 Go). On peut observer que les stratégies qui requièrent une inférence sur l'ensemble des données annotées (à savoir `diverse_pred` et `uncertainty_mean_min3`) sont les plus longues. Ce n'est pas une surprise, mais ce paramètre est très rarement mentionné dans les travaux sur l'*active learning*. Que ce soit en termes de temps d'attente ou d'impact carbone, l'amélioration des performances étant relativement limitée, il est pourtant permis de remettre en question leur utilisation, au profit de stratégies basées sur le vocabulaire, qui ne nécessitent pas d'être relancées à chaque itération.

Classes rares L'*active learning* a fait l'objet d'études pour améliorer la performance sur les classes

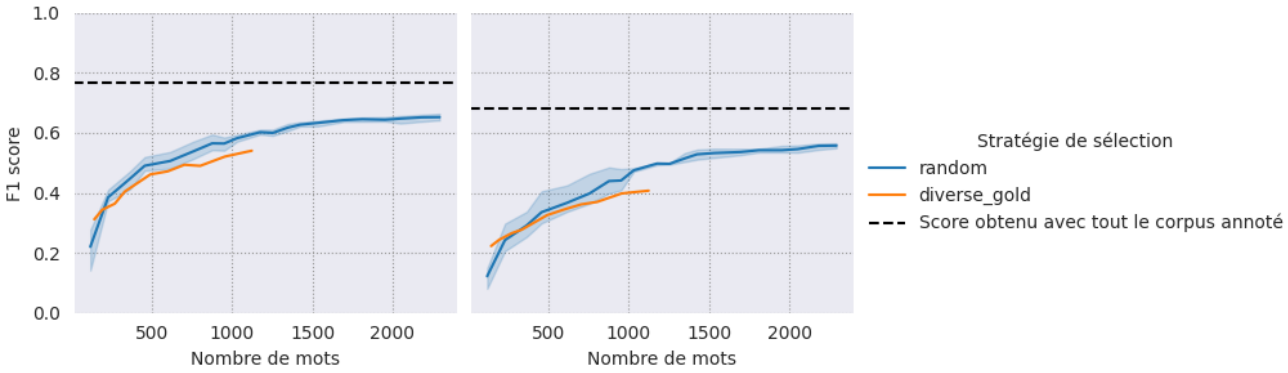

FIGURE 5 – Comparaison entre `random` et `diverse_gold` sur MERLOT

rares (Zhu & Hovy, 2007). Nous avons examiné une stratégie supplémentaire, `diverse_gold` qui imite ce contexte. Elle sélectionne un exemple de façon aléatoire, puis, itérativement, cherche les phrases qui maximisent le nombre d'apparition de l'entité la plus rarement sélectionnée. Cette méthode cherche donc à représenter toutes les entités de manière équitable et donc à sur-représenter les types rares. Les résultats obtenus n'ont pas été encourageants. À titre d'exemple, la figure 5 montre les performances de cette stratégie, comparée à `random`, sur MERLOT (cf. tableau 1 pour la distribution des types d'entités).

Stabilité Enfin, remarquons que les variations du modèle `random` sont parfois plus fortes que celles des autres stratégies (corpus E3C, et dans une moindre mesure EMEA et MERLOT). Certaines sélections aléatoires peuvent conduire à de très mauvais résultats, ce qui n'est pas le cas des stratégies d'*active learning*, beaucoup plus stables. Ce point peut prendre de l'importance lorsque la campagne d'annotation utilise une technique de pré-annotation des textes à partir d'un modèle entraîné avec peu de données, pour faciliter le travail humain.

Conclusion Nous avons étudié cinq stratégies d'*active learning* pour la reconnaissance d'entités nommées dans quatre corpus médicaux en français. Nos résultats suggèrent que les stratégies de représentativité sont particulièrement intéressantes sur des petits corpus en terme de temps de calcul et de stabilité des performances.

6 Remerciements

Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen qui nous ont permis d'utiliser le corpus LERUDI pour cette étude.

Références

AMBATI V., VOGEL S. & CARBONELL J. (2011). Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of Machine Translation Summit XIII : Papers*, Xiamen, China.

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. DOI : 10.1016/j.jbi.2022.104073, HAL : hal-03655039.

BLOODGOOD M. & CALLISON-BURCH C. (2010). Bucking the trend : Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 854–864, Uppsala, Sweden : Association for Computational Linguistics.

BRINKER K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, p. 59–66 : AAAI Press.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2017). 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. DOI : 10.1007/s10579-017-9382-y, HAL : hal-01631743.

CHEN C., PALMER A. & SPORLEDER C. (2011). Enhancing active learning for semantic role labeling via compressed dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, p. 183–191, Chiang Mai, Thailand : Asian Federation of Natural Language Processing.

CHEN J., SCHEIN A., UNGAR L. & PALMER M. (2006). An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, p. 120–127, New York City, USA : Association for Computational Linguistics.

CHEN Y., LASKO T. A., MEI Q., DENNY J. C. & XU H. (2015). A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, **58**, 11–18. DOI : https ://doi.org/10.1016/j.jbi.2015.09.010.

CHENG Y., CHEN Z., LIU L., WANG J., AGRAWAL A. & CHOUDHARY A. (2013). Feedback-driven multiclass active learning for data streams. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, p. 1311–1320, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2505515.2505528.

CLAVEAU V. & KIJAK E. (2015). Stratégies de sélection des exemples pour l'apprentissage actif avec des champs aléatoires conditionnels. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 13–24, Caen, France : ATALA.

COHEN R., ELHADAD M. & ELHADAD N. (2013). Redundancy in electronic health record corpora : Analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, **14**, 10. DOI : 10.1186/1471-2105-14-10.

COHN D., GHAHRAMANI Z. & JORDAN M. (1994a). Active learning with statistical models. In G. TESAURO, D. TOURETZKY & T. LEEN, Éds., *Advances in Neural Information Processing Systems*, volume 7 : MIT Press.

COHN D. A., ATLAS L. E. & LADNER R. E. (1994b). Improving generalization with active learning. *Machine Learning*, **15**, 201–221.

CULOTTA A. & MCCALLUM A. (2005). Reducing labeling effort for structured prediction tasks. In *AAAI Conference on Artificial Intelligence*.

ECK M., VOGEL S. & WAIBEL A. (2005). Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

ERDMANN A., WRISLEY D. J., ALLEN B., BROWN C., COHEN-BODÉNÈS S., ELSNER M., FENG Y., JOSEPH B., JOYEUX-PRUNEL B. & DE MARNEFFE M.-C. (2019). Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. 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)*, p. 2223–2234, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1231.

FORT K., NAZARENKO A. & ROSSET S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In *Proceedings of COLING 2012*, p. 895–910, Mumbai, India : The COLING 2012 Organizing Committee.

GEIFMAN Y. & EL-YANIV R. (2017). Deep active learning over the long tail. *CoRR*, **abs/1711.00941**.

GROUIN C., LAVERGNE T. & NÉVÉOL A. (2014). Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, p. 54–58, Dublin, Ireland : Association for Computational Linguistics and Dublin City University. DOI : 10.3115/v1/W14-4907.

HAERTEL R. A., SEPPI K. D., RINGGER E. K. & CARROLL J. L. (2008). Return on investment for active learning. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 72 : Citeseer.

HOULSBY N., HUSZAR F., GHAHRAMANI Z. & LENGYEL M. (2011). Bayesian active learning for classification and preference learning. *CoRR*, **abs/1112.5745**.

JIANG Z., GAO Z., DUAN Y., KANG Y., SUN C., ZHANG Q. & LIU X. (2020). Camouflaged Chinese spam content detection with semi-supervised generative active learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3080–3085, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.279.

KARAMCHETI S., KRISHNA R., FEI-FEI L. & MANNING C. (2021). Mind your outliers ! investigating the negative impact of outliers on active learning for visual question answering. 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)*, p. 7265–7281, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.564.

KIRSCH A., VAN AMERSFOORT J. & GAL Y. (2019). Batchbald : Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, **32**.

LEWIS D. D. & CATLETT J. (1994). Heterogeneous uncertainty sampling for supervised learning. In W. W. COHEN & H. HIRSH, Éds., *Machine Learning Proceedings 1994*, p. 148–156. San Francisco (CA) : Morgan Kaufmann. DOI : https ://doi.org/10.1016/B978-1-55860-335-6.50026-X.

LEWIS D. D. & GALE W. A. (1994). A sequential algorithm for training text classifiers. In B. W. CROFT & C. J. VAN RIJSBERGEN, Éds., *SIGIR '94*, p. 3–12, London : Springer London.

LIU M., TU Z., ZHANG T., SU T., XU X. & WANG Z. (2022). Ltp : A new active learning strategy for crf-based named entity recognition. *Neural Process. Lett.*, **54**(3), 2433–2454. DOI : 10.1007/s11063-021-10737-x.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2021). The e3c project : European clinical case corpus. *Language*, **1**(L2), L3.

MARGATINA K., VERNIKOS G., BARRAULT L. & ALETRAS N. (2021). Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 650–663, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.51.

MARTIN L., MULLER B., SUÁ REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MCCALLUM A. & NIGAM K. (1998). Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, p. 350–358, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

MCCRAY A. T., BURGUN A. & BODENREIDER O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, **84**(0 1), 216.

MIRROSHANDEL S. A. & NASR A. (2011). Active learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies*, p. 140–149, Dublin, Ireland : Association for Computational Linguistics.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

RADMARD P., FATHULLAH Y. & LIPANI A. (2021). Subsequence based deep active learning for 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)*, p. 4310–4321, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.332.

REN P., XIAO Y., CHANG X., HUANG P.-Y., LI Z., GUPTA B. B., CHEN X. & WANG X. (2021). A survey of deep active learning. *ACM Comput. Surv.*, **54**(9). DOI : 10.1145/3472291.

ROY N. & MCCALLUM A. (2001). Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 441–448, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

SCHEFFER T., DECOMAIN C. & WROBEL S. (2001). Active hidden markov models for information extraction. In F. HOFFMANN, D. J. HAND, N. ADAMS, D. FISHER & G. GUIMARAES, Éds., *Advances in Intelligent Data Analysis*, p. 309–318, Berlin, Heidelberg : Springer Berlin Heidelberg.

SCHEIN A. & UNGAR L. (2007). Active learning for logistic regression : An evaluation. *Machine Learning*, **68**, 235–265. DOI : 10.1007/s10994-007-5019-5.

SEARLE T., IBRAHIM Z., TEO J. & DOBSON R. (2021). Estimating redundancy in clinical text. *Journal of Biomedical Informatics*, **124**, 103938. DOI : https ://doi.org/10.1016/j.jbi.2021.103938.

SENER O. & SAVARESE S. (2018). Active learning for convolutional neural networks : A core-set approach. In *International Conference on Learning Representations*.

SETTLES B. & CRAVEN M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 1070–1079, Honolulu, Hawaii : Association for Computational Linguistics.

SHEN D., ZHANG J., SU J., ZHOU G. & TAN C.-L. (2004). Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, p. 589–es, USA : Association for Computational Linguistics. DOI : 10.3115/1218955.1219030.

SHEN Y., YUN H., LIPTON Z. C., KRONROD Y. & ANANDKUMAR A. (2018). Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

TANG M., LUO X. & ROUKOS S. (2002). Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 120–127, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073105.

TOMANEK K. & HAHN U. (2010). A comparison of models for cost-sensitive active learning. In *Coling 2010 : Posters*, p. 1247–1255.

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université. HAL : tel-03624928.

WEI Q., CHEN Y., SALIMI M., DENNY J. C., MEI Q., LASKO T. A., CHEN Q., WU S., FRANKLIN A., COHEN T. *et al.* (2019). Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, **26**(11), 1314–1322.

WU F., HUANG Y. & YAN J. (2017). Active sentiment domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1701–1711, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1156.

YU Y., KONG L., ZHANG J., ZHANG R. & ZHANG C. (2022). AcTune : Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1422–1436, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.102.

ZHAN X., WANG Q., HUANG K.-H., XIONG H., DOU D. & CHAN A. B. (2022). A comparative survey of deep active learning. *arXiv preprint arXiv :2203.13450*.

ZHANG S., GONG C., LIU X., HE P., CHEN W. & ZHOU M. (2022a). ALLSH : Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 1328–1342, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.99.

ZHANG Z., STRUBELL E. & HOVY E. (2022b). A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 6166–6190, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

ZHAO Y., ZHANG H., ZHOU S. & ZHANG Z. (2020). Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1796–1806, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.162.

ZHDANOV F. (2019). Diverse mini-batch active learning. *CoRR*, **abs/1901.05954**.

ZHOU B., CAI X., ZHANG Y., GUO W. & YUAN X. (2021). Mtaal : Multi-task adversarial active learning for medical named entity recognition and normalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(16), 14586–14593. DOI : 10.1609/aaai.v35i16.17714.

ZHU J. & HOVY E. (2007). Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 783–790, Prague, Czech Republic : Association for Computational Linguistics.