Reconnaissance d'entités nommées avec peu de supervision dans le domaine biomédical

Florent Charrier
Aix-Marseille Unversité
florent.charrier@etu.univ-amu.fr

Léo Bouscarrat
Euranova, Aix-Marseille Université
Cécile Capponi, Carlos Ramisch
Aix-Marseille Université
leo.bouscarrat@euranova.eu
cecile.capponi@lis-lab.fr
carlos.ramisch@lis-lab.fr

Abstract

La reconnaissance d'entités nommées (NER) est un processus fondamental dans beaucoup de tâches de traitement du langage naturel. Les meilleures méthodes aujourd'hui utilisent de l'apprentissage supervisé c'est-à-dire qu'elle nécessite des corpus annotés. Or, cette étude se place dans un contexte plus large où l'accès à de tels corpus est difficile, un projet de détection de nouveauté dans la veille sanitaire.

Dans ce travail nous voulons donc savoir à quel point les méthodes semi-supervisées existantes sont efficaces comparées aux méthodes supervisées et si l'utilisation de ces méthodes était envisageable pour le projet dans lequel ce travail s'inscrit.

Pour cela on va étudier en particulier deux méthodes, une conçu pour le domaine biomédical et une autre conçue pour les domaines non spécialisés. On les testera sur des corpus classiques de NER dans le domaine biomédical.

Les résultats obtenus indiquent que les méthodes conçues pour des domaines non spécifiques s'adaptent difficilement au biomédical, mais que celles conçu pour ce domaine obtiennent des résultats satisfaisants au regard de l'état de l'art sur les méthodes supervisées, en effet, on obtient des résultats à 7 points de F-score de l'état de l'art supervisé. L'utilisation de ces méthodes est donc envisageable pour le projet de détection de nouveauté dans la veille sanitaire.

Mots-clef : TAL, Biomédical, NER, Machine learning, semi-supervisé, LSTM, NCBI, BC5CDR, veille sanitaire

1 Introduction

La reconnaissance d'entités nommées (NER) est la tâche d'extraction d'information dans du texte consistant à identifier les mots ou groupe de mots dénotant un objet nommé spécifique, tels que les noms de personnes, endroits, institutions, etc. C'est un processus fondamental qui est en amont de nombreuse autres tâches de traitement du langage naturel tel que l'extraction de relation, la classification et catégorisation de documents, le topic modeling, la compréhension de documents. Par exemple, prenons la phrase "Pierre et son frère se rendent dans la ville de Marseille", on identifiera ici deux entités nommées, Pierre qui est un prénom et Marseille qui est un lieu.

La reconnaissance d'entités nommées ne comprend généralement pas les domaines de spécialité tels que le biomédical et les noms de maladies, mais le problème et les méthodes pour le résoudre sont similaires. Cette tâche nécessite généralement une grande quantité de données sous la forme de corpus de texte. De telles quantités de données sont disponibles et de plus en plus apparaissent chaque jour, MEDLINE par exemple indexe plus de 904 000 articles et citations provenant de 5 200 revues scientifiques, PubMed 24 millions de citations publiées depuis 1950 dans environ 5000 revues biomédicales.

Une des difficultés majeures pour le NER provient du fait que les méthodes efficaces actuelles sont basées sur l'apprentissage supervisé, c'est-à-dire qu'elles nécessitent du texte annoté, et que bien qu'une grande quantité de données soit disponible, l'annotation de ces données est difficile et extrêmement coûteuse, dans le domaine général et particulièrement dans le domaine biomédical. En effet, contrairement aux domaines plus simples tels que l'identification de lieux ou de personnes, l'annotation de ces données requiert un avis expert (médecin, chercheur par exemple). C'est pour cette raison que l'on cherche à développer des solutions basées sur l'apprentissage non supervisé ou semi-supervisé, car ces méthodes ne nécessitent pas ou peu d'annotations.

Ce travail s'inscrit dans le projet de thèse de Léo Bouscarrat et est géré par trois entités : le LIS, l'entreprise Euranova et le CESPA (Centre d'épidémiologie et de santé publique des armées). Il s'agit d'une thèse sur la reconnaissance de la nouveauté pour de la veille sanitaire. En effet les institutions gérant ces données-là, emploient des experts qui sont bien souvent surchargés, l'idée est de proposer un système capable de prioriser les news les plus importantes et les plus "nouvelles". Ce projet a besoin de NER mais l'on ne dispose pas de corpus annoté, d'où l'intérêt de ce travail pour évaluer les méthodes non et peu supervisés.

L'objectif de ce travail est d'évaluer l'efficacité des méthodes actuelles de NER semi-supervisé appliqué au domaine biomédical et essayer d'analyser et de trouver les points faibles et points forts de ces méthodes.

Le domaine du bio-médical présente certaines caractéristiques qui vont rendre cette tâche difficile, par exemple :

— les entités sont souvent "descriptives" (ex : normal thymic epithelial cells) ce qui peut poser des problèmes d'ambiguïté. (exemple : "dilated cardiomyopathy" qui est une entité et "cardiomyopathy" qui est en une aussi).
— les entités ont parfois plusieurs écritures valides. (ex : "N-acetylcysteine", "N-acetyl-cysteine", ou "NAcetylCysteine")
— les corpus contiennent beaucoup d'abréviations, souvent ambiguës. (ex : "TCF" qui peut signifier "T cell factor" ou "Tissue Culture Fluid")

Les solutions d'apprentissage semi-supervisé dans le domaine suivent à peu près toutes le même schéma : on dispose d'un dictionnaire de base qu'on essaie d'étendre et d'un corpus. À partir du dictionnaire, on annote le texte avec du dictionnary matching, c'est un dire un algorithme qui cherche la chaîne de mots dans le texte la plus longue correspondante exactement à une entrée du dictionnaire. Puis on applique les méthodes de l'état de l'art du NER supervisé sur ce texte annoté automatiquement (généralement des Bilstm-crf ou plus récemment BERT).

Nous allons essayer de répondre aux questions suivantes :

1. Est-ce que les méthodes semi-supervisées plutôt efficaces dans des textes non spécialisés s'adaptent bien dans le domaine biomédical ? Au vu de la complexité du domaine on peut s'attendre à des performances légèrement inférieures à celles obtenues dans des textes non-spécialisés.
2. Ces méthodes sont-elles capables d'étendre le dictionnaire de façon non triviale, i.e sont-elles capables de trouver des mots nouveaux (non présents dans le dictionnaire de base) ou trouver des entités complexes (par exemple, les acronymes ou les entités composées de plusieurs mots) ?
3. Quel est l'impact sur les performances du système de la taille et des caractéristiques des données d'entrée (i.e le dictionnaire et son origine, le corpus sur lequel on applique la méthode d'extension de dictionnaire) ?

2 État de l'art

Au cours de ces dernières années, les méthodes basées sur des réseaux de neurones récurrents tels que l'architecture LSTM-CRF ont permis de grandement faire progresser la tâche de reconnaissance d'entités nommées dans le domaine biomédical (Habibi et al., 2017; Giorgi and Bader, 2018; Lample et al., 2016) Cependant, comme dans d'autres domaines du traitement du langage naturel, l'état de l'art actuel dans le domaine emploie des méthodes basé sur BERT (Devlin et al., 2018). BERT applique l'entraînement d'un transformé bidirectionnel (un modèle d'attention) pour obtenir un modèle de langue, la plupart des modèles de langue était obtenu de façon unidirectionnelle en passant sur le texte de gauche à droite ou de droite à gauche (ou en combinant les deux). La particularité de BERT vient de son modèle de langue masqué qui permet à BERT de capter plus de contexte en analysant la phrase de façon bidirectionnel. BioBERT, une version de BERT pré-entraîné sur une grande quantité de documents biomédical, est l'état de l'art actuel sur les jeux de données de détection d'entités biomédical (Lee et al., 2019).

Si ces méthodes sont efficaces, elles ne sont pas adaptées à la tâche qui nous concerne, en effet, ce sont avant tout des méthodes supervisés, or nous nous trouvons dans une situation qui ne nous permet pas d'obtenir des données à grande échelle. De plus, le projet se situant dans le domaine de la veille sanitaire, la capacité à détecter de nouveaux noms de maladie (comme le covid-19) est importante, ce que les méthodes supervisées ont du mal à faire.

Beaucoup d'approches ont été proposées pour la tâche de reconnaissance d'entité nommées avec peu de supervision, mais on remarque assez vite que la plupart des méthodes proposent un schéma similaire : on annote le corpus en utilisant soit du dictionnary matching avec une étape d'expansion du dictionnaire, soit des modèles NER classique, puis on utilise le corpus ainsi annoté dans un modèle plus sophistiqué.

Une des méthodes que nous étudions est basée sur le PU Learning (Peng et al., 2019)

C'est une méthode utilisant seulement des exemple positifs (P) et des exemples non étiqueté (U) contenant à la fois des exemples positifs et négatifs. En partant d'un dictionnaire de base, ils étiquettent le corpus à partir d'un dictionary matching. Les entités ainsi annotées seront la classe P, tout le reste est considéré comme faisant parti de U, ensuite un classifieur binaire est utilisé sur ces données avec une perte légèrement modifiée pour empêcher le système de tout étiqueter vers une seule classe. Ici, ils ont utilisé une architecture BiLSTM-CRF.

Pour que cette méthode fonctionne correctement, il faut que la classe P générée possède une distribution de probabilité proche de la vraie classe que l'on cherche à identifier, ce qui est difficile à obtenir en utilisant un simple dictionnaire qui ne sera pas suffisant pour l'estimer. Pour régler ce problème, ils proposent une méthode inspirée de l'AdaSampling : un classifieur est entraîné, puis pour une entité prédite, si cette entité apparaît un certain nombre de fois et que pour chaque apparition de cette entité sont prédite comme appartenant à la classe P, alors on ajoute cette entité dans le dictionnaire. C'est un processus itératif qui ne s'arrête que lorsqu'on n'ajoute plus de mots dans le dictionnaire.

Une autre méthode évaluée dans notre travail est pro-

posée par (Wang et al., 2019). Leur solution est divisée en trois étapes distinctes :

— Une phase de dictionary matching (avec une adaptation préalable du dictionnaire pour réduire le nombre de faux positif).
— Une étape d'expansion d'entités : pour chaque type, ils définissent un set d'entité lui correspondant tiré du texte, les phrases non sélectionnées pour définir le set d'entité constitue ainsi le set candidat, enfin, selon un score de proximité sémantique ils décident si oui ou non le candidat est ajouté au set.
— Après ces deux étapes, on obtient un corpus annoté sur lequel est appliqué un LSTM. Comme on l'a vu précédemment, la plupart des méthodes supervisées utilisent une méthode BiLSTM. On a cependant deux différences ici : tout d'abord on n'a pas de couche CRF, et le tagging utilisé diffère du tagging habituel (IOB, IOBES).

On suit un schéma Tie Or Break : le classifieur va d'abord être entraîné à trouvé le lien entre deux mots (Tie si les deux appartiennent à la même entité, sinon Break). Ceci leur permet d'obtenir les frontières de l'entité et le vecteur la représentant passe par une couche softmax pour déterminer à quelles catégories il appartient. Ils expliquent que ce schéma leur permet de mieux exploiter l'information du dictionnaire grâce à deux observations : tout d'abord, même si le modèle ne trouve pas précisément la position complète d'une entité, la plupart des liens entre les mots d'une même entité restent inchangé. Deuxièmement, ils ont remarqué que beaucoup d'entités monotoken était des faux positifs. Ce tagging leur permet d'atténuer cet effet : puisque qu'importe si le monotoken est un vrai positif ou un faux positif, dans les deux cas, il sera entouré de deux Break.

Enfin, il existe aussi des méthodes de bootstrap (Mathew et al., 2019), mais il existe aussi d'autres approches plus original basées sur de l'apprentissage par renforcement (Yang et al., 2018) ou encore qui tente de transformer le problème de NER en problème de Question Answering pour pouvoir utiliser les récents modèles très efficaces de ce domaine (Banerjee et al., 2019)

Nous avons choisi la méthode basée sur le PU learning et AutoBioNER car elles semblaient intéressantes pour répondre à nos questions : la première proposait des résultats proche des méthodes supervisés, mais sur un domaine totalement différent (f mesure de 82 sur CoNLL par exemple comparé aux 90 de la méthode supervisé) la seconde a été pensé pour le domaine biomédical.

3 Méthode et données

Dataset

— BC5CDR a été créé à partir d'article PubMed. Il contient 1500 articles avec 12 852 mention de maladie et 15 933 mentions d'élément chimique. Le corpus est déjà divisé en train, dev et test, chacun contenant 500 articles.
— NCBI-Disease est un dataset benchmark classique dans le domaine biomédical, il est obtenu à partir d'abstract Pubmed. Il contient 793 abstract avec 6 881 mentions de maladie. Il est lui aussi déjà séparé en trois, un train de 593 abstract, et un dev et test de 100 abstract chacun.

Dictionnaire

Pour les différentes expérimentations, on a utilisé plusieurs dictionnaires :

— Train : Un dictionnaire généré en récupérant un pourcentage des entités présentent dans le train.
— MeSH : Un dictionnaire crée à partir de la base de donnée de la NLM (National Library of Medicine), il contient 6877 entités.
— Ontologie : Un dictionnaire extrait d'une ontologie médicale, il contient 32 562 entités.

Métrique

Pour mesurer l'efficacité de ces méthodes, on utilisera le f-score F, calculé à partir de la précision P et du rappel R selon les formules suivantes :

$$F = 2 \cdot \frac{P * R}{R + P} \tag{1}$$

$$P = \frac{Vraipositif}{vraipositif + fauxpositif} \tag{2}$$

$$R = \frac{Vraipositif}{vraipositif + fauxnegatif} \tag{3}$$

Méthode

Pour répondre à nos questions, nous allons tout d'abord tester la méthode basé sur le PU learning, que l'on nommera juste PU Learning par la suite, puis la méthode conçu pour le biomédical que l'on nommera AutoBioNER. Ces deux méthodes sont basées sur une phase d'extension du dictionnaire pour annoter le corpus, puis l'utilisation d'un réseau neuronal ensuite. Au cours de nos expérimentations, nous nous sommes rendu compte que l'étape importante pour notre projet était surtout celle d'extension du dictionnaire. Ainsi, pour la 1re question, nous testerons entièrement les deux méthodes, mais pour les deux questions suivantes nous vérifierons nos hypothèses seulement sur l'étape d'expansion du dictionnaire. Pour la 1re question, on utilisera le dictionnaire issue du train, et le corpus utilisé sera celui spécifié dans les tableaux correspondant. Il est important de mentionner le script ConLL eval. ConLL eval est un script utilisé pour mesurer les performances de reconnaissance d'entités nommées sur

Dataset	BC5DR	NCBI
Origine	Article	Abstract
type d'entité	Chimie, Maladie	Maladie
Nombre de phrases	20 217	7 286
Nombre d'entité maladie	12 852	6 881

TABLE 1 – Dataset

un dataset connu, le dataset ConLL. Les valeurs affichées par la suite auront été obtenues avec ce script. La seule différence se situe au niveau du PU Learning, en effet nous nous somme rendue compte lors de nos expérimentations que les valeurs renvoyées par cette méthode était très différentes de celle obtenue par le script standard. Ainsi pour la 1re question, les deux métriques sont présenté, mais par la suite seul le script ConLL sera utilisé pour mesurer la performance. Pour la deuxième question, les tests ont été effectué sur NCBI-Disease seulement et sur les résultats obtenus par le dictionary matching après extension du dictionnaire par la méthode de PU learning. Et enfin, pour la 3e question, les tests ont été effectués sur AutoBioNER, seulement sur la phase d'extension du dictionnaire. Tous les dictionnaires cités plus haut sont utilisés ici, et les tests sont effectués sur BC5CDR.

4 Expériences et résultats

Efficacité des méthodes étudiées

On observe dans la table 2 que les méthodes étudiées permettent d'obtenir un gain assez significatif sur l'extension du dictionnaire. C'est particulièrement vrai pour AutoBioNER, pour lequel on observe un gain de 5 points de F1 sur BC5CDR et de 9 points sur NCBI.

Dans la table 3, on remarque plusieurs choses :

— Dans le cas de AutoBioNER, les résultats obtenus sont très encourageant, on est à 10 points seulement de la méthode complètement supervisé sur NCBI et 7 points sur BC5CDR

— On remarque aussi que les deux étapes de chaque méthode ont bien un effet sur la reconnaissance d'entités nommées, l'extension du dictionnaire permet d'ajouter des entités, et le LSTM dans les deux cas ici permet d'ajouter encore des entités que la simple extension du dictionnaire ne peut trouver.

— Enfin, on peut voir une ligne supplémentaire mentionnant CoNLL script, CoNLL est un dataset souvent utilisé dans la reconnaissance d'entité nommée. On s'est rendu compte au cours de nos expérimentations que le F-score renvoyé par la méthode de PU learning n'était pas calculé de façon standard comme expliquée plus haut. En effet, en NER, on peut souvent tomber sur des réponses "partielles", par exemple l'entité est détectée correctement, mais les frontières de cette entité sont trop large ou trop courte. Ainsi, il faut choisir entre une mesure exacte, c'est à dire que seule entité identifiée entièrement est valide, et une mesure où l'on peut décidé que simplement avoir trouvé une partie de l'entité suffit à donner des points.

Ces résultats nous permettent de répondre à notre première question : les méthodes semi-supervisés sont assez efficace pour être utilisés dans le projet, et les méthodes issue de domaines non spécialisé semble rencontrer des difficultés à s'adapter au domaine biomédical.

Analyse des résultats

On peut faire plusieurs constats sur ces résultats : Les méthodes semi-supervisées obtiennent des résultats plutôt acceptables sur la tâche qui nous intéresse. En comparant les résultats de PU learning, une méthode conçue pour la reconnaissance d'entités nommées classique (personne, location, etc...) avec AutoBioNER, méthode pensée pour le domaine biomédical, on voit que le domaine biomédical possède des spécificités qui rendent le passage d'un domaine à l'autre difficile.

On va donc analyser certains des résultats obtenu par PU learning pour mettre en évidence ces difficultés.

On s'intéresse ici aux différents types d'entités que l'on peut rencontrer : les entités constituées d'un seul token (ex : cancer), celle constituée de plusieurs tokens (Deux ou plus donc, ex : hyperammonemic encephalopathy) et enfin les acronymes (ex : HIV).

Dans la table 4, on observe tout d'abord que dans tous les cas le rappel est inférieur à la précision. Dans les méthodes basées sur le dictionary matching c'est généralement l'inverse (car le dictionnaire ne peut pas être exhaustif) et c'est ce qu'on observe d'ailleurs dans la table 3. Ici la méthode basée sur le PU learning semble donc faire chuter la précision au profit du rappel. Premièrement, les entités monotoken semble plus difficile à identifier, on s'attendait plutôt au contraire. De façon assez surprenante cette méthode fonctionne le mieux sur les acronymes. Cela pourrait venir du fait que le modèle se contente de souvent tagger les entités entièrement en majuscule, on peut en effet s'attendre à ce que dans un texte biomédical, ces tokens là soit des acronymes. De manière générale, les résultats obtenus sont assez décevant, et la méthode PU learning semble avoir du mal à trouver des entités non-triviales.

Impact de la nature, taille du dictionnaire et du corpus d'entraînement

On va ici s'intéresser aux résultats de la seconde méthode, AutoBioNER, plus précisément les résultats de la phase de dictionary expansion, car c'est avant tout ce que l'on cherche à améliorer ici. On utilisera les dictionnaires cité plus haut, et les tests sont effectués sur BC5CDR

La figure 1 représente l'évolution du f-score en fonction de la taille du corpus d'entraînement. "Corpus" représente BC5CDR et le pourcentage de ce corpus que l'on a utilisé. À partir de 100% on utilise PubMed, un autre corpus bien plus gros que BC5CDR (BC5 étant à peu près égale à 14% de pubmed). On observe ici que la taille du corpus d'entraînement n'impacte que très peu la phase d'expansion du dictionnaire. On passe d'un 44% initial à 47% pour le meilleur résultat. Cela n'est pas forcement surprenant, le corpus de base est exploité en utilisant un dictionnaire de base et ce sont les annotations obtenues à partir de ce dictionnaire qui permettent son expansion.

Dataset	Méthode	Précision	Rappel	F1
ncbi	Dictionary Matching	65	58	61
	PU learning	63.47	62.21	62.83
	AutoBioNER	95.95	57.89	72.21
BC5CDR	Dictionary Matching	92.28	57.47	70.80
	PU learning	71.05	64.89	67.83
	AutoBioNER	91.89	63.56	75.14

TABLE 2 – Résultats sur le dictionary matching apres la phase de dictionary expansion

Dataset	Methode	Rappel	Précision	F1
ncbi	PU learning	58.85	69.56	63
	PU learning (conll script)	33	65	44
	AutoBioNER	75.31	77.98	77.58
	BioBert (supervisé)	88.02	86.76	87.38
BC5CDR	PU learning	60.24	70.76	65.07
	AutoBioNER	87.34	84.53	85.91
	BioBert (supervisé)	92.76	92.52	92.64

TABLE 3 – Résultats finaux obtenu sur le tagging obtenu des réseaux de neurones de chaque méthode

mono token precision	mono token rappel	mono token f1
0.28	0.58	0.37
acronyme precision	acronyme rappel	acronyme f1
0.55	0.56	0.56
multi token precision	multi token rappel	multi token f1
0.37	0.71	0.49

TABLE 4 – PU learning résultats en fonction du type d'entité

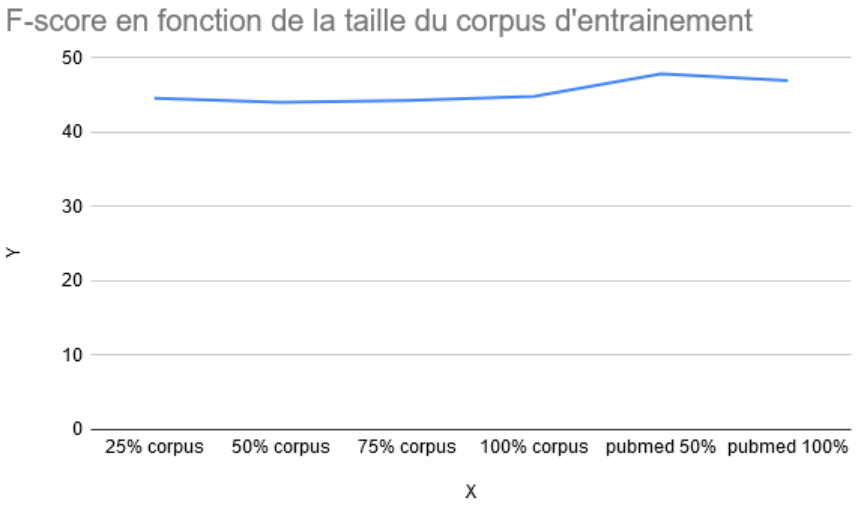

FIGURE 1 – F1 en fonction de la taille du corpus

Dictionnaire	Précision	Rappel	F1	Taille dictionnaire
Extrait de NCBI	92.72	13.15	23.04	1757
MeSH	46.61	82.08	59.46	7425
Ontologie	91.17	29.72	44.82	32 626
50% Ontologie	91.30	8.64	15.79	16 305

TABLE 5 – Impact du type et de la taille du dictionnaire

On peut voir cette information de façon bien plus prononcée dans la table 5. L'origine du dictionnaire semble avoir un impact bien plus important que la taille de celui-ci. En effet, le dictionnaire construit à partir de MeSH obtient les meilleurs résultats bien qu'il soit plus petit que les dictionnaires issu de l'ontologie. Cependant, cela ne signifie pas forcement que la taille n'a pas d'effet. On voit en effet qu'en n'utilisant que 50% du dictionnaire issue de l'ontologie, on perd presque 30 points sur notre f score.

Ainsi, la nature du dictionnaire semble être le facteur le plus important pour améliorer les performances de la phase de dictionnary expansion sur ces méthodes semi-supervisé. Bien que la taille du corpus et du dictionnaire semble aussi avoir un impact positif sur les performances, ce n'est pas le facteur le plus important.

5 Conclusions et perspectives

Nous avons vu qu'il existait des méthodes semi-supervisées plutôt efficace pour la reconnaissance d'entités nommées dans le domaine biomédical. Ces méthodes nécessitent cependant l'obtention d'un dictionnaire choisi de façon soigneuse.

Les méthodes qui fonctionnent bien dans les domaines non spécialisés semblent avoir des difficultés à s'adapter au domaine biomédical, on aimerait ainsi dans le futur étudier plus en détail pourquoi et quels sont précisément les caractéristiques du domaine biomédical qui pose problème.

Enfin, la plupart des méthodes actuelles semblent être basées sur une phase d'extension de dictionnaire pour annoter faiblement un corpus, puis d'appliquer un réseau de neurones dessus. Il serait intéressant dans le futur d'essayer ces méthodes en appliquant BERT, état de l'art actuel dans le domaine, puisque les méthodes utilisées ici ont été conçues avant/pendant l'arrivée de BERT.

References

[Banerjee et al.2019] Pratyay Banerjee, Kuntal Pal, Murthy Devarakonda, and Chitta Baral. 2019. Knowledge guided named entity recognition. 11.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT : pre-training of deep bidirectional transformers for language understanding. volume abs/1810.04805.

[Giorgi and Bader2018] John Giorgi and Gary Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. volume 34, 06.

[Habibi et al.2017] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. volume 33, pages i37–i48, 07.

[Lample et al.2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. volume abs/1603.01360.

[Lee et al.2019] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert : a pre-trained biomedical language representation model for biomedical text mining. volume abs/1901.08746.

[Mathew et al.2019] Joel Mathew, Shobeir Fakhraei, and José Luis Ambite. 2019. Biomedical named entity recognition via reference-set augmented bootstrapping. volume abs/1906.00282.

[Peng et al.2019] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. volume abs/1906.01378.

[Wang et al.2019] X. Wang, Y. Zhang, Q. Li, X. Ren, J. Shang, and J. Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503.

[Yang et al.2018] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.