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

Extraction d'entités nommées à partir de descriptions d'espèces

Maya Sahraoui[1, 2] Vincent Guigue[3] Regine Vignes-Lebbe[2] Marc Pignal[2]
(1) ISIR, Sorbonne Université, Paris , France
(2) MNHN, Sorbonne Université, Paris , France
(3) AgroParisTech, Paris-Saclay, Paris, France
sahraoui@isir.upmc.fr, vincent.guigue@isir.upmc.fr,
regine.vignes_lebbe@sorbonne − universite.fr, marc.pignal@mnhn.fr

RÉSUMÉ

Les descriptions d'espèces contiennent des informations importantes sur les caractéristiques morphologiques des espèces, mais l'extraction de connaissances structurées à partir de ces descriptions est souvent chronophage. Nous proposons un modèle texte-graphe adapté aux descriptions d'espèces en utilisant la reconnaissance d'entités nommées (NER) faiblement supervisée. Après avoir extrait les entités nommées, nous reconstruisons les triplets en utilisant des règles de dépendance pour créer le graphe. Notre méthode permet de comparer différentes espèces sur la base de caractères morphologiques et de relier différentes sources de données. Les résultats de notre étude se concentrent sur notre modèle NER et démontrent qu'il est plus performant que les modèles de référence et qu'il constitue un outil précieux pour la communauté de l'écologie et de la biodiversité.

ABSTRACT

Named Entity Recognition for species descriptions.

Species descriptions contain important information about the morphological characteristics of species, but extracting structured knowledge from these descriptions is often time consuming. We propose a text-graph model adapted to species descriptions using weakly supervised named entity recognition (NER). After extracting the named entities, we reconstruct the triplets using dependency rules to create the graph. Our method allows us to compare different species based on morphological characters and link different data sources. The results of our study focus on our NER model and demonstrate that it outperforms benchmark models and is a valuable tool for the ecology and biodiversity community.

MOTS-CLÉS : Extraction d'entités nommées$_1$, Supervision distante$_2$, Bases de connaissance$_3$..

KEYWORDS: Named entity recognition$_1$, Distant supervision$_2$, Knowledge base$_3$..

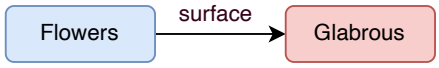
Actes de la 18e Conférence en Recherche d'Information et Applications (CORIA), pages 111–126.

1 Introduction

Les descriptions d'espèces sont une source d'information cruciale pour les études sur la biodiversité et l'identification des espèces. Elles fournissent des informations détaillées sur les caractéristiques morpho-anatomiques d'une espèce donnée, qui sont essentielles pour la distinguer des autres espèces. Ces descriptions sont généralement rédigées sous forme de texte dans les revues scientifiques, les livres ou les bases de données. Pour faciliter l'étude et la comparaison de différentes espèces, les descriptions d'espèces sont souvent représentées sous la forme de graphes de connaissances qui sont utilisés pour identifier les relations entre différentes espèces et modéliser l'évolution des espèces. Cependant, l'extraction de ces informations à partir des descriptions d'espèces peut être une tâche longue et laborieuse, généralement réalisée par des experts (Saucède *et al.*, 2021). L'automatisation de ce processus est donc un enjeu pour les chercheurs de cette communauté.

Les modèles d'apprentissage profond ont montré des performances remarquables dans les tâches d'extraction de connaissances (Miwa & Bansal, 2016). Cependant, ces modèles sont lourds et nécessitent une quantité importante de données annotées dans le domaine visé pour atteindre une grande précision, ce qui est très couteux. Nous nous focalisons ainsi sur les approches d'extraction d'entités nommées (NER) en faisant l'hypothèse que le domaine particulier des documents visés nous permettra de créer les étiquettes spécifiques pour les sources –*organes*– et les cibles –*types spécifiques d'attributs*–. Le fait que les attributs soient identifiés avec leur nature permet la reconstruction du triplet comme dans l'exemple :

<ORGAN> Flowers </ORGAN> 4-merous. Calyx aestivation valvate, campanulate, 2-3.6mm long, abaxially <DESC-SURFACE> glabrous </DESC-SURFACE>



La plupart des corpus annotés et des modèles pré-entraînés en NER sont conçus pour des données générales (e.g. wikipedia). De plus, le transfert entre domaines est difficile (Taillé *et al.*, 2021). A l'exception des documents financiers ou biomédicaux (Magge *et al.*, 2018), peu de domaines spécifiques sont abordables en exploitant seulement les ressources académiques disponibles. L'enjeu de cet article est donc très appliqué : il s'agit de proposer une méthodologie globale pour l'extraction d'information dans le domaine de la description d'espèces.

Notre approche repose sur 3 étapes : (1) Segmenter le texte par description d'organe à l'aide de règles ; (2) Identifier les organes et les descripteurs par extraction d'entités nommées ; (3) Reconstruire les triplets à l'aide de règles. Cet article décrit principalement le processus de l'étape (2) qui se décompose lui-même en différentes étapes : (a) Récupérer un glossaire (mots-clés et classes) auprès d'experts du domaine ; (b) Exploiter les termes du glossaire pour annoter le corpus de manière distante ; (c) Entrainer un premier modèle NER ; (d) Améliorer le modèle (teacher-student self-training). La contribution applicative de cet article est donc d'étudier un cas d'usage spécifique de l'extraction d'entités nommées en supervision distante avec les outils de l'état de l'art.

Après une rapide bibliographie sur la gestion et l'extraction de connaissance, en particulier sur les approches de NER apprise en supervision distante (section 2), nous décrirons en section 3 le processus d'annotation et le modèle utilisé pour l'extraction. La section 4 décrit les résultats quantitatifs et qualitatifs obtenus lors de la campagne d'expériences.

2 Travaux connexes

La représentation des connaissances est un enjeu important pour l'étude des espèces en biologie. Les systèmes actuels reposent principalement sur des experts (Vignes-Lebbe *et al.*, 2017), ce qui est très intéressant pour des études ciblées mais qui ne permet pas le traitement en masse des documents disponibles. L'enjeu consiste à trouver les bons outils dans la littérature pour faire face à ce défi. Les modèles d'apprentissage profond ont montré des performances remarquables dans les tâches d'extraction de connaissances mais nécessitent en général un large corpus de données annotés dans le domaine pour l'apprentissage (Miwa & Bansal, 2016; Taillé *et al.*, 2021).

En réduisant la tâche d'extraction de connaissances à de la reconnaissance d'entités nommées (NER), comme illustré en introduction, les architectures sont plus simples. Les architectures de NER ont beaucoup évoluées ces dernières années : les chaines de Markov cachées (HMM) puis les champs aléatoires conditionnels (CRF) ont permis des avancées significatives dans la modélisation et l'analyse des séquences de mots au début des années 2000 (Malouf, 2002; McCallum & Li, 2003). Les approches neuronales (convolutionnelles puis récurrentes) ont ensuite largement contribué à l'amélioration des performances (Collobert *et al.*, 2011) mais c'est la combinaison de différentes représentations des mots (Mikolov *et al.*, 2013) à l'entrée de ces architectures neuronales qui a engendré les gains de performances les plus importants (Tai *et al.*, 2015; Lample *et al.*, 2016). L'avènement de l'architecture Transformer (Vaswani *et al.*, 2017) et des modèles dérivés (Devlin *et al.*, 2019) ont entretenu la dynamique d'amélioration des perfromances à la fin des années 2010. Malgré cette dynamique remarque, il reste nécessaire d'avoir des données étiquetées du domaine du fait des capacités de transfert limitées des approches NER (Taillé *et al.*, 2020).

Cette limite sur le transfert explique l'émergence d'approches (et de corpus) spécifiques pour différentes langues (Souza *et al.*, 2019; Jia *et al.*, 2020) ou différents type de données comme les données biomédicales (Cho & Lee, 2019), légales (Leitner *et al.*, 2019) ou financières (Lee *et al.*, 2022). Cela montre aussi à quel point l'extraction d'entités nommées reste encore aujourd'hui une tâche particulièrement délicate en TAL malgré les avancées récentes.

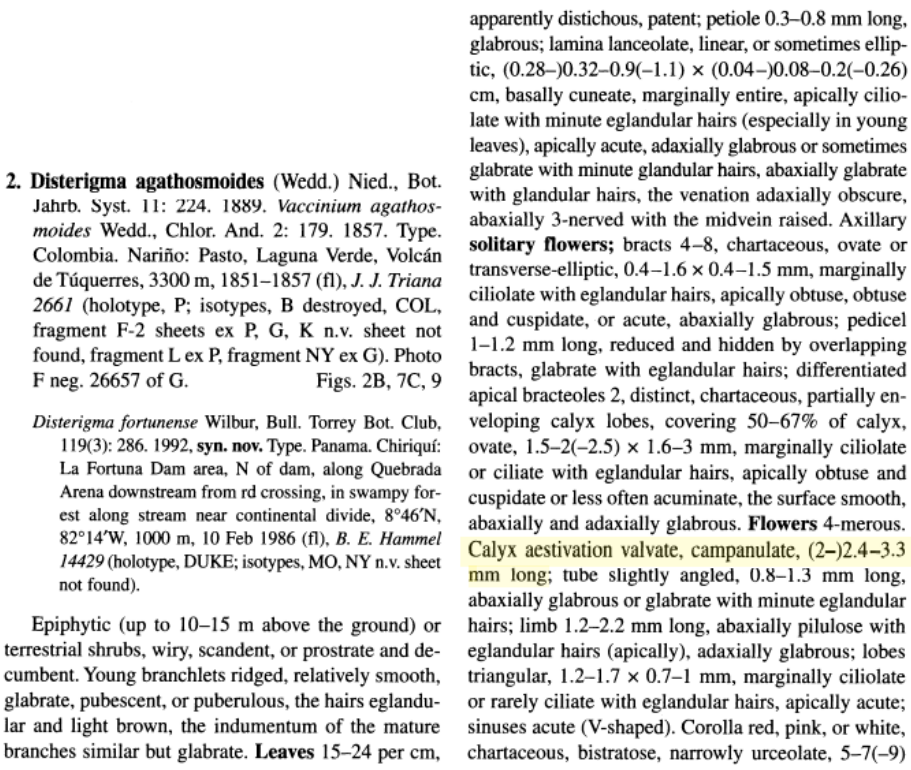
Le coût d'un corpus annoté en NER est conséquent, ce qui nous a poussé à étudier les possibilités de supervision distante (Wang *et al.*, 2020, 2021). A partir d'une liste de termes catégorisés (ie un glossaire réalisé par un expert), nous utilisons des expressions régulières pour annoter un corpus d'apprentissage. La supervision distante est par essence imparfaite : certaines annotations sont manquantes, d'autres ambigues voire erronées. Ce type de corpus implique donc des approches robustes qui seront en mesure de moins pénaliser certaines erreurs pour améliorer la généralisation. L'auto-apprentissage est actuellement l'approche la plus efficace pour éviter le sur-apprentissage d'une part et forcer le modèle à découvrir de nouveaux termes afin d'améliorer le rappel d'autre part. L'intégration séquentielle d'étiquettes prédites avec une forte confiance dans l'ensemble d'apprentissage pose cependant un grand risque de dérive du modèle qui peut être contenu en recourant à une architecture teacher-student (Liang *et al.*, 2020). Il est également pertinent de ré-entrainer le modèle de langue, à découvrir des mots masqués dans le contexte du domaine cible (Meng *et al.*, 2021). Nous allons exploiter ces deux stratégies dans le modèle proposé dans la section suivante.

3 Travaux et méthodes

Nous décrivons d'abord les données brutes et leur préparation, c'est-à-dire le processus d'annotation distante pour les entités. Nous détaillerons ensuite l'architecture envisagée pour la tâche de NER et la mise en œuvre de l'auto-apprentissage.

3.1 Création du jeu de données

Les corpus de faune et de flore contiennent des descriptions textuelles détaillées des espèces, y compris les caractéristiques morphologiques de divers groupes taxonomiques tels que les espèces, les genres et les familles. Notre travail se concentrera principalement sur les flores et en particulier le corpus *Flora Neotropica*, qui comprend des clés et des descriptions de différentes espèces, ainsi que des informations géographiques, cf Figure 1.



2. Disterigma agathosmoides (Wedd.) Nied., Bot. Jahrb. Syst. 11: 224. 1889. *Vaccinium agathosmoides* Wedd., Chlor. And. 2: 179. 1857. Type. Colombia. Nariño: Pasto, Laguna Verde, Volcán de Túquerres, 3300 m, 1851–1857 (fl), *J. J. Triana 2661* (holotype, P; isotypes, B destroyed, COL, fragment F-2 sheets ex P, G, K n.v. sheet not found, fragment L ex P, fragment NY ex G). Photo F neg. 26657 of G. Figs. 2B, 7C, 9

Disterigma fortunense Wilbur, Bull. Torrey Bot. Club, 119(3): 286. 1992, **syn. nov.** Type. Panama. Chiriquí: La Fortuna Dam area, N of dam, along Quebrada Arena downstream from rd crossing, in swampy forest along stream near continental divide, 8°46′N, 82°14′W, 1000 m, 10 Feb 1986 (fl), *B. E. Hammel 14429* (holotype, DUKE; isotypes, MO, NY n.v. sheet not found).

Epiphytic (up to 10–15 m above the ground) or terrestrial shrubs, wiry, scandent, or prostrate and decumbent. Young branchlets ridged, relatively smooth, glabrate, pubescent, or puberulous, the hairs eglandular and light brown, the indumentum of the mature branches similar but glabrate. **Leaves** 15–24 per cm, apparently distichous, patent; petiole 0.3–0.8 mm long, glabrous; lamina lanceolate, linear, or sometimes elliptic, (0.28–)0.32–0.9(–1.1) × (0.04–)0.08–0.2(–0.26) cm, basally cuneate, marginally entire, apically ciliolate with minute eglandular hairs (especially in young leaves), apically acute, adaxially glabrous or sometimes glabrate with minute glandular hairs, abaxially glabrate with glandular hairs, the venation adaxially obscure, abaxially 3-nerved with the midvein raised. Axillary **solitary flowers;** bracts 4–8, chartaceous, ovate or transverse-elliptic, 0.4–1.6 × 0.4–1.5 mm, marginally ciliolate with eglandular hairs, apically obtuse, obtuse and cuspidate, or acute, abaxially glabrous; pedicel 1–1.2 mm long, reduced and hidden by overlapping bracts, glabrate with eglandular hairs; differentiated apical bracteoles 2, distinct, chartaceous, partially enveloping calyx lobes, covering 50–67% of calyx, ovate, 1.5–2(–2.5) × 1.6–3 mm, marginally ciliolate or ciliate with eglandular hairs, apically obtuse and cuspidate or less often acuminate, the surface smooth, abaxially and adaxially glabrous. **Flowers** 4-merous. Calyx aestivation valvate, campanulate, (2–)2.4–3.3 mm long; tube slightly angled, 0.8–1.3 mm long, abaxially glabrous or glabrate with minute eglandular hairs; limb 1.2–2.2 mm long, abaxially pilulose with eglandular hairs (apically), adaxially glabrous; lobes triangular, 1.2–1.7 × 0.7–1 mm, marginally ciliolate or rarely ciliate with eglandular hairs, apically acute; sinuses acute (V-shaped). Corolla red, pink, or white, chartaceous, bistratose, narrowly urceolate, 5–7(–9)

FIGURE 1 – Echantillon du corpus *Flora Neotropica* contenant la description de l'espèce *Disterigma agathosmoides*

Les descriptions morphologiques des espèces dans les flores suivent un format semi-structuré qui passe progressivement d'une description d'un organe à ses sous-organes apparentés. Chaque organe ou sous-organe est décrit à l'aide de plusieurs descripteurs, notamment la couleur, la forme, la position et la disposition. Après avoir analysé les descriptions morphologiques, nous avons identifié trois schémas distincts dans la syntaxe des descriptions. La ponctuation (point et point-virgule) joue un rôle prépondérant dans la segmentation des descriptions disponibles :

Schéma 1 : Organ 1 description. Organ 2 description.

Schéma 2 : Organ 1 description ; Organ 2 description.
Schéma 3 : Main organ 1 description ; sub organ 1 ; sub organ 2 ; sub organ 3. Main organ 2 description ; sub organ 1 ; sub organ 2 ; sub organ 3.

Nous faisons l'hypothèse qu'en divisant les descriptions selon ces schémas, chaque segment de texte est centrée sur un organe ou un sous-organe. Après la segmentation, notre système va extraire l'organe en question (parfois malheureusement non unique) et les descripteurs. L'astuce, mentionnée en introduction, consiste à typer les entitées descripteurs à la fois avec une catégorie (couleur, forme, position,...) et à considérer ces entités comme des valeurs (vert, acuminé, alterné,...). Ainsi, les triplets modélisant les connaissances seront formés, dans chaque segment, de la manière suivante : le sujet sera l'organe principal, le prédicat correspondra à la catégorie et l'objet sera la valeur du descripteur extrait.

Soit G un glossaire contenant des mots liés à des descriptions morphologiques. Chaque mot w_j du glossaire est associé à une étiquette spécifique y_j.

$$G = \{(w_1, y_1), \dots (w_j, y_j), \dots, (w_N, y_N)\}, \qquad (w_j, y_j) \in \mathcal{W} \times \mathcal{Y} \tag{1}$$

Par soucis de simplicité, l'expert n'a dans un premier temps lister que des mots simples : nous n'abordons donc pas ici la problématique des entitées composées de plusieurs mots. L'ensemble $\mathcal{Y} = \mathcal{Y}_0 \cup \mathcal{Y}_1$ est constitué de deux sous-ensembles d'étiquettes correspondant respectivement aux organes et aux descripteurs :

$$\mathcal{Y}_0 = \{Flower, Fruit, Habit, Leaf, Part\text{-}of, Stem\text{-}root\} \tag{2}$$

$$\mathcal{Y}_1 = \{Color, Disposition, Form, Position, Surface\text{-}texture\} \tag{3}$$

Attention, dans toute la suite, les y désigneront en réalité des vecteurs de scores sur les C classes ($y \in \mathbb{R}^C$). Dans la version initiale, y prend la forme d'un *one-hot* sur la classe visée. Dans la suite de l'article, pour faire simple, nous parlerons de classe y alors que formellement, la classe serait plutôt $\text{argmax}(y)$. Cette subtilité est importante pour pouvoir introduire les mécanismes d'auto-supervision dans la suite.

Processus d'annotation distante. Après application des schémas sus-mentionnés, nous considérons le corpus comme un ensemble de phrases $S = \{s_0, s_1, ..., s_M\}$, chaque phrase $s_m = \{w_0, w_1, ..., w_{N_m}\}$ étant un ensemble de mots w_n.

Les mots w du corpus S et du glossaire G sont lemmatisés puis comparés : chaque appariement permet d'annoter un mot w_n avec l'étiquette y du glossaire. Les mots non reconnus sont affectés à la classe O. Nous obtenons ainsi un ensemble de séquences d'étiquettes $L = \{\ell_1, \dots, \ell_M\}$, $\ell_m = \{y_1, \dots, y_{N_m}\}$ alignées avec le corpus de phrases S.

Les statistiques du glossaire et de sa projection sur le corpus sont données dans le tableau 1.

TABLE 1 – Statistiques du jeu de données : classes considérées, nombres de mots distincts dans claque classe et nombre d'occurences dans le corpus.

Ensemble	Classe	Nombre d'occurences	Nombre de mots
$\mathcal{Y}_0$	Flower	22890	23
	Fruit	4968	10
	Habit	1920	3
	Leaf	4364	5
	Part-of	23849	25
	Stem-root	3296	7
$\mathcal{Y}_1$	Color	18342	15
	Disposition	8405	21
	Form	24816	64
	Position	10936	13
	Surface-texture	18325	23

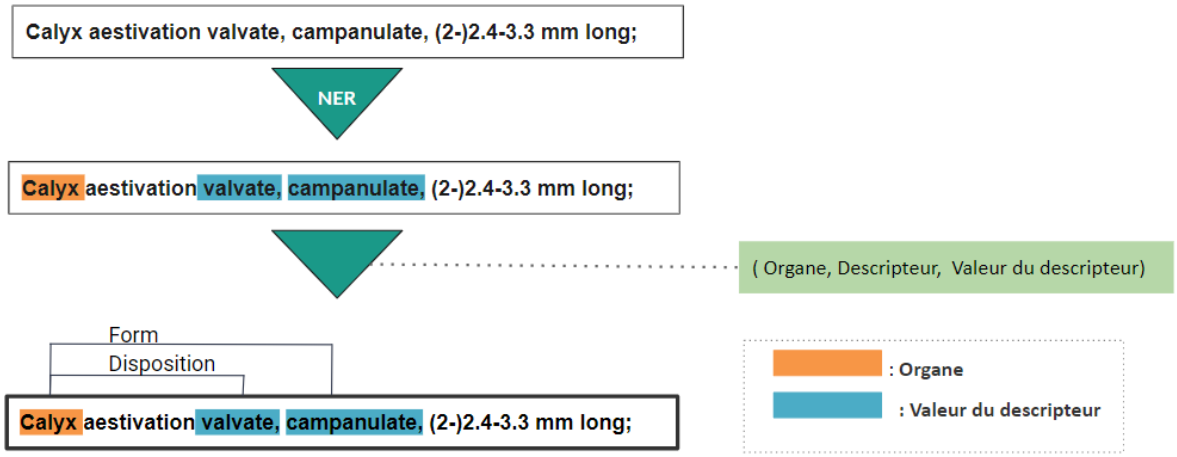


FIGURE 2 – Processus d'extraction de triplets à partir d'une phrase centrée sur un organe

3.2 Méthode d'extraction d'entitées nommées proposée

L'extraction d'entités nommées est une tâche difficile, ainsi, le développement d'une telle approche à partir de données aux annotations **bruitées** et **incomplètes** est un défi scientifique. La question même de l'évaluation dans un tel cadre est non triviale, nous y reviendrons plus tard. Pour résoudre ce problème, nous avons adopté l'auto-apprentissage, une technique qui intègre itérativement les prédictions les plus fiables du modèle dans la vérité terrain pour améliorer progressivement la couverture du modèle.

Le second défi auquel nous avons été confrontés réside dans la spécificité du vocabulaire et des structures de phrases associée à un domaine aussi pointu. Comme le montre la Figure 1, la distribution du vocabulaire et l'organisation des séquences de mots diffèrent trop du langage naturel usuel pour bénéficier de l'apport des modèles de langue pré-entrainés. Nous avons étudié deux stratégies pour palier ce problème : (1) l'utilisation d'un modèle de langue pré-entrainé, en misant sur la robustesse de celui-ci et (2) l'affinage du modèle de langues pré-entrainé en refaisant des prédictions de mots masqués sur les documents issus des flores.

Architecture. Le modèle de reconnaissance des entités nommées repose sur un encodeur BERT pré-entraîné avec une couche entièrement connectée pour la classification par mots-clés. Nous désignons ce modèle par f_θ, où θ représente l'ensemble des paramètres (modèle de langue et couche de classification). Pour chaque séquence s_m, $f_\theta(s_m) \in \mathbb{R}^{C \times N_m}$ est une estimation de $p(Y_j = c|s_m)$ pour tous les mots w_j de s et pour les C classes considérées. Nous noterons $f_{\theta,j,c}(w_j)$ la prédiction associée au mot j et à la classe c. L'entrainement du modèle est effectué sur des paires de séquences alignées (s_m, ℓ_m). La fonction de coût est classiquement une entropie croisée :

$$\mathcal{L} = - \sum_{(s,\ell) \in S,L} \sum_{(w_j,y_j) \in (s,\ell)} \sum_{c=0}^{C} y_{j,c} \log \frac{\exp(f_{\theta,j,c}(s))}{\sum_{c'=1}^{C} \exp(f_{\theta,j,c'}(s))} \tag{4}$$

Note : rappelons comme indiqué en section 3.1, que les y sont en fait des vecteurs de score sur les classes. $y_{j,c}$ désigne ainsi le score du mot j pour la classe c. Dans cette première partie, $y_{j,c} = 1$ pour la classe annotée de w_j et 0 pour toutes les autres classes.

Ré-entrainement de toutes les couches du modèle. Au cours du processus d'apprentissage, nous mettons à jour toutes les couches du modèle de bout en bout : les couches de l'encodeur ne sont pas figées. Cette approche améliore la qualité des résultats et l'efficacité d'apprentissage en modifiant les dernieres couches de l'encodeur pour les adapter à la tâche de classification.

Pré-entrainement du modèle de langue. Le pré-entraînement de l'encodeur BERT sur la tâche non supervisée de prédiction de mots masqués peut améliorer de manière significative la qualité des représentations apprises lorsque le domaine textuel est très particulier, comme c'est le cas dans notre application. Nous avons donc mis en place une précédure par entraîner un classifieur de mots $g_{\theta'}$ sur un corpus large et diversifié de textes biologiques, en particulier sur des données de descriptions d'espèces (différentes de celles du corpus utilisé pour le NER). Une fois quelques itérations effectuées, les paramètres de la couche de classification sont éliminés et les paramètres du modèle de langue θ' sont transférés en initialisation du modèle NER qui devient $f_{\theta'}$. Cette procédure permet au modèle NER de bénéficier du pré-entraînement.

3.3 Apprentissage auto-supervisé

La procédure d'auto-supervision est une approche itérative. Nous utilisons la stratégie décrite dans (Liang *et al.*, 2020).

Initialisation du *teacher*. Définissons d'abord un modèle de NER de référence f_θ^T entrainé sur notre jeu de données : ce modèle est appelé *teacher* (T). A l'itération 1, $f_\theta^{(T)}$ génère une liste de prédictions $\hat{Y} = \text{Softmax}(f_\theta^{(T)}(S)))$ associées aux phrases du corpus. Les prédictions dont le score de confiance dépassent un seuil fixé γ sont utilisées pour corriger les labels Y. Nous notons ce nouvel étiquetage $Y^{(1)}$.

$$\forall j,\ y_j^{(1)} = \begin{cases} y_j & \text{si argmax}(\hat{y}_j) = \text{argmax}(y_j) \text{ [bonne classification]} \\ \hat{y}_j & \text{si } \max(\hat{y}_j) > \gamma \text{ et argmax}(y_j) = 0 \\ y_j & \text{sinon} \end{cases} \tag{5}$$

Initialisation du *student*. Le modèle *student* $f_{\theta'}^{(S)}$ est appris sur ce jeu de données *corrigées*. En reprenant l'équation (4), l'intérêt de la procédure est plus clair : le fait de considérer la distribution

des scores sur y_j permet d'éviter des changements trop brusques dans l'étiquetage et de stabiliser l'évolution des modèles [1].

Itérations du *student*. Le modèle *student* $f_{\theta'}^{(S)}$ prédit successivement de nouvelles étiquettes $Y^{(t)}$, selon la procédure décrite en équation (5), puis met à jour ses poids θ' en exploitant $Y^{(t)}$.

Cette procédure présente un risque évident de dérive, le modèle se confortant progressivement dans des propositions fausses émanant de lui-même. Le modèle BOND (Liang *et al.*, 2020) propose de réinitialiser régulièrement les poids θ' du modèle à θ (les poids du *teacher*) : les étiquettes évoluent continuellement mais le risque de dérive est limité par un retour périodique au modèle d'origine.

Le nombre d'itérations $N_{self\text{-}training}$ et la période de retour à l'origine N_{reinit} sont des hyperparamètres très sensibles pour éviter la divergence.

3.4 Module de reconstruction des triplets

Sur la base de nos hypothèses concernant les schémas trouvés dans les descriptions d'espèces dans les corpus Flora neotropica (cf Section 3), nous avons conçu un module de reconstruction de triplets qui exploite les sorties du modèle d'extraction d'entités nommées. Nous supposons que, grâce à notre méthode d'échantillonnage, chaque phrase est centrée sur un organe ou un sous-organe particulier : il suffit alors de rattacher les descripteurs de la même phrase à cet organe.

$$\begin{aligned}&\{(w_{org}, y_{desc_0}, w_{desc_0}),\\ &(w_{org}, y_{desc_1}, w_{desc_1}), ...,\\ &(w_{org}, y_{desc_{N'}}, w_{desc_{N'}})\}\end{aligned} \tag{6}$$

Dans les rares cas où plusieurs organes sont détectés dans la même phrase, la meilleure solution consiste simplement à retenir la première détection dans l'ordre de la phrase.

4 Expériences

Dans cette section, nous présentons la campagne d'expériences concernant l'extraction des entités nommées. L'évaluation des triplets extraits est une tâche très importante pour les experts mais coûteuse en interventions humaines. Par conséquent, elle ne sera pas traitée dans cette étude.

Nous envisageons dans cette section plusieurs variantes de notre modèle pour la reconnaissance d'entités nommées (NER) sur les descriptions d'espèces.

Baseline : L'architecture de référence pour la reconnaissance des entités nommées utilisée est un modèle BERT-base pré-entraîné sur lequel est superposée une couche de classification avec une adaptation complète du modèle. Le pas d'apprentissage a été fixé à 10^{-6} et le nombre d'itérations nécessaire à la convergence du modèle est de 26. Le critère de convergence est calculé sur le score f1 en validation.

Baseline with pre-trained language model : Pour cette variante, nous pré-entraînons le modèle de langue de BERT-base sur l'ensemble des données de descriptions d'espèces en utilisant

1. Il s'agit d'une des variantes étudiées dans (Liang *et al.*, 2020), elle s'est révélée la plus performante sur nos données.

la tâche de prédiction de mots masqués. Nous ré-entrainons ensuite le modèle pour l'extraction d'entités nommées. Le pas d'apprentissage a été fixé à 10^{-6} et le nombre d'itérations nécessaire à la convergence du modèle est de 26 (score f1 maximal en validation).

Baseline with self training : Le modèle initial *teacher* est initialisé avec les paramètres du modèle **Baseline**. Le processus d'auto-apprentissage décrit dans la section 3.3 est ensuite appliqué pour entraîner le modèle. Le pas d'apprentissage a été fixé à 10^{-6} et le nombre d'itérations nécessaire à la convergence du modèle est de 4, le seuil de confiance γ a été fixé à 0.9 et le modèle *student* est réinitialisé 2 fois par epoch.

Baseline with self-training and language model pre-training : Le modèle initial *teacher* est cette fois initialisé avec les paramètres du modèle **Baseline with pre-trained language model** avant d'appliquer le processus d'auto-apprentissage. Le pas d'apprentissage a été fixé à 10^{-6} et les performances du modèle en validation ont commencé à chuter au bout de 3 itérations, le seuil de confiange γ a été fixé à 0.9 et le modèle *student* est réinitialisé 2 fois par epoch.

Afin d'évaluer l'efficacité des méthodes , nous calculons le score F1, la précision et le rappel pour chaque variante. Pour une classe de données c, nous calculons classiquement :

$$F1_c = 2 \cdot \frac{\text{Precision} \cdot \text{Rappel}}{\text{Precision} + \text{Rappel}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

en utilisant les abbréviations anglaise ; TP (vrai positif), FP (faux positif), FN (faux négatif). Nous calculons ensuite le score F1 micro pondéré par le poids des classes :

$$F1_{weighted} = \frac{\sum_{c=1}^{N_{cl}} N_{occ_c} \times F1_c}{\sum_{c=1}^{N_{cl}} N_{occ_c}} \tag{8}$$

Où N_{cl} représente le nombre total de classes et N_{occ_c} représente le nombre d'occurences de la classe c dans le jeu de données de test.

Classe 0 *vs* organes et descripteurs. Comme c'est toujours le cas en détection d'entités nommées, la classe 0, ultra-majoritaire, est bien entendu exclue du calcul $F1_{weighted}$.

Détection *vs* classification. Chaque métrique peut être calculée à deux niveaux : la **détection** (niveau le plus lâche) consiste à quantifier les termes qui ont été détectés, indépendemment de la classe affectée. La **classification** représente dans cet article la métrique la plus stricte, intégrant à la fois la détection et la bonne classification des termes.

Ces mesures fournissent une évaluation de la capacité du modèle à identifier et à extraire avec précision des entités à partir de données textuelles, ainsi que sa sensibilité aux entités nouvelles ou inédites. Il faut cependant garder à l'esprit l'étiquetage distant et par conséquent le risque –léger– d'ambiguité sur les étiquettes qui devrait affecter le rappel ainsi que le risque –très fort– d'entités non annotées qui devrait affecter le rappel.

La plupart des ensembles de données de référence en NER présentent un biais de chevauchement entre les termes apparaissant à la fois en entraînement et en test. Ce défaut rend l'interprétation des résultats ambigüe puisque nous ne pouvons pas conclure si le modèle est capable de détecter de nouvelles entités ou si il ne fait que de la mémorisation des entités déjà vues (Taillé *et al.*, 2021).

Afin de limiter ce phénomène, nous proposons de tester nos modèles sur deux ensembles : l'ensemble X composé de phrases nouvelles mais contenant des entités déjà vues en apprentissage (biais de

TABLE 2 – Nombres d'occurences des classes sur les jeux de test X, X_c (hors distribution) ainsi que pour le jeu d'apprentissage

Ensemble	Classe	X	X_c	Jeu d'apprentissage
$\mathcal{Y}_1$	Flower	2988	2800	11242
	Fruit	753	0	3004
	Habit	345	0	1270
	Leaf	618	0	2257
	Part-of	2950	1689	11110
	Stem-root	536	1887	2046
$\mathcal{Y}_2$	Color	1760	5174	6536
	Disposition	929	1210	3537
	Form	2415	745	8630
	Position	1627	0	5932
	Surface-texture	2024	0	7543

chevauchement) et l'ensemble X_c contenant des phrases nouvelles et exclusivement des entitées nommées qui ne figurent pas dans le jeu d'apprentissage. Nous considérons l'ensemble X_c comme un "ensemble de test hors distribution", cet ensemble permettra de mesurer la quantité de nouvelles entités détectées par le modèle. Nous faisons l'hypothèse que cette mesure est très corrélée avec le nombre d'annotations manquantes corrigées par le modèle. Le tableau 2 représente les statistiques des deux jeux de test proposés. Certaines classes ne sont pas représentées dans le jeu de test hors distribution X_c par manque de représentativité dans le jeu de données initial (voir Tableau 1).

4.1 Capacité du modèle à généraliser en fonction du contexte

Dans cette partie les différentes variantes du modèle sont testées sur l'ensemble X qui contient des phrases qui n'ont pas été vues pendant la phase d'apprentissage, mais qui contiennent toujours des entités présentes dans l'ensemble d'apprentissage. Cela nous permet de mesurer la capacité du modèle à s'adapter à de nouveaux contextes et à de nouvelles phrases.

TABLE 3 – Capacité du modèle à détecter et à classifier des entités vues en apprentissage dans un contexte différent (scores en Détection/Classification)

Modèles	Rappel	Précision	Score F1
Baseline	100/92.25	83.87/77.19	91.22/83.07
Baseline w/ lm	100/93.30	84.94/78.99	91.86/84.70
Baseline w/self-train	100/96.34	88.13/84.67	93.69/89.48
Baseline w/ lm_self-train	100/95.30	85.90/81.61	92.41/87.29

Le tableau 3 montre les performances de notre modèle NER avec et sans pré-entraînement du modèle de langue. Nous avons observé une amélioration significative des trois mesures d'évaluation avec le pré-entrainement, y compris une augmentation de 1.63 % du score F1, une augmentation de 1.05 % du rappel et une augmentation de 1.8% du de la précision, en classification. Ces résultats indiquent

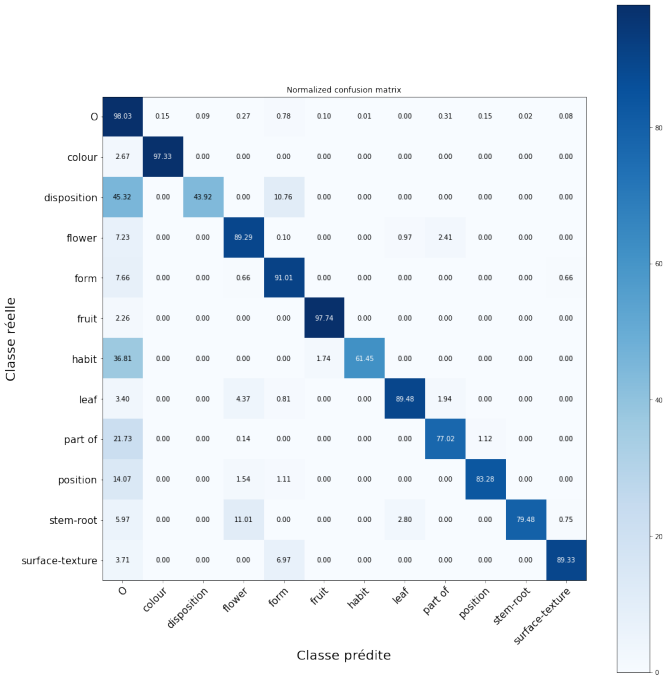


FIGURE 3 – Matrice de confusion du modèle avec auto-apprentissage sur le jeu de données X

que le pré-entraînement du modèle de langue sur un corpus large et diversifié de textes biologiques est une méthode prometteuse pour améliorer les performances du modèle NER sur des données en domaine clos, comme c'est le cas pour les descriptions d'espèces.
Nos expériences sur l'effet de l'auto-apprentissage ont également donné des résultats encourageants. Notre modèle auto-entraîné a atteint un score F1 de 89.48%, surpassant le score F1 du modèle de référence de 4.22%. Ces résultats soulignent l'efficacité de l'auto-apprentissage pour surmonter l'impact négatif de l'annotation distante (bruit et silence sur les étiquettes). Nous pouvons en conclure que l'auto-apprentissage est une technique utile pour améliorer la performance des modèles NER dans les scénarios où les données étiquetées sont rares ou de mauvaise qualité.
Cependant la combinaison d'auto-apprentissage et pré-entrainement du modèle de langue ne semble pas apporter de gain notable sur les performances en détection et en classification, ce qui peut être expliqué par le fait que la combinaison des deux algorithmes conduit à un sur-ajustement aux données.

Sur l'ensemble des expériences, nous notons un écart très significatif entre la détection et la classification. Nous attribuons cet écart au manque de données étiquetées mais nous sommes convaincus qu'il serait possible de le résorber en utilisant par exemple de l'augmentation de données. Il s'agit d'une perspective intéressante pour ce travail.

La matrice de confusion de la figure 3 montre en particulier des erreurs de classification pour la classe **Disposition**, qui a le plus de faux négatifs. Cette classe est souvent confondue avec la classe **O**, la plus représentée du jeu de données d'apprentissage, ainsi qu'avec la classe **Form**, qui est la mieux représentée parmi les classes d'entités. Cependant, il est important de noter que les confusions entre classes de descripteurs et classes d'organes sont très faibles, ce qui indique que le modèle a réussi à assimiler ces notions. Ce dernier point est crucial pour la qualité des triplets extraits par la suite.

FIGURE 4 – Distributions de probabilités du modèle de référence et du modèle avec auto-apprentissage pour la détection d'une entité appartenant à la classe **Part-of**. Jeu de données X_c

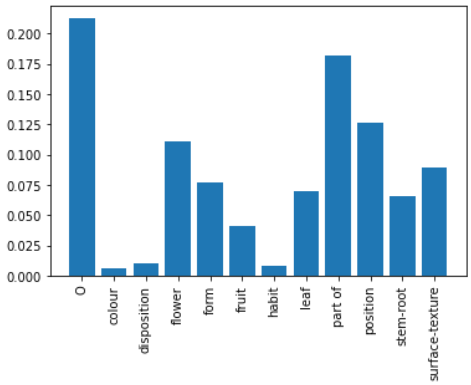


FIGURE 5 – Modèle de référence

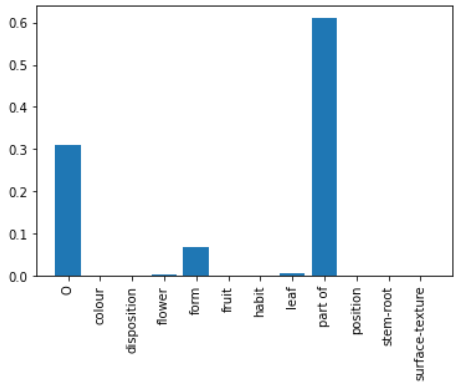


FIGURE 6 – Modèle avec auto-apprentissage

4.2 Capacité des modèles à généraliser sur de nouvelles entités

Le tableau 4 représente les performances des différentes variantes sur l'ensemble X_c qui contient des entités entièrement nouvelles qui n'étaient pas présentes dans l'ensemble d'apprentissage (**hors distribution**). Ce cadre expérimental est très difficile mais il est particulièrement important car il permet d'évaluer la capacité du modèle à détecter et à classifier de nouvelles entités.

TABLE 4 – Capacité des modèles à détecter et à classifier de nouvelles entités, hors de la distribution de l'ensemble d'apprentissage. (Scores en Détection/Classification)

Modèles	Rappel	Précision	Score F1
Baseline	100/71.07	76.79/65.45	86.87/66.62
Baseline w/ lm	100/71.03	83.87/65.16	91.22/66.46
Baseline w/self-train	100/82.11	58.38/47.86	73.75/50.54
Baseline w/ lm_self-train	100/81.54	65.55/55.32	79.19/58.66

Dans le cas où l'ensemble de test ne contient que des entités nommées nouvelles, qui n'ont pas été vues pendant l'apprentissage, le rappel est l'indicateur clé pour mesurer la capacité du modèle à détecter ces nouvelles entités. En effet, le rappel mesure la proportion d'entités nommées prédites par le modèle. Ainsi, la mesure du rappel est plus pertinente que la précision ou le score F1 dans ce contexte particulier.

Dans l'étude mentionnée, le modèle de référence a obtenu un rappel de 71,07% sur l'ensemble de test contenant de nouvelles entités.. Lorsque l'auto-apprentissage a été appliqué au modèle, le rappel a augmenté de manière significative pour atteindre 82.11%. Cependant, le pré-entrainement du modèle de langue n'a pas apporté de gain en rappel. Ce qui démontre l'innefficacité de cette technique pour généraliser sur de nouvelles entités.

Le modèle avec auto-apprentissage est donc celui qui obtient la meilleure précision sur les entités nouvelles.

Qualitativement nous observons sur la figure 4 un exemple d'entité nommée dont la détection a été corrigée par l'auto-apprentissage, l'entité **lobe** appartenant à la classe **Part-of** n'était pas présente dans le jeu d'apprentissage. Nous pouvons clairement observer la baisse d'entropie engendrée par l'auto-apprentissage et son effet bénéfique sur cet exemple.

Il convient de rappeler que tous les modèles ont été entraînés sur un ensemble de données annoté de manière distante, ce qui pourrait entraîner un biais de rétention plus important vers les entités les plus fréquentes. Par conséquent, les améliorations observées dans les modèles avec auto-apprentissage et pré-entraînement du modèle de langue sont encore plus impressionnantes, car elles montrent la capacité de ces techniques à améliorer les performances du modèle sur des entités nouvelles et non vues.

5 Discussion

Dans le contexte des deux ensembles de test, l'un contenant uniquement de nouvelles entités et l'autre contenant de nouvelles phrases avec des entités vues lors de la phase d'apprentissage, il convient de noter que le préentraînement du modèle de langue et l'auto-apprentissage ont tous deux permis d'améliorer les performances du modèle d'extraction d'entités nommées comparé au modèle de référence sur les deux ensembles de données. Cela s'explique par la spécificité du vocabulaire spécialisé et de la forme des phrases. L'auto-apprentissage est plus difficile à régler mais il permet clairement d'augmenter les détections et d'améliorer le rappel au fil des itérations. La question du critère d'arrêt sur cette phase est difficile car la précision n'est pas complètement fiable, un certain nombre de termes pertinents étant probablement étiquetés 0 du fait de l'annotation distante. Il sera nécessaire d'échanger avec les experts du domaine sur la base des prédictions pour trouver le meilleur compromis.

La combinaison du pré-entraînement et de l'auto-entraînement n'a pas permis d'améliorer davantage les performances sur l'un ou l'autre des ensemble de tests et les performances étaient même légèrement inférieures à celles du modèle utilisant uniquement l'auto-apprentissage. En analysant les performances des deux modèles sur de nouvelles entités, les résultats laissent penser qu'une des explication possibles serait que le pré-entrainement du modèle de langue bien qu'apportant une notion de contexte permettant une bonne précision, il ne permet pas généraliser sur des entités nouvelles et combiné à l'auto-apprentissage il risque d'apporter une redondance d'informations et mener à un sur-apprentissage.

Dans l'ensemble, le modèle le plus performant sur les deux ensembles de tests est celui qui a été formé avec l'auto-apprentissage seul. Cela indique que dans le contexte de l'apprentissage supervisé à distance avec des données d'apprentissage limitées, l'auto-apprentissage peut être une technique puissante pour améliorer les performances des modèles de langue dans des contextes de supervision distante et ce même sur des données de domaine clos telles que les descriptions d'espèces.

6 Conclusion

Notre étude visait à améliorer les performances des modèles d'extraction de connaissances pour l'analyse des descriptions d'espèces biologiques. Nous avons proposé un modèle supervisé à distance pour la reconnaissance des entités nommées et décrit un protocole pour la construction de graphes de connaissances à partir de l'étiquetage des entités. Pour évaluer nos modèles avec précision, nous avons proposé un protocole de test consistant en deux ensembles de données, l'un contenant les entités vues pendant l'entraînement et l'autre contenant de nouvelles entités.

Nous avons identifié deux défis scientifiques : la spécificité du vocabulaire et des tournures de phrases, et les annotations manquantes. Pour résoudre le problème du vocabulaire, nous avons proposé une technique de pré-entraînement du modèle de langage qui a amélioré les performances de notre modèle NER le premier ensemble, sans apporter de gain sur le second. Pour le problème des annotations manquantes, nous avons proposé une architecture teacher-student formulée sous forme d'auto-apprentissage, qui a permis d'obtenir le rappel le plus élevé sur les deux ensembles de données.

Pour ce dernier cas de figure, il est d'une part difficile de bien régler ce modèle qui tend à diverger. D'autre part, il est difficile d'interpréter finement les résultats car l'absence d'annotation fiable débouche malheureusement sur différentes interprétations.

En conclusion, notre étude démontre l'apport des modèles de langue récents pour l'analyse de textes complexes et spécifiques. Cette application est vraiment critique pour les chercheurs qui étudient la diversité et l'évolution des espèces, avec des applications potentielles en morphologie comparative et en informatique de la biodiversité. Si l'étude présente ne permet pas de lever tous les verrous scientifiques, elle a contribué largement à inciter les chercheurs du domaine à approfondir ce sujet de recherche : il semble clair aujourd'hui que la construction automatique d'un graphe de connaissances en biologie des espèces est un objectif atteignable à moyen terme.

Au niveau des perspectives, l'auto-apprentissage semble perfectible en introduisant de nouvelles hypothèses et en testant de nouvelles approches. Nous envisageons aussi d'utiliser des modèles de langues plus larges qui semblent encore plus performants en extraction d'entités nommées.

Références

CHO H. & LEE H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. **20**(1), 735. DOI : 10.1186/s12859-019-3321-4.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. Number : arXiv :1810.04805.

JIA C., SHI Y., YANG Q. & ZHANG Y. (2020). Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6384–6396. DOI : 10.18653/v1/2020.emnlp-main.518.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, p. 260–270.

LEE J., PHAM L. H. & UZUNER O. (2022). Mnlp at fincausal2022 : Nested ner with a generative model. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, p. 135–138.

LEITNER E., REHM G. & MORENO-SCHNEIDER J. (2019). Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs : 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings*, p. 272–287 : Springer.

LIANG C., YU Y., JIANG H., ER S., WANG R., ZHAO T. & ZHANG C. (2020). BOND : BERT-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 1054–1064 : ACM. DOI : 10.1145/3394486.3403149.

MAGGE A., SCOTCH M. & GONZALEZ-HERNANDEZ G. (2018). Clinical ner and relation extraction using bi-char-lstms and random forest classifiers. In *International workshop on medication and adverse drug event detection*, p. 25–30 : PMLR.

MALOUF R. (2002). Markov models for language-independent named entity recognition. In *COLING-02 : The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 188–191.

MENG Y., ZHANG Y., HUANG J., WANG X., ZHANG Y., JI H. & HAN J. (2021). Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. Number : arXiv :2109.05003.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics : Human language technologies*, p. 746–751.

MIWA M. & BANSAL M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1105–1116.

SAUCÈDE T., ELÉAUME M., JOSSART Q., MOREAU C., DOWNEY R., BAX N., SANDS C., MERCADO B., GALLUT C. & VIGNES-LEBBE R. (2021). Taxonomy 2.0 : computer-aided identification tools to assist antarctic biologists in the field and in the laboratory. **33**(1), 39–51. Publisher : Cambridge University Press, DOI : 10.1017/S0954102020000462.

SOUZA F., NOGUEIRA R. & LOTUFO R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv :1909.10649*.

TAI K. S., SOCHER R. & MANNING C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. DOI : 10.48550/arXiv.1503.00075.

TAILLÉ B., GUIGUE V. & GALLINARI P. (2020). Contextualized embeddings in named-entity recognition : An empirical study on generalization. In *Advances in Information Retrieval : 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, p. 383–391 : Springer.

TAILLÉ B., GUIGUE V., SCOUTHEETEN G. & GALLINARI P. (2021). Separating retention from extraction in the evaluation of end-to-end relation extraction. Number : arXiv :2109.12008.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VIGNES-LEBBE R., BOUQUIN S., KERNER A. & BOURDON E. (2017). Desktop or remote knowledge base management systems for taxonomic data and identification keys : Xper2 and xper3. **1**, e19911. Publisher : Sofia : Pensoft Publishers, 2017-, DOI : 10.3897/tdwgproceedings.1.19911.

WANG X., HU V., SONG X., GARG S., XIAO J. & HAN J. (2021). Chemner : fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

WANG X., SONG X., LI B., ZHOU K., LI Q. & HAN J. (2020). Fine-grained named entity recognition with distant supervision in covid-19 literature. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 491–494 : IEEE.