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

Reconnaissance d'Entités Nommées fondée sur des Modèles de Langue Enrichis avec des Définitions de Types d'Entités

Jesús Lovón-Melgarejo[♠] Jose G. Moreno[♠] Romaric Besançon[♢]
Olivier Ferret[♢] Lynda Tamine[♠]
[♠]Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France
[♢]Université Paris-Saclay, CEA, List, Palaiseau, France

RÉSUMÉ

Des études récentes ont identifié de nouveaux défis dans la tâche de reconnaissance d'entités nommées (NER), tels que la reconnaissance d'entités complexes qui ne sont pas des syntagmes nominaux simples et/ou figurent dans des entrées textuelles courtes, avec une faible quantité d'informations contextuelles. Cet article propose une nouvelle approche qui relève ce défi, en se basant sur des modèles de langues pré-entraînés par enrichissement des définitions des types d'entités issus d'une base de connaissances. Les expériences menées dans le cadre de la tâche MultiCoNER I de SemEval ont montré que l'approche proposée permet d'atteindre des gains de performance par rapport aux modèles de référence de la tâche.

ABSTRACT

Named Entity Recognition based on Language Models Enriched with Entity Type Definitions

Recent studies have identified new challenges in the Named Entity Recognition (NER) task, such as recognizing complex entities that are not simple noun phrases, and/or occur in short text inputs with limited context information. This paper proposes a novel approach relying on pretrained language models (PLM) that leverage entity type definitions from knowledge bases. The experiments conducted on the MultiCoNER task of SemEval showed that the proposed approach enhances the model's performance and shows consistent gains compared to the task baselines.

MOTS-CLÉS : reconnaissance d'entités nommées, entités complexes, SemEval, type d'entité.

KEYWORDS: named entity recognition (NER), complex entities, SemEval, entity type.

1 Introduction

La Reconnaissance d'Entités Nommées (NER, *Named Entity Recognition*) (Grishman & Sundheim, 1996) consiste à détecter des groupes de mots en tant qu'entités nommées dans une phrase donnée et à identifier leur type à partir d'une liste a priori de types d'entités. Selon la taille de cette liste, la tâche de NER est classée en i) *granularité grossière* lorsque la liste des types est petite (comme les noms de personnes, d'organisations et de lieux, tels que proposés dans la tâche CoNLL (Tjong Kim Sang & De Meulder, 2003)) ou ii) *granularité fine* pour une liste plus étendue (Ling & Weld, 2021). De plus, nous pouvons distinguer différentes classes d'entités nommées, qualifiées comme traditionnelles (ex. Personne, Lieu) et non traditionnelles (ex. titres d'œuvres, tels qu'un livre ou une chanson).

Des études récentes ont identifié de nouveaux défis dans la tâche de NER (Meng *et al.*, 2021), particulièrement pour les entités non traditionnelles, dites complexes (Ashwini & Choi, 2014) car très évolutives et non identifiables par des expressions nominales (ex. titres de films comme *Avatar*) ou alors présentes dans une phrase textuelle courte présentant un faible contexte (Jayarao *et al.*, 2018). Les systèmes classiques de NER ne sont pas performants dans ces cas de figure, ne permettant pas ainsi de relever ces défis (Fetahu *et al.*, 2022). Par conséquent, de nouveaux cadres d'évaluation, dont MultiCoNER dans SemEval, ont été proposés pour évaluer les performances des modèles dans ces conditions difficiles de NER (Malmasi *et al.*, 2022b). Parmi les approches proposées dans le but de relever ce défi, on trouve l'utilisation de techniques d'augmentation de données (Gan *et al.*, 2022) ou l'augmentation de la phrase d'entrée en s'appuyant sur des bases de connaissances (Wang *et al.*, 2022). Ces travaux se concentrent sur l'exploitation de représentations contextualisées de la phrase d'entrée, qui impliquent des coûts élevés pour l'encodage de représentations textuelles de longueur importante. Cependant, à notre connaissance, les travaux à ce jour ne se sont pas focalisés sur la représentation des types d'entités. Dans ce travail, nous soutenons l'idée que les représentations contextualisées des types d'entités peuvent influencer positivement les performances d'un modèle de NER, en particulier pour une taxonomie d'entités à granularité fine, où certaines classes sont généralement sous-représentées.

Cet article propose ainsi une nouvelle approche qui exploite des *définitions des types d'entités* contextualisées, riches et pertinentes, pour enrichir un modèle de langue pré-entraîné (PLM) utilisé comme une base d'un classifieur NER. Nous créons manuellement ces définitions pour plusieurs langues et proposons une architecture de modèle pour exploiter ces représentations. Notre intuition est d'associer les entités de la phrase d'entrée à une définition détaillée et contextualisée de leurs types d'entités, conduisant à un espace de représentation adapté où les entités partageant le même type auront des représentations plus proches. Notre modèle comprend deux configurations différentes pour l'entraînement et le test. Pour la configuration d'entraînement, nous avons entraîné un PLM, XLM-Roberta, qui calcule des représentations contextualisées pour la phrase et les *définitions des types d'entités* associés obtenus à partir des annotations. Ensuite, nous avons aligné et agrégé les deux types de représentations. Les représentations finales des mots (*token*) sont ensuite transmises en entrée d'une couche de type champ aléatoire conditionnel en chaîne linéaire (CRF, *Conditional Random Fields)* (Lafferty *et al.*, 2001) pour la prédiction des entités nommées. Pour la configuration de test, nous évaluons notre modèle en n'utilisant que le PLM enrichi et la couche CRF. Nous avons testé nos modèles pour quatre langues : l'anglais, l'espagnol, le français et le portugais. Nos expérimentations ont montré que l'injection de cette connaissance aide à améliorer la performance du modèle et montre un gain constant par rapport à un modèle standard entraîné pour la tâche.

2 Travaux connexes

Ces dernières années, les PLM tels que BERT (Devlin *et al.*, 2018) ou XLM-RoBERTa (Conneau *et al.*, 2020) ont démontré leur efficacité pour améliorer les performances de la reconnaissance d'entités nommées dans plusieurs cadres d'évaluation (Jayarao *et al.*, 2018; Wang *et al.*, 2022; Jayarao *et al.*, 2018). Cependant, des études récentes ont montré que ces méthodes présentent des performances limitées pour faire face aux défis du monde réel (Meng *et al.*, 2021), tels que les entrées textuelles courtes avec un contexte limité (Jayarao *et al.*, 2018) et les entités complexes, dont les nouvelles entités émergentes telles que les titres de livres, films et chansons qui sont publiés chaque semaine (Fetahu *et al.*, 2022). Par conséquent, des travaux récents ont proposé de nouveaux cadres d'évaluation

pour relever ces défis, dont la tâche MultiCoNER dans SemEval (Malmasi *et al.*, 2022a).

L'ajout de contexte pertinent à la phrase d'entrée avec des ressources lexicales externes, telles que des bases de connaissances, a contribué à créer des représentations de *tokens* améliorées qui ont eu un impact positif sur les performances de la tâche de NER (Jayarao *et al.*, 2018; Wang *et al.*, 2022). De même, des travaux antérieurs ont proposé de pré-entraîner des PLM enrichis en fusionnant des représentations avec celles issues des bases de connaissances pour améliorer les représentations de *tokens*, ce qui s'est avéré utile dans des tâches de NLP connexes (Zhang *et al.*, 2019; Peters *et al.*, 2019). Néanmoins, ces techniques sont coûteuses en termes de calcul en raison des architectures neuronales supplémentaires impliquées.

Une autre ligne de travaux a exploré l'utilisation de PLM comme bases de connaissances pour extraire (Petroni *et al.*, 2019) et injecter des faits (Talmor *et al.*, 2020) de ces ressources externes en les transformant en énoncés textuels et en appliquant la tâche de pré-entraînement du PLM à ces énoncés. Sur la base de ces approches, nous proposons une nouvelle méthode qui utilise une *définition de type d'entité* comme représentation textuelle d'un fait de base de connaissances. Nous visons à extraire puis injecter cette information pour enrichir un PLM dans le cadre d'une configuration d'entraînement peu coûteuse.

Classe	Lang.	Définition du type d'entité
Medicine-Symptom	EN	A symptom is any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease. A symptom is a medical term.
	ES	Un síntoma es cualquier sensación o cambio en la función corporal que experimenta un paciente y que se asocia a una enfermedad concreta. Un síntoma es un término médico.
	FR	Un symptôme est toute sensation ou modification d'une fonction corporelle ressentie par un patient et associée à une maladie particulière. Un symptôme est un terme médical.
	PT	Um sintoma é qualquer sensação ou mudança na função corporal que é experimentada por um paciente e está associada a uma determinada doença. Um sintoma é um termo médico.

TABLE 1 – Définitions des types d'entités créés pour les classes de granularité fine Symptôme en anglais (EN), espagnol (ES), français (FR) et portugais (PT).

3 Méthodologie

Cette section présente le système global de notre modèle NER ainsi que sa mise en œuvre détaillée, y compris la construction de définitions de types d'entités et l'architecture du modèle.

3.1 Motivations

Dans une tâche de NER, chaque mot de la phrase d'entrée est associé à un type d'entité appartenant à une liste de candidats prédéfinis. Ces candidats sont généralement associés à des catégories d'entités

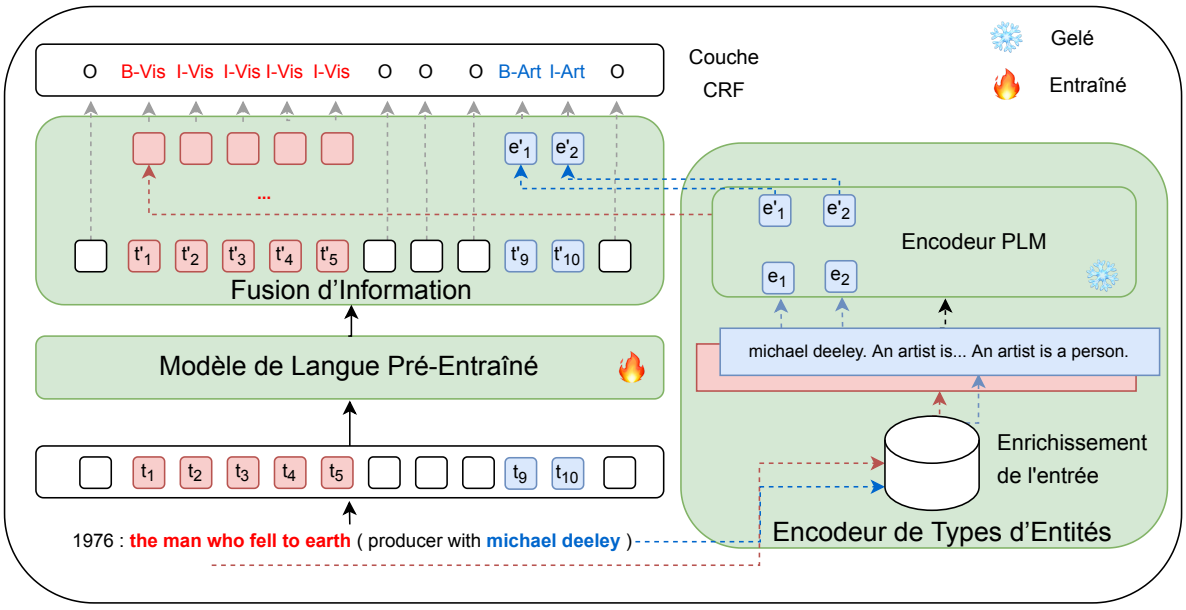

FIGURE 1 – Illustration de l'architecture du modèle.

bien définies. Cependant, quand le nombre de candidats augmente, les différences entre les catégories deviennent moins importantes, ce qui rend plus difficile leur distinction. Par exemple, il est plus facile de distinguer une entité nommée catégorisée comme *Personne* parmi une petite liste de trois autres types (ex. *Lieu*, *Organisation*, *Alimentation*) que de distinguer une entité étiquetée comme *Maladie* parmi une liste de 40 types où d'autres types, tels que *Symptôme*, sont présents.

Alors que les travaux antérieurs se sont principalement concentrés sur l'amélioration de la phrase d'entrée et des caractéristiques associées, le présent travail vise à intégrer des caractéristiques essentielles des candidats de type d'entité dans le processus d'apprentissage, ce qui pourrait profiter aux systèmes de NER. Dans ce travail, nous proposons d'utiliser des *définitions des types d'entités* (cf. Tableau 1). Une telle *définition de type d'entité* prend ici la forme d'une description textuelle du type. De plus, dans le cas des taxonomies à granularité fine, nous étendons cette définition avec une description fondée sur l'hyperonymie pour capturer la structure hiérarchique de la taxonomie. En exploitant cette information, nous visons à coder des représentations sémantiques plus riches et donc plus faciles à catégoriser comparativement à des représentations non contextualisées.

3.2 Architecture du modèle

Nous décrivons maintenant notre système NER, qui se compose de trois modules : le module *Modèle de Langue Pré-Entraîné* (PLM$_{NER}$), le module *Encodeur de Types d'Entités* et le module *Fusion d'Informations*, comme le montre la Figure 1. Notre approche consiste à adapter un PLM selon l'architecture de la Figure 1 pour améliorer les représentations du modèle pour la tâche. Mais nous nous appuyons directement sur ce seul PLM adapté pour effectuer l'inférence. Nous prenons en entrée une phrase (t) composée de n *tokens* $t = \{t_1, t_2, ..., t_n\}$ pour le module PLM$_{NER}$ et le module d'encodage des types d'entités. Cet encodeur étend d'abord les mentions d'entités de la phrase d'entrée en les concaténant avec les *définitions des types* construits manuellement, puis calcule leurs représentations. Ces représentations sont alignées avec les représentations des mots issues du module PLM$_{NER}$ puis combinées à l'aide du module de fusion d'information. Les représentations combinées

sont alors transmises à une couche CRF qui produit les prédictions des types d'entités.

Nous décrivons ci-après la nature et le rôle de chacun des modules composant le modèle proposé.

Modèle de langue pré-entraîné Ce module se réduit à son seul constituant, en l'occurrence un PLM. Étant donné une phrase d'entrée t, nous utilisons ce PLM pour calculer un ensemble de représentations contextualisées $\{t'_1, t'_2, ..., t'_n\}$, où chaque t'_i correspond à la représentation contextualisée du *token* t_i de la phrase en entrée t.

Encodeur de types d'entités À partir de l'entrée annotée de l'ensemble d'entraînement, nous identifions les mentions d'entités (représentée par une suite de *tokens* $\{e_1, e_2, .., e_m\}$) et leur type d'entité annotée. Nous étendons ensuite chaque mention d'entité en concaténant la *définition de type d'entité* correspondant (représentée par les *tokens* $\{w_1, w_2, .., w_n\}$). Enfin, le nom du type d'entité et la définition construite créent une entrée de la forme : $\{e_1, e_2, ..., w_1, ..., w_n\}$. Nous fournissons cette entrée à l'encodeur PLM de ce module pour calculer les représentations des *tokens* et ne sélectionner que les *tokens* appartenant à la mention d'entité $\{e'_1, e'_2, ...\}$.

À titre d'exemple, la Figure 1 considère la phrase d'entrée q=*"1967 : the man who fell to earth (producer with Michael Deeley)"*. Cette phrase a deux types d'entités reconnus : *Œuvre Visuelle (Vis)*, qui est un type de *Travail Créatif*, et *Artiste (Art)*, qui est un type de *Personne*, correspondant aux sous-textes *"the man who fell to earth"* et *"michael deeley"*, respectivement. L'encodeur de types d'entités génère deux phrases en concaténant les définitions des types aux entités. Plus précisément, les phrases générées sont q_1=*"the man who fell to earth. A visual work is [définition]. A visual work is a creative work."* et q_2=*"michael deeley. An artist is [définition]. An artist is a person."*. Ensuite, q_1 et q_2 sont passées dans un encodeur PLM pour obtenir les représentations correspondant aux mentions d'entités enrichies.

Ce module vise à améliorer les représentations de *tokens* en intégrant des informations contextuelles à partir des types d'entités annotés. Comme les représentations de l'encodeur PLM de ce module sont généralement informatives, nous n'entraînons pas les éléments de ce module en gelant les paramètres de ce PLM.

Fusion d'information Nous calculons une représentation agrégée en utilisant le module PLM$_{NER}$ et l'encodeur de types d'entités. Tout d'abord, pour chaque mention d'entité, nous alignons les représentations des *tokens* en sortie de l'encodeur de types avec celles en sortie du module PLM$_{NER}$. Nous ajoutons des vecteurs nuls pour les sous-*tokens* qui n'appartiennent pas aux entités. Nous effectuons ensuite une moyenne pondérée pour calculer les représentations finales :

$$(1 - b) * t'_i + b * e'_i \tag{1}$$

où *b* est un hyperparamètre avec des valeurs entre $[0, 1]$.

Enfin, les représentations agrégées sont fournies en entrée d'une couche CRF pour prédire la meilleure séquence de types d'entités parmi toutes les séquences possibles.

4 Cadre expérimental

4.1 Jeux de données

Nous avons utilisé deux jeux de données récents pour nos évaluations, les ensembles MultiCoNER I (Malmasi *et al.*, 2022a) et MultiCoNER II (Fetahu *et al.*, 2023). Les ensembles de données sont issus de trois domaines : des phrases d'encyclopédie, des questions de QA et des requêtes Web. Ces ensembles de données fournissent des phrases à faible contexte et difficiles, dans plusieurs langues, en incluant des entités complexes. La principale différence entre ces ensembles de données réside dans la taille de l'ensemble de types. L'ensemble de données MultiCoNER I fournit un ensemble de types d'entités à granularité grossière de six types, tandis que l'ensemble de données MultiCoNER II fournit une taxonomie de granularité fine avec 36 types. Chaque ensemble de données suit le format CoNLL et les annotations des types d'entités suivent un schéma BIO.

4.2 Modèles de référence

Nous avons utilisé les modèles de référence fournis avec les jeux de données, qui s'appuient sur le modèle XLM-Roberta avec une couche CRF. Nous adoptons la dénomination *XLM-RoB* dans le reste de l'article pour désigner ces modèles de référence. *XLM-RoB* est une variante multilingue du modèle RoBERTa qui a été pré-entraînée pour plus de 100 langues. *XLM-RoB* fonctionne en générant une représentation pour chaque *token*, qui est ensuite utilisée pour prédire le type du *token* à l'aide d'une couche de classification CRF. Il convient de noter que ces modèles de référence ont été affinés avec des hyperparamètres partagés pour toutes les langues [1]. Nous avons affiné notre version de XLM-RoB, sous le nom de *XLM-RoB*$_{nous}$, en utilisant la version *large* du modèle ainsi que des hyperparamètres spécifiques pour chaque langue.

Nous avons également considéré comme référence le système le plus performant pour *MultiCoNER I*, *DAMO-NLP* (Wang *et al.*, 2022). Ce dernier est fondé sur un modèle XLM-RoBERTa large et sur un extracteur de faits à partir de bases de connaissances pour ajouter du contexte pertinent à l'entrée, par exemple un paragraphe de *Wikipédia*. À l'heure actuelle, les systèmes *MultiCoNER II* ne sont pas encore disponibles, ce qui ne nous permet pas de considérer un modèle plus performant.

4.3 Modèles proposés

À l'instar de Wang *et al.* (2022), notre système utilise XLM-Roberta comme modèle de base en raison de ses performances pour cette tâche (Malmasi *et al.*, 2022c), plus précisément dans sa version *xlm-roberta-large*, disponible au niveau du *model hub* de Hugging Face. Nous avons entraîné six modèles, un pour chaque ensemble de données monolingue : anglais et espagnol pour *MultiCoNER I*, et anglais, espagnol, portugais et français pour *MultiCoNER II* [2]. Nous avons utilisé des valeurs de $b = 0{,}15$ pour l'entraînement, avec un taux d'apprentissage de 2×10^{-5} pour l'espagnol, le français et le portugais, et 1×10^{-5} avec $b = 0{,}1$ pour l'anglais. Nous avons entraîné pendant 5 époques

1. Les scores rapportés ont été extraits de https://competitions.codalab.org/competitions/36044 et https://github.com/modelscope/AdaSeq/tree/master/examples/SemEval2023_MultiCoNER_II

2. Le français et le portugais ne sont pas disponibles dans MultiCoNER I.

avec un *batch size* de 32, sur une carte graphique Nvidia RTX6000. Le temps d'entraînement était d'environ 3 heures par modèle. Nous avons utilisé une valeur de $b = 0$ pour les tests. Conformément aux travaux antérieurs (Malmasi *et al.*, 2022a), nous avons principalement rapporté le score F1, calculé pour l'ensemble complet des étiquettes.

5 Résultats

Nous avons analysé nos résultats sur les ensembles de test *MultiCoNER I* et *II*. Pour le premier, nous avons entraîné cinq modèles avec des graines aléatoires différentes (*random seeds*) et avons rapporté la moyenne de leurs scores. En revanche, pour le second, nous n'avons entraîné qu'un seul modèle en raison d'un accès limité et d'un nombre limité d'évaluations au moment de la rédaction de cet article. Les performances de notre modèle sur les deux ensembles de données et les améliorations apportées par rapport aux modèles de référence sont présentées dans le Tableau 2. Nous considérons deux modèles de référence distincts fondés sur le modèle RoBERTa : *XML-RoB* est le modèle de référence officiel de l'évaluation *MultiCoNER I* ; *XLM-Rob*$_{nous}$ est un modèle que nous avons ré-entraîné et appliqué à *MultiCoNER II* et qui constitue une référence plus robuste.

Nous avons obtenu des améliorations pour les deux jeux de données grâce à notre approche. Par rapport au modèle de référence, *XLM-RoB*, nous avons observé une amélioration de $+4{,}5$, $+6{,}3$ pour *MultiCoNER I* en anglais et en espagnol, respectivement. De même, dans les deux versions de *MultiCoNER*, nous avons constaté des améliorations entre $+0{,}3$ et $+1{,}1$ par rapport au modèle de référence plus robuste, *XLM-RoB*$_{nous}$, sauf pour la langue anglaise.

La tâche NER en espagnol a connu l'impact le plus important, avec une augmentation allant jusqu'à $+1{,}1$ points pour le score F1. Ces résultats impliquent que notre approche améliore de manière effective les représentations de modèle en utilisant des *définitions des types d'entités* dans les classes d'ensemble de données à granularité grossière et fine. Cependant, même si notre approche a produit de meilleurs scores sur l'ensemble de données *MultiCoNER II* avec la taxonomie à granularité fine (une amélioration globale de $+0{,}38$), ils étaient inférieurs à ceux obtenus sur *MultiCoNER I* (une amélioration globale de $+0{,}45$). Cette différence suggère que la difficulté liée à la taxonomie à granularité fine de la nouvelle version de l'ensemble de données pose un défi et l'ajout de définitions textuelles d'hyperonymie n'aide pas à capturer des représentations hiérarchiques riches de façon optimale.

En outre, notre modèle est nettement moins performant que le modèle *DAMO-NLP*. Une différence

	MultiCoNER I		**MultiCoNER II**			
Modèle	**EN**	**ES**	**EN**	**ES**	**FR**	**PT**
XLM-RoB$_{nous}$	65,9	62,6	62,3	66,1	64,7	65,6
Notre modèle	65,7(-0,2)	63,7(+1,1)	62,2(-0,1)	67,1(+1,0)	65,0(+0,3)	65,9(+0,3)
XLM-Rob	61,2	57,4	-	-	-	-
DAMO-NLP	91,2	89.9	-	-	-	-

TABLE 2 – Scores F1 macro-moyens obtenus pour les jeux des données MultiCoNER I et II dans les quatre langues EN, ES, FR et PT.

notable entre les approches de *DAMO-NLP* et la nôtre réside dans la modification de l'entrée lors de l'inférence. En effet, *DAMO-NLP* étend la phrase initiale, ce qui augmente considérablement le temps d'inférence pour ce modèle (10 *tokens* en moyenne à encoder dans l'entrée originale versus 218 *tokens* avec l'entrée étendue (Wang *et al.*, 2022)). L'objectif de notre modèle est d'explorer l'impact des *définitions des types d'entités*, en utilisant seulement l'entrée originelle, ce qui favorise également une inférence plus performante en termes de temps d'exécution.

Classe	**EN**	**ES**	**FR**	**PT**
OtherLOC	+3,6	+1,7	+8,6	+3,6
HumanSettl	+2,35	+2,2	+0,7	+2,3
Station	+7,0	+10,6	+8,8	+7,0
MusicalWork	+3,1	+0,5	+3,9	+3,1
WrittenWork	+0,1	+5,3	+1,5	+0,1
OtherPER	+3,9	+2,9	+3,6	+3,9
Symptom	+28,6	+1,7	+5,8	+28,6

TABLE 3 – Classes fines améliorées par notre modèle par rapport à notre modèle de base *XLM-Rob*$_{nous}$.

Dans le Tableau 3, nous montrons les classes fines de *MultiCoNER II* ayant présenté une amélioration constante. Nos résultats révèlent une amélioration substantielle allant jusqu'à $+28{,}6$ points pour le score F1 macro dans différentes langues, indiquant que notre modèle peut incorporer des informations plus pertinentes pour certains types en particulier. Néanmoins, une limitation significative de notre approche est que nous avons obtenu des scores nuls pour les types qui n'étaient pas présents dans les annotations de l'ensemble d'entraînement, ce qui montre une capacité de généralisation limitée.

Compte tenu de ces analyses, nos résultats suggèrent globalement que l'incorporation d'informations contextuelles pertinentes à propos des types d'entités dans les modèles améliore leur performance. Cependant, un travail plus conséquent est nécessaire pour dépasser les meilleurs modèles, tels que *DAMO-NLP*, qui utilisent des ressources externes pour exploiter les représentations du modèle.

6 Conclusion

Dans cet article, nous avons présenté notre approche visant à améliorer les performances de la tâche NER dans des cas de figures présentant une complexité liée au contexte limité ou multiplicité et variabilité des types d'entités. Nous avons évalué l'efficacité de l'injection d'informations contextuelles sur les définitions des types d'entités afin d'améliorer les représentations d'un PLM. Nos évaluations ont montré des améliorations par rapport aux modèles de référence, mais ont également montré ses limites, particulièrement pour des types non vus à l'entraînement. Dans les travaux futurs, nous évaluerons comment extraire de meilleures représentations de la taxonomie hiérarchique à granularité fine et adapterons cette version améliorée de XLM-RoBERTa comme modèle de base pour d'autres approches, par exemple, comme complément à d'autres modèles exploitant les représentations des phrases en entrée.

Remerciements

Ce travail a été financé par le projet ANR-19-CE23-0028 MEERQAT. Il a en outre bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011012638R1 attribuée par GENCI.

Références

ASHWINI S. & CHOI J. D. (2014). Targetable named entity recognition in social media. *arXiv preprint arXiv :1408.0782*.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

FETAHU B., CHEN Z., KAR S., ROKHLENKO O. & MALMASI S. (2023). MultiCoNER v2 : a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

FETAHU B., FANG A., ROKHLENKO O. & MALMASI S. (2022). Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2777–2790, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.200.

GAN W., LIN Y., YU G., CHEN G. & YE Q. (2022). Qtrade AI at SemEval-2022 task 11 : An unified framework for multilingual NER task. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, p. 1654–1664, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.semeval-1.228.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

JAYARAO P., JAIN C. & SRIVASTAVA A. (2018). Exploring the importance of context and embeddings in neural ner models for task-oriented dialogue systems. *arXiv preprint arXiv :1812.02370*.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

LING X. & WELD D. (2021). Fine-grained entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **26**(1), 94–100. DOI : 10.1609/aaai.v26i1.8122.

MALMASI S., FANG A., FETAHU B., KAR S. & ROKHLENKO O. (2022a). MultiCoNER : a Large-scale Multilingual dataset for Complex Named Entity Recognition.

MALMASI S., FANG A., FETAHU B., KAR S. & ROKHLENKO O. (2022b). SemEval-2022 Task 11 : Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* : Association for Computational Linguistics.

MALMASI S., FANG A., FETAHU B., KAR S. & ROKHLENKO O. (2022c). SemEval-2022 task 11 : Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, p. 1412–1437, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.semeval-1.196.

MENG T., FANG A., ROKHLENKO O. & MALMASI S. (2021). GEMNET : Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1499–1512.

PETERS M. E., NEUMANN M., LOGAN R., SCHWARTZ R., JOSHI V., SINGH S. & SMITH N. A. (2019). Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 43–54, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1005.

PETRONI F., ROCKTASCHEL T., MILLER A. H., LEWIS P., BAKHTIN A., WU Y. & RIEDEL S. (2019). Language models as knowledge bases ? In *In : Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

TALMOR A., TAFJORD O., CLARK P., GOLDBERG Y. & BERANT J. (2020). Leap-of-thought : Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, **33**, 20227–20237.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.

WANG X., SHEN Y., CAI J., WANG T., WANG X., XIE P., HUANG F., LU W., ZHUANG Y., TU K., LU W. & JIANG Y. (2022). DAMO-NLP at SemEval-2022 task 11 : A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, p. 1457–1468, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.semeval-1.200.

ZHANG Z., HAN X., LIU Z., JIANG X., SUN M. & LIU Q. (2019). ERNIE : Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1441–1451, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1139.