Étude de méthodes d'augmentation de données pour la reconnaissance d'entités nommées en astrophysique

Atilla Kaan Alkan, Cyril Grouin, Pierre Zweigenbaum

▶ **To cite this version:**

Atilla Kaan Alkan, Cyril Grouin, Pierre Zweigenbaum. Étude de méthodes d'augmentation de données pour la reconnaissance d'entités nommées en astrophysique. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2023, Paris, France. pp.1-13. hal-04130190

HAL Id: hal-04130190
https://hal.science/hal-04130190

Submitted on 20 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

Étude de méthodes d'augmentation de données pour la reconnaissance d'entités nommées en astrophysique

Atilla Kaan Alkan[1, 2] Cyril Grouin[1] Pierre Zweigenbaum[1]
(1) Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France.
(2) IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
{atilla.alkan, cyril.grouin, pz}@lisn.upsaclay.fr

RÉSUMÉ

Dans cet article nous étudions l'intérêt de l'augmentation de données pour le repérage d'entités nommées en domaine de spécialité : l'astrophysique. Pour cela, nous comparons trois méthodes d'augmentation en utilisant deux récents corpus annotés du domaine : DEAL et TDAC, tous deux en anglais. Nous avons générés les données artificielles en utilisant des méthodes à base de règles et à base de modèles de langue. Les données ont ensuite été ajoutées de manière itérative pour affiner un système de détection d'entités. Les résultats permettent de constater un effet de seuil : ajouter des données artificielles au-delà d'une certaine quantité ne présente plus d'intérêt et peut dégrader la F-mesure. Sur les deux corpus, le seuil varie selon la méthode employée, et en fonction du modèle de langue utilisé. Cette étude met également en évidence que l'augmentation de données est plus efficace sur de petits corpus, ce qui est cohérent avec d'autres études antérieures. En effet, nos expériences montrent qu'il est possible d'améliorer de 1 point la F-mesure sur le corpus DEAL, et jusqu'à 2 points sur le corpus TDAC.

ABSTRACT

Investigating Data Augmentation Methods for Astrophysical Named Entity Recognition.

In this paper, we investigate the effectiveness of data augmentation for named entity recognition in astrophysics. To this end, we compare three augmentation methods using two recent annotated corpora in the domain : DEAL and TDAC, both in English. We generated artificial data using rule-based and language model-based approaches. The data was then iteratively added to finetune an entity detection system. The results show a threshold effect : adding artificial data beyond a specific quantity is no longer beneficial and can decrease F-measure. The threshold varies for each method and depends on the language model employed. This study also highlights that data augmentation is more effective for small corpora, consistent with previous studies. Indeed, our experiments demonstrate the potential to improve the F-measure by 1 point in the DEAL corpus and up to 2 points in the TDAC corpus.

MOTS-CLÉS : Repérage d'entités nommées, Augmentation de données, Annotation, Astrophysique.

KEYWORDS: Named Entity Recognition, Data Augmentation, Annotation, Astrophysics.

1 Introduction

Tâche de base en Traitement Automatique des Langues (TAL), la reconnaissance d'entités nommées (REN) consiste à repérer des mentions d'entités dans un texte afin de les catégoriser dans des classes

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 1–13.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

pré-définies. Depuis son introduction en 1996 (Grishman & Sundheim, 1996), cette tâche s'est avérée utile pour la recherche d'information (Banerjee *et al.*, 2019) ou encore pour la constitution de systèmes de questions-réponses (Mollá Aliod *et al.*, 2006). Cependant, l'entraînement des systèmes de repérage d'entités nommées par apprentissage requièrent une quantité significative de données annotées, qui ne sont pas toujours disponibles pour certaines langues, peu dotées en ressources linguistiques, ou encore en domaine de spécialité.

Dans cet article, nous étudions l'intérêt de l'augmentation de données pour le repérage d'entités nommées, dans un domaine de spécialité peu étudié en TAL et pour lequel il existe peu de ressources annotées : l'astrophysique. Pour cela, nous proposons de comparer trois approches d'augmentation de données en utilisant les deux seuls corpus du domaine disponibles, tous deux en anglais : DEAL (Grezes *et al.*, 2022) et TDAC (Alkan *et al.*, 2022).

Notre étude a mis en evidénce la présence d'un effet de seuil, au-delà duquel l'ajout de données artificielles peut dégrader les performances des systèmes de détection d'entités nommées. L'évaluation des approches d'augmentation sur les deux corpus du domaine astrophysique montrent que les performances varient en fonction de la méthode employée ainsi que du modèle de langue utilisé pour générer les données artificielles. Nous décrivons notre méthode dans la Section 3. Nous présentons nos résultats expérimentaux dans la Section 4, et les analysons en Section 5.1.

2 Recherches connexes

2.1 Méthodes existantes

Explorée principalement en vision par ordinateur (Shorten & Khoshgoftaar, 2019), l'augmentation de données consiste à générer de nouvelles données artificielles pour l'entraînement des modèles, sans avoir à en recueillir davantage de façon manuelle. Ces dernières années, cette technique fait l'objet d'une attention croissante par les chercheurs en TAL, notamment pour la traduction (Wang *et al.*, 2018), les systèmes de questions-réponses (Yang *et al.*, 2019), la classification (Claveau *et al.*, 2021) et le résumé automatique de texte (Pasunuru *et al.*, 2021).

Les méthodes d'augmentation peuvent se diviser en plusieurs familles (Feng *et al.*, 2021). Une première famille regroupant les méthodes à base de règles, utilisant des ressources linguistiques comme WordNet (Miller, 1994) pour remplacer des mots d'une phrase, ou encore la permutation de mots (Wei & Zou, 2019). La deuxième famille repose sur les modèles de langue, utilisant des techniques de rétrotraduction par exemple (Yu *et al.*, 2018). Le remplacement des mots d'une phrase peut également s'effectuer en utilisant un modèle de langue masqué (MLM). L'entraînement d'un modèle de langue masqué constitue la phase de pré-entraînement des modèles de langue contextualisés de type BERT (Devlin *et al.*, 2019). Cette tâche consiste à masquer aléatoirement un mot dans le texte d'entrée avec le token `[MASK]`, puis à prédire le mot masqué. L'insertion du mot prédit dans le texte permet alors de constituer une séquence artificielle à partir de la séquence d'origine. Une troisième famille d'augmentation existe, reposant cette fois sur des techniques d'interpolation (Zhang *et al.*, 2018, 2020). Ces méthodes consistent à générer des données artificielles en interpolant entre des échantillons de données existants. Cela peut se faire en prenant par exemple deux séquences d'entrée (sélectionnées selon leur similarité ou leur pertinence), puis en générant un échantillon artificiel qui se situe entre eux.

2.2 Attention moindre pour la reconnaissance d'entités nommées

Si elle améliore les performances de certaines tâches de TAL telles que le traduction automatique (Nguyen *et al.*, 2020) et les systèmes de questions-réponses (Yu *et al.*, 2018), l'augmentation de données est plus difficile à mettre en œuvre pour la détection d'entités. De récentes études (Dai & Adel, 2020) soulignent que ces méthodes peuvent générer des données artificielles erronées lorsqu'elles sont appliquées sur des tâches au niveau du token. En effet, l'utilisation d'un MLM pour remplacer une entité peut conduire à des erreurs lors de l'affinage de systèmes de reconnaissance d'entités nommées si le type du mot prédit ne correspond pas au type du mot masqué. En raison de cette difficulté, l'augmentation de données pour la REN a fait l'objet d'une attention moindre comparée à d'autres tâches. Pour pallier ce problème, Zhou *et al.* (2022) ont proposé la méthode MELM : *Masked Entity Language Modeling*, qui vise à insérer autour de chaque token masqué le type qui lui correspond lors de la phase de pré-entraînement (par exemple, le token spécial `[B-Ins]` ou `[I-Ins]` autour des tokens d'une entité de la classe `Instrument`). Par conséquent, lors de l'apprentissage, la prédiction du token masqué est conditionnée à la fois par son contexte et par son type, réduisant le risque de non-correspondance entre le token prédit et la classe.

L'astrophysique est un domaine de spécialité générant une quantité significative de documents à analyser, mais possédant en revanche très peu de corpus annotés en entités nommées. L'un de ces corpus, publié très récemment, a servi dans la campagne d'évaluation internationale DEAL sur la reconnaissance d'entités nommées dans des articles d'astrophysique (Grezes *et al.*, 2022). Parmi les travaux qui se sont intéressés à la détection d'entités nommées en astrophysique (Becker *et al.*, 2005; Hachey *et al.*, 2005; Murphy *et al.*, 2006), une seule méthode d'augmentation de données a été proposée par l'un des participants de cette campagne d'évaluation (Huang, 2022). L'auteur utilise des modèles pré-entraînés spécifiques (He *et al.*, 2021; Berquand *et al.*, 2021) à base d'adapteurs (Houlsby *et al.*, 2019) pour l'augmentation de données, lui permettant d'atteindre une F-mesure de 0,7799 sur le jeu de test du corpus DEAL.

3 Protocole expérimental pour l'augmentation de données

3.1 Présentation générale des corpus

Nous utilisons les deux seuls corpus annotés et disponibles du domaine astrophysique : le corpus DEAL (Grezes *et al.*, 2022) et le corpus TDAC (Alkan *et al.*, 2022). Ces ensembles de données proviennent de sources différentes et sont en anglais, car la communauté astrophysique, composée d'amateurs et de professionnels utilise principalement l'anglais pour communiquer.

Le corpus DEAL Ce corpus est constitué de fragments d'articles scientifiques en astrophysique générale. Il a été annoté en entités nommées pour la campagne d'évaluation DEAL (*Detecting Entities in the Astrophysics Literature*) et se compose de trois ensembles : entraînement, développement et test, comprenant respectivement 1753, 1366 et 2505 documents. Le corpus est accessible sur HuggingFace [1].

1. https://huggingface.co/datasets/adsabs/WIESP2022-NER/tree/main

Le corpus TDAC Ce corpus se compose de rapports d'observation (courts messages textuels) qui constituent l'une des sources premières de partage d'informations entre astronomes. A la différence du corpus DEAL, ce corpus se focalise uniquement sur les phénomènes cosmiques dits « transitoires [2] » (Neronov, 2019), possédant ainsi un vocabulaire et un discours spécifiques que nous ne retrouvons pas nécessairement dans le corpus DEAL. Le corpus TDAC est accessible sur GitHub [3]. L'une des limites de ce corpus concerne son nombre limité de documents annotés disponibles : 75 rapports d'observations, dont 59 pour l'entraînement et 16 pour le test. Le tableau 1 fournit quelques statistiques concernant ces deux corpus.

Corpus	Nb classes	Nb tokens	tokens annotés	Long. moyenne (tokens)
DEAL	31	1 815 237	337 663	322
TDAC	28	26 133	4526	256

TABLE 1 – Statistiques des deux corpus d'étude.

Entités nommées et particularités des textes en astrophysique Bien que ces corpus soient de sources différentes, les catégories définies ainsi que les schémas d'annotation sont identiques. Le guide d'annotation comprend 31 entités nommées au total et couvre les entités d'intérêt du domaine : installations astronomiques, objets célestes, coordonnées, formules ou encore techniques d'observation. Une liste détaillée des classes est disponible sur le dépôt HuggingFace [4]. La figure 1 montre la proportion des entités nommées annotées dans chacun des deux corpus.

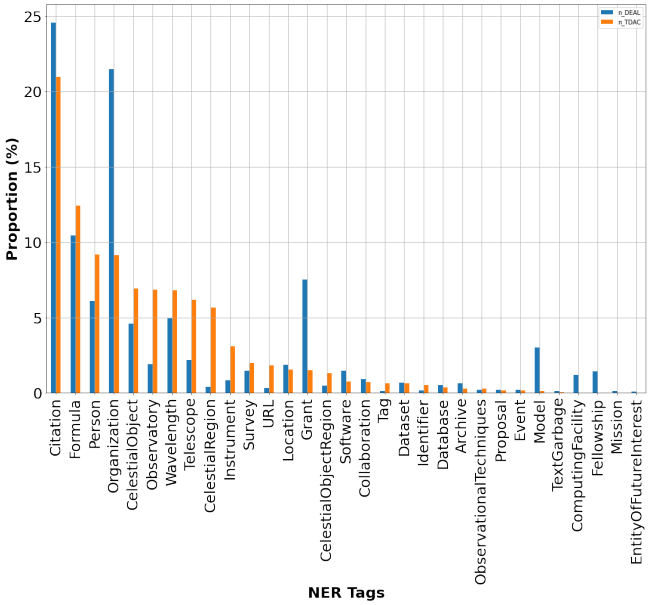


FIGURE 1 – Proportions des entités nommées dans le corpus DEAL (en bleu) et dans TDAC (en orange).

2. Les phénomènes transitoires sont de violentes explosions de courtes durées telles que les explosions de supernovae, les sursauts gamma, ou encore les jets de particules en provenance de certaines galaxies.

3. https://github.com/AtillaKaanAlkan/TDAC

4. https://huggingface.co/datasets/adsabs/WIESP2022-NER/blob/main/tag_definitions.md

La répartition des classes au sein des deux corpus n'est pas similaire. En effet, dans le corpus TDAC, les catégories les plus fréquentes sont par exemple `Formula`, `CelestialObject`, `Observatory` ou `CelestialRegion`. Il s'agit de catégories spécifiques au domaine astrophysique. La plupart de ces classes spécifiques sont moins présentes dans le corpus DEAL, dans lequel on retrouve principalement des classes du type : `Citation`, `Organization`, `Grant` ou `Person`, qui semblent être des catégories d'entités nommées plus génériques dans les articles scientifiques.

3.2 Description des méthodes d'augmentation

Nous masquons aléatoirement 70 % des entités nommées d'une séquence sans changer les tokens de type « O ». La valeur de 70 % a été suggérée par Zhou *et al.* (2022) pour la méthode MELM à l'issue d'une recherche par grille (*grid search*). Nous avons en plus fait le choix que si un token faisant partie d'une portion annotée est masqué, alors l'intégralité de cette portion est remplacée. La figure 2 schématise nos trois méthodes.

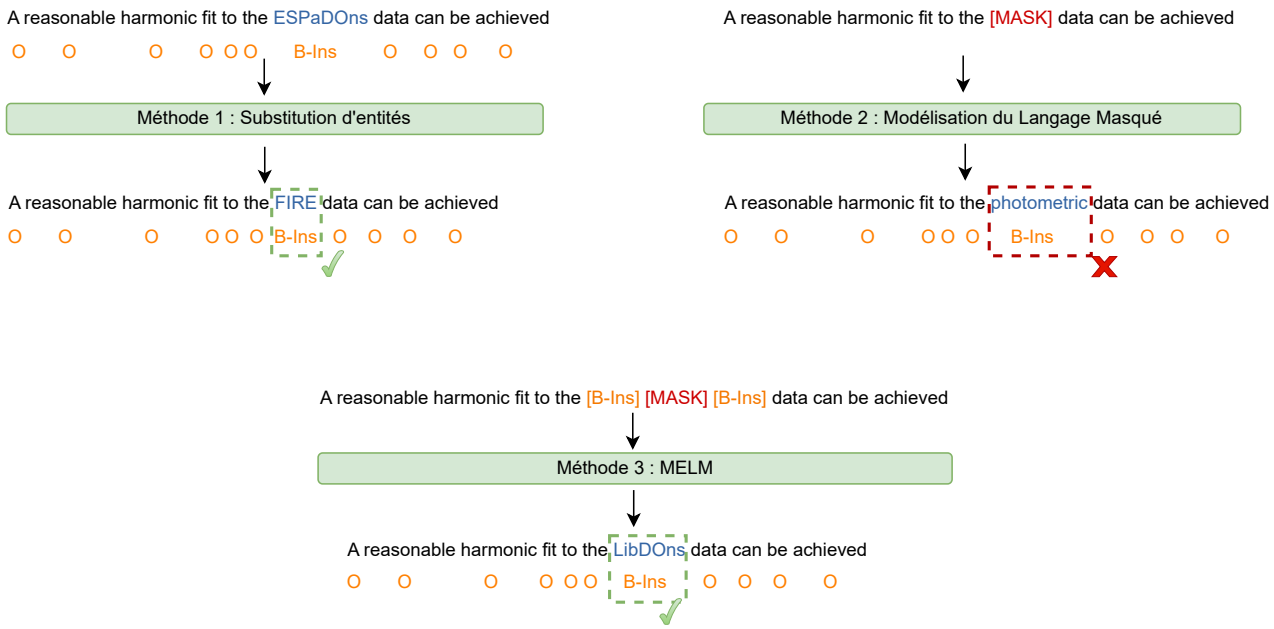


FIGURE 2 – Illustration des trois méthodes d'augmentation proposées.

Méthode 1 : Substitution d'entités à base de règles La première méthode que nous explorons consiste à remplacer aléatoirement une entité d'une classe donnée par une autre entité de cette même classe du corpus d'entraînement. Par exemple, nous remplaçons un nom d'instrument de mesure comme « ESPaDOnS » de la classe `Instrument` par « NIRCam » de cette même classe, créant ainsi une donnée augmentée. Il s'agit donc d'une méthode simple d'augmentation à base de règles.

Méthode 2 : Modèle de langue masqué (MLM) La deuxième méthode que nous proposons consiste à utiliser des modèles de langue masqué. Afin d'évaluer l'impact de l'augmentation de données, nous utilisons trois modèles : astroBERT (Grezes *et al.*, 2021), SciBERT (Beltagy *et al.*, 2019) et RoBERTa (Zhuang *et al.*, 2021). Ces modèles se distinguent par leur corpus de pré-entraînement.

Alors qu'astroBERT est conçu pour l'analyse de textes en astrophysique, SciBERT repose sur des textes scientifiques, tandis que RoBERTa est un modèle pré-entraîné sur des corpus de langue générale. Nous masquons aléatoirement des entités du corpus et utilisons la prédiction du mot masqué pour produire nos données artificielles.

Méthode 3 : Modèle de langue d'entités masquées (MELM) Comme évoqué dans les travaux connexes, l'usage d'un modèle de langue masqué peut générer des erreurs de type. Dans cette troisième méthode, nous proposons d'appliquer la méthode MELM pour la prédiction de nouveaux mots, et constituer ainsi nos données artificielles. Comme illustré sur la figure 2, nous ajoutons autour des tokens masqués les étiquettes correspondantes. Le modèle astroBERT étant spécifique au domaine de spécialité et la méthode MELM s'appuyant sur les classes d'entités relatives au domaine astrophysique, nous avons étudié l'impact de cette méthode au moyen d'astroBERT principalement. Nous pré-entraînons astroBERT sur la tâche de prédiction de mot masqué avec les tokens spéciaux ajoutés en utilisant les échantillons d'entraînement respectifs des deux corpus : 1753 documents pour DEAL, et 59 rapports d'observation pour TDAC pendant 20 époques, comme conseillé par Zhou *et al.* (2022).

4 Expériences pour la reconnaissance d'entités nommées

4.1 Configurations

Modèles et hyperparamètres Nous avons affiné astroBERT (Grezes *et al.*, 2021) sur une tâche de REN sur 15 époques, avec un taux d'apprentissage $\alpha = 2 \times 10^{-5}$ et une taille de lot d'entraînement (*training batch size)* de 4. Les expériences ont été réitérées cinq fois avec des valeurs d'amorces différentes ($seeds = [0, 123, 762, 5000, 6822]$) choisies aléatoirement.

Variation du taux d'augmentation Amalvy *et al.* (2022) ont montré qu'au delà d'un certain seuil, l'augmentation de données ne présentait plus d'intérêt et engendrait une baisse des performances, principalement due à un sur-apprentissage. C'est pourquoi nous avons analysé l'impact de l'augmentation de données sur les performances des systèmes de détection d'entités nommées en faisant varier la quantité d'exemples artificiels ajoutés au corpus d'entraînement d'origine. Plus précisément, nous avons examiné des augmentations par incrément de 25 %.

4.2 Résultats

Afin d'évaluer nos systèmes de détection d'entités nommées, nous calculons la précision (P), le rappel (R) et la F-mesure (F) en suivant la méthode CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003). Nous présentons les résultats obtenus sur les jeux de test des corpus DEAL et TDAC par un modèle de détection d'entités astroBERT entraîné sur des données étendues avec chacune des trois méthodes d'augmentation de données que nous avons présentées : substitution d'entités (tableau 2), modèle de langue masqué (tableau 3) et MELM (tableau 4). Pour évaluer l'intérêt de l'augmentation de données, nous rappelons sur la première ligne de chaque tableau les performances du système entraîné uniquement sur les données d'origine (sans augmentation).

Taux	DEAL			TDAC		
	P	R	F	P	R	F
∅	0,799	**0,834**	**0,816** (0,002)	0,666	0,737	0,700 (0,003)
25 %	**0,805**	0,827	0,816 (0,003)	0,641	0,726	0,681 (0,018)
50 %	0,800	0,823	0,811 (0,003)	0,663	0,728	0,694 (0,014)
75 %	0,801	0,825	0,813 (0,001)	0,667	0,734	0,699 (0,009)
100 %	0,789	0,826	0,807 (0,003)	**0,677**	**0,749**	**0,711 (0,004)**

TABLE 2 – Impact de l'augmentation de données par la méthode de substitution d'entités, évaluée sur une tâche de repérage d'entités nommées, avec comparaison sans augmentation de données (∅). Pour chaque configuration, nous affichons la moyenne (et l'écart type) des 5 affinages réalisés. Pour des raisons de visibilité, nous indiquons l'écart type uniquement pour la F-mesure.

Modèle et taux d'augmentation		DEAL			TDAC		
		P	R	F	P	R	F
	∅	0,799	0,834	0,816 (0,002)	0,666	0,737	0,700 (0,003)
astroBERT	25 %	0,799	0,829	0,814 (0,005)	0,647	0,728	0,685 (0,007)
	50 %	0,795	0,824	0,809 (0,005)	0,664	0,729	0,695 (0,009)
	75 %	0,789	0,824	0,806 (0,004)	0,670	0,748	0,706 (0,016)
	100 %	0,788	0,820	0,804 (0,002)	0,680	0,749	0,713 (0,009)
SciBERT	25 %	0,796	0,831	0,813 (0,004)	0,631	0,718	0,671 (0,014)
	50 %	0,794	0,828	0,810 (0,002)	**0,693**	**0,749**	**0,720 (0,016)**
	75 %	0,788	0,821	0,804 (0,006)	0,677	0,744	0,709 (0,013)
	100 %	0,776	0,822	0,798 (0.003)	0,688	0,745	0,715 (0,008)
RoBERTa	25 %	0,799	0,836	0,817 (0,001)	0,624	0,710	0,664 (0,011)
	50 %	0,794	**0,839**	0,816 (0,003)	0,635	0,711	0,671 (0,014)
	75 %	**0,806**	0,835	**0,820 (0,003)**	0,674	0,728	0,700 (0,014)
	100 %	0,798	0,831	0,814 (0,003)	0,639	0,707	0,671 (0,008)

TABLE 3 – Impact de l'augmentation de données par la méthode MLM, en fonction du modèle de langue masqué utilisé (astroBERT, SciBERT, RoBERTa) pour générer les données artificielles et du taux d'augmentation (de 25,0 % à 100 % par incrément de 25,0 %), évaluée sur une tâche de repérage d'entités nommées, avec comparaison sans augmentation de données (∅). Pour chaque configuration, nous affichons la moyenne (et l'écart type) des 5 affinages réalisés. Pour des raisons de visibilité, nous indiquons l'écart type uniquement pour la F-mesure.

5 Analyse et discussion

5.1 Impact de la méthode d'augmentation et du modèle de langue

Les trois méthodes d'augmentation proposées fournissent des résultats assez proches. Néanmoins, dans le cadre de nos expériences, nous constatons que les approches basées sur l'utilisation de modèles de langue (méthodes 2 et 3) permettent d'obtenir de meilleurs scores qu'une approche de substitution à base de règles. Nous gagnons jusqu'à deux points de F-mesure avec les méthodes 2 et 3 (tableaux 3 et 4) sur le corpus TDAC, et 1 point sur le corpus DEAL, tandis que la méthode à base de règles (tableau 2) nous permet de gagner 1 point de F-mesure sur le corpus TDAC, et aucune amélioration

Modèle et taux d'augmentation		DEAL			TDAC		
		P	**R**	**F**	**P**	**R**	**F**
	$\emptyset$	0,799	0,834	0,816 (0, 002)	0,666	0,737	0,700 (0, 003)
astroBERT	25 %	0,798	0,835	0,816 (0, 005)	0,664	0,732	0,696 (0, 008)
	50 %	0,796	0,834	0,815 (0, 003)	0,667	0,738	0,701 (0, 004)
	75 %	0,797	0,835	0,816 (0, 004)	0,668	0,738	0,701 (0, 004)
	100 %	0,797	0,836	0,816 (0, 004)	0,669	0,738	0,702 (0, 005)

TABLE 4 – Impact de l'augmentation de données par la méthode MELM, en fonction du taux d'augmentation (de 25,0 % à 100 % par incrément de 25,0 %),évaluée sur une tâche de repérage d'entités nommées, avec comparaison sans augmentation de données ($\emptyset$). Pour chaque configuration, nous affichons la moyenne (et l'écart type) des 5 affinages réalisés. Pour des raisons de visibilité, nous indiquons l'écart type uniquement pour la F-mesure.

n'est constatée sur le corpus DEAL.

Cette observation peut s'expliquer par la diversité générée grâce aux modèles de langue. En effet, ces derniers permettent d'offrir une plus grande diversité dans la génération de nouvelles entités, contrairement à la méthode 1 qui, reste limitée aux entités présentes dans le corpus d'entraînement lors de la génération de données.

Toujours en lien avec la diversité, nous constatons dans le tableau 3 que l'utilisation d'un modèle de langue générale (RoBERTa) sur DEAL et d'un modèle entraîné sur des textes scientifiques (SciBERT) sur TDAC pour générer des données artificielles, permettent d'obtenir de meilleurs résultats que l'utilisation d'astroBERT. Il semble donc y avoir un impact du corpus d'entrainement du modèle de langue sur l'augmentation de données : un modèle de langue pré-entraîné sur un corpus plus général, permet de générer une plus grande diversité. Nous envisageons donc de poursuivre la comparaison de la méthode MELM avec les deux autres modèles RoBERTa et SciBERT.

De plus, nous estimons que la capacité de prise en compte du contexte des modèles de langue lors de la génération de données peut également être une des raisons conduisant à de meilleurs résultats.

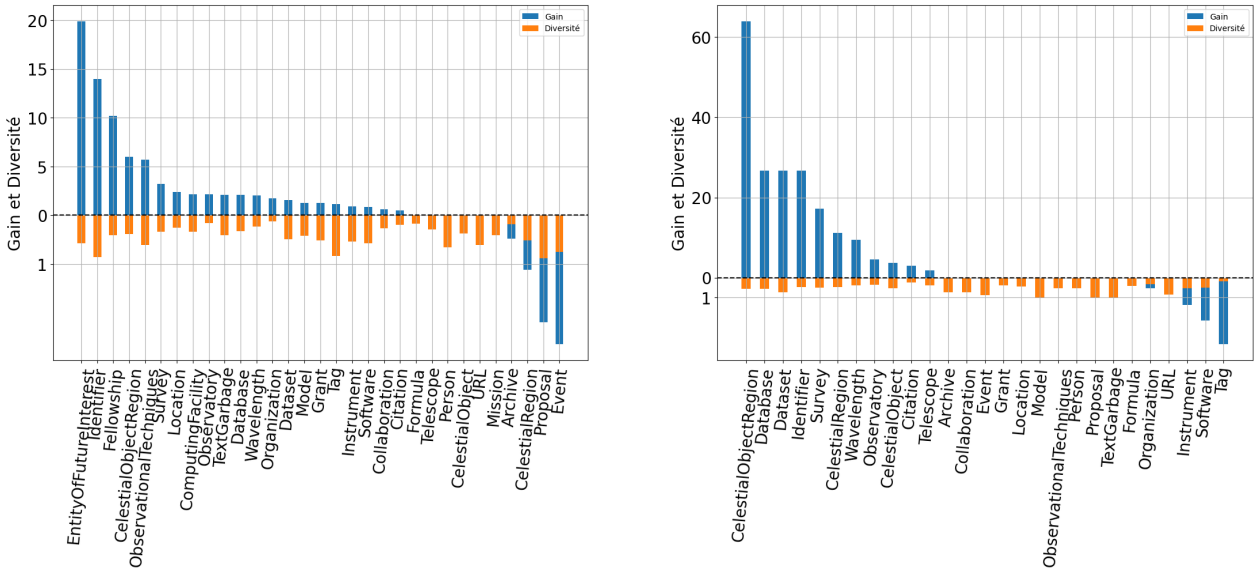
5.2 Effet de seuil

L'augmentation par incrément nous permet d'observer un effet de seuil, qui est également constaté dans les travaux de Amalvy *et al.* (2022). En effet, peu importe la méthode d'augmentation proposée, l'ajout de données artificielles au delà d'une certaine quantité ne présente plus d'intérêt et dégrade la F-mesure. Dans nos expériences, ce seuil varie selon la méthode employée et en fonction du modèle de langue utilisé. Il semblerait que plus le modèle de langue est général, plus le seuil est élevé. Ceci peut s'expliquer par le fait que les données générées avec astroBERT sont très proches des données d'origines et moins diversifiées que celles produites par RoBERTa, conduisant à un sur-apprentissage.

Par ailleurs, au vu des résultats obtenus, l'augmentation de données semble plus efficace lorsque appliquée sur de petits corpus, ce qui rejoint également la conclusion de Dai & Adel (2020). En effet, il est possible de gagner jusqu'à plus de 2 points de F-mesure sur le corpus TDAC (tableau 3), alors que sur le corpus DEAL, plus grand, nous gagnons seulement 1 point (tableaux 3 et 4) toutes méthodes confondues.

5.3 Répartition des gains et diversité des classes

En considérant les méthodes ayant donné les meilleurs scores globaux en terme de F-mesure, nous avons analysé pour chacune des classes (types d'entités) s'il existait un lien entre le gain obtenu (diminution ou augmentation de la F-mesure) et la diversité de la classe (nombre de formes de surface différentes). Le calcul du coefficient de corrélation de Pearson [5] (c_p) montre qu'il y a une faible corrélation négative entre le gain et la diversité. En effet, les coefficients sont relativement faibles : $c_p = -0{,}062$ pour le corpus DEAL, et $c_p = 0{,}023$ pour le corpus TDAC. La figure 3 montre que le gain est réparti sur différentes classes.



(a) Corpus DEAL

(b) Corpus TDAC

FIGURE 3 – Répartition des gains (en bleu) et diversité (en orange) sur les deux corpus.

Dans l'ensemble, la majeure partie des classes bénéficient d'une augmentation de la F-mesure avec une amélioration significative sur les classes du domaine astrophysique. En effet, l'ajout de 25 % de données artificielles au corpus DEAL avec la méthode MELM permet au système de REN de mieux repérer certains concepts dans le texte tels que les coordonnées d'objets célestes et les techniques utilisées lors des observations : 6 points pour la classe `CelestialObjectRegion`, et 5,7 pour la classe `ObservationalTechniques`. L'augmentation de données présente également un intérêt pour le repérage d'installations astronomiques (`Observatory` :2,13) et des longeurs d'ondes (`Wavelength` : 2,02). Nous constatons également des améliorations dans la détection des coordonnées sur le corpus TDAC (`CelestialObjectRegion` : 63,9) et des noms d'objets astrophysiques (`CelestialObject` :3,71).

6 Conclusion et perspectives

Dans cet article, nous avons étudié l'intérêt de l'augmentation de données pour entraîner des systèmes de repérage d'entités nommées. Les méthodes proposées ont été appliquées à un domaine de spécialité

5. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

peu étudié en TAL : l'astrophysique, en utilisant deux corpus existants. Nous avons exploré l'efficacité de trois méthodes d'augmentation permettant de limiter les erreurs token-étiquette lors de la génération de données.

Notre étude a permis d'identifier les classes qui bénéficient le plus de l'augmentation et de mettre en évidence que les performances varient selon la méthode d'augmentation employée, en fonction du modèle de langue utilisé.

Ces résultats ouvrent la voie à plusieurs perspectives de recherche. Tout d'abord, il pourrait être bénéfique de cibler spécifiquement les classes à augmenter plutôt que de procéder à une augmentation aléatoire. En effet, cibler les classes à augmenter sur le corpus DEAL pourrait éviter une diminution des performances sur deux classes importantes du domaine astrophysique : `CelestialRegion` ($-5{,}59$), et `CelestialObject` ($-0{,}29$).

En outre, il serait intéressant de trouver une approche permettant de diversifier la structure des documents et de modifier le contexte local autour des entités. En effet, bien que les méthodes comparées dans cet article permettent d'augmenter le nombre d'entités du corpus et leur diversité (selon le modèle utilisé), elles ne changent pas la structure des documents et ne modifient pas le contexte, ce qui génère des données artificielles proches des données d'origine, pouvant conduire à un risque de sur-apprentissage qui expliquerait pourquoi les performances des systèmes peuvent se dégrader au-delà d'un certain seuil.

Une solution serait de remplacer les tokens de type « O » (les adjectifs et les adverbes par exemple) autour des entités annotées afin de pouvoir modifier le contexte localement. Enfin, il pourrait être utile de tester les méthodes à base d'interpolation (Zhang *et al.*, 2020) et de combiner plusieurs approches d'augmentation pour améliorer davantage les performances.

Remerciements

Nous remercions chaleureusement Fabian Schüssler (IRFU, CEA, Université Paris-Saclay), astrophysicien, pour son accompagnement, son expertise et ses nombreux conseils dans le cadre de ces travaux de recherches.

Références

ALKAN A. K., GROUIN C., SCHUSSLER F. & ZWEIGENBAUM P. (2022). TDAC, the first corpus in time-domain astrophysics : Analysis and first experiments on named entity recognition. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 131–139, Online : Association for Computational Linguistics.

AMALVY A., LABATUT V. & DUFOUR R. (2022). Remplacement de mentions pour l'adaptation d'un corpus de reconnaissance d'entités nommées à un domaine cible (Mention replacement for adapting a named entity recognition dataset to a target domain). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 198–205, Avignon, France : ATALA.

BANERJEE P. S., CHAKRABORTY B., TRIPATHI D., GUPTA H. & KUMAR S. S. (2019). A information retrieval based on question and answering and ner for unstructured information without using sql. *Wirel. Pers. Commun.*, **108**(3), 1909–1931. DOI : 10.1007/s11277-019-06501-z.

BECKER M., HACHEY B., ALEX B. & GROVER C. (2005). Optimising selective sampling for bootstrapping named entity recognition. In *In Proceedings of the ICML Workshop on Learning with Multiple Views*, p. 5–11.

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. 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. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

BERQUAND A., DARM P. & RICCARDI A. (2021). Spacetransformers : Language modeling for space systems. *IEEE Access*, **9**, 133111–133122. DOI : 10.1109/ACCESS.2021.3115659.

CLAVEAU V., CHAFFIN A. & KIJAK E. (2021). La génération de textes artificiels en substitution ou en complément de données d'apprentissage (Generating artificial texts as substitution or complement of training data). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 37–49, Lille, France : ATALA.

DAI X. & ADEL H. (2020). An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 3861–3867, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.343.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

FENG S. Y., GANGAL V., WEI J., CHANDAR S., VOSOUGHI S., MITAMURA T. & HOVY E. H. (2021). A survey of data augmentation approaches for NLP. *CoRR*, **abs/2105.03075**.

GREZES F., BLANCO-CUARESMA S., ACCOMAZZI A., KURTZ M. J., SHAPURIAN G., HENNEKEN E., GRANT C. S., THOMPSON D. M., CHYLA R., MCDONALD S., HOSTETLER T. W., TEMPLETON M. R., LOCKHART K. E., MARTINOVIC N., CHEN S., TANNER C. & PROTOPAPAS P. (2021). Building astroBERT, a language model for astronomy & astrophysics. DOI : 10.48550/ARXIV.2112.00590.

GREZES F., BLANCO-CUARESMA S., ALLEN T. & GHOSAL T. (2022). Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 1–7, Online : Association for Computational Linguistics.

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

HACHEY B., ALEX B. & BECKER M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, p. 144–151, Ann Arbor, Michigan : Association for Computational Linguistics.

HE P., GAO J. & CHEN W. (2021). Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, **abs/2111.09543**.

HOULSBY N., GIURGIU A., JASTRZEBSKI S., MORRONE B., DE LAROUSSILHE Q., GESMUNDO A., ATTARIYAN M. & GELLY S. (2019). Parameter-efficient transfer learning for NLP. In K. CHAUDHURI & R. SALAKHUTDINOV, Éds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, p. 2790–2799 : PMLR.

HUANG P.-W. (2022). Domain specific augmentations as low cost teachers for large students. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 84–90, Online : Association for Computational Linguistics.

MILLER G. A. (1994). WordNet : A lexical database for English. In *Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

MOLLÁ ALIOD D., VAN ZAANEN M. & SMITH D. (2006). Named entity recognition for question answering. In L. CAVEDON & I. ZUKERMAN, Éds., *Proceedings of the Australasian Language Technology Workshop, ALTA 2006, Sydney, Australia, November 30-December 1, 2006*, p. 51–58 : Australasian Language Technology Association.

MURPHY T., MCINTOSH T. & CURRAN J. R. (2006). Named entity recognition for astronomy literature. In *Proceedings of the Australasian Language Technology Workshop 2006*, p. 59–66, Sydney, Australia.

NERONOV A. (2019). Introduction to multi-messenger astronomy. *Journal of Physics : Conference Series*, **1263**(1), 012001. DOI : 10.1088/1742-6596/1263/1/012001.

NGUYEN X., JOTY S. R., WU K. & AW A. T. (2020). Data diversification : A simple strategy for neural machine translation. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

PASUNURU R., CELIKYILMAZ A., GALLEY M., XIONG C., ZHANG Y., BANSAL M. & GAO J. (2021). Data augmentation for abstractive query-focused multi-document summarization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, p. 13666–13674 : AAAI Press.

SHORTEN C. & KHOSHGOFTAAR T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data*, **6**, 60. DOI : 10.1186/s40537-019-0197-0.

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., PHAM H., DAI Z. & NEUBIG G. (2018). SwitchOut : an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 856–861, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1100.

WEI J. & ZOU K. (2019). EDA : Easy data augmentation techniques for boosting performance on text classification tasks. 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. 6382–6388, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1670.

YANG W., XIE Y., TAN L., XIONG K., LI M. & LIN J. (2019). Data augmentation for BERT fine-tuning in open-domain question answering. *CoRR*, **abs/1904.06652**.

YU A. W., DOHAN D., LUONG M., ZHAO R., CHEN K., NOROUZI M. & LE Q. V. (2018). QANet : Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* : OpenReview.net.

ZHANG H., CISSÉ M., DAUPHIN Y. N. & LOPEZ-PAZ D. (2018). mixup : Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* : OpenReview.net.

ZHANG R., YU Y. & ZHANG C. (2020). SeqMix : Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 8566–8579, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.691.

ZHOU R., LI X., HE R., BING L., CAMBRIA E., SI L. & MIAO C. (2022). MELM : Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2251–2262, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.160.

ZHUANG L., WAYNE L., YA S. & JUN Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, p. 1218–1227, Huhhot, China : Chinese Information Processing Society of China.