Extraction d'entités nommées décrivant des chaînes de traitement bioinformatiques dans des articles scientifiques en anglais

Clémence Sebe, Sarah Cohen-Boulakia, Olivier Ferret, Aurélie Névéol

Extraction d'entités nommées décrivant des chaînes de traitement bioinformatiques dans des articles scientifiques en anglais

Clémence Sebe[1] Sarah Cohen-Boulakia[1] Olivier Ferret[2] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr

RÉSUMÉ

Les chaînes de traitement d'analyses de données biologiques utilisées en bioinformatique sont une solution pour la portabilité et la reproductibilité des analyses. Ces chaînes figurent à la fois sous forme descriptive dans des articles scientifiques et/ou sous forme de codes dans des dépôts. L'identification de publications scientifiques décrivant de nouvelles chaînes de traitement et l'extraction de leurs informations sont des enjeux importants pour la communauté bioinformatique. Nous proposons ici d'étendre le corpus *BioToFlow* ayant trait aux articles décrivant des chaînes de traitement bioinformatiques et de l'utiliser pour entraîner et évaluer des modèles de reconnaissance d'entités nommées bioinformatiques. Ce travail est accompagné d'une discussion critique portant à la fois sur le processus d'annotation du corpus et sur les résultats de l'extraction d'entités.

ABSTRACT

Extracting named entities describing bioinformatic workflows from the literature in English.

Workflows used in bioinformatic are a solution for analysis portability and reproducibility. These workflows are either described in publications and/or are in source code in repositories. Identifying new workflows in scientific articles and extracting related information is a challenge for the bioinformatics community. Herein, we propose to extend a corpus of articles describing bioinformatic workflows (*BioToFlow*) and to use it to train and evaluate bioinformatics named entity recognition models. We also engage in a critical discussion of both the corpus annotation process and the results of information extraction.

MOTS-CLÉS : Chaînes de traitement bioinformatiques, Annotation, Reconnaissance d'entités nommées.

KEYWORDS: Bioinformatic workflows, Annotation, Named entity recognition.

1 Introduction

La biologie est un domaine dans lequel l'arrivée de nouvelles technologies dites à haut-débit a permis l'acquisition de très grands volumes de données biologiques. Ces données brutes sont nombreuses mais aussi très hétérogènes et l'enjeu de la bioinformatique est de croiser, intégrer et analyser ces données pour faire avancer les connaissances en biologie. Face aux masses de données disponibles, la communauté bioinformatique a développé un grand nombre d'outils bioinformatiques. Une analyse de données bioinformatique consiste en l'enchaînement d'un ensemble d'outils bioinformatiques,

chaque outil consommant les données brutes en entrée et générant de nouvelles données, consommées à leur tour par l'outil suivant. Ces analyses peuvent être facilement implémentées via un script python ou un notebook pour des analyses rapides et impliquant de petits volumes de données. Mais lorsqu'il convient d'automatiser des analyses complexes pouvant impliquer l'utilisation de plusieurs dizaines d'outils et des volumes importants de données, il est nécessaire d'utiliser des solutions adaptées, plus faciles à déployer sur des grappes de calculs et offrant des solutions pour la portabilité et la reproductibilité de l'analyse. Les *systèmes de workflows scientifiques* ont été conçus pour répondre à ces besoins. Deux systèmes sont en particulier de plus en plus utilisés dans la communauté bioinformatique : Nextflow (Di Tommaso *et al.*, 2017) et Snakemake (Mölder *et al.*, 2021). Dans ces systèmes, une chaîne de traitement (ou *workflow*) est un code structuré où les étapes d'analyse sont bien distinguables. Lorsqu'une chaîne de traitement originale est conçue, le code est mis à disposition de la communauté dans un dépôt github et la description de la chaîne de traitement est publiée sous la forme d'un article scientifique dans une revue bioinformatique. Au 9 février 2024, le nombre d'articles ayant un lien github vers une chaîne de traitement dans le système de gestion Nextflow est ainsi de 89 articles et 91 pour le système Snakemake [1].

Un enjeu important pour la communauté bioinformatique est d'identifier les articles scientifiques décrivant une nouvelle chaîne de traitement et d'extraire les informations qui lui sont associées (par exemple, les outils bioinformatiques ou les données utilisées). À terme, la perspective de ce travail est, pour une même chaîne de traitement, de pouvoir comparer et fusionner des informations extraites des articles scientifiques d'une part et des codes issus de dépôts publics d'autre part.

Dans ce contexte, nous avons récemment proposé une première méthode de modélisation et d'extraction des composants des chaînes de traitement avec un schéma décrivant un ensemble d'entités nommées et de relations (Sebe *et al.*, 2023). Nous avons par ailleurs introduit le corpus *BioToFlow*, composé de 24 articles décrivant des chaînes de traitement (20 Nextflow et 4 Snakemake) et annotés par 3 annotateurs. Des expériences de reconnaissance d'entités ont été réalisées sur ce corpus et ont généré de premiers résultats prometteurs.

Cet article présente la suite de ce travail : nous nous focalisons sur l'extraction d'entités nommées à l'aide d'un ensemble plus important et plus varié d'articles, annotés par un ensemble plus important d'annotateurs. Plus précisément, nos contributions sont les suivantes :

— l'introduction et l'annotation de façon croisée par quatre annotatrices d'un corpus de 52 articles en anglais étendant *BioToFlow* et décrivant des chaînes de traitement bioinformatiques issues à la fois des systèmes Nextflow (26 articles) et Snakemake (26 articles) ;
— l'étude de l'utilisation de ce corpus pour l'entraînement et l'évaluation de modèles pour l'extraction automatique d'entités relatives aux chaînes de traitement ;
— une discussion critique des résultats de cette étude et des performances des modèles obtenus au niveau des différentes classes d'entités nommées.

1. Ce qui correspond à l'extraction d'articles de PubMed Central via la requête *(nextflow[Abstract] OR snakemake[Abstract] OR nextflow[Title] OR snakemake[Title]) AND github[All Fields]*

2 Extension du corpus annoté *BioToFlow*

2.1 Corpus *BioToFlow*

BioToFlow [2] est un corpus contenant 24 articles scientifiques (issus de revues telles Bioinformatics ou F1000Research) annotés manuellement à l'aide de 16 entités modélisant la composition des chaînes de traitement. C'est un corpus de petite taille avec une répartition déséquilibrée des articles relatifs aux chaînes de traitement issues des différents systèmes de gestion (4 Snakemake et 20 Nextflow).

Schéma d'annotation. L'ensemble des entités considérées dans *BioToFlow* est décrit à la figure 1. Les entités sont relatives à la chaîne de traitement elle-même (partie gauche) et à ses composants (partie droite). Chaque chaîne de traitement est désignée par un nom (*WorkflowName*). L'analyse qu'elle effectue fait référence à des méthodes algorithmiques (*Method*) implémentées par des outils bioinformatiques (*Tool*). Une chaîne de traitement met en jeu des données (*Data*) attendues dans un format de fichier particulier (*file*) et/ou issues d'une base de données (*Database*). Sur un plan technique, une chaîne de traitement est implémentée dans un langage de programmation (*ProgrammingLanguage*) et peut faire appel, pour s'exécuter, à un environnement d'exécution (*Environnement*), un système de conteneur (*Container*), un système de gestion de workflow (*ManagementSystem*) et nécessiter l'utilisation de bibliothèques (*LibraryPackage*) ou d'infrastructures particulières (*Hardware*). Chaque composant de la chaîne de traitement peut avoir un numéro de version (*Version*), une description (*Description*), des paramètres de lancement (*Parameter*) et des informations bibliographiques (*Biblio*).

Les annotations issues de ce schéma réalisées sur les articles du corpus *BioToFlow* sont considérées dans ce qui suit comme le *gold standard* pour les processus de reconnaissance d'entités nommées.

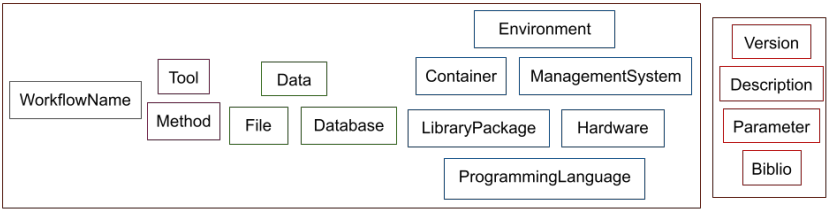

FIGURE 1 – Entités caractéristiques d'un workflow bioinformatique. À gauche : entités constitutives d'un workflow. À droite : entités liées à ses composants.

2.2 Nouveaux articles annotés

Choix des nouveaux articles. Nous avons suivi la méthodologie d'extraction des articles issus de PubMed Central de Sebe *et al.* (2023) pour augmenter la taille du corpus avec l'ajout de 28 nouveaux articles rééquilibrant le nombre d'articles décrivant des chaînes de traitement Nextflow et Snakemake :

2. https://doi.org/10.5281/zenodo.10650467

nous avons choisi de manière aléatoire parmi l'ensemble des publications disponibles 22 nouveaux articles décrivant des chaînes de traitement sous Snakemake et 6 sous Nextflow. La nouvelle version du corpus contient donc un total de 52 articles, avec 26 articles décrivant des chaînes de traitement de chaque système (Nextflow et Snakemake).

Démarche d'annotation des 28 nouveaux articles. Les 28 nouveaux articles ont été annotés en termes d'entités par quatre annotatrices (bio)informaticiennes selon la méthodologie décrite par Fort (2016). Le logiciel BRAT (Stenetorp *et al.*, 2012) et l'outil BRAT-Eval (Verspoor *et al.*, 2013) ont été respectivement utilisés pour l'annotation de ces articles et l'évaluation de la qualité de l'annotation (accord inter-annotatrices). La tâche d'annotation s'est effectuée en deux phases.

Phase 1 d'annotation : formation des annotatrices. L'objectif de la première phase était de former les annotatrices. Cinq articles du corpus initial *BioToFlow* ont été utilisés pour ce faire en enlevant les annotations de référence mais en utilisant comme préannotation les résultats d'un modèle entraîné sur le reste de ce corpus. Les quatre annotatrices ont annoté ces cinq articles à l'aide du guide d'annotation de Sebe *et al.* (2023). La qualité de leur annotation a été évaluée en comparant les annotations obtenues et celles du gold standard (score F1) issu de la version initiale de *BioToFlow*. La durée d'annotation des cinq articles varie entre deux et trois heures selon les annotatrices. Cette première phase a été concluante, avec un score F1 de 73 % à 83 % en mode strict (deux portions de texte doivent être strictement identiques et avoir la même étiquette) et 84 % à 88 % en mode relâché (où le recouvrement entre deux entités avec la même étiquette est accepté).

Phase 2 d'annotation : annotation des nouveaux articles. La seconde phase d'annotation a porté sur les 28 nouveaux articles à annoter. Des exemples de phrases annotées sont présentés dans l'Annexe A. Chaque annotatrice a annoté entre 8 et 17 articles. Il est à noter que ces articles sont plus longs et de nature plus variée que ceux du corpus initial *BioToFlow*. Alors que la version initiale de *BioToFlow* avec 24 articles contient 29 577 tokens et 10 125 tokens annotés (34 %), le nouvel ensemble de 28 articles regroupe 48 842 tokens, dont 17 661 sont annotés (36 %).

Statistiques. La version étendue de *BioToFlow* proposée est composée de 52 articles pour un ensemble de 78 419 tokens (dont 27 786 tokens annotés). Les entités sont réparties selon le tableau 1.

Entités	Occurrences	Entités	Occurrences
Data	2 434	Version	454
Tool	1 482	Hardware	429
Description	1 300	Database	288
Biblio	1 251	ManagementSystem	243
Method	936	Container	108
WorkflowName	851	ProgrammingLanguage	104
File	780	LibraryPackage	101
Parameter	464	Environment	83

TABLE 1 – Nombre d'entités par catégorie dans la version étendue de *BioToFlow*.

Les entités sont distribuées de façon très variable avec un nombre important d'occurrences pour les entités générales de la chaîne de traitement qui décrivent les données (gènes, protéines...), indiquent les noms d'outils utilisés ou encore le nom de cette chaîne de traitement. Au contraire, des informations plus techniques n'apparaissent pas toujours dans les articles scientifiques. Il en résulte un nombre plus réduit d'occurrences pour des entités telles que *Version* ou *Environnement* par exemple.

2.3 Qualité des annotations

Le tableau 2 donne les accords inter-annotateurs obtenus entre les différentes annotatrices deux à deux pour les nouveaux articles annotés. Les scores calculés sont tous supérieurs à 70 % en mode relâché : ceci signifie que le guide d'annotation est suffisamment intelligible, complet et non ambigu et qu'il existe peu de divergences dans la manière d'annoter des quatre annotatrices.

	A2	**A3**	**A4**
A1	*66,7* 72,7	*83,2* 86,3	*79,8* 80,6
A2		*69,7* 75,1	*66,8* 73,0

TABLE 2 – Accord inter-annotateur en *mode strict* et en mode relâché entre annotatrices ayant annoté des articles communs (en pourcentage).

Toutefois, le tableau 3 présente le détail des accords pour chaque type d'entité. L'hétérogénéité observée suggère que certaines entités sont plus simples à annoter que d'autres (Fort *et al.*, 2012).

Entités	**P**	**R**	**F1**	**Entités**	**P**	**R**	**F1**
Biblio	*96,8* 96,8	*99,6* 99,6	*98,2* 98,2	Environment	*57,1* 57,1	*52,6* 52,6	*54,8* 54,8
Container	*80,0* 80,0	*100,0* 100,0	*88,9* 88,9	ManagementSystem	*90,7* 91,5	*95,5* 96,4	*93,0* 93,9
Data	*65,8* 70,4	*62,7* 67,3	*64,2* 68,8	Method	*57,7* 61,9	*61,7* 66,2	*59,6* 64,0
Description	*56,5* 62,8	*68,1* 75,5	*61,7* 68,6	Tool	*76,8* 80,5	*79,8* 84,0	*78,2* 82,2

TABLE 3 – Détail des scores moyens en pourcentage obtenus pour certaines entités en *mode strict* et en mode relâché.

Tandis que des entités telles que *Biblio* ou *Container* ont des scores d'accord élevés, démontrant que la tâche d'annotation pour ces entités est simple, d'autres entités obtiennent au contraire des scores très inférieurs. Nous analysons ci-après trois causes possibles de ce constat.

La première cause identifiée est celle de l'ambiguïté dans le tagset (Fort *et al.*, 2012). C'est le cas des entités *Description* et *Method*. Après échange avec les annotatrices, il ressort que certains articles

scientifiques décrivent les méthodes bioinformatiques sans nécessairement les nommer. Les entités *Description* et *Method* sont alors souvent imbriquées et pas toujours délimitées de façon identique. Dans Zhang & Jonassen (2020) par exemple, la phrase « When the user is satisfied with the quality of the reads, the workflow proceeds to the next step : *quantification of read abundance or expression level* for transcripts or genes », l'entité *quantification of read abundance or expression level* a été annotée soit en tant que *Description*, soit en tant que *Method* selon les annotatrices.

Une deuxième cause identifiée est relative au critère de discrimination de Fort *et al.* (2012). Par exemple, les entités *Data* et *Description* ont une quantité d'annotations variant très fortement d'une annotatrice à l'autre. Certains articles sont rédigés par des bioinformaticiens, décrivant de façon précise les aspects méthodologiques et techniques, tandis que d'autres sont centrés sur le résultat biologique fourni. Dans cette seconde catégorie d'articles, les termes désignant des objets biologiques (gènes, RNA, SNP…) ont été annotés en *Data* par certaines annotatrices puisqu'il s'agit de la désignation de données. D'autres annotatrices les ont annotés en tant que *Description* ou pas annoté du tout, considérant qu'il s'agissait du contexte (biologique) de l'article.

Une troisième cause identifiée est le domaine d'expertise de l'annotatrice, qui influe sur son choix d'annotation. Par exemple, des bibliothèques telles que *Numpy* ou *Scikit-Learn* sont parfois annotées comme des outils bioinformatiques par des annotatrices ayant une formation initiale en biologie tandis que les bioinformaticiennes issues de l'informatique vont les annoter comme des *LibraryPackage*.

3 Expériences d'extraction d'entités nommées

3.1 Cadre expérimental

Choix du modèle et de son implémentation. Pour l'extraction de nos entités cibles, nous avons choisi le modèle neuronal biLSTM-CRF de Wajsbürt (2021), implémenté par l'outil NLStruct [3], dans sa version 0.2.0. Ce modèle est en effet capable de prendre en compte les entités imbriquées, ce qui est nécessaire pour certaines de nos entités, et a par ailleurs montré de bonnes performances dans le domaine biomédical, proche du nôtre. Nous avons plus précisément utilisé ce modèle dans deux configurations : d'une part avec le modèle de langue BERT (Devlin *et al.*, 2019), entraîné en domaine général ; d'autre part avec le modèle de langue SciBERT (Beltagy *et al.*, 2019), plus spécifiquement entraîné à partir d'articles scientifiques et a priori plus adapté à notre cas de figure.

Expériences réalisées. Nous avons classiquement choisi de répartir les articles en deux grands ensembles : un premier ensemble regroupant 75 % du corpus pour constituer le jeu d'entraînement (soit 39 articles) et un second ensemble correspondant aux 25 % restants pour le jeu de test (soit 13 articles), la séparation entre ces deux ensembles se faisant par tirage aléatoire. Au sein du jeu d'entraînement, 2/3 des articles (soit 26 articles) sont utilisés pour l'entraînement proprement dit des modèles et 1/3 pour leur validation (soit 13 articles). La particularité ici est que nous avons construit quatre volets pour ce découpage, par tirage aléatoire, afin de limiter la dépendance à un découpage particulier. Pour chaque découpage, quatre versions du modèle de reconnaissance d'entités sont produites avec des graines aléatoires différentes. Les hyperparamètres sont donnés en Annexe B.

3. https://github.com/percevalw/NLStruct

3.2 Résultats obtenus

Les scores obtenus par les modèles de langue testés BERT et SciBERT en faisant varier les graines aléatoires et les différents jeux d'entraînement et de validation sont présentés dans le tableau 4. L'empreinte carbone de tous les entraînements et évaluations est équivalent à 413 grammes de CO_2, calculée à l'aide de Green Algorithms [4]. Les scores entre chacun de nos volets sont globalement proches, la différence entre les volets extrêmes ne dépassant pas 1,7 point de F1 dans les deux configurations. Nous avons utilisé le test *Almost Stochastic Order* (Dror *et al.*, 2019) avec un niveau de confiance de 0,05 pour mesurer la significativité entre les deux modèles et obtenons que les modèles entraînés sur SciBERT sont stochastiquement dominants par rapport à ceux entraînés sur BERT ($\epsilon_{min} = 0$). Dans chacun des cas, utiliser un modèle de langue pré-entraîné à partir d'un corpus d'articles scientifiques issus des domaines biomédical et informatique (SciBERT) est plus performant.

	BERT			**SciBERT**		
	P	**R**	**F1**	**P**	**R**	**F1**
V1	65,2 ± 0,7	68,5 ± 0,6	66,8 ± 0,3	70,4 ± 0,3	70,2 ± 0,6	70,3 ± 0,4
V2	66,7 ± 0,3	66,8 ± 0,2	66,8 ± 0,1	70,7 ± 0,4	68,8 ± 0,6	69,7 ± 0,1
V3	66,9 ± 0,2	67,6 ± 0,6	67,3 ± 0,3	70,5 ± 0,5	71,0 ± 0,4	70,7 ± 0,4
V4	68,0 ± 0,3	69,0 ± 0,4	68,5 ± 0,2	71,2 ± 0,5	71,6 ± 1,0	71,4 ± 0,6
All	66,7 ± 0,9	68,0 ± 0,9	67,3 ± 0,6	70,7 ± 0,5	70,4 ± 1,1	70,5 ± 0,7

TABLE 4 – Moyenne des scores (et écarts-types) en pourcentage obtenus avec l'outil NLStruct en mode relâché pour chaque volet de découpage entraînement/validation. La moyenne des scores est calculée en fonction des résultats obtenus pour chaque graine aléatoire.

Le détail des scores de certaines entités sont données en *mode strict* et en mode relâché dans le tableau 5. Les scores F1 varient de 30 % à 98 %. Les scores faibles s'accordent avec les difficultés mises en exergue lors de l'annotation manuelle. L'entité *Method* a, en particulier, plus de mal à être extraite. Au contraire, d'autres (*Biblio* ou *Container*) obtiennent des scores F1 supérieurs à 80 %.

Les premiers résultats obtenus sur ce nouveau corpus étendu sont légèrement inférieurs (67,3 % avec le modèle BERT et 70,5 % avec SciBERT) à ceux obtenus par Sebe *et al.* (2023), dont les scores sont de 70,7 % avec un modèle entraîné sur BERT et 72,4 % sur un modèle entraîné avec SciBERT. Par ailleurs, la stabilité des résultats obtenus pour nos différents volets suggère que la taille du corpus annoté est maintenant suffisante pour l'entraînement des modèles de reconnaissance d'entités bioinformatiques. Pour mieux comprendre le fonctionnement des modèles testés, nous avons étudié la capacité des modèles à mémoriser.

3.3 Mémorisation et généralisation

Afin de déterminer l'impact de la mémorisation sur les performances de l'extraction d'entités, nous avons évalué les performances d'une baseline réalisant une simple projection des entités du corpus d'entraînement sur le corpus de test, en fonction de nos différents jeux d'entraînement avec un outil

4. http://calculator.green-algorithms.org/

Entités	P	R	F1	Entités	P	R	F1
Biblio	*94,1*	*96,4*	*95,2*	Environment	*64,1*	*90,9*	*75,1*
	96,1	98,1	97,1		65,2	92,5	76,3
Container	*92,9*	*81,9*	*86,9*	ManagementSystem	*65,0*	*81,5*	*72,3*
	92,9	81,9	86,9		66,7	83,6	74,1
Data	*51,5*	*45,7*	*48,4*	Method	*26,0*	*54,5*	*35,2*
	62,6	56,4	59,3		30,0	63,3	40,7
Description	*37,3*	*36,5*	*36,9*	Tool	*65,7*	*63,7*	*64,7*
	58,7	58,1	58,4		72,7	69,1	70,9

TABLE 5 – Détail des scores moyens en pourcentage obtenus pour certaines entités en *mode strict* et en mode relâché pour le premier volet de l'outil NLStruct (moyenne des résultats obtenus en fonction des différentes graines aléatoires).

proposé par Grouin (2016). Le tableau 6 présente les performances obtenues, qui sont globalement très inférieures à celles des modèles d'extraction des entités.

	P	R	F1
Propagation	*23,6* $\pm$ *0,3*	*13,83* $\pm$ *0,7*	*17,4* $\pm$ *0,6*
	45,8 $\pm$ 0,4	30,6 $\pm$ 1,7	36,6 $\pm$ 1,1

TABLE 6 – Moyenne des scores (et écarts-types) en pourcentage obtenus avec un outil de propagation d'annotation en *mode strict* et en mode relâché en fonction de chaque jeu d'entraînement.

Entités	P	R	F1	Entités	P	R	F1
Biblio	*67,5*	*33,5*	*44,8*	Environment	*39,1*	*37,7*	*33,1*
	81,1	43,4	56,5		75,0	61,0	59,6
Container	*92,9*	*100,0*	*96,3*	ManagementSystem	*89,6*	*96,1*	*92,2*
	92,9	100,0	96,3		89,6	96,1	92,2
Data	*17,6*	*6,2*	*9,1*	Method	*8,6*	*7,5*	*8,0*
	61,6	26,3	36,5		25,6	23,7	24,6
Description	*3,7*	*3,6*	*3,7*	Tool	*29,4*	*41,1*	*34,2*
	21,8	22,3	22,0		34,2	49,4	40,4

TABLE 7 – Moyenne des scores obtenus en pourcentage avec l'outil de propagation d'annotation en *mode strict* et en mode relâché en fonction des différents jeux d'entraînement sur certaines entités.

Dans le tableau 7 relatif aux résultats de certaines entités, on observe une grande hétérogénéité dans la distribution des scores. Ainsi en *mode strict*, trois catégories d'entités semblent se distinguer.

La première catégorie correspond aux entités pour lesquelles la mémorisation fonctionne très bien. Il s'agit des entités prenant un nombre restreint de valeurs, telles les entités *Container* ou *Management-*

System. Les performances obtenues dans ce cas par la simple mémorisation sont meilleures que celles des modèles neuronaux.

La deuxième catégorie regroupe les entités pour lesquelles la mémorisation présente des performances moyennes. C'est le cas de l'entité *Tool* : certains outils bioinformatiques génériques sont communs à de nombreuses chaînes de traitement alors que d'autres correspondent à des outils spécifiques à chaque domaine bioinformatique. Le modèle sait reconnaître les outils génériques souvent cités car ils ont été annotés précédemment dans le corpus *BioToFlow* mais ne peut distinguer les outils spécifiques si ces derniers n'y figuraient pas. Ainsi, comme pour les entités où la mémorisation est efficace, on peut penser que l'injection de connaissances spécifiques au domaine dans les modèles neuronaux pourrait être bénéfique pour la reconnaissance d'entités. La contribution potentielle d'une telle injection est d'autant plus intéressante dans le domaine bioinformatique que des bases de connaissances recensant des informations sur les outils comme BioTools [5] (Ison *et al.*, 2016), les conteneurs et systèmes de gestion des chaînes de traitement existent déjà.

La dernière catégorie correspond aux entités que le modèle ne peut mémoriser, par exemple *Description* et *Method*. Ces entités sont complexes car pouvant être composées d'un ou plusieurs tokens. Les modèles devront posséder la capacité de généraliser pour extraire ce type d'entités.

3.4 Comparaison des chaînes de traitement Nextflow et Snakemake

Une autre perspective à explorer pour l'amélioration des scores d'extraction consiste à traiter les articles relatifs aux deux systèmes de gestion Nextflow et Snakemake de manière séparée. De fait, même s'il s'agit de deux systèmes très utilisés dans la communauté bioinformatique, il semblerait qu'ils soient utilisés par des communautés d'utilisateurs distinctes, ce qui pourrait impliquer des différences de style et de structure des articles.

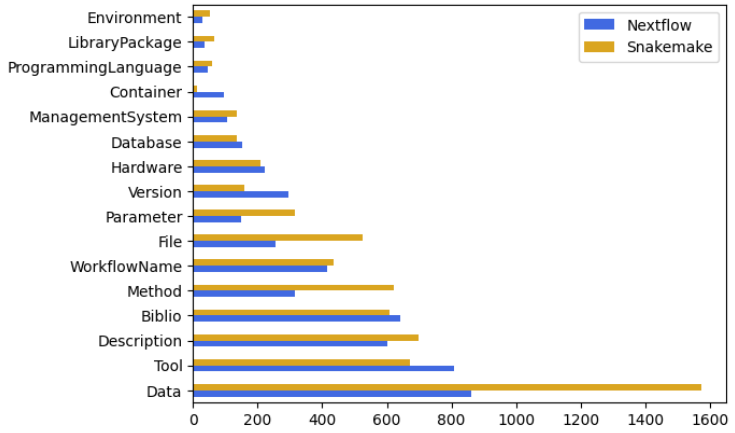

FIGURE 2 – Occurrences des entités figurant dans le corpus enrichi selon les systèmes de gestion Nextflow et Snakemake.

5. https://bio.tools/

La figure 2 représente la répartition de chacun des types d'entités dans les articles relatifs au système Nextflow et dans ceux relatifs au système Snakemake. On constate que les articles décrivant des chaînes de traitement sous Snakemake comportent presque le double de mentions d'entités *Data* et *Method* que ceux se rapportant à Nextflow. A contrario dans les articles relatifs à Nextflow, figurent plus souvent les noms des outils bioinformatiques utilisés. Tester deux modèles de langue différents est donc une piste à explorer pour l'amélioration des scores d'extraction.

4 Conclusion et perspectives

L'intersection entre le domaine du traitement automatique des langues et celui de la bioinformatique ouvre de nouvelles perspectives pour la recherche et l'analyse des chaînes de traitement bioinformatiques contenues dans la littérature. Ainsi, nous présentons ici deux contributions portant sur l'introduction d'un corpus de 52 articles étendant *BioToFlow* et sur l'utilisation de ce corpus pour l'entraînement et l'évaluation de modèles d'extraction d'entités.

Constitution du corpus *BioToFlow* étendu. Nous proposons une nouvelle version du corpus *BioToFlow* [6] composée de 52 articles variés annotés en termes d'entités nommées. Ce corpus est riche en entités, tant en nombre qu'en variété.

Lors de l'annotation manuelle par de nouvelles annotatrices, de nouveaux questionnements ont émergé sur la définition de certaines entités, notamment la distinction entre les entités *Method* et *Description*. Ces points devront être rediscutés entre annotatrices afin de diminuer les désaccords et le guide d'annotation devra être mis à jour en conséquence. Une fois ce travail réalisé, nous pouvons espérer un accroissement des performances de nos modèles d'extraction d'entités nommées.

Méthodes d'extraction d'entités nommées. Les résultats d'extraction des entités nommées obtenus sur notre corpus d'articles sont hétérogènes. Certaines entités obtiennent de très bons scores d'extraction, supérieurs à 80 %, alors que d'autres présentent des scores très inférieurs. Ceci s'explique par la combinaison des facteurs suivants :

- l'ambiguïté citée dans le paragraphe précédent concernant la distinction entre les entités *Method* et *Description* ;
- les inégalités élevées en matière de fréquence d'occurrence des entités ;
- les limites des modèles de langue utilisés en matière de connaissance du vocabulaire spécifique à la bioinformatique et aux chaînes de traitement, en particulier en ce qui concerne les noms des outils bioinformatiques.

Ce dernier point ne pourra être résolu que par l'injection de connaissances spécifiques au domaine des chaînes de traitement bioinformatiques dans les modèles de langue.

Futurs travaux L'extraction des chaînes de traitement dans les articles scientifiques nécessite non seulement l'extraction de leurs constituants, ce qui est l'objet du travail présenté, mais également des relations qu'ils entretiennent. La prochaine étape de ce travail est donc la définition de modèles pour

6. https://doi.org/10.5281/zenodo.11204427

l'extraction de ces relations. À plus long terme, la liaison référentielle entre les composants extraits des chaînes de traitement dans la littérature scientifique en anglais et ceux se trouvant dans les codes issus de dépôts publics sera nécessaire pour fusionner les informations venant des deux sources.

Remerciements

Nous remercions Noémie Bossut et Marie Schmit pour leur précieuse contribution à l'annotation des publications scientifiques. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-22-PESN-0007.

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A Pretrained Language Model for Scientific Text. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *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.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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.

DI TOMMASO P., CHATZOU M., FLODEN E. W., BARJA P. P., PALUMBO E. & NOTREDAME C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319. Number : 4 Publisher : Nature Publishing Group, DOI : 10.1038/nbt.3820.

DROR R., SHLOMOV S. & REICHART R. (2019). Deep Dominance - How to Properly Compare Deep Neural Models. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2773–2785, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1266.

FORT K. (2016). *Collaborative annotation for reliable natural language processing : Technical and sociological aspects*. John Wiley & Sons.

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

GROUIN C. (2016). Controlled Propagation of Concept Annotations in Textual Corpora. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4075–4079, Portorož, Slovenia : European Language Resources Association (ELRA).

ISON J., RAPACKI K., MÉNAGER H. & AL (2016). Tools and data services registry : a community effort to document bioinformatics resources. *Nucleic Acids Research*, **44**(D1), D38–D47. DOI : 10.1093/nar/gkv1116.

MÖLDER F., JABLONSKI K. P., LETCHER B., HALL M. B., TOMKINS-TINCH C. H., SOCHAT V., FORSTER J., LEE S., TWARDZIOK S. O., KANITZ A., WILM A., HOLTGREWE M., RAHMANN S., NAHNSEN S. & KÖSTER J. (2021). *Sustainable data analysis with Snakemake*. Rapport interne 10 :33, F1000Research. Type : article, DOI : 10.12688/f1000research.29032.1.

SEBE C., NÉVÉOL A., COHEN-BOULAKIA S. & GAIGNARD A. (2023). Extraction d'informations sur les workflows scientifiques à partir de la littérature. volume Extraction et Gestion des Connaissances, RNTI-E-39, p. 313.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a Web-based Tool for NLP-Assisted Text Annotation. In F. SEGOND, Éd., *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

VERSPOOR K., JIMENO YEPES A., CAVEDON L., MCINTOSH T., HERTEN-CRABB A., THOMAS Z. & PLAZZER J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, **2013**, bat019. DOI : 10.1093/database/bat019.

WAJSBÜRT P. (2021). *Extraction et normalisation d'entités simples et structurées dans les documents médicaux*. These de doctorat, Sorbonne université.

ZHANG X. & JONASSEN I. (2020). RASflow : an RNA-Seq analysis workflow with Snakemake. *BMC bioinformatics*, **21**(1), 110. DOI : 10.1186/s12859-020-3433-x.

A Exemple d'annotations

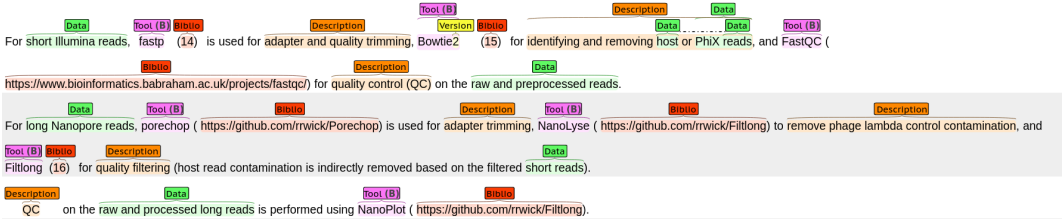

FIGURE 3 – Exemple d'annotation provenant de l'article PMID35118380.

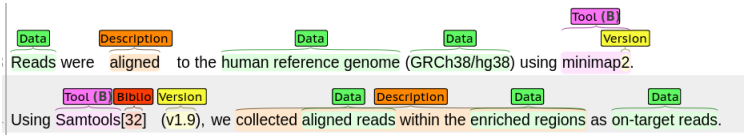

FIGURE 4 – Exemple d'annotation provenant de l'article PMID34103501.

B Hyperparamètres utilisés

Paramètre	Valeur
Encodeur de base	BERT-base-uncased SciBERT_scivocab_uncased
Longueur des séquences	256
Nombres d'itérations max.	5000
Optimiseur	AdamW
Taux d'apprentissage	1e-3
Dropout	0,1
Warmup ratio	0,1
Graines aléatoires	1 - 8 - 22 - 42

TABLE 8 – Hyperparamètres utilisés pour Nlstruct.