Où la frugalité rejoint l'éthique : utilisation de données synthétiques pour la reconnaissance d'entités cliniques

Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

Où la frugalité rejoint l'éthique : utilisation de données synthétiques pour la reconnaissance d'entités cliniques

Nicolas Hiebel[1] Olivier Ferret[2] Karën Fort[3] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, 54506, Vandœuvre-lès-Nancy, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr,
[3]karen.fort@loria.fr

Contexte

Dans certains domaines du traitement automatique des langues (TAL), comme le domaine médical, le fait de ne pas pouvoir partager des données crée une contrainte de frugalité. Les corpus cliniques dont l'accès est relativement facile en français (E3C (Magnini *et al.*, 2020), CAS (Grabar *et al.*, 2018)) ne sont pas tout à fait représentatifs des documents confidentiels présents dans les hôpitaux. Le partage des connaissances au sein de la communauté scientifique est compliqué. Aucune reproductibilité n'est possible, tout comme les comparaisons avec d'autres méthodes ou données. Une piste pour s'affranchir de ce manque de données partageables est de générer des documents synthétiques similaires aux documents réels mais ne présentant pas d'informations sensibles. Cela pourrait permettre à des personnes ayant accès à un corpus protégé de générer un corpus librement distribué à partir du premier. En partageant la méthode de génération, il serait également possible de reproduire l'expérience sur d'autres données confidentielles. La mise à disposition des données générées donnerait alors à la communauté scientifique un terrain de test, de comparaison, de discussion et d'entraide dans la recherche en TAL biomédical, tout en préservant la confidentialité des données.

Cependant, l'utilité des données synthétiques générées n'est pas facile à évaluer, particulièrement dans le cas où le corpus dont s'inspirent les données synthétiques n'est pas annoté. Dans le cadre de l'évaluation de la qualité des données synthétiques, nous présentons ici un protocole pour mesurer la pertinence des données synthétiques afin d'entraîner des modèles de TAL.

Objectifs À la base de ce travail, nous générons des textes synthétiques à partir de cas cliniques réels selon plusieurs configurations et nous souhaitons mesurer leur qualité. L'objectif est d'obtenir des données synthétiques partageant suffisamment de caractéristiques avec les données réelles, de manière à ce qu'il soit aussi pertinent de travailler sur les données synthétiques que sur les données réelles. Dans le même temps, les données synthétiques doivent être suffisamment éloignées des données réelles pour limiter le risque de divulguer des informations sensibles présentes dans les données réelles. Dans cette étude, nous évaluons la pertinence de ces cas cliniques synthétiques en nous intéressant à leur utilité pour une tâche classique en TAL : la reconnaissance d'entités nommées (REN), ici dans le domaine clinique. Parallèlement, nous évaluons la proximité des données synthétiques avec les données d'origine à l'aide de recouvrement de ngrammes et analysons manuellement les types d'erreurs présentes dans les cas cliniques générés.

Méthodologie

Corpus Deux corpus biomédicaux interviennent dans cette étude. Le premier est le corpus multilingue librement disponible E3C. Nous sélectionnons ici uniquement les cas cliniques en français, sous-ensemble que nous appellerons $\texttt{E3C}_{FR}$. Ce corpus est utilisé pour générer les documents synthétiques. Le second est le corpus MERLOT (Campillos *et al.*, 2018), un corpus clinique de 500 documents dont l'accès est restreint. Ce corpus contient des annotations manuelles en entités. Ces annotations sont utilisées à la fois pour entraîner des modèles de reconnaissance d'entités cliniques et pour évaluer les modèles entraînés sur les autres corpus.

Génération Nous avons ici choisi d'explorer la capacité des modèles neuronaux auto-régressifs pré-entraînés, comme GPT2 (Radford *et al.*, 2019) et BLOOM (Scao *et al.*, 2022), à s'adapter au domaine médical et à générer des cas cliniques. Nous avons sélectionné un modèle français que nous appellerons LLF [1] et un modèle multilingue, en l'occurrence une des versions du modèle BLOOM [2]. Pour pouvoir comparer les résultats, nous avons choisi deux modèles de taille à peu près équivalente (environ 1 milliard de paramètres). Les modèles sont entraînés à générer des documents cliniques entiers, en indiquant seulement le début et la fin d'un document par des balises. À la génération, seule une balise de début de document est donnée en amorce. Deux configurations d'entraînement sont testées pour chaque modèle. Pour la première, les modèles sont entraînés sur les données d'$\texttt{E3C}_{FR}$ brutes. Pour la seconde, les données d'$\texttt{E3C}_{FR}$ sont d'abord annotées par les modèles REN entraînés sur MERLOT et les annotations sont ensuite intégrées aux textes sous forme de balises de début et de fin d'entités. Les modèles sont entraînés sur les textes contenant les annotations. Ainsi, les modèles apprennent à générer les annotations en même temps que le texte. Les corpus synthétiques générés seront nommés $\texttt{Bloom}_{E3C}$ et $\texttt{LLF}_{E3C}$ pour les versions entraînées sans annotation et $\texttt{Bloom}_{E3C+T}$ et $\texttt{LLF}_{E3C+T}$ pour les autres.

Filtrage Les textes générés avec des modèles auto-régressifs peuvent présenter des défauts, comme des répétitions de tokens, ou bien un manque de diversité entre les différentes générations. C'est pourquoi nous générons beaucoup plus de tokens que le nombre de tokens visé (pour obtenir pour chaque configuration autant de tokens que dans le jeu d'entraînement). Nous avons choisi ici de sélectionner des documents divers et d'éliminer les cas où la génération présente une erreur facilement repérable (tokens extrêmement longs, boucle de répétition de tokens...).

Trois niveaux d'évaluation Les textes générés sans annotation sont annotés de la même manière qu'$\texttt{E3C}_{FR}$ à l'aide des modèles REN entraînés sur MERLOT. Ensuite, de nouveaux modèles REN sont entraînés sur chaque corpus et tous les modèles REN sont testés sur les jeux de test de tous les corpus annotés obtenus.

Nous ajoutons à cela une mesure du pourcentage de recouvrement de ngrammes entre chaque corpus généré et le corpus d'origine $\texttt{E3C}_{FR}$ pour évaluer la proximité des corpus.

Enfin, nous avons réalisé une analyse manuelle de 15 documents pour chaque configuration (pour un total de 60 documents et 13 809 tokens) visant à évaluer la grammaticalité et la cohérence clinique

1. https://huggingface.co/asi/gpt-fr-cased-base
2. https://huggingface.co/bigscience/bloom-1b1

	Test					
	E3C$_{FR}$			**MERLOT**		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C$_{FR}$	**89,5**	**91,0**	**90,2**	64,4	78,0	70,5
MERLOT	87,1	90,8	88,9	**85,2**	**85,8**	**85,5**
Bloom$_{E3C}$	87,6	87,9	87,8	63,1	74,9	68,5
LLF$_{E3C}$	87,6	87,2	87,4	64,5	76,4	70,0
Bloom$_{E3C+T}$	83,4	68,9	75,4	71,7	55,1	62,3
LLF$_{E3C+T}$	84,4	68,1	75,4	75,7	46,4	57,5

TABLEAU 1 – Résultats de la tâche de REN sur les corpus réels (P=précision, R=rappel, F=F-mesure).

des documents générés.

Résultats

Le tableau 1 présente les résultats de la reconnaissance d'entités cliniques sur les jeux de test des corpus réels. Ce tableau montre la pertinence des corpus générés pour notre tâche de reconnaissance d'entités cliniques, notamment avec les résultats sur le corpus MERLOT. Comme il a été annoté manuellement, contrairement aux autres corpus, nous considérons MERLOT comme une référence de haute qualité. Bien que les modèles entraînés sur E3C$_{FR}$ soient significativement moins performants que les modèles entraînés sur MERLOT, nous pouvons observer que les performances des modèles entraînés sur Bloom$_{E3C}$ et LLF$_{E3C}$ sont proches de celles des modèles entraînés sur E3C$_{FR}$. Ainsi, dans notre contexte, l'utilisation d'un corpus généré à partir d'un corpus réel est relativement équivalente à l'utilisation du corpus réel. Cela se confirme avec les résultats sur le corpus E3C$_{FR}$.

Le tableau 2 présente le recouvrement de ngrammes entre le corpus E3C et les corpus générés et un autre corpus réel de cas cliniques, CAS.

	Corpus	1gram	...	4gram	...	8gram
Synthétique	Bloom$_{E3C}$	0,16419		0,00368		0,00011
	Bloom$_{E3C+T}$	0,13887		0,00531		0,00020
	LLF$_{E3C}$	0,11740		0,00447		**0,00023**
	LLF$_{E3C+T}$	0,11935		0,00505		0,00013
Réel	CAS	**0,20373**		0,00899		0,00013

TABLEAU 2 – Recouvrement de ngrammes entre les corpus générés et E3C. Chaque ligne correspond à la comparaison entre le corpus et E3C. La comparaison entre CAS et E3C sert de baseline.

On constate que le corpus réel CAS présente le plus de 1-grammes, donc de vocabulaire, en commun avec le corpus E3C. En revanche, CAS est l'un des corpus comparés avec le moins de séquences longues (8-grammes) en commun avec E3C. Parmi les corpus générés, le corpus Bloom$_{E3C}$ est prometteur car il contient le plus grand nombre de 1-grammes en commun et le plus petit nombre de 8-grammes en commun. À l'inverse, le corpus LLF$_{E3C}$ possède le plus petit nombre de 1-grammes en commun tout en ayant le plus grand nombre de 8-grammes en commun, ce qui est moins souhaitable.

Conclusion

Nous présentons dans ce travail quatre modèles de génération de cas cliniques synthétiques en français et nous montrons qu'entraîner des modèles de reconnaissance d'entités cliniques sur ces textes synthétiques est pratiquement équivalent à entraîner des modèles sur les données réelles dont ils sont issus, sans pour autant les copier.

Références

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2018). A French clinical corpus with comprehensive semantic annotations : development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, **52**(2), 571–601. DOI : 10.1007/s10579-017-9382-y.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 de *CEUR Workshop Proceedings* : CEUR-WS.org.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). *Language Models are Unsupervised Multitask Learners*. Rapport interne, OpenAI.

SCAO T. L. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. DOI : 10.48550/ARXIV.2211.05100.