Les textes cliniques français générés sont-ils dangereusement similaires à leur source ? Analyse par plongements de phrases

Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

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

Les textes cliniques français générés sont-ils dangereusement similaires à leur source ? Analyse par plongements de phrases

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, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr,
[3]karen.fort@loria.fr

RÉSUMÉ

Les ressources textuelles disponibles dans le domaine biomédical sont rares pour des raisons de confidentialité. Des données existent mais ne sont pas partageables, c'est pourquoi il est intéressant de s'inspirer de ces données pour en générer de nouvelles sans contrainte de partage. Une difficulté majeure de la génération de données médicales est que les données générées doivent ressembler aux données originales sans compromettre leur confidentialité. L'évaluation de cette tâche est donc difficile. Dans cette étude, nous étendons l'évaluation de corpus cliniques générés en français en y ajoutant une dimension sémantique à l'aide de plongements de phrases. Nous recherchons des phrases proches à l'aide de similarité cosinus entre plongements, et analysons les scores de similarité. Nous observons que les phrases synthétiques sont thématiquement proches du corpus original, mais suffisamment éloignées pour ne pas être de simples reformulations qui compromettraient la confidentialité.

ABSTRACT

Are Synthesized Clinical Texts in French Dangerously Similar to Their Source ? An Analysis Using Sentence Embeddings.

Textual resources available in the biomedical field are scarce due to confidentiality issues. This problem can be addressed by automatically generating shareable data from existing restricted data. However, in the medical field, generated data should not contain sensitive information from restricted training data while remaining as close as possible to it. This paradox makes the evaluation of synthetic text challenging. In this study, we extend the evaluation of generated clinical corpora in French with semantic representations of sentences using sentence embeddings. We use cosine similarity between sentence embeddings to find similar sentences and analyze the similarity scores. We observe that the generated sentences are thematically close to those of the original corpus while being distant enough to avoid compromising confidentiality.

MOTS-CLÉS : Génération, Évaluation, Similarité, Texte clinique, Texte synthétique, Français.

KEYWORDS: Generation, Evaluation, Similarity, Clinical Text, Synthetic Text, French.

1 Introduction

Le manque de ressources est l'un des problèmes les plus communément rencontrés dans le Traitement Automatique des Langues (TAL), que ce soit en termes de domaine, de tâche, ou les deux. C'est le

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 46–54.

cas dans le domaine biomédical, où les ressources disponibles dans des langues autres que l'anglais sont rares (Névéol *et al.*, 2018). Ainsi, les données stockées dans les hôpitaux ne sont accessibles que par un nombre très restreint de personnes. Les données ne pouvant pas être diffusées, le partage des connaissances au sein de la communauté scientifique est difficile. Les possibilités de reproduction d'expériences et de comparaisons méthodologiques sont limitées.

Une piste pour résoudre ce problème est de générer de nouvelles données similaires aux données privées tout en préservant la confidentialité. La mise à disposition des données générées pourraient alors devenir un terrain de test, de comparaison, de discussion et d'entraide dans la recherche en TAL biomédical. Les modèles de type *transformer* (Vaswani *et al.*, 2017) ont montré leur efficacité dans différentes tâches de génération de textes (traduction, résumé, etc.). Les modèles auto-régressifs pré-entraînés comme GPT-2 (Radford *et al.*, 2019), GPT-3 (Brown *et al.*, 2020) et plus récemment InstructGPT (Ouyang *et al.*, 2022) et son successeur ChatGPT ont la capacité de générer des textes bien écrits. Les connaissances acquises lors du pré-entraînement de ces modèles pourraient être utiles dans des domaines peu dotés où il n'est pas possible d'entraîner un modèle génératif en partant de zéro.

L'évaluation automatique de la génération repose essentiellement sur des mesures de similarité avec une référence (Frisoni *et al.*, 2022) que l'on considère comme la réponse idéale attendue dans un contexte donné. La référence est comparée avec la ou les hypothèses du système. Les plus connues sont les mesures BLEU (Papineni *et al.*, 2002) et ROUGE (Lin, 2004) qui se basent sur des recouvrements de n-grammes. Cependant, dans le cas d'une génération ouverte, il n'y a pas de référence.

Nous proposons ici des alternatives à ces mesures pour évaluer une génération ouverte dans le domaine médical. Nous utilisons la méthode de génération mise en oeuvre par Hiebel *et al.* (2023). Dans ce travail, la génération est faite à l'aide d'un ajustement (*fine-tuning*) de modèles auto-régressifs pré-entraînés sur un corpus clinique dans plusieurs configurations.

Nous utilisons ici les corpus générés dans ce travail et nous proposons une évaluation sémantique des corpus en utilisant des similarités phrastiques obtenues à l'aide du modèle de plongements de phrases SENTENCE-BERT (SBERT) (Reimers & Gurevych, 2019) ajusté selon plusieurs configurations. Cette évaluation automatique supplémentaire permet d'estimer la proximité sémantique entre le contenu des corpus générés et le corpus réel. Nos contributions pour l'évaluation de la génération dans le domaine clinique sont les suivantes :

— nous présentons deux utilisations des plongements contextuels de phrases pour déterminer la distance des phrases des corpus générés avec les phrases du corpus original ;
— nous confirmons la qualité de la génération d'un point de vue sémantique ;
— nous identifions une piste d'amélioration dans la génération pour faciliter l'évaluation.

2 Méthode

2.1 Corpus utilisés

E3C (Magnini *et al.*, 2020) est un corpus multilingue librement disponible composé de documents médicaux provenant de différentes sources. Comme le travail de Hiebel *et al.* (2023), nous sélectionnons ici les cas cliniques en français du corpus.

CAS (Grabar *et al.*, 2018) est un corpus médical français contenant des cas cliniques français dé-identifiés dont la publication a été consentie par les patients concernés.

DEFT STS (Cardon & Grabar, 2020) est un corpus français de paires de phrases annotées en scores de similarité. Les phrases proviennent du corpus CLEAR (Grabar & Cardon, 2018), un corpus de phrases parallèles ayant pour but d'associer des phrases complexes avec leur version simplifiée. Sous ensemble de ce corpus, le corpus DEFT STS contient 1 010 paires de phrases qui ont été annotées en se reposant sur l'intuition des annotateurs. Certaines de ces paires portent sur le domaine biomédical.

CLISTER (Hiebel *et al.*, 2022) est un corpus clinique français de 1 000 paires de phrases annotées en scores de similarité. Ce corpus contient des phrases du corpus CAS. Les annotations ont été faites spécialement pour s'adapter au domaine clinique, avec une notion de compatibilité clinique entre les phrases, qui consiste à observer si les phrases peuvent ou non correspondre au même patient.

Est Républicain (ATILF & CLLE, 2020) est un corpus journalistique composé d'articles parus dans le quotidien éponyme. Nous avons ici sélectionné une sous-partie du corpus de manière à obtenir un corpus hors domaine de même taille que le corpus clinique E3C.

2.2 Génération des textes cliniques synthétiques

La génération des textes cliniques synthétiques est faite en ajustant des modèles auto-régressifs sur le corpus E3C.

Deux modèles différents ont été testés, le modèle multilingue BLOOM (Scao *et al.*, 2022) et un modèle français que nous appelons ici LLF (Simoulin & Crabbé, 2021). Pour une comparaison équitable, nous avons choisi deux modèles d'environ un milliard de paramètres.

Chaque modèle a été entraîné avec deux configurations différentes, dont l'une où des annotations en entités cliniques sont ajoutées au texte sous la forme de balises XML pour que le modèle génère directement des annotations. Les documents ont été générés avec comme unique contrainte une amorce (*prompt*) sous forme d'un token marquant le début d'un document. La génération a été faite en utilisant des paramètres de décodage favorisant la diversité des documents générés avec une température de 1,5 pour la génération annotée, 1,2 pour la génération non annotée, ainsi qu'une pénalité de répétition de 10.

La table 1 présente les statistiques des corpus réels et générés que nous souhaitons comparer, ne comprenant donc pas les corpus de similarité phrastique. Les corpus générés en incluant des annotations sont notés avec le suffixe « $_{+T}$ ».

	Toks	Docs	Toks/doc	Phrases/doc	Long. phrases	Self-Bleu	Perplexité
Est Républicain	306 866	8 226	37,3	2,0	18,4	0,52	77,8
CAS	231 662	717	323,1	15,8	20,4	0,66	17,0
E3C	328 645	1 009	325,7	15,2	21,4	0,68	22,0
Bloom$_{E3C}$	329 328	943	349,2	1,8	194	0,70	9,97
Bloom$_{E3C+T}$	346 413	1 997	173,5	3,1	56,0	0,73	9,27
LLF$_{E3C}$	328 498	1 028	319,6	6,9	46,3	0,68	10,03
LLF$_{E3C+T}$	336 154	978	343,7	7,2	47,7	0,67	11,9

TABLE 1 – Statistiques des corpus réels et des corpus générés, hors corpus de similarité phrastique.

2.3 Utilisation de plongements de phrases

Modèles de plongement de phrases Les plongements de phrases ont été obtenus en utilisant un modèle pré-entraîné multilingue de l'outil SBERT [1]. Afin d'observer différentes formes de similarité, nous avons calculé les plongements de phrases de tous les corpus selon trois configurations : une version avec le modèle pré-entraîné SBERT sans ajustement, une version avec le modèle ajusté sur le corpus DEFT STS et une version du modèle ajusté sur CLISTER.

Calcul des scores de similarité Pour obtenir une représentation de la proximité des phrases des différents corpus générés avec le corpus original E3C, nous recherchons pour chaque phrase du corpus généré les 100 phrases les plus proches dans le corpus E3C ainsi que les scores de similarité avec une similarité cosinus. Nous utilisons pour cela la bibliothèque FAISS (Johnson *et al.*, 2021). Nous obtenons pour chaque phrase générée 100 scores de similarité. Nous répétons le processus avec le corpus de cas cliniques CAS et le corpus hors domaine de l'Est Républicain.

Scores BLEU Nous souhaitons observer la proximité des phrases générées avec le corpus original à l'aide de la mesure BLEU. Cependant, contrairement au calcul des plongements de phrases en amont qui permet de de rechercher rapidement les éléments similaires dans la matrice de plongements, la mesure BLEU nécessite de calculer les similarités des phrases directement deux à deux, ce qui pose un problème de complexité dans notre cas où il faut comparer toutes les phrases du corpus source à toutes les phrases du corpus cible. C'est pourquoi nous observons ici uniquement les scores BLEU des phrases similaires déjà trouvées à l'aide d'un modèle de plongements de phrases.

Recherche des phrases du corpus E3C Comme deuxième approche pour observer la distance entre les corpus et E3C, nous essayons à partir des phrases d'E3C de retrouver dans la combinaison des phrases d'E3C et du corpus comparé les autres phrases du corpus E3C. Ainsi, plus il est simple de retrouver les autres phrases d'E3C, plus la distance entre les corpus est grande. Nous utilisons pour cela des mesures de recherche d'information.

3 Résultats

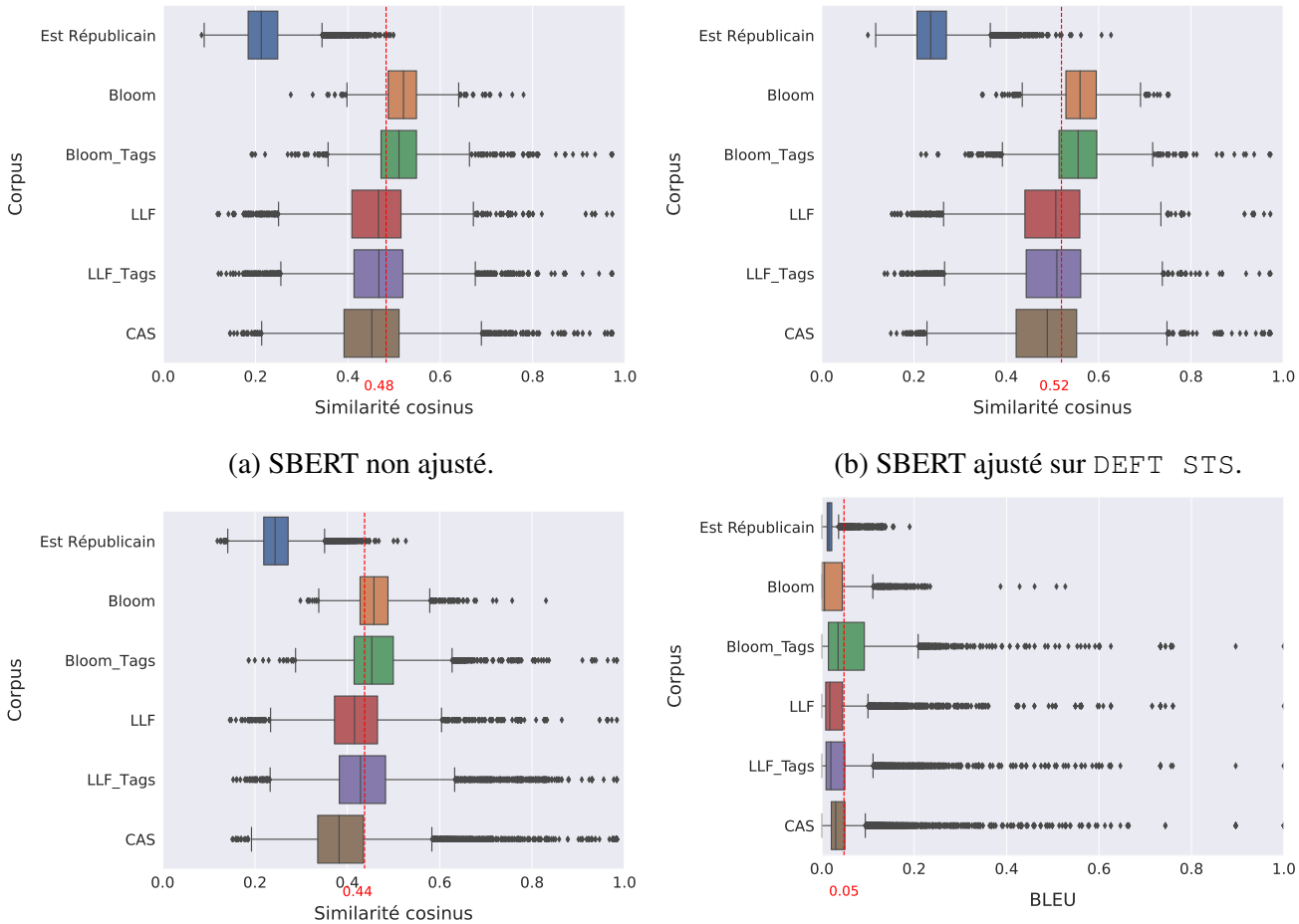
3.1 Distribution des similarités sémantiques des phrases

La figure 1 présente sous forme de diagrammes en boîtes les distributions des scores de similarité des phrases des corpus comparées aux phrases du corpus E3C. On y remarque facilement la mise à l'écart du corpus hors domaine de l'Est Républicain (en haut), surtout sur les figures 1a, 1b et 1c.

En ce qui concerne les modèles pour les corpus cliniques, les similarités les plus hautes sont obtenues avec le modèles SBERT ajusté sur le corpus DEFT STS. Cela peut s'expliquer à la fois par le fait qu'une portion des paires de phrases de ce corpus provient du domaine médical, et par la définition de similarité dans ce corpus qui se base sur des principes de simplification. Nous cherchons à repérer une thématique commune entre les phrases, et cela explique une similarité plus forte entre les corpus que celle obtenue avec le modèle SBERT non ajusté. Par ailleurs, les similarités obtenues avec le modèle SBERT ajusté sur CLISTER sont les plus faibles. Le modèle a été spécialisé sur le domaine clinique avec un critère de compatibilité clinique. Les similarités plus faibles des corpus cliniques avec ce modèle montrent que les phrases ont une thématique commune mais ne parlent pas forcément des

1. *distiluse-base-multilingual-cased-v1*

mêmes patients, ce qui est encourageant pour la qualité de la génération. Avec ce modèle, le corpus réel CAS se démarque un peu plus des corpus générés qu'avec les autres modèles, ce qui souligne une ressemblance entre E3C et CAS davantage surfacique que sémantique.



(a) SBERT non ajusté.

(b) SBERT ajusté sur DEFT STS.

(c) SBERT ajusté sur CLISTER.

(d) Scores BLEU.

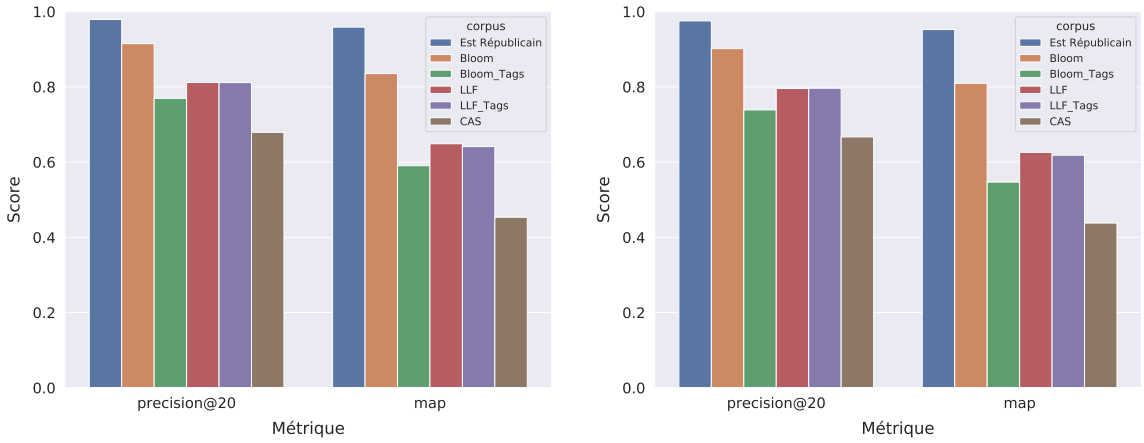
FIGURE 1 – Moyennes des scores de chaque phrase du corpus E3C avec les phrases les plus proches dans les corpus générés, CAS et un corpus hors domaine (Est Républicain). Pour les figures 1a, 1b et 1c, les similarités des 100 phrases les plus proches sont obtenues par similarité cosinus des plongements de phrases, issus de modèles SBERT. Pour la figure 1d, nous récupérons les 20 scores BLEU les plus élevés parmi les 100 phrases les plus proches trouvées avec les plongements de SBERT ajustés sur CLISTER. La ligne pointillée rouge correspond à la moyenne des scores des corpus cliniques.

La figure permet également d'observer la proximité des corpus générés ayant pour base le même modèle pré-entraîné. On constate que le corpus généré $\texttt{Bloom}_{E3C}$ se démarque particulièrement des autres corpus générés, avec une distribution plus resserrée et des similarités hors distributions moins extrêmes que dans les autres corpus. Cela s'explique sûrement par les très longues phrases du corpus $\texttt{Bloom}_{E3C}$. De par leur longueur, celles-ci contiennent mécaniquement plus d'informations, et chaque élément va influer sur l'encodage de la phrase dans un espace vectoriel réduit. Cela va donc

gommer les très fortes similarités de certaines sections de la phrase avec les phrases plus courtes de E3C, et inversement, il sera plus facile de trouver des phrases de E3C avec des points de similarité. Enfin, on observe sur la figure 1d que les scores BLEU des phrases sont en proportion beaucoup plus faibles que les scores obtenus avec les modèles de plongements de phrases, malgré l'étude des 20 phrases les plus similaires au lieu de 100. Le corpus $\texttt{Bloom}_{E3C+T}$ reste le plus similaire à E3C avec cette mesure, mais les autres corpus cliniques sont difficiles à différencier, illustrant le fait que cette mesure n'est pas la plus adaptée dans ce contexte.

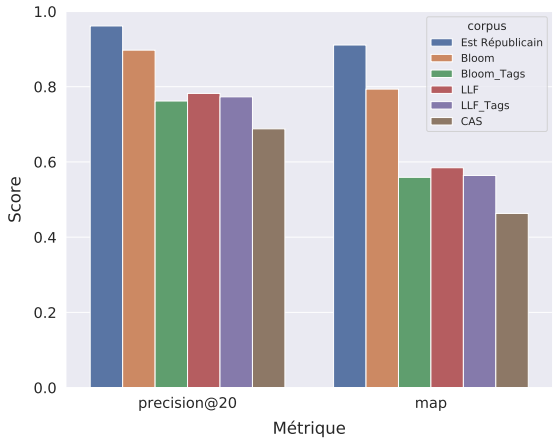
3.2 Recherche des phrases d'E3C

On observe sur la figure 2 les résultats de la recherche des phrases du corpus E3C dans les différentes combinaisons des phrases d'E3C et des corpus comparés.



(a) SBERT non ajusté.

(b) SBERT ajusté sur DEFT STS.



(c) SBERT ajusté sur CLISTER.

FIGURE 2 – Scores MAP et Précision@20 de la recherche des phrases de E3C parmi la combinaison des phrases de E3C et des phrases du corpus comparé. Les phrases sont représentées par des plongements de mots issus de Sentence-BERT. Nous recherchons parmi les 100 phrases les plus similaires.

On distingue logiquement le corpus de l'Est Républicain, dans lequel les phrases de E3C sont presque systématiquement retrouvées avec les trois modèles de plongements de phrases. Pour les corpus cliniques, on observe que Bloom_{E3C} présente des scores nettement plus élevés, CAS présente les scores les plus faibles et les scores des corpus générés Bloom_{E3C+T}, LLF_{E3C} et LLF_{E3C+T} se situent entre les deux.

Avec les trois modèles, les phrases de E3C sont retrouvées beaucoup plus facilement lorsqu'elles sont mélangées avec celles du corpus Bloom_{E3C}, au point que l'on se rapproche presque des scores obtenus avec le corpus hors domaine, particulièrement dans les premières positions (précision@20). Cela peut probablement s'expliquer par la différence de longueur des phrases, qui signifie aussi un nombre de phrases beaucoup plus faible. Proportionnellement, même en sélectionnant des phrases au hasard, il y a beaucoup plus de chance de tomber sur une autre phrase d'E3C qu'une phrase de Bloom_{E3C}. Concernant les modèles, on constate que les écarts entre les corpus cliniques sont moins importants lorsqu'on utilise les plongements obtenus par SBERT ajusté sur CLISTER.

Il est intéressant de constater qu'il est plus difficile de retrouver les autres phrases d'E3C en les mélangeant aux phrases du corpus CAS avec les trois modèles. Malgré les scores de similarité plus faibles pour CAS dans la figure 1, on constate ici que ce sont les phrases du corpus CAS qui sont les plus proches de celles d'E3C. Une première explication pourrait être les longueurs des phrases, très proches entre les corpus cliniques réels et beaucoup plus longues dans les corpus générés, spécialement pour Bloom_{E3C}, avec lesquelles les phrases d'E3C sont le plus facilement différenciées, se rapprochant presque du corpus hors domaine.

4 Conclusion

Nous avons présenté une évaluation automatique de corpus générés dans le domaine clinique à l'aide de mesures de similarité entre plongements de phrases issus de modèles implémentant plusieurs définitions de similarité sémantique. Les analyses montrent que les phrases synthétiques sont thématiquement proches des phrases du corpus dont elles sont inspirées, sans pour autant concerner les mêmes patients. Ces résultats suggèrent que les corpus générés sont proches du corpus original tout en apportant suffisamment d'innovation pour ne pas facilement pouvoir retrouver des patients. Cependant, une grande différence de longueur de phrases entre les corpus générés et le corpus original peut fausser ces comparaisons. Une perspective intéressante serait donc de contraindre la taille des phrases. Cela permettrait d'augmenter la proximité avec les corpus réels et d'améliorer l'évaluation de la génération.

Remerciements

Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, CODEINE (artificial text COrpus DEsIgNed Ethically), ANR-20-CE23-0026-01. Nous remercions par ailleurs Natalia Grabar (Université de Lille, CNRS, STL) qui nous a permis d'utiliser les corpus CAS et DEFT STS pour cette étude.

Références

ATILF & CLLE (2020). Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CARDON R. & GRABAR N. (2020). A French corpus for semantic similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 6889–6894, Marseille, France : European Language Resources Association.

FRISONI G., CARBONARO A., MORO G., ZAMMARCHI A. & AVAGNANO M. (2022). NLG-metricverse : An end-to-end library for evaluating natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3465–3479, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

GRABAR N. & CARDON R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7002.

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.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023). Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatia : Association for Computational Linguistics.

HIEBEL N., FORT K., NÉVÉOL A. & FERRET O. (2022). CLISTER : Un corpus pour la similarité sémantique textuelle dans des cas cliniques en français (CLISTER : A corpus for semantic textual similarity in French clinical narratives). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 287–296, Avignon, France : ATALA.

JOHNSON J., DOUZE M. & JÉGOU H. (2021). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, **7**, 535–547.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

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.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, **9**(1), 12. DOI : 10.1186/s13326-018-0179-8.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. DOI : 10.48550/ARXIV.2203.02155.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

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

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. 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 2019, Hong Kong, China, November 3-7, 2019*, p. 3980–3990 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

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

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.