Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation de la Similarité Textuelle : Entre Sémantique et Surface dans les Représentations Neuronales

Julie Tytgat[1, 2] Guillaume Wisniewski[1] Adrien Betrancourt[2]
(1) Université de Paris Cité, LLF, CNRS 75 013 Paris, France
(2) IPSIDE, 31 100 Toulouse
julie.tytgat@etu.u-paris.fr, guillaume.wisniewski@u-paris.fr,
a.betrancourt@ipside.com

RÉSUMÉ

La mesure de la similarité entre textes, qu'elle soit basée sur le sens, les caractères ou la phonétique, est essentielle dans de nombreuses applications. Les réseaux neuronaux, en transformant le texte en vecteurs, offrent une méthode pratique pour évaluer cette similarité. Cependant, l'utilisation de ces représentations pose un défi car les critères sous-jacents à cette similarité ne sont pas clairement définis, oscillant entre sémantique et surface. Notre étude, basée sur des expériences contrôlées, révèle que les différences de surface ont un impact plus significatif que les différences de sémantique sur les mesures de similarité entre les représentations neuronales des mots construites par de nombreux modèles pré-entrainés. Ces résultats soulèvent des questions sur la nature même de la similarité mesurée par les modèles neuronaux et leur capacité à capturer les nuances sémantiques.

ABSTRACT

Evaluating Text Similarity : Between Semantics and Surface in Neural Representations

Measuring similarity between texts, whether based on meaning, characters, or phonetics, is essential in many applications. Neural networks, by transforming text into vectors, provide a practical method for assessing this similarity. However, the use of these representations is challenging because the criteria underlying this similarity are not clearly defined, oscillating between semantics and superficiality. Our study, based on controlled experiments, shows that surface differences have a more significant impact than semantic differences on measures of similarity between neural representations of words. These results raise questions about the nature of similarity measured by neural models and their ability to capture semantic nuance.

MOTS-CLÉS : Similarité textuelle, Analyse des Représentations Neuronales, Analyse Comparative de Modèles Pré-entraînés.

KEYWORDS : Text similarity, Neural Representations Analysis, Comparative Analysis of Pretrained Models.

1 Introduction

La mesure de la similarité entre deux textes est au cur de nombreuses applications, que la définition de la similarité soit basée sur le sens des textes (comme pour un moteur de recherche), la similarité des chaines de caractères (par exemple, pour permettre des correspondances floues) ou même la similarité phonétique (par exemple, dans les méthodes d'indexation basées sur la phonétique telles

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 85–96.

que `Soundex`).

Les réseaux neuronaux permettent de représenter le texte, qu'il s'agisse de mots, de phrases ou de paragraphes, sous forme de vecteurs, ce qui offre une méthode simple pour évaluer leur similarité. La mesure de la similarité entre deux vecteurs est en effet au cur de la fouille de données, et de nombreuses mesures de distance et de similarité aux propriétés bien établies ont été proposées dans la littérature (Duda *et al.*, 2001). Cette méthode est d'autant plus intéressante qu'il existe de nombreux modèles de langue pré-entraînés simplifiant la mise en uvre de celle-ci.

Un exemple concret d'utilisation de la similarité entre vecteurs contextuels, au cur aujourd'hui de nombreux travaux et développements, est le *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2020). Le RAG consiste à fournir à un giga-modèle (LLM) une base de documents pour enrichir sa production. Dans l'étape de récupération de documents, afin d'abonder la requête avec les passages pertinents, il est nécessaire de trouver les documents les plus similaires en fonction de leur produit scalaire avec le vecteur de requête dans un espace de vecteurs de haute dimension.

Toutefois, l'utilisation de représentations construites par les modèles de langue pour mesurer la similarité entre deux textes soulève un problème majeur. Bien qu'ils puissent faire de très bonnes prédictions dans de nombreuses tâches et en particulier générer des textes sémantiquement cohérents et syntaxiquement corrects, les réseaux neuronaux restent des modèles de type *boite noire* et les informations qu'ils encodent dans leur représentation ne sont pas clairement identifiées : s'il est facile d'utiliser des représentations neuronales pour mesurer la similarité, les critères sur lesquels cette similarité est fondée ne sont pas clairs.

Notre travail s'inscrit dans une longue série d'études visant à comprendre et à analyser les représentations apprises de manière auto-supervisée par les modèles de langue et en particulier les fameux *giga modèles* (Rogers *et al.*, 2020). Plus particulièrement, notre étude continue les travaux sur la pertinence des différentes mesures de similarité et leurs conditions d'utilisation sur les représentations tirées de ces modèles (Timkey & van Schijndel, 2021). Dans ce cadre, nous soulevons une nouvelle question : la similarité entre des représentations neuronales de texte est-elle fondée sur des critères sémantiques (comme on pourrait s'y attendre vues les bonnes performances des modèles utilisant ces représentations dans de nombreuses tâches) ou des critères de surface — deux alternatives qui ne s'excluent pas mutuellement.

Pour cela, suivant la proposition de (Isabelle *et al.*, 2017) d'évaluer les modèles sur des ensembles de données spécifiques construits autour de propriétés linguistiques clairement identifiées, nous proposons de mesurer la similarité entre une phrase et une version soigneusement modifiée de celle-ci, où les mots (noms, adjectifs et verbes) sont remplacés par des synonymes, des antonymes ou des paronymes (mot dont la prononciation est similaire à celle d'un autre mot, mais dont le sens est différent). Cette approche est détaillée dans la section 2. Grâce à ces expériences contrôlées, nous espérons pouvoir déterminer les critères sur lesquels se base la similarité mesurée entre les représentations neuronales des mots en établissant un lien entre la relation sémantique entre les mots échangés et la similarité entre les phrases.

Nos expériences, détaillées dans la section 3, montrent que les différences de surface ont un impact plus important sur les mesures de similarité que les différences de sémantique. Cette conclusion est particulièrement surprenante dans la mesure où de nombreuses études ont montré que les réseaux de neurones sont capables de construire des représentations abstraites des mots et des phrases (Li *et al.*, 2023). Nos expériences montrent également que la mesure de la similarité entre des représentations neuronales peut donner des résultats inattendus : des mots ayant des significations très différentes

peuvent être identifiés à tort comme étant très similaires.

2 Distinguer la similarité sémantique de la similarité de surface

Corpus Pour comprendre le fonctionnement interne de la similarité entre les représentations neuronales, nous avons créé un corpus de phrases en français et en anglais dans lesquelles un mot est remplacé par un autre mot dont la relation (sémantique) avec le mot d'origine est clairement identifiée [1]. Nous mesurons ensuite la similarité entre la phrase originale et la phrase modifiée en utilisant soit la similarité cosinus, soit la distance euclidienne.

Nous avons commencé, à l'aide de ressources en ligne, par compiler deux listes, une en anglais et une en français, de respectivement 373 et 354 mots associés à leur paronyme (des paires de mots, comme *irruption* et *éruption* dont la prononciation est similaire, mais pas le sens). Nous avons ensuite cherché dans différents corpus une phrase contenant un de ces mots.

Un mot et son paronyme ont des formes de surface et des prononciations très similaires, mais des significations très différentes. Il est courant, même pour un être humain, de confondre un mot avec son paronyme et d'utiliser l'un au lieu de l'autre. Si deux phrases dont la seule différence est l'utilisation d'un mot ou de son paronyme sont très similaires, la similarité repose davantage sur des informations de surface que sur la sémantique. On définit donc ici la similarité de surface comme un nombre de caractères en commun.

Pour chaque mot de notre liste, nous avons également cherché un antonyme et un synonyme, en nous assurant manuellement qu'ils pouvaient être utilisés à la place du mot d'origine dans la phrase (notamment en sélectionnant la forme correcte de ce dernier ou en sélectionnant parmi tous les synonymes ou antonymes d'un dictionnaire ceux qui pouvaient être utilisés dans le contexte). Nous avons ensuite répété la même expérience : en mesurant la similarité entre la phrase et la phrase dans laquelle le mot a été remplacé par un synonyme ou un antonyme, nous espérons pouvoir à la fois mieux comprendre quand une mesure de similarité appliquée à des représentations neuronales détecte que deux phrases sont similaires et déterminer la dynamique de ces mesures (notamment dans quel domaine la mesure varie). Bien que les sens distributionnels d'un mot et de son antonyme sont proches dans le cadre de la sémantique lexicale, nous nous plaçons du point de vue des utilisateurs finaux des LLMs, pour lesquels il est plus naturel de voir un antonyme comme particulièrement lointain du mot original.

Au final, notre corpus est constitué de 727 phrases, chacune apparaissant dans 5 versions différentes : la phrase originale et 4 versions modifiées dans lesquelles un mot a été successivement remplacé par un synonyme, un antonyme, un paronyme et un synonyme du paronyme. Le tableau 1 donne quelques exemples de phrases de notre corpus.

Test ABX Pour déterminer sur quel type d'informations repose la similarité entre deux représentations neuronales d'une phrase, nous utilisons un test ABX (Carlin *et al.*, 2011 ; Schatz *et al.*, 2013). Ce test s'appuie sur les représentations vectorielles construites par un modèle pré-entraîné de trois textes : deux textes, notés A et B sont proches sémantiquement et le troisième, noté X, a un sens différent. Le test ABX consiste simplement à vérifier si la similarité $s(A, B)$ est plus grande que $s(A, X)$.

1. Notre code et notre corpus seront diffusés lors de la publication.

①	originale	Le chirurgien procède à l'**ablation** du poumon.
	paronyme	Le chirurgien procède à l'**ablution** du poumon.
	synonyme	Le chirurgien procède à la **résection** du poumon.
	synonyme du paronyme	Le chirurgien procède à la **toilette** du poumon.
	antonyme	Le chirurgien procède à la **greffe** du poumon.
②	originale	Seasickness made him **retch** over the side of the boat.
	paronyme	Seasickness made him **wretch** over the side of the boat.
	synonyme	Seasickness made him **vomit** over the side of the boat.
	synonyme du paronyme	Seasickness made him **beggar** over the side of the boat.
	antonyme	Seasickness made him **swallow** over the side of the boat.

TABLE 1 – Exemples de phrases issues de nos corpus et les variantes considérées dans nos expériences.

Le score ABX correspond à la proportion de triplets pour lesquels $s(A, B) > s(A, X)$. Un score ABX proche de 50 % (ou inférieur) indique qu'en moyenne, la similarité entre A et X est plus grande que la similarité entre A et B, ce qui suggère que la similarité ne repose pas sur des informations sémantiques.

Modèles de langue Nous considérons [2] cinq modèles de langue multilingues pré-entrainés différents pour construire la représentation vectorielle des phrases de notre corpus : mBERT (Devlin *et al.*, 2019), le modèle ń sentence BERT ż (sBERT) de la phrase introduit dans (Reimers & Gurevych, 2019), le modèle d'OpenAI ADA, le modèle de Meta LLaMA-2 (Touvron *et al.*, 2023) et un modèle monolingue français, FlauBERT (Le *et al.*, 2020).

Différents *tokenizers* sont utilisés par ces modèles pour segmenter leur entrée : mBERT utilise WordPiece, sBERT utilise SentencePiece (Kudo & Richardson, 2018), tandis que ADA, LLaMA-2 et FlauBERT s'appuient sur BPE (Sennrich *et al.*, 2016).

Représentation de la phrase Si dans le cas de sBERT, la représentation construite par le réseau de neurones est directement celle de la phrase, ce n'est pas le cas pour les autres modèles qui construisent une représentation pour chaque token. Il existe plusieurs stratégies éprouvées pour construire la représentation d'une phrase à partir des représentations de ses tokens, la première étant d'utiliser la représentation du token spécial [CLS] comme représentation de la phrase, ce que nous faisons pour mBERT, ADA et FlauBERT. En revanche, l'approche choisie pour LLaMA consiste à calculer la représentation en moyennant [3] les *embeddings* de la sortie.

Certains de nos modèles ne peuvent être utilisés que par le moyen d'une API n'offrant accès qu'à certaines informations. Typiquement, nous ne pouvons accéder qu'aux représentations de la dernière couche d'ADA. Dans la mesure où ce modèle est utilisé dans de nombreuses applications, il nous a

2. Plus précisément, nous avons utilisé les modèles suivants via HuggingFace 🤗 (Wolf *et al.*, 2020) : bert-base-multilingual-cased pour mBERT, all-MiniLM-L6-v2 pour sBERT, Llama-2-7b-hf pour LLaMA-2 et flaubert_base_cased pour FlauBERT. Pour ADA, nous avons utilisé le modèle text-embedding-ada-002, via l'API fournie par OpenAI.

3. On présente en annexe, figure 5, différentes stratégies pour obtenir une représentation de la phrase. Si les résultats du test ABX sont sensiblement les mêmes dans tous les cas, les distributions non, les variations diminuant dans les cas avec normalisation (moyenne et standardisation), justifiant donc notre choix.

paru important de l'inclure dans notre comparaison. Par soucis de cohérence, nous avons décidé de considérer la dernière couche pour tous les modèles, même si plusieurs travaux récents (voir, par exemple, (Bordes *et al.*, 2023)) ont montré que, suivant les tâches, cette dernière couche ne permet pas toujours d'obtenir les meilleures représentations.

3 Résultats expérimentaux

La table 2 (resp. 3) reporte les résultats du test ABX pour l'anglais (resp. le français) pour les différentes combinaisons de substitutions décrites à la section 2. Ces tests sont effectués sur les similarités cosinus entre phrases. Les tables 5 et 6, en annexe, montrent que la différence obtenue en comparant des paires de phrases est minime. FlauBERT, bien qu'entrainé uniquement sur des données francophones, est également utilisé pour l'anglais comme expérience de contrôle, et les résultats sont décrits ici à titre indicatif.

Synonymes et antonymes Dans notre première expérience, nous comparons l'effet d'une substitution d'un mot par un synonyme (AB) avec une substitution par un antonyme (AX). Cette comparaison permet de s'assurer que le modèle différencie bien des phrases exprimant un sens contraire : plus le score ABX se rapproche de 100 %, plus les représentations de deux phrases ayant le même sens sont proches.

En ce qui concerne l'anglais, tous les modèles établissent une plus grande proximité dans le cas du synonyme que de l'antonyme. Néanmoins, si ADA le fait de manière quasi systématique, FlauBERT (pourtant francophone) est plus proche du hasard. En revanche, l'interprétation des résultats est moins évidente pour le français : mBERT et sBERT obtiennent un score ABX proche de 50 %, indiquant que les substitutions par un synonyme sont, en moyenne, aussi proches que celles par des antonymes. Par contre, ADA et LLaMA-2 restent eux capables de faire la distinction. De manière surprenante, ce n'est pas le cas pour un modèle monolingue français, FlauBERT. Il serait nécessaires de pouvoir mieux contrôler les données d'apprentissage (et notamment la proportion de données en français et de données en anglais) pour identifier les causes de cette différence de comportement.

Paronymes et synonymes Dans une seconde expérience, nous comparons les effets du remplacement d'un mot par un synonyme ou un paronyme. Dans notre corpus, les altérations paronymiques produisent souvent des phrases ayant des sens complètement différents, et parfois même des constructions n'ayant aucun sens, même si le paronyme a une forme de surface très proche du mot d'origine. À l'inverse, le sens de la phrase cible et de son homologue synonyme sont intrinsèquement similaires, voire identiques. La seconde colonne des tables 2 et 3 reporte les résultats obtenus en mesurant la similarité entre la phrase originale et la phrase comportant un paronyme, désignée par AX, et la similarité entre la phrase originale et la phrase comportant un synonyme, désignée par AB. Dans ce contexte, plus le score ABX est proche de 0%, plus le modèle a tendance à identifier le paronyme comme plus similaire au mot d'origine que le synonyme, suggérant que la similarité mesurée dépend fortement des informations lexicales.

On observe que pour le corpus anglais (table 2), mBERT, et dans une moindre mesure sBERT, capturent la proximité sémantique sans être influencés par la proximité des formes de surface. LLaMA-2 et ADA sont plus indécis, et FlauBERT, lui, favorise nettement la version paronymique.

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
Attendu pour une similarité sémantique	100	100	50
mBERT	67,7	64,2	69,7
sBERT	78,6	67,7	69,4
ADA	94,5	58,9	92,3
LLaMA-2	83,6	57,8	89,0
FlauBERT	52,2	32,3	73,2

TABLE 2 – Résultat des tests ABX (en %) sur le corpus en anglais (similarité cosinus). A désigne systématiquement le mot d'origine.

De manière surprenante, les résultats du corpus français présentés table 3 sont différents : la similarité mesurée pour tous les modèles semble impactée par la proximité de surface à des degrés divers. Chez certains, mBERT, sBERT et LLaMA-2, cette influence est particulièrement nette : la représentation de la phrase dans laquelle le nom est remplacé par son paronyme est presque systématiquement plus proche de la représentation de la phrase originale que lorsque le nom est remplacé par un synonyme indiquant clairement que la similarité mesurée repose essentiellement sur la similarité des formes de surface.

Paronymes et synonymes des paronymes Dans cette troisième expérience, les deux variations considérées, avec le paronyme AB ou un synonyme du paronyme AX, ont la même signification — la différence étant que la phrase avec le paronyme présente une plus grande similarité de surface. Si nos modèles ne considèrent que le sens des mots, le score ABX devrait tendre vers 50% aucune des substitutions n'ayant de raison d'être plus similaire que l'autre à la phrase originale.Si le résultat s'approche de 100%, le paronyme est préféré suggérant une influence de la surface.

Dans le cas de l'anglais comme du français, la version avec un paronyme est en moyenne toujours préférée, de manière assez nette, suggérant une préférence pour une forme avec une surface similaire, et donc une influence de cette dernière. L'écart de cette influence varie entre les expériences 2 et 3, en particulier pour LLaMA-2 et ADA : en l'absence de critères sémantiques, la proximité de surface reste capturée.

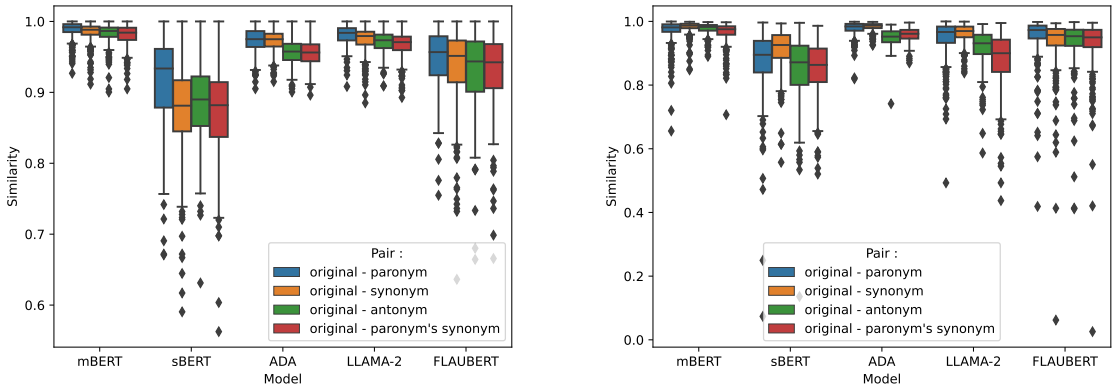
3.1 Distribution des distances

Pour compléter les résultats décrits dans la section précédente, nous avons représenté à la figure 1 (resp. figure 2) les distributions des similarités (resp. distances) entre les représentations des phrases d'origine et les phrases après substitution. Si les résultats varient d'une langue à l'autre, ces observations corroborent celles déjà décrites précédemment.

Pour le français, la similarité entre la phrase originale et la phrase comportant une substitution par un paronyme est toujours au dessus des autres substitutions, ce qui corrobore l'idée d'une influence de la surface, notamment car cette similarité est plus petite que celle reposant sur la substitution par un synonyme. Avec ADA et LLaMA-2, la similarité entre la phrase d'origine et la phrase avec un

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
Attendu pour une similarité sémantique	100	100	50
mBERT	58,9	32,4	74,7
sBERT	56,0	21,5	80,8
ADA	83,1	49,4	86,2
LLaMA-2	67,6	33,7	89,4
FlauBERT	54,6	43,6	60,3

TABLE 3 – Résultat des tests ABX (en %) sur le corpus français (similarité cosinus). A désigne systématiquement le mot d'origine.



(a) Corpus français

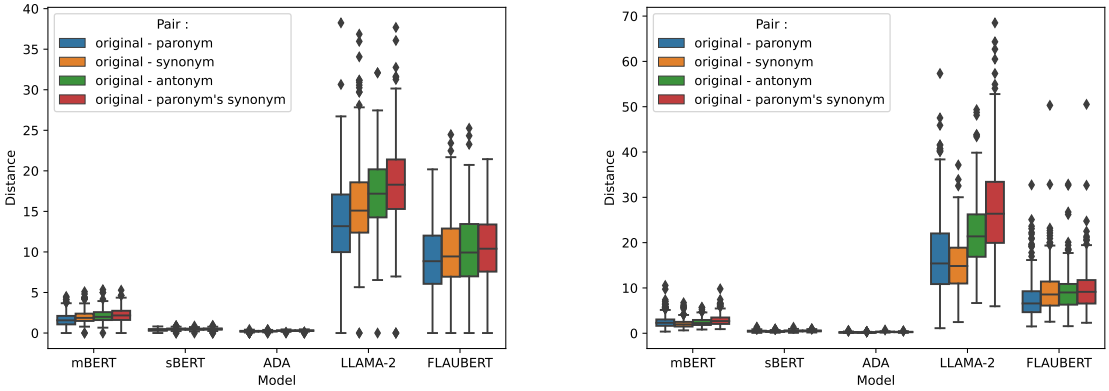
(b) Corpus anglais

FIGURE 1 – Distribution de mesures de similarité cosinus entre les phrases originales et après une substitution.

antonyme est systématiquement plus faible que lorsque la similarité est mesurée avec n'importe quel autre type de substitution. Mais sBERT considère parfois qu'une substitution par un synonyme donne une phrase moins similaire à la phrase originale que lorsque le mot est remplacé par un antonyme, une observation qui nous incite à la prudence dans l'interprétation des similarités entre les représentations neuronales.

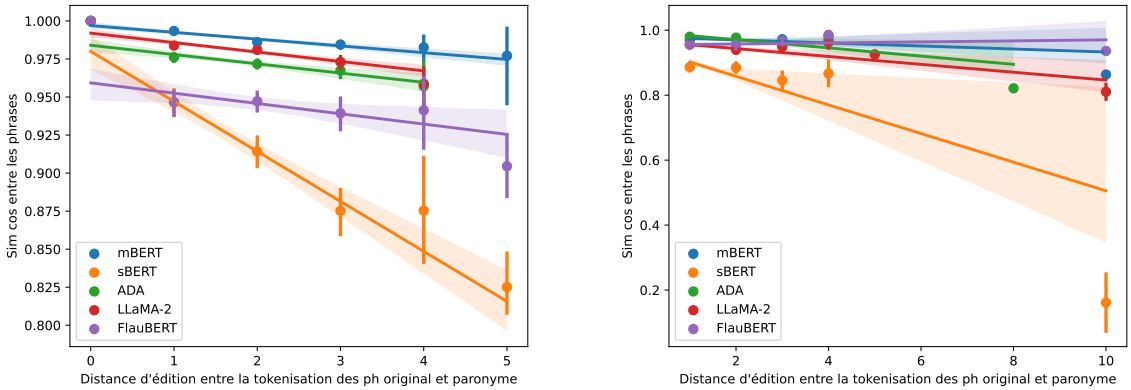
Pour l'anglais, la distribution des similarités est, pour tous les modèles, meilleure dans le cas d'une substitution avec un synonyme. Les antonymes sont eux correctement discriminés. Les modèles semblent dans ce cas bien hiérarchiser les phrases en fonction de leur proximité sémantique et non de surface. Néanmoins, les paronymes sont au dessus des synonymes des paronymes et non à égalité, n'excluant pas tout à fait l'influence d'une proximité de surface.

Ces observations montrent également que les similarités mesurées sont toujours très fortes et que les distances observées varient toujours dans un domaine très restreint. La valeur absolue de la distance ou de la similarité entre deux représentations est donc difficile à interpréter et devrait toujours être considérée avec précaution, par exemple si l'on souhaite définir un seuil pour filtrer des éléments à partir de celle-ci.



(a) Corpus français (b) Corpus anglais

FIGURE 2 – Distribution de mesures de distance euclidienne entre les phrases originales et après une substitution.



(a) Corpus français (b) Corpus anglais

FIGURE 3 – Relation entre la distance d'édition entre la tokenisation d'une phrase et de la version avec paronyme, et leur similarité cosinus.

3.2 Impact de la segmentation en sous-mots

Les résultats présentés dans la section précédente peuvent être faussés par la tokenisation : si notre analyse est effectuée au niveau du mot, le modèle ne manipule que les unités sous-lexicales et il est possible que la segmentation en unités sous-lexicales impacte la similarité et la magnitude de l'influence de la surface. Par exemple, si *acceptation* et son paronyme *acception* ont une distance d'édition de deux, leur segmentation en unités sous-lexicales (*accept##ation* et *accept##ion* respectivement) ne diffère que d'un seul token. La distance d'édition calculée au niveau du token n'est donc que de 1, et la tokenisation des sous-mots crée également un token supplémentaire identique entre les deux phrases.

Pour analyser cette possibilité, nous représentons dans la figure 3 l'évolution de la similarité entre la phrase originale et la phrase avec une substitution par un paronyme en fonction de la distance d'édition calculée au niveau des unités sous-lexicales. Cette figure montre clairement que le nombre

	Français	Anglais
mBERT	-0,36	-0.12
sBERT	-0,48	-0.43
ADA	-0,28	-0.43
LLaMA-2	-0,36	-0.19
FlauBERT	-0,15	0.03

TABLE 4 – Coefficients de Pearson entre la distance d'édition de la tokenisation d'une phrase et de la version avec paronyme, et leur similarité cosinus.

d'unités sous-lexicales a un impact sur la similarité : la similarité est d'autant plus faible que le nombre d'unité sous-lexicales différentes est important, mettant en évidence l'impact des informations de surface. Une mesure directe de la corrélation entre ces deux grandeurs (table 4) montre toutefois que cet effet est faible.

4 Conclusion

Dans cet article, nous avons décrit plusieurs expériences utilisant des mesures de similarité pour déterminer et mesurer la présence d'une interférence entre la surface (nombre de caractères en commun) et la représentation vectorielle de différents LLMs. Nos résultats montrent que le calcul d'une similarité entre deux représentations neuronales d'un texte repose essentiellement sur des informations de surface. Nos expériences montrent également que, quelle que soit la métrique considérée, les représentations neuronales détectent de très fortes similarités même entre des phrases de sens opposés, un résultat surprenant alors que de nombreux travaux ont mis en évidence la capacité de celles-ci à capturer le sens d'une phrase ou à générer un texte sémantiquement cohérent. Cette observation a des implications pratiques importantes, la détection de similitudes entre des textes étant au cur de nombreuses applications.

Ces travaux préliminaires peuvent être enrichis sur de nombreux aspects. Outre l'ajout de modèles et de langues pour consolider nos résultats et mesurer d'éventuels écarts, il pourrait également être intéressant d'étudier d'autres types de plongements, par exemple des *embeddings* de position, pour évaluer le rôle joué par la position du mot altéré. De même, la comparaison entre les couples de mots et leurs tokenisations pourrait être plus parlante avec un *embedding* statique, par exemple de type `fasttext`.

Remerciements

Nous remercions les relecteurs anonymes pour leur temps et précieux conseils, et pour avoir contribué à l'amélioration de ce papier. Ce travail a été financé par le projet DIAGNOSTIC soutenu par l'Agence de l'Innovation de Défense (subvention nř 2022 65 007).

Références

BORDES F., BALESTRIERO R., GARRIDO Q., BARDES A. & VINCENT P. (2023). Guillotine regularization : Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*.

CARLIN M. A., THOMAS S., JANSEN A. & HERMANSKY H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, p. 4171–4186. DOI : 10.18653/v1/N19-1423.

DUDA R. O., HART P. E. & STORK D. G. (2001). *Pattern Classification*. New York : Wiley, 2 édition.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486–2496, Copenhagen, Denmark. DOI : 10.18653/v1/D17-1263.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : des modèles de langue contextualisés pré-entraînés pour le français. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278 : ATALA.

LEWIS P. S. H., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, **abs/2005.11401**.

LI B., WISNIEWSKI G. & CRABBÉ B. (2023). Assessing the capacity of transformer to abstract syntactic representations : A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, **11**, 18–33. DOI : 10.1162/tacl_a_00531.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *EMNLP*, p. 3982–3992. DOI : 10.18653/v1/D19-1410.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2020). A primer in BERTology : What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866. DOI : 10.1162/tacl_a_00349.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association*, p. 1–5.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

TIMKEY W. & VAN SCHIJNDEL M. (2021). All bark and no bite : Rogue dimensions in transformer language models obscure representational quality. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4527–4546, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.372.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

A Résultats supplémentaires

Les figures 5 et 6 présentent les médianes des résultats obtenus en performant les différents tests ABX, tandis que la figure 5 montre la différence dans les distributions de distance euclidienne obtenues avec différentes stratégies de représentation de la phrase à partir des représentations des tokens : prendre le dernier, faire la moyenne ou bien effectuer une standardisation de la représentation du dernier token.

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
mBERT	0.003873	0.003484	0.005049
sBERT	0.042417	0.024965	0.022611
ADA	0.030399	0.002901	0.016763
LLaMA-2	0.029717	0.004966	0.053019
FlauBERT	0.001616	-0.011391	0.015375

TABLE 5 – Test ABX sur corpus anglais (similarité cosinus) : médiane de la différence $AB - AX$

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
mBERT	0.001941	-0.003248	0.005874
sBERT	0.005882	-0.038234	0.040888
ADA	0.017039	-0.000001	0.015579
LLaMA-2	0.004425	-0.004446	0.011270
FlauBERT	0.002520	-0.006072	0.009486

TABLE 6 – Test ABX sur corpus français (similarité cosinus) : médiane de la différence $AB - AX$

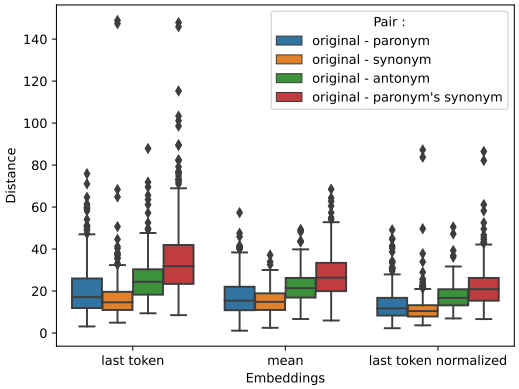


FIGURE 4 – Corpus anglais

FIGURE 5 – Comparaison des distances euclidienne en fonction de la stratégie de création du plongement de la phrase pour LLaMA-2