Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Plongement des constituants pour la représentation sémantique des phrases

Eve Sauvage†[1, 2] Iskandar Boucharenc†[1] Thomas Gerald[1] Julien Tourille[2]
Sabrina Campano[2] Cyril Grouin[1] Sophie Rosset[1]
(1) LISN, bât.507 Rue du Belvédère, 91405 Orsay, France
(2) EDF Lab, 7 boulevard Gaspard Monge, 91120 Palaiseau, France
eve.sauvage@edf.fr, iskandar.boucharenc@lisn.fr

† Les auteur.ice.s ont contribué de manière équivalente

RÉSUMÉ

Les méthodes d'apprentissage profond en traitement automatique des langues reposent souvent sur une segmentation des textes en tokens avant leur vectorisation. Cette segmentation produit des sous-unités lexicales offrant une grande flexibilité. Toutefois, la réutilisation de tokens identiques dans des mots de sens différents peut favoriser des représentations basées sur la forme plutôt que sur la sémantique. Ce décalage entre la forme de surface et le sens peut induire des effets indésirables dans le traitement de la langue. Afin de limiter l'influence de la forme sur la sémantique des représentations vectorielles, nous proposons une représentation intermédiaire plus compacte et plus fidèle au sens des mots.

ABSTRACT

Towards a better representation of meaning in the tokenization of language models

Deep learning methods in Natural Language Processing make extensive use of tokenization before building a vector representation. Tokenization offers great flexibility by producing subword units. However, using the same units in diverse contexts induces representations that tend to favor surface form over semantics. This phenomenon unbalances the processing of textual data. To reduce the impact of surface form on vector semantics, we propose an intermediate representation that is more compact and semantically faithful.

MOTS-CLÉS : Modèle pré-entraînés, Apprentissage profond, Représentation latente, Analyse en constituants, Tokénisation.

KEYWORDS: Deep learning, Pre-Trained Models, Latent Representation, Constituent Analysis, Tokenization.

1 Introduction

Les méthodes d'apprentissage profond pour le Traitement Automatique des Langues (TAL) reposent sur la vectorisation du texte. La méthode la plus couramment utilisée consiste à entraîner des algorithmes de segmentation statistique (*tokenizers*) tel que SentencePiece (Sennrich *et al.*, 2016) afin de découper le texte en sous-unités exploitables par les modèles. Ce processus aboutit à une segmentation à l'échelle du sous-mot produisant des *tokens* qui présentent l'avantage d'être assemblables et d'éviter des erreurs liées aux mots hors-vocabulaire.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 614–628.

Plusieurs travaux récents (Tao *et al.*, 2024; Huang *et al.*, 2025) suggèrent que la taille du vocabulaire d'un modèle suit une loi d'échelle. Leurs expériences indiquent que l'augmentation de la taille du vocabulaire pourrait être bénéfique aux performances des modèles ayant un grand nombre de paramètres. Notamment, Gee *et al.* (2023) fusionnent les tokens fréquemment vus ensemble et ajoutent ces fusions au vocabulaire, compressant les séquences.

Tytgat *et al.* (2024) montrent que les modèles profonds ont tendance à privilégier la forme de surface du lexique plutôt que la sémantique dans le calcul de similarité textuelle. Cette préférence pour la forme de surface peut entraîner une baisse de performance dans des tâches nécessitant une distinction fine sur le sens des phrases, comme c'est le cas dans les tâches de similarité sémantique (*Semantic Textual Similarity*, STS) ou la détection de paraphrases (Muennighoff *et al.*, 2023). Pour pallier ce problème, nous proposons d'améliorer l'information sémantique contenue dans les plongements lexicaux. Dans cet article, nous explorons trois hypothèses : (i) la prédominance de la forme de surface sur la sémantique est liée à une décorrélation entre segmentation statistique et information sémantique ; (ii) une segmentation tenant compte de la sémantique pourrait réduire la dépendance des représentations vectorielles à la forme de surface ; enfin, (iii) la fusion des tokens appartenant à un même groupe sémantique pourrait aider à la désambiguïsation des représentations vectorielles.

La suite de cet article est organisée comme suit : la section 2 présente les travaux liés à notre démarche et nos motivations. La section 3 détaille les méthodes employées et le protocole expérimental. Enfin, nous exposons nos résultats et analyses en section 4, avant de conclure en section 5.

2 État de l'art

En Traitement Automatique des Langues (TAL), les méthodes issues de l'apprentissage profond (AP) dominent aujourd'hui la plupart des tâches. Ces méthodes nécessitent de représenter le texte sous forme de vecteurs. Les premières approches de vectorisation reposaient sur un encodage 1 parmi n (*one-hot encoding*), une méthode simple qui souffrait de deux limitations majeures : d'une part elle produisait des vecteurs de très grande dimension et d'autre part, elle ne tenait pas compte du contexte d'apparition des mots.

Avec l'essor des réseaux de neurones artificiels, des solutions plus efficaces ont émergé, permettant d'intégrer la notion de contexte dans les représentations vectorielles à travers l'apprentissage non-supervisé (Mikolov *et al.*, 2013; Pennington *et al.*, 2014). Ces approches exploitent des tâches telles que la prédiction de mots masqués (*e.g.* texte à trou) pour extraire des représentations contextualisées. Cependant, si ces méthodes améliorent la modélisation du contexte, elles peinent encore à capturer efficacement la sémantique des mots rares ou hors-vocabulaires.

Pour pallier cette limitation, Sennrich *et al.* (2016) proposent une segmentation en sous-mots basée sur des règles statistiques, une technique permettant d'encoder des mots absents du vocabulaire d'entraînement, bien que d'une manière imparfaite. Nous désignerons ces sous-mots « tokens » dans la suite de cet article. Toutefois, cette segmentation repose sur des critères purement statistiques, sans fondement linguistique, ce qui introduit un biais sur la forme de surface des textes (Tytgat *et al.*, 2024).

Aujourd'hui les modèles d'AP les plus performants s'appuient sur l'architecture *Transformer* (Vaswani *et al.*, 2017). Bien que hautement parallélisables, ces modèles présentent une complexité quadratique en fonction de la longueur des séquences traitées. Dès lors, la segmentation en sous-mots pose un problème supplémentaire en termes de consommation de ressources, aussi bien en calcul

qu'en mémoire (Dettmers *et al.*, 2023). Bien que ces limites soient connues, la littérature s'est principalement focalisée sur l'optimisation des modèles (Dao *et al.*, 2022; Dettmers *et al.*, 2023). Néanmoins, certaines études ont exploré des stratégies de compression des entrées afin d'améliorer l'efficacité des modèles. Par exemple, en vision par ordinateur (VO), l'approche de la fusion des tokens visuels (*TokenMerging* (ToME)) (Bolya *et al.*, 2023) a montré son efficacité pour accélérer le traitement. En TAL, Gee *et al.* (2023) ont étudié l'accroissement du vocabulaire des modèles de langue par adjonction d'expressions polylexicales, en se basant sur la fréquence d'apparition des n-grammes. Notre travail s'inscrit dans cette démarche, mais se distingue par l'utilisation d'un critère de fusion non statistique. En effet, les approches basées sur un critère fréquentiel peuvent entraîner des pertes de performances sur certaines tâches (Kumar & Thawani, 2022). Cependant, nous conservons l'approche qui consiste à augmenter la taille du vocabulaire, même virtuellement, pour améliorer la représentation des textes. Une telle augmentation ne semble pas poser de contraintes excessives, ni en termes de performance, ni en termes de coût computationnel. En effet, Tao *et al.* (2024) rappellent que la taille optimale du vocabulaire est souvent sous-estimée, tandis que Huang *et al.* (2025) démontrent que l'adaptation du vocabulaire peut être réalisée de manière suffisamment flexible pour découpler le vocabulaire d'entrée de celui de sortie.

3 Protocole expérimental

L'objectif de cette étude est d'examiner l'impact des modifications de représentation des tokens sur les modèles encodeurs. Pour cela, il est nécessaire d'adapter les poids des modèles via un processus d'entraînement spécifique. Cette section détaille les modèles et les jeux de données sélectionnés (3.1), les critères et les motivations guidant la modification des représentations (3.2) ainsi que les choix méthodologiques adoptés pour l'entraînement des modèles (3.3).

3.1 Tâches, corpus et modèles

Tâches Nous évaluons la pertinence des représentations obtenues en nous appuyant sur la tâche de similarité sémantique, souvent calculée via la similarité cosinus, dans la continuité des travaux de Tytgat *et al.* (2024). L'intérêt principal de cette tâche dans notre cadre expérimental est de vérifier que la représentation d'un synonyme d'un terme est plus proche de celle du terme en question que celle de son homonyme. Pour cela, nous utilisons le corpus proposé par Tytgat *et al.* (2024), que nous désignerons sous le nom de corpus SYNPAR.

Corpus d'évaluation Le corpus SYNPAR issu de Tytgat *et al.* (2024) est un corpus d'ensembles de phrases dont l'un des mots est décliné en 4 versions : original, synonyme, paronyme et antonyme. La paire original-synonyme est sémantiquement proche tandis que la paire original-paronyme est graphiquement proche. Le corpus est constitué de 372 ensembles de phrases pour l'anglais.

original	paronyme	synonyme
Please **accept** this gift as a sign of my gratitude.	Please **except** this gift as a sign of my gratitude.	Please **receive** this gift as a sign of my gratitude.

TABLE 1 – Exemple issu du dataset SYNPAR

Corpus d'entraînement Nous entraînons nos modèles sur deux corpus distincts du corpus SYNPAR, afin d'éviter toute contamination des données d'entraînement.

1. Pré-entraînement sur un sous-ensemble, tiré uniformément, de Wikipédia EN [1] en raison de la richesse et de la diversité des sujets abordées, faisant de ce corpus une base adéquate pour l'entraînement initial des modèles.
2. Affinage sur STSBenchmark [2] issu de la suite d'évaluation MTEB (Muennighoff *et al.*, 2023) qui rassemble de nombreuses suite d'évaluations préalables et suivantes sur la similarité de phrases. Le corpus STSBenchmark offre une large quantité de paires de phrases, annotées avec des scores de similarité, ce qui permet un ajustement fin des représentations.

Modèles Nous utilisons les modèles étudiés par Tytgat *et al.* (2024), à savoir `BERT` (Devlin *et al.*, 2019) [3] et `all-miniLM-L6-v2` [4]. Le choix de ces modèles repose sur plusieurs critères :

— leur légèreté, qui facilite l'adaptation et la reproductibilité des expériences
— la disponibilité et la transparence de leurs données d'entraînement
— leur entraînement préalable sur des quantités de données restreintes, limitant ainsi l'impact des biais acquis

Par ailleurs, ces deux modèles sont spécifiquement adaptés aux tâches de similarité de textes et offrent une base optimale pour les expérimentations.

Langue Nous réalisons les expériences exclusivement sur des corpus anglophones pour tester notre approche. Une généralisation des résultats sur des corpus multilingues est envisagée en travaux futurs. Le choix de l'anglais comme langue de travail est dû à la disponibilité des ressources dans cette langue. Une telle disponibilité facilite la mise en place d'expériences préliminaires permettant de vérifier l'efficacité de notre approche.

3.2 Modifications de la représentation des tokens

Afin de tester nos hypothèses, nous proposons une approche de fusion des représentations de certains tokens en nous basant sur une analyse sémantique. Cette section présente la méthode employée, les critères retenus et les détails d'implémentation.

Réduction de l'impact de la forme de surface La motivation principale derrière la fusion des tokens réside dans la prise en compte prépondérante de la forme de surface dans l'encodage des textes. Un mot ou un groupe de mots appartenant à une unité sémantique cohérente peut être segmenté en plusieurs tokens, tandis que des mots de sens distincts peuvent partager des tokens identiques (cf. figure 1). Cette proximité artificielle dans l'espace des représentations peut conduire au rapprochement de phrases sémantiquement éloignées.

Pour remédier à ce problème, nous proposons une représentation intermédiaire qui atténue l'impact de la forme de surface en fusionnant les tokens appartenant à une même unité sémantique. La méthode retenue par Gee *et al.* (2023) identifie les tokens les plus fréquemment observés ensemble et les fusionnent en les ajoutant au vocabulaire afin de compresser l'entrée. Par ailleurs, Tao *et al.* (2024) montrent qu'une augmentation de la taille du vocabulaire peut améliorer l'entraînement et les

1. https://huggingface.co/datasets/legacy-datasets/wikipedia
2. https://huggingface.co/datasets/mteb/stsbenchmark-sts
3. https://huggingface.co/google-bert/bert-base-uncased
4. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

$Seg_{synonyme}$: 'certain', 'exercises', 'can', 'help', 'to', 'stimulate', 'the', 'muscles', '.'
$Seg_{originale}$: 'certain', 'exercises', 'can', 'help', 'to', 'inner', '##vate', 'the', 'muscles', '.'
$Seg_{paronyme}$: 'certain', 'exercises', 'can', 'help', 'to', 'en', '##er', '##vate', 'the', 'muscles', '.'

FIGURE 1 – Exemple de tokénisation produite par le modèle `all-miniLM-l6-v2` avec une proximité forte entre la tokénisation de la phrase originale (2e ligne) et celle contenant le paronyme (3e ligne). Cet exemple est issu du corpus SYNPAR

performances des modèles. Ainsi, en intégrant une étape de fusion par moyenne dans le modèle, nous cherchons à évaluer son impact sur la performance et la compression des données en entrée.

Diminution de la taille des entrées La fusion des tokens dans un texte permet de réduire le nombre de tokens à traiter par le modèle et, par conséquent, de diminuer la longueur des séquences. Or, les modèles Transformers ayant une complexité quadratique en fonction de la taille des séquences, cette diminution représente un avantage computationnel significatif.

Critères de fusion Compte tenu de l'importance de la représentation sémantique dans les modèles de langue et du prisme sémantique adopté par Mikolov *et al.* (2013) ou Peters *et al.* (2018) pour expliquer les plongements lexicaux, il semble pertinent d'adopter un critère de fusion basé sur la sémantique. Un découpage en lexèmes, les plus petites unités de sens, pourrait sembler une approche logique. Toutefois, il n'existe ni corpus de référence ni modèles de segmentation en lexèmes facilement accessibles. De plus, une telle approche ne permettrait pas de prendre en compte des expressions polylexicales et nécessiterait une modification en profondeur des représentations utilisées par les modèles.

Nous proposons donc une approche basée sur l'analyse en constituants. Cette analyse permet d'extraire les syntagmes, les léxèmes, des différentes phrases, lesquels sont souvent considérés comme des unités sémantiques indépendantes autant que des unités syntaxiques (Dany Amiot, 2001; Camara, 2019). Ils sont identifiables grâce à des tests de constituance mettant en évidence leur autonomie structurelle (topicalisation, substitution, dislocation, clivage, etc.). L'analyse en constituant permet de récupérer ces unités sémantiques indépendante grâce à l'ajout d'information par rapport à l'analyse en dépendance. En effet, il manque à l'analyse en dépendance les informations d'appartenance à des catégories grammaticales comme théorisé par (Chomsky, 1956). Contrairement à une approche se limitant uniquement aux entités nommées, la segmentation en constituants offre une couverture plus large sur l'ensemble du texte. Bien que cette méthode permette d'identifier un grand nombre de plongements à fusionner, elle entraîne une modification substantielle de la distribution des vecteurs, ce qui pourrait être préjudiciable à la performance du modèle.

Pour l'analyse en constituants, nous utilisons le modèle fourni par Stanza (Qi *et al.*, 2020), l'un des rares analyseurs accessibles pour ce type de traitement. Nous choisissons de fonder la fusion sur les groupes nominaux (NP), prépositionnels (PP) et verbaux (VP) les plus proches des feuilles de l'arbre syntaxique. Cette sélection permet d'obtenir des constituants de taille significative, réduisant ainsi fortement la longueur des séquences en entrée des modèles. Par souci de simplification, nous excluons les groupes adverbiaux et adjectivaux bien que ceux-ci fassent partie des cinq types de syntagmes généralement identifiés. En effet, ces groupes sont souvent inclus dans d'autres syntagmes ce qui permet de les prendre en compte sans les traiter séparément.

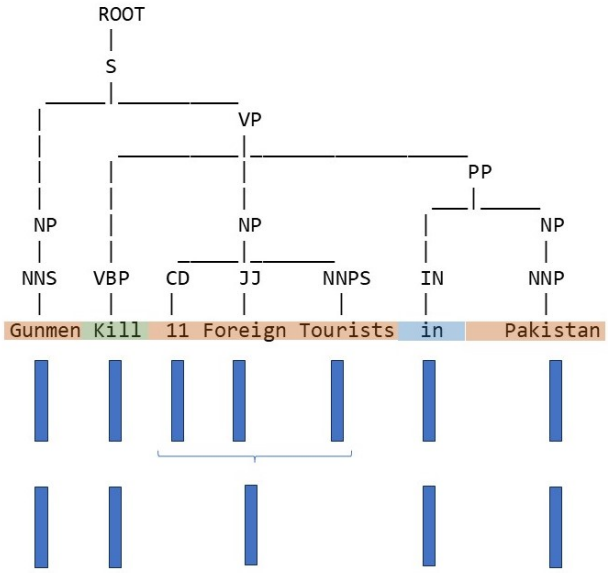


FIGURE 2 – Exemple de fusion des tokens sur un textes du corpus *STSbenchmark*. Les barres bleues représentent les représentations vectorielles tandis que les différents syntagmes sont représentés grâce aux groupes de couleurs

3.3 Entraînements des modèles

Pour permettre aux modèles de s'adapter à la représentation fusionnée, nous modifions leurs poids. Notre approche repose sur deux étapes : (i) un pré-entraînement basé sur une tâche analogue au MLM (*Masked Language Modeling*) et (ii) un affinage sur une tâche de similarité de phrases, comme résumé dans le tableau 2.

Modèle	nom du modèle	fusion	pré-entraîné	affiné
`all-miniLM-L6-v2`	miniLM-zs	✗ / ✓	✗	✗
	miniLM-pt	✗	✓	✗
	miniLM-ft	✗ / ✓	✗	✓
	miniLM-pt-ft	✗ / ✓	✓	✓
`bert-base-uncased`	bert-zs	✗ / ✓	✗	✗
	bert-pt	✗	✓	✗
	bert-ft	✗ / ✓	✗	✓
	bert-pt-ft	✗ / ✓	✓	✓

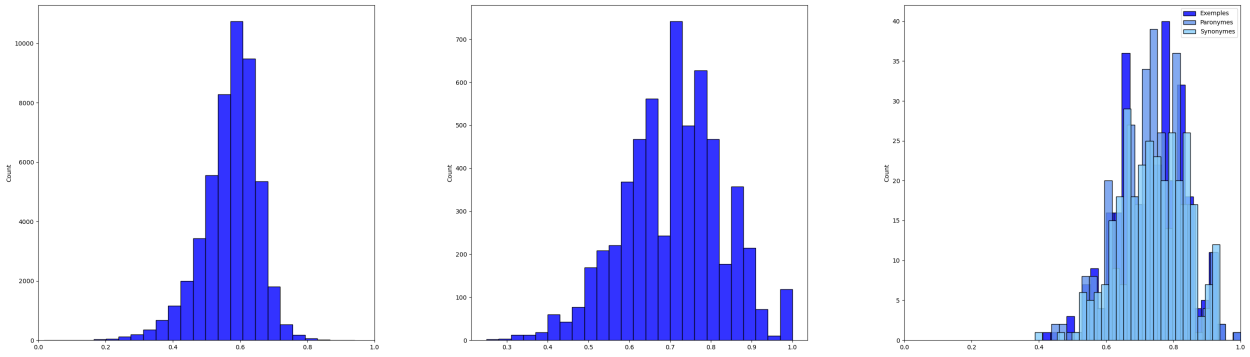
TABLE 2 – Récapitulatif des expériences réalisées. Le pré-entraînenment (pt) se fait sur 50 000 textes de Wikipédia, l'affinage (ft) sur le jeu *STSBenchmark*. L'abréviation `zs`, désigne l'évaluation *zero-shot*. Nous utiliserons ces appellations dans le reste de l'article pour plus de clarté

3.3.1 Pré-entraînement

La fusion des représentations entraîne mécaniquement une modification de la distribution de la taille des vecteurs de représentation (cf. Figure 3), réduisant en moyenne la longueur des séquences de $42,83\%$ sur le corpus Wikipédia. Afin d'atténuer l'impact de ce changement, une approche courante (appliquée par exemple lors d'un changement de domaine) consiste à poursuivre le pré-entraînement du modèle. Étant donné que les modèles utilisés ont été initialement pré-entraînés sur une tâche de MLM, nous adaptons cette tâche à notre représentation. Cette adaptation est nécessaire car les représentations fusionnées ne correspondent plus directement à des tokens du vocabulaire rendant inapplicable une résolution classique de la tâche de pré-entraînement par entropie croisée sur les tokens. Nous optons donc pour une approche continue de cette tâche. L'objectif de ce pré-entraînement est de permettre au modèle d'acquérir une compréhension sémantique des groupes de tokens fusionnés. Plutôt que d'optimiser une entropie croisée entre les probabilités de prédiction des tokens masqués, nous entraînons le modèle à reconstruire le vecteur issu de la fusion avant son passage dans le réseau : les représentations sont d'abord fusionnées puis les groupes obtenus sont masqués pour être reconstruits par le modèle. La représentation produite par la couche finale est traitée par un perceptron, puis comparée au vecteur d'entrée correspondant. La fonction de perte retenue est la distance L^2 entre la sortie du perceptron correspondant à une représentation masquée et la représentation fusionnée comme présenté ci-dessous.

$$\mathcal{L}(X, \tilde{X}) = \mathrm{MSE}(X, \mathcal{F}(X))$$

avec $\mathcal{F} = p \circ \mathcal{M}$ ou p est un perceptron dont les dimensions d'entrées et sorties sont la dimension du modèle $\mathcal{M}$, $\tilde{X} = \mathcal{F}(X)$. Le modèle et le perceptron sont entraînés sur un sous-ensemble de Wikipédia sélectionné aléatoirement et représentant 14M de tokens. Nous sauvegardons le modèle produisant la meilleure perte moyenne sur 3200 textes.



(a) corpus Wikipédia (b) corpus STSBenchmark (c) corpus SynPar

FIGURE 3 – Taille des textes fusionnés par rapport aux textes originaux. On lit le ratio de taille sur l'axe des abscisses (pourcentage) et le nombre de séquences ayant ce ratio sur l'axe des ordonnées

3.3.2 Adaptation

Parallèlement au pré-entraînement des modèles pour intégrer la fusion des tokens, nous proposons d'adapter ces modèles à la tâche de similarité de phrases en prenant en compte cette fusion. Contrairement à la tâche de tokens masqués, la perte du modèle peut être calculée de manière similaire à

l'entraînement des réseaux siamois proposés par Reimers & Gurevych (2019). Plus précisément, nous évaluons la similarité cosinus entre la représentation du `pooler` appliquée à la paire de phrases dont les représentations ont été fusionnées puis nous calculons une perte MSE (erreur quadratique moyenne) entre la similarité attendue et celle prédite par notre approche. Nous utilisons comme critère d'arrêt la corrélation de Spearman sur l'ensemble de validation.

3.4 Evaluation

Nous évaluons les performances de nos modèles sur le corpus SYNPAR, en nous intéressant plus particulièrement aux paires original-synonyme et original-paronyme. Dans les paires original-synonyme, un mot de la phrase originale est remplacé par un synonyme dans la phrase synonyme. Cette substitution entraîne une modification de la forme de surface tout en conservant (ou avec un changement minimal) la sémantique de la phrase. Dans les paires original-paronyme, un changement est appliqué au même mot original, mais sans modification notable de la forme de surface (ou avec un changement minimal), tout en entraînant une modification majeure du sens de la phrase.

Test ABX Afin de comparer nos résultats à ceux de Tytgat *et al.* 2024, nous utilisons le test ABX décrit par Carlin *et al.* (2011) et Schartz *et al.* (2013). Ce test mesure, entre 0 et 100, la similarité entre un élément A et un élément B ayant des différences minimes avec lui, en comparaison avec un élément X, qui présente des différences majeures (équation 1).

$$100 \times \frac{\text{Card}(\cos(A, B) \geq \cos(A, X))}{\#\text{Exemples}} \tag{1}$$

Corrélation de Spearman Comme dans l'évaluation du Benchmark MTEB contenant *STSBenchmark*, nous utilisons la corrélation de Spearman (1904) pour l'évaluation sur ce corpus. Nous excluons la corrélation de Pearson, conformément aux recommandations de Reimers *et al.* (2016).

4 Résultats & Analyses

4.1 Résultats sur STSBENCHMARK

Pour évaluer les résultats sur le corpus d'entraînement, conformément à la littérature, nous calculons la corrélation de Spearman sur les différents modèles dont les résultats sont reportés dans le tableau 3 en reprenant les notations définies dans le tableau 2. Le tableau est divisé en deux parties, une pour chaque modèle de base. Les deux premières lignes représentent à chaque fois le modèle sans fusion des constituants. La deuxième ligne correspond à un affinage sans fusion à titre de contrôle. Les trois lignes suivantes peuvent se lire comme une étude d'ablation en rajoutant successivement une phase d'affinage (ft) et de pré-entraînement et d'affinage (pt-ft).

Le maintien du cadre zéro-shot n'est plus possible à cause du changement de distribution des représentations comme le montre la baisse importante de performance. Dans le cas du modèle BERT, la fusion donne de bons résultats, y compris face à l'affinage du modèle sans fusion. La hausse de la métrique de corrélation avec le taux de compression de 29,91% valide nos hypothèses pour le modèle BERT. Cependant dans le cas de MiniLM, toute tentative de fusion s'accompagne d'une baisse des performances. Bien que le pré-entraînement relativise cette baisse et que le taux de compression moyen reste intéressant, il est plus difficile de se prononcer sur ce modèle. Notons que les objectifs visés par les deux modèles diffèrent légèrement. La version `all-miniLM-L6-v2` est conçue pour

	fusion	spearmanr validation	spearmanr test	Δ (test)
bert-zs	✗	59,32	47,29	–
bert-ft	✗	77,35	69,74	+22,45
bert-zs	✓	48,46	40,13	-7,16
bert-ft	✓	83,96	79,05	+31,76
bert-pt-ft	✓	**84,21**	**79,65**	**+32,36**
miniLM-zs	✗	**86,72**	82,03	–
miniLM-ft	✗	84,34	**84,40**	**+2,37**
miniLM-zs	✓	59,24	42,81	-39,22
miniLM-ft	✓	85,05	78,10	-3,93
miniLM-pt-ft	✓	86,08	81,03	-1,0

TABLE 3 – Corrélation de Spearman (spearmanr) pour les différents modèles et différence de performance avec l'évaluation zéro-shot (Δ). Les résultats sont grisés si la nature de la représentation (fusionnée ou non) change entre le pré-entraînement et l'évaluation (*e.g.* modèle original sur représentation fusionnée ou inversement). *cf.* tableau 2 pour l'explication du nom des modèles.

générer des plongements généralistes performants, ce qui explique la difficulté à rattraper la baisse de performance. BERT est, en revanche, destiné à être affiné sur des tâches aval (Devlin *et al.*, 2019)

4.2 Résultats sur le corpus SYNPAR

Dans le tableau 4, on peut noter que les représentations fusionnées donnent de meilleurs résultats lorsque l'on ne considère que la couche de plongement. Pour les autres méthodes de représentation (*mean*, *pooler*, `CLS`), nous retrouvons des résultats similaires à la partie précédentes : les gains sur BERT sont bien plus importants.

Tandis que MiniLM accuse une diminution des scores ABX, BERT dépasse la performance de MiniLM en prenant la moyenne des vecteurs comme représentation (voir bert-pt-ft dans le tableau 4). L'amélioration du score de 2,6 points par rapport à la version zero-shot nous permet d'affirmer que fusionner les représentations présente un intérêt pour le modèle BERT tant pour compresser les séquences que pour éviter les confusions avec la paronymie. Sur MiniLM, une amélioration est visible sur les données non-fusionnées après avoir pré-entraîné le modèle sur les fusions. Il y a donc une piste d'amélioration possible.

4.3 Discussion

Partant du constat que les modèles de langue calculent une similarité plus élevée avec les paronymes que les synonymes, nous avons proposé une méthode pour fusionner les tokens sur des critères linguistiques. L'évaluation a pris deux critères en compte. Le premier sur les performances sur une tâche commune de TAL, la STS. Le second, plus qualitatif, sur la distinction entre paronyme et synonyme dans le test ABX. Dans les deux cas, la méthode s'est montrée concluante après une phase de pré-entraînement et d'affinage sur la tâche STS. Si l'on considère l'évaluation sur le jeu de donnée STSBenchmark, le modèle BERT pré-entraîné et affiné donne les meilleurs résultats avec un taux de compression de 29,91 %. MiniLM montre une baisse de performance de 1 point mais conserve la

	représentation	configurations			
		mean	pooler	cls	plongement
miniLM-zs	fusionnée	59,74	64,29	63,31	**50,00**
	non fusionnée	**67,53**	**68,83**	**68,83**	49,03
miniLM-pt	fusionnée	58,17	62,99	62,99	50,00
	non fusionnée	65,91	67,53	69,16	49,02
miniLM-pt-ft	fusionnée	58,77	62,66	64,94	**50,32**
	non fusionnée	64,29	65,91	66,88	49,35
bert-zs	fusionnée	60,39	55,84	58,44	56,82
	non fusionnée	66,88	56,17	66,56	52,60
bert-pt	fusionnée	61,04	57,80	57,80	56,82
	non fusionnée	62,01	64,61	**65,26**	52.60
bert-pt-ft	fusionnée	**69,48**	**65,26**	66,23	56,82
	non fusionnée	67,53	63,64	**67,21**	52.60

TABLE 4 – Résultats des tests ABX sur le corpus SYNPAR comparant l'ajout de la fusion de tokens. Les résultats sont grisés lorsque la représentation d'inférence ne correspond pas à la représentation d'entraînement des modèles (modèle original sur représentation fusionnée ou inversement). Les meilleurs scores sont mis en gras.

propriété de compression. Dans les deux cas, les expériences menées ont montré une assez bonne stabilité aux changements de taux d'apprentissage (*learning rate*).

En réponse à notre constat initial, l'évaluation qualitative s'est montrée concluante uniquement pour BERT. En effet, ces résultats montrent que la tâche de STS joue un rôle dans la distinction des paronymes avec ce modèle, puisque l'affinage améliore les résultats. En revanche, MiniLM voit une baisse du score ABX, à l'exception de la version pt-ft évaluée sans fusion. Ce résultat peut être symptomatique d'une mauvaise méthode d'entraînement. Nous supposons qu'une amélioration pourrait être obtenue en augmentant la taille du corpus de pré-entraînement qui ne comporte que 14 millions de tokens. Notons que nos résultats sur STS ne peuvent être considérés comme une amélioration de la tâche puisque contrairement aux modèles originaux nous sortons du cadre zero-shot. Ainsi, la fusion de représentations permet d'atteindre des résultats similaires aux modèles originaux tout en induisant une forte diminution des séquences d'entrée. Nous pouvons observer sur la figure 3 qu'une fusion des tokens sur un critère d'analyse en constituants permet, sur les trois corpus, une réduction moyenne respective de $42{,}83\,\%$, $29{,}91\,\%$, et $26{,}42\,\%$ de la taille de l'entrée du modèle. Les textes de Wikipédia sont plus longs et donnent lieu à une compression plus importante. Comme expliqué en section 3, cette réduction permet donc de prendre des contextes plus longs en entrée sans augmenter la complexité des modèles.

5 Conclusions

Dans le but d'améliorer la représentation sémantique des textes, nous avons proposé une méthode de fusion des représentations vectorielles. En se basant sur l'hypothèse qu'une analyse en constituant peut fournir de l'information sémantique, nous fusionnons les représentations des tokens des constituants.

Cette méthode donne des résultats intéressants sur BERT et plus mitigés sur MiniLM. Dans les deux cas, la compression des séquences est importante, invitant à des expériences plus approfondies.

Nous devons reconnaître certaines limites à notre approche. La fusion des tokens est conditionnée à la proximité séquentielle des tokens. Il est difficile de considérer la fusion de deux tokens disjoints dans le cadre des modèles de langue. Or, même si les méthodes d'analyse en constituants sélectionnées ici ne les prennent pas en compte, certains constituants sont discontinus (Coavoux, 2020). La sélection d'un critère linguistique pour la fusion des tokens entraîne un pré-traitement lourd des textes qui amenuisent les gains de performance dus à la diminution de la taille des séquences, l'utilisation d'analyseurs syntaxiques plus rapides pourrait grandement améliorer notre méthode.

Les expériences réalisées ici prennent en compte une segmentation sur les syntagmes nominaux, verbaux et prépositionnels. D'autres types de segmentations peuvent également avoir un intérêt et nécessitent d'être testés. En effet, l'analyse syntaxique par constituant est surtout motivée en anglais et manque de littérature dans d'autres langues. La méthode proposée devrait être agnostique à la langue. Toutefois, les expériences n'ont été réalisées que sur des corpus anglais et pourraient bénéficier à être explorées dans d'autres langues, notamment en français. Les résultats obtenus sur MiniLM interrogent sur notre procédure de pré-entraînement. Un corpus contenant plus de tokens ou plus similaire à celui de pré-entraînement du modèle seraient intéressant. L'impact de la nature des corpus peut influencer les résultats obtenus. Il serait donc pertinent de construire un corpus similaire au corpus SYNPAR à partir des données du *STSBenchmark* pour s'assurer de la similarité des corpus.

Références

BOLYA D., FU C.-Y., DAI X., ZHANG P., FEICHTENHOFER C. & HOFFMAN J. (2023). Token merging : Your vit but faster. *ICLR*, **abs/2210.09461**.

CAMARA M. (2019). Syntagme nominal expansif ou l'expression de l'hétérogénéite et/ou de l'homogénéite dans *Petit bodiel* de Hampate Ba. *Multilinguales [En ligne]*. DOI : 10.4000/multilinguales.3962.

CARLIN M., 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*.

CHOMSKY N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, **2**(3), 113–124. DOI : 10.1109/TIT.1956.1056813.

COAVOUX M. (2020). Qu'apporte BERT à l'analyse syntaxique en constituants discontinus ? une suite de tests pour évaluer les prédictions de structures syntaxiques discontinues en anglais (what does BERT contribute to discontinuous constituency parsing ? a test suite to evaluate discontinuous constituency structure predictions in English). In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e 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. 189–196, Nancy, France : ATALA et AFCP.

DANY AMIOT, WALTER DE MULDER N. F. (2001). *Le Syntagme nominal : syntaxe et sémantique*. Artois Presses Université.

DAO T., FU D. Y., ERMON S., RUDRA A. & RÉ C. (2022). Flashattention : Fast and memory-efficient exact attention with io-awareness. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D.

BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems 35 : Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

DETTMERS T., PAGNONI A., HOLTZMAN A. & ZETTLEMOYER L. (2023). QLoRA : Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

GEE L., RIGUTINI L., ERNANDES M. & ZUGARINI A. (2023). Multi-word tokenization for sequence compression. In M. WANG & I. ZITOUNI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing : Industry Track*, p. 612–621, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-industry.58.

HUANG H., ZHU D., WU B., ZENG Y., WANG Y., MIN Q. & ZHOU X. (2025). Over-tokenized transformer : Vocabulary is generally worth scaling. *CoRR*, **abs/2501.16975**. DOI : 10.48550/ARXIV.2501.16975.

KUMAR D. & THAWANI A. (2022). BPE beyond word boundary : How NOT to use multi word expressions in neural machine translation. In S. TAFRESHI, J. SEDOC, A. ROGERS, A. DROZD, A. RUMSHISKY & A. AKULA, Éds., *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, p. 172–179, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.insights-1.24.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In Y. BENGIO & Y. LECUN, Éds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

MUENNIGHOFF N., TAZI N., MAGNE L. & REIMERS N. (2023). MTEB : Massive text embedding benchmark. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2014–2037, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.148.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A python natural language processing toolkit for many human languages. In A. CELIKYILMAZ & T.-H. WEN, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 101–108, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-demos.14.

REIMERS N., BEYER P. & GUREVYCH I. (2016). Task-oriented intrinsic evaluation of semantic textual similarity. In Y. MATSUMOTO & R. PRASAD, Éds., *Proceedings of COLING 2016, the 26th*

International Conference on Computational Linguistics : Technical Papers, p. 87–96, Osaka, Japan : The COLING 2016 Organizing Committee.

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)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

SCHARTZ 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*.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In K. ERK & N. A. SMITH, Éds., *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.

SPEARMAN C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1), 72–101.

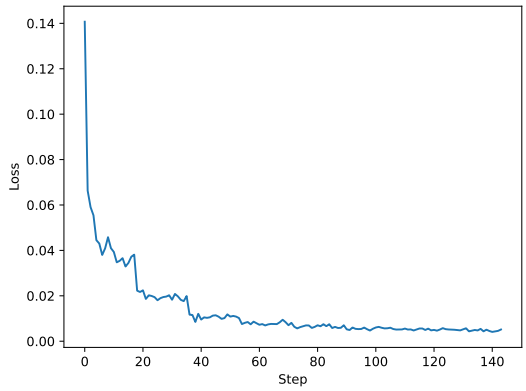
TAO C., LIU Q., DOU L., MUENNIGHOFF N., WAN Z., LUO P., LIN M. & WONG N. (2024). Scaling laws with vocabulary : Larger models deserve larger vocabularies. In A. GLOBERSON, L. MACKEY, D. BELGRAVE, A. FAN, U. PAQUET, J. TOMCZAK & C. ZHANG, Éds., *Advances in Neural Information Processing Systems*, volume 37, p. 114147–114179 : Curran Associates, Inc.

TYTGAT J., WISNIEWSKI G. & BETRANCOURT A. (2024). Évaluation de la similarité textuelle : Entre sémantique et surface dans les représentations neuronales. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, p. 85–96, Toulouse, France : ATALA and AFPC.

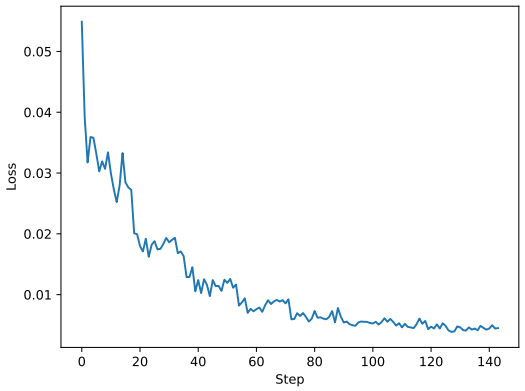
VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran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

Annexes

A Courbes d'entraînements

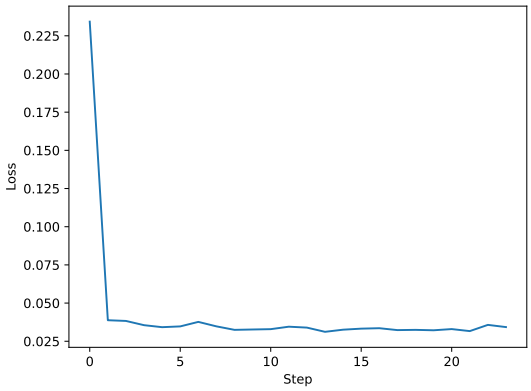


(a) `bert-base-uncased`

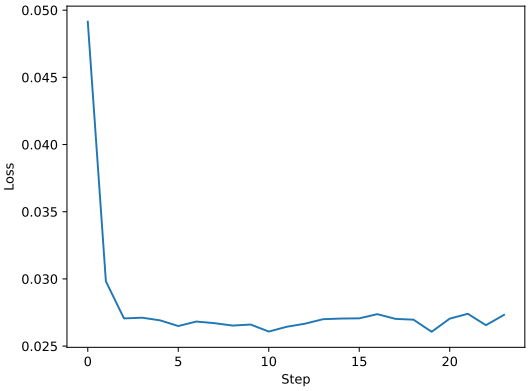


(b) `all-MiniLM-l6-v2`

FIGURE 4 – Évolution de la perte d'entraînement des modèles lors de l'affinage sur *STSbenchmark*

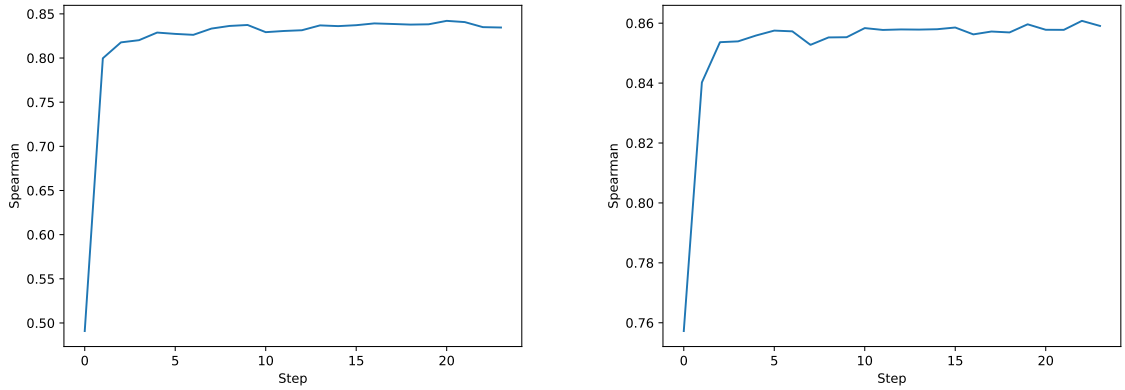


(a) `bert-base-uncased`



(b) `all-MiniLM-l6-v2`

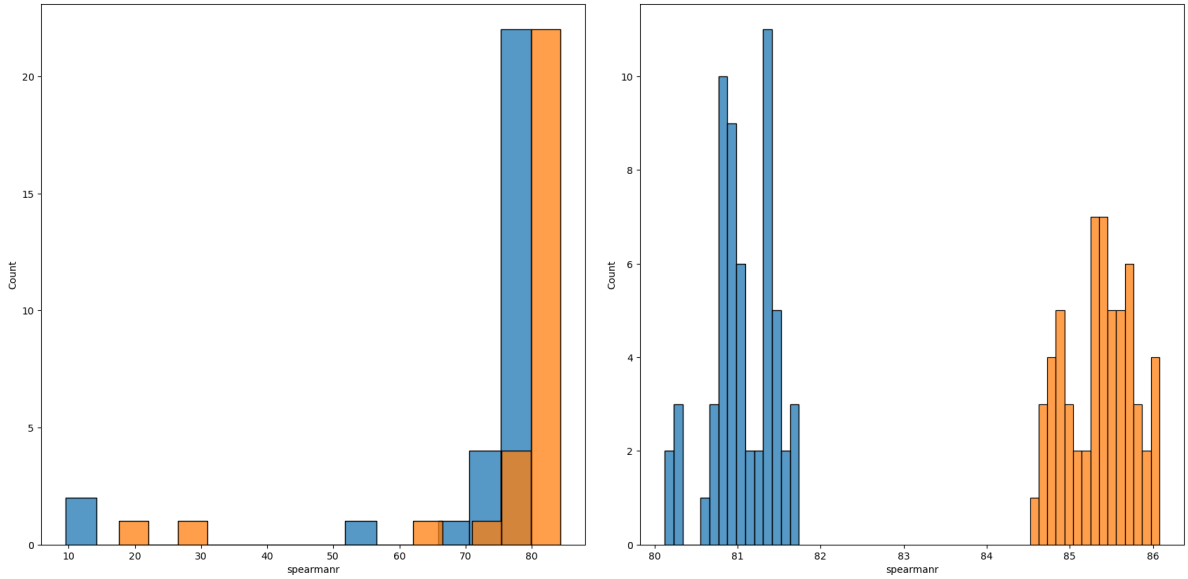
FIGURE 5 – Évolution de la perte de validation des modèles lors de l'affinage sur *STSbenchmark*



(a) `bert-base-uncased` (b) `all-MiniLM-l6-v2`

FIGURE 6 – Évolution de la corrélation de Spearman sur le jeu de validation lors de l'affinage des modèles sur *STSbenchmark*

B Distribution des résultats STS



(a) `bert-base-uncased` (b) `all-MiniLM-l6-v2`

FIGURE 7 – Distribution des résultats val (orange) et test (bleu) sur *STSBenchmark*