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

ParaPLUIE - une mesure automatique d'évaluation de la qualité sémantique des systèmes de paraphrases

Quentin Lemesle[1] Jonathan Chevelu[1] Damien Lolive[1] Arnaud Delhay[1]
Philippe Martin[1]
(1) Univ Rennes, IRISA, CNRS, 22300 Lannion, France
{quentin.lemesle, jonathan.chevelu, damien.lolive,
arnaud.delhay, philippe.martin}@irisa.fr,

RÉSUMÉ

L'évaluation des systèmes de production automatique de paraphrases est une tâche difficile car elle implique, entre autre, d'évaluer la proximité sémantique entre deux phrases. Les mesures traditionnelles s'appuient sur des distances lexicales, ou au mieux des alignements de plongements sémantiques. Dans cet article nous étudions certaines de ces mesures sur des corpus de paraphrases et de non-paraphrases reconnus pour leurs qualités ou difficultés sur cette tâche. Nous proposons une nouvelle mesure, ParaPLUIE, s'appuyant sur l'utilisation d'un grand modèle de langue. D'après nos expériences, celui-ci est plus à même de trier les paires de phrases par proximité sémantique.

ABSTRACT

ParaPLUIE : ParaPhrase, Llm Used for Improved Evaluation

Evaluating automatic paraphrase production systems is a difficult task because it involves, among other things, assessing the semantic proximity between two sentences. Usual measures are based on lexical distances, or at least on semantic embedding alignments. In this article we study some of these measures on datasets of paraphrases and non-paraphrases known for their quality or difficulty on this task. We propose a new measure, ParaPLUIE, based on the use of a large language model. According to our experiments, this one is better to sort pairs of sentences by semantic proximity.

MOTS-CLÉS : paraphrase, évaluation sémantique, grand modèle de langue.

KEYWORDS: paraphrase, semantic evaluation, large language model.

1 Introduction

Dans le domaine de la production automatique de paraphrases, de nombreuses définitions d'une paraphrase ont été proposées (Mel'čuk, 1997; Barzilay & McKeown, 2001; Sekine, 2005; Zhao *et al.*, 2009; Fabre *et al.*, 2021). Toutes ces définitions, tout comme les travaux de linguistiques traitant de paraphrase (Leeman, 1973), contiennent une notion de conservation du sens, celle-ci étant par nature ambiguë.

Malgré cela, les systèmes ont besoin de mesures automatiques de proximité sémantique pour s'entraîner ou se comparer. Généralement, celles utilisées fonctionnent par comparaison de lexiques (Papineni *et al.*, 2002) ou de plongements (Zhang *et al.*, 2020). Par construction, les approches par lexiques

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 605–616.

ont des difficultés à rapprocher des transformations simples comme le remplacement d'un mot par son synonyme (Banerjee & Lavie, 2005). Elles auront aussi du mal à rejeter deux phrases proches syntaxiquement même si elles ont un sens opposé. D'un autre coté, les mesures utilisant des plongements sémantiques reposent sur une notion d'alignement sous-phrastique sans vision globale des phrases. Ces deux points ont été mis en évidence par Zhang *et al.* (2019) et ont conduit à construire le corpus PAWS.

L'architecture *Transformer* et l'émergence des grands modèles de langues (*LLM*) ont permis de nombreuses avancées dans le domaine du traitement automatique du langage (Vaswani *et al.*, 2017). En particulier, le mécanisme d'auto-attention permet de capturer, sur un grand contexte, des relations sémantiques. Nous proposons d'explorer l'utilisation d'un *LLM* pour la mise au point d'une nouvelle mesure de similarité sémantique nommée ParaPLUIE.

Nous commençons par présenter, dans la section 2, les mesures communément utilisées, puis les corpus d'évaluation en section 3. Dans la section 4, nous présentons une expérience préliminaire d'utilisation d'un *LLM* en tant que classifieur de paraphrases avant de définir ParaPLUIE en section 5. Nous comparons ensuite ParaPLUIE aux autres mesures, en terme de dynamique des scores en section 6, puis en terme de re-classement de paraphrases en section 7.

2 Mesures sémantiques automatiques pour les paraphrases

L'état de l'art des mesures d'évaluation automatique de conservation de sens entre deux phrases peut être séparé en deux groupes. Le premier groupe concerne les mesures estimant la distance lexicale qui ont pour but d'évaluer à quel point les structures de deux phrases sont similaires. Le second regroupe les mesures estimant la proximité sémantique entre deux phrases.

Dans le premier groupe, on peut inscrire la distance de Levenshtein (LEV.) (Levenshtein, 1965), WER (Woodard & Nelson, 1982), BLEU (Papineni *et al.*, 2002) et METEOR (Banerjee & Lavie, 2005).

LEV. donne une mesure de différence entre deux chaînes de caractères. Cette mesure repose sur la détermination du nombre minimal de suppressions, insertions et remplacements pour passer de la chaîne de caractères hypothèse à la chaîne référence. LEV. augmentant avec la taille des chaînes de caractères considérées, elle est généralement normalisée par la taille de la chaîne de caractères la plus grande parmi l'hypothèse et la référence.

WER est un dérivé de LEV., travaillant au niveau des mots et non des caractères. Cette mesure correspond au rapport entre le nombre de mots communs entre la phrase hypothèse et celle de référence, et le nombre de mots de la plus longue des deux phrases.

BLEU a été conçu pour être une mesure de qualité de traduction. Elle consiste en la comptabilisation de la présence de ngrammes d'une phrase hypothèse dans une ou plusieurs phrases de référence. Généralement, tous les ngrammes de longueur 1 à 4 mots sont considérés. BLEU est associé à un score de rappel des ngrammes de l'hypothèse dans la référence. Par la suite, dans cet article, la version de BLEU que nous utilisons est l'implémentation de *torchtext* [1] avec les paramètres par défaut.

METEOR reprend les principes de BLEU en calculant une moyenne harmonique, à partir de la précision et du rappel de l'apparition d'un ngramme hypothèse parmi les références. De plus, METEOR prend

1. https://pytorch.org/text/stable/data_metrics.html

en compte une correspondance synonymique lors du calcul du score. METEOR a montré une meilleure corrélation avec le jugement humain que BLEU.

On pourrait argumenter que si deux phrases ont une structure lexicale très proche alors elles sont plus probablement des paraphrases. La faiblesse de cette hypothèse est que deux phrases peuvent partager une structure commune sans véhiculer le même sens. Pour pallier ce problème, un effort de recherche a été employé à la création d'un second groupe de mesures qui reposent sur une distance sémantique. Ces métriques s'appuient sur les plongements de symboles représentant des mots au sein d'un *LLM*. Dans ce groupe, on peut considérer notamment $\text{BERT}_{\text{score}}$ (Zhang *et al.*, 2020) et ParaScore (Shen *et al.*, 2022).

$\text{BERT}_{\text{score}}$ est un score de similarité de chaque plongement de symbole, composant une phrase hypothèse, avec les plongements de symboles d'une phrase de référence. Sa définition repose sur l'hypothèse que, s'il existe un appariement entre deux phrases tel que, tous les plongements qui les composent sont proches, alors leur sens est proche. Par la suite, dans cet article, nous utilisons la version de $\text{BERT}_{\text{score}}$ provenant de *Hugging Face* [2]. Celle-ci utilise le modèle BERT (Devlin *et al.*, 2019), nous spécifions le type du modèle en tant que *"bert-base-uncased"*.

Shen *et al.* (2022) observent que lorsque les distances lexicales séparant deux paraphrases augmentent, les performances des mesures diminuent. Ils proposent donc ParaScore, une mesure qui étend $\text{BERT}_{\text{score}}$, en ajoutant au calcul de similarité une distance de Levenshtein normalisée.

Il est à noter que les mesures de similarité sémantiques considèrent un alignement mot-à-mot, sans prise en compte de relations sémantiques de plus hauts niveaux. Ainsi un risque quant à la qualité de la classification de paraphrase existe.

Dans la suite de ce document, nous proposons une évaluation de ces différentes métriques dans le cadre de la similarité sémantique, sur deux corpus de paraphrases.

3 Corpus

L'étude des mesures automatiques, pour mesurer la proximité sémantique de paires de phrases, implique l'utilisation d'un corpus annoté en paraphrase/non paraphrase. Idéalement, afin de mesurer la pertinence des mesures dans des cas difficiles, les couples étiquetés non-paraphrases doivent être proches lexicalement ou sémantiquement (sans toutefois être considérés comme paraphrases par des évaluateurs humains). Notre choix s'est donc porté sur deux corpus en langue anglaise : PAWS (Zhang *et al.*, 2019), construit pour tromper les mesures lexicales, et MRPC (Dolan & Brockett, 2005) contenant des exemples d'inférence sémantique (mais asymétrique).

Pour PAWS, nous utilisons ici le sous-ensemble *dev*. Celui-ci comprend 8 000 couples dont 3 539 sont des paraphrases, soit 44% du corpus. L'entièreté du corpus est formée de 108 463 couples et a été générée de façon semi-automatique par inversion de mots et par traduction inverse. Lors de l'annotation de ces données, pour chaque couple de phrases, cinq juges ont été interrogés pour déterminer de façon binaire si les deux phrases sont des paraphrases. PAWS a été conçu pour être un challenge pour les modèles automatiques de détection de paraphrases. En effet, la génération de phrases par inversion de mots génère souvent des non-paraphrases, tout en maintenant une forte similarité lexicale. Voici un exemple de non-paraphrase caractéristique de PAWS : « *flights from New*

2. https://huggingface.co/spaces/evaluate-metric/bertscore

York to Florida » et « *flights from Florida to New York* ».

Le corpus MRPC utilisé est disponible sur *HuggingFace* [3] et comprend 5 801 couples dont 3 900 paraphrases, soit 67% du corpus. Ce corpus a été créé de façon automatique, depuis un grand corpus d'articles de presse regroupés par thème. Lors de l'annotation de ces données, le protocole a été le suivant : pour chaque couple de phrases, deux juges ont été interrogés pour savoir si les deux phrases pouvaient être considérées comme sémantiquement équivalentes ; ils ne pouvaient répondre que de façon binaire, par oui ou par non, et en cas de désaccord entre les deux jugements, un troisième juge répondait à la même consigne. Voici un exemple de non-paraphrase caractéristique de MRPC : « *Last year, Bush appointed him to the Homeland Security Advisory Council.* » et « *He has also served on the president's Homeland Security Advisory Council.* ».

L'ensemble des deux corpus comporte 54% de paraphrases. Les distributions, des mesures présentées à la section 2 et appliquées sur chacune des classes de ces deux corpus, sont présentées dans la table 1. On constate effectivement que les paires de phrases de PAWS sont très proches contrairement à MRPC.

Au sein d'un même corpus, pour toutes les mesures considérées (sauf ParaScore sur PAWS), on constate que les moyennes des différentes classes sont bien cohérentes avec l'étiquette de référence. En revanche, les écarts-types laissent penser qu'une classification ou qu'un tri des phrases en fonction de leur score, ne permettraient pas de bien identifier les paraphrases. De plus, si on croise les résultats des corpus, on constate que le score moyen des non-paraphrases de PAWS est meilleur que le score moyen des paraphrases de MRPC, quelle que soit la mesure. Rappelons toutefois que nous traitons volontairement des corpus très difficiles pour des mesures de proximité sémantique.

Corpus	*Para.*	Taille	LEV. ↓	WER ↓	BLEU ↑	METEOR ↑	$\text{BERT}_{\text{score}}$ ↑	ParaScore ↑
MRPC	Oui	3900	0,38 ±0,16	0,51 ±0,20	0,40 ±0,21	0,69 ±0,14	0,82 ±0,07	0,83 ±0,07
	Non	1901	0,51 ±0,13	0,67 ±0,19	0,28 ±0,18	0,56 ±0,15	0,74 ±0,08	0,76 ±0,09
PAWS	Oui	3539	0,20 ±0,15	0,26 ±0,18	0,62 ±0,18	0,91 ±0,06	0,94 ±0,04	0,92 ±0,03
	Non	4461	0,32 ±0,15	0,37 ±0,18	0,49 ±0,19	0,88 ±0,07	0,91 ±0,04	0,92 ±0,04
Total	Oui	7439	0,30 ±0,18	0,39 ±0,23	0,51 ±0,22	0,80 ±0,16	0,88 ±0,08	0,88 ±0,07
	Non	6362	0,38 ±0,17	0,46 ±0,23	0,43 ±0,21	0,79 ±0,18	0,86 ±0,10	0,87 ±0,09

TABLE 1 – Moyenne des scores de chaque mesure sur les corpus MRPC et PAWS. Les corpus ont été découpés en sous-corpus séparant les paraphrases et non-paraphrases (colonne *Para.*). La taille des corpus dénote le nombre de couples de phrases. Le signe ↑ associé à une mesure indique que plus sa valeur est élevée, meilleur est son score, et ↓ signale l'inverse.

4 Expérience préliminaire : classer avec un *LLM*

Les mesures actuelles se focalisent sur la notion de proximité lexicale ou au mieux d'alignement entre plongements de mots. En conséquence, elles ne peuvent pas prendre en compte des relations complexes entre paraphrases. Récemment, les progrès des architectures de type *Transformer* ont montré qu'il était possible d'avoir une meilleure prise en compte de relations internes à un texte, grâce

3. https://huggingface.co/docs/datasets/v1.13.0/about_dataset_features.html?highlight=mrpc

aux mécanismes d'auto-attention (Vaswani *et al.*, 2017). Pour la création d'une mesure de proximité sémantique, notre intuition est que ces derniers seraient plus à même « d'aligner » les parties des phrases ayant un sens proche.

Afin de vérifier la capacité d'un *LLM* à détecter la relation de paraphrase, une expérience préliminaire consiste à l'utiliser en mode génératif sur le principe d'un agent conversationnel. Pour cette expérience, le modèle Mistral (Jiang *et al.*, 2023), un modèle de taille intermédiaire parmi les grands modèles de langue est utilisé. Précisément, nous utilisons la version *7B Instruct v0-2* [4] au format demi-précision. Ce modèle est basé sur l'architecture *Transformer* et utilise une fenêtre d'attention glissante, dans le but de réduire le coût de calcul. Comme son nom de version l'indique, il possède 7 milliards de paramètres. Le corpus utilisé lors de son entraînement n'est pas divulgué. Dans cette configuration, l'empreinte mémoire du modèle est de 15 gigaoctets. Les expériences présentées ici ont été réalisées sur une machine équipée d'une carte carte graphique Nvidia RTX 4090.

On pose la paraphrase hypothétique (H) et la phrase de référence (R). Le classifieur binaire considéré ici consiste à appliquer un patron sur cette paire (Patron(H, R)) et à utiliser le modèle pour produire les symboles suivants les plus probables. Puisque ce modèle fonctionne sur un principe d'agent conversationnel, le patron simule le début d'un échange entre un utilisateur ($user$), et assistant ($assistant$). Un premier patron testé consiste à présenter les deux phrases et à demander si elles ont le même sens. Si le premier symbole produit en anglais est « *yes* », alors les paires sont considérées comme paraphrases. Toute autre réponse (« *no* », « *they are similar* », ...) est associée à un label négatif. Ce patron noté **Patron**$_{\text{Direct}}$ est précisément construit comme ceci :

Patron$_{\text{Direct}}(H, R)$:
($user$) : « You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no". »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »

Qiao *et al.* (2023) relèvent le fait que les performances d'un *LLM* sont améliorées si l'on utilise des étapes intermédiaires lors de la génération. Cela permet de simuler un cheminement de pensée (Wei *et al.*, 2022). De part la nature auto-régressive des *LLM*, ajouter des informations dans le patron aide à la résolution de la tâche. Nous proposons les deux patrons détaillés ci-dessous utilisant la génération d'une explication intermédiaire (E) avant la classification de relation entre deux phrases.

Patron$_{\text{Expliqué}}(H, R)$
($user$) : « You will receive two sentences A and B. Do these two sentences mean the same thing? »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »
($assistant$) : « (E) »
($user$) : « Summarize your answer with only one word "yes" or "no". »

De peur que le système ne favorise la catégorisation en paraphrase, nous proposons un patron demandant si les phrases ont exactement le même sens.

Patron$_{\text{Exact}}(H, R)$
($user$) : « You will receive two sentences A and B. Do these two sentences mean **exactly** the same

4. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

thing? »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »
($assistant$) : « (E) »
($user$) : « Summarize your answer with only one word "yes" or "no". »

Corpus	Patron	VP	VN	FP	FN	Exactitude	F1	Rappel	Précision
MRPC	Direct	3 550	956	945	350	0,78	0,84	0,91	0,79
MRPC	Expliqué	3 089	1 343	558	811	0,76	0,82	0,79	0,85
MRPC	Exact	1 965	1 661	240	1 935	0,62	0,64	0,50	0,89
PAWS	Direct	3 381	1 297	3 164	158	0,58	0,67	0,95	0,52
PAWS	Expliqué	3 177	1 919	2 542	362	0,64	0,68	0,90	0,55
PAWS	Exact	2 453	3 074	1 387	1 086	0,69	0,66	0,69	0,64
Total	Direct	6 931	2 253	4 109	508	0,66	0,75	0,93	0,63
Total	Expliqué	6 266	3 262	3 100	1 173	0,69	0,74	0,84	0,67
Total	Exact	4 418	4 735	1 627	3 021	0,66	0,65	0,59	0,73

TABLE 2 – Performance du classifieur par *LLM* en fonction du patron utilisé. Le comptage des résultats est donné en tant que Vrai Positif (VP), Vrai Négatif (VN), Faux Positif (FP) et Faux Négatif (FN).

Les résultats de classification de ces approches sur les corpus PAWS et MRPC, sont rapportés dans la table 2. L'utilisation du patron direct donne une exactitude (*accuracy*) supérieure au hasard (0,66 contre 0,54). Comme supposé, l'ajout de l'explication permet d'améliorer les performances de l'approche directe. En revanche, imposer une relation d'équivalence sémantique exacte dégrade les résultats. On notera qu'une variation, même minime de la formulation du patron, peut influencer grandement les résultats.

Cette expérience montre la capacité du *LLM* à discriminer les paraphrases des non-paraphrases, malgré un contexte difficile. En revanche, compte-tenu du caractère continu de la relation de paraphrase, il semble plus judicieux de travailler avec un degré de proximité sémantique plutôt qu'une classification binaire. L'usage d'un *LLM* semble donc pertinent pour construire une mesure plus performante que celles traditionnellement utilisées.

5 Proposition : ParaPLUIE

Pour rappel, les modèles de langues sont avant tout une modélisation des probabilités d'apparition d'un symbole textuel, sachant un historique. Il est donc possible de comparer deux séquences pour calculer un degré d'appartenance à une classe comme Chen *et al.* (2023). Ainsi, nous proposons ParaPLUIE (*ParaPhrase, Llm Used for Improved Evaluation*), une mesure de proximité sémantique reposant le modèle probabiliste d'un *LLM*. ParaPLUIE est définie comme le logarithme des rapports de vraisemblances, sur le fait que le patron appliqué à la paraphrase hypothétique (H) et à la référence (R) est suivi du symbole « yes » ou « no », c'est-à-dire :

$$\text{ParaPLUIE}(H, R) \quad = \quad \log \left(\frac{p\left(\text{yes}|\text{Patron}(H, R)\right)}{p\left(\text{no}|\text{Patron}(H, R)\right)} \right)$$

Si les patrons sont identiques et les mots « yes » et « no » ne sont codés que sur un unique symbole, alors ce ratio de probabilité est égale au ratio des perplexités (ppl), à une puissance près. La perplexité reflétant justement la « surprise » du modèle lors de l'apparition des symboles. De plus, généralement, les *LLM* sont justement appris en utilisant la perplexité comme fonction objectif ($loss$). Ainsi, le calcul de la mesure devient :

$$\begin{aligned} \text{ParaPLUIE}(H,R) &= \log\left(\frac{ppl\left(\text{Patron}(H,R)\circ\text{no}\right)^{T+1}}{ppl\left(\text{Patron}(H,R)\circ\text{yes}\right)^{T+1}}\right) \\ &= (T+1)\times\left[loss_{LLM}(\text{Patron}(H,R)\circ\text{no}) - loss_{LLM}(\text{Patron}(H,R)\circ\text{yes})\right] \end{aligned} \tag{1}$$

où T est le nombre de symboles dans le patron et « $\circ$ » l'opération de concaténation de deux textes.

La mesure proposée est donc à valeurs réelles. Plus le score est élevé et plus le système estime que les deux phrases sont vraisemblablement des paraphrases alors qu'un score inférieur à zéro indiquerait que le sens des deux phrases est différent. Notons que cette propriété aide à l'interprétation des résultats contrairement à d'autres scores.

Comme pour la section 4, nous utilisons le *LLM* Mistral pour ParaPLUIE. Puisque les résultats de la table 2 sont relativement proches sur la tâche de classification, et par soucis de simplicité, nous utilisons le $\textbf{Patron}_{\text{Direct}}$.

6 Dynamique des scores

Nous nous intéressons ici à la dynamique des scores ParaPLUIE au regard des étiquettes paraphrases/non-paraphrases fournis dans les corpus de la section 3.

La distribution des scores ParaPLUIE sur les deux corpus de référence est présentée table 3. Contrairement à ce qui a été observé dans la table 1, le score moyen des paraphrases est bien supérieur à celui des non-paraphrases en intra-corpus et en inter-corpus. On regrettera que le score moyen des non-paraphrases ne soit pas négatif. Encore une fois, cela peut s'expliquer par le caractère volontairement trompeur des corpus considérés dans cette expérience.

Corpus	*Para.*	ParaPLUIE ↑
MRPC	Oui	20,02 ±8,94
	Non	4,41 ±15,43
PAWS	Oui	22,04 ±6,64
	Non	12,80 ±13,46
Total	Oui	20,98 ±7,99
	Non	10,29 ±14,59

TABLE 3 – Moyenne et écart-type des scores ParaPLUIE sur les corpus MRPC et PAWS.

La figure 1 propose une comparaison graphique des distributions des scores en fonction de l'étiquette des phrases. Puisqu'aucune mesure ne classe parfaitement les corpus, on observe un chevauchement

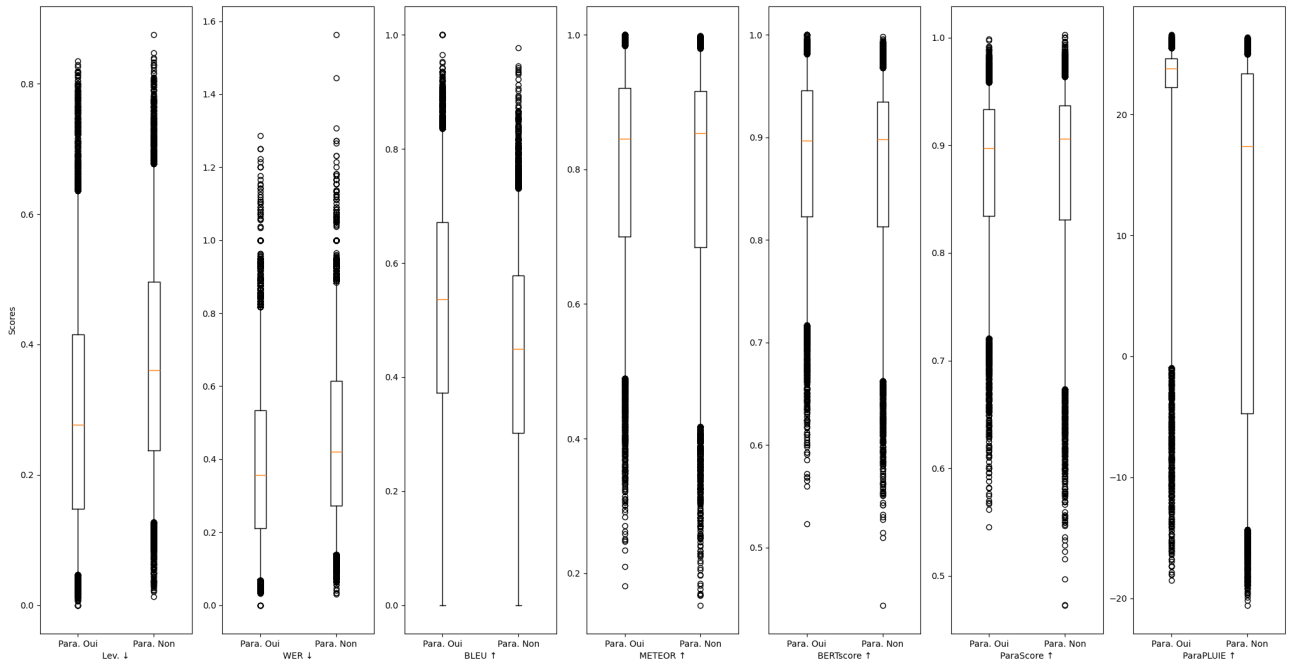


FIGURE 1 – Boîte à moustache des scores pour les différentes mesures. Les cercles correspondent aux 5% des valeurs les plus extrêmes.

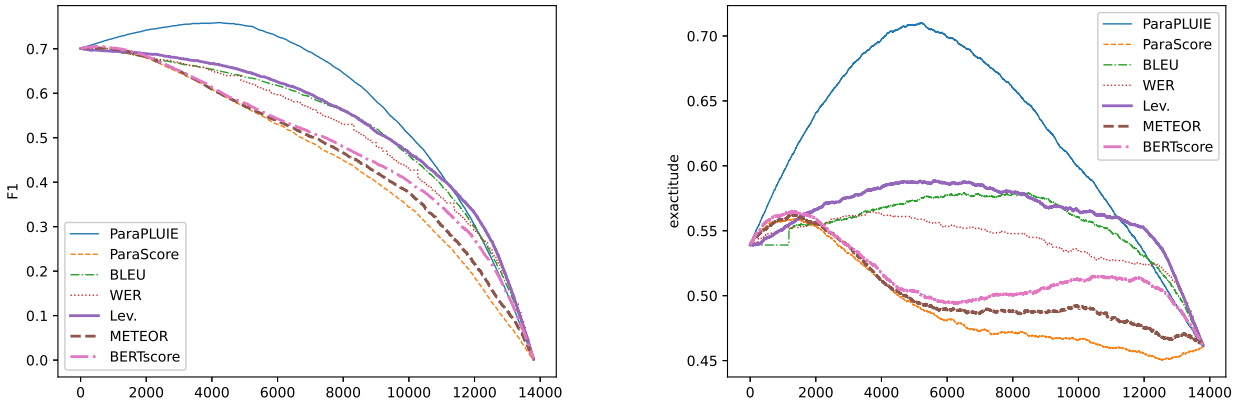
des scores. Mais pour ParaPLUIE, contrairement aux autres mesures, le score des non-paraphrases semble avoir une dynamique bien différente du score des paraphrases. Notons aussi que les mesures lexicale semble meilleurs sur ces corpus que les mesures sémantiques, hormis ParaPLUIE.

7 Re-classement de paraphrases

Comparons la capacité qu'ont les différentes mesures à ordonner les paraphrases et les non-paraphrases des corpus. Les paires de phrases sont triées par score décroissant pour LEV. et WER, et par score croissant pour les autres. Ainsi, plus une paire de phrases a un rang élevé, plus ces dernières sont considérées comme proches l'une de l'autre par la mesure. En faisant varier le rang à partir duquel on considère qu'une paire est effectivement paraphrase, et sachant l'étiquette d'une paire, il est possible de calculer la variation du score F1 (moyenne harmonique du rappel et de la précision), ainsi que l'exactitude de classification. Ces résultats sont présentés en figure 2.

Comme l'illustre la figure 2a, le score F1 de toutes les mesures chute rapidement, sauf celui de ParaPLUIE. On remarque figure 2b que l'exactitude reste basse, sauf pour ParaPLUIE, et qu'elle chute rapidement pour METEOR, $\text{BERT}_{\text{score}}$ et ParaScore. Sur ce type de corpus, étonnamment, les mesures lexicales exactes surpassent les mesures plus sémantiques autres que ParaPLUIE. Cette dernière est meilleure que les mesures lexicales, pour la très grande majorité des rangs, et surpasse toujours celles plus sémantiques.

Si l'on se concentre sur le maximum a posteriori de score F1, on constate qu'à l'exception de ParaPLUIE, l'optimal consiste approximativement à classifier toutes les paires comme paraphrases. Pour le maximum d'exactitude a posteriori, avec 0,71, ParaPLUIE est sensiblement meilleur que



(a) F1 en fonction du rang.

(b) Exactitude en fonction du rang.

FIGURE 2 – Évolution de l'exactitude et de la F1 des différentes mesures en fonction du rang.

les autres mesures, comprises entre 0,56 et 0,59. Ces résultats sont détaillés dans la table 4. Notons que, compte-tenu du léger déséquilibre des classes, étiqueter toutes les phrases comme paraphrases donnerait une exactitude de 0,54.

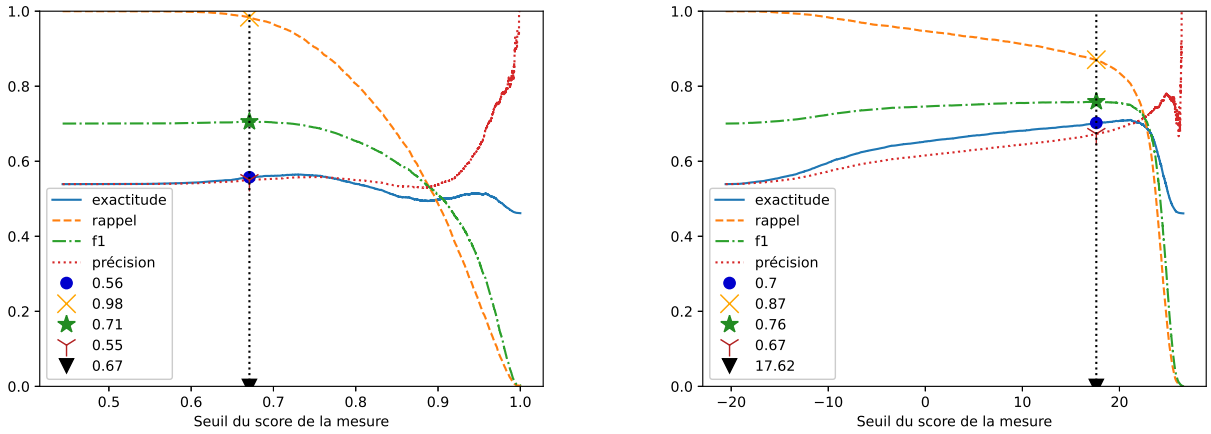
Score	LEV.	WER	BLEU	METEOR	$\text{BERT}_{\text{score}}$	ParaScore	ParaPLUIE
F1 max.	0,70	0,70	0,70	0,70	0,71	0,70	**0,76**
$\text{Rappel}_{\text{F1 max.}}$	**1,00**	**1,00**	**1,00**	0,99	0,98	0,99	0,87
$\text{Précision}_{\text{F1 max.}}$	0,54	0,54	0,54	0,55	0,55	0,55	**0,67**
Exact. max.	0,59	0,56	0,58	0,56	0,56	0,56	**0,71**

TABLE 4 – Récapitulatif des scores de rappel et de précision selon la meilleure F1 *(F1 max.)* ainsi que de l'exactitude la plus haute *(Exact. max.)* pour chaque mesure sur les deux corpus. Les valeurs les plus élevées sont en gras.

Les figures 3a et 3b permettent de comparer plus en détail les classements produits par $\text{BERT}_{\text{score}}$ et ParaPLUIE. On constate la chute précoce du rappel pour $\text{BERT}_{\text{score}}$ alors qu'elle intervient beaucoup plus tard pour ParaPLUIE. Pour la précision, celle de $\text{BERT}_{\text{score}}$ n'augmente que pour les valeurs maximales – un score très proche de 1 semble être un indicateur fiable de la relation de paraphrase – alors que celle de ParaPLUIE augmente beaucoup plus tôt. Malgré cela, on notera la présence de non-paraphrases dans les meilleurs scores de ParaPLUIE.

ParaPLUIE étant définie par une soustraction (voir l'équation 1), il existe un point de bascule entre les deux probabilités en 0. En fixant un seuil de classification sur cette valeur, l'exactitude est de 0,65. Cette performance reste supérieur à toutes les autres mesures, quelque soit le seuil choisi. Autrement dit, le seuil de classification fixé à priori pour ParaPLUIE est meilleur que toutes les autres mesures, même en fixant leurs seuils de classification à posteriori.

Ces expériences, sur des corpus complexes semblent indiquer que ParaPLUIE est une bonne mesure de proximité sémantique plus performante que l'état l'art.



(a) BERT$_{\text{score}}$ (b) ParaPLUIE

FIGURE 3 – Évolution de la qualité du tri des paraphrases/non-paraphrases en fonction de la mesure. Les valeurs mises en avant correspondent au maximum de score F1.

8 Conclusion et perspectives

ParaPLUIE est une nouvelle mesure de proximité sémantique utilisant un *LLM*. Nous avons évalué la qualité de notre mesure par rapport aux autres mesures communément utilisées. Cette évaluation a été effectuée sur deux corpus de paraphrases en anglais, annotés par des humains, reconnus pour leurs qualités ou difficultés. Notre analyse montre que ParaPLUIE obtient une meilleure exactitude que les autres mesures de l'état de l'art. Nous proposons des variantes de patrons et montrons que s'appuyer sur une étape d'explication intermédiaire générée par *LLM* améliore l'exactitude.

Cette étude a été menée sur une quantité limitée de données, dû au temps de calcul important nécessaire à l'utilisation d'un *LLM*. Dans ces expériences, nous n'avons pas réalisé de *few-shots prompting*, c'est-à-dire, ajouté un exemple de résolution de la tâche dans le patron utilisé. Nous sommes confiants dans le fait que cela améliorerait les résultats de ParaPLUIE mais nous avons souhaité éviter de créer un biais sur les corpus que nous avons étudiés.

Être capable d'estimer si deux phrases sont des paraphrases permet l'avancée sur d'autres terrains de recherche, comme la création de petit modèles de langue dédiés à la tâche de production de paraphrase, en coopération avec des méthodes d'apprentissage par distillation de connaissances (Hsieh *et al.*, 2023).

Nous souhaitons étendre cette étude sur des corpus de paraphrases de grande distance lexicale, là où cette étude s'est tournée sur un ensemble de couples avec une faible distance lexicale. Nous souhaitons également concevoir un petit modèle de langue dédié à l'identification de paraphrase.

Enfin nous appelons à ne pas utiliser ParaPLUIE avec un *LLM* de très grande taille. Intuitivement, l'utilisation d'un tel modèle améliorerait les résultats de ParaPLUIE ; néanmoins, rien ne le garantit et leur utilisation est particulièrement coûteuse. En revanche l'étude du patron utilisé semble être une piste prometteuse pour améliorer la mesure.

Remerciements

Recherche soutenue financièrement par le Ministère des Armées - Agence de l'Innovation de la Défense.

Références

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS, Éds., *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 50–57, Toulouse, France : Association for Computational Linguistics. DOI : 10.3115/1073012.1073020.

CHEN Y., WANG R., JIANG H., SHI S. & XU R. (2023). Exploring the use of large language models for reference-free text quality evaluation : An empirical study. In J. C. PARK, Y. ARASE, B. HU, W. LU, D. WIJAYA, A. PURWARIANTI & A. A. KRISNADHI, Éds., *Findings of the Association for Computational Linguistics : IJCNLP-AACL 2023 (Findings)*, p. 361–374, Nusa Dua, Bali : Association for Computational Linguistics.

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

DOLAN W. B. & BROCKETT C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

FABRE B., URVOY T., CHEVELU J. & LOLIVE D. (2021). Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2100–2111, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.180.

HSIEH C.-Y., LI C.-L., YEH C.-K., NAKHOST H., FUJII Y., RATNER A., KRISHNA R., LEE C.-Y. & PFISTER T. (2023). Distilling step-by-step ! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics : Association for Computational Linguistics 2023*, p. 8003–8017, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.507.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

LEEMAN D. (1973). La paraphrase. *Langages*, p. 43–54.

LEVENSHTEIN V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, **163**, 845–848.

MEL'ČUK I. M. (1997). *Vers une linguistique sens-texte : leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France.

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.

QIAO S., OU Y., ZHANG N., CHEN X., YAO Y., DENG S., TAN C., HUANG F. & CHEN H. (2023). Reasoning with language model prompting : A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5368–5393, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.294.

SEKINE S. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

SHEN L., LIU L., JIANG H. & SHI S. (2022). On the evaluation metrics for paraphrase generation. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 3178–3190, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.208.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

WOODARD J. & NELSON J. (1982). An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZHANG Y., BALDRIDGE J. & HE L. (2019). PAWS : Paraphrase Adversaries from Word Scrambling. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

ZHAO S., LAN X., LIU T. & LI S. (2009). Application-driven statistical paraphrase generation. In K.-Y. SU, J. SU, J. WIEBE & H. LI, Éds., *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 834–842, Suntec, Singapore : Association for Computational Linguistics.