Questionnaire - Compte Rendu des Résultats

Abdurrahman Shahid

L3 MIASHS - Sciences Cognitives

Février 2021

1 Introduction

Le but de ce résumé des résultats est d'expliquer la méthodologie de l'étude d'une part, et les critères de performance utilisés d'autre part. Cependant, ce document n'étant pas la production final, il sera court et n'entrera pas dans les détails du modèle (notamment son mécanisme d'*attention*) utilisé ni de la méthodologie utilisée pour entraîner le modèle.

L'objectif était de créer un modèle, utilisant des réseaux de neurones artificiels, qui apprend à corriger les erreurs liées à certaines paires d'homophones. Pour désigner le fait que les modèles utilisent des réseaux de neurones artificiels, on parle plus communément de *Deep Learning* (**DL**). De plus, lorsque l'on travaille aves des données textuelles, on parle généralement de *Natural Language Processing* (**NLP**). Dans le domaine du traitement du langage, une avancée majeure a eu lieu en 2017 lorsque Vaswani et al. (article ici) ont introduit le modèle appelé *Transformer*. Bien qu'une traduction française du terme existe, le "Transformateur", nous utiliserons son nom d'origine en anglais le **Transformer**. Le modèle en question qui a été entraîné pour corriger les homophones, est le modèle Transformer de base qui a été introduit en 2017, avec toutefois une taille très largement réduite. En ce qui concerne le fonctionnement du modèle, on procède de la façon suivante. En entrée du modèle on donne une phrase. Cette phrase peut comporter ou non une erreur sur un homophone. Le modèle doit alors remplacer l'homophone par sa forme correcte s'il y a une erreur, et ne rien modifier si la phrase est déjà correcte. Pour ce faire, cette étude s'appuie sur les progrès récents accomplis aussi bien dans le domaine du DL que du NLP.

Enfin les **homophones**, ce sont des mots qui se prononcent de la même façon, mais qui sont écrits différemment et ont des significations différentes. Pour cette étude, 10 paires d'homophones ont été utilisées (voir Table 1).

("a","à")	("est","et")	("ces","ses")	("ce","se")	("ou","où")
("la","là")	("tout","tous")	("leur","leurs")	("ceux","ce")	("cette","cet")

Table 1: Les 10 paires d'homophones utilisées

2 Mesure des performances

Pour mesurer et comparer les performances des humains avec le modèle Transformer, on va utiliser 3 mesures/probabilités (Precision, Recall, Accuracy), qui se déduisent à partir de 4 autres mesures (TP, FP, TN, FN). Intéressons nous d'abord à ces 4 mesures.

2.1 Les 4 situations possibles: TP, FP, TN, FN

Nous allons voir le raisonnement avec une paire d'homophones en particulier. Le raisonnement sera ensuite généralisé à tout les autres paires. Prenons la paire (a, à). Ici l'objectif pour l'humain et le modèle est de pouvoir détecter une erreur s'il y en a. Tout d'abord, il y a 2 possiblités: soit la phrase que l'on propose à l'humain ou au modèle comporte une erreur sur un homophone, soit la phrase ne comporte pas d'erreur. Ensuite il y a encore 2 possibilités, l'humain ou le modèle va choisir de modifier ou pas l'homophone présent dans la phrase qu'on lui propose.

Par exemple, on peut proposer la phrase "Il *a* fait valoir que les tarifs du câble ont été réglementés aux États-Unis" (1). Cette phrase est correcte car c'est bien "a" qu'il faut mettre et non "à". Cette phrase est ce que l'humain ou le modèle prend en entrée. Cependant, même si la phrase en entrée est correcte, l'humain ou le modèle peut très bien modifier l'homophone (ici "a") et le remplacer par sa forme associé (ici "à"), et ainsi se tromper. Lorsque la phrase en entrée est correcte et que le modèle ou l'humain modifie la phrase, et donc se trompe, on parle de *faux positif*, puisque une erreur est détectée (d'où le terme *positif*) là où il y en a pas (d'où le terme *faux*). On utilise alors l'abbréviation **FP** (qui vient de l'anglais *False Positive*) pour désigner cette situation. Par ailleurs, si l'humain ou le modèle ne change pas la phrase correcte qu'on lui a proposée (ici "Il *a* fait valoir que les tarifs du câble ont été réglementés aux États-Unis."), alors il ne commet pas d'erreur et on parle de *vrai négatif*, puisque aucune erreur n'est détectée (d'où le terme *négatif*) et il n'y a avait vraiment aucune erreur (d'où le terme *vrai*). On utilise alors l'abbréviation **TN** (qui vient de l'anglais *True Negative*) pour désigner cette situation. On vient de voir les 2 possibilités qui en découlent lorsque la phrase présentée est correcte. Regardons maintenant ce qui se passe, si la phrase présentée à l'humain ou au modèle contient une erreur au niveau d'un homophone.

On peut prendre une phrase en y introduisant volontairement une erreur (c'est

d'ailleurs comme cela que les données d'entraînement ont été créées pour entraîner le modèle Transformer). On propose à l'humain ou au modèle, par exemple, la phrase "Mais quand t'es embarqué dans l'engrenage, tu ne penses même pas *a* ça" (2). Cette phrase est incorrecte car ça devrait être "à" et non "a". L'objectif pour l'humain ou le modèle est alors de détecter cette erreur et remplacer "à" par "a" pour donner en sortie la correction "Mais quand t'es embarqué dans l'engrenage, tu ne penses même pas *à* ça". Toutefois, l'humain ou le modèle peut échouer à détecter cette erreur, et ainsi ne modifiera pas la phrase et laissera la forme incorrecte "a". Lorsque la phrase en entrée est incorrecte, et que l'humain ou le modèle échoue à détecter l'erreur (ne corrige pas l'erreur et ainsi se trompe en acceptant de laisser la forme incorrecte), on parle de *faux négatif*, puisqu'une erreur n'est pas détectée (d'où le terme *négatif*) alors qu'il y a vraiment une erreur (d'où le terme *faux*). On utilise alors l'abbréviation **FN** (qui vient de l'anglais *False Negative*) dans cette situation. Par contre, si l'humain ou le modèle arrive à détecter l'erreur et remplacer "a" par la forme correcte "à" (et donc en sortie donne "Mais quand t'es embarqué dans l'engrenage, tu ne penses même pas *à* ça."), alors on parle de *vrai positif*, puisqu'une erreur est détectée (d'où le terme *positif*) et qu'il y avait vraiment une erreur (d'où le terme *vrai*). On utilise alors l'abbréviation **TP** (qui vient de l'anglais *True Positive*) dans cette situation. Ces 2 possibilités viennent s'ajouter aux 2 possibilités vu au dernier paragraphe, ce qui nous fait donc 4 situations possibles, à savoir TP, FP, TN et FN. Le tableau 2 résume les 4 possibilités.

	Sortie correcte :	Sortie incorrecte :
Entrée incorrecte :	TP	FN
Entrée correcte :	TN	FP

Table 2: Les 4 situations possibles, lorsqu'une phrase comporte un homophone.

Ainsi, avec les phrases d'exemples (1) et (2) nous venons de voir les 4 cas possibles lorsque'on propose au modèle ou à l'humain une phrase comportant l'homophone "a". Chaque phrase conduit à deux possibilités, TN/FP pour l'une, et TP/FN pour l'autre. De plus, chaque homophone est associé à un autre homophone, formant une paire d'homophones. L'homophone associé est alors ce que le modèle ou l'humain met, s'il décide de modifier la phrase donnée en entrée. De la même manière, on peut donner en entrée une phrase comportant "à" et qui est correcte, et une phrase comportant "à" mais

incorrecte (le modèle ou l'humain devra alors changer le "à" en "a"). Donc, pour la paire d'homophones ("a","à"), on a besoin de 4 phrases différentes pour couvrir tous les cas possibles.

Plus généralement, on a besoin de 4 phrases pour représenter tout les cas possibles en entrée pour chaque paire d'homophones. Par conséquent, pour avoir tous les cas possibles pour toutes les paires (on en a 10 dans cette étude) on a besoin de 40 phrases. C'est pourquoi dans le questionnaire en ligne il y avait 40 questions. En utilisant ces 4 mesures pour chaque homophone, on construit 3 autres mesures qui peuvent être interprétées comme des probabilités.

2.2 Precision, Recall et Accuracy

Les 3 critères qui seront utilisés pour comparer les performances sont appelés en anglais, *Precision*, *Recall* et *Accuracy*. Ces mesures sont obtenues en effectuant les calculs suivants:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

La mesure du **Precision** prend en compte toutes les fois où l'homophone présent dans la phrase proposée au modèle ou à l'humain a été changé en sa forme associée par ce dernier. Soit il l'a changé et il y avait vraiment une erreur (TP), soit il l'a changé mais il n'y avait pas d'erreur (FP, il s'est trompé en changeant). Le *Precision* donne la probabilité que le modèle ou l'humain ait eu raison s'il change l'homophone présent dans la phrase. Autrement dit, si le modèle ou l'humain change l'homophone présent dans la phrase et met la forme associée à la place, la probailité qu'il ait raison est donnée par le *Precision*.

$$\text{Recall } = \frac{TP}{TP + FN} \tag{2}$$

La mesure du **Recall** prend en compte toutes les fois où l'homophone présent dans la phrase proposée au modèle ou à l'humain doit être changé en sa forme associée par ce dernier. Toutes les phrases présentées pour mesurer ce critère contiennent donc une erreur. Par exemple, la phrase "Mais quand t'es embarqué dans l'engrenage, tu ne penses même pas *a* ça" peut être utilisée pour calculer le *Precision* de l'homophone "a". Par la suite, soit l'humain ou le modèle change l'homophone et le remplace par sa forme associée et on sait qu'il y avait vraiment une erreur (TP), soit il ne détecte pas l'erreur et

ne change rien (FN, il s'est trompé en laissant la forme incorrecte i.e. a échoué à détecter l'erreur). Le *Precision* donne la probabilité, pour un homophone donné, que le modèle ou l'humain va détecter l'erreur s'il y en a. Autrement dit, si l'homophone présent dans la phrase n'est pas de la forme correcte (et donc doit être remplacé par sa forme associée), la probailité que le modèle ou l'humain arrivera à détecter l'erreur est donnée par le *Recall*.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

La mesure du **Accuracy** prend en compte toutes 4 les situations possibles pour chaque homophone lorsqu'il est présent dans une phrase. C'est la mesure à laquelle on pense intuitivement lorsqu'on mesure un score par exemple. On compte toutes les fois où le modèle ou l'humain à donné la bonne réponse ($TP + TN$). On en déduit alors toutes les fois où le modèle ou l'humain à donné la mauvaise réponse ($FP + FN$). Puis on divise le nombre de bonnes réponses par le nombre total de phrases qui ont été présentées ($TP + TN + FP + FN$). L'*Accuracy* donne une mesure globale des performances toute situation confondu. Cependant, lorsque il y a un déséquilibre entre la fréquence des phrases incorrectes et la fréquence des phrases correctes, la mesure de l'*Accuracy* est moins informative. En effet, imaginons une situation où une erreur survient que dans 5% des cas, et on souhaite détecter l'erreur lorsqu'elle survient. Une règle simple serait de décider qu'il n'y a pas d'erreur à chaque fois, et dans ce cas on donnera la bonne réponse dans 95% des cas puisque de base 95% des cas sont exempts d'erreurs. On aura alors un *Accuracy* de 0.95 (95%) et l'on pourrait se dire que l'on a un bon système. Toutefois, le *Recall* (qui mesure la probabilité de détecter une erreur) sera de 0 (0%), puisque jamais on va prendre la décision, avec la règle simple, de dire qu'il y a une erreur - et donc quand il y a vraiment une erreur (dans 5% des cas) on ne le détectera pas - alors que le but de notre système est, avant tout, de détecter les erreurs !

2.3 Test statistique

Cette section explique brièvement la signification des tests statistiques utilisés pour évaluer les différences entre les performances humaines et celles du Transformer. Tout d'abord, il faut préciser comment les mesures sont obtenues. En ce qui concerne les performances humaines, elles ont été obtenues à travers le questionnaire en ligne. Les réponses aux 40 questions données par les répondants ont été utilisées pour le décompte des TP, FP, TN

et FN puis le calcul des Precision, Recall et Accuracy. Les performances du Transformer ont été obtenues après l'avoir testé sur 20 000 phrases. Pour chaque paire d'homophones il y avait 2 000 phrases. Dans ces 2 000 phrases, chaque homophone de la paire était présent dans 1 000 phrases. Parmi ces 1 000 phrases, dans 500 phrases sa forme était la forme correcte (on mesure ainsi le nombre de FP et TN) et, dans les autres 500 phrases il était une erreur et donc le modèle devait le détecter et le remplacer par sa forme associée (on mesure ainsi le nombre de TP et FN).

Pour avoir une idée de la pertinence du modèle, on compare ses performances avec les performances humaines. Toutefois, il nous reste à savoir quand est-ce l'écart entre les performances pourra être jugé significatif. Par ailleur, le Precicion, le Recall, et l'Accuracy sont des mesures comprises entre 0 et 1 puidqu'elles représentent des probabilités. Dans la suite de notre discussion on utilisera plutôt les pourcentages (%) pour illustrer les différences, puisque c'est une échelle qu'on a tous l'habitude d'utiliser. Pour la suite, considérons l'exemple suivant: si le modèle à un Precision de 97.25% pour un certain homophone, et les humains ont un Precision de 97.75% pour ce même homophone, est-ce que l'écart de 0.50% en faveur des humains est suffisant pour dire que l'humain est significativement meilleur que le modèle ? Pour répondre - rigoureusement - à cette question, on utilisera des tests statistiques. Pour faire court, dans un test statistique on a 2 hypothèses. Notons, le Precision du modèle p_0 et le Precision des humains p_h. La première hypothèse, appelé hypothèse nulle et noté H_0, est l'hypothèse que l'on considère vraie par défaut. Ici, H_0 est l'hypothèse que le Precision du modèle est égale au Precision des humains: $H_0 : p_0 = p_h$. La deuxième hypothèse, appelé hypothèse alternative et noté H_1, est l'hypothèse contre laquelle on test H_0. Le but du test est de voir si l'on peut rejeter ou pas l'hypothèse H_0 au profit de H_1. Par contre, il est primordial de comprendre, que si on arrive pas à rejeter H_0 au profit de H_1, ça ne veut pas dire qu'on a montré que H_0 est vraie. Tout ce que le test permet de dire, c'est est-ce qu'on peut rejeter H_0 au profit de H_1. Le test ne permet en aucun cas de prouver que H_0 est vraie. Ceci étant dis, pour chaque test, on ne peut jamais être sûr à 100% de prendre la bonne décision. Il existe toujours un risque de se tromper. Ce risque est donné par le *p-value*. Le **p-value** indique, si H_0 est vraie, alors quelle est la probabilité d'observer un écart égale ou plus grand que l'écart qu'on a observé sur nos données. Ainsi, plus le p-value est petit, plus la probabilité que H_0 soit vraie au regard de nos données est petite. On décide alors d'une

valeur limite du p-value, en dessous de laquelle on décide de rejer H_0. Par exemple, si l'on fixe la valeur limite à p-value = 0.05, alors si un test à un p-value plus petit, 0.02 par exemple, alors on va rejeter H_0, et on dira que l'on rejette H_0 et on décide de prendre H_1 avec un risque (i.e. une probabilité) de se tromper de 0.02 . Plus généralement, on peut noter p_0 la performance du Transformer pour une mesure donnée, et p_h la performance des humains pour cette même mesure. Enfin pour toutes les comparaisons, l'hypothèse alternative sera $H_1 : p_0 < p_h$ lorsqu'on va tester si les humains sont significativement meilleur que le Transformer (quand l'écart des performances est en faveur des humains), et $H_1 : p_0 > p_h$ lorsqu'on va tester si les humains sont significativement plus mauvais que le Transformer (quand l'écart des performances est en faveur du Transformer). Si, un test a un p-value supérieur au valeur limite qu'on s'est fixé, alors on ne peut ni rejeter H_0 ni accepter H_0, on dira que le test n'est pas significatif, ou encore non concluant. Dans notre cas, si un test est non significatif, on dira qu'on est indécis entre le Transformer et les humains. En effet, intuitivement, si l'écart des performances est très petit, alors il faut un grand échantillon de répondants pour pouvoir conclure, or la taille de notre échantillon est limitée par le nombre de personnes qui ont répondu au questionnaire.

3 Présentation des résultats

3.1 Résultats de l'ensemble des participants

Au total, 83 personnes ont répondu au questionnaire, avec une moyenne d'âge de 28 ans (l'écart-type est de 15 ans), un minimum de 10 ans et un maximum de 73 ans. Voir figure 1 pour des informations plus détailés sur l'âge des répondants. Le niveau d'étude des répondants quant à lui varie du collège au niveau supérieur à BAC + 5 (voir figure 2 pour plus de détails).

En termes de scores (mesurés sur 40 points, chaque bonne réponse donnant 1 point supplémentaire), le score moyen est de 36 points (l'écart-type est de 4 points), la médiane est de 38 points, le minimum est de 20 points et le maximum est de 40 points. Le Transformer a obtenu un score de 37. Pour plus de détails sur les scores, voir figure 3.

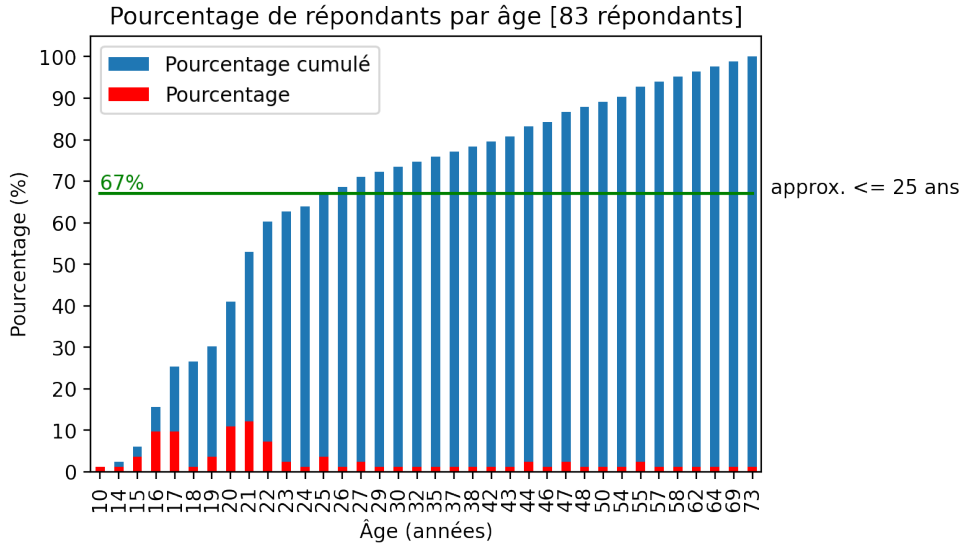

Figure 1: Détails sur l'âge

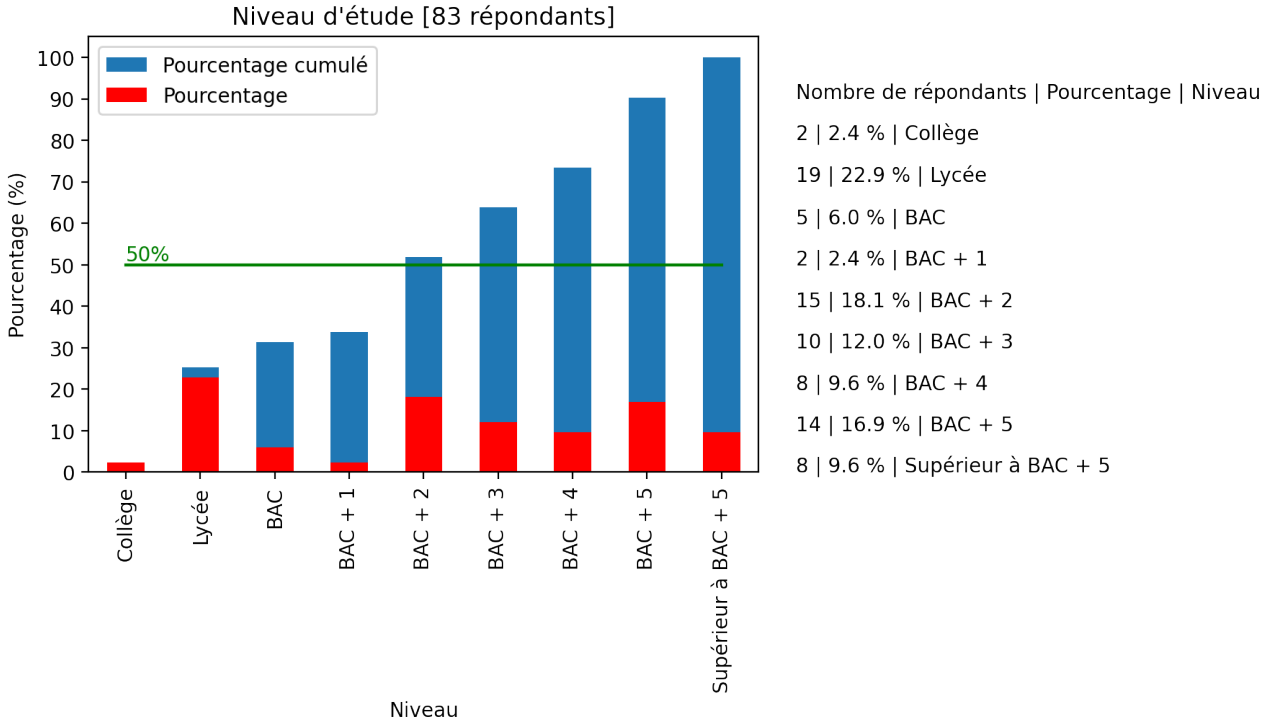

Figure 2: Détails sur le niveau d'étude

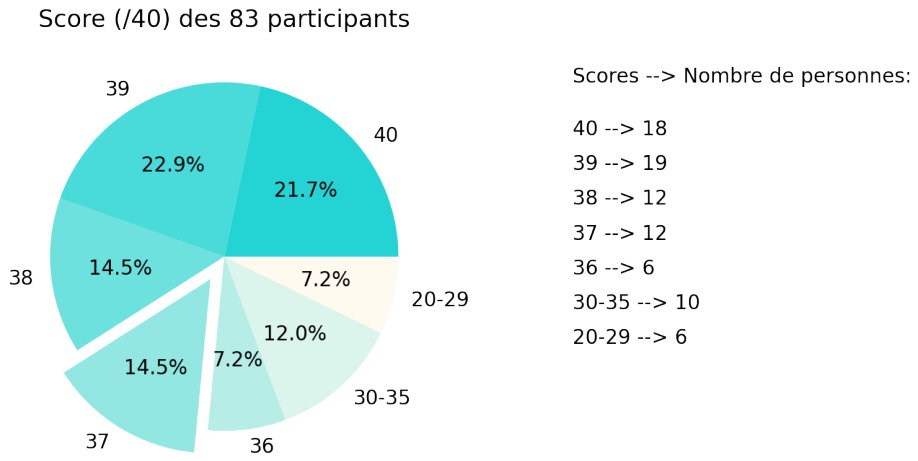

Figure 3: Les scores de tous les répondants. Le Transformer a obtenu 37 points.

Le cadre global étant en place, nous pouvons passer à la comparaison des performances du modèle avec les performances humaines. Nous utiliserons le Precision, le Recall et l'Accuracy. Les résultats sont présentés dans la tableau 3, où les écarts significatifs (p-value < 0.05) sont en gras.

L'une des critiques souvent formulées à l'égard des modèles d'apprentissage en profondeur (DL) est qu'ils doivent s'entraîner sur une grande quantité de données et qu'ils nécessitent donc également une grande puissance de calcul. Cependant, un autre objectif de ce TER, est de montrer qu'il est possible d'entraîner un modèle relativement petit et qui ne nécessite pas une grande quantité de données ni une grande puissance de calcul. Les performances que vous allez voir ont été obtenus avec un modèle qui a pris seulement *une demi heure* à converger (i.e. a appris le maximum qu'il pouvait apprendre avec les données qui lui ont été fournies) sur un PC avec GPU . En outre, le modèle a été entraîné sur *100 000 phrases* au total, le fichier contenant ces phrases occupe *6 Mo* sur le disque - soit la taille d'environ 2 photos prises avec un smartphone, ce qui est remarquablement petit. Enfin, le modèle a *431 490 paramètres*, et occupe seulement *5 Mo* sur le disque. Pour vous donner une idée, les modèles en DL, et particulièrement ceux en NLP, qui font la une des journaux grâce à leurs prouesses, sont souvent entraînés pendant plusiseurs semaines d'affilée sur des machines puissantes, avec des données allant de quelques Gigaoctets à plusieurs centaines de Gigaoctets, et ces modèles ont un nombre de paramètres allant de quelques dizaines de millions à plusieurs centaines de millions, certains atteignant même des dizaines de milliards de paramètres !

	Precision (en %)		Recall (en %)		Accuracy (en %)	
	Humains	Transformer	Humains	Transformer	Humains	Transformer
a/à	91.76	**98.31**	**93.98**	82.39	92.77	89.60
à/a	94.05	**99.53**	**95.18**	87.17	94.58	92.10
est/et	96.20	92.29	**91.57**	81.61	**93.98**	85.90
et/est	91.86	**98.70**	**95.18**	62.35	**93.37**	79.8
ces/ses	82.02	87.41	**87.95**	74.34	84.34	80.80
ses/ces	**98.59**	90.00	**84.34**	62.07	**91.57**	76.20
ce/se	96.25	**99.54**	92.77	88.10	94.58	92.50
se/ce	96.25	**99.30**	92.77	87.30	94.58	92.30
ou/où	94.74	94.40	86.75	**92.67**	90.96	92.30
où/ou	93.02	93.12	**96.39**	89.58	94.58	91.10
la/là	96.34	**99.78**	95.18	92.68	95.78	94.70
là/la	98.70	97.40	91.57	**98.19**	95.18	97.10
tout/tous	92.86	95.04	78.31	**89.09**	86.14	**91.20**
tous/tout	86.17	**94.30**	**97.59**	90.67	90.96	91.40
leur/leurs	86.57	**96.45**	69.88	**93.33**	79.52	**94.20**
leurs/leur	90.91	94.82	84.34	**95.97**	87.95	**94.80**
ceux/ce	97.53	93.50	95.18	98.19	96.39	95.00
ce/ceux	98.67	98.30	89.16	**94.48**	93.98	94.50
cette/cet	91.57	**97.57**	91.57	**96.79**	91.57	**96.6**
cet/cette	96.51	98.18	100.0	97.39	98.19	97.5
Moyenne	93.53	95.90	90.48	87.72	92.05	90.98
Avantage	1	9	8	7	3	4

Table 3: Comparaison des performances du Transformer avec celles de tous les participants. La lettre en gras au début de chaque ligne indique l'homophone dont les mesures sont reportées sur la ligne, et la forme associée est indiquée après le slash. Les résultats statistiquement significatifs (p-value < 0.05) sont en gras.

Il est également intéressant de savoir que, pour la phase phase d'entraînement et de test, toutes les phrases sont d'abord converties en minuscules, puis données en entrée

au modèle. Cela permet de limiter la taille du vocabulaire que le modèle doit utiliser, et donc réduire les ressources de calculs et le temps nécessaire pour entraîner le modèle. Le tableau 3 se lit de la façon suivante: sur la première ligne, par exemple, **a**/à signifie que, la phrase proposée au modèle ou à l'humain contient la lettre **a**. Le Precision sur cette ligne indique alors, le pourcentage du temps où le modèle ou l'humain a raison (la confiance que l'on peut avoir sur la modification apportée) lorsqu'il change le **a** en à. Le Recall indique, parmis toutes les fois où **a** devait être changé en à (car la phrase n'est pas correcte), le pourcentage de temps le modèle ou l'humain détecte l'erreur. Enfin, l'Accuracy indique le pourcentage de bonnes réponses que le modèle ou l'humain obtient lorsque **a** est présent dans la phrase proposée (toutes situations confondues). Pour la deuxième ligne, on s'intéresse aux résultats quand c'est **à** qui est donné à l'entrée du modèle ou de l'humain. Plus généralement sur chaque ligne, les résultats sont donnés par rapport à l'homophone mis en évidence en gras au début de la ligne. L'homophone après la barre oblique indique le mot avec lequel le modèle ou l'humain remplace le mot en gras (lorsqu'il le remplace, ce qui n'est pas toujours le cas).

Pour la comparaison des résultats, on pourrait seulement comparer les performances moyennes. Cependant, une mesure plus fine serait de comparer le modèle et l'humain pour chaque condition d'une paire d'homophones. Chaque paire à deux conditions (un des deux homophones est présent dans la phrase proposée dans chaque condition). Pour les 10 paires il y a donc au total 20 conditions correspondant aux 20 lignes dans le tableau 3. Les performances pris condition par condition (chaque ligne du tableau 3) nous indique que, pour le Precision, le modèle performe mieux que les humains dans 9 conditions, l'humain performe mieux que le modèle dans 1 condition, et l'écart est statistiquement non significatif dans 10 conditions. Pour le Recall, le modèle performe mieux que les humains dans 7 conditions, l'humain performe mieux que le modèle dans 8 conditions, et l'écart est statistiquement non significatif dans 5 conditions. Enfin, pour l'Accuracy, le modèle performe mieux que les humains dans 4 conditions, l'humain performe mieux que le modèle dans 3 conditions, et l'écart est statistiquement non significatif dans 13 conditions. Avec cette manière de comparer, on a 60 points de comparaisons. Ce qui ressort de ces résultats, est un net avantage (9-1) pour le modèle en ce qui concerne le Precision, des résultats serrés en ce qui concerne le Recall (7-8) et l'Accuracy (4-3). Au total le modèle remporte plus de points (20) que les humains (12), et "gagne" contre

les humains sur un score de 20-28-12, 28 indiquant le nombre de points qui n'ont pu être attribués à l'un ou l'autre parce que la différence n'était pas statistiquement significative (voir 2.3 pour un rappel sur les tests statistiques).

3.2 Discussion des résultats

Le modèle, performe globalement au moins aussi bien que les humains. Toutefois, sur certaines paires d'homophones commme ("est","et") et ("ces","ses") le modèle performe beaucoup moins bien que les humains. Cela peut s'expliquer par la nécéssité pour ces paires d'avoir un contexte plus large pour lever toute ambiguïté. En effet, la paire ("ces","ses") pose aussi des problèmes aux humains comparé aux autres paires. Pour améliorer les performances, on pourrait entraîner le modèle en présentant plusieurs phrases (au lieu d'une seule) formant un contexte plus grand. Par ailleurs, le modèle performe conidérablement mieux que les humains pour la paire ("leurs","leur") et remporte 5 points sur les 6 possibles pour cette paire.

Enfin, il convient de discuter des limites concernant l'évaluation des performances humaines. En effet, il y a plusieurs biais qui ont, sans exagérer, largement favorisé les résultats des humains.

- Le principal biais est la façon dont les réponses ont été receuillis par l'intermédiaire du questionnaire en ligne. Lorsque l'on commet une faute, on le fait en l'écrivant une première fois, puis si on se relit on peut ne pas détecter l'erreur (ou modifier un mot qui était en fait bon), et ainsi laisser l'erreur (ou introduire une nouvelle erreur) après vérification. D'une part, le questionnaire simulait davantage la 2e condition, à savoir une relecture de ce qu'on a déjà écrit, puisqu'une phrase était présentée et le participant devait juger si elle était correcte ou pas. Ce qui favorise de meilleurs performances, puisque toute l'attention est focalisée sur la détection des erreurs. D'autant plus que, dans le cadre du questionnaire, les répondants savaient qu'il y aurait des erreurs (en effet cela a été explicitement indiqué dans la présentation du questionnaire) et ont donc été encore plus vigilant. D'autre part, la verification des erreurs a été rendu encore plus facile, car les humains n'avaient que deux options de réponse, "oui, la phrase est correcte" et "non, la phrase n'est pas correcte, *il faut remplacer x par y*" où x est un homophone présent dans la phrase et y son homophone associé. Là encore, les humains savaient exactement où porter leur attention

dans la phrase grâce à l'indication, ce qui s'éloigne très largement des conditions réelles lorsqu'on corrige des erreurs lors d'une relecture. Ce qui peut potentiellement expliquer la bonne performance des humains pour la mesure du Recall qui indique la capacité à détecter une erreur lorsqu'elle est présente. Tandis que le modèle, n'a reçu en entrée que la phrase (convertie en minuscule) sans aucune autre indication, et devait sortir la phrase en y apportant des modifications/corrections si nécessaire. Ces biais sont liés à la façon dont les performances humaines ont été évaluées.

- Un second biais, mais tout aussi important, découle de la composition de l'échantillon des répondants. En effet parmis les répondants, 48% ont un niveau d'éducation supérieur à BAC+2 (et 10% ont même un niveau d'éducation supérieur à BAC+5) (voir figure 2), tandis que d'après le site de l'Insee (voir ici), seulement 23% des français ont un niveau d'étude supérieur à BAC+2. De plus, il est également indiqué que 20% des français entre 25 et 64 ans n'ont aucun diplôme (maximum le brevet des collèges), alors que dans l'échantillon des répondants, tout les répondants entre 25 et 64 ans ont minimum le BAC. Ces deux éléments combinés suggèrent que l'échantillon est biaisé en faveur des personnes très instruites. Il est raisonnable de supposer que ce biais a favorisé les performances humaines.

Avec ces deux biais en tête, on peut aussi ajouter un autre argument en faveur de la pertinence du modèle: le temps. En effet, pour évaluer les 40 phrases du questionnaire, il n'a fallu que 38 secondes (soit un peu moins d'une seconde par phrase) au Transformer - avec un score de 37/40. Tandis que, les humains prennent bien plus de temps, les retours de quelques participants ont fait état de 10 à 15 minutes pour terminer le questionnaire (le temps mis par le répondant n'a pas été mesuré) - soit 15 à 20 fois plus que le Transformer. En outre, l'implémentation du Transformer n'a pas été optimisée, ainsi, on peut diminuer encore plus le temps d'évaluation, si on devait l'utiliser, par exmple, pour une application commerciale. Pour résumer, le Transformer performe globalement mieux que les humains et prends considérablemet moins de temps que ces derniers, même lorsque les humains sont avantagés dans la réalisation de la tâche (corriger des erreurs). Dans la section sivante, les performances du Transformer sont comparés aux répondants âgés de 25 ans ou moins (tableau 4) et les répondants ayant un niveau d'éducation inférieur ou égale à BAC+2 (tableau 5).

4 Comparaison avec les répondants de 25 ans ou moins, puis avec ceux ayant un niveau d'éducation inférieur ou égal à BAC+2

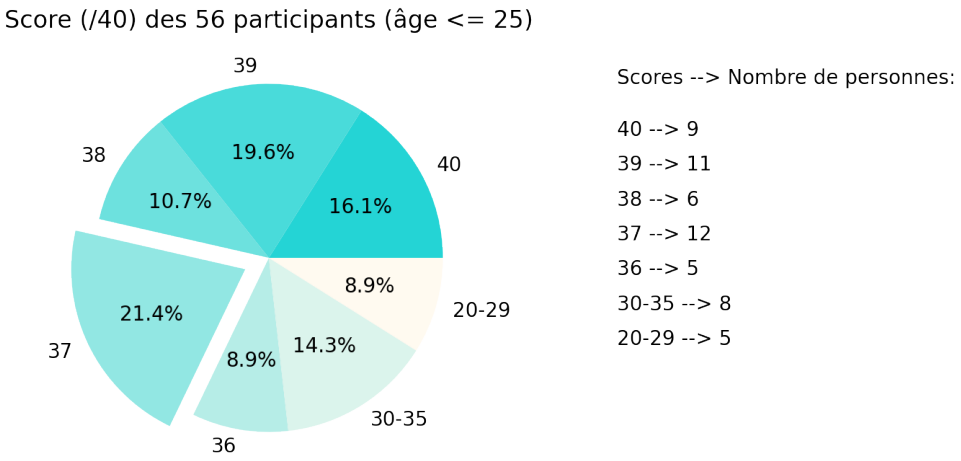

Figure 4: Les scores de tous les répondants de 25 ans ou moins. Le Transformer a obtenu 37 points.

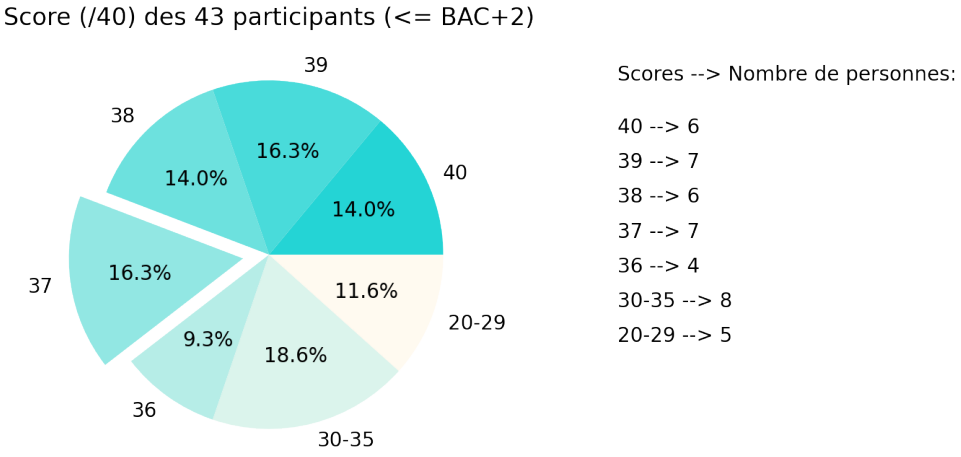

Figure 5: Les scores de tous les répondants ayant un niveau d'éducation inférieur ou égal à BAC+2. Le Transformer a obtenu 37 points.

	Precision (en %)		Recall (en %)		Accuracy (en %)	
	Humains	Transformer	Humains	Transformer	Humains	Transformer
a/à	89.66	**98.31**	**92.86**	82.39	90.35	89.60
à/a	91.38	**99.53**	94.64	87.17	92.11	92.10
est/et	96.08	92.29	87.50	81.61	**91.18**	85.90
et/est	89.93	**98.70**	**94.64**	62.35	**91.23**	79.8
ces/ses	81.36	87.41	**85.71**	74.34	82.44	80.80
ses/ces	**97.87**	90.00	**82.14**	62.07	**89.40**	76.20
ce/se	94.44	**99.54**	91.07	88.10	92.08	92.50
se/ce	98.04	99.30	89.29	87.30	92.94	92.30
ou/où	92.31	94.40	85.71	92.67	88.56	92.30
où/ou	93.10	93.12	96.43	89.58	93.86	91.10
la/là	94.55	**99.78**	92.86	92.68	92.97	94.70
là/la	98.04	97.40	89.29	**98.19**	92.94	**97.10**
tout/tous	91.49	95.04	76.79	**89.09**	84.13	**91.20**
tous/tout	83.08	**94.30**	96.43	90.67	87.74	91.40
leur/leurs	82.61	**96.45**	67.86	**93.33**	76.22	**94.20**
leurs/leur	88.46	94.82	82.14	**95.97**	85.04	**94.80**
ceux/ce	96.30	93.50	92.86	**98.19**	93.84	95.00
ce/ceux	98.00	98.30	87.50	**94.48**	92.05	94.50
cette/cet	90.74	**97.57**	87.50	**96.79**	88.57	**96.6**
cet/cette	94.92	98.18	100.0	97.39	96.50	97.5
Moyenne	92.11	95.90	88.66	87.72	89.71	90.98
Avantage	1	**8**	4	**7**	3	**5**

Table 4: Comparaison des performances du Transformer avec celles des participants âgés de 25 ans ou moins. La lettre en gras au début de chaque ligne indique l'homophone dont les mesures sont reportées sur la ligne, et la forme associée est indiquée après le slash. Les résultats statistiquement significatifs (p-value < 0.05) sont en gras.

	Precision (en %)		Recall (en %)		Accuracy (en %)	
	Humains	Transformer	Humains	Transformer	Humains	Transformer
a/à	84.78	**98.31**	90.70	82.39	87.21	89.60
à/a	88.89	**99.53**	93.02	87.17	90.70	92.10
est/et	92.50	92.29	86.05	81.61	**89.53**	85.90
et/est	85.11	**98.70**	**93.02**	62.35	**88.37**	79.8
ces/ses	82.61	87.41	**88.37**	74.34	84.88	80.80
ses/ces	**100.0**	90.00	**81.40**	62.07	**90.70**	76.20
ce/se	97.44	99.54	88.37	88.10	93.02	92.50
se/ce	97.37	99.30	86.05	87.30	91.86	92.30
ou/où	89.74	94.40	81.40	**92.67**	86.05	**92.30**
où/ou	89.13	93.12	95.35	89.58	91.86	91.10
la/là	97.50	99.78	90.70	92.68	94.19	94.70
là/la	97.30	97.40	83.72	**98.19**	90.70	**97.10**
tout/tous	88.89	95.04	74.42	**89.09**	82.56	**91.20**
tous/tout	83.67	**94.30**	95.35	90.67	88.37	91.40
leur/leurs	75.68	**96.45**	65.12	**93.33**	72.09	**94.20**
leurs/leur	91.67	94.82	76.74	**95.97**	84.88	**94.80**
ceux/ce	95.12	93.50	90.70	**98.19**	93.02	95.00
ce/ceux	97.30	98.30	83.72	**94.48**	90.70	94.50
cette/cet	87.80	**97.57**	83.72	**96.79**	86.05	**96.60**
cet/cette	93.48	98.18	100.0	97.39	96.51	97.50
Moyenne	90.80	95.90	86.40	87.72	88.66	90.98
Avantage	1	**6**	3	**8**	2	**6**

Table 5: Comparaison des performances du Transformer avec celles des participants ayant un niveau d'éducation inférieur ou égal à BAC+2. La lettre en gras au début de chaque ligne indique l'homophone dont les mesures sont reportées sur la ligne, et la forme associée est indiquée après le slash. Les résultats statistiquement significatifs (p-value < 0.05) sont en gras.