Écouter parler une IA ou comment mesurer la diversité expressive des modèles de langage automatiques

Mathias Aurand-Augier
School of Computer Engineering
TELECOM Nancy
Villers-les-Nancy, France
mathias.aurand-augier@telecomnancy.eu

Théo Hornberger
School of Computer Engineering
TELECOM Nancy
Villers-les-Nancy, France
theo.hornberger@telecomnancy.eu

Terry Tempestini
School of Computer Engineering
TELECOM Nancy
Villers-les-Nancy, France
terry.tempestini@telecomnancy.eu

***Résumé*—The abstract goes here.**

I. Introduction

This demo file is intended to serve as a "starter file" for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

mds
August 26, 2015

A. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

II. Motivation

Considérons un professeur de français corrigeant une rédaction produite par un élève sur un sujet quelconque, comment évalue-t-il la qualité de cette rédaction ? Celui-ci dirait sûrement qu'il se base sur la richesse du vocabulaire, la syntaxe, la cohérence, la pertinence, la clarté, la concision, la précision, la variété des structures de phrases, autant de concept relativement difficile à mesurer quantitativement et souvent basé sur l'interprétation de chacun.

Pourtant, la compréhension de ces concepts conditionnent dans une certaine mesure la faculté de l'élève à progresser, comment peut il à l'avenir améliorer sa "cohérence" si la définition du terme n'est pas claire à ses yeux.

En ce sens, remplaçons le professeur par un programme informatique et l'élève par un modèle de langage automatique, comment un programme pourrait il évaluer la qualité d'un texte produit par un autre programme ? Il faudrait pour cela définir des critères quantitatif de "qualité" d'un texte.

Une méthode possible d'évaluation pourrait se baser sur des références : il ne s'agirait alors pas d'évaluer la qualité d'un texte en lui même, mais plutôt de se demander si un texte semble "de meilleure qualité" qu'un autre. Pour illustrer ce principe, nous avons à notre disposition 5 générateurs de texte, chacun ayant une qualité de sortie différente. Chaque générateur produit un commentaire relatif à une image donnée. Le but est de mettre en évidence certains signes de qualités des textes par rapport aux autres.

III. La notion de "qualité" d'un texte

A. Richesse du vocabulaire

Pour évaluer la "qualité" d'un texte, on peut commencer par évaluer la richesse du vocabulaire. Par richesse du vocabulaire, on peut entendre plusieurs choses, en effet plusieurs méthodes peuvent être utilisé pour évaluer la richesse du vocabulaire. Dans un premier temps, on peut compter le nombre de mots différents généré par chaque intelligence artificielle.

On obtient les résultats suivants :

```
1    Number of words in BART_1: 433
2    Number of words in BART_2: 483
3    Number of words in T5_1: 295
4    Number of words in T5_2: 310
5    Number of words in FST: 194
6
7    Number of words in total: 863
```

Avec les résultats obtenus, on observe que BART_2 et BART_1 ont un lexique plus variés, suivi de T5_2, T5_1 et FST. Cela pourrait indiquer que BART_2 et BART_1 ont un vocabulaire plus riche que les autres, et qu'elles emploient alors des mots plus variés.

Néanmoins, cette méthode n'est que très superficielle, car elle ne prend pas en compte la fréquence d'apparition des mots ni la longueur de la phrase.

Aussi tentons d'évaluer des caractéristiques statistiques basiques comme la moyenne, la médiane et l'écart type du nombre de mots par phrase générée par chaque

IA. Cela pourrait éventuellement expliquer la richesse du vocabulaire de chaque IA.

En effet, la longueur d'une phrase peut cacher une certaine complexité syntaxique, ainsi qu'un contenu explicatif de l'image commenté plus fourni : le fond serait alors de meilleur qualité. Voyons ce que cela donne :

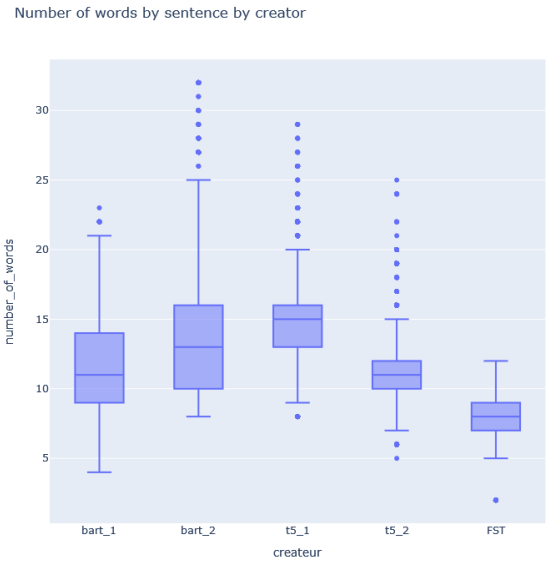

FIGURE 1. Boîte à moustache de la longueur des phrases générées par chaque IA

Les résultats sont très interessants puisque l'on peut voir que la génération T5_1 a les phrases les plus longues en moyenne et pourtant un des vocabulaire le plus pauvre, ce qui pourrait être le signe d'une répétition accru des mêmes mots. Les différences de longueurs entre BART_1 et BART_2 sont également interessantes, puisqu'elle pourrait expliquer le différentiel de vocabulaire entre les deux modèles. Ensuite, la longueur des phrases de FST est la plus faible, ce qui pourrait expliquer la pauvreté de son vocabulaire.

Il peut être aussi interessant d'étudier la fréquence d'utilisation des mots les plus fréquents. On peut pour cela utiliser un diagramme à barres empilées (on met ainsi en évidence la contribution de chaque IA à l'utilisation globale de chaque mot).

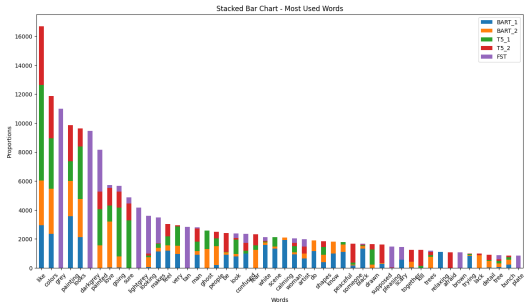

FIGURE 2. Matrice de similarité entre les phrases générées par BART_1

Un fait particulièrement interessant visible sur ce diagramme est que certains des mots les plus fréquents tel que "grey", "darkgrey", "lightgrey" ne sont utilisé que par FST. Cela révèle une focalisation excessive sur la couleur grise qui n'est apriori une couleur plus présente en générale dans les oeuvres. La pertinence des réponses de FST peut donc remise en question sachant qu'elle semble concentrée sur les mêmes caractéristiques, cela peut également expliquer la pauvreté du vocabulaire chez FST.

B. Diversité syntaxique au sein d'une même IA

Une manière interessante de mesurer la capacité d'une intelligence à générer des réponses différentes est de comparer ses propres réponses entre elles : si les phrases sont globalement proche, cela signifie que l'IA a tendance à répéter les mêmes structures de phrases, et donc à manquer de diversité syntaxique. Voici un exemple :

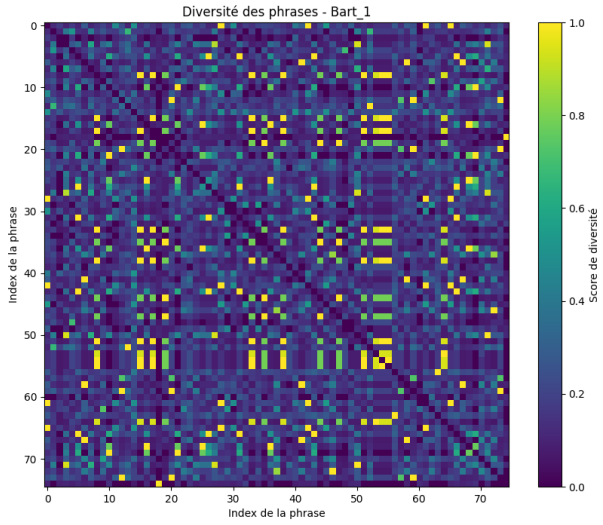

FIGURE 3. Matrice de similarité entre les phrases générées par BART_1

On remarque ici que la matrice de diversité est relativement sombre, donc proche de 0, ce qui signifie que les phrases générées par BART_1 sont relativement différentes les unes des autres, démontrant ainsi une certaine capacité à renouveler ses structures phrastiques.

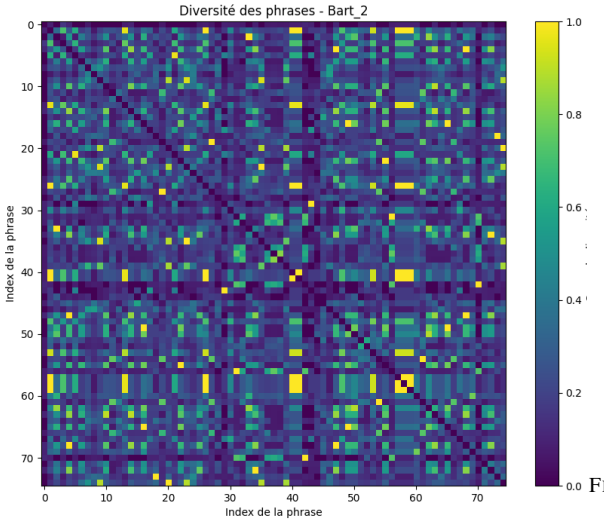

FIGURE 4. Matrice de similarité entre les phrases générées par BART_2

Ce qui est remarquable pour cette matrice est qu'elle est globalement plus claire que la précédente mais reste néanmoins assez sombre : les phrases sont donc relativement différentes.

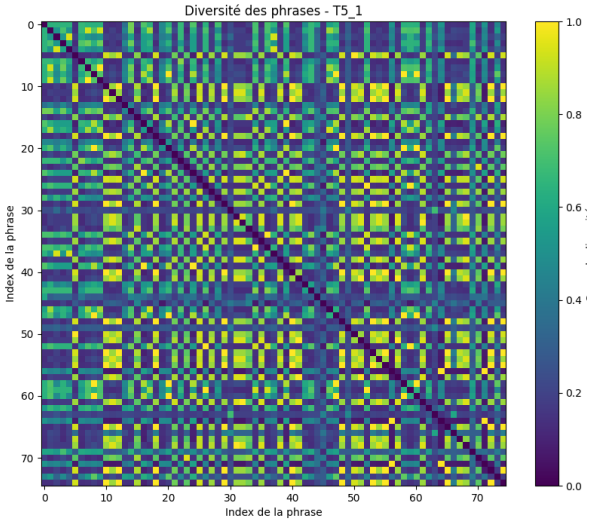

FIGURE 5. Matrice de similarité entre les phrases générées par T5_1

En revanche, ici il est clair que la matrice a une couleur beaucoup plus claire, ce qui montre que les réponses générées par T5_1 sont répétitives. Pourtant, quelques points sombres montre qu'elle est capable de se renouveler mais cela semble ici ponctuel et non récurrent.

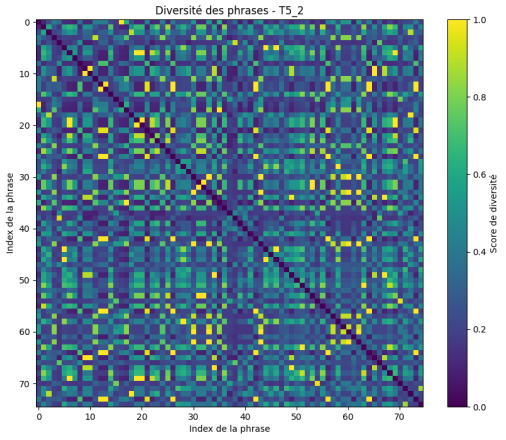

FIGURE 6. Matrice de similarité entre les phrases générées par T5_2

La matrice de T5_2 est très similaire à celle de T5_1, ce qui montre que T5_2 a également tendance à répéter les mêmes structures de phrases.

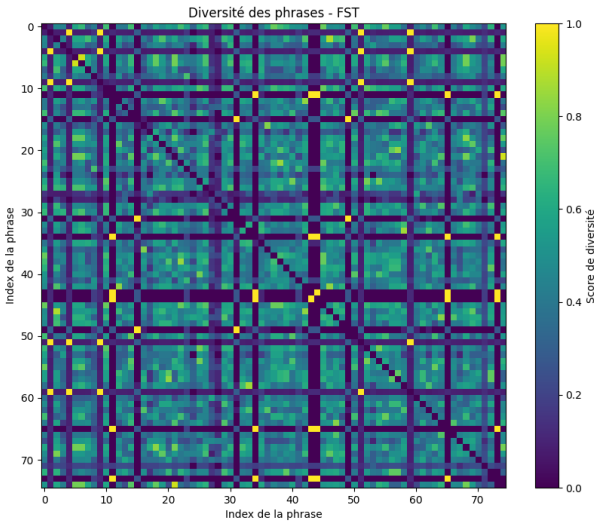

FIGURE 7. Matrice de similarité entre les phrases générées par FST

De même, la matrice de FST est très similaire à celle de T5_1 et T5_2, ce qui montre que FST a également tendance à répéter les mêmes structures de phrases.

C. Mesure de similitude entre les différentes intelligences artificielles

D. Mesure de similitude au sein des phrases générées par une même IA

E.

F.

G.

IV. Introduction

V. Analyse de la proximité des phrases générées par les IA

Dans un premier temps, nous avons effectué une comparaison des réponses des IA concernant une image donnée. L'objectif principal est de déterminer le degré de similarité ou de divergence entre les phrases générées par les différentes IA. Cette comparaison permet non seulement de mesurer la proximité entre les IA deux à deux, mais aussi de déterminer si certaines IA se distinguent des autres.

Cependant, il est important de noter que ces mesures ne reflètent pas la "qualité" intrinsèque d'une phrase donnée. Elles ne fournissent qu'une indication de la proximité entre les phrases générées par les différentes IA. Néanmoins, ces mesures sont utiles pour étudier le comportement des IA et identifier d'éventuelles similitudes ou différences marquées entre elles.

A. Utilisation de la métrique ROUGE

La première métrique utilisée dans notre étude est ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Cette technique compare la similarité entre une phrase de référence et celle générée par l'IA. Elle se base sur le comptage des chevauchements de mots (appelés n-grammes) ou de groupes de mots, selon les paramètres définis. La métrique ROUGE utilise le rappel, la précision et la F-mesure pour évaluer la similarité.

Dans notre cas, nous avons utilisé ROUGE pour comparer les mots simples et les groupes de mots (en utilisant ROUGE-N). Le choix de la phrase de référence a été fait en prenant la phrase de chaque modèle et en la comparant à toutes les autres phrases générées par les différentes IA. Cette approche nous a permis d'obtenir une matrice de comparaison, où chaque IA est confrontée aux autres, et qui nous donne la proximité des phrases générées pour une image donnée.

Les résultats présentés dans la figure 11 illustrent la comparaison effectuée à l'aide de la métrique ROUGE-1 sur un échantillon de 1000 phrases. On peut observer que les réponses générées par les modèles t5_2 et t5_1 sont très similaires (couleurs bleu), avec un score de similarité élevé (indice de 0.6). En revanche, les réponses générées par les modèles bart et FST sont nettement différentes, avec un score de similarité faible, ce qui indique l'absence de lien significatif entre leurs réponses respectives.

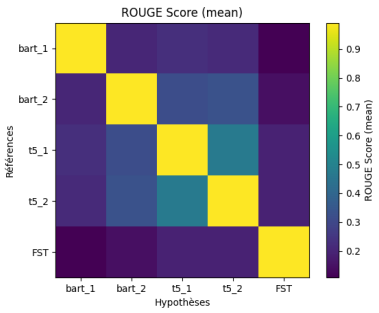

FIGURE 8. Moyenne de la métrique ROUGE sur un échantillon de 1000 phrases

B. Analyse des résultats avec ROUGE-N

Dans cette section, nous avons calculé la moyenne des résultats de la métrique ROUGE-4 sur un ensemble de 1000 images commentées, ce qui permet de comparer les 4-grammes entre les différentes IA.

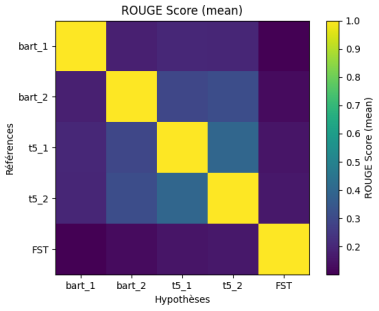

FIGURE 9. Résultats de la métrique ROUGE-4 sur 1000 phrases

Les résultats présentés dans la figure 9 confirment les tendances observées précédemment avec ROUGE-1. Les modèles t5_1 et t5_2 génèrent des phrases très similaires, ce qui peut s'expliquer par le fait qu'ils utilisent souvent les mêmes structures de génération.

C. Analyse avec les métriques BLEU et METEOR

En plus de la métrique ROUGE, nous avons également calculé les métriques BLEU et METEOR pour évaluer la similarité entre les phrases générées par les différentes IA.

La métrique BLEU (Bilingual Evaluation Understudy) mesure la qualité des reformulations, en se basant sur la précision des N-grammes tout en pénalisant la brièveté. Toutefois, cette métrique a tendance à pénaliser les phrases qui utilisent des mots moins courants. Dans notre étude, nous avons considéré une phrase de référence et nous avons comparé les phrases générées par les différentes IA pour une même image.

La métrique METEOR (Metric for Evaluation of Translation with Explicit ORdering) se base sur la moyenne harmonique et prend en compte la

correspondance radicale et synonymique. Cette métrique est généralement plus précise que BLEU.

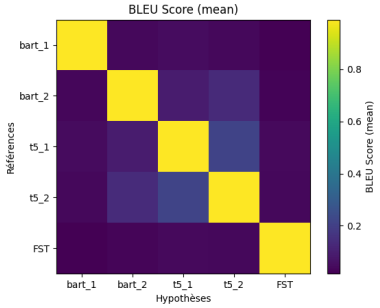

FIGURE 10. Moyenne des scores BLEU pour les phrases générées par les différentes IA

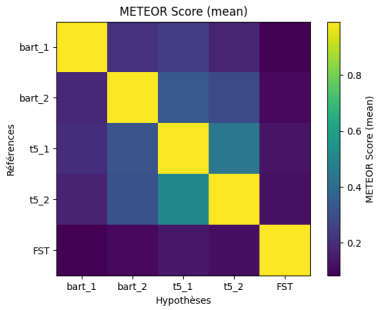

FIGURE 11. Moyenne des scores METEOR pour les phrases générées par les différentes IA

Nous avons réalisé ces mesures sur un grand ensemble de phrases, ce qui nous a permis de moyenner les résultats et d'en tirer des conclusions générales. Les résultats obtenus avec les métriques BLEU et METEOR sont similaires à ceux obtenus avec ROUGE. On peut observer que les modèles t5_1 et t5_2 génèrent des phrases très proches pour une image donnée, tandis que les modèles bart_1 et FST se démarquent avec des phrases moins similaires.

D. Conclusion de l'analyse

En résumé, notre analyse basée sur les métriques ROUGE, BLEU et METEOR a permis de mettre en évidence des similarités et des différences entre les phrases générées par les différentes IA pour une même image. Les modèles t5_1 et t5_2 génèrent des phrases très proches, tandis que les modèles bart_1 et FST se distinguent par des phrases moins similaires. Ces résultats fournissent des indications précieuses sur le comportement des différentes IA et peuvent aider à sélectionner le modèle le plus adapté à une tâche spécifique.

VI. CONCLUSION

The conclusion goes here.

ACKNOWLEDGMENT

The authors would like to thank...

RÉFÉRENCES

[1] H. Kopka and P. W. Daly, A Guide to LaTeX, 3rd ed. Harlow, England : Addison-Wesley, 1999.