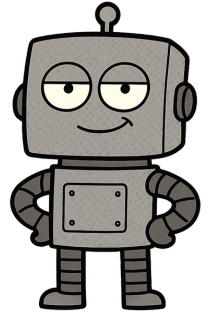


Anti-surprise : Une métrique complémentaire pour évaluer l'apprentissage lexical des (grands) modèles de langue

Nazanin Shafiabadi¹ & Guillaume Wisniewski²

1. ISIR, 4 place Jussieu, 75 005 Paris, France

2. LLF, 8 Rue Albert Einstein, 75 013 Paris, France



TL;DR

L'**anti-surprise** dévoile comment les modèles de langue évitent les faux pas lexicaux—complétant le récit que la surprise seule ne peut raconter.

1

INTRODUCTION

Les modèles de langue sont souvent évalués en analysant leurs courbes de **surprise**, qui mesurent la probabilité d'un mot dans un contexte donné. Toutefois, cette approche néglige un aspect essentiel : la capacité du modèle à **éviter** les erreurs de lexicalisation.

Nous introduisons une nouvelle mesure, l'**anti-surprise**, qui permet de compléter cette évaluation et d'obtenir une meilleure compréhension des dynamiques d'apprentissage des modèles de langue.

2

DÉFINITIONS CLÉS

Surprise : Indique dans quelle mesure un mot est attendu dans son contexte.

Anti-surprise : Évalue la capacité du modèle à ne pas utiliser un mot dans un contexte inapproprié.

3

MÉTHODOLOGIE

Cadre expérimental

- **Modèle** : **MultiBERT**, évalué à différents stades d'apprentissage.
- **Corpus** : Extraction de mots cibles à partir de **WikiText-103**, avec génération de contextes appropriés (c^+) et inappropriés (c^-).
- **Métriques analysées** : Courbes d'évolution de la surprise et de l'anti-surprise, modélisées au niveau lexical par régression linéaire pour capturer la tendance générale propre à chaque mot.

4

EXEMPLE

Mot cible : "contract"

c^+ : "They signed a [MASK]. It was legally binding."

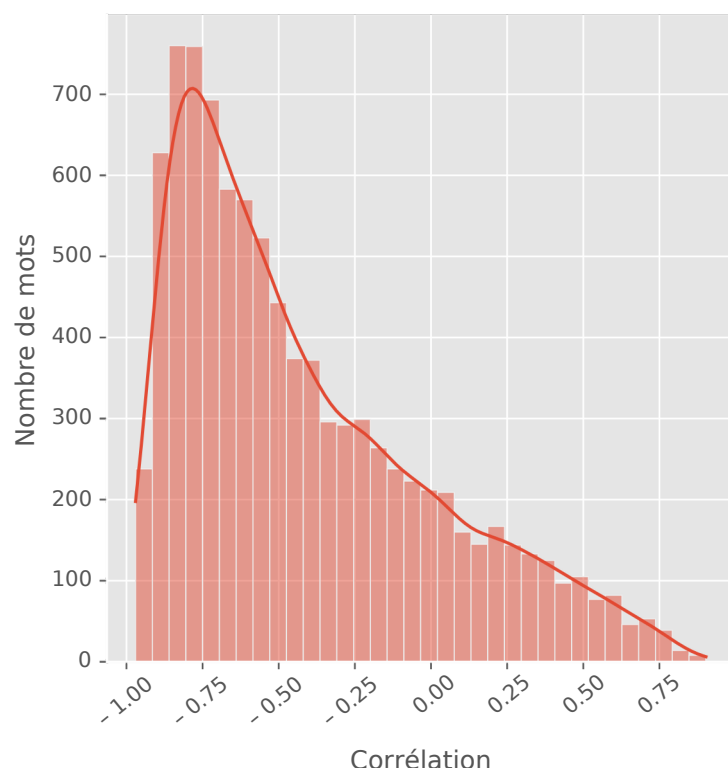
c^- : "They signed a contract. It was [MASK] binding."

ANTI-SURPRISE	SURPRISE	
	croissante	décroissante
croissante	3%	91,1%
décroissante	0,3%	5,4%

Les moyennes masquent la diversité : $\approx 8,7\%$ des mots suivent des trajectoires atypiques, d'où l'intérêt d'une analyse mot à mot.

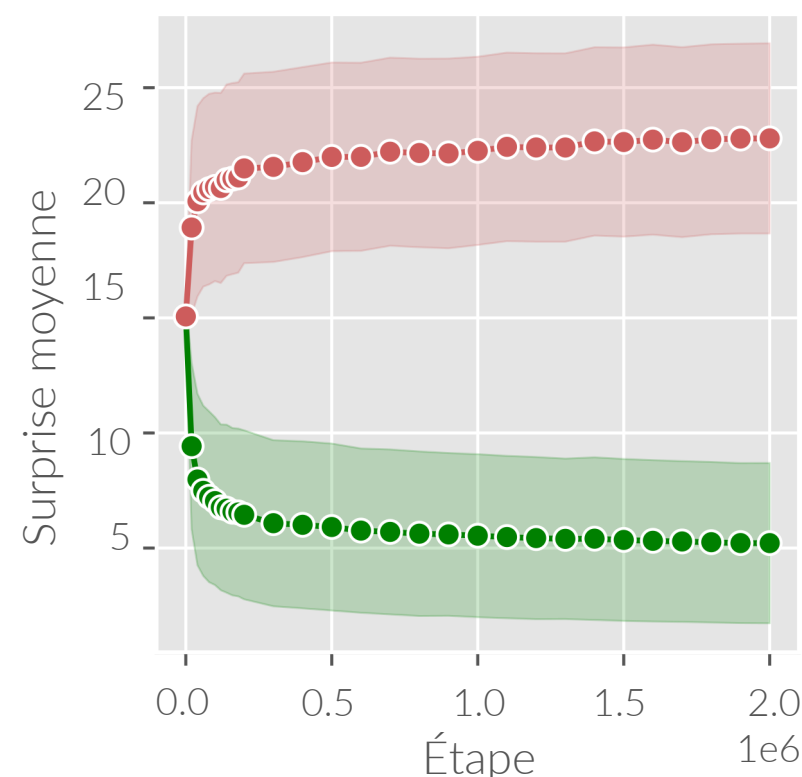
5

RÉSULTATS



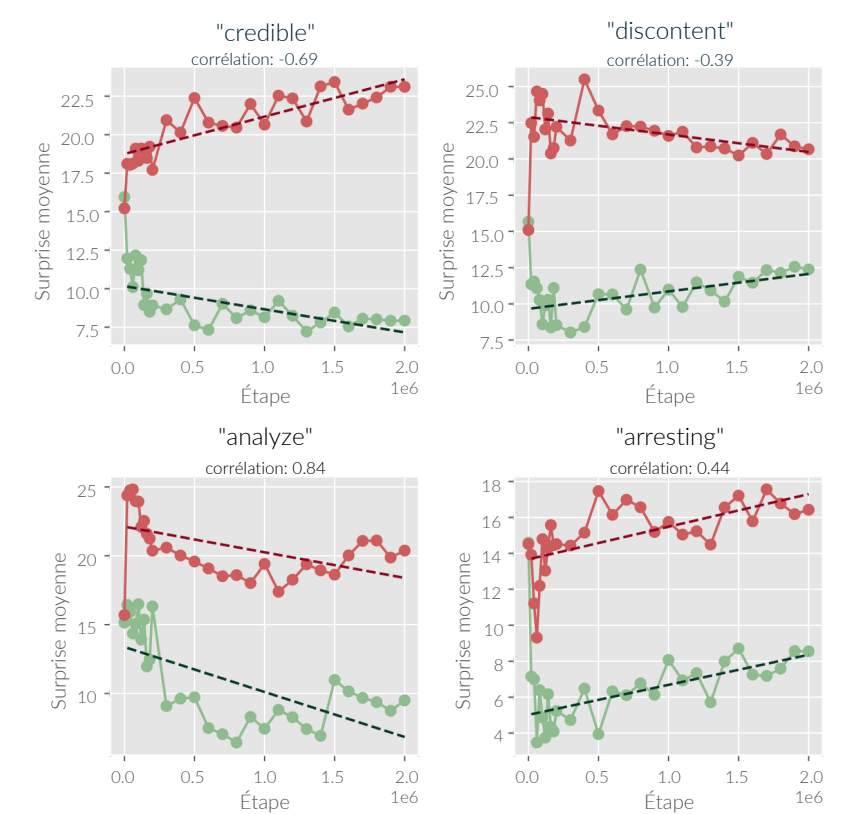
Corrélation surprise/anti-surprise :

Une forte corrélation négative entre les deux métriques suggère qu'un meilleur usage des mots est souvent associé à une meilleure capacité d'évitement des contextes erronés.



Tendance générale :

La surprise **diminue** au fil de l'apprentissage, tandis que l'anti-surprise **augmente**, indiquant une amélioration du modèle dans l'évitement des erreurs lexicales.



Analyse individuelle des mots :

L'acquisition du lexique varie selon les mots ; certains suivent des trajectoires d'apprentissage cohérentes, tandis que d'autres présentent des tendances atypiques.

6

CONCLUSION

L'**anti-surprise** enrichit l'analyse de l'acquisition lexicale en permettant une vision plus complète des performances d'un modèle. L'intégration de cette métrique ouvre des perspectives pour l'amélioration des stratégies d'apprentissage et de généralisation des modèles de langue.

REMERCIEMENTS

Nous remercions Benoît Crabbé et Timothée Bernard pour leurs retours précieux sur une version antérieure de ce travail. Ce projet a été partiellement financé par les projets Diagnostic (AID, n° 2022 65 007) et DeepTypo (ANR-23-CE38-0003-01).

CONTACT

Nazanin Shafiabadi
Doctorante, ISIR, Sorbonne Université
nazanin.shafiabadi@isir.upmc.fr



PAPER