



Normalisation lexicale de contenus générés par les utilisateurs sur les réseaux sociaux

Lydia Nishimwe

ALMAAnCH, Inria Paris

lydia.nishimwe@inria.fr

PR[AI]RIE

PaRis Artificial Intelligence Research InstitutE

6 juin 2023



- Soumission : RECITAL 2023
- 2^{ème} année de thèse
 - Sujet : **Traduction automatique neuronale robuste**
 - Encadrants : Rachel Bawden et Benoît Sagot
 - Financement : PRAIRIE
 - Équipe ALMAnaCH, Inria Paris
 - Sorbonne Université





1. Une relation amour-haine : le TAL et les **C**ontenus **G**énérés par les **U**tilisateurs

Les Contenus Générés par les Utilisateurs (CGU)

Phénomènes transversaux

Phénomènes ergographiques

demain c'est l'ete

tous mes examen son normaux

je suis oqp

je sé

a 2m1

atu compris ?

lamour

c a dire

nimp

mdr

stp

N. U. L.

Marques d'expressivité

superrrr !!!!

<3



!d10t

Influence de langues étrangères

Cette fête a l'air fun, let's go !

liker

burkini

(Seddah et al., 2012)
(Sanguinetti et al., 2020)

L'impact négatif de CGU sur le TAL

Exemple (Modèles HuggingFace)

Source Sentence

see you tomorrow

Similarité sémantique

see u tomorrow

see u 2morrow

c u tomorrow

c u 2morrow

c u 2moro

Traduction anglais-français

0.911

Rendez-vous demain

0.673

A voir demain

0.641

Voir le 2 de demain

0.457

C à 2 demain

0.189

C à 2 m2

<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

<https://huggingface.co/spaces/Iker/Translate-100-languages>



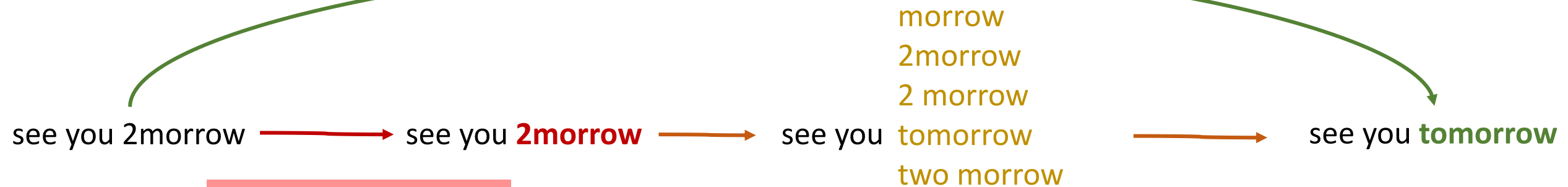
2. La normalisation lexicale : le chevalier blanc ?

Méthodes (cf. état de l'art de l'article)

1. Adaptation des modèles de TAL aux données CGU (ex. BERTweet) **+**

2. Normalisation lexicale des données CGU en amont des modèles de TAL **✓**

Traduction



Détection :

- classifieur
- modèle de langue

Génération et sélection de candidats :

- lexique
- distance lexicale
- similarité phonétique
- règles
- modèle de langue

Évaluation (van der Goot, 2019)

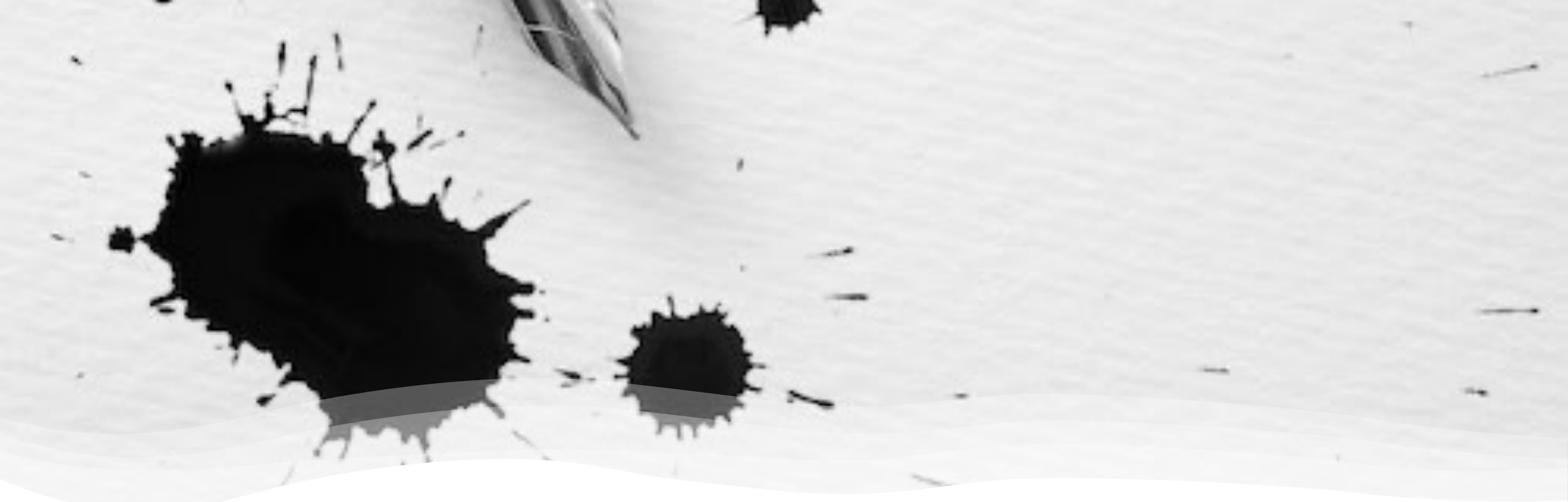
Métriques

- WER (*Word Error Rate*)
- SER (*Sentence Error Rate*)
- Taux de couverture (*Coverage*)
- Exactitude et précision sur les mots hors-vocabulaire (*OOV*)
- **Exactitude, précision, rappel, F1**
- **BLEU** (métrique de traduction)
- **ERR (*Error Reduction Rate*)**

Limites

- Métriques trop **complexes**
- Métriques trop **difficiles à interpréter**
- **Différentes méthodes ↔ différentes métriques ↔ différentes données**

$$\text{ERR} = \frac{\% \text{exactitude} - \% \text{mots non normalisés}}{100 - \% \text{mots non normalisés}}$$
$$= \frac{TP - FP}{TP + FN}$$



3. La normalisation lexicale : une tâche difficile

Étude expérimentale préliminaire

Données

LexNorm2015

Jeu de données	# tweets	% mots normalisés	dont		
			% 1-1	% 1-N	% N-1
Entraînement	2950	8,85	73,25	26,55	0,20
Test	1967	9,40	73,92	25,68	0,40

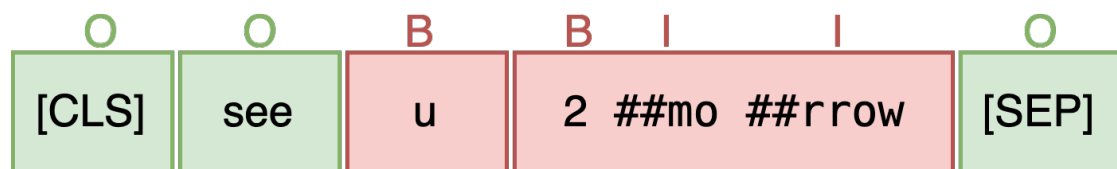
Modèles

Pré-entraînés

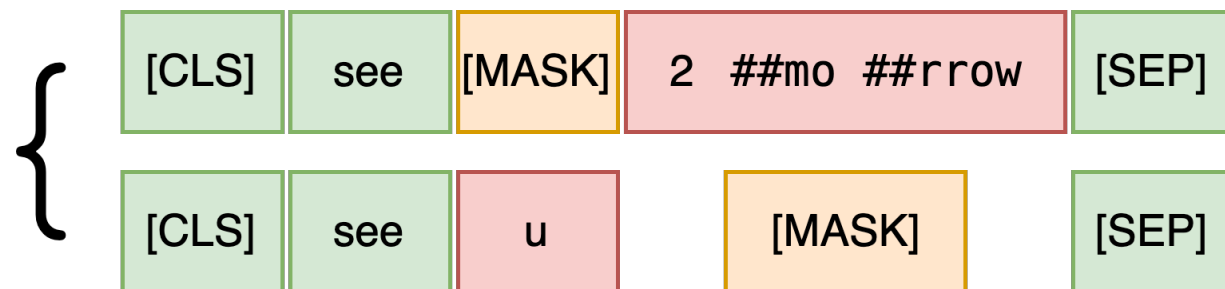
Modèle	Données d'entraînement	Taille du vocabulaire
BERT	<i>BookCorpus, Wikipedia</i>	30 522
RoBERTa	= BERT + <i>CC-News, OpenWebText, Stories</i>	50 265
ELECTRA	= BERT	30 522
BERTweet	<i>Twitter</i>	64 000
Twitter RoBERTa	= RoBERTa + <i>Twitter</i>	50 265

Modèle : détecteur-correcteur

Détection



Masquage



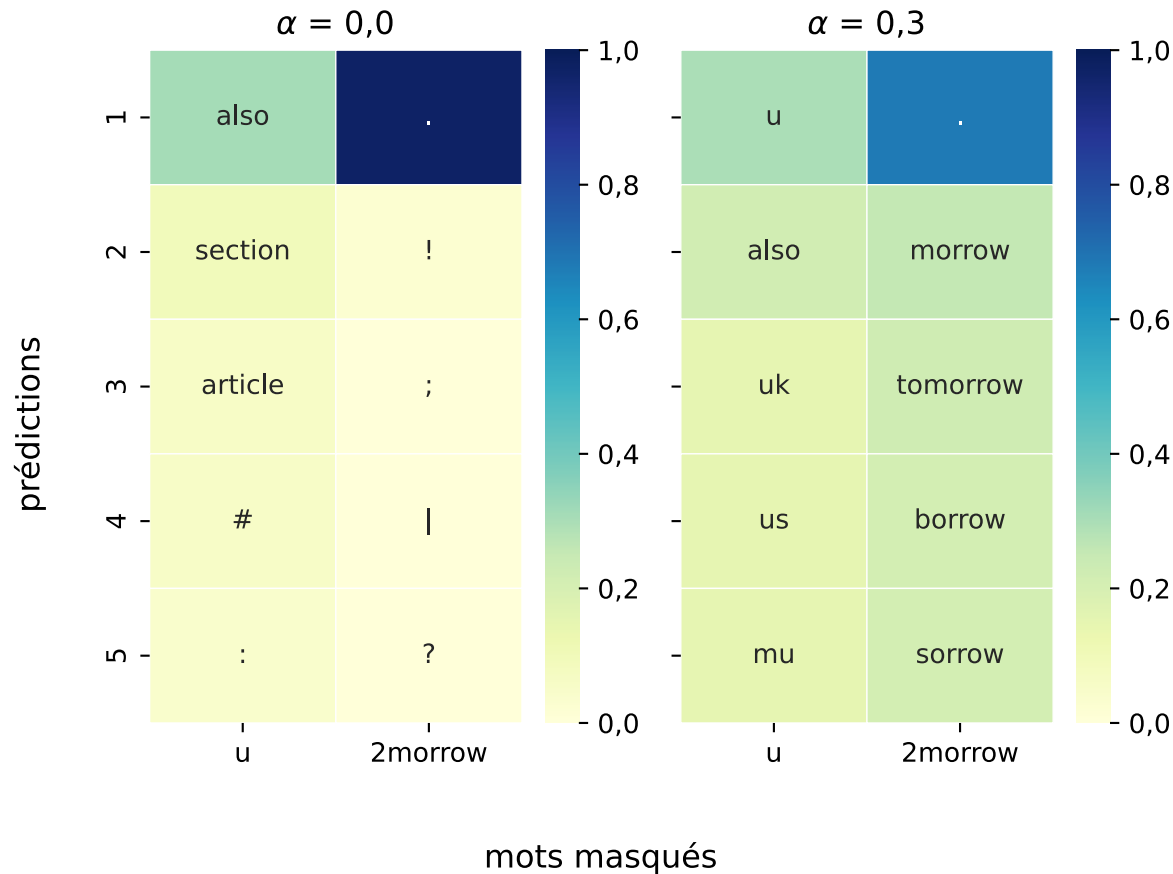
Correction

$$\hat{x}_i = \arg \max_{\mathcal{V}} \left[(1 - \alpha) \text{MLM}(\bar{\mathbf{x}}) + \alpha (1 - \text{Lev}_{\text{norm}}(x_i, \mathcal{V})) \right]$$

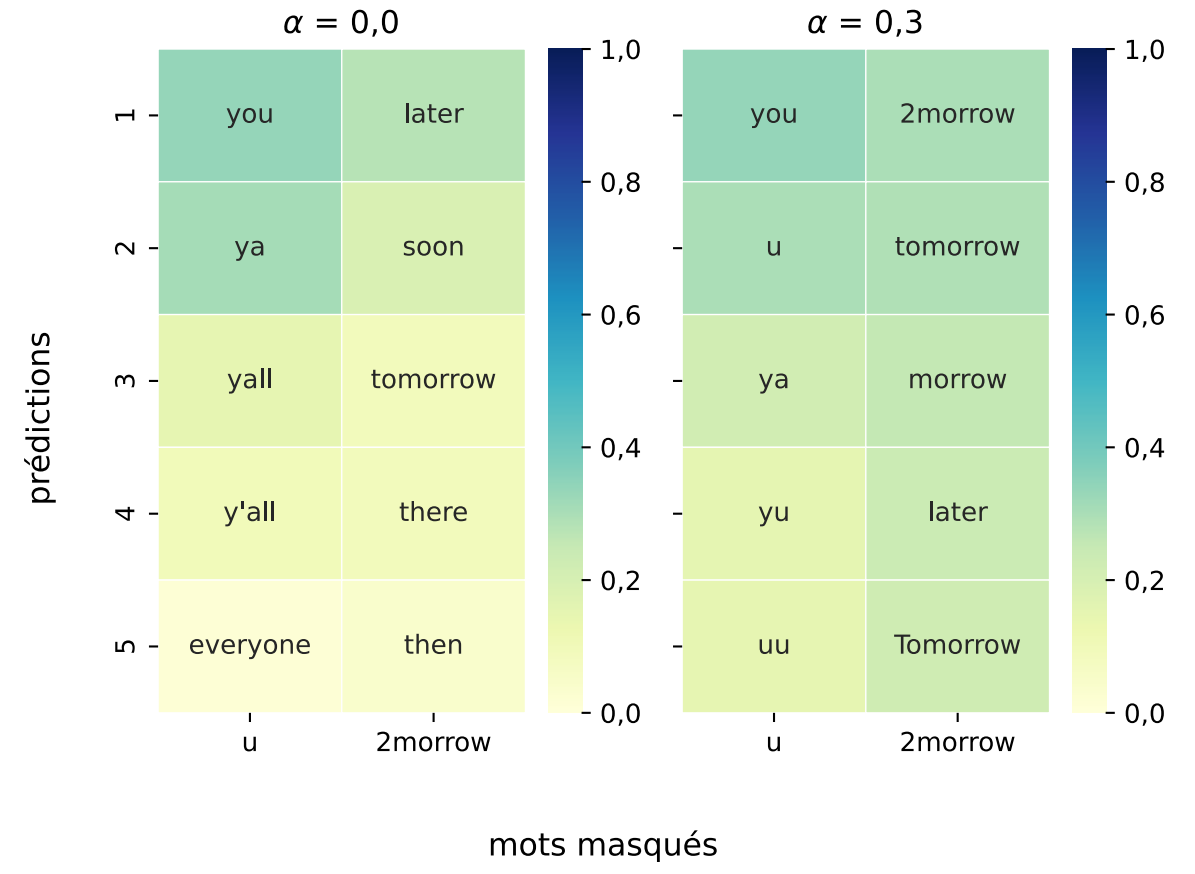
modèle de langue distance d'édition

Analyse qualitative : BERT vs. BERTweet

Détecteur BERT affiné



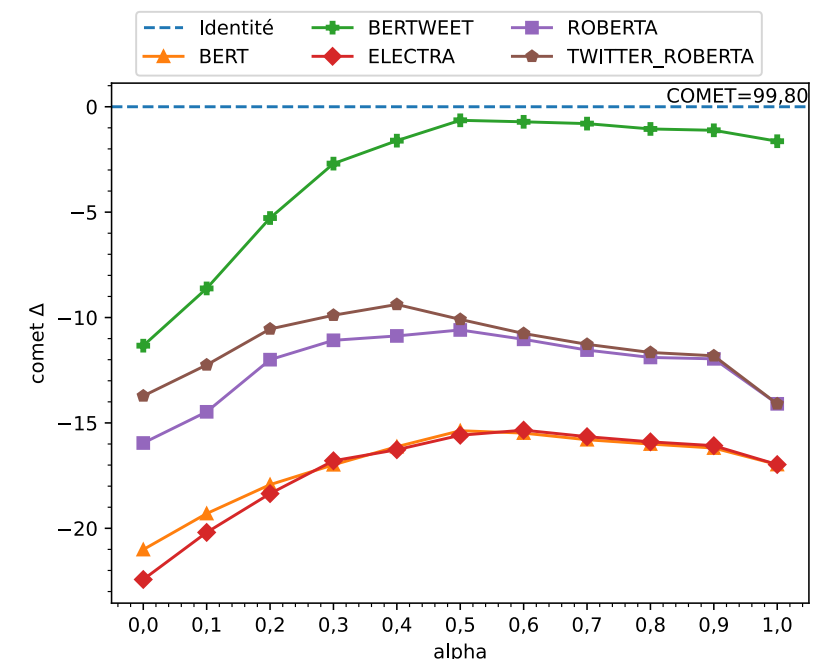
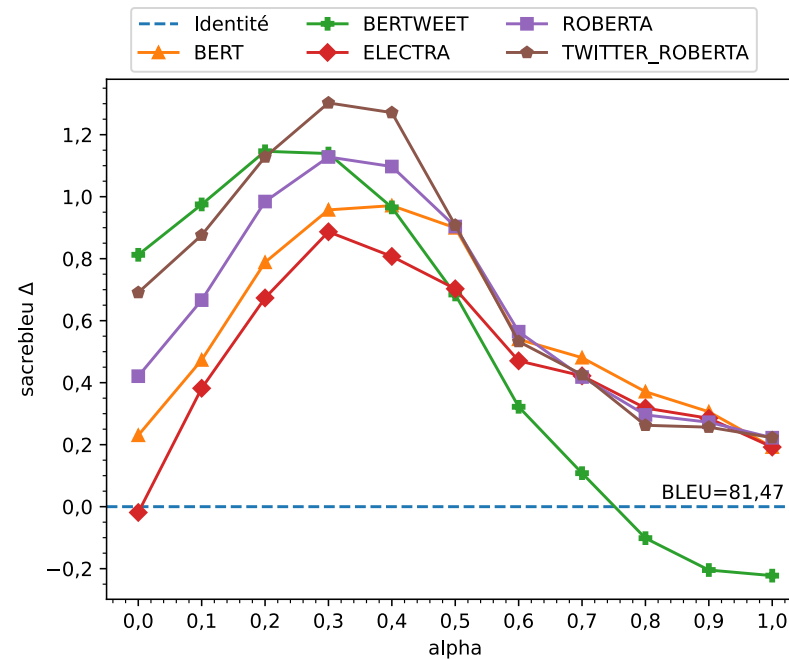
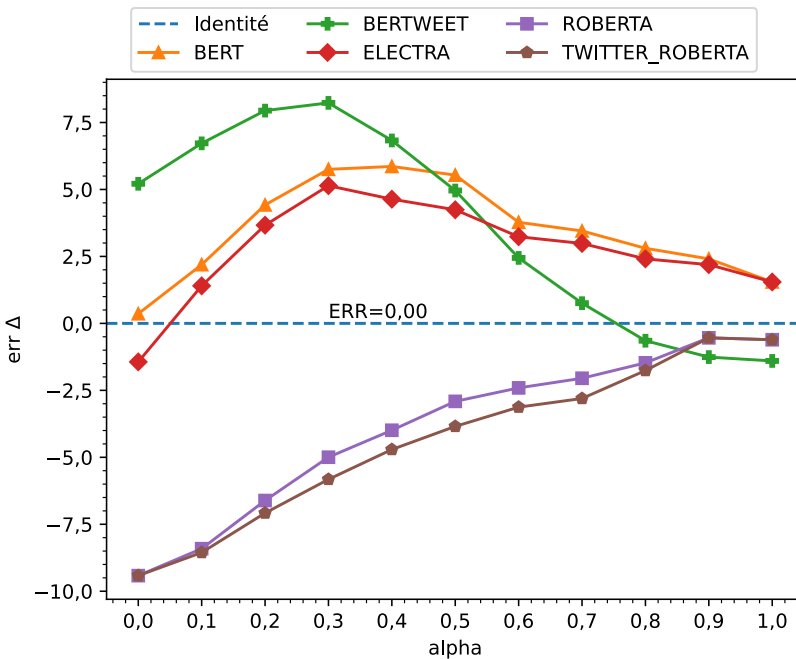
Correcteur BERT



Correcteur BERTweet

Analyse quantitative

Détecteur : exactitude 97,82%, précision 90,14%, rappel 86,41%, F-mesure 88,24 %



$$ERR = \frac{TP - FP}{TP + FN}$$

BLEU

COMET

Conclusion

1. Problématique :

les modèles de TAL ont du mal avec les CGU non standards

2. Solution (?) :

la normalisation lexicale en amont des modèles de TAL

3. Mais...

c'est une tâche difficile à réaliser et à évaluer !

Amélioration du modèle

- Stratégies de masquage plus larges (*whole-word/span masking*)
- Module de similarité phonétique
 - Normalisation autorégressive
- Apprendre la valeur optimale de α

...

Évaluation

Métrique BLEU + COMET ?

Perspectives



Merci !
Questions ?


Bibliographie

- Goot, Rob van der. « Normalization and parsing algorithms for uncertain input ». *PhD thesis*. University of Groningen, 2019.
https://pure.rug.nl/ws/portalfiles/portal/78256480/Complete_thesis.pdf.
- Sanguinetti, Manuela, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, et Amir Zeldes. « Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies ». In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5240-50. Marseille, France: European Language Resources Association, 2020.
<https://aclanthology.org/2020.lrec-1.645>.
- Seddah, Djamé, Benoit Sagot, Marie Candito, Virginie Moulleron, et Vanessa Combet. « The French Social Media Bank: a Treebank of Noisy User Generated Content ». In *Proceedings of COLING 2012*, 2441-58. Mumbai, India: The COLING 2012 Organizing Committee, 2012.
<https://aclanthology.org/C12-1149>.



Annexes

L translate "see you tomorrow" to french

 "See you tomorrow" can be translated to French as "À demain".




L translate "see u tomorrow" to french

 "See you tomorrow" can be translated to French as "À demain" or "A demain" (informal).



L translate "see u 2morrow" to french

 "See you tomorrow" can be translated to French as "À demain" or "A demain" (informal).
However, if you want to translate "see u 2morrow" with the numerical substitution for "tomorrow," you can say "À 2m1" or "A 2m1" (informal).



L translate "c u tomorrow" to french

 "C u tomorrow" can be translated to French as "À demain" or "A demain" (informal).



L translate "c u 2morrow" to french

 "C u 2morrow" can be translated to French as "À demain" or "A demain" (informal).




L translate "c u 2moro" to french

 "C u 2moro" can be translated to French as "À demain" or "A demain" (informal).



ChatGPT

L translate "wen ur at a restaurant nd u c ur food comin" to french

 "When you're at a restaurant and you see your food coming" can be translated to French as
"Quand vous êtes au restaurant et que vous voyez votre nourriture arriver."



Tâches connexes

- Traduction
- Normalisation phonétique
- Correction post-OCR et post-ASR
- Correction d'erreurs grammaticales
- Normalisation de variantes linguistiques
 - Dialectales
 - Historiques
- Simplification de textes

Données LexNorm 2015

Jst read a tweet **lol** and **l o v e** it



just read a tweet **laughing out loud** and **love** it

'input': ['Jst', 'read', 'a', 'tweet', 'lol', 'and', 'l', 'o', 'v', 'e', 'it'],

'output': ['just', 'read', 'a', 'tweet', 'laughing out loud', 'and', 'love', '', '', '', 'it']

Métriques d'évaluation (van der Goot, 2019)

- True negative (TN) = Annotators did not normalize, system did not normalize
- False positive (FP) = Annotators did not normalize, system normalized
- False negative (FN) = Annotators normalized, but system did not find the correct normalization. This could be because it kept the original word, or proposed a wrong candidate.
- True positive (TP) = Annotators normalized, systems normalized correctly

Analyse qualitative : BERT vs. BERTweet

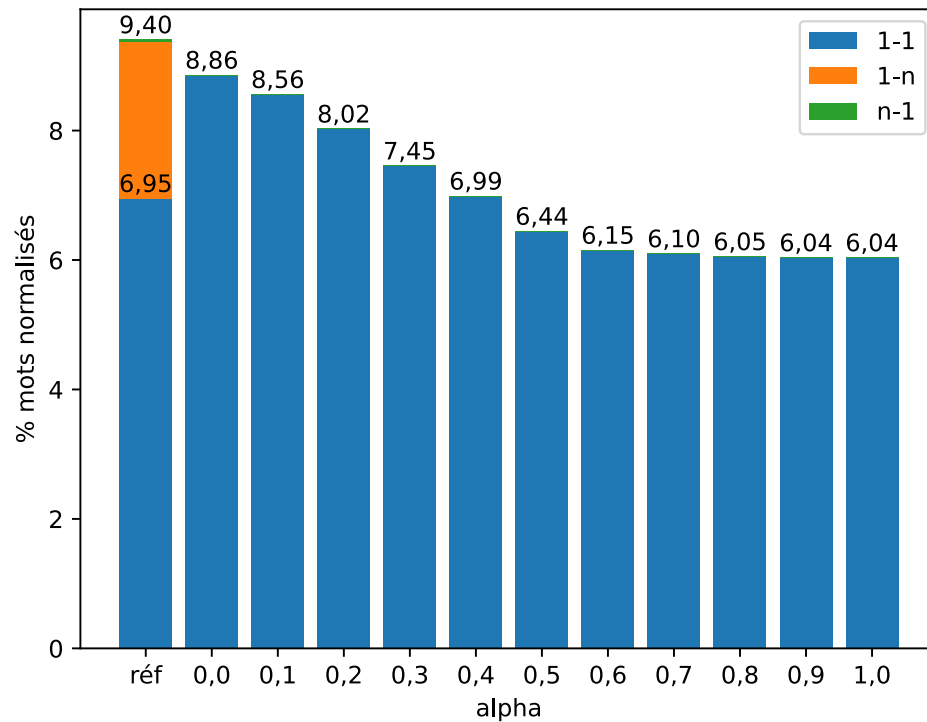
BERT

Source	rt @tehreelhov : wen ur at a restaurant nd u c ur food comin http://t.co/ducpxt7dry
Cible	rt @tehreelhov : when you're at a restaurant and you see your food coming http://t.co/ducpxt7dry
$\alpha = 0$	rt @tehreelhov : <u>r ##d</u> at a restaurant <u>. : : a food .</u> http://t.co/ducpxt7dry
$\alpha = 0,3$	rt @tehreelhov : wen ur at a restaurant and u c ur food coming http://t.co/ducpxt7dry
$\alpha = 1$	rt @tehreelhov : wen ur at a restaurant and u c ur food coming http://t.co/ducpxt7dry

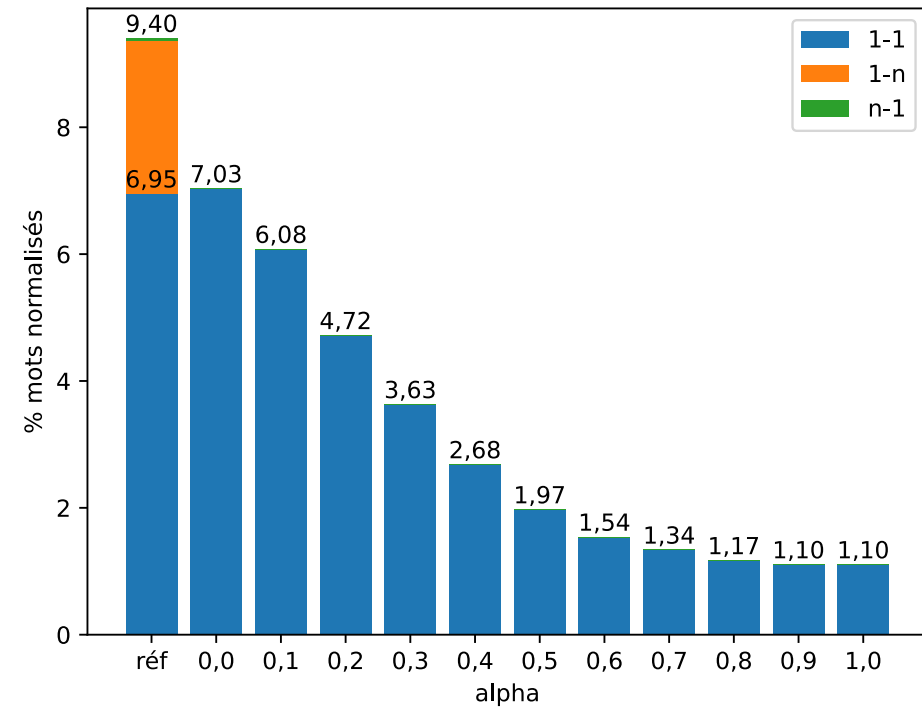
BERTweet

Source	rt @tehreelhov : wen ur at a restaurant nd u c ur food comin http://t.co/ducpxt7dry
Cible	rt @tehreelhov : when you're at a restaurant and you see your food coming http://t.co/ducpxt7dry
$\alpha = 0$	rt @tehreelhov : when ur at a restaurant and u see ur food <u><@@</u> http://t.co/ducpxt7dry
$\alpha = 0,3$	rt @tehreelhov : when ur at a restaurant and u see ur food comin http://t.co/ducpxt7dry
$\alpha = 1$	rt @tehreelhov : wen ur at a restaurant nd u c ur food comin http://t.co/ducpxt7dry

Analyse qualitative : BERT vs. BERTweet

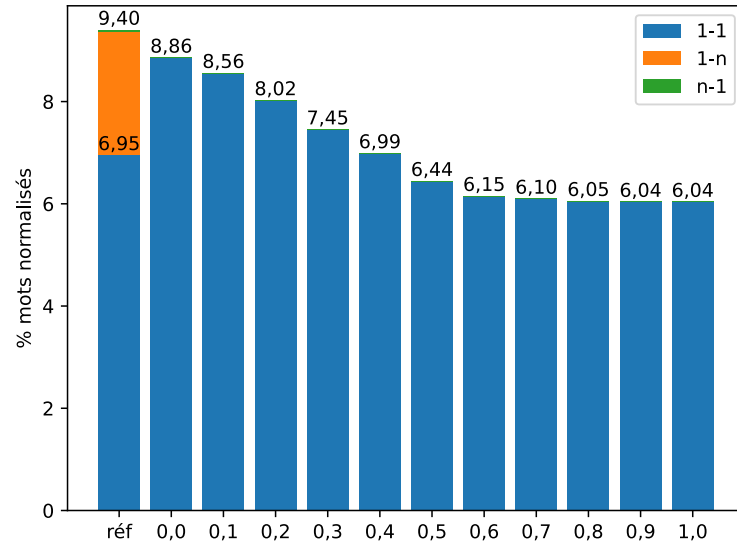


BERT

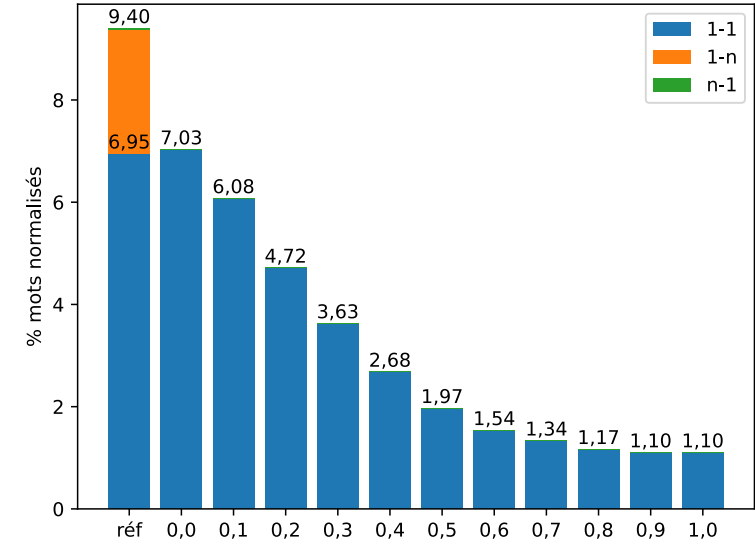


BERTweet

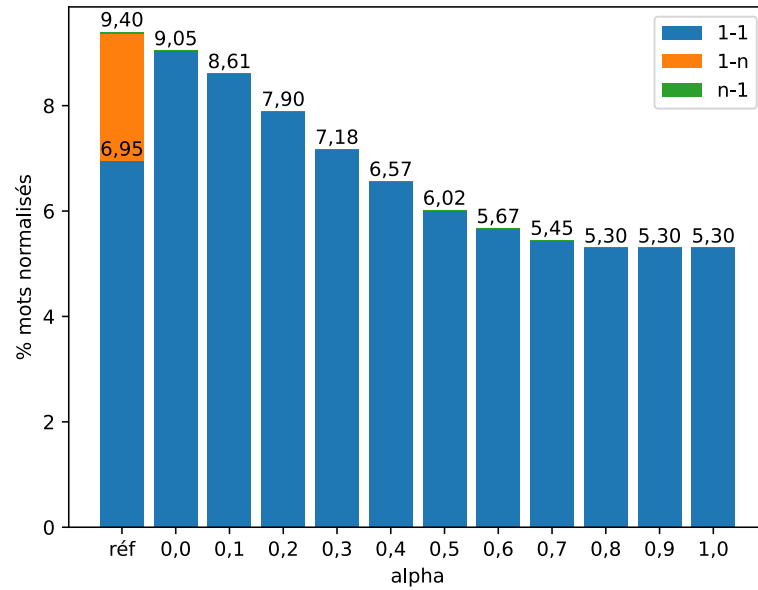
BERT



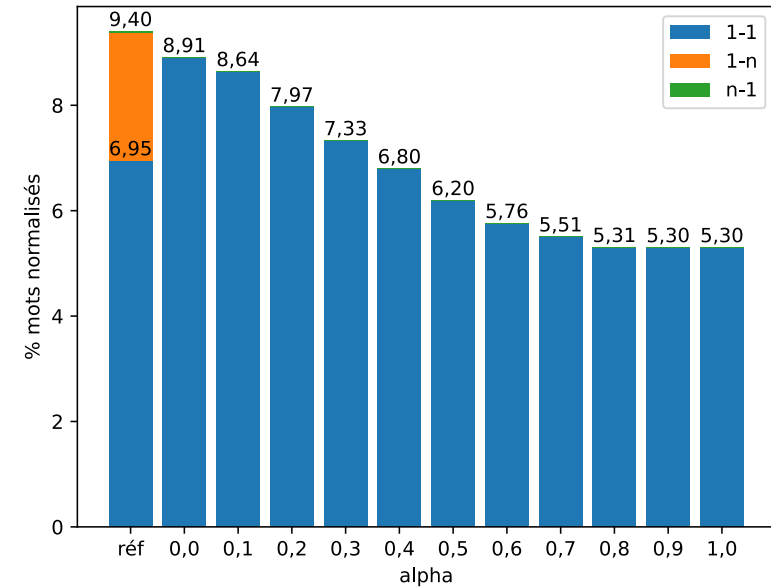
BERTWEET



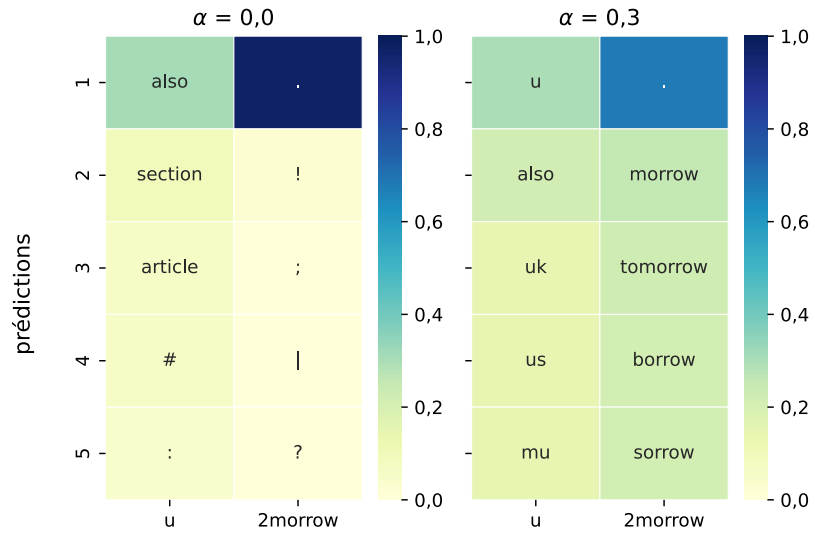
ROBERTA



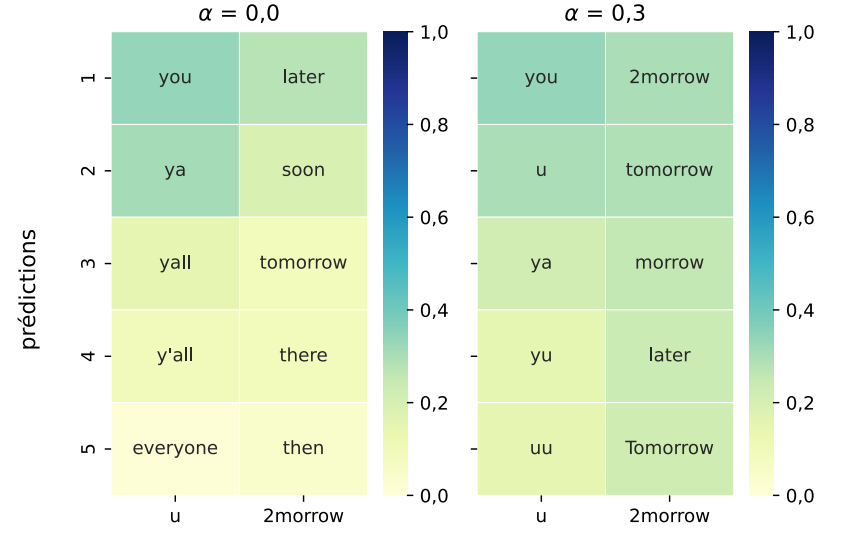
TWITTER ROBERTA



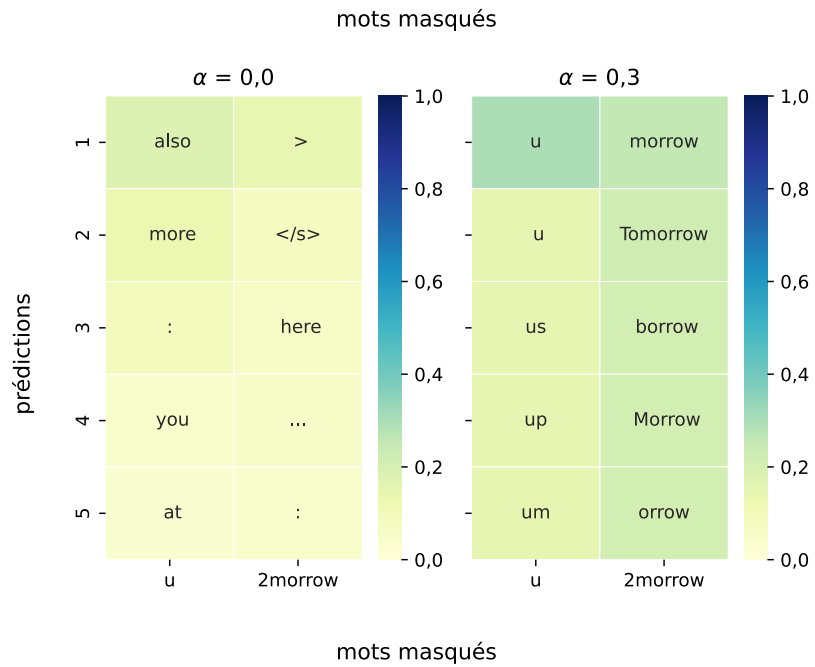
BERT



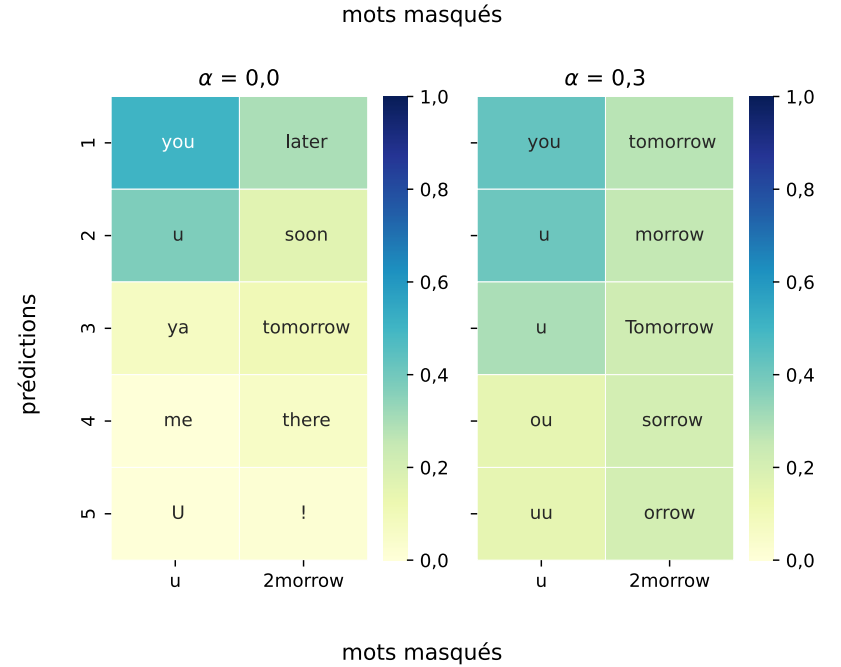
BERTWEET



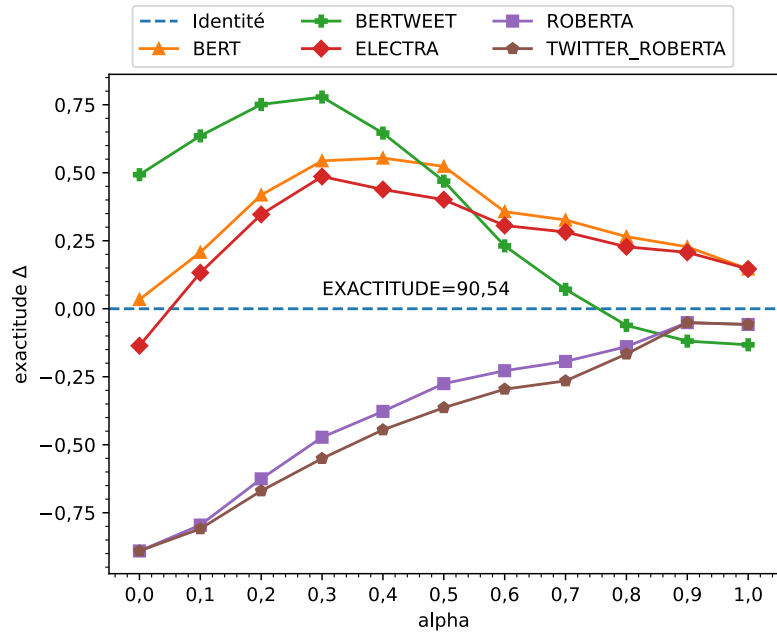
ROBERTA



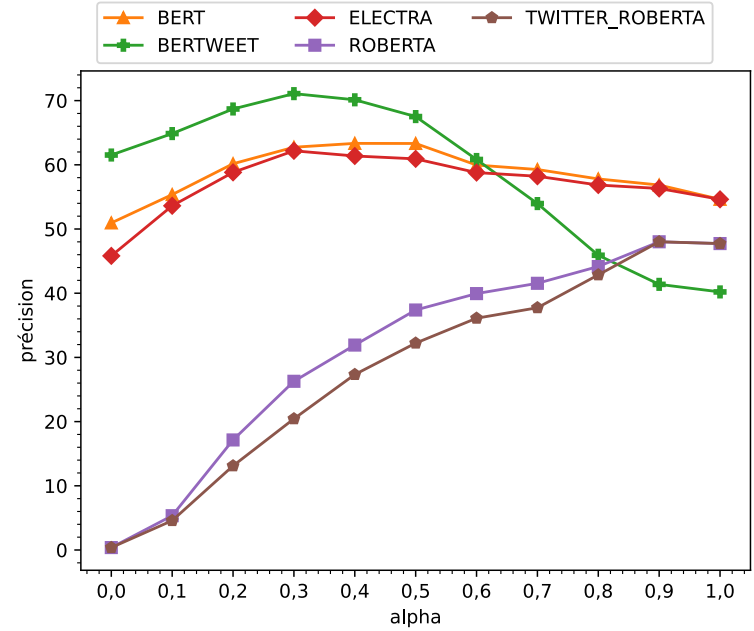
TWITTER ROBERTA



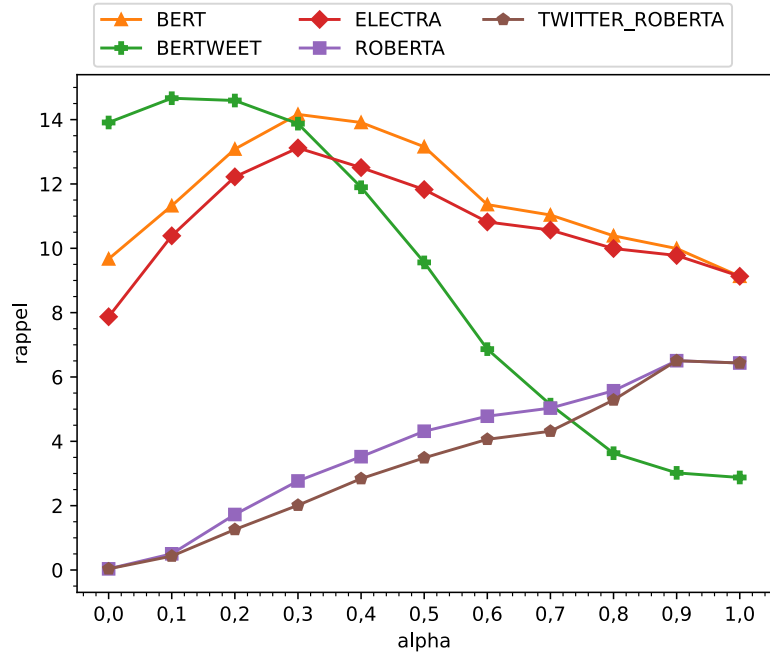
EXACTITUDE



PRECISION



RAPPEL



F-MESURE

