



HAL
open science

QuickNorm, une méthode rapide et peu coûteuse pour la normalisation d'entité

Arnaud Ferré, Louise Deleger

► **To cite this version:**

Arnaud Ferré, Louise Deleger. QuickNorm, une méthode rapide et peu coûteuse pour la normalisation d'entité. Journée GdR TAL - accès à l'information, Oct 2022, Rennes, France. 2022, 10.1186/s12859-023-05350-9 . hal-04318360

HAL Id: hal-04318360

<https://hal.inrae.fr/hal-04318360v1>

Submitted on 1 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QUICKNORM, UNE MÉTHODE RAPIDE ET PEU COÛTEUSE POUR LA NORMALISATION D'ENTITÉ

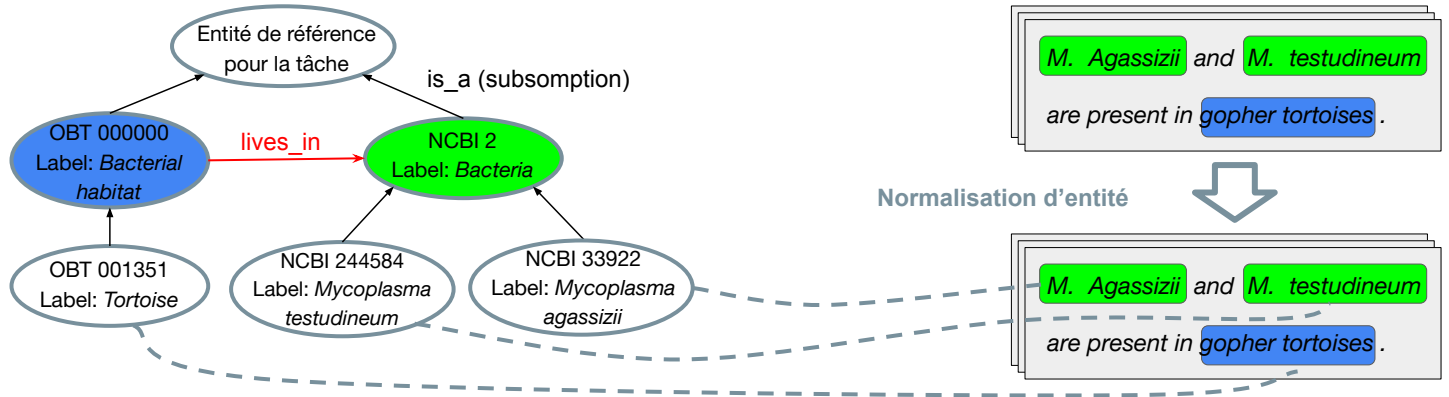


Arnaud Ferré
Louise Deléger

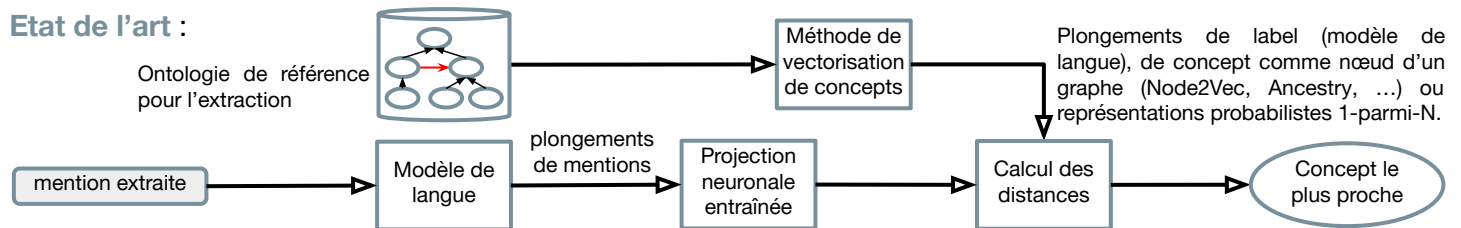


Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

Introduction : De nombreuses informations d'intérêt sont disséminées dans la littérature des domaines de spécialité, en particulier dans les publications scientifiques. L'extraction d'information vise à extraire ces informations et à les structurer de façon à les rendre exploitables. Nous aborderons ici une des étapes de l'extraction d'information : la tâche de normalisation d'entité.

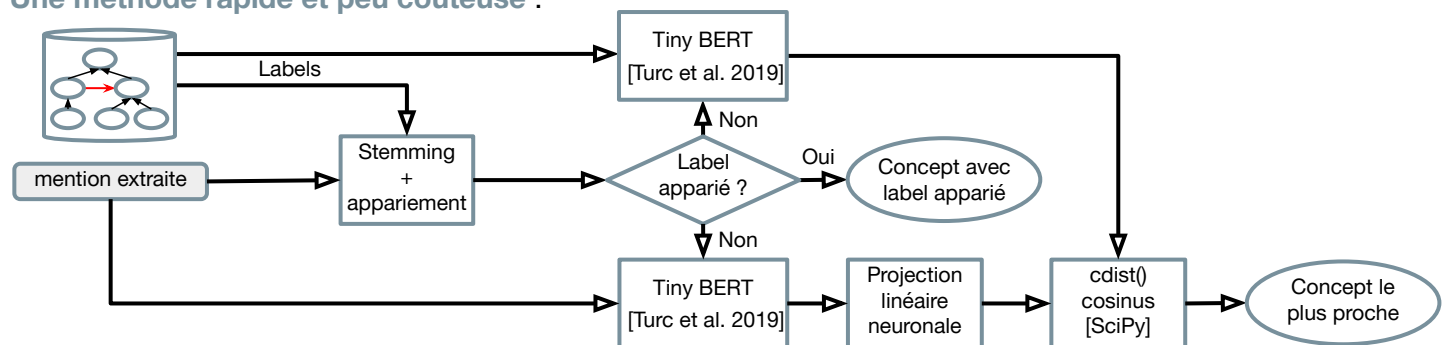


Etat de l'art :



Problématique : Les performances n'ont pas arrêté de croître, proportionnellement au nombre de paramètres à entraîner et donc aux besoins en puissance et en temps de calcul. Cela pose de plus en plus de problèmes en termes d'accessibilité, de reproductibilité, ainsi qu'en termes écologiques.

Une méthode rapide et peu coûteuse :



Résultats : La méthode a été évaluée sur un jeu de données standard d'évaluation, celui de Bacteria Biotope 4 [Bossy et al. 2019]. Comparée à des méthodes de l'état de l'art, la méthode réussit à ne pas être trop consommatrice de RAM et à s'exécuter (chargement des données, entraînement et prédiction) au moins 10 fois plus rapidement sur CPU que les méthodes état de l'art performantes sur GPU.

	Tiny BERT	BERT	PubMed BERT
Taille vecteur	128	768	768
Volume (Mb)	15,97	389,71	386,20
Jeu de données	Wikipedia and BooksCorpus	Wikipedia and BooksCorpus	MEDLINE/ PubMed

	Tiny BERT	BERT	PubMed BERT
Accuracy	48,93	50,96	51,44
Ecart-type	0,59	0,54	0,29
Temps (sec.)	82	1352	1339

(calcul de la variabilité sur 5 exécutions, CPU d'ordinateur portable de 32 Go de RAM)

Références :

- Bossy et al. 2019. Bacteria biotope at BioNLP open shared tasks 2019. BioNLP open shared tasks workshop.
- Turc et al., 2019. Well-read students learn better: On the importance of pre-training compact models. arXiv:1908.08962.
- SciPy: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cdist.html>