kNN-*Vote* : Ensemble kNN pour la recherche contextuelle et l'appariement sémantique d'adresses postales

El Moundir Faraoun[*,**], Nédra Mellouli[*,***]
Stéphane Millot[**] Myriam Lamolle[*]

[*]LIASD, Université de Paris 8, 2 rue de la liberté, Saint-Denis, France
el.moundir.faraoun@gmail.com,
[**]TEDIES, Talk Solutions, 45 Av. de Paris, Monéteau, France
[***]Léonard de Vinci Pôle universitaire, research center, Paris La défense, France

1 Introduction

Le problème de *correspondance d'adresses* est un défi central du TALN (Traitement automatique de langage naturel). Cette tâche consiste à traiter et aligner avec précision une adresse souvent erronée/bruitée avec une adresse correcte de référence dans le but de savoir si elles réfèrent au même objet du monde réel. Les erreurs peuvent être orthographiques, aléatoires (ajout de noms ou numéros de téléphone), ou sémantiques (abréviations, attributs d'adresses remplacés). Ces anomalies déforment la structure des adresses, rendant difficile leur appariement avec des données de référence.

2 *k*NN-*Vote* : système de recherche et de recommandation contextuelle d'adresses postales

Notre approche, kNN-*Vote*, transforme la tâche de *correspondance d'adresses* en une tâche de *recherche sémantique d'adresses*. En exploitant les plongements de phrases issus de plusieurs modèles de Transformeurs bi-encodeurs, nous représentons les adresses dans différents espaces vectoriels. Un algorithme kNN, combiné à un système de vote, nous permet de récupérer les adresses correctes les plus similaires, améliorant ainsi la robustesse et la pertinence des résultats. Nous décrivons le système de vote comme suit (*Cf.* Figure 1). Les lecteurs sont invités à se référer à la version longue de ce papier (Faraoun et al., 2024).

3 Résultats

Trois bi-encodeurs différents [1] ont été ajustés sur 147 000 paires d'adresses réelles collectées chez un transporteur du secteur privé. Les trois bi-encodeurs ainsi que d'autres types

1. Les modèles Camembert, XLM-Roberta et Roberta (pré-entraîné de zéro) ont été utilisés pour l'encodage.

d'encodage [2] sont utilisés individuellement à titre de comparaison. Les mesures de comparaison sont le ratio d'adresses correcte récupérées au rang k ($\%@k$) et MRR (*Rang réciproque moyen*).

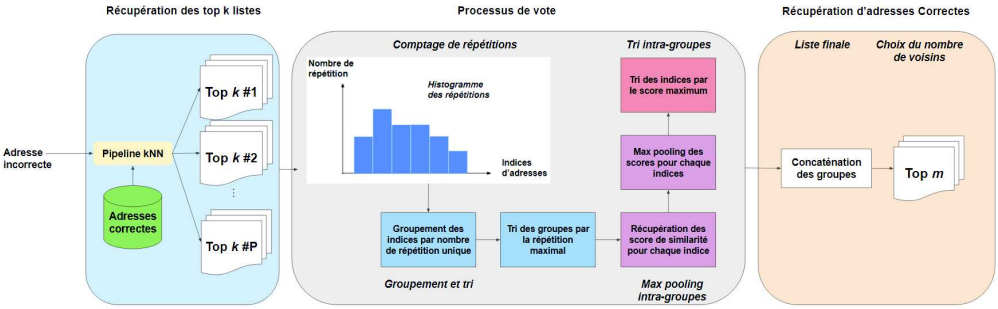

FIG. 1 – *Un pipeline de récupération via kNN est utilisé. Le kNN fournit à chaque fois le top k d'adresses correctes similaires selon un type de distance et une base d'encodage d'un bi-encodeur donnée. Par la suit, les différents top k sont agrégés en une liste finale à travers un processus de vote qui se base sur le tri via le nombre de répétition des adresses correctes récupérées dans les différents top k et le Max pooling des scores de similarité associés.*

Système	$\%@1$ exact	$\%@1$	$\%@10$ exact	$\%@10$	MRR
kNN$_{Token_set_ratio}$	0.740	0.829	0.866	0.889	0.791
kNN$_{Bi_DistilBert}$	0.763	0.793	0.918	0.939	0.826
kNN-*Vote*$_{Tous}$	0.760	0.872	0.950	0.962	0.830
kNN-*Vote*$_{Transf+texte}$	0.801	0.896	0.957	0.967	0.859
kNN-*Vote*$_{Transf}$	0.852	0.920	0.959	0.972	0.894
kNN-*Vote*$_{Camembert+XLM}$	**0.862**	**0.921**	**0.960**	**0.972**	**0.900**

TAB. 1 – *Les résultats montrent la supériorité des modèles de vote en terme de ratios et MRR. Plusieurs types d'encodages hétérogènes (recherche via des mesures de similarité textuelle) ont été ajoutés dans le vote pour analyser leur impact sur la performance du système. Nous notons que les votes à base de Transformeurs seulement sont nettement supérieurs.*

Références

Duarte, A. et A. Oliveira (2023). Improving address matching using siamese transformer networks. In *Proceedings of the Conference on Artificial Intelligence EPIA*, pp. 413–425.

Faraoun, E., N. Mellouli, S. Millot, et M. Lamolle (2024). Contextual knn ensemble retrieval approach for semantic postal address matching. In *Proceedings of the Intl. Worksh. & Tutorial on Interactive Adaptive Learning (IAL@ECML-PKDD)*.

2. Le modèle Bi_DistilBert utilisé par Duarte et Oliveira (2023) pour l'encodage ainsi que des recherches kNN via des similarités textuelles ont été testés.