



**HAL**  
open science

## Résolution de toponymes par apprentissage profond à partir de cooccurrences et de relations spatiales.

Jacques Fize, Ludovic Moncla, Bruno Martins

► **To cite this version:**

Jacques Fize, Ludovic Moncla, Bruno Martins. Résolution de toponymes par apprentissage profond à partir de cooccurrences et de relations spatiales.. 16th Spatial Analysis and Geomatics Conference (SAGEO 2021), May 2021, La Rochelle, France. hal-03225106

**HAL Id: hal-03225106**

**<https://hal.science/hal-03225106v1>**

Submitted on 12 May 2021

**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.

# Résolution de toponymes par apprentissage profond à partir de cooccurrences et de relations spatiales

Jacques Fize<sup>1</sup> Ludovic Moncla<sup>1</sup> Bruno Martins<sup>2</sup>

<sup>1</sup>LIRIS UMR 5205, INSA Lyon, France <sup>2</sup>Instituto Superior Técnico - INESC-ID, University of Lisbon, Portugal

## Résolution de toponymes

La **résolution de toponymes** (ou *geocoding*) fait partie de la tâche de *geoparsing* (processus permettant l'**identification** et la **géolocalisation des lieux mentionnés dans un texte** [Gritta et al., 2018]). Elle est généralement utilisée en complément de l'étape de repérage des noms de lieux :

1. **Geotagging**. Identification des noms de lieux (reconnaissance d'entités nommées)
2. **Geocoding**. Association d'une localisation (coordonnées géographiques).

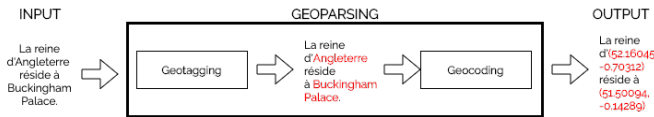
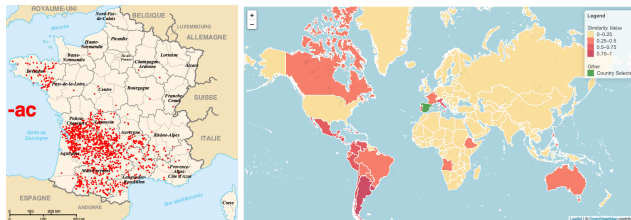


Figure 1. Processus de geoparsing [Gritta et al., 2018]

Dans ce travail, nous proposons une approche pour entraîner des modèles de geocoding [DeLozier et al., 2015, Cardoso et al., 2019] afin de ne pas avoir à utiliser de gazetteers lors de l'étape de *geocoding*. L'architecture que nous proposons pour l'entraînement des modèles prend en entrée des paires de toponymes. Le premier est celui que l'on cherche à localiser, le second sert de contexte.

## Hypothèse

Notre proposition s'appuie sur l'utilisation de **plusieurs types de contextes** pour la construction des paires de jeu de données d'entrée (cooccurrences et relations spatiales) et sur la **régionalité de certains affixes dans les noms de lieux** (Voir Fig. 2).



(a) dont le suffixe est -ac (en France) (b) d'Espagne et les autres pays

Figure 2. Similarité entre les noms de lieux

## Les données

Nous utilisons deux sources de données différentes pour constituer les paires de toponymes en fonction du contexte.

- **Contexte spatial**. Les données de **Geonames** sont utilisées pour constituer les paires de toponymes en s'appuyant sur la proximité ou l'inclusion qui existe entre les deux lieux.
- **Contexte textuel**. Les paires de toponymes sont générées à l'aide des cooccurrences existantes entre les lieux au sein des pages de **Wikipedia**.

Dans le but d'améliorer les performances du modèle, différentes configurations impliquant des combinaisons de ces contextes ont été testées et sont présentées dans la section *Expérimentation et résultats*.

## L'architecture

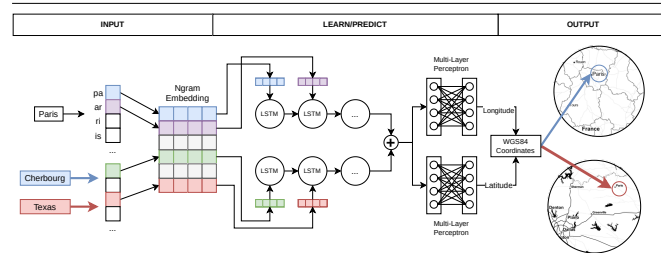


Figure 3. Architecture pour l'entraînement d'un modèle de geocoding.

Nous proposons une architecture s'appuyant sur un réseau de neurones récurrents (Bi-LSTM) avec :

- **En entrée**. Un couple de toponyme où le premier est le toponyme que l'on souhaite géolocaliser et le second est utilisé comme contexte.
- **En sortie**. Les coordonnées (en latitude-longitude) prédites par le modèle du premier toponyme.

Chaque toponyme est décomposé en séquence de ngram de caractères. Par exemple, pour  $n = 4$  le toponyme **La Rochelle**, on obtient la séquence suivante :  $\{\$ \$ \$ \$, \$ \$ \$ L, \$ L a, \$ L a, \$ R, a \$ R o, \$ R o c, R o c h, o c h e, h e l l, e l l e, l l e \$, l e \$ \$, e \$ \$ \$\}$ . Une comparaison des performances de notre modèle selon la taille de n-gram indique que de meilleurs résultats sont obtenus avec  $n > 3$ .

## Protocole d'évaluation

- Différents jeux de données d'apprentissage et d'évaluation selon le pays : France, Royaume-Uni, États-Unis, Japon, Argentine, Nigeria
- Score :  $Accuracy@k$ , ou  $A@k(F, T)$  où  $F$  correspond au coordonnées prédites et  $T$  les coordonnées attendues.

$$A@k(F, T) = \frac{1}{|F|} \times \sum_{i=0}^k \begin{cases} 1, & \text{if } distance(F_i, T_i) < k \\ 0, & \text{sinon} \end{cases}$$

- Jeux de données d'évaluation : *Geocoding* de pages de lieux sur Wikipedia.

## Expérimentations et résultats

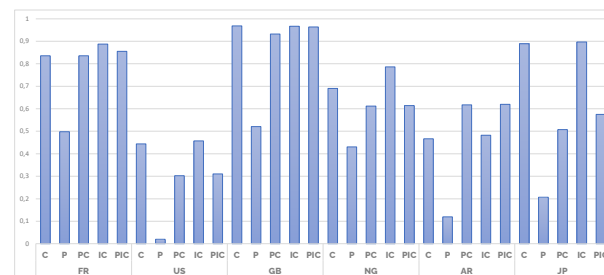


Figure 4. Exactitude de notre modèle selon la combinaison de contextes utilisée (C = Cooccurrences; P = Proximité; I = Inclusion)

## Apport du contexte spatial pour certains pays

Pour les pays avec peu d'articles dans Wikipedia, l'ajout de paires de toponymes basées sur les relations spatiales améliore les résultats.

	# paires		Accuracy @100km	
	C	P	C	PC
FR	714 974	985 090	0,8359	0,8359
AR	9 378	3 052 000	0,4669	0,6176

Table 1. Impact de la combinaison de contextes sur les performances (P= Proximité, C= Cooccurrence)

## Impact du sampling

Le nombre total de paires de toponymes pour un pays pouvant être très élevé, nous effectuons un échantillonnage sur l'ensemble des paires disponibles. La Fig. 5 montre l'impact positif d'un échantillonnage plus grand sur l'entraînement du modèle.

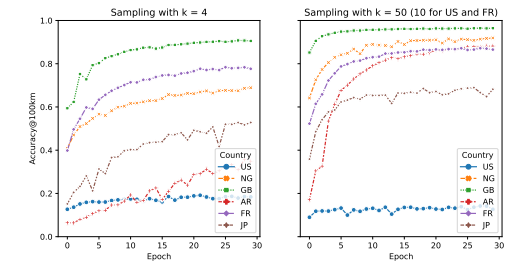


Figure 5. Évaluation lors de l'entraînement des modèles.

## Remerciements

Ces travaux s'inscrivent dans le projet HextGEO réalisé grâce au soutien financier du projet IDEXYLON de l'Université de Lyon, dans le cadre du programme Investissement d'Avenir (ANR-16-IDEX-0005). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2020-AD011011551R1 attribuée par GENCI.

## Références

[Cardoso et al., 2019] Cardoso, A. B., Martins, B., and Estima, J. (2019). Using Recurrent Neural Networks for Toponym Resolution in Text. In *Proceedings of the EPIA Conference on Artificial Intelligence*, pages 769–780. Springer International Publishing.

[DeLozier et al., 2015] DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

[Gritta et al., 2018] Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

