Extraction de connaissances géographiques et localisation incertaine à partir d'un corpus d'annonces immobilières



Introduction

Les annonces immobilières regorgent de connaissances géographiques et spatiales qui mettent en avant la localisation d'un bien et qui peuvent enrichir les dictionnaires géographiques (e.g., IGN, OpenStreetMap, GeoNames). Ces connaissances peuvent être des noms de lieux connus, mais aussi des lieux non-officiels (e.g., la « banane » à Cannes) ou non-nommés (e.g., la plage, la gare, etc.). De plus, des relations spatiales qualitatives (e.g., proche, non loin, à deux pas) sont souvent employées pour localiser le bien mais sont difficile à modéliser. Enfin, les limites des lieux mentionnés sont souvent incertaines et ne correspondent pas aux limites officielles mais plutôt à la perception que les agents immobilier en ont. Dans cette étude, nous proposons une méthode pour extraire les connaissances géographiques et spatiales dans les annonces immobilières écrites en français. Ensuite, nous développons une méthode pour représenter de manière floues les lieux et relations spatiales. Enfin, nous utilisons les informations extraites pour retrouver une approximation de la localisation du bien.

Méthode

Corpus d'annonces immobilières écrites en français

Extraction des connaissances géographiques

- **Définition** des catégories d'entités à extraire
- **Annotation** de 1200 annonces
- **Entraînement** d'un modèle Bi-LSTM + CRF
- **Extraction de relations** entre entités : règles à partir de la nature des entités et des dépendances grammaticales

Création de la base de données de lieux

- Nettoyage des entités (fautes d'orthographe, abréviations)
- Regroupement toponyme, entité géographique et spatiotemporelle (e.g., « proche place Massena ») **Extraction** d'annonces
- géolocalisées **Conservation** des lieux mentionnés au moins dans 10 annonces

différentes

Représentation « floue » des lieux

- **Estimation** par noyau (KDE) d'un lieu à partir d'annonces géolocalisées
- **Application** de la fonction aux parcelles cadastrales
- Création de zones floues à partir de seuils de probabilités (0.99, 0.9, etc.)
- Stockage des zones dans un graphe RDF

Fusion des informations et localisation du bien

- Représentation des zones comme des ensembles flous
- Regroupement des ensembles flous à partir de l'opérateur moyenne arithmétique
- Représentation du nouvel ensemble flou avec des alpha-coupes



Application de la méthode

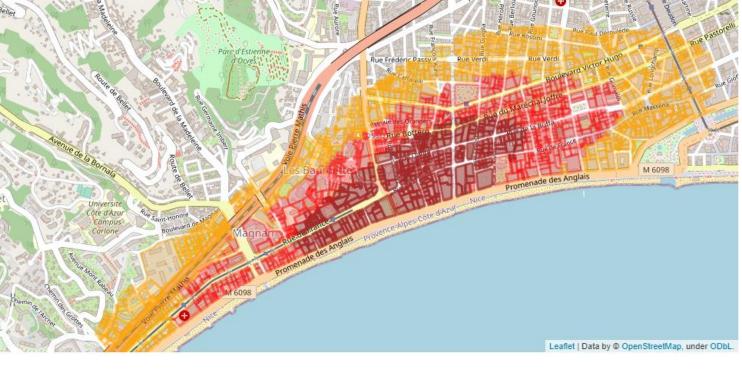
Annonce immobilière

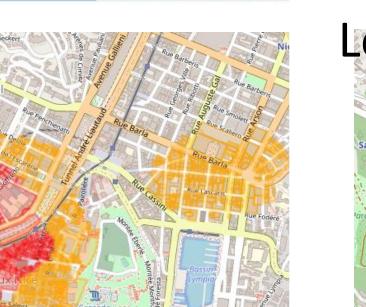
Lieux extraits

Représentation floue

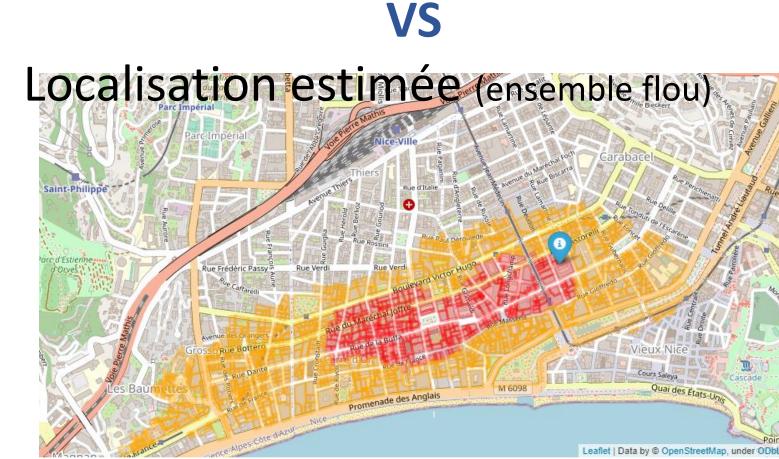


Appartement situé la célèbre Promenade Lieu 1 : « Promenade des Anglais » des <mark>Anglais</mark> à deux pas de la place Massena. La plage est à 5 minutes à pied.







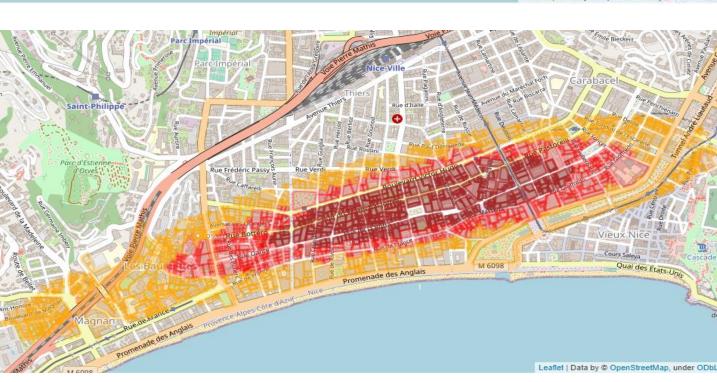




Catégories d'entités :

- Toponyme
- Entité géographique
- Entité Spatio-temporelle
- Mode de transport





Conclusion

Dans ce travail, nous avons développé une méthode qui permet d'extraire, représenter et fusionner des informations géographiques et spatiales floues dans un corpus d'annonces immobilières écrites en français. Cette méthode peut s'appliquer à d'autres types de textes. Aussi, l'extraction et la résolution des lieux peuvent contribuer à l'enrichissement des bases de données géographiques.

Contact et Partenaires

