Connaissances géospatiales dans les annonces immobilières : détection et extraction d'information spatiale à partir du texte

Lucie Cadorel, Alicia Blanchi, Andrea G. B. Tettamanzi

Connaissances géospatiales dans les annonces immobilières : détection et extraction d'information spatiale à partir du texte

Lucie Cadorel[1,3], Alicia Blanchi[2,3], Andrea G. B. Tettamanzi[1]

[1] Université Côte d'Azur, Inria, CNRS, I3S
[2] Université Côte d'Azur, ESPACE, CNRS
[3] KCityLabs

lucie.cadorel@inria.fr

Résumé

Nous avons proposé un modèle d'extraction de connaissances géospatiales à parir du texte appliqué au cas des annonces immobilières. La première étape consiste à extraire les entités géographiques et spatiales à l'aide d'un modèle basé sur une architecture BiLSTM-CRF *et la concaténation de plusieurs embeddings. Ensuite, nous avons réalisé l'extraction de relations, notamment spatiales, pour créer une base de connaissance géospatiale structurée stockée dans un graphe de connaissance RDF.*

Mots-clés

Extraction d'information, connaissance géographique, reconnaissance d'entités nommées, extraction de relation

Abstract

We proposed a workflow to extract geospatial knowledge from text applied to Real Estate advertisements. We first extracted geographic and spatial entities using a model based on a BiLSTM-CRF *architecture with a concatenation of several text representations. Secondly, we performed relations extraction, particularly spatial relations extraction, to build a structured Geospatial knowledge base that we stored in a RDF Knowledge Graph.*

Keywords

Information extraction, geographical knowledge, named entity recognition, relationship extraction

1 Introduction

La reconnaissance d'entités géospatiales dans les textes a largement été développée par les avancées en traitement du langage naturel, et a été appliquée à divers types de textes tels que les blogs de voyage [2], les réseaux sociaux [3] ou bien les annonces immobilières [4]. L'approche traditionnelle consiste à utiliser des règles linguistiques et des dictionnaires géographiques (gazetteer). Cependant, cette approche donne des résultats limités puisqu'elle dépend de la complétude des règles et des dictionnaires. Ainsi, les modèles utilisant du Deep Learning sont de plus en plus développées et obtiennent de très bons résultats. Néanmoins, la plupart des études detectent seulement les lieux-nommées alors que des termes géographiques (e.g., la gare, la plage, les écoles, etc.) peuvent être aussi utilisés pour mentionner un lieu. De plus, les relations spatiales sont aussi source d'information et permettent de mieux localiser un lieu (e.g, "à 10 minutes", "proche", "à deux pas", etc.) mais sont rarement extraites. On retrouve notamment ce type de connaissances dans les annonces immobilières. En effet, les agents immobiliers décrivent de façon vague les lieux qui permettent de situer une propriété (e.g, "L'appartement est situé dans un quartier résidentiel proche de l'université de Nice. Commodités à deux pas."). Cependant, ces lieux flous donnent des informations à la fois sur la localisation du logement et sur le quartier et ses équipements. Il est donc important de reconnaître et d'extraires ces connaissances afin de mieux comprendre le marché immobilier, les éléments influents les prix ou encore la perception des lieux résidentiels . Ainsi, les agents immobiliers pourront mieux connaître les prix, les tendances du marché d'un quartier et les biens similairs à celui vendu, notamment lorsque celuici est situé hors du secteur habituel de l'agent.

Nous avons donc proposé un modèle d'extraction de connaissances géospatiales à parir du texte appliqué au cas des annonces immobilières. Cet article est un résumé traduit et mis à jour de l'article [1] que nous avons publié à K-CAP '21.

Le reste de cet article est organisé de la manière suivante. Dans la section 2, nous présentons le pipeline d'extraction mis en place pour retrouver les entités géospatiales et les relations puis les stocker de manière structurée. La section 3 détaille l'évaluation et la comparaison du modèle proposé.

2 Extraction d'information géospatiale

2.1 Reconnaissance d'entités géospatiales

La reconnaissance d'entités nommées géospatiales consiste à identifier les termes d'un texte faisant référence à des entités géographiques et spatiales telles que les lieux-nommés ("Nice", "Place Masséna", etc.). Nous avons identifié quatre catégories à extraire : Lieu-nommé (Toponym), type de lieu (Feature), entité spatio-temporelle (Spatiotemporal) et le mode de transport (Mode of transporation). Les deux pre-

mières catégories définissent explicitement un lieu à différents niveaux de précision, tandis que les entités spatio-temporelles et le mode de transport décrivent une relation spatiale permettant de localiser ce lieu.
Pour extraire ces informations, nous avons mis au point un modèle basé sur une architecture *BiLSTM+CRF* prenant en entrée un embedding du texte. Nous l'avons entraîné sur un corpus d'environ 1200 annonces immobilières préalablement annotées en utilisant le format de tag BIESO. L'embedding utilisé est un vecteur composé de la concaténation de trois représentations différentes du texte. La première représentation est un Word Embedding classique entraîné sur notre corpus d'annonces immobilières. La seconde est basé sur le modèle de langage pré-entraîné Flair pour le français que nous avons réentraîné sur notre corpus. Enfin, nous utilisons le modèle de langage CamemBERT sans réentraînement dû au manque d'un corpus de taille suffisante. Ces trois représentations permettent de capturer les spécificités et la variabilité du style de langage utilisé dans les annonces immobilières.

2.2 Extraction de relations

La deuxième partie de notre travail vise à obtenir une représentation structurée des informations extraites. Pour cela, nous avons extrait trois types de relations entre les entités retrouvées : Attribut, Type de lieu nommé et Spatiale.
Nous avons fait plusieurs hypothèses pour extraire les relations. D'abord, une relation a lieu seulement entre deux entités d'une même phrase. Il existe donc toujours un lien direct ou indirect entre les deux entités qui peut être ainsi retrouvé à l'aide d'un graphe de dépendance grammaticale. Pour obtenir ce graphe de dépendance, nous utilisons un modèle d'analyse de dépendance qui renvoie la structure syntaxique d'une phrase à partir de la grammaire. Ce modèle détermine les connections grammaticales entre les mots suivant le schéma <*Sujet, Fonction grammaticale, Objet*> qui est adapté à la structure syntaxique des annonces immobilières. En effet, celles-ci ne suivent pas toujours la grammaire standard avec un ordre des mots différents, un sujet ou un verbe absent, etc. Le modèle utlisé est l'analyseur syntaxique de Stanza pour le français basé sur la taxonomie universelle des dépendances (Universal Dependencies taxonomy) et pré-entraîné sur un grand corpus. Néanmoins, ce modèle ne donnait pas des résultats satisfaisants, notamment pour la partie étiquetage morpho-syntaxique (Part-of-Speech). Nous avons donc décidé d'entraîner notre propre modèle d'étiquetage morpho-syntaxique sur nos annonces immobilières.
A partir des dépendances syntaxiques, nous construisons le graphe de dépendances pour chaque phrase. Nous avons ensuite extrait le plus court chemin entre chaque paire d'entités candidates à une relation. Enfin, grâce à des règles prédéfinies, nous déterminons si les chemins extraits correspondent à une relation ou non.

2.3 Représentation des connaissances

La dernière étape de notre travail porte sur la manière de représenter et d'interroger la connaissance extraite. Nous avons choisi d'utiliser un graphe de connaissance car il offre une manière flexible de représenter les entités (nœuds) et les relations (arcs) mais aussi un langage de requête pour naviguer et raisonner sur les informations. Le modèle RDF et le langage de requête GeoSPARQL ont été choisis pour décrire et stocker les données.

3 Evaluation

Nous avons évalué le modèle d'extraction de connaissances à partir d'un jeu de données d'environ 1200 annonces immobilières préalablement traitées, nettoyées et découpées en 10 échantillons pour faire une validation croisée. Nous avons comparé plusieurs architectures de notre modèle avec le modèle Spacy pré-entraîné pour le français. Le meilleur modèle, qui utilise l'architecture *BiLSTM+CRF* avec l'embedding décrit dans 2.1, obtient un F1-Score de 0.876 soit 5.5 points au-dessus du modèle de Spacy pré-entraîné.

4 Conclusion et perspectives

Nous avons décrit dans cet article une méthode pour extraire des informations géospatiales des textes appliquée aux annonces immobilières écrites en français. Nous avons créé un modèle de reconnaissance d'entités pour extraire des lieux-nommés mais aussi les types de lieu, les entités spatio-temporelles et les modes de transport. Nous avons aussi conçu une méthode pour extraire des relations entre les entités et plus particulièrement des relations spatiales. Enfin, nous avons representé les connaissances extraites à l'aide d'un graphe de connaissance RDF.
Par la suite, nous envisageons de retrouver la localisation des lieux mentionnés et de les relier à des graphes de connaissances existants (e.g., GeoNamnes, DBpedia, etc.). Aussi, nous aimerions prendre en compte l'incertitude et l'imprécision des termes spatio-temporels afin d'améliorer la fiabilité de la localisation d'un lieu.

Références

[1] L. Cadorel et al., *Geospatial Knowledge in Housing Advertisements : Capturing and Extracting Spatial Information from Text.* In Proceedings of the 11th on Knowl- edge Capture Conference, K-CAP '21, page 41–48, New York, NY, USA, 2021. Association for Computing Machinery.

[2] B. Adams and K. Janowicz, *On the geo-indicativeness of non-georeferenced text.* In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012. The AAAI Press, 2012

[3] R. Grace, *Toponym usage in social media in emergencies.* International Journal of Disaster Risk Reduction, 52 :101923, 2021

[4] Y. Hu, et al., *A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements.* Int. J. Geogr. Inf. Sci., 33(4) :714–738, 2019.