Qualifying and quantifying uncertainty of geolocation information extracted from french real estated ads

Lucie Cadorel

Localisation sur le territoire et prise en compte de l'incertitude lors de l'extraction des caractéristiques de biens immobiliers à partir d'annonces

Lucie CADOREL

Centre Inria d'Université Côte d'Azur, CNRS, I3S
Equipe Wimmics

Présentée en vue de l'obtention du grade de docteur en Informatique
d'Université Côte d'Azur
Dirigée par : Andrea G. B. TETTAMANZI, Professeur des Universités, Université Côte d'Azur
Invité : Denis OVERAL**,** Responsable Innovation, SepteoProptech
Soutenue le : 24/01/2024

Devant le jury, composé de :
Président du Jury :
Examinateurs :
Elena CABRIO, Professeure des Universités, Université Côte d'Azur
Ludovic MONCLA, Maître de conférences, INSA Lyon, LIRIS
Rapporteurs :
Nathalie AUSSENAC-GILLES, Directrice de Recherche, IRIT, CNRS
Ross PURVES, Professeur des Universités, Université de Zurich

Acknowledgments

First I would like to thank my supervisors, Andrea G. B. Tettamanzi and Denis Overal, for their advice, patience, and support throughout the duration of this thesis. Thank you Andrea, for your support and research guidance over these three years. It has been a pleasure working with you and learning what the word 'research' means. I cannot think it is the end of our collaboration but rather the beginning. Thank you Denis, for your mentorship, and providing me the opportunity to join your team at SepteoProptech. This experience gave another dimension to my thesis by connecting the industrial and academic world. It showed me that the two sides are not disjointed and, they can and should collaborate together. I would be lying if I said it was an easy journey, but thanks to you, Andrea and Denis, I managed to juggle between the industrial and academic world.

I would like to express gratitude to each member of the jury. Nathalie Aussenac-Gilles and Ross Purves for taking the time to review my thesis. Elena Cabrio and Ludovic Moncla for agreeing to be my thesis examiners and for providing me advice and encouragement. All the insightful remarks offered me new perspectives on my work.

I am thankful for every member of the WIMMICS team I have met since I joined in May 2020, and in particular Fabien Gandon for always trying to make this team as awesome as possible. I am very happy to say that some of my colleagues have become valuable friends, and without their support, laughter, sports sessions, coffee breaks and afterwork events, this thesis would have not been the same. I also thank my colleagues from SepteoProptech and the ESPACE research team who taught me a new research field namely geography. A special thank to Alicia who has not been only my colleague but also my friend. I can never thank you enough for all the support you have given me in my professional but also personal life.

Last but not least, I would like to thank my family for their presence and constant support. Thank you to Nicholas for his help, support and listening even when I was grumpy and desperate. Thank you to my parents for always believing in me, and my brother, Julien, who is happier than me to address me as 'Dr. Cadorel'.

Résumé

Ces dernières années, de nombreuses applications basées sur des données textuelles générées par des utilisateurs ont vu le jour sur le Web grâce aux évolutions faites dans le domaine du traitement automatique du langage (TAL). Plus particulièrement, les données textuelles ont joué un rôle important dans de nombreuses applications géographiques telles que l'enrichissement des systèmes d'information géographique (SIG) ou la géolocalisation d'événements (accidents, catastrophes naturelles, etc.). Ces données peuvent provenir de différentes sources, telles que les blogs de voyage, les réseaux sociaux ou encore les annonces immobilières, et se réfèrent toutes qualitativement à des lieux. Cependant, les descriptions sont souvent vagues et imprécises (par exemple, 'proche du centre-ville', 'à l'ouest de Nice') rendant plus difficile la géolocalisation des lieux. Par exemple, dans les annonces immobilières, les agents utilisent souvent des termes vagues et exagèrent les limites afin de rendre plus attractif un bien et de le promouvoir. Bien que les annonces immobilières constituent une excellente source de données pour l'analyse du marché immobilier, qui repose essentiellement sur une connaissance très locale et approfondie des biens, des prix et des quartiers, les agents français ont tendance à cacher la localisation spatiale du bien (latitude/longitude) pour garder leur mandat exclusif. Par conséquent, la description textuelle du bien est souvent essentielle pour estimer la localisation d'une annonce afin d'exploiter cette information pour une analyse plus fine. Motivée par une collaboration industrielle avec SepteoProptech, une société opérant dans le domaine de l'immobilier, cette thèse explore la combinaison d'applications de plusieurs disciplines, telles que le TAL, les SIG, et le Web Sémantique, toutes appliquées au domaine de l'immobilier. En premier lieu, nous proposons une méthode pour détecter et extraire des informations spatiales à partir d'annonces immobilières écrites en français. Ensuite, nous proposons d'apprendre et de représenter les limites des descriptions spatiales imprécises et, en particulier, des lieux vernaculaires et des relations spatiales floues, grâce à une estimation de la densité et à la théorie des ensembles flous. A partir de ces résultats, nous proposons une méthode pour agréger des informations imprécises provenant de différentes sources afin de reconstruire la localisation approximative d'un bien immobilier. Nous évaluons notre approche sur un

jeu de données collecté auprès de différents annonceurs français et localisé sur la Côte d'Azur. Enfin, nous proposons une méthode pour construire un graphe de connaissances afin de faciliter l'exploitation des données immobilières, puisque les annonces ne sont pas structurées et il peut être difficile de raisonner dessus. En outre, nous présentons une ontologie créer pour représenter des biens immobiliers ainsi que des informations spatiales imprécises. Ainsi, nous avons créé et publié un nouveau jeu de données appelé SURE-KG, pour lequel nous proposons des applications potentielles dans le domaine de l'immobilier ainsi qu'à un plus large éventail d'utilisateurs, tels que les utilisateurs des SIG, les géographes et les chercheurs en TAL.

Mots clés : Extraction d'Information, TAL, Incertitude, Géographie, Immobilier

Abstract

In recent years, many applications based on user-generated free text have arisen on the Web, driven by significant evolutions in the field of Natural Language Processing (NLP). In particular, textual data have played an important role in many geographic applications, including Geographic Information Systems (GIS) enrichment or the geolocation of events (e.g., accidents, natural disasters). These data might come from different sources such as travel blogs, social media or Real Estate advertisements, but they all qualitatively refer to locations. However, descriptions are often vague and uncertain (e.g., 'Near the city center', 'West of Nice, France') and make it challenging to geocode places. For instance, in real-estate advertisements, the agents often use vague terms and exaggerate boundaries in order to hype a property. Altough real-estate advertisements are a great source of data to analyse the market, which relies upon a very local and deep knowledge of the properties, the prices and the neighborhoods, French Real Estate agents often hide the spatial location (e.g., latitude/longitude) since the deal between their agency and the owner might not be exclusive. Therefore, the textual description is often essential to estimate the location of an advertisement in order to exploit such information for further analysis. Motivated by an industrial collaboration with SepteoProptech, a company operating in the Real Estate domain, this thesis explores the combination of application of several disciplines, such as NLP, GIS, and the Semantic Web, to the Real Estate domain. First, we propose a method to detect and extract spatial information from Real Estate advertisements written in French. Additionally, we propose to learn and represent boundaries of vague spatial descriptions and, in particular, vernacular places and vague spatial relations, thanks to a density estimation and fuzzy set theory. As a result, we propose a method to aggregate imprecise information from different sources in order to reconstruct the approximate location of a Real Estate property. We evaluate our approach on a dataset collected from different French advertisers and located in the French Riviera. Finally, we propose a method to build a knowledge graph to make it easy exploit Real Estate data, since the advertisements are unustructured and, it might be difficult to reason over them. We design an ontology to represent Real Estate as well as uncertain spatial

information, and we create and publish a new dataset called SURE-KG. We show the potential applications of this dataset to the Real Estate domain, but also its interest for wider range of users, such as the GIS community, geographers, and NLP researchers.

Keywords : Information Extraction, NLP, Uncertainty, Geography, Real Estate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1 Context and Motivations

In recent years, many applications based on user-generated free text have arisen on the Web, driven by significant evolutions in the field of Natural Language Processing (NLP). In particular, geographic and spatial information are often qualitatively described in unstructured (textual) data such as travel blogs, social media or Real Estate advertisements. While Geographic coordinate systems, such as the World Geodetic System (WGS84) used by GPS, have been mainly used to represent and interpret places in information systems (e.g., digital gazetteers), textual data remains the main source of spatial data. Indeed, people often refer to a place by using natural language, such as place names (e.g., *Nice*, *Paris*), and linking them to spatial footprints is useful in many applications (e.g., geocoding accidents or natural disasters). Nevertheless, these place names are often vernacular and local and might not be recorded in existing gazetteers which makes the geocoding difficult. A vernacular place is a place name with vague boundaries defined by regional culture (e.g., *Downtown*). Thus, disagreement might arise between official boundaries and what the locals consider as the actual boundaries. For instance, Montello et al. [Montello, 2003] discussed the difficulties to represent vague cognitive places in precise coordinates. They took the example of the Downtown of Santa Barbara in California to show how each inhabitant has his own representation, and proposed an empirical approach to draw vague boundaries. In addition to vernacular places, people think and talk with vague spatial terms to describe

a location (e.g., near). For example, the preposition *next to*, in the spatial description *next to the Riquier train station*, gives a precision of the location and could improve the accuracy of the geocoding. Although several models have been developed to define an area in which a preposition could validly be used, it is very challenging to generalize it since it largely depends on the context (e.g., the use of *near* could be different between two regions) [Carlson, 2005; Herskovits, 1985; Stock, 2018; Tyler, 2003; Aflaki, 2022]. Therefore, it remains difficult to delimit boundaries for vague spatial objects. Finally, these textual data play more and more an important role in many geographic applications, including Geographic Information Systems (GIS) enrichment, better describing our environment [Adams, 2012], or the location of events (e.g., natural disasters) [Hu, 2021], but they are not always adapted to geographic information systems. Digital gazetteers provide organized collections of place names, place types, and their spatial footprints, and fill the critical gap between formal computational representation and informal human discourse. Nevertheless, the spatial objects are often represented with sharp and well defined boundaries (e.g., point, lines, polygons), which might not be suitable for vague places and relations [Goodchild, 2000]. Several methods have been proposed to overcome this issue and represent and reason over imprecise spatial data such as the fuzzy set theory [Goodchild, 1998] or the *Egg-yolk* model [Cohn, 2020], but their storage remains challenging. To unify the different geospatial representations and data access, Knowledge Graph (KG) and Linked Open Data have been more and more studied. A KG could be defined as 'a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities' [Hogan, 2021], and gives a more flexible manner to design and maintain data. Linked Open Data refers to the collection of interrelated datasets on the Web that can be reused for wider applications. These technologies allow to exploit and share unstructured data, and bridge the gap between the machines and humans.

Application to the Real Estate Domain

The work presented in this thesis has been conducted in collaboration with a French company, SepteoProptech[1], operating for the Real Estate domain. The real estate domain plays a relevant role in the French economy, and outperforms the combined weight of the industry and agriculture [Bosvieux, 2018]. It also gathers several subfields such as Finance, Legal or Building industry. Moreover, the real estate industry is currently undergoing a digital transformation, also called PropTech [Siniak, 2020; Starr, 2021], which aims at improving customer experience (e.g., property management, smart home, listing services). Baum [Baum, 2017] described the different categories

[1] https://septeo-proptech.fr/

of Proptech according to several actors of the domain. For instance, Venture Scanners classifies the Real Estate technologies into 12 categories such as Property Management, Construction Management, Home Services, Real Estate Agent Tools or IoT Home. SepteoProptech could be classified in the category of *Real Estate Agent Tools*. Indeed, the company sells a software, called CityScan[2], which uses real estate data to give more insights to real estate professionals in their expertise (e.g., by comparing a property for sale to similar ones) and help sellers to highlight the strengths and weaknesses of their property. Baum also classifies the Proptech companies into three sub-sectors (i.e., Real Estate FinTech, Shared Economy and Smart Real Estate) and three drivers (i.e., Information, Transactions, and control) that are factors that influence the outcome of an activity. SepteoProptech could be seen as a Real Estate FinTech with Information as its driver.

Furthermore, the Real Estate data have been widely used to study the Real Estate market. For instance, the data science competition platform Kaggle[3] hosts a lot of challenges dealing with Real Estate price predictions. The proposed dataset is often structured and represents transactions with property's attributes. Nevertheless, other types of data could be very useful in the study of the real estate domain, and provide much more information, such as the real estate advertisements. A Real Estate advertisement is written by a real estate agent or a seller, to promote a property, that is for sale or for rent, to a potential buyer. It is mainly compound of attributes, images and a textual description which give the opportunity to study the real estate market in different ways and with different methods. Moreover, the advertisements are a fairly exhaustive and updated source of data, and are published online which facilitates data collection (e.g., by crawling housing websites). For all these reasons, SepteoProptech decided to study the real estate market through the real estate advertisements.

Although the real estate advertisements provide many information and have several benefits, the French real estate agents often hide the exact location (i.e., latitude/longitude) to keep the selling mandate since the deal between their agency and the owner might not be exclusive. The use of the advertisements might be limited because of this lack of exact coordinates. Indeed, studying the market implies to deeply know the territory and the price at a fine scale. Thus, the main goal of this thesis is to develop a method to estimate the location of each advertisement using the textual description which often gives clues to the location of a property. For instance, Figure 1.2 shows the example of a real estate advertisement where the location is set to the whole city of Nice while spatial information is given in the text (toponyms are highlighted in blue, geographic feature in purple and spatial relation in green). All the information could be used to retrieve an approximation of the location of the property described by the

[2] https://www.cityscan.fr/
[3] https://www.kaggle.com/

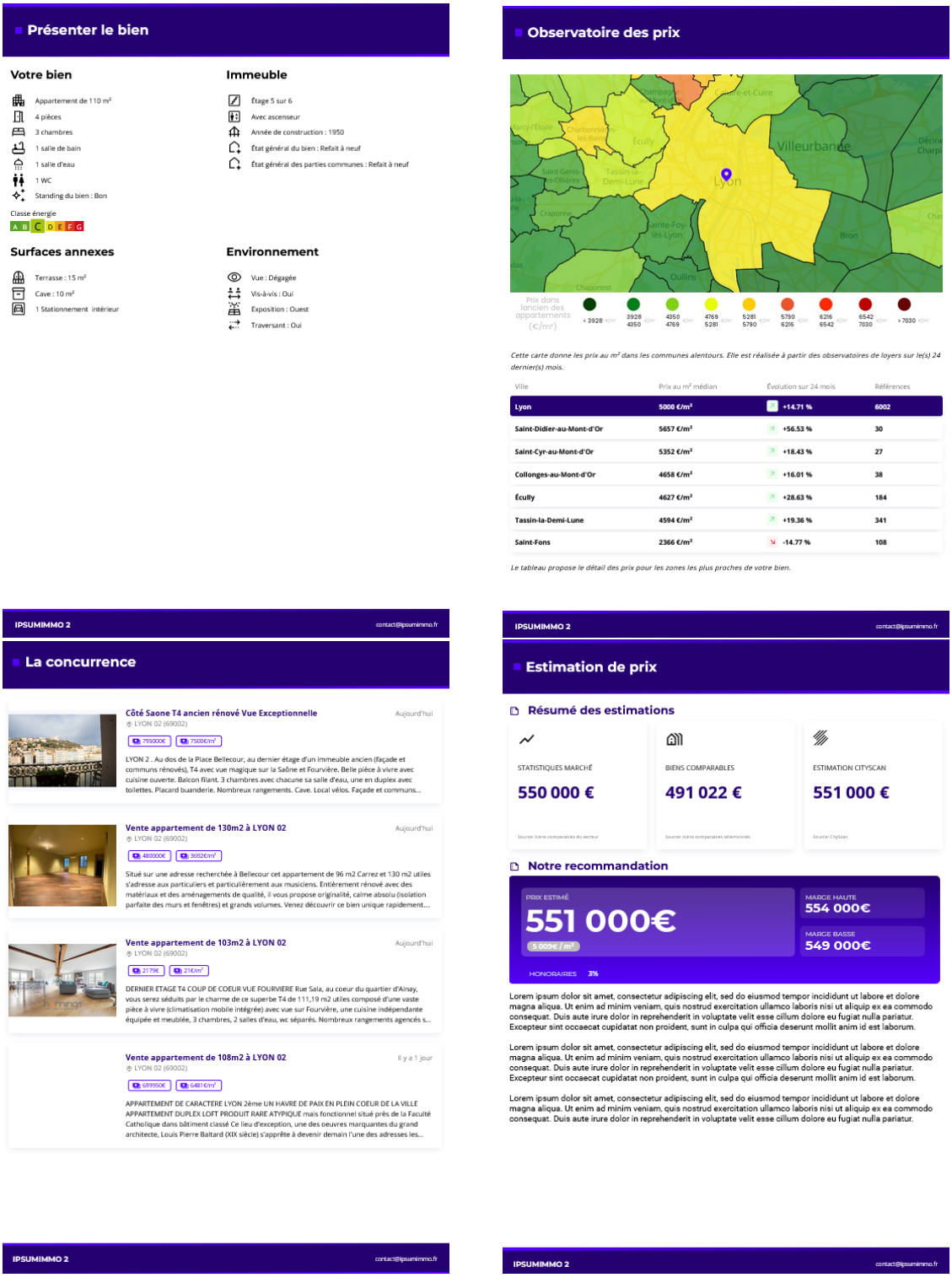

Figure 1.1: Examples of information (e.g., characteristics, price, similar properties) given by CityScan about a property.

advertisement.

Figure 1.2: Example of a real estate advertisement without latitude/longitude coordinates.

2 Challenges and Research Questions

This thesis aims at retrieving the location of an advertisement through the spatial description of its environment found in the text. Nevertheless, this study raises several challenges that go beyond the business issues of SepteoProptech.

First of all, this work is part of a bigger project called Incertimmo[4], started in 2017 between SepteoProptech and the University Côte d'Azur. The goal was to develop an innovative method to create a map of real estate values in the city, integrating a rigorous treatment of the uncertainty of the knowledge and the fusion of data from multiple sources at multiple scales (street, neighborhood, etc.). While the results in the French Riviera were promising, the method relied on transactions data bought for the project, which would have been too expensive to deploy at the national scale. On the other hand, the location is one of the most important purchasing factors of a real estate property, which implies to have a good knowledge of the territory. Several spatial indicators could be used to evaluate a real estate property such as the proximity to schools and transports, the security or the type of neighborhood (e.g., residential). The

[4] https://imredd.fr/en/projet/incertimmo-en/

study of the real estate values in the city helps to increase the knowledge about the territory and the environment of a property. Nevertheless, the project Incertimmo was based on official or corporate data that give a limited view of the territory. Therefore, the real estate advertisements seem to be a good alternative to the transactions and official data. Indeed, they provide the information about the property (e.g., price, number of rooms, etc.) as well as its environment through the real estate professionals or sellers. The data collection is also facilitated by the online version and could help to deploy the method nationally.

However, the study of the real estate advertisements raises several research questions that we try to answer in this thesis. First, a real estate ad is a short text with a particular language and the textual analysis is very challenging because of several specificities, typical vocabulary or errors that we further detail in Chapter 2. Moreover, this thesis focuses on real estate advertisements written in French, which has been less studied than the English language. These particularities pose challenges to automatically extract information, and in particular spatial information, from the text. Thus, we identified the following research questions answered in Chapter 3:

RQ1: What are the typical languages and structures used to write a real estate advertisement ?

RQ2: How is spatial information described in the real estate advertisements ?

RQ3: How to adapt the specificity of the real estate advertisements to existing information extraction methods ?

Secondly, the real estate agents have a very good knowledge of the territory and the advertisements provide a spatial description that reflects their point of view. They often mention the neighborhood, its attributes (e.g., residential, quiet, etc.) and the proximity to amenities (e.g., near the schools). Nevertheless, the real estate agents aims at selling a property to a potential buyer, which implies to use the same place names and exaggerate the proximity to some points of interest. Thus, the spatial description is often compound of vernacular and local places (e.g., *Dowtownn*, *La Banane, Cannes*, etc.). Furthermore, the real estate agents use vague and imprecise spatial relations to describe and exaggerate the location (e.g., near). McKenzie and Hu [McKenzie, 2017b] argue that the proximity-related terms such as *near* are used with a more relaxed definition in Real Estate advertisements than other domains. They compare the use of *Nearby* in the real estate as the one of *Natural* in the food industry in the United States. Since no regulations exist, the meaning of *Natural* does not give any guarantee, which means that the use of *Nearby* might not ensure the proximity. The vagueness and imprecision of the spatial description raise issues to estimate the location of each advertisement. Indeed, we need to retrieve the spatial boundaries of each place

mentioned in the advertisement before approximating the location, but their vagueness and imprecision might lead to disagreement about their spatial boundaries. Thus, we defined the following research questions answered in Chapter 4:

RQ4: How to geocode places extracted from the advertisements, and in particular local places ?

RQ5: How to take into account the spatial relations to geocode a spatial description ?

RQ6: How to deal with the vagueness of the spatial relations ?

RQ7: How to combine imprecise and vague spatial information ?

Finally, the study of the real estate advertisements produces new knowledge, which is valuable for SepteoProptech, but also for broader communities, such as the GIS community, that could exploit it. However, this knowledge is unstructured which is difficult to reason over and share it. Representing knowledge is crucial to effectively solve complex tasks, share, and discover new knowledge. Knowledge graphs are one of the possible structured representations and have several benefits such as a more flexible manner to design and maintain data, reasoning with ontologies or discovery of hidden patterns. Furthermore, the Semantic Web promotes the publication and linking of data on the Web in order to be reused by users for wider applications. However, the information extracted from the real estate advertisements is uncertain and vague and needs a suitable representation to exploit it. Therefore, we answered the following research questions in Chapter 5:

RQ8: How to represent uncertain and vague information to facilitate its interoperability ?

RQ9: How to query and reason over vague spatial boundaries ?

RQ10: What are the potential applications using the produced knowledge ?

3 Contributions and Thesis Outline

In this thesis, we propose to divide the problem of automatically retrieving the imprecise location of the real estate advertisements from the text into three sub-problems: (1) extracting spatial information, (2) geocoding places and combining imprecise information, and (3) representing, storing and reasoning over the data by building a knowledge graph. Figure 1.3 presents the pipeline which summarizes the contributions of this thesis and their links. The remaining chapters of this thesis are organised as follows.

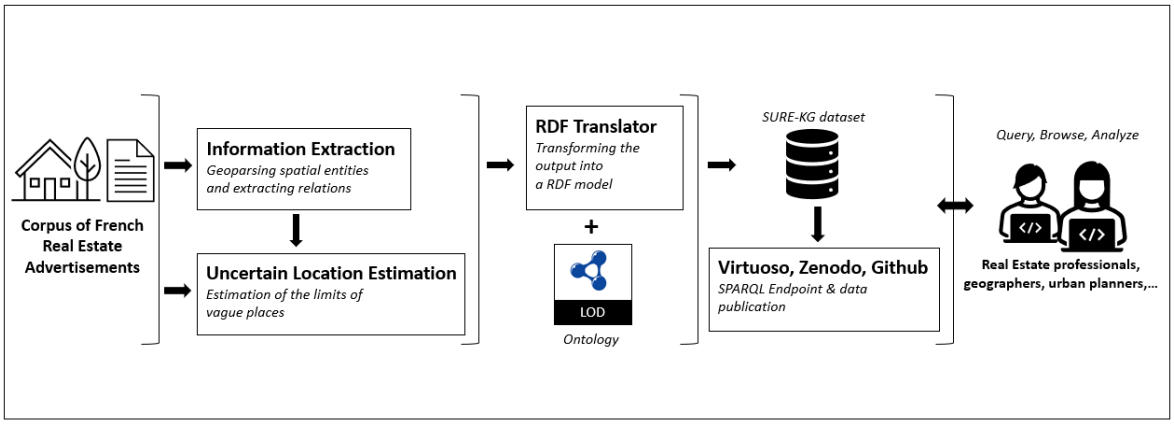

Figure 1.3: Overview of the pipeline and contributions presented in this thesis.

Chapter 2 describes the background and relevant works with our concern of extracting and representing vague spatial information from real estate advertisements. Firstly, we detail the particularities of the real estate advertisements as well as the spatial language used to describe a location. Secondly, we explore the different methods to automatically retrieve information in a structured format. Then, we outline how to tackle imprecision with a focus on fuzzy set theory and spatial vagueness. Finally, we focus on representing, storing and reasoning over data thanks to a knowledge graph.

Chapter 3 addresses the problem of automatically extract and represent spatial information in the Real Estate advertisements which describes the location of a property. We detail the specifities of the language used in the real estate advertisements and their challenges. Then, we present the two steps of our workflow to automatically extract spatial information as well as the annotation guidelines used to create a training dataset. Finally, we discuss the results of our experiments.

Chapter 4 presents the method to estimate the boundaries of the vague spatial descriptions. We first propose to quickly study how to geocode places thanks to existing gazetteers, and we show their limitations given our data. Then, we focus on empirically estimating the boundaries of the different places and the spatial relations, and how to represent their imprecision. Lastly, we describe and evaluate our method to combine the imprecise spatial information to geolocate each real estate advertisement.

Chapter 5 describes SURE-KG, a new knowledge graph built from a real dataset at the core of the industrial application of SepteoProptech. First, we define a new ontology to represent real estate and uncertain spatial information given

several motivating scenarios. Then, we give an overview of the pipeline set up to process the initial corpus and generate the RDF dataset. We also detail the characteristics of the knowledge graph and services made available to exploit it. Finally, we illustrate and discuss potential applications and use cases.

Chapter 6 summarizes the contributions done in this thesis, presents the lessons we learned during the realization of this work, and discusses some research perspectives.

The list of all our publications is available in Appendix A.

Chapter 2

Background and Related Work

Objectives

This chapter aims at describing the background and relevant works with our concern of extracting and representing vague spatial information from Real Estate advertisements, which should be sufficient to guide the reader throughout the remaining of this thesis. Firstly, we detail the particularities of the real estate advertisements as well as the spatial language used to describe a location. Secondly, we explore the different methods to automatically retrieve information in a structured format. Then, we outline how to tackle imprecision with a focus on fuzzy set theory and spatial vagueness. Finally, we focus on representing, storing and reasoning over data thanks to a knowledge graph.

Contents

1 Introduction

This thesis involves connections between several disciplines such as Natural Language Processing (NLP), Geographic Information Science (GIS) and the Semantic Web, all applied to the Real Estate domain. These different disciplines have been widely studied alone or in pairs (NLP/GIS, NLP/Semantic Web, GIS/Semantic Web). However, very few focus on the Real Estate domain and its specifity, such as the uncertainty and vagueness of information.

In this chapter, we describe the background and relevant works with our concern of extracting and representing uncertain spatial information from real estate advertisements. To tackle our problem, we first need to understand how spatial language is used in real estate advertisements. Then, we need to explore the different models to automatically retrieve information in a structured format. Finally, we need to represent spatial information and its vagueness and store it into a reusable knowledge base.

The remainder of this chapter is structured as follows. Section 2 provides an overview of works dealing with real estate advertisements and spatial language. More specifically it shows how cognitive places and qualitative spatial relations are represented in language and how spatial language has been studied in other type of textual data. Section 3 provides an overview of NLP methods for named entity recognition, relation extraction and their application to spatial information. Then, Section 4 describes how to tackle imprecision with a focus on fuzzy set theory and spatial vagueness. Section 5 focuses on how to store and share the knowledge extracted from the advertisements thanks to a Knowledge graph. Finally, Section 6 summarizes and concludes this chapter.

2 Real Estate Advertisements

2.1 Overview

A Real Estate advertisement is written by a Real Estate agent or a seller, to promote a property, that is for sale or for rent, to a potential buyer. It is mainly composed of:

- Property information such as the price, the floor area or the street address;
- Property pictures;
- A description of the property.

A Real Estate advertisement gathers a high variety of data (text, images, semi-structured) which gives the opportunity to study it in different ways and with different methods. Moreover, with the digital transformation of the Real Estate domain, the

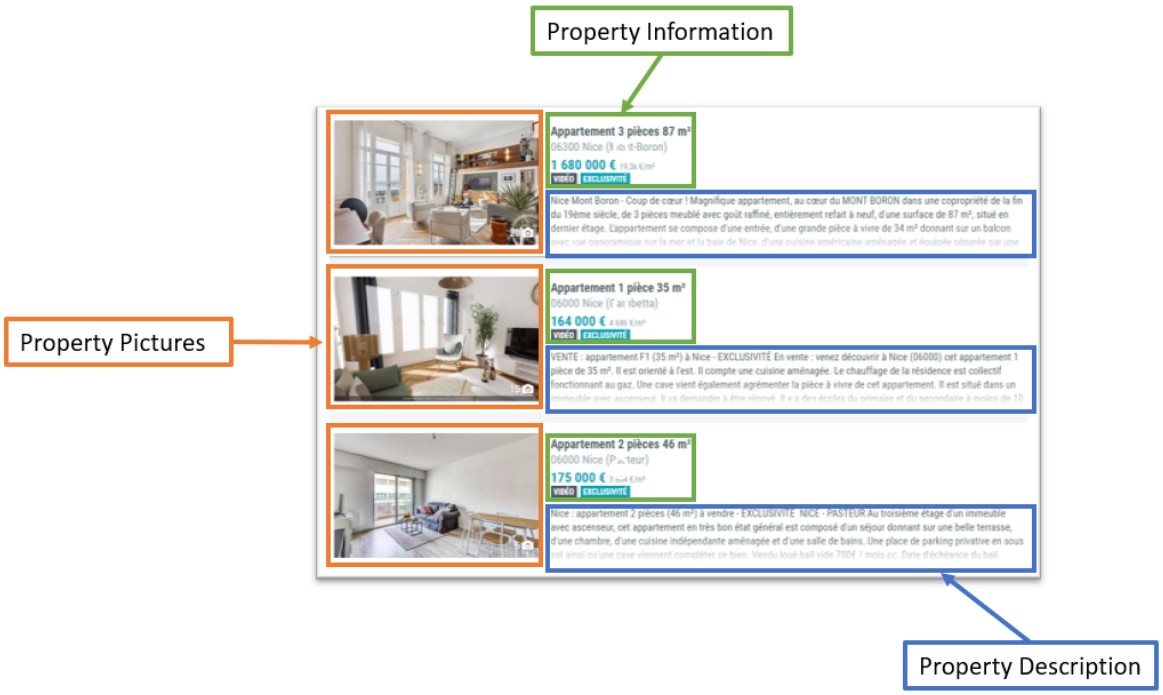

Figure 2.1: Example of an advertisement from the website Bien'ici (http://www.bienici.com/)

number of online advertisements has skyrocketed, their collection has become easy and more and more studies and applications have been carried out. For instance, Kolbe et al. [Kolbe, 2021] use listings, as a substitute for transaction data, to estimate the price. The main advantages to use listings are their abundance, availability and frequent update. The authors perform hedonic regressions by extracting the property information provided by the realtor (i.e., price, floor area, building age, house type, district), but they find a difference between the ask price and the sale price which is a limitation to predict market. Nowak and Smith [Nowak, 2017], and Abdallah and Khashan [Abdallah, 2017] go further in the use of listings into the hedonic pricing model by incorporating textual data. They tokenize the text to add it into the regression model, and show that the price estimation error decreases. They also estimate and quantify the impact of certain words and phrases on the price. Other studies [Carrillo, 2008; Benefield, 2011; Yu, 2021] focus on the pictures to better understand a property price and time-on-market.

While the property information and pictures could be used in the price estimation, the property descriptions have wider applications in other domains. A property description is often a short text dealing with the characteristics of a home, its location, conditions of sale and contact details. A potential buyer will first select a property according to the basic information and pictures, but then will read the description to understand why a property is more interesting than other homes on the market.

Therefore, the description is really important to encourage a buyer to visit a home. Moreover, the realtors have a good knowledge of the market and often target a subgroup of buyer (e.g., families, single person, etc.) by putting forward attributes or location information (e.g., proximity to specific amenities) of interest to their target. These two types of information could be extracted to recommend a real property which best fits the buyers' needs. Bekoulis et al. [Bekoulis, 2018] propose to extract and structure useful characteristics of a property from the text. They develop a Named Entity Recognition model to retrieve information and use Dependency Parsing to find relations between the characteristics. To do so, they can rebuild a home and its subspace (e.g., kitchen, bedrooms, etc.), and keep the human reading effort limited. Harrison and Khazane [Harrison, 2022] also focus on the recommendation of Real Estate properties using textual data from Zillow's listings. They construct a hierarchical taxonomy of the attributes but also of the amenities surrounding the property. Indeed, the location is one of the most important purchasing factors and should be included in the recommendation.

Regarding the location description, other studies have been conducted in other domain such as Geographic Information Science. For instance, Hu et al. [Hu, 2019] harvest vernacular place names from geotagged advertisements. They detect local place names and then estimate their location comparing different geospatial clustering methods. This work could enrich gazetteers that do not often contain vernacular places, while the latter play an important role in many applications (e.g., locating victims in disaster response). Despite the presence of local place names in other data sources, the authors argue that these data are often noisy and not focusing on the description of the place. For example, they point out that a tweet geotagged to a neighborhood can deal with any topic, not necessarily linked to the local neighborhood. On the other hand, a real estate agent highlights the location convenience by describing the neighborhood and the nearby amenities. Moreover, the real estate agent can be seen as a local person and the advertisement can be seen as reflecting the perspectives of the inhabitants.

2.2 Spatial Language in Property Descriptions

Many studies aim at detecting location in textual data including social media [Grace, 2021; Hu, 2021], web pages [Jones, 2008] or fictional novels [Moncla, 2017]. Most of these works typically focus on the extraction of toponyms. A toponym, or place name, is the name used to define a spatial named entity, and that refers to the proper name of a place (e.g., Paris, Nice, etc.). One of the main reasons to detect a toponym is to easily retrieve its corresponding latitude/longitude coordinates, as it is often stored in geographical databases such as gazetteers (e.g., GeoNames or OpenStreetMap) detailed in Section 5. However, toponyms form a vary small part of how a location is described.

For example, in the sentence *'the beach close to the Promenade des Anglais'*, the location description consists of a geographic feature (*beach*), a spatial relation (*close to*) and a toponym (*Promenade des Anglais*). The sole extraction of the toponym could lead to significant errors in georeferencing the place. Furthermore, Purves et al. [Purves, 2018] point out that differences may occur in the way to describe a location according to the purpose, the language or the type of communication. For instance, the purpose of a location description in a real estate advertisement is not the same as the purpose of an itinerary description. The former aims at promoting a neighborhood while the latter gives instructions to travelers to find their way.

Therefore, in the literature on spatial natural language ([Herskovits, 1986; Talmy, 2000; Coventry, 2004]), the linguistic representation of a place has been studied and defined by three elements: the object to be located (or the figure), a spatial relation and the reference object (or the ground). Lesbegueries et al. [Lesbegueries, 2006] classify a place into two main categories: Absolute and Relative. An absolute place is a place-name that could be associated with its type (e.g., 'Riquier train station', 'Nice Castle'). A relative place is defined as an absolute place linked to a spatial relation (e.g., 'Next to Riquier train station'). Similarly, Bennett and Agarwal [Bennett, 2007] identify four ways to describe a place:

- Place-Names: the easiest way to describe a place is to use its proper name (e.g., Nice, Promenade des Anglais);
- Place-Like Count Noun: a common noun that could be regarded as a place and is capable of locating other objects (e.g., city, neighborhood, area);
- Locative Property Phrases: sentences compound of a preposition referring to a spatial relation and a reference to one or more objects (e.g., in Nice, between the train station and the university);
- Definite Descriptions: nominal expressions referring to places (e.g., the train station, the beach close to the Promenade des Anglais);

Their categories also include non-named places (e.g., the train station) that are often used to locate a place in the real property descriptions. The following examples show the diversity of place descriptions in real estate advertisements.

1. *Quartier Cimiez - 2 pièces de 45 m^2 [...] .*

   Cimiez District - 1-bedroom of 45 sqm [...] .

2. *Proche de la place Garibaldi et des commerces, transports, restaurants.*

   Close to Garibaldi Square and all shops, transportation and restaurants.

3. *Situé à 10 minutes à pied du centre-ville et de l'avenue Jean-Médecin.*

   Located 10 minutes walk from the city center and Jean-Médecin Avenue.

In the first sentence, the property is located through an absolute place (*Cimiez District*) whereas the two other phrases refer to the proximity to toponyms or amenities. Moreover, nominal expressions are used to describe places (e.g., city center, shops) and different type of spatial relations (e.g., close to, 10 minutes). In addition, the last sentence (3) mentions the mode of transportation (i.e., walk) to clarify the spatial relation. Thus, since our goal is to locate a real property, we have to study spatial relations, non-named places as well as toponyms.

Spatial Relations

A spatial relation describes the location of an object by specifying its direction with respect to a reference object whose location is known [Landau, 1993; Carlson-Radvansky, 1999]. A spatial relation could express different configurations in space such as proximity (e.g., close to), adjacency (e.g., next to), overlap (e.g., in) or orientation (e.g., north of). The term used to indicate the spatial relation is often a preposition. Landau and Jackendoff [Landau, 1993] identify approximately 80 to 100 spatial prepositions in English to specify a spatial configurations. Dittrich et al. [Dittrich, 2015] and Stock et al. [Stock, 2022] also pinpoint verbs, adverbs or adjectives as possible spatial relation terms. Table 2.1 provides some examples of spatial relations in real estate advertisements. We find the classic type of relations (proximity, overlap, etc.) and the visibility that is often cited to promote a property (e.g., sea view). Moreover, the terms are mainly prepositions or adverbs although some terms are transformed into proper nouns. For instance, *North Nice* refers to the north of Nice but can be seen as a new toponym. Last, the real estate agents may not use a spatial relation to refer to the neighborhood or street where the property is located. They only write the name of the place (e.g., Cimiez District) followed by the attributes of the property whereas it consists of an overlap relation. The inclusion in the place is implied by its mention.

As mentioned before, a spatial relation gives the location of an object by specifying its direction with respect to a reference object. Therefore, we need to represent the spatial relation in terms of geometric or topological features to find the coordinates of the object to be located. For instance, Randell et al. [Randell, 1992] and Egenhofer [Egenhofer, 2005] propose several models for topological relations (Region Connection Calculus and 9-intersection-model) that refer to relations of overlapping or adjacency. Clementini [Clementini, 2019] describes a conceptual model to represent spatial relations according to the level of representation, the geometric space or cardinality. The geometric space is classified in three hierarchical categories: topological, projective and

Class of spatial relations	Example
Proximity	***Proche de*** *la plage* (**close to** the beach), *Les commerces sont* ***à proximité*** (the shops are **nearby**) ***A 10 km de*** *l'aéroport* (**10 km from** the airport) *L'université est* ***à 5 minutes*** (the university is **5 minutes away**)
Overlap	***Au cœur du*** *quartier Gambetta* (**In the heart of** Gambetta district) , ***Dans*** *le centre-ville* (In the downtown), ∅ *Quartier Cimiez* (∅ Cimiez District)
Adjacency	***A côté de*** *la gare* (**next to** the train station), ***Entre*** *la pharmacie et l'école* (**between** the drugstore and the school)
Orientation	***Au sud de*** *Nice* (**south of** Nice), *Nice Nord* (**North** Nice)
Visibility	***Vue*** *sur la mer* (sea **view**)

Table 2.1: Examples of spatial relations in the real estate advertisements

metric relations. Besides topological relations, the projective and metric relations allow to respectively represent orientation and cardinal relations, and distances (e.g., 100 meters). However, this representation assumes that the spatial relations are independent of the context. Carlson et al. [Carlson-Radvansky, 1999] demonstrate that the reference object significantly affects the meaning of the spatial relation and its representation could be different. Stock et al. [Stock, 2022] focus on the spatial relation terms that might have multiple sense, vary in meaning by context or contain vagueness and ambiguity. They differentiate these spatial relations to the classic ones found in geographical information systems such as inside or overlap. Moreover, in the context of real estate, the location description is often exaggerated and the spatial relations used more liberally. The real estate agents prefer mentioning the proximity to popular and well-reputed places in order to arouse positive reactions from potential buyers. Thus, McKenzie and Hu [McKenzie, 2017b] suggest that proximity-related terms such as *nearby* are used with a more relaxed definition in Real Estate advertisements than other domains. They compare the use of *nearby* in the real estate as the one of *natural* in the food industry in United States. Since no regulation exists, the meaning of *natural* does not give any guarantee. Overall, the context in real estate advertisements is very important to represent the spatial relations. This challenge is discussed in Section 4.

3 Information Extraction

The goal of automatic extraction of information is to transform unstructured data into a structured format. It often focuses on the detection of named entities, relationships or attributes in a text [Sarawagi, 2008] and, it is highly related with Natural Language Processing (NLP). The growth of online textual resources has contributed to its development, and particularly for the English language since NLP progress has been mainly made for this language. Nevertheless, more and more languages have been covered, including French [Abeillé, 2003; Ortiz Suárez, 2020; Martin, 2019; Le, 2019].

In our work, we faced three challenges: (1) the French language, (2) the specific language style, and (3) the spatial entities. Several works deal with the French language in other domains [Barrière, 2019; Jabbari, 2020; Copara, 2020]. For instance, Barrière and Fouret[Barrière, 2019] compare different deep neural network models to retrieve Named Entities applied to French Legal texts. They demonstrate that a BiLSTM-CRF architecture combined with text representations gives the best results for this task. Furthermore, a property description has a specific language style that might not follow usual grammar rules such as social media messages [Grace, 2021; Hu, 2021]. It is often a succession of facts that may not contain a subject or a verb. Bekoulis et al. [Bekoulis, 2018] extract a structured description of real estate properties and their attributes by annotating a corpus of real estate advertisements to train traditional models. Finally, the third challenge is to deal with spatial and geographic terms that could be seen as a specific case of Information Extraction (IE). For instance, Hu et al. [Hu, 2019] propose a model to extract local place names in real estate advertisements in English. Similarly, Moncla et al. [Moncla, 2014] extract spatial entities from hiking descriptions written in French.

In the following, we detail the Named Entity Recognition and Relation Extraction tasks and then, we discussed their application to geospatial descriptions.

3.1 Named Entity Recognition

Named entity recognition (NER) is the task of detecting and classifying named entities in text into predefined categories. It is often the first step in information extraction (IE). The term "Named Entity" (NE) was first used at the sixth Message Understanding Conference (MUC-6) [Grishman, 1996] to retrieve the proper nouns of the people, locations and organizations. They also proposed to recognize time, currency, and percentage expressions. Nadeau and Sekine [Nadeau, 2007] point out that the definition of NE is quite restrictive and could not handle specific domains. Sekine and Nobata [Sekine, 2004] try to cover most of the names in newspapers and define about 200 hierarchical categories. Nowadays, NER is widely applied in different domains such

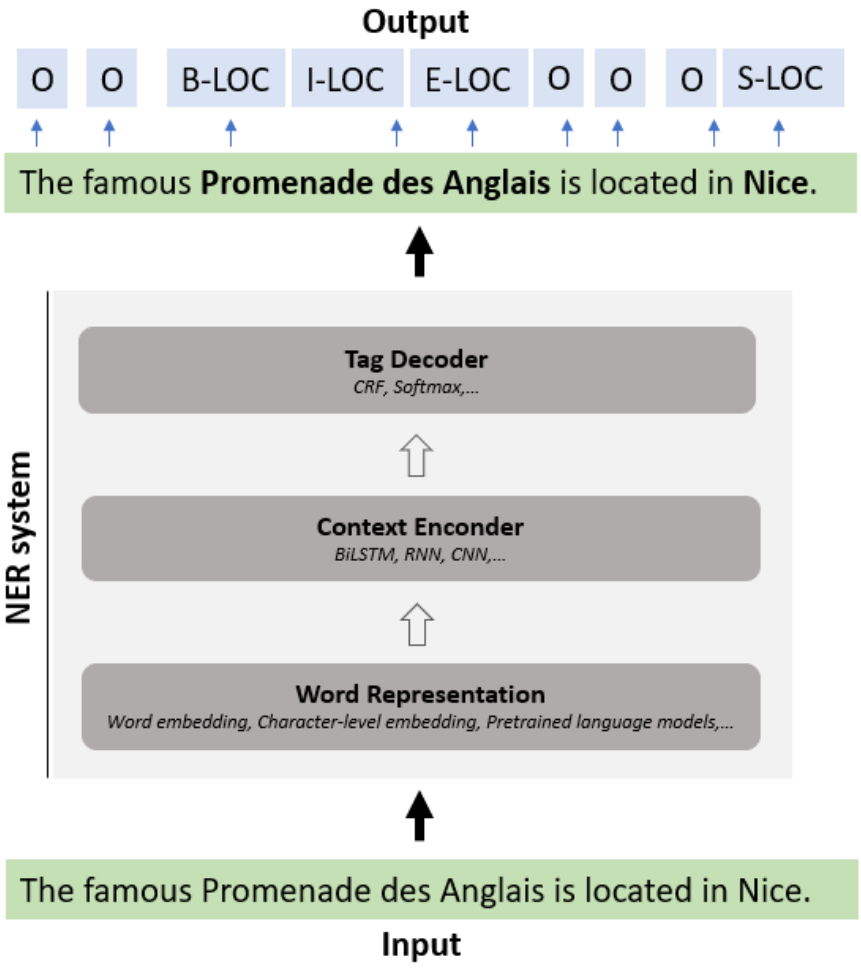

Figure 2.2: NER System Overview

as biomedical, newspapers or social media, which has led to classify NEs into two categories: generic (e.g., person, location) and domain-specific (e.g., proteins, genes).

The NER systems were first based on hand-crafted rules designed by knowledge dictionaries or syntactic-lexical patterns [Kim, 2000; Humphreys, 1998; Aone, 1998; Krupka, 1998]. This approach does not require any annotated data as they rely on lexicon resources and domain-specific knowledge. However, the results are often limited due to the incompleteness of dictionaries, and a system cannot be used in other domains because the rules are domain-specific. Other traditional approaches are data-driven and based on learning techniques. Two types of systems can be distinguished: unsupervised and supervised. Unsupervised techniques [Nadeau, 2006; Collins, 1999], typically clustering, use lexical resources, lexical patterns and statistics to gather named entities into groups based on their similarity. On the other hand, supervised systems learn a model from annotated data to recognize some patterns. Many different models have been tested, such as Decision Trees [Szarvas, 2006], Support Vector Machines [Takeuchi, 2002] or Hidden Markov Models [Bikel, 1997], and give good results in many domains. Nevertheless, both unsupervised and supervised techniques rely on feature-engineering (e.g., word vector representation, list lookup features, document and corpus features), which is crucial to make good predictions but it requires a domain expertise and engineering skills.

In recent years, models based on deep learning have been more and more developed, and achieved top performance [Huang, 2015; Habibi, 2017; Xu, 2018]. One of the best advantages of deep learning is the capability to automatically learn complex features and representations from raw data and, thus to skip the feature-engineering step. The NER task could be seen as a sequence labeling problem (i.e., the text is a sequence of words to be labeled with tags). A typical architecture of sequence labeling is a bidirectional long-short term memory (BiLSTM) with a sequential conditional random layer [Lample, 2016].

BiLSTM-CRF Model

Long-short term memory (LSTM) [Hochreiter, 1997] is a type of recurrent neural network (RNN), mainly used on sequential data (e.g., time series, video, speech recognition) which allows information to persist by learning order dependencies. RNNs tend to foster most recent dependencies while, in theory, they are able to learn long dependencies. Therefore, the LSTM architecture has been developed to overcome RNNs' failure by adding a memory cell and using input, output and forget gates to determine information to give to the memory cell and to forget across time steps. Formally, the implementation of the LSTM is defined by:

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1}), \tag{2.1}$$

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1}), \tag{2.2}$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1}), \tag{2.3}$$

$$N_t = \tanh(W_c * x_t + U_c * h_{t-1}), \tag{2.4}$$

$$c_t = f_t * c_{t-1} + i_t * N_t, \tag{2.5}$$

$$h_t = o_t * \tanh(c_t), \tag{2.6}$$

where x_t is the input vector at the current timestamp t, i_t, o_t, f_t, respectively the input, output and forget gates, N_t is the new information passed to the cell state c_t, and h_t is the current hidden state. σ denotes the sigmoid function, and W and U are the weight matrices associated with the input and hidden state. The gates constitute the three parts of the LSTM: (1) the first step decides if the information from the previous time step should be kept or forgotten by the forget gate (Equation 2.1), (2) then the input gate (Equation 2.2) quantifies the importance of new information (Equation 2.4) carried by the input, and the memory cell (Equation 2.5) tries to learn it, (3) finally the current hidden state (Equation 2.6) is updated by using the memory cell and the output gate (Equation 2.3). The hidden state might be seen as the short term memory and the memory cell as the long term memory.

A bidirectional LSTM (BiLSTM) is a combination of two LSTM layers that use the inputs from both directions simultaneously to better capture the context. The first layer uses past information via forward pass, while the other one uses future information via backward pass. The outputs from both LSTM layers are combined by operations such as average, sum, multiplication, or concatenation, and yield to a representation of the context surrounding each token.

The final stage of the NER model is the tag decoder that takes the BiLSTM output as input to predict a sequence of tags. Conditional random field (CRF) [Lafferty, 2001] is a discriminative model to label sequence data by taking into account context instead of predicting label by label independently. The predictions are modelled as a graph which represents the presence of dependencies between the predictions. During the training phase, the log-probability of the correct tag sequence is maximized [Lample, 2016]. To do so, the score of each sequence of predictions y, for an input sequence X, is computed as follows:

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}, \tag{2.7}$$

where A_{y_i, y_j} is the transition score from tag i to tag j, and $P_{i,j}$ is the score of the j^{th} tag for the i_{th} token. Then, a softmax function and a log-operation are applied to get a probability. Finally, the output sequence is the one that obtains the maximum score.

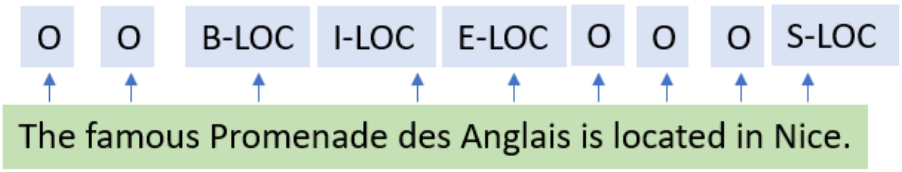

Figure 2.3: Example of IOBES tagging scheme

The IOBES Format

To assign a tag to every token in a sentence, as the BiLSTM-CRF does, the so-called IOB format (Inside, Outside, Beginning) is often used. A named entity could span several tokens within a sentence and need a special tagging scheme. The IOB format decides whether a token begins (B) a span, is inside (I) or outside (O) a span. A variant of IOB format, is the IOBES tagging scheme that includes singleton entities (S) and specifies the end of a named entity (E). Using this format, tagging a token as inside (I) a named entity with high confidence will ensure that the subsequent token is inside (I) or the end (E). For instance, Figure 2.3 shows an example of the IOBES format to tag location (LOC) in a sentence. The named entity *Promenade des Anglais* is compound of three tokens that are tagged as beginning (B-LOC), inside (I-LOC) and end (E-LOC) whereas *Nice* is a unique token and tagged as a singleton (S-LOC).

Text Representation

The NER systems take as input a vector representing the sentence to be machine-readable and understandable. For instance, the easiest representation is the one-hot vector, which consists of zeros for any components and one for the component corresponding to the word. However, this representation gives completely different representations for two similar words and suffers from the curse of dimensionality in that it needs as many components as the size of the dictionary. On the other hand, distributed representations represent each word as a low-dimensional vector where each dimension represents a latent feature, automatically learned from the text. This representation captures semantic and syntactic properties of the word, and could be used to compare the semantic similarity of two words (e.g., using the cosine similarity function). For instance, Mikolov et al. [Mikolov, 2013] develop two word-level representations, namely the continuous bag-of-words (CBOW) and skip-gram (SG), that belong to the Word2Vec algorithms. These representations use a neural network model that is trained on a corpus to learn word associations. Some studies [Yao, 2015; Zhai, 2017; Zhou, 2017] use this type of word representations combined with their NER model. Instead of only considering word-level representations, other NER systems [Tran, 2017; Kuru, 2016] take a character-based word representation as input. Indeed, the character-level representation is very useful to represent sub-word-level information such as prefix and

suffix and handles out-of-vocabulary words unlike Word2Vec. Nevertheless, both word and character-level representations are non-contextual dependent which means a word could not have a different representation and meaning according to its surrounding and context. Akbik et al. [Akbik, 2018] propose a contextualized character-level representation based on a LSTM architecture. This representation allows a word, in particular polysemous words, to have several embeddings depending on the context where they occur. A critical limitation of this model is the fixed length of the vector, which leads the network to overlook large sentences. Recently, Transformers have been developed and rapidly showed their efficiency compared to the other models. They are very useful to process sequential data and their parallel attention layer provides context for any position in the input sequence. Moreover, Transformer Language Models [Devlin, 2018; Yang, 2019; Brown, 2020] are trained on huge corpora usually through self-supervised learning. Self-supervised learning is a type of training that creates labels directly from the input data. For example, some large language models generate embeddings by asking the model to predict masked tokens in a sentence (MLM) or to find which sentence follows the other one between two sentences (NSP). While these models reach high performance, some limitations have arisen: the need of a huge training set on the one hand and limitation of their application to specific domains or tasks due to pre-trained models on general domain on the other hand. Thus, the model has to be fine-tuned in a supervised way on a specific task.

3.2 Relation Extraction

Relation extraction (RE) is the task of detecting and extracting semantic relations between different entities. It often follows NER and gives a structured representation of the information found in a text. For instance, in our work we mainly focus on the spatial relations (i.e., located in, near, etc.). The RE systems are often divided into two categories: rule-based and ML-based. In recent years, a large number of works [Kambhatla, 2004; Culotta, 2004; Zhou, 2007] has used ML-based methods often combined with textual features, including Part-of-Speech (POS) extraction and dependency parsing. In grammar, a POS is a category of words that have similar grammatical properties (e.g., verb, noun), and dependency parsing is the process to analyze and extract the grammatical structure in a sentence in the form of grammatical dependencies. A grammatical relation contains a head, a dependent, which modifies the head, and a tag, such as the Universal Dependency Relations [Nivre, 2016], describing the nature of the grammatical function. The output of the dependency parsing is a dependency tree that starts with a root node and where each arc indicates the grammatical relation between two words. Several methods have been developed to extract relations from the dependency tree. Bunescu and Mooney [Bunescu, 2005]

stipulate that if two entities in a same sentence have a semantic relation then it is mostly concentrated in the shortest path of the dependency tree between the two entities. They have proven that the shortest path offers a very condensed representation of the information needed to estimate their relationship. A similar approach [Hassan, 2014] has been developed to find frequent sub-graph from the dependency tree in order to discover patterns for each relation. A drawback of these methods is the pre-processing part based on a linguistic analysis, which could not be perfect and introduces an error in the classifier. Therefore, the use of deep learning is more and more frequent in the field of RE. For instance, Nguyen and Grishman [Nguyen, 2015] replace the linguistic features analysis by a convolutional neural network (CNN), which automatically learns features from sentences and minimizes the dependence on external NLP resources. Nevertheless, neural networks need huge annotated corpora for training, and might not be effective for relationships occurring in few examples. Finally, other approaches exist, such as ontology-based [Schutz, 2005] or unsupervised methods [Hasegawa, 2004], but they are beyond the scope of this work.

3.3 Spatial Information Extraction

Geoparsing, also known as Toponym Recognition [Leidner, 2008], is a subtask of NER applied to geographic terms in various types of text such as travel blogs [Adams, 2012], social media in emergencies [Grace, 2021; Hu, 2021], real estate advertisements [Hu, 2019], or fictional novels [Moncla, 2017]. Leidner and Lieberman [Leidner, 2011] list three main approaches: (1) Gazetteer Lookup Based, (2) rule-based and (3) ML-based methods. The first approach searches for occurrences of toponyms found in a dictionary of place names, also called a gazetteer, and has been broadly used in many studies [Li, 2002; Stokes, 2008; Lieberman, 2011; Alex, 2019]. Gazetteers, such as GeoNames or OpenStreetMap, have been more and more developed thanks to open collaboration, and are easily available over Web services and linked data. These resources contain millions of place names around the world, including France. However, gazetteers do not contain all places because of their relative insignificance to a gazetteer covering a large area (e.g., world gazetteer) or their vernacular nature (e.g., abbreviations, non-official names). Moreover, a place name is not the only way to describe a place, as pointed out in Section 2. Gaio and Moncla [Gaio, 2017] proposed the concept of Extended Named Entity (ENE), that is composed of proper names (e.g., Nice) or descriptive proper names (e.g., the city of Nice) and several levels of overlapping (e.g., the castle of the city of Nice). The authors used a hybrid solution combining POS, finite-state transducers and gazetteers. Syed et al. [Syed, 2022] extract relative spatial information (e.g., north Nice, 5 km from Cannes) using the SpaCy Python library combined with hand-crafted rules according to the type of spatial relations (i.e., cardinal, ordinal and topological).

Finally, a recent work [Wang, 2019] points out that geoparsers are often off-the-shelf NER systems and, they are not designed for the language irregularities often found in social media messages. The authors propose a model, called NeuroTPR, based on a BiLSTM-CRF architecture specifically designed to retrieve toponyms in social media messages.

The task of geoparsing is often followed by geocoding, which is the challenging task of retrieving the spatial representation of a text-based description of a location. Buscaldi [Buscaldi, 2011] classifies the geocoding approaches into three categories: (1) map-based, (2) knowledge-based and (3) data-driven approach. The knowledge-based approaches have been mainly developed [Overell, 2008; Buscaldi, 2008; Lieberman, 2012] but are limited with the completeness of the gazetteers and could not handle unknown places. On the other hand, DeLozier et al. [DeLozier, 2015] and Alex et al. [Alex, 2019] design methods, without using gazetteers, that rely on the geographic distributions of words over the surface of the earth using Wikipedia and travel blog articles. Jones et al. [Jones, 2008] propose to enrich gazetteers with vague places extracted from web pages and spatial density estimation methods. Vague places are commonly employed by users and do not correspond to the official place names recorded within typical gazetteers. Therefore, their enrichment is important to improve the quality of place name-based information retrieval. However, modelling vague places might need a suitable representation, which deals with uncertainty and vague boundaries, as discussed in Section 4.

4 Modelling Vague Places

Imperfection in data mainly arises from uncertainty and imprecision [Smets, 1997], and could be removed, tolerated by a robust algorithm or modeled. Uncertainty results from ignorance and describes the degree of knowledge required to decide if a statement is true or false. It arises when there are multiple possibilities or when there is limited information available. For instance, when predicting the weather, there is uncertainty because meteorologists cannot be absolutely certain about the exact conditions in the future. Uncertainty is often associated with probability. On the other hand, imprecision refers to the lack of exactness or precision in measurements, data, or descriptions. It occurs when there is inherent variability or limitations in the measurement process. We distinguish two types of imprecision: with and without error. The first one refers to inaccuracy of the information and could be measured whereas the imprecision without error is seen as vagueness and could not be quantified. Vagueness is defined as a lack of clear or precise boundaries, definitions, or meanings of an object or concept. It is often inherent to natural language and depends on the context. For instance, terms

like 'tall' or 'small' are vague because their meanings can vary depending on subjective or contextual factors. A popular solution to represent vagueness is fuzzy set theory, introduced by Zadeh [Zadeh, 1965], whose main idea is to define membership in a set as gradual instead than all-or-nothing.

4.1 Fuzzy Set Theory

Fuzzy set theory has been first introduced by Zadeh [Zadeh, 1965] in 1965, to extend classical set theory, which is to rigid to represent classes, in particular natural language terms since they do not have precisely defined criteria of membership. In classic set theory, an object either belongs to a set, or it does not, and this is represented by a degree of membership equal to 0 or 1. Given S a crisp set and x, y its elements, a subset of S, noted A, is defined by its characteristic function χ_A (2.8). Fuzzy set theory extends this binary function as a continuous one in the [0,1] interval.

$$\chi_A(x) = \begin{cases} 1 & \text{if x} \in \text{A} \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Definition 4.1 (Fuzzy Set)**.** A fuzzy set F of S is defined by its membership function μ_F, which maps all element x of S to the value $\mu_F(x)$ in the [0,1] interval. This value represents the degree of membership of x in fuzzy set F. Thus, the closer the value of $\mu_F(x)$ to 1, the higher the degree of membership of x in F. A fuzzy subset is denoted by $\{(x, \mu_F(x)), x \in F\}$.

A membership function could be used to define several crisp sets in S such as the support $supp(F)$, the core $core(F)$, the height $h(F)$ and the α-cuts F_α, that describe a fuzzy subset [Dubois, 2012].

Definition 4.2 (Support)**.** The support of a fuzzy set F of S, denoted $supp(F)$, is the set where all the elements belongs to F with a certain degree higher than 0:

$$Supp(F) = \{x \in F, \mu_F(x) > 0\} \tag{2.9}$$

Definition 4.3 (Core)**.** The core of a fuzzy set F of S, denoted $core(F)$, is the set where all the elements completely belong to F:

$$Core(F) = \{x \in F, \mu_F(x) = 1\} \tag{2.10}$$

Definition 4.4 (Height)**.** The height of a fuzzy set F of S, denoted $h(F)$, is the largest value reached by the support:

$$h(F) = \sup_{x \in F} \mu_F(x) \tag{2.11}$$

It is not always equal to 1. A fuzzy set is normalized if its height is equal to 1.

Definition 4.5 (α-cut)**.** An α-cut, denoted F_α, is a crisp set where all the elements belongs to F with a degree higher or equal to α:

$$F_\alpha = \{x \in F, \mu_F(x) \geq \alpha\} \tag{2.12}$$

A strong α-cut, denoted F_α^+, refers to a set where all the elements belogs to F with a degree strictly higher than α:

$$F_\alpha^+ = \{x \in F, \mu_F(x) > \alpha\} \tag{2.13}$$

The support is a particular strong α-cut with $\alpha = 0$.

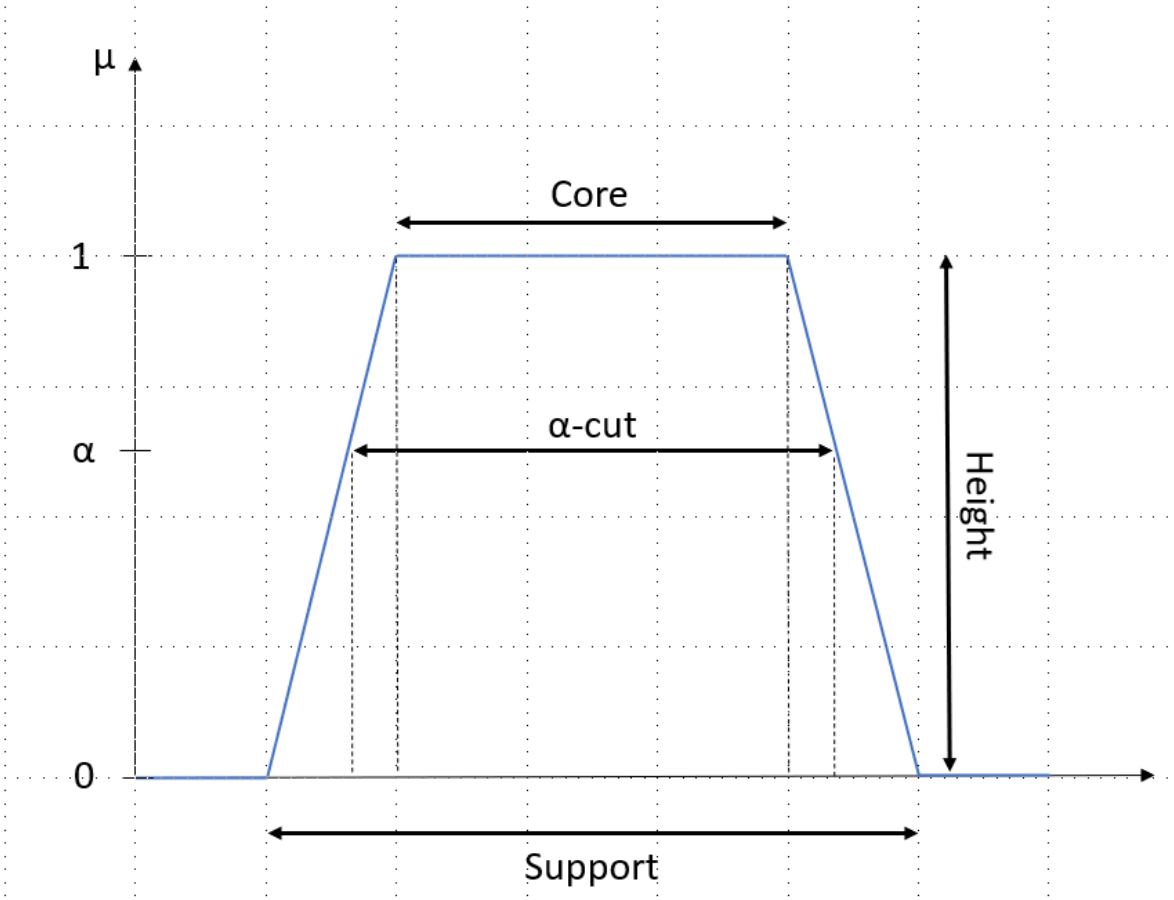

Figure 2.4: Support, core, height, and α-cut of a fuzzy set.

Fuzzy Set Operations

In classical set theory, operators are defined to combine information from multiples sources (e.g., conjunction and disjunction) but their generalization to fuzzy set theory

is not unique. If two information, A and B, have a degree of truth equal to 0.3 and 0.7 respectively, what is the degree of truth of the conjunction 'A and B' ? A large number of operators have been defined [Zimmermann, 2011; Dubois, 1985] but, in this work, we only focus on classic conjunction, disjunction and ordered weighted averaging [Yager, 1988].

First, the conjunction (AND) and disjunction operators (OR) are often modelled by a t-norm (triangular norm) and t-conorm (triangular cornorm). Table 2.2 summarizes the most popular t-norms and their associated t-conorms.

Definition 4.6 (T-norm)**.** A triangular norm [Klement, 2004] is a mapping function $T : [0,1]^2 \to [0,1]$ such that for all $x, y, z \in [0,1]$ the following properties are satisfied:

1. Commutativity: $T(x,y) = T(y,x)$;
2. Associativity: $T(x, T(y,z)) = T(T(x,y), z)$;
3. Monotonicity: $x \geq y \implies T(x,z) \geq T(y,z)$;
4. Boundary Condition: $T(x,1) = x$.

Definition 4.7 (T-conorm)**.** A triangular conorm [Klement, 2004] is a mapping function $S : [0,1]^2 \to [0,1]$ such that for all $x, y, z \in [0,1]$ the following properties are satisfied:

1. Commutativity: $S(x,y) = S(y,x)$;
2. Associativity: $S(x, S(y,z)) = S(S(x,y), z)$;
3. Monotonicity: $x \geq y \implies S(x,z) \geq S(y,z)$;
4. Boundary Condition: $S(x,0) = x$.

The relation between t-norm and its associated t-cornorm is given by:

$$S(x,y) = 1 - T(1-x, 1-y). \tag{2.14}$$

The t-norm and t-conorm generalize the conjunction and disjunction operators to fuzzy sets. One of the most popular t-norm is the minimum, which combines two information by taking the lowest degree of truth, while its dual t-conorm is the maximum, which combines two information by taking the highest degree of truth. They are defined as follows:

$$\begin{aligned} T_M(x,y) &= min(x,y), \\ S_M(x,y) &= max(x,y). \end{aligned} \tag{2.15}$$

T-norm	Dual T-conorm
$T_M(x,y) = min(x,y)$	$S_M(x,y) = max(x,y)$
$T_P(x,y) = x \cdot y$	$S_P(x,y) = x + y - x \cdot y$
$T_L(x,y) = max(x+y-1,0)$	$S_L(x,y) = min(x+y,1)$
$T_W(x,y) = \begin{cases} min(x,y) & \text{if } max(x,y) = 1 \\ 0 & \text{otherwise} \end{cases}$	$S_W(x,y) = \begin{cases} max(x,y) & \text{if } min(x,y) = 0 \\ 1 & \text{otherwise} \end{cases}$

Table 2.2: Popular t-norms and their dual t-conorms

The conjunction and disjunction operators are either too or not enough restrictive, that is why other operators have been defined as a compromise. For instance, Yager [Yager, 1988] proposes the ordered weighted average (OWA).

Definition 4.8 (OWA)**.** The operator OWA is a mapping function $OWA : [0,1]^2 \rightarrow [0,1]$ defined by:

$$OWA(x_1, ..., x_n) = \sum_{i=1}^{n} w_i x_{(i)} \tag{2.16}$$

where $w = \{w_1, .., w_n\}$ is a vector of weight such as $w_i \in [0,1]$ and $\sum_{i=1}^{n} w_i = 1$. The notation $x_{(i)}$ defines the i_{th} ordered information such as $x_{(1)} \leq ... \leq x_{(n)}$.

Remark. We can retrieve some particular operators by choosing the weights of the OWA operator:

- If $w_1 = 1$ and $w_i = 0$ for $i > 1$, the OWA is the *min*.
- If $w_n = 1$ and $w_i = 0$ for $i \neq n$, the OWA is the *max*.
- If $w_i = \frac{1}{n}, \forall j \in [1,n]$, the OWA is the *average*.
- $min(x_1, ..., x_n) \leq average(x_1, ..., x_n) \leq max(x_1, ..., x_n)$

4.2 Spatial Vagueness

Spatial objects do not escape vagueness, and Fisher [Fisher, 2000] argues that it is endemic to our condition and profoundly embedded in our natural language. He takes the concept of proximity, which does not have a proper definition and could be seen as a sorites paradox [Sainsbury, 2009]. The sorites paradox is a paradox resulting from vague predicates and, its formulation often involves a heap of sand, from which grains are removed individually. We do the assumption that removing a single grain does not cause a heap to become a non-heap, then the paradox is to consider what happens when the process is repeated enough times that only one grain remains. Is it still a heap ? If not, when did it change from a heap to a non-heap ? In the case of the concept of

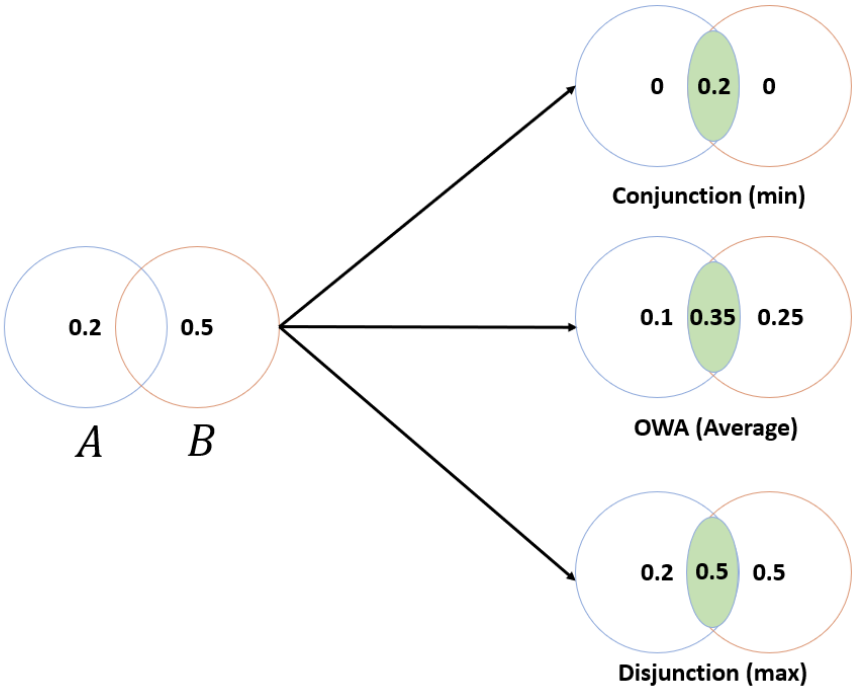

Figure 2.5: Example of the combination of two pieces of fuzzy information with the minimum, maximum and OWA operators.

proximity, if we model the spatial relation *near* as a distance between two points, A and B. Then, if we move one meter away, can we say that A is still near B ? If we move one meter by one meter away, we could say that A is always near B, which is similar to the sorites paradox. Montello et al. [Montello, 2003] classify vague spatial terms into two categories: spatial relations and spatial regions. Vague spatial relations refer to terms as *near*, *around* or *to the north* while vague spatial regions describe cognitive or perceptual regions, such as *Downtown*, different from administrative regions (e.g., states, city).

Earlier approaches have been user-centered on asking humans to draw vague places [Fisher, 1991; Worboys, 2001] . For instance, Montello et al. [Montello, 2003] present the spatial extent of *Downtown Santa Barbara* by aggregating polygons drawn by pedestrians (the 50% and 100% confidence borders). Similarly, Carlson and Covey [Carlson, 2005] study the spatial relations, such as *near* or *left*, and their dependency on the context. They show that the size of an object has an impact on the distance associated to the spatial relation. These experiments are very time-consuming to conduct and analyse, and could not be carried out on a large scale. Therefore, data-driven approaches have been developed to automatically analyze texts. The most readily available source of text is the Web. Jones et al. [Jones, 2008] extract vague places from web pages, assign coordinates and estimate spatial density to approximate boundaries. Their techniques are similar to Montello et al. [Montello, 2003], but the use of the

Web allows for quick data collection. Derungs and Purves [Derungs, 2016] and, Aflaki et al. [Aflaki, 2022] compare distance thresholds for spatial relations by using spatial descriptions, and draw the conclusion that the reference object influences the distance thresholds of the spatial relations.

Despite the numerous studies about vague places and relations, geographic information systems mainly represent spatial objects with sharp and well-defined boundaries (e.g., point, lines, polygons). However, Goodchild [Goodchild, 2000] gives the example of a caller to an emergency dispatcher to show the necessity of efficiently converting a location description to a quantitative spatial position since that could be a matter of life and death. Several models have been proposed to represent and reason about imprecise spatial objects, and could be classified into three categories [Erwig, 1997]: (1) fuzzy, (2) exact, and (3) probabilistic models. A well-known exact model is the *Egg-yolk*, proposed by Cohn and Gotts [Cohn, 2020], and consisting of two regions. The first region is the precise one ('yolk') which is included in the second one ('white') which represents the imprecision. The authors define topological relations based on the RCC-8 model. This approach could be expressed and improved by rough sets [Bittner, 2002]. Fuzzy theory has been more and more studied and used to represent vague spatial objects [Goodchild, 1998; Altman, 1994; Schneider, 2001] since more than two regions could be used to represent the vague object compared to *Egg-yolk* and rough sets. Schockaert et al. [Schockaert, 2011] propose a model to fuzzify spatial and topological relations, and vague regions. The authors also give a model to gather several vague spatial relations and reason over them. Finally, probabilistic models are mainly used to model uncertainty, which is out of the scope of our work.

5 Knowledge Graphs, Ontology and the Semantic Web

5.1 Overview

Knowledge involves the understanding, interpretation and application of meaningful information extracted from data [Schreiber, 2000]. Representing knowledge is crucial to effectively solve complex tasks, share, and discover new knowledge. Knowledge graph is one of the possible structured representations that has gained popularity since the 2012 announcement of Google Knowledge Graph [Singhal, 2012], and have been widely used in applications such as search engine enhancement, question answering, or product recommendation. Graphs have several benefits, compared to other relational or NoSQL databases, such as a more flexible manner to design and maintain data, reasoning with ontologies and discovery of hidden patterns.

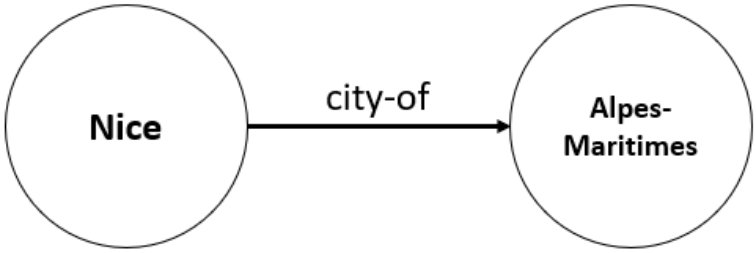

Figure 2.6: Graph representation of '*Nice is a city of the Alpes-Maritimes*'.

Knowledge Graphs

The definition of *knowledge graph* (KG) is not unique and remains contentious [Paulheim, 2017; Ehrlinger, 2016; Ji, 2021]. Recently, Hogan et al. [Hogan, 2021] has proposed to define a KG as '*a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities*'. The structure of a graph could also be modeled in different ways (e.g., directed edge-labelled graph, heterogeneous graph or property graph). In this work, we focus on the directed edge-labelled graph, which is defined as a set of nodes and a set of directed labelled edges between those nodes. For instance, in the statement '*Nice is a city of the Alpes-Martimes*', the nodes are **Nice** and **Alpes-Martimes** and the labelled edge could be *city of*.

A particular directed graph is the Resource Description Framework (RDF) graph [Cyganiak, 2014] which is a standard of the World Wide Web Consortium (W3C) and used by the Semantic Web community. The RDF model expresses statements in the form of triples: $<$ *subject, predicate, object* $>$. The subject is a (web) resource to describe and represented as *Internationalized Resource Identifier* (IRI) [Dürst, 2005] or a blank node. The predicate denotes a property (e.g., attribute, characteristics or relationship between resources) defined by a IRI. The object is the value of the property represented as a IRI, a blank node or a literal (e.g., strings, integers, dates). A RDF graph is composed of a set of RDF triples. A limitation of RDF is that it is a very abstract language and does not provide the semantics of the resources which is crucial for reasoning. Consequently, two standards, RDF Schema (RDFS) [Brickley, 2014] and Web Ontology Language (OWL) [Hitzler, 2009], have been built upon RDF to encode ontologies.

Ontology

An *ontology* is a formal representation of the meaning of concepts and their properties within a domain. The main components of an ontology are individuals, classes and properties. An individual is an instance of a class which could be concrete (e.g., a

```
SELECT DISTINCT ?label
WHERE {
      ?city rdf:type sure:City;
            rdfs:label ?label.
}
```

Figure 2.7: Example of a SPARQL query to retrieve the name of the entities that instantiate the class *City*.

label
"andon"
"antibes"
"ascros"
"aspremont"
"auribeau sur siagne"
"bairols"
"beaulieu sur mer"
"beausoleil"
"belvedere"
"bendejun"
"berre les alpes"
"beuil"
"bezaudun les alpes"
"biot"
"blausasc"
"bollene"
"bonson"
"bouyon"
"breil sur roya"
"briancon"
"brianconnet"
"cabris"
"cagnes sur mer"
"caille"
"cannes"

Figure 2.8: SPARQL query output

person) or abstract (e.g., a number or a word). A class is a concept that gathers a set of individuals with similar characteristics (e.g, the class *Person* gathers all individuals that are a person). A property describes an attribute (e.g., label) or a relationship between concepts to specify how an object is linked to another one in the ontology. The domain and the range define the classes that can be deduced as the type of the subject and the object of a property. As a result, ontologies make it possible to automatically reason about data by providing the relationships between concepts. They also facilitate the querying of data by giving a coherent schema of the concepts. Simple Protocol and RDF Query Language (SPARQL) [Harris, 2013] is a standard to query, retrieve and manipulate information from RDF graphs. SPARQL queries are based on triple patterns and would return resources that match these patterns. Figures 2.7 and 2.8 show an example of a SPARQL query and its output to retrieve the name of entities that belong to the class *City*.

Linked Data

The technologies previously described (i.e., RDF, RDFS, OWL, SPARQL) are at

the core of the Semantic Web which aims at promoting the publication and linking of data on the Web. *Linked data* refers to the collection of interrelated datasets on the Web that can be reused by users for wider applications. To achieve and create Linked Data, Berners-Lee [Berners-Lee, 2006] recommends to follow four principles, called the Linked Data principles, before publishing:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Berners-Lee [Berners-Lee, 2006] also provides a "five-star" scheme to create Linked Open Data (i.e., linked data are also open data) and assess its quality. According to this shcema, a dataset should be:

⋆ Available on the Web with an open licence;

⋆⋆ Available in a machine-readable structured format;

⋆⋆⋆ Available in a non-proprietary structured format (e.g., .csv);

⋆⋆⋆⋆ Published using W3C standards (e.g., RDF);

⋆⋆⋆⋆⋆ Linked to other Linked Open Data datasets.

Figure 2.9 shows an overview of the Linked Open Data cloud[1] gathering all datasets and their interlinks. On June 2023, the LOD cloud recorded 1,600 knowledge graphs including generic and cross-domain KGs such as DBpedia [Lehmann, 2015], YAGO [Suchanek, 2007] or Wikidata [Vrandečić, 2014]. The two first datasets have been automatically built by extracting information from semi-structured data whereas Wikidata has been manually and collaboratively created. These three KGs are major nodes of the LOD cloud. Nevertheless, their general knowledge is also a limit to more specific applications in speciliazed domains such as Real Estate or Geographic Information Systems (GIS).

[1] https://www.lod-cloud.net/

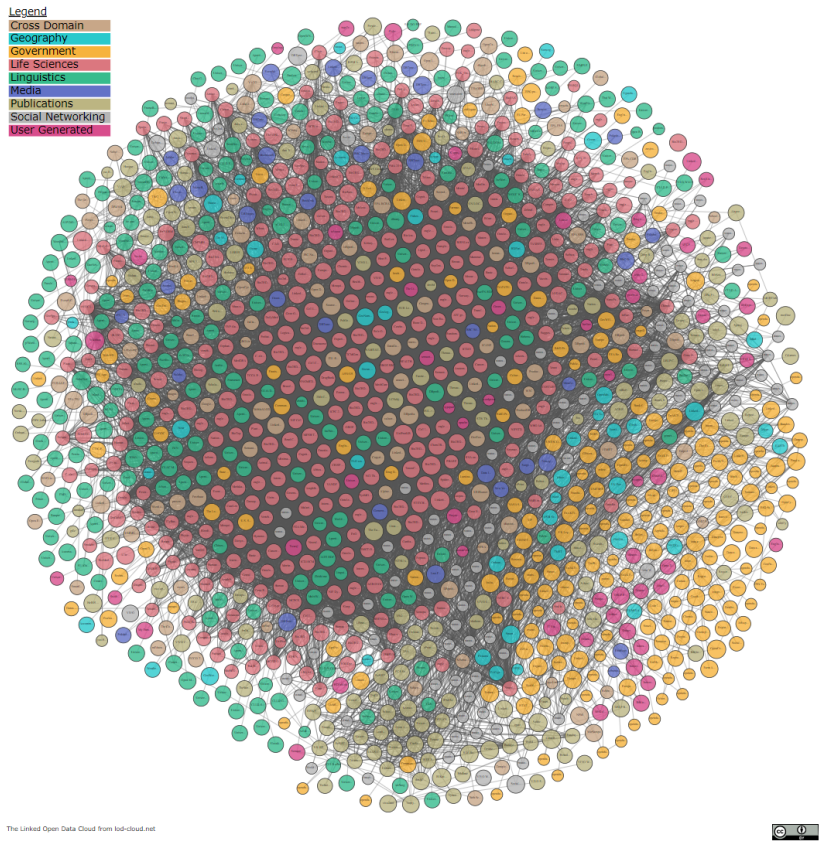

Figure 2.9: The Linked Open Data cloud from https://lod-cloud.net in June, 2023 (1600 knowledge graphs).

5.2 Geospatial Data

In the GIS community, geospatial data have been widely collected and stored in gazetteers that are simple dictionaries or more advanced geographical ontology and, provide place names, geographical features (e.g., city, beach, school) and their spatial footprint (e.g., point or polygon). Gazetteers are often often developed by government agencies from structured data (e.g., the National Geographic Institute (IGN) large databases[2] in France) or collaboratively built (e.g., OpenStreetMap[3]), focusing on a specific domain, application or region, and without any links between them. Consequently, Knowledge Graph and Linked Open Data have also been studied in the realm of GIS to provide more open and cross-domain data. For instance, popular KGs, such as DBpedia and Wikidata, have already integrated spatial entities but they are too generic and lack of coverage. More specialized geographic KGs have been developed such as GeoNames[4] or YAGO2geo [Karalis, 2019] which is an extension of YAGO mainly focused on administrative regions and based on OpenStreetMap (OSM) and other reference datasets such as Greek Administrative Geography (GAG) and Global Administrative Areas dataset (GADM). LinkedGeoData [Auer, 2009] converts OSM data into a RDF graph by using the tags and keys of OSM to create its ontology. Dsouza et al. [Dsouza, 2021] point out that geographic KGs lack coverage of geographic classes. They also propose to build a knowledge graph from OSM data and to create a more comprehensive ontology that could be aligned with the classes of Wikidata and DBpedia. Nevertheless, these graphs are all based on very generic data that could not be suitable for some applications or domains. For instance, Janowicz et al. [Janowicz, 2022] develop their own geographic KG, called KnowWhereGraph, to support environmental applications. They include 27 different data layers from 16 major data sources covering the environmental applications (e.g., climate hazard, wildfire, and air quality). Regarding our application, the location of a real property is one of the major factors in the purchasing decision but the real estate agents often mention vernacular places to describe a location, which are not always included in these graphs (e.g., "city center") [Keßler, 2009].

To unify geospatial representations and data access, the reuse as much as possible of existing vocabularies and ontologies is determining while the plurality of the definitions of a place according to the domain (e.g., place cognition vs place engineering) makes it difficult to design a single ontology [Ballatore, 2016]. Atemezing and Troncy [Atemezing, 2012] review different modeling approaches and show the diversity of representations. For instance, GeoNames uses the SKOS [Miles, 2009] concepts to defines high-level codes (A, H, L, P, R, S, T, U, V) where each letter corresponds to

[2] http://geoservices.ign.fr/
[3] http://www.openstreetmap.fr/donnees/
[4] http://www.geonames.org/about.html

a precise category and classes are attached to these codes. On the other hand, the IGN has developed its own ontology, which does not reuse existing vocabularies and is based on the dataset BDTopo. The ontology classifies spatial entities according to the topographic and administrative use (buildings, road network, green area, etc.). The existing geographic vocabularies always classify the spatial entities according to their nature and topography, which limits their use for a more perceptual and cognitive representation. The Open Geospatial Consortium (OGC) had adopted the GeoSPARQL standard [Battle, 2012] to represent and query geospatial data on the Semantic Web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF and provides an extension to the SPARQL query language for processing geospatial data. The ontology, based on OGC's Simple Features model, is composed of an upper-classes `geo:SpatialObject` and two subclasses, `geo:Feature` and `geo:Geometry`. Feature and Geometry are two core components of geospatial science. A feature is any entity with a spatial location (e.g., train station, beach, school) and a geometry is any geometric shape (e.g., point, polygon, line) used to represent the feature's spatial footprint. A feature is linked to a geometry thanks to the property `geo:hasGeometry`, and could have several geometries. GeoSPARQL proposes two different ways to represent geometry: WKT and GML formats. This standard also extends the SPARQL query language by including topological relationships (e.g., overlaps, touches), between spatial entities. Nevertheless, GeoSPARQL is still limited to represent and query vague places and spatial relations (e.g., near).

5.3 Real Estate Data

The study of the Real Estate domain often involves other aspects such as Finance, Law or Geography, and could focus on different levels (e.g., land or buildings). Previous works have shown that ontology formalization and knowledge graphs of the Real Estate domain depend on the data and use cases. Shi and Roman [Shi, 2018] compare several Real Estate ontologies focusing on different aspects and levels: the land with cadastral data [Sladić, 2013], the legal domain [Paasch, 2005] and the transactions [Stubkjaer, 2017]. The *proDataMarket* ontology [Shi, 2017] gathers these three sub-domains and studies the Real Estate market through the land and the transactions. However, this ontology is not based on up-to-date data and does not study the building and its environment. In other words, it is not possible to search for a real property for sale according to its attributes (floor size, floor level, etc.) and its location. The *NAREO* ontology [Laddada, 2020] tries to answer one of these challenges by describing the neighborhood and the proximity to amenities to recommend a neighborhood according to location and environment criteria. Nevertheless, the authors do not represent the real property itself. Also, they only use official data such as OpenStreetMap and

the French national institute for statistics and economic studies *INSEE*, which do not contain local and vernacular places and the real estate agents' point of view on the environment (e.g., "residential", "quiet", etc.). The NAREO ontology is close to our approach but has not been populated and only focuses on one of our use cases (i.e., retrieving a place according to its proximity to facilities), as presented in Chapter 1.

6 Summary

In this chapter, we presented the background and the relevant literature with our concern of extracting and representing vague spatial information from real estate advertisements. We introduced the real estate advertisements, their particularities and an overview of the studies dealing with this data. We detailed the spatial language used to describe a place (e.g., toponym, geographic feature, spatial relations) and, how it has been studied in other types of text, such as travel blogs or social media. Then, we explored methods to automatically retrieve information in a structured format, and in particular we described the tasks of Named Entity Recognition and Relation Extraction. We focused on the BiLSTM-CRF architecture, which is widely used in NER systems. Moreover, we compared the models used to retrieve spatial information and, we showed that the proposed methods have limits when the spatial description is vague. Thus, we discussed how to overcome vagueness to represent a spatial object. We identified two approaches: user-centered and data-driven. We also detailed fuzzy set theory, which is a popular method to tackle imprecise information. Finally, we presented a brief overview of knowledge graphs and the Semantic Web, their application to geospatial and real estate data, and the challenges to store and reason about vague spatial information. The content of this chapter should be sufficient to guide the reader throughout the remaining of this thesis, which presents (1) our model to extract spatial information, (2) the methodology to estimate and combine the boundaries of vague spatial descriptions and, (3) the pipeline to build a knowledge graph.

Chapter 3

Geospatial Knowledge in Real Estate Advertisements: Capturing and Extracting Spatial Information from Text

Objectives

This chapter addresses the problem of automatic extraction and representation of spatial information in the Real Estate advertisements, which describes the location of a property. We detail the specifities of the language used in the real estate advertisements and their challenges. Then, we present the two steps of our workflow to automatically extract spatial information as well as the annotation guidelines used to create a training dataset. Finally, we discuss the results of our experiments.

Contents

1 Introduction

Text-based geospatial information found in various documents (e.g social media, newspapers, housing advertisements) plays an important role in many geographic applications such as Geographic Information Systems (GIS) enrichment, better understanding and description of our environment [Adams, 2012], or the location of events (e.g., natural disasters) [Hu, 2021]. To capture and extract spatial information from text, Geoparsing applications have been widely developed and are mainly focused on Place-name extraction. Indeed, names are often used by people to refer to places and they can be linked to existing digital gazetteers. However, the gazetteers mostly record official Place names, whereas text documents may contain non-official and local Place-names. Also, non-named entities such as place types might be preferred and used to locate a place. For example, in the Real Estate advertisements, non-named entities give more information about a neighbourhood and its facilities (e.g., near the shops and schools). Another challenge of extracting spatial information is to create a structured knowledge base in order to exploit and reason over it. Relation Extraction is the task of extracting relationships between entities found in the text. It provides a structured representation of information.

In this chapter, we address the problem of automatically extracting and representing spatial information in the Real Estate advertisements, which describes the location of a property. This problem involves the annotation and the detection of spatial entities, and the extraction of relationships between entities. The objective is to propose an automatic workflow to support the extraction of spatial entities and their structured representation. This workflow is divided in two main stages: (1) Named Entity Recognition applied to spatial entities and (2) Relationship Extraction to provide a structured knowledge.

The main contribution of this chapter is the Named Entity Recognition model designed for Real Estate advertisements written in French and which relies on the annotation of different type of entities based on our definitions. The Relation Extraction algorithm is mainly based on the study of the grammatical structure of the advertisements.

The remainder of this chapter is structured as follows. Section 2 proposes an overview of the language used in Real Estate advertisements and its challenges. Section 3 and Section 4 describe the two steps of our proposed workflow to automatically extract geospatial information. Finally, Section 5 presents and discusses the results of experiments based on a corpus of French Real Estate advertisements, and Section 6 summarises and concludes this chapter.

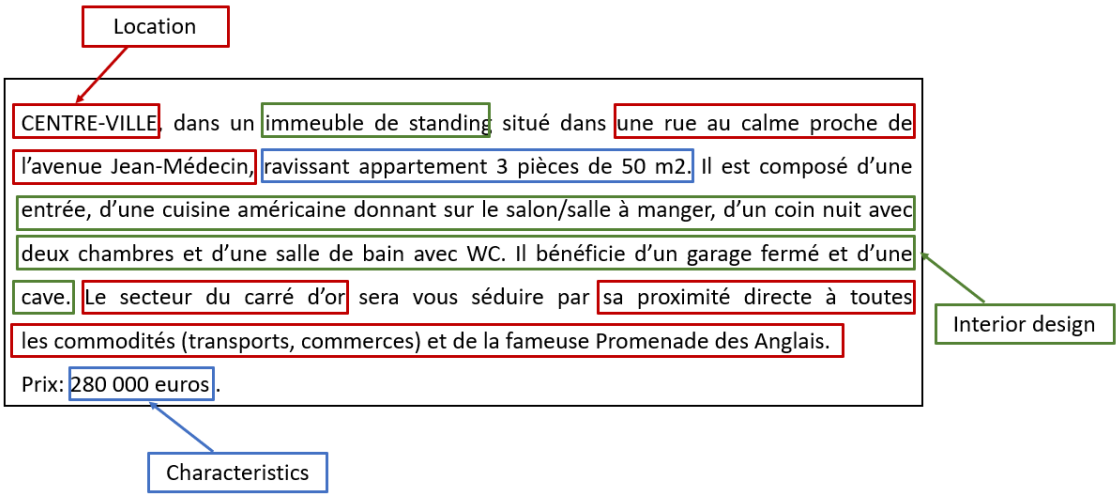

Figure 3.1: Example of the information found in an advertisement.

2 Natural Language in the Real Estate Advertisements

A Real Estate advertisement aims at selling a property to a potential buyer by providing a thorough description of the characteristics and location convenience. This short text has an important impact on the buyer's final choice of visiting a real property. Moreover, the agents have a good knowledge of the market and often target a sub-group of buyers (e.g., families, single people, etc.) by putting forward attributes or location information (e.g., proximity to specific amenities) as explained in Chapter 2 Section 2. For instance, the following example (Fig. 3.1) shows the different kinds of information found in an advertisement: the location (red), the typical characteristics (blue) and the description of the interior design (green). The agent first highlights the main location (*Centre-ville* – Downtown, *proche de l'avenue Jean Medecine* – Near the Avenue Jean-Médecin), and the characteristics of the real property (*appartement 3 pièces de 50* m^2 – 3-room apartment (50 sqm)), which are the most important information, to catch the buyer's attention. Then, the advertisement compiles the interior design of the home before finishing by promoting the neighborhood. The description of the neighborhood suggests a targeted sub-group of buyers, which could be people who are pedestrians and city-dwellers (transports, proximity to amenities).

Although the above example gives a lot of information in a well-structured manner, there is a high variability in the writing of real estate advertisements. The textual analysis is very challenging because of several specificities, typical vocabulary or errors:

- **Different structures**: although some information is mandatory and a specific vocabulary is often used, it does not exist an official template to write an adver-

tisements. There are as many structures as realtors.

  - *Nice – Appartement 2 pièces situé dans un petit immeuble à proximité de toutes les commodités, des écoles, lycées et universités [...].*
    Nice – 2-room apartment located in a small building near all the amenities, schools, high schools and universities [...].
  - *Bonjour, le studio est situé dans le vieux Villefranche-sur-mer.*
    Hello, the condo is located in the old town of Villefranche-sur-mer.

- **Short text**: the description, and in particular the location, is based on few sentences or words, which could be difficult to analyze.
  - *Antibes – Quartier résidentiel, 3 pièces de 55 m^2 .*
    Antibes – Residential area, 3-room of 55 sqm.

- **Poor grammar and vocabulary**: the basic grammar rules are not always followed (e.g., there are not subjects or verbs). The sentences are stereotypical and the vocabulary is limited.
  - *Vends villa, construction 1884, avenue des arènes de Cimiez.*
    Sell villa, building 1984, Avenue des Arènes de Cimiez.

- **Advertising language**: the language is used to promote a property which implies mentioning only its positive characteristics. Also, the realtors often exaggerate, in particular the location, in order to sell quickly.
  - *RARE A CANNES ! SUPERBE APPARTEMENT situé sur LA CROISETTE*
    RARE IN CANNES ! FABULOUS APARTMENT located in LA CROISETTE.

- **Spelling errors and abbreviations**: the authors are not always real estate professionals and neglect the spelling. Moreover, the advertisements are very short and written in a telegram style.
  - *Antibes - Centre Ville - Etage élevé - 3 pièces de 61 m^2 - vue mer.*
    Antibes- Downtown - Upper stairs - 3-room of 61 sqm - sea view.

Furthermore, as we have seen in Section 2 of Chapter 2, the spatial language used in real estate advertisements is particular. Indeed, the aim of the location description is to promote a neighborhood which leads to the use of vague and exaggerated terms. Moreover, a location is often described by nominal and ambiguous entities (e.g., near downtown) that are difficult to capture. We identified six ways to describe a place in an advertisement:

- Toponym (e.g., Nice, Cannes)
- Feature (e.g., the train station, the university)
- Toponym + Feature (e.g., Riquier station, Avenue Jean Médecin)
- Spatial Relation + Toponym (e.g., Close to Nice)
- Spatial Relation + Feature (e.g., Nearby the university)
- Spatial Relation + Toponym + Feature (e.g., Riquier station 5 minutes away)

In a nutshell, the natural and spatial language used in the real estate advertisement is very specific and complex, and differs from typical texts in NLP (e.g., tweets, scientifc papers, news). Therefore, standard models and datasets might not be suitable for our work.

3 Named Entity Recognition

The first stage of our pipeline is the Named Entity Recognition module, which detects and classifies entities in text into predefined categories. In this work, we focus on the extraction of geospatial terms, also called Geoparsing, which has been widely studied in various types of text, but mainly written in English [Adams, 2012; Grace, 2021; Hu, 2021; Moncla, 2017]. Moreover, Geoparsing methods only extract toponyms (i.e., place-names) which is very limited for our research. Indeed, we would like to extract all the entities referring to a location. These entities could be named (e.g., Nice) or non-named (e.g., the train station). For instance, Medad et al. [Medad, 2020] proposed to define and extract French spatial nominal entity (i.e., different from named entity) which refers to physical objects in a spatial context (i.e., a geographic term which is not linked to a Toponym). This approach is similar to our proposal.

To face the mentioned challenges (French language, the specific language style, and the spatial entities), we first propose to define categories for the extraction of spatial information. Then, our proposed approach includes the annotation and creation of a new dataset based on a corpus of real estate advertisements and the predefined categories. Finally, according to the annotation process, we build a model trained on the new dataset.

3.1 Annotation Guidelines

To help machine learning model to detect meaningful information, the process of data annotation is crucial. In NER task, the annotation aims at assigning a tag, which

belongs to predefined categories (e.g., Person, Location, Event), to entities within a block of text, identified by the IOBES format. In this work, we deal with specific types of text, categories and languages as explained in Section 2. Moreover, a French standard dataset is not available for our task and use case. Hence, we decided to create our own dataset. To annotate and create a new dataset, we first collected about 1,200 advertisements[1] written in French and gathered from various online advertisers in the French Riviera. The average length of the ads is about 650 words after a preprocessing stage. The preprocessing stage involves a cleaning process which was necessary since the advertisements were full of noisy, repetitive words and abbreviations. We also removed new lines, URLs, special symbols and characters (e.g., &, #, *) using regular expressions.

The annotation task involved two annotators, who are PhD students respectively in Geography and Computer Science, and native French speakers which prevents any language barrier. Each annotator processed around 600 texts relying on guidelines and using *doccano* [Nakayama, 2018], an open-source text annotation tool (see Appendix B). Writing guidelines was an iterative process between the two annotators to confirm which entities should be annotated. Indeed, we noticed ambiguous and borderline cases which were difficult to classify. Moreover, we paid attention to define categories that were specific enough for our use case, and different enough for the NER model. For instance, increasing the number of small categories (e.g., natural feature, administrative boarder, building, etc.) could be meaningful for our use case, but the model could perform poorer since it is difficult to differentiate such categories. Hence, we defined four categories (i.e., Toponym, Feature, Spatial Relation and Mode of Transportation) described in the following annotation guidelines consisting of a definition, positive and negative examples, and special cases:

Toponym: A toponym, or place-name, is an entity referring to a place's proper name which is the easiest way to describe a place.

Positive Examples

- ***Cannes*** *– 3 pièces [...] Prix: 300 000 euros.*
  Cannes – 3-room [...] Price: 300,000 euros.
  Justification: The name '**Cannes**' refers to a city in the French Riviera, and is easily recognized thanks to its position in the text and its uppercase letter.

- *Immeuble situé sur le boulevard* ***Sadi Carnot****, à deux pas du lycée* ***Carnot****.*
  Building located on the Boulevard **Sadi Carnot**, a short distance away from the

[1] https://github.com/lcadorel/GeoInformationRealEstate

Carnot high-school.
Justification: The two proper names '**Sadi Carnot**' and '**Carnot**' are classified as Toponym because they identify a type of geographic feature (i.e., 'boulevard', 'high-school'). Moreover, the uppercase letter and the spatial relations are also a hint.

- *Villa située place **masséna**.*
  Villa located in the place **masséna**.
  Justification: The proper name does not always contain an uppercase letter such as this example ('**masséna**'). Nevertheless, the context surrounding this word ('located in', 'place') allows to identify the toponym.

- ***SOPHIA ANTIPOLIS** – Situé à dans la technopole [...].*
  SOPHIA ANTIPOLIS – Located in the technology park [...].
  Justification: '**SOPHIA ANTIPOLIS**' is classified as Toponym because of its position in the text. Also, the uppercase letters are a typical writing of the real estate agents to emphasize a location.

- ***NICE CIMIEZ** – 2 pièces [...].*
  NICE CIMIEZ – 2-room [...].
  Justification: On the contrary of the previous example, this one has two toponyms ('**Nice**' and '**Cimiez**'), which are more difficult to identify. Indeed, '**Nice**' is a city while '**Cimiez**' is a neighborhood, and the difference only relies on the knowledge of the annotator. Nevertheless, it is common to find the name of the city followed by the name of a neighborhood in the advertisements.

Negative Examples and Special Cases

- *Belle villa avec piscine [...] Agence Immobilière **Mandelieu La Napoule**.*
  Beautiful villa with a swimming pool [...] Estate agency **Mandelieu La Napoule**.
  Justification: This example shows a place name used to identify an estate agency, which does not give information about the property's location. This proper name should not be annotated.

- ***CENTRE VILLE** – 3 pièces [...].*
  DOWNTOWN – 3-room [...].
  Justification: The location is at the beginning of the text, such as the positive examples, but we chose not to classify '**Downtown**' as a Toponym.

- *NICE **OUEST** – 2 pièces [...].*
  WEST NICE – 2-room [...].

Justification: Although '**WEST NICE**' could be seen as a proper name of a neighborhood, we chose to classify '**WEST**' as a spatial relation which specifies the orientation.

Feature: A feature is an entity representing natural features, constructions and subdivisions of land which is located on, or near the surface of the earth. It includes both representations that exist physically (e.g. a building) and those that are conceptual or social creations (e.g., a neighborhood).

Positive Examples

- *Immeuble situé sur le **boulevard** Sadi Carnot, à deux pas du **lycée** Carnot.*
  Building located on the **Boulevard** Sadi Carnot, a short distance away from the Carnot **high-school**.
  Justification: The two entities are classified as Feature because they define the type of geographic amenities of the proper names '**Sadi Carnot**' and '**Carnot**'.

- *Proche des **commerces, transports, écoles et de la gare**. [...].*
  Close to the **shops, transports, schools and the train station**. [...].
  Justification: This example shows a list of amenities which are '**close to**' the property. The list is a typical writing of the real estate agents. Also, the features are cited without proper names, but they give information about the location.

- ***CENTRE-VILLE** d'Antibes – Proche de la **vieille ville** [...].*
  DOWNTOWN of Antibes – Near the **old town** [...].
  Justification: We chose to classify the entities '**Downtown**' and '**Old twon**' as Feature because they could be find in any city and are not really a proper name. Moreover, the name of the city could be added after these words, which is the same structure as the first example.

- *Appartement avec vue sur la **mer**.*
  Apartment with **sea** view.
  Justification: The entity '**sea**' is a natural feature and is associated with the relation of visibility '**view**', which gives location information.

Negative Examples and Special Cases

- *L'appartement est proche du centre. [...] En plein **centre** du quartier des Musiciens [...].*
  The apartment is close to the center. [...] In the **center** of the Musiciens neighborhood [...].

Justification: In this example, the word '**center**' is used twice but we only classified it as Feature once. Indeed, the first one refer to the city center, or downtown, and could be seen as a feature. However, the second one is a spatial relation between the property and the neighborhood.

Spatial Relation: A spatial relation describes the location of an object by specifying its direction with respect to a reference object whose location is known (see chapter 2). A spatial relation could express different configurations in space such as proximity (e.g., close to), adjacency (e.g., next to), overlap (e.g., in) or orientation (e.g., north of). The term used to indicate the spatial relation is often a preposition.

Positive Examples:

- *L'appartement est **proche** de la place Masséna et de l'avenue Jean Médecin.*
  The apartment is **near** the Masséna Square and the Avenue Jean Médecin.
  Justification: The preposition '**near**' specifies the spatial relation between the apartment and two spatial objects ('**Masséna Square**' and '**Avenue Jean Médecin**').

- *La gare est à 10 **minutes** à pied. [...] La plage se trouve à 200 **mètres**.*
  The train station is 10 **minutes** away by walk. [...] The beach is 200 **meters** away.
  Justification: In this example, the spatial relations are expressed by a distance ('**meters**') and a temporal relation ('**minutes**'). It should be noted that we did not include the numbers (10 and 200) in the tag, which will be extracted as attributes in the next step.

- *Jolie villa avec **vue** sur la mer.*
  Beautiful villa with a sea **view**.
  Justification: This annotation refers to a visibility ('**view**') relation between the property and the sea.

- ***Au coeur** du Vieux Nice, [...].*
  In the heart of the Vieux Nice, [...].
  Justification: Although this spatial relation is a vague expression, it represents an overlap between the property and '**Vieux Nice**'.

- *NICE **NORD** – Le bien est dans une résidence de standing [...].*
  NORTH NICE – The property is in a luxury residence [...].
  Justification: As mentioned before, '**North Nice**' could be seen as a Toponym but we chose to annotate '**North**' as a spatial relation since it describes the orientation.

Negative Examples and Special Cases

- *Quartier Carré d'Or – L'appartement est situé **dans** la rue Massenet.*
  Carré d'Or district – The apartment is located **in** the Massenet Street
  Justification: We only tagged the spatial relation different from the small prepositions '**in**' or '**on**'. Indeed, the example shows two spatial objects where the property overlaps. However, the first objects ('**Carré d'Or district**) is not linked to a preposition. Hence, we chose not to annotate the small preposition referring to the overlap. Therefore, each spatial object without a spatial relation is considered as an overlap.

Mode of Transportation: A mode of transportation is an entity which describes the travel mode between two places. It is often associated with a spatial relation described with a temporal entity.

Positive Examples

- *Le bien est à 10 minutes en **voiture** de l'université.*
  The property is 10 minutes away from the university by **car**.
  Justification: The entity '**car**' precises the spatial relation between the property and the university. Indeed, '**10 minutes by car**' is different from by walk.

- *Tout à **pied**.*
  Everything by **walk**.
  Justification: This example shows a common expression to describe the proximity to the amenities. The entity '**walk**' suggests that the travel mode between the property and the main amenities is walking.

Negative Examples and Special Cases

- *Garage pour 2 **voitures**.*
  Garage for 2 **cars**.
  Justification: In this example, the entity '**car**' does not refer to the mode of transportation, but it is only a characteristic of the property.

In Table 3.1, we summarize the categories and the number of tagged entities in our dataset. Feature is the most represented category in the dataset since a lot of amenities are used to describe the location. On the other hand, the Mode of transportation has very few examples because we did not retrieve a lot of different words (e.g., walk and car). Finally, we also annotated misspellings of entities, as the misspelled words are still a true signal of where a real entity would appear. Moreover, a post-processing step could help to retrieve the correct word from the misspelled words.

Category	Examples	Count
Feature	- This apartment reveals a magnificent view over the **sea**. - Excellent location, close to the **local shops**, the **university** and the **tram**.	3313
Toponym	- **Nice Vinaigrier**, neo provençal style property in excellent condition - In **Cannes**, in the residential and sought after area of **La Californie**	2313
Spatial Relation	- a **stone's throw** from the sea and local shops - For sale **near** the Croisette in Cannes	1476
Mode of transportation	- 5 minutes **walking distance** from Place Masséna - Nice Côte d'Azur Airport a 20 minute **driving** away	160

Table 3.1: Number of annotations for each category.

3.2 Model

The analysis of the language in Real Estate advertisements and the annotations of a new dataset led to propose an approach based on a *BiLSTM-CRF* architecture [Lample, 2016], which has achieved very good results on NER tasks, and implemented by *Flair*, a Python NLP package. Indeed, linguistic rules or gazetteers are often used for Geoparsing [Moncla, 2017; Lieberman, 2010], but they give limited results and depend on the completeness of the rules and gazetteers. Hence, we chose to use a deep learning method to automatically learn complex features and representations, and to skip the feature-engineering step. Moreover, the advertisements have language and stylistic specificity and variability, an informal format, and are written in French. All these aspects could be difficult to learn for a simple *BiLSTM-CRF* model. Hence, a specific text representation can be suitable to tackle the complex and French language [Barrière, 2019]. We added an embedding to the *BiLSTM-CRF*, which is a global vector composed of the concatenation of three different text representations, to capture features at different levels (see Figure 3.2).

The first text representation is a basic Word Embedding architecture [Mikolov, 2013] trained on our corpus. This embedding represents each word as a low-dimensional vector where each dimension constitutes a latent feature, automatically learned from text. It captures semantic properties of a word but does not take into account the context.

Secondly, we fine-tuned the two pre-trained French Flair Language Models [Akbik, 2018] with our corpus, which corresponds to a *BiLSTM* Language Model. This Language Model has a context-based and character-level representation, which is well-suited for complex tasks. Also, the specificity of only keeping the first and last state helps handle out-of-vocabulary words and small dictionaries. Finally, its French pre-trained model is a good advantage for our task since few models are available for French.

Finally, we applied *CamemBERT* [Martin, 2019], a French Transformer Language Model. *Transformers* are faster and more efficient than *BiLSTM* or *CNN* architec-

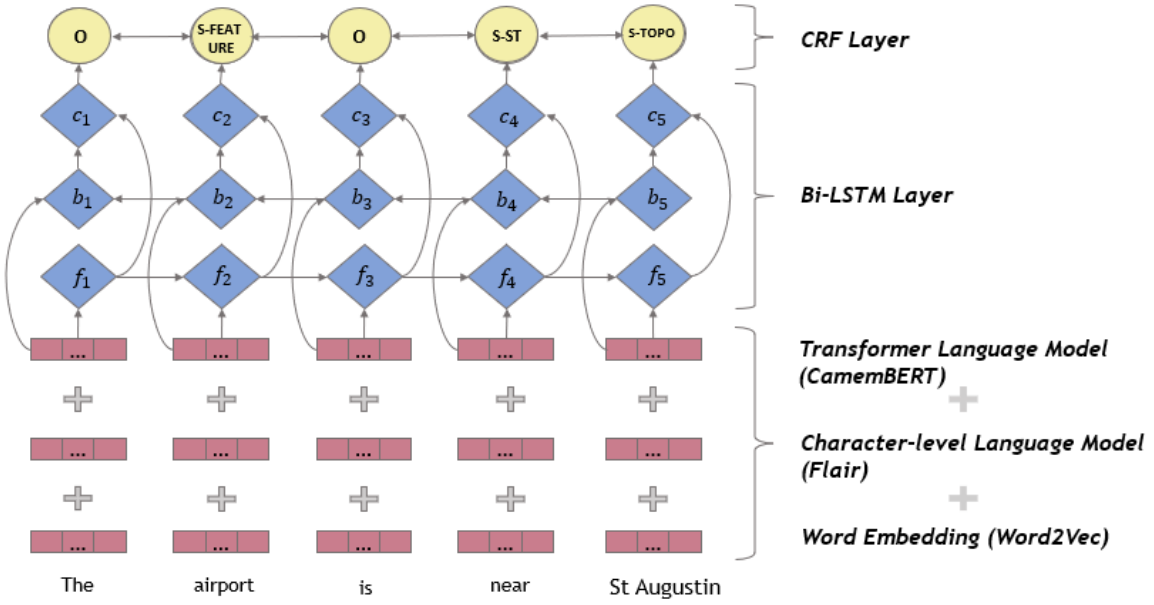

Figure 3.2: Our NER architecture

tures, since they use parallel attention layers [Vaswani, 2017]. Moreover, Transformer Language Models are trained on huge corpora with two specific tasks: masked language model (MLM) and next sentence prediction (NSP). For MLM, some tokens are masked and the model has to predict them in a sentence. The other task (NSP) aims at predicting, of two sentences, which one follows the other. While these models reach high performance, some limitations have arisen: the need of a huge training set on the one hand and limitation of their application to specific domains or tasks due to pre-trained models on general domain on the other hand. Nevertheless, we decided not to fine-tune the French pre-trained model, *CamemBERT*, due to a lack of a huge training corpus.

Figure 3.3 shows the output of the model applied to the text of the previous example (Fig. 3.1). The output returns all entities corresponding to a pre-defined category, their position in the text and the confidence of the prediction. In our annotation, we gave French names for the category: EG stands for Feature, TOPONYME stands for Toponym and ET stands for Spatial Relation. Also, an entity can be composed of several words such as '**CENTRE VILLE**' or '**carré d or**' and the model returns all the position in the text (e.g., [1,2]). Finally, the model extracted all the spatial entities with a high confidence. Although the prediction is correct, the confidence of the entity '**rue**' is a bit low. This low confidence might be explained by the lack of a toponym following the feature '**rue**', which the model could expect.

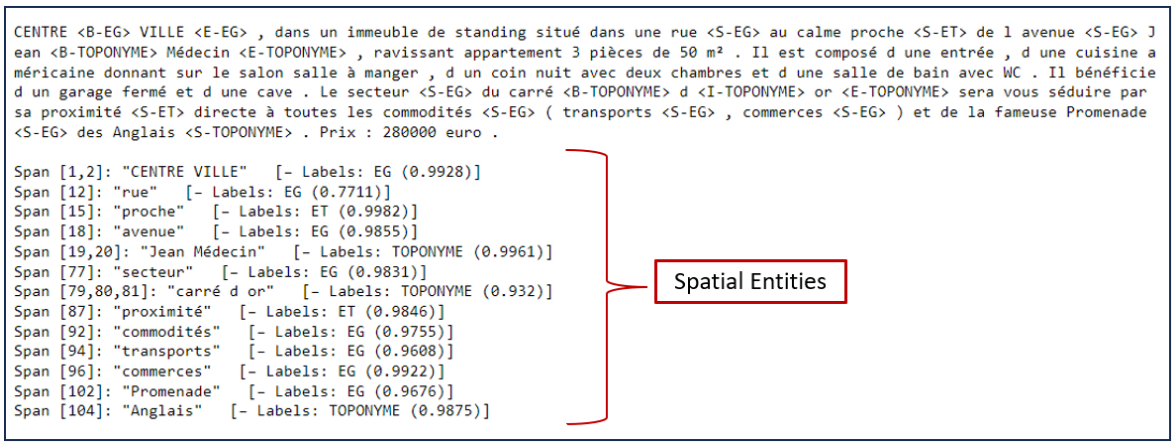

Figure 3.3: Example of the output of the NER model

4 Relation Extraction

The second stage relies on providing a structured representation of the information, such as a knowledge graph, by extracting relationships between spatial entities. Relation extraction (RE) is the task of detecting and extracting semantic relations between different entities found in the NER model. As stated in the literature review (see Chapter 2), the RE methods are mainly based on handcrafted patterns or Machine Learning approaches. ML-based approaches usually use a labeled corpus to train a model, or focus on verbs in the sentence to describe the relation. However, in our work, we did not have a labeled corpus, and the advertisements often contain a free word order and lack verbs. Hence, we chose to design handcrafted patterns to retrieve relationships between entities.

First of all, we identified and defined four types of relationship:

- **Attribute**: an attribute describes or gives more information about a spatial entity. It could be an adjective, a numeral or a noun object (e.g., **5** minutes, **residential** neighborhood, **famous** Promenade des Anglais, etc.);
- **Type of Place**: the affiliation link between a feature and a toponym (e.g., Avenue Jean Médecin, city of Cannes, Audiberti High School, etc.);
- **Spatial**: the spatial relationship is a ternary relationship between an object (e.g., a property), a spatial relation entity (e.g., near, close to), and a referent object described by a feature, a toponym or both (e.g., the university, Place Masséna, Nice);
- **Mode of transportation**: a relation between a mode of transportation (e.g., by car) and the spatial relation it specifies (e.g., 5 minutes).

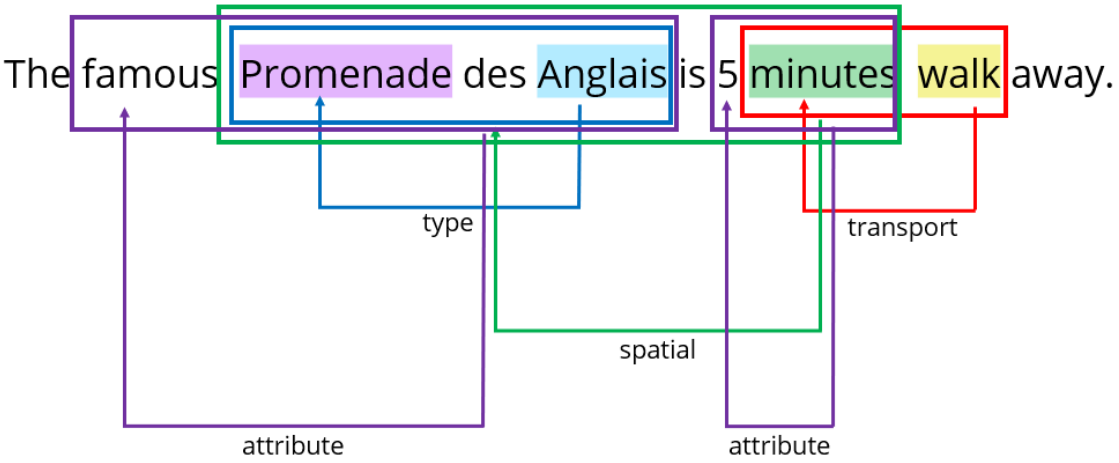

Figure 3.4: Example of the four types of relationship.

Figure 3.4 shows the example of the four types of relationship in a sentence. For instance, the toponym '**Anglais**' is linked to the feature '**Promenade**', which is its type. The two terms form a place, which is characterized by the adjective '**famous**'. Moreover, the spatial relation entity '**minutes** is specified by the numerical adjective '**5**' and the mode of transportation '**walk**'. Finally, a spatial relationship exists between the referent object '**Promenade des Anglais** and the spatial relation entity '**minutes**'. The object is unknown in this sentence, but we can assume it is the property since we deal with a real estate advertisement.

To retrieve and extract all these relationships, we made some assumptions. First, we considered that a relationship occurs only between two entities of the same sentence. Hence, we assumed that a direct or indirect grammatical connection between the two entities always exists in the sentence, which can be captured by the dependency graph. Bunescu et al. [Bunescu, 2005] argued that if two entities in a given sentence have a semantic relation, then it is mostly concentrated in the shortest path of the dependency graph between the two entities. The shortest path seems to offer a very condensed representation of the information needed to estimate a relationship. We proposed to use the shortest path and the grammatical analysis to design rules to extract relationship between two entities. We first computed a dependency graph using a pre-trained dependency parser. Then, we improved the dependency parser by fine-tuning a Part-of-Speech tagger and using the previously extracted NER tags. Finally, we extracted the shortest path between two entities and design rules to detect relationships.

4.1 Dependency Parsing

The study and the extraction of the grammatical structure from a text, represented as a parse tree, is frequently based on two methods: Constituency and Dependency

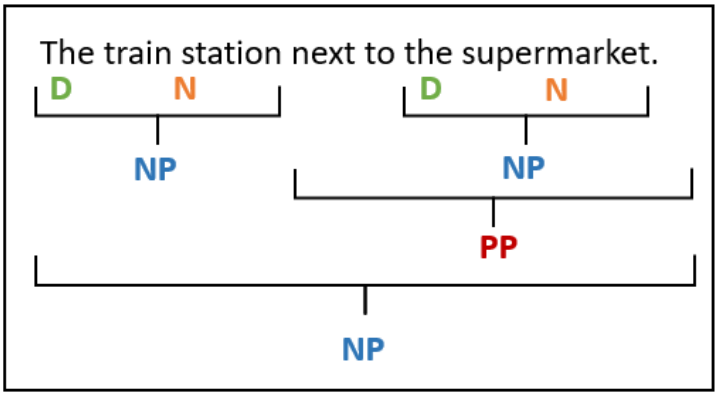

Figure 3.5: Example of context-free grammar rules

parsing. On the one hand, Constituency parsing consists of dividing a sentence into sub-phrases that belong to the same grammatical category, called constituents. The formalism of context-free grammars is the most common technique to gather words and is based on a set of grammar rules, which specify how individual words can be grouped to form constituents. For instance, a constituent can be a noun phrase (NP), where a rule could be the association of a determiner (DT) and a noun (N), or a noun phrase (NP) followed by a prepositional phrase (PP) as shown in Figure 3.5. Nevertheless, this method faces several limits, especially in the study of real estate advertisements. Indeed, the free word order and the complex sentences in real estate advertisements lead to an inaccurate parsing.

On the other hand, Dependency parsing is based on another formalism called dependency grammar, where the syntactic structure is represented by directed binary grammatical relations between two words. A binary relation consists of a head word, a dependent word modifying the head, and a tag describing the nature of the grammatical function. The Universal Dependency Relations [Nivre, 2016] is a taxonomy to capture grammatical relations across world's languages, often used to label the relationship between the head and the dependent. Figure 3.6 shows the main dependencies in the same sentence as the constituency parsing example. The grammatical labels are defined as follows:

- *det*: the determiner relation holds between a nominal term and its determiner;
- *compound*: this relation is used to group words belonging to a multiword expression (MWE);
- *advmod*: an adverbial modifier is an adverb or adverbial phrase that usually modifies a predicate (i.e., a verb). In this case, the sentence does not have a predicate and the adverb modifies a nominal expression;

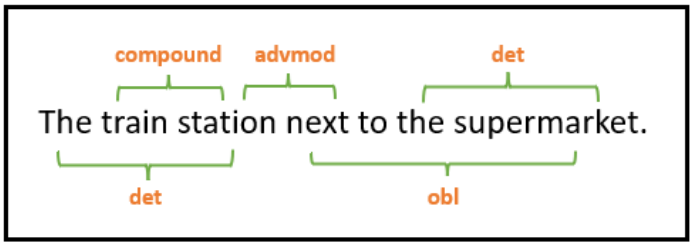

Figure 3.6: Example of dependency parsing.

- *obl*: the oblique relation is used for a nominal expression functioning as an adverbial attaching to another adverb. It specifies the adverb *next to*.

This approach is more suitable for our analysis since the order of words does not always follow the classical grammar (e.g., advertisements do not always contain a verb or a subject). Also, the head-dependent relation offers a good approximation of predicate-argument relation for Information Extraction. Finally, the Universal Dependencies taxonomy is general enough to easily capture dependencies in real estate advertisements, and can be easily adapted to other languages. The majority of dependency parsing models are supervised methods and lie into two types: graph-based and transition-based, both of which have been largely studied under the traditional statistical and the neural approaches [Zhang, 2020]. Currently, neural network architectures achieve the state-of-the-art performance.

In our work, as we have not got a labeled corpus available, we used *Stanza* [Qi, 2020], a Python NLP package providing tools (e.g., tokenization, Part-of-Speech, lemmatization, etc.), which can be used in a pipeline to analyse a text and its structure. The toolkit is designed to support more than 70 languages and is based on a neural network architecture, using the Universal Dependencies formalism and is already trained on a large corpus. In particular, the dependency parser is a BiLSTM-based deep biaffine neural network [Dozat, 2017] augmented with two linguistic features that achieves good results [Qi, 2019]. The model takes as input pretrained word embeddings, frequent word and lemma embeddings, character-level word embeddings and Part-of-Speech embeddings. The output is a document where each word is represented as a dictionary entry containing the label and the id of the head of the relationship. In addition, grammatical information is given for each word, such as the Part-of-Speech or the lemma, as shown in Figure 3.7.

We applied the French pre-trained model to our data to compute the dependency graph for each text. However, after a quick analysis, we noticed that some relationships were not accurate, especially when the POS tagging failed. Indeed, the model has been trained on other types of text very different from Real Estate advertisements. Despite its high performance, the model might fail on predicting POS because of the

language specificity of the advertisements. Nevertheless, it is essential to correctly predict the POS since the performance of the dependency parser is very related to these good predictions. *Stanza* gives the possibility to use our own POS tagger and to pass a pretagged document to the dependency parser as input. Hence, we proposed to improve the performance of the POS tagger and to pass the output to the dependency parser to get more accurate results.

4.2 Part-of-Speech Tagging

A Part-of-Speech (POS) is a category of words that have similar grammatical properties, such as a noun, an adjective or a verb. The process of tagging aims at assigning the correct POS to each word in a text, and is similar to the NER approach described in Section 3. A word could have more than one possible POS and the correct one depends on its use in a sentence. The Universal Dependencies [Nivre, 2020] defined 16 universal POS, including noun, verb, determiner, numeral, adjective, etc. Figure 3.8 shows the POS tags assigned to each word of the sentence, and defined as follows:

- *DET*: a determiner is a word that modifies a noun and expresses the reference of the noun in a certain context (e.g., definite or indefinite determiner),
- *NOUN*: a noun is only used to denote a common noun (proper nouns and pronouns have their own tags);
- *ADV*: an adverb typically modify a verb related to time, place, direction or manner. It may also modify adjectives and other adverbs.
- *ADP*: an adposition is a preposition attached to a noun phrase called complement;
- *PUNCT*: a punctuation mark is a non-alphabetical character and delimits linguistic units in a text.

As mentioned in Section 4.1, the tagger implemented in *Stanza* often fails to predict the correct POS for our corpus because of the language used in the advertisements. For instance, in Table 3.2, one of the tag predicted by *Stanza* is wrong. Indeed, '*recherché*' (sought-after) is an adjective that specifies the word '*quartier*' (neighborhood), but has been predicted as a verb. This mistake might come from the ambiguity of the word '*recherché* since it could be a verb or an adjective depending on the context. Moreover, the grammar and structure of the sentence are not common because the sentence does not contain a verb. Therefore, we trained our own POS tagger that better fits our corpus.

The POS-tagging models are similar to the NER model since the aim is to tag each word with a predefined label (e.g., ADJ, NOUN, VERB, etc.). Training a model from

```
{
  "id": 1,
  "text": "Dans",
  "lemma": "dans",
  "upos": "ADP",
  "head": 3,
  "deprel": "case",
  "misc": "start_char=0|end_char=4"
},
{
  "id": 2,
  "text": "un",
  "lemma": "un",
  "upos": "DET",
  "feats": "Definite=Ind|Gender=Masc|Number=Sing|PronType=Art",
  "head": 3,
  "deprel": "det",
  "misc": "start_char=5|end_char=7"
},
{
  "id": 3,
  "text": "quartier",
  "lemma": "quartier",
  "upos": "NOUN",
  "feats": "Gender=Masc|Number=Sing",
  "head": 9,
  "deprel": "nmod",
  "misc": "start_char=8|end_char=16"
},
{
  "id": 4,
  "text": "résidentiel",
  "lemma": "résidentiel",
  "upos": "ADJ",
  "feats": "Gender=Masc|Number=Sing",
  "head": 3,
  "deprel": "amod",
  "misc": "start_char=17|end_char=28"
},
{
  "id": 5,
  "text": "d'",
  "lemma": "de",
  "upos": "ADP",
  "head": 6,
  "deprel": "case",
  "misc": "start_char=29|end_char=31"
},
{
  "id": 6,
  "text": "Antibes",
  "lemma": "Antibes",
  "upos": "PROPN",
  "head": 3,
  "deprel": "nmod",
  "misc": "start_char=31|end_char=38"
},
```


Figure 3.7: Example of the output of the dependency parser of Stanza.

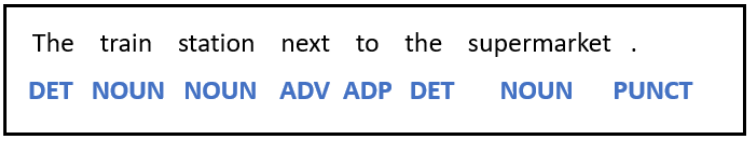

Figure 3.8: Example of Part-of-Speech tagging.

scratch is expensive and requires a labeled corpus. In our approach, we applied a POS-tagger, developed by *Flair*, to our corpus to get labeled data that we manually corrected afterwards. Then, we fine-tuned the pre-trained model of *Flair* with the corrected labeled corpus to quickly improve the predictions. In Table 3.2, we can notice that the new model tagged the word '*recherché*' with the correct label. Finally, we made two other improvements to help the model to better predict the POS: (1) we cut the text into sentences to reduce the ambiguity, and (2) we grouped the entities extracted by the NER model that belong to the same tag as a single term. For example, in Figure 3.8, the entity '*train station*' is divided in two distinct words, both tagged as a noun. In our approach, we gathered these two words into a single noun in order to reduce the number of grammatical relations that could make the dependency parsing more complex. Finally, these modifications are incorporated into the pretagged document that is passed to the dependency parser as input.

Dans un quartier recherché avec vue mer, 3 pièces à Nice.
In a sought-after neighborhood with a sea view, 3-room in Nice.

Word	**Stanza**	**Fine-tuned tagger**
Dans	ADP	ADP
un	DET	DET
quartier	NOUN	NOUN
recherché	**VERB**	**ADJ**
avec	ADP	ADP
vue	NOUN	NOUN
mer	NOUN	NOUN
,	PUNCT	PUNCT
3	NUM	NUM
pièces	NOUN	NOUN
à	ADP	ADP
Nice	PROPN	PROPN
.	PUNCT	PUNCT

Table 3.2: Comparison of POS taggers

4.3 Shortest Path Dependency

The final step of Relation Extraction aims at retrieving the shortest path between two entities in order to confirm their relationship. The dependency parsing produces a dependency graph where nodes are the words and the grammatical labels represent the edges. From this graph, we can extract a path, in particular the shortest path, that links a word to another. The hypothesis of the shortest path refers to the high probability to find a predicate that links two entities in the shortest path between them

[Bunescu, 2005]. In our work, we know the predicate (e.g., attribute, type of place, spatial and mode of transportation) and we proposed to use the shortest path to accept or reject a relationship between two entities. Indeed, we stipulated that if the shortest path between two entities has a small length, then we should accept the relationship between these entities.

For instance, Table 3.3 shows the shortest path extracted from the dependency graph for different entities and types of relationship. In the sentence, we detect four entities: '*coeur*' as a Spatial Relation, '*quartier*' and '*arènes*' as Feature and '*Cimiez*' as Toponym. These entities are represented as nodes in the graph. Then, according to the type of relationship, we can look for a path between these nodes and other types of node. For example, to retrieve the attributes of the two Feature entities, we can extract the shortest path between these nodes and any adjective in the sentence. In this example, all the shortest paths have a length of one, except the spatial relationship between '*arènes*' and '*coeur*' that has a length of two. This path is longer because '*arènes*' is linked to '*coeur*' thanks to the entity '*quartier*' and the conjunction '*et*'. Therefore, the length of the shortest path plays an important role to determine if a relation is acceptable. Indeed, it is possible to find a very long path between two words if they are connected by other words. But the longer the length, the less reliable the relationship. We analysed the paths found in our corpus for each type of relationships, and set a maximum length of three for the Attribute relationships and four for the others. Finally, all paths that match the rule are kept as a relationship.

L'appartement se situe au ***coeur*** *du* ***quartier*** *résidentiel de* ***Cimiez*** *et des célèbres* ***arènes***.
The apartment is located in the heart of the residential Cimiez district and the famous arenas.

Entities	**Type of Relation**	**Shortest Path Dependency**
(quartier, résidentiel)	Attribute	FEATURE ↓ amod ADJECTIVE
(cœur, quartier)	Spatial	SPATIAL ↓ nmod FEATURE
(quartier, Cimiez)	Type of Place	FEATURE ↓ nmod TOPONYM
(cœur, arènes)	Spatial	SPATIAL ↓ nmod FEATURE (*quartier*) ↓ conj FEATURE
(arènes, célèbres)	Attribute	FEATURE ↓ amod ADJECTIVE

Table 3.3: Example of relationships and the shortest paths.

5 Evaluation

In this section, we describe the evaluation of the proposed workflow. We first define the classic metrics used to assess the performance of NER and POS-tagging models. Then, we present and discuss the results for the NER and RE tasks.

5.1 Evaluation Metrics

To quantify the performance of the models, the Precision, Recall and F1-Score metrics are widely used in NLP. We used these metrics to evaluate the results of the different steps of our processing chain (i.e., NER, POS tagging).

Definition 5.1 (Precision)**.** Precision is the ratio between the correctly identified positive results (true positives) and the total number of positive predictions. It evaluates how many of the positive predictions are correct. In the context of NER, Precision is the ratio between the number of correctly annotated entities and the total number of entities annotated by the model.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{3.1}$$

Definition 5.2 (Recall)**.** Recall is the ratio between the correctly identified positive results (true positives) and the total number of positive results that should have been

returned. It measures the model's ability to predict the positive classes, and provides an indication of missed positive predictions. In the context of NER, Recall is the ratio between the number of correctly annotated entities and the total number of entities labeled in the dataset.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}, \tag{3.2}$$

Definition 5.3 (F1-Score)**.** F1-Score combines Precision and Recall to provide a single metric that weights the two ratios in a balanced way. The use of the Harmonic mean implies to have high values for both metrics to rise the F1-Score value.

$$\text{F1-Score} = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}. \tag{3.3}$$

5.2 Named Entity Recognition

We evaluated the Named Entity Recognition model by applying the model to the annotated dataset described in Section 3.1. We split the dataset into 10-folds and evaluated our pipeline through a cross-validation. We compared our model to a fine-tuned model implemented by *SpaCy*[2], and we tested several combinations of our embeddings. We chose to perform the comparison with *SpaCy* because the neural architecture is different from the BiLSTM-CRF model (CNN and RNN respectively). We fine-tuned a pre-trained model for French with our entities and trained another from scratch.

Table 3.4 reports the average Precision, Recall, F1-Score and their standard deviation, based on the 10-folds cross-validation. Table 3.5 presents the average metrics for each category predicted by our proposed NER model. Finally, we also computed a Welch's t-test, with t defined as

$$t = \frac{\overline{X_i} - \overline{X_j}}{\sqrt{s^2_{\overline{X}_i} + s^2_{\overline{X}_j}}}, \tag{3.4}$$

where $\overline{X_i}$ and $s_{\overline{X}_i}$ are, respectively, the i^{th} sample mean and its standard error, to measure statistical significance of the difference between F1-Scores at a 1% significance level. The results are summarized in Table 3.6.

First, we can notice that the *BiLSTM+CRF* architecture mostly outperforms *SpaCy* models except with *Word2Vec* embedding at a 1% significance level. Regarding *SpaCy*, the model trained from scratch gets slightly better results compared to a pre-trained model. Then, the results show that among the *BiLSTM+CRF* models with a single embedding, the one with *CamemBERT* achieves the best performance. The one with

[2] http://spacy.io/

character-level embeddings implemented by *Flair* gets a similar Recall but the Precision is lower. Also, we can underline that the combination of the embedding computed by *Flair* and *CamemBERT* increases the performance of the three metrics. The combinations of *Flair* and *Word2Vec* or *CamemBERT* and *Word2Vec* increase the Precision but decreases the Recall. Moreover, there is no evidence of statistically significant differences between those models. Finally, the combination of the three representations achieves the best results while the Welch's t-test does not show any difference with *Flair + CamemBERT*. Nevertheless, we chose to keep this model for our application.

Furthermore, Table 3.5 shows that the predictions made by our model for Feature and Toponym are very good, since F1-Score is very close to 0.9. However, the results fall off for the Spatial Relation and Mode of Transportation. In particular, the Precision is a bit low (i.e., less than 0.8) while the Recall remains pretty high. It may mean that the model predicts too many positive entities that are not correct (i.e., high false positives). This drawback of the model could be explained by the lower number of annotated entities in the dataset.

	Model	**Precision**	**Recall**	**F1-Score**
Spacy				
	Pre-trained French model	0.830 (0.03)	0.822 (0.03)	0.821 (0.02)
	Own training	0.828 (0.02)	0.845 (0.01)	0.835 (0.01)
Bi-LSTM - CRF				
	Word2Vec	0.786 (0.02)	0.741 (0.02)	0.763 (0.02)
	Flair	0.833 (0.01)	0.876 (0.01)	0.854 (0.01)
	CamemBERT	0.851 (0.01)	0.877 (0.005)	0.865 (0.01)
	Flair + Word2Vec	0.837 (0.01)	0.864 (0.01)	0.85 (0.01)
	Camembert + Word2Vec	0.860 (0.01)	0.872 (0.004)	0.866 (0.005)
	Flair + CamemBERT	0.861 (0.01)	0.884 (0.02)	0.872 (0.01)
	Flair + CamemBERT + Word2Vec	**0.863 (0.005)**	**0.889 (0.01)**	**0.876 (0.01)**

Table 3.4: Performance of NER models.

	Precision	**Recall**	**F1-Score**
Feature	0.900 (0.02)	0.892 (0.02)	0.896 (0.02)
Toponym	0.879 (0.03)	0.898 (0.03)	0.890 (0.02)
Spatial Relation	0.785 (0.02)	0.919 (0.02)	0.848 (0.01)
Mode of Transportation	0.743 (0.08)	0.954 (0.06)	0.834 (0.07)

Table 3.5: Performance of NER models by type of entity.

5.3 Relation Extraction

To investigate the performance of the Relation Extraction process, we first evaluated the POS-tagging model and then the extracted relationships. As explained in Section

Model 1	Model 2	p-value
Spacy (Own training)	Spacy (Pre-trained)	0.02
Flair	Flair + Word2Vec	0.3
CamemBERT	CamemBERT + Word2Vec	0.31
CamemBERT	Flair + CamemBERT	0.02
CamemBERT + Word2Vec	Flair + CamemBERT	0.04
Flair + CamemBERT	Flair + CamemBERT + Word2Vec	0.14

Table 3.6: Level significance > 1%

4, we fine-tuned a POS-tagger to help the dependency parser to get better results. Hence, we compared the fine-tuned model to the one implemented by *Stanza* and the pre-trained model developed by *Flair*. The metrics used to evaluate the performance are the same as for the NER evaluation since the POS-tagging is similar to the NER task. Then, we manually evaluated the Relation Extraction process by extracting relationships from 150 texts. This part has not been automatically done because we did not have a labeled dataset available.

Table 3.7 summarizes the performance of the POS-tagger according to the three metrics. Table 3.8 reports the number of extracted relations from 150 texts and the number of relations identified as correct. We also added the number of identified missing relations, that is to say, we identified relations that have not been extracted during the manual evaluation process. However, the number of missing relations is not exhaustive and could be higher.

The evaluation of the POS-tagger gives good promises as the ability to increase the performance of the dependency parser. Indeed, as mentioned in Section 4, the POS-tagger implemented by *Stanza* fails to predict some ambiguous POS, such as adjective and verb, because of the language used by the real estate agents. Table 3.7 shows that Precision for the VERB, PROPN and ADJ is very low for *Stanza* and the pre-trained model of *Flair*. After our fine-tuning, the performance skyrockets for these three POS. All in all, we increased the overall performance of the POS-tagging task by a rise in the F1-Score of eight points when compared to *Stanza*.

Finally, we evaluated the Relation Extraction process by checking the correctness of each relation extracted from 150 texts. We can see that the number of correct relations is high for all the types of relationship. It means that our workflow is able to extract correct relations from unstructured data. However, the extraction is not complete since we highlighted missing relations. For instance, there are at least 40 spatial relations that have not been extracted. If we add these relations to the total number of relations, then the performance decreases to 75 %. Also, these numbers are not exhaustive and might be higher. Moreover, the performance depends on the other part of the workflow (NER, POS-tagging, Dependency parsing) and suffers from the propagation of errors.

In a nutshell, the method performs well enough to extract relations from the text and to provide a structured representation of the information for our use case. A limit is the lack of labeled data to automatically retrieve and evaluate the model. This limit could be overcome in future work.

	Precision			**Recall**			**F1-Score**		
	Stanza	Flair	Own model	Stanza	Flair	Own model	Stanza	Flair	Own model
PROPN	0.57	0.47	**0.79**	0.84	0.87	**0.92**	0.68	0.61	**0.85**
PRON	0.92	0.96	0.96	0.96	1	0.96	0.94	0.98	0.96
VERB	0.64	0.68	**0.94**	0.96	1	0.92	0.77	0.81	**0.93**
ADP	0.99	1	0.99	0.97	0.93	0.99	0.98	0.96	0.99
NOUN	0.90	0.95	0.96	0.93	0.90	0.99	0.92	0.92	0.98
DET	0.98	0.88	0.98	1	1	1	0.99	0.94	0.99
PUNCT	0.94	0.94	0.99	1	1	0.99	0.97	0.97	0.99
ADJ	0.89	0.89	**0.91**	0.6	0.72	**0.89**	0.72	0.80	**0.90**
CCONJ	1	1	1	1	1	1	1	1	1
ADV	1	1	1	0.9	1	0.85	0.95	1	0.92
NUM	0.96	0.98	1	0.98	0.98	0.98	0.97	0.98	0.99
AUX	1	1	1	1	1	0.82	1	1	0.90
SYM	0.88	1	1	1	1	1	0.93	1	1
X	0.92	1	0.95	0.23	0.36	0.83	0.37	0.53	0.89
SCONJ	0.67	1	1	1	1	0.8	1	1	
Total	0.88	0.92	0.97	0.89	0.92	0.94	0.87	0.90	**0.95**

Table 3.7: Performance of POS taggers

Type of Relation	Nb of Extracted Relations	Nb of Correct Relations	Nb of Identified Missing Relations
Spatial	209	183 (88 %)	40
Attribute	116	109 (94 %)	5
Type of Place	87	82 (94 %)	7
Mode of Transportation	22	21 (95 %)	0

Table 3.8: Performance of the Relation Extraction.

6 Summary and Perspectives

This chapter presented the first contribution of this thesis: a method to automatically extract and represent the spatial information from the real estate advertisement. We first analyzed the language used in the real estate advertisements such as the typical vocabulary, errors or the structure since it is different from traditional text studied in NLP (e.g., tweets, scientific papers, news). We detailed five challenging specificities, with few examples, that have to be tackled before performing a textual analysis. We

also identified the different ways to describe a place in an advertisement to show that standard NLP models are not suitable for this task. Then, we presented our method based on a Named Entity Recognition to detect and classify entities in the text into predefined categories. We described the guidelines used to annotate spatial information in the advertisements and to create a dataset to train our model. The model is based on a BiLSTM-CRF architecture combined with three text representations to capture the language specificities: a basic word embedding architecture, a context-based and character-level language model and a transformer language model. As the output of a NER model is not structured, we also proposed to extract relations between entities. We identified four types of relationship that have been retrieved thanks to the analysis of the grammatical structure. Finally, we performed an evaluation of the pipeline and we achieved good results for the NER model as well as the RE approach.

For further improvements, we identified several directions to expand this work in short-term. First, the quality of the annotated dataset used to train our NER model has not been assessed, despite many tasks NLP highly depend on high-quality, manually-labeled text data. Although we performed a lot of iterations to define categories that were suitable for our use case as well as the NER model, we did not check the inter- and intra-annotators agreements (e.g., Fleiss' and Cohen's Kappa) to ensure the consistency of the annotations. Indeed, it was difficult to compute the metrics since we divided the dataset between the two annotators and, we did not label the same texts. Moreover, the number of annotators was limited which led to a small corpora. Therefore, the number of annotators should be increased to create a bigger corpora, especially for some categories (e.g., mode of transportation). A possible direction to increase the size of the labeled corpora without extra human annotators, would be the use of generative large language models (LLMs) such as ChatGPT. For instance, the annotation guidelines presented in Section 3.1 could be the prompt given to the LLM. Nevertheless, the performance of the LLM varies across annotation tasks due to prompt quality, text data particularities, and conceptual difficulty. Thus, an evaluation should be carried out to compare the predictions of the LLM against the human labels and, the prompt should be refined to emphasize incorrect classifications until reaching a high performance. This method relies on a good prompt engineering.

Furthermore, the NER model takes text representations as input and, in particular, the representation obtained from *CamemBERT*, a French Transformer Language Model. We showed that this representation helps to reach the best results combined with the BiLSTM-CRF architecture. However, we did not fine-tune the French pre-trained model although this may increase the performance. Fine-tuning is an approach to transfer learning in which the weights of a pre-trained model are trained on new data in order to become more specific to the given domain or task. This approach allows to keep and apply a general knowledge extracted from a bigger dataset to a smaller

target dataset. It also improves models' generalization ability. Hence, fine-tuning *CamemBERT* should improve the performance of the NER model.

Finally, regarding the RE process, the lack of data is also a limitation. Indeed, our approach is based on the analysis of grammatical dependencies and handcrafted rules. This approach depends on NLP ressources, such as POS tagger and Dependecy Parsing, that could increase the propagation of errors. On the other hand, deep learning methods automatically learn features from sentences, minimize the dependence on external NLP resources, and have shown promising results in other works. Therefore, a possible improvement would be to create a labeled dataset in order to apply a deep learning model.

Chapter 4

Geocoding and Spatial Representation of Real Estate Advertisements from Vague Spatial Descriptions

Objectives

The goal of this chapter is to estimate the boundaries of the Real Estate property from its vague spatial description. We first propose to quickly study how to geocode places thanks to existing gazetteers, and we show their limitations given our data. Then, we focus on empirically estimating the boundaries of the different places and the spatial relations, and how to represent their imprecision. Lastly, we describe and evaluate our method to combine the imprecise spatial information to geolocate each Real Estate advertisement.

Contents

1 Introduction

The real estate market analysis is based on a very local and deep knowledge about the properties, the prices and the neighborhoods. Although the real estate advertisements are a good source of data to explore the market, the French real estate agents often hide the spatial location (e.g., latitude/longitude) since the deal between their agency and the owner might not be exclusive. For instance, in our dataset, 81% of the advertisements do not provide latitude/longitude coordinates. Moreover, among the 19% advertisements providing the coordinates, 12% are not reliable because they are located in the centre of Nice (i.e., same latitude/longitude). Therefore, geocoding the spatial entities extracted in Chapter 3 is essential to estimate the location of an advertisement. Geocoding is the task of retrieving the spatial coordinates of a text-based description. This task often refers to Toponym resolution, which aims at assigning unambiguous coordinates to place names (i.e., toponyms) referenced within documents such as gazetteers. Nevertheless, we showed in Chapters 2 and 3 that the location descriptions, in the advertisements, often use vernacular and local places that are not recorded in gazetteers, as well as vague spatial relations (e.g., near). Therefore, it remains difficult to delimit boundaries for these vague spatial objects.

In this chapter, we present the method to estimate boundaries of a vague spatial description. This problem involves the estimation of the different places and spatial relations found in the text, and the combination of the imprecise boundaries. The objective is to propose an automatic pipeline to delimit an area where the property described in the advertisement should be within. We first propose to empirically estimate the boundaries of the different places and the spatial relations, and we focus on how to represent their imprecision. We also propose a method to combine vague places by using fuzzy set theory in order to retrieve an approximate location for each advertisement.

The remainder of this chapter is structured as follows: Section 2 details a preliminary study to geocode places thanks to existing gazetteers, and shows their limitations given our data. Sections 3 and 4 describe the method to estimate the boundaries of the vernacular places and spatial relations as well as the representation of the vague boundaries to display them on a map, and apply spatial operations. Section 5 presents the aggregation of the different places to approximately geolocate each real estate advertisement. Finally, Section 6 evaluates the method to combine the vague spatial information and discusses the results.

2 Toponyms and Spatial Prepositions Analysis

In Chapter 3, we proposed to automatically extract spatial entities from free-text descriptions. The second step is to convert the unstructured description of places into unambiguous spatial location (e.g., coordinates, polygons), also known as geocoding. Most of the previous works has focused on toponym resolution, which is a complementary task to Topoynm recognition, to match a toponym with an unambiguous spatial footprint [Leidner, 2008; Lieberman, 2012]. The traditional approach of Toponym Resolution tends to leverage gazetteers such as GeoNames[1] or OpenStreetMap[2] (OSM) to match the extracted toponyms to existing knowledge [Overell, 2008; Buscaldi, 2008; Lieberman, 2012]. Nevertheless, this approach faces several challenges such as the completeness of the gazetteers or the referent ambiguity. The referent ambiguity arises when a place name does not refer to a unique location (e.g., Paris is the capital of France and a city of Texas, USA) [Leidner, 2008]. In our study, we proposed to query OpenStreetMap and GeoNames to retrieve the spatial footprints for the extracted toponyms in Nice, Cannes, Antibes and Grasse, that are the 4 biggest cities in the Alpes-Maritimes. We aimed at verifying the completeness of these gazetteers and the ease to disambiguate toponyms. OpenStreetMap is a Volunteered Geographic Information (VGI) [Goodchild, 2007] and provided by a large community over the Web, while GeoNames mostly gathers data from official public agencies around the world (e.g., IGN[3], INSEE[4], Open Data France[5], etc.). We used the Python client GeoPy[6], which facilitates the access to popular geocoding web services, to browse OSM and GeoNames data. Altough they provide a structured form of the search query, we processed free-form queries formed of the name of the toponym, the feature if it exists, and the city (e.g., 'The Promenade des Anglais in Nice' is searched as 'promenade anglais, Nice, France'). The output can contain several possible locations, and each location is composed of its name, its spatial footprint (e.g., coordinates, polygons, etc.) and other information such as its identifier or its class in OSM or GeoNames. In order to reduce the number of possible locations, we filtered the locations by only keeping the ones that are within the footprint of the city. Figures 4.1 and 4.2 present the final number of locations found for each toponym. First, we notice that OSM matched more toponyms than GeoNames for almost all the city. GeoNames is a database mostly provided by official public agencies and does not contain all places because of their relative insignificance to its coverage (e.g., worldwide coverage). Furthermore, although we removed

[1] https://www.geonames.org/
[2] https://www.openstreetmap.fr/
[3] http://www.ign.fr/
[4] http://www.insee.fr
[5] http://www.data.gouv.fr/
[6] https://geopy.readthedocs.io/

the locations outside the city, only a few toponyms match with a unique location which implies to develop a method to disambiguate them. In addition to the disambiguation process, the knowledge-based approach does not seem suitable for our study since we have extracted a lot of toponyms that are not recorded in these two data sources.

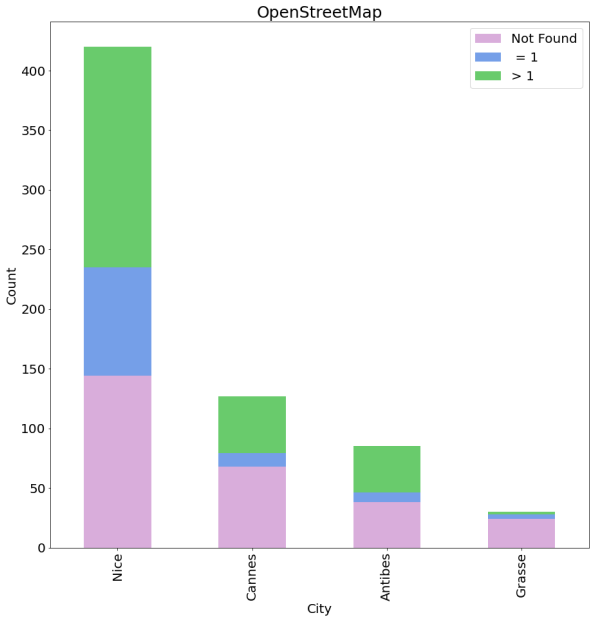

Figure 4.1: Number of toponyms retrieved by OpenStreetMap by city.

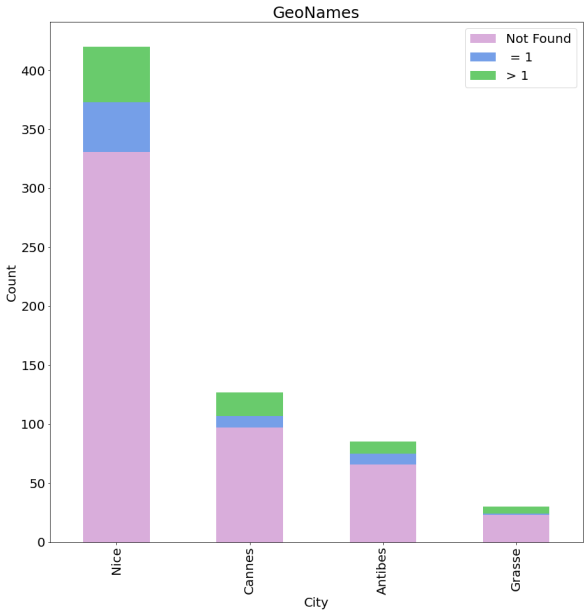

Figure 4.2: Number of toponyms retrieved by GeoNames by city.

On the other hand, a spatial description is also composed of a spatial relation term (e.g., near, next to) and the geocoding process should take it into account. For instance, the preposition *next to*, in the spatial description *next to the Riquier train station*, gives a precision of the location and could improve the accuracy of the geocoding. To consider spatial prepositions, previous works have developed methods such as acceptance models [M Hall, 2011; Platonov, 2018; Schockaert, 2008; Moratz, 2006; Skoumas, 2016; Aflaki, 2022]. The acceptance models are often probabilistic or predictive, and define a threshold or describe an area in which a preposition could validly be used. Moreover, several studies have pointed out that the context is very important to interpret a spatial preposition [Carlson, 2005; Herskovits, 1985; Stock, 2018; Tyler, 2003]. Recently, Aflaki et al. [Aflaki, 2022] extracted spatial descriptions from the Google search engine and studied nine spatial prepositions across three locations. They compared the acceptance thresholds and the variations according to the context. They showed that some prepositions are used to describe larger distance, and the reference object influences the selection of the prepositions.

In our work, we compared the use of the preposition *near* in the real estate advertisements according to the type of objects and the territory. Indeed, the previous

works mainly focused on the context of the object (e.g., size, type, etc.) but the use of a preposition could also be influenced by the culture or the territory (e.g., countryside and city). We gathered real estate advertisements with accurate coordinates in different cities (see the description of the dataset in Chapter 5) mentioning the proximity of either school, train station or airport with the preposition *near*. The referent object is not a place name but only a type of spatial object (e.g., '*the apartment is near the school*'). We also extracted all the official schools, train stations and airports as well as their coordinates. Then, we computed the shortest distance between the real estate advertisements and the referent object, and represented the distribution of the distance with a boxplot.

The first experience, presented Figure 4.3, describes the distribution of the distance for the three object types (i.e., school, train station and airport) in the city of Nice. As expected, the size of the object plays an important role in the acceptance threshold. Indeed, the object *Airport*, which is bigger than *Train Station* and *School*, has a higher median and range of distance. Moreover, the scarcity of the object and the frequency of use are also important. For instance, there is only one airport in Nice whereas we extracted 94 schools. The second experience compared how the preposition *near* associated with the amenity *School* is used for different cities in the Alpes-Maritimes. We classified the cities according to their location in the territory: *Coast, Moyen-Pays and Haut-Pays*. The topography of the Alpes-Maritimes is very particular since it is surrounded by the Alps and the Mediterranean sea. The coastal area is very urbanized and populated and included almost all the cities. The terms *Moyen-Pays* and *Haut-Pays* are used to describe the cities located in the mountains according to their distance to the coast. Figure 4.4 shows the distribution of the distance for each selected city. We noticed that there is a difference between the coast and the mountains, and in particular the *Haut-Pays*. Indeed, this area is very far from the coast and sparsely populated. Puget-Theniers was the only village with enough advertisements to compare with the other cities. Nevertheless, the properties are often outside the centre of the village and the preposition *near* is used very differently compared to the other cities (i.e., the median is around 7.5 km). Furthermore, the size of the city and the scarcity of the object play again an important role: the city of Mougins is located in the coast but smaller than the city of Grasse, which is shown by the higher median and range of distance. Finally, the third experience confirms the results of the two other experiences: the type of object, its scarcity and the type of city influence the distance. We also noticed that similar cities (i.e., Cannes, Antibes and Grasse) have similar distributions. In a nutshell, these experiences have pointed out that geocoding a spatial description implies to create a model for each preposition, each type of object and each type of city. Several research questions arise from this study and are addressed in the following section:

- **RQ1**: How to geocode locations extracted from the advertisements, and in particular local places ?
- **RQ2**: How to take into account the prepositions to geocode a spatial description ?
- **RQ3**: How to deal with the vagueness of the prepositions ?

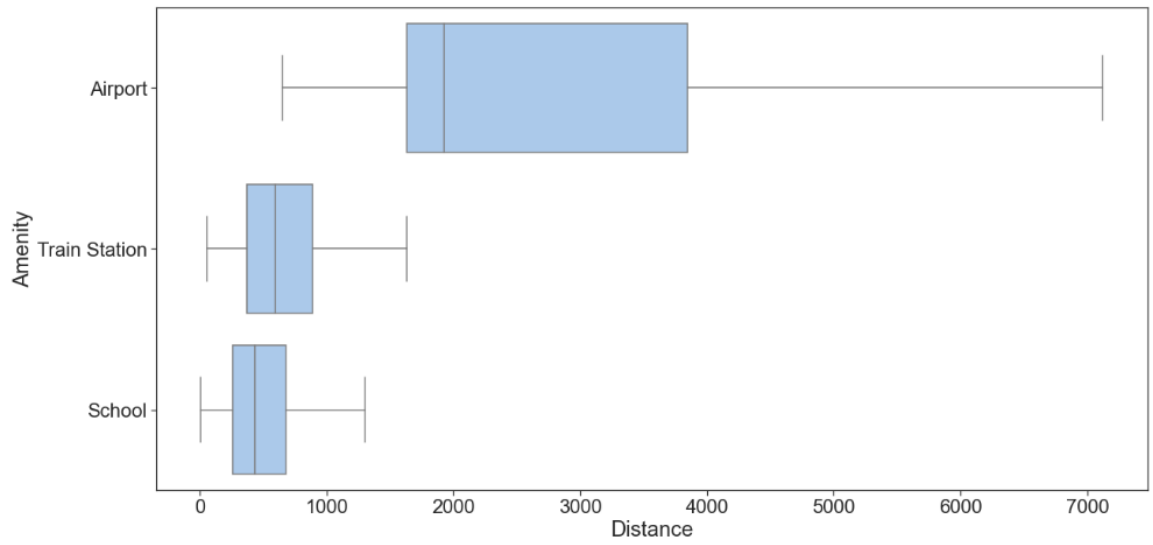

Figure 4.3: Boxplot of the distance (in meter) for the preposition *near* and three amenities in Nice, France.

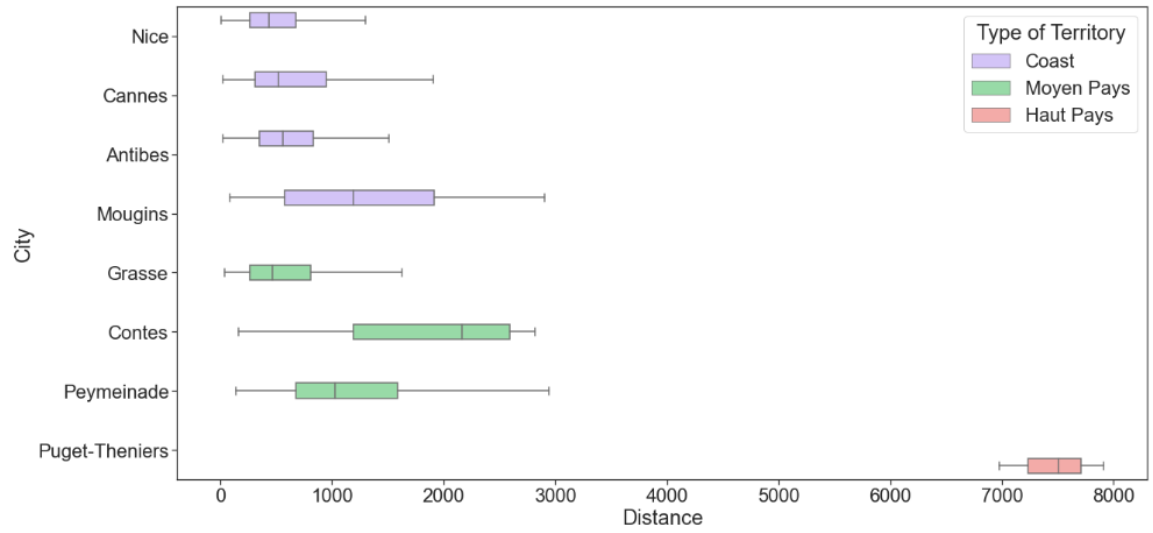

Figure 4.4: Boxplot of the distance (in meter) for the preposition *near* and the amenity *school* in several cities located in the Alpes-Maritimes.

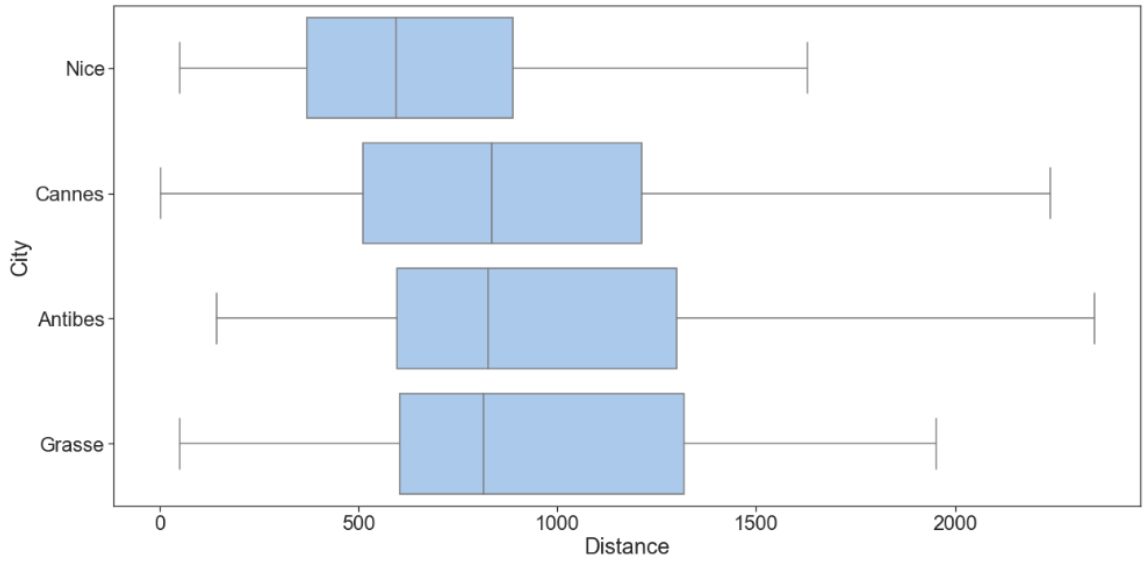

Figure 4.5: Boxplot of the distance (in meter) for the preposition *near* and the amenity *train station* in several cities located in the Alpes-Maritimes.

3 Learning Density Estimation of Vague Spatial Descriptions

Although the description of the location and environment of a real estate advertisement in the text is often short, it gives clues about the approximate property's location. The different pieces of information could be crossed to refine the approximate location and find an area in which the property should be within. Nevertheless, the first step is to retrieve and estimate the boundaries of each spatial information extracted in the text.

In this section, we focus on the estimation of the boundaries of the spatial descriptions extracted in the text. In Section 2, we showed that the method, presented in Chapter 3, has extracted a lot of place names that are not found in official gazetteers since their relative insignificance to the purpose of these data sources. Moreover, place names form a very small part of how a location is described, and their sole geocoding could lead to errors. Several studies proposed to estimate boundaries to enrich gazetteers with local place names [Hu, 2019], thematic places [McKenzie, 2017a] or vague locations [Jones, 2008] by using geotagged data dealing with the same spatial description. While a small part of our real estate advertisements are precisely geolocated (i.e., reliable latitude/longitude pair), they could be used to approximate the places.

One approach to construct regions from point data is to use kernel density estimation (KDE). This method generates a smooth surface by inferring the shape from a sample of point data, and gives a value for each point of the support by using a probability density function. The kernel estimator is defined as:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{4.1}$$

where n is the number of points, h is the bandwidth that determines the amount of smoothing, K is the kernel function, x is the location to be estimated, and x_i is the coordinates of a real estate advertisement mentioning the location x.

In this work, we proposed to extend this method to estimate the boundaries of the spatial descriptions extracted in the real estate advertisements. In Chapter 3, we identified and presented the different ways to describe a location in a real estate advertisement:

- Toponym (e.g., Nice, Cannes)
- Feature (e.g., the train station, the university)
- Toponym + Feature (e.g., Riquier station, Avenue Jean Médecin)
- Spatial Relation + Toponym (e.g., Close to Nice)
- Spatial Relation + Feature (e.g., Nearby the university)
- Spatial Relation + Toponym + Feature (e.g., Riquier station 5 minutes away)

Most of the previous works using KDE focus on the estimation of the three first type of spatial descriptions, and do not take the spatial relations into account. However, the inhabitants have often a vague idea of where a location described by a spatial relation (e.g., 'Nearby the beach') is (or is not) in the city. Therefore, we also proposed to applied this method to this type of spatial description.

To generate regions for each extracted place, we first collected all the geotagged advertisements and only kept the reliable one. Indeed, the real estate agents often hide the exact position since the deal between their agency and the owner might not be exclusive. Thus, we designed a rule based on the frequency of the latitude/longitude pair in the dataset in order to detect the very frequent coordinates that correspond to the centre of the city or neighborhood. We set a threshold to 20 occurrences, and we only kept 15% of the advertisements in the dataset. Secondly, we post-processed the information extraction to clean the texts from misspelling, plural and abbreviations. We replaced the well-known abbreviations by its correct terms (e.g., 'min' for 'minutes', 'm' for 'meters'). We also applied the Jaro-Winkler distance to retrieve very similar terms and correct misspelling.

The last treatment focused on removing outliers since some spatial footprints are very far from the distribution. Indeed, we noticed that some advertisements, mentioning a place, are geolocated very far from the other advertisements mentioning the

same place. This issue might arise because of errors during the information extraction stage. For instance, our model presented Chapter 3 could have extracted only a place name or feature without its spatial relation (e.g., 'Riquier' instead of 'Near Riquier'). Moreover, it could be due to a detection of an inaccurate geolocation since our rule to keep reliable coordinates is only based on the number of occurrences. Therefore, we computed the Mahalanobis distance, which is a measure of the distance between a point P and a distribution Q. It calculates the distance to the centroid of the data, and the larger the value, the more likely the point is to be an outlier. To decide if a point is an outlier, the square of Mahalanobis distance is compared against a Chi-Qquare (χ^2) distribution with degrees of freedom equal to the number of variables (i.e., in our case the number of variables is two: latitude and longitude). We removed all the point where the value is higher than the 99^{th} percentile of the χ^2 distribution. The formula to compute Mahalanobis distance D is given as follows:

$$D = \sqrt{(x-m)^T \cdot C^{-1} \cdot (x-m)}, \tag{4.2}$$

where x is the vector of the observation, m the vector of the centroid, and C the covariance matrix.

Finally, we applied KDE on the geotagged advertisements for each place. We only computed the density estimation for locations having more than 10 geotagged advertisements mentioning it. A drawback of this method is the amount of data required to get a reliable estimation. Moreover, two important parameters must be decided in kernel density: the kernel bandwidth h and the kernel function K. We chose a Gaussian kernel function, defined as follows:

$$K(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{-1}{2}(\frac{x-x_i}{h})^2} \tag{4.3}$$

Regarding the bandwidth, we applied the rule of thumb designed by Scott **??**. The bandwidth is proportional to $n^{(-1/(d+4))}$ where n is the number of points and d is the number of dimensions. This rule is very fast to compute and typically useful for estimating gradual trend.

4 Fuzzy Representation

The estimated density function with the KDE method gives the shape of the surface of the region. However, the representation of the boundaries as a spatial object (e.g., polygon) remains important to display them on a map, and apply spatial operations (e.g., union, intersection, etc.). First, a real estate property is always within a parcel which led us to choose parcels as spatial representation of the estimated locations.

Then, we applied the density function to each parcel to get a value of membership to the estimated place. Sharp boundaries can be generated by selecting a value that serves as a threshold. The output is a set of parcels that verify the criteria. Nevertheless, to represent imprecise regions, traditional methods are often based on the supervaluation method [Kulik, 2001] where a vague region is defined by a set of sharp regions, which means to define several thresholds. This method is similar to fuzzy set theory since each sharp boundary could be seen as an α-cut.

Definition 4.1 (α-cut)**.** An α-cut, denoted F_α, is a crisp set where all the elements belongs to F with a degree higher or equal to α:

$$F_\alpha = \{x \in F, \mu_F(x) \geq \alpha\}, \tag{4.4}$$

where $\mu_F(x)$ is the membership function.

In our work, we proposed to transform the density function to a membership function of a fuzzy set in order to define several α-cuts as the representation of an imprecise location. Moreover, fuzzy set theory allows us to represent uncertain as well as imprecise objects. The Gaussian distribution estimated in Section 3 is easily convertible since it is a frequent membership function. Therefore, we normalized the values of the density function between 0 and 1. Then, we set thresholds and created 5 different α-cuts for each place (i.e., $\alpha \in [0.2, 0.4, 0.6, 0.8, 1]$).

The method of estimating the vague places has not been evaluated because we did not have all the official geometries of each place. Indeed, some places are vernacular and do not exist in official databases. Moreover, the estimation represents the real estate agents' knowledge that could be different from the official one but not necessarily incorrect. Therefore, we checked some places to ensure their validity (see Appendix C). The two first examples, in Figures 4.6 and 4.7, represent two place names. The first one is *La Vieille Ville* (i.e., the old town) in Nice which is an official toponym, while the second one is a local toponym called *La Banane* in Cannes. Although the official boundary of the first place incorporates the brown, purple and a part of the red areas, our estimation has its core in the very old town of the city where the buildings are very small and close to each other. For the second place, there is not official boundary since it is a vernacular place. However, the term *La Banane* refers to the shape of the area which is a banana, as shown by our estimation.

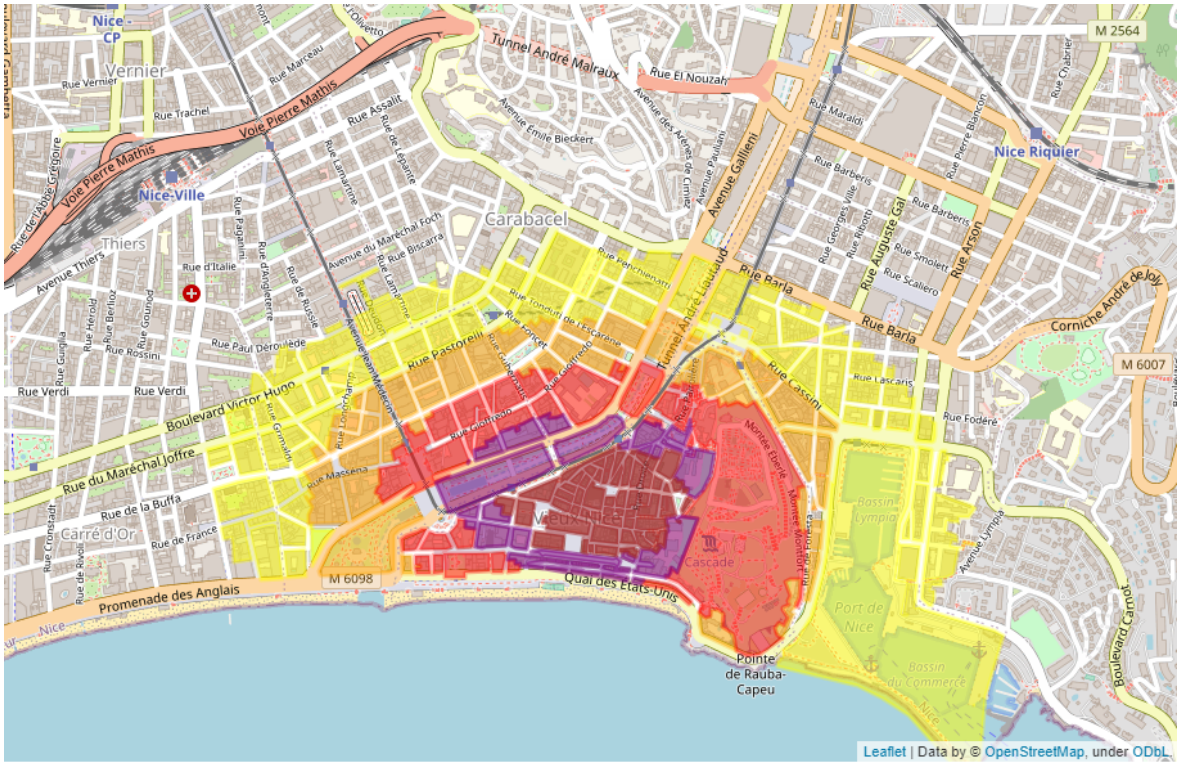

Figure 4.6: La Vieille-Ville, Nice.

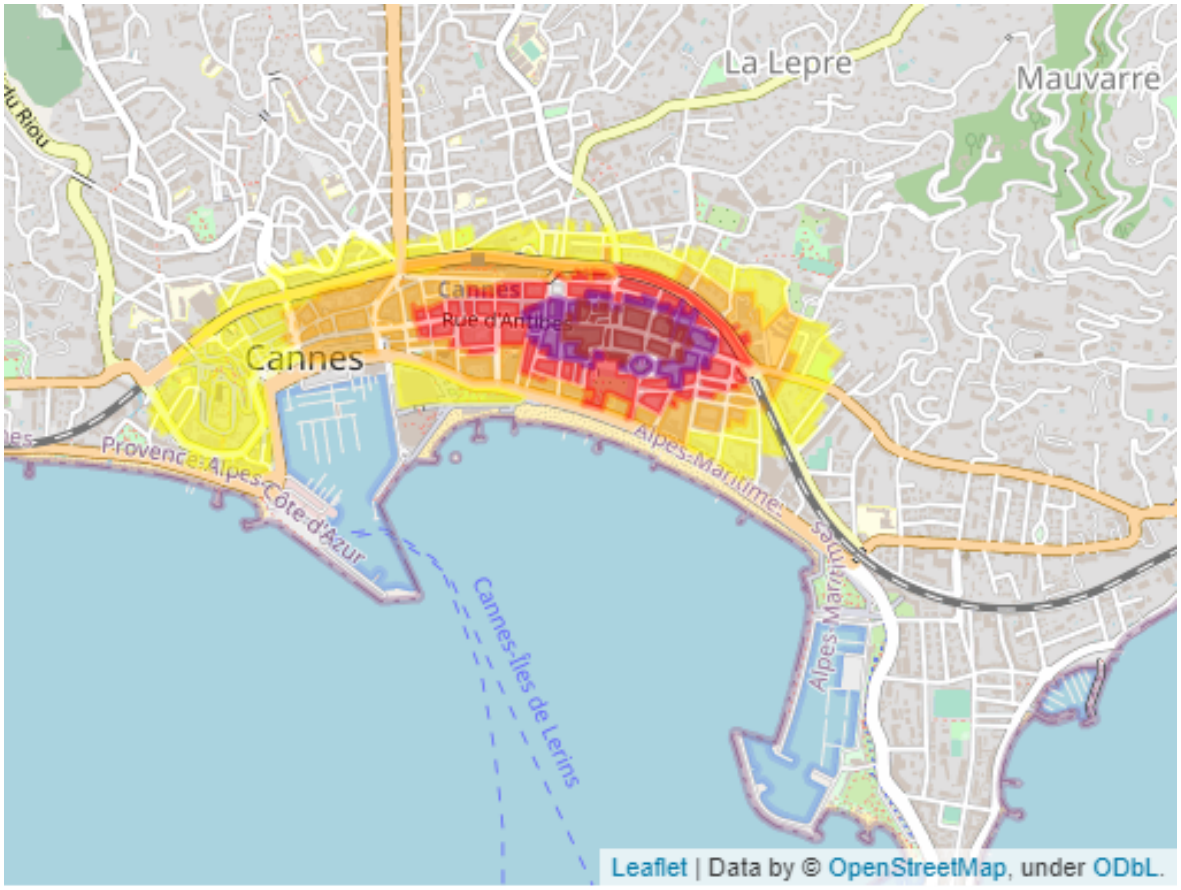

Figure 4.7: La Banane, Cannes.

The second example shows the fuzzy representation of the uncertain place *Downtown* in four large cities in the Alpes-Maritimes. *Downtown* is a very vague place since it does not exist any official limit, and everyone has their own opinion on where it starts and where it ends. A high number of studies in GIS have been conducted to automatically estimate the position of *Downtown* in different cities [Montello, 2003]. Our method gives a partial answer of what and where *Downtown* is in Antibes, Cannes, Nice, and Grasse according to the real estate agents. We can see that the downtown in Nice is pretty well defined around a main avenue or square. On the other hand, Antibes and Cannes seem to have two downtowns since both have two neighborhoods, far from the centre, that might be considered as smaller cities (Juan-les-Pins and Cannes La Bocca) and where a small downtown could be found. Finally, Grasse has a small downtown and it seems more difficult to estimate a reliable region. Indeed, the brown area is correctly located in the downtown but it is also a bit large. Therefore, the other α-cuts are too big and cover a large area of the city. This example shows a limitation of our approach for smaller cities.

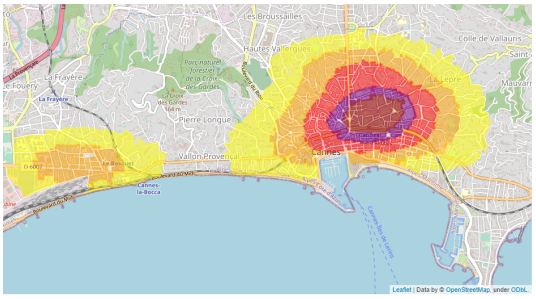

Figure 4.8: Downtown Cannes.

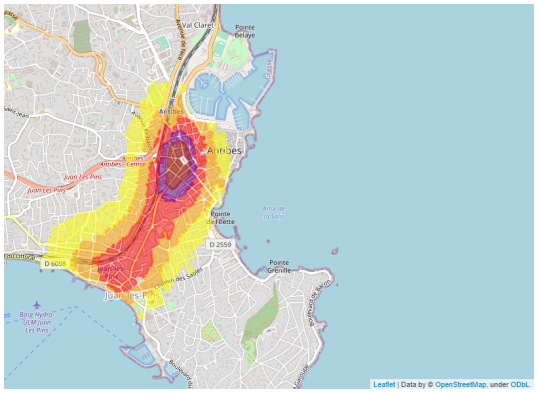

Figure 4.9: Downtown Antibes.

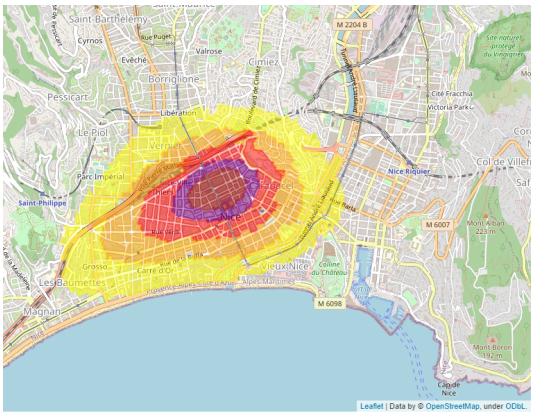

Figure 4.10: Downtown Nice.

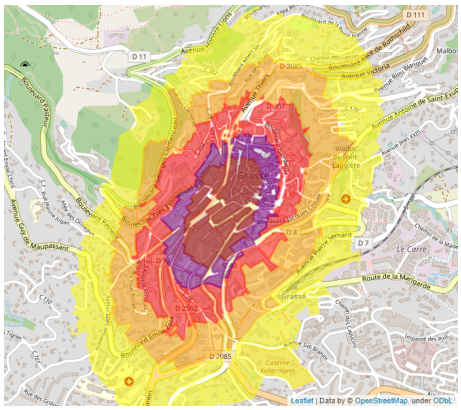

Figure 4.11: Downtown Grasse.

Lastly, Figures 4.12, 4.13 and 4.14 show the estimated areas of the location *'Near the train station'* in Cannes, Antibes and Nice. In Section 2, we studied this place and pinpointed the difficulty to represent the spatial relation *near* because of the context. In this example, we added all the train stations found in the three cities (i.e., blue icons) to compare their location to our estimation. In Antibes, we identified two train stations that are both busy places. Figure 4.13 shows that our estimated area is around the two train stations. On the other hand, Nice and Cannes have train stations that are far from the estimated area. In Cannes, our boundaries are around the main train station while the second one is a very small station. In Nice, we identified two areas around the central station and the second most important station. The other train stations are too small to be mentioned in a Real Estate advertisement. This example shows that the estimation of *'Near the train station'* is difficult since it highly depends on the context.

In a nutshell, our method seems to be pretty accurate for all the presented examples but some limitations arise because of the lack of data or the size of the city. As a perspective, an user evaluation should be conducted to extend our analysis to potential users and experts (e.g., real estate agents) and validate our estimations.

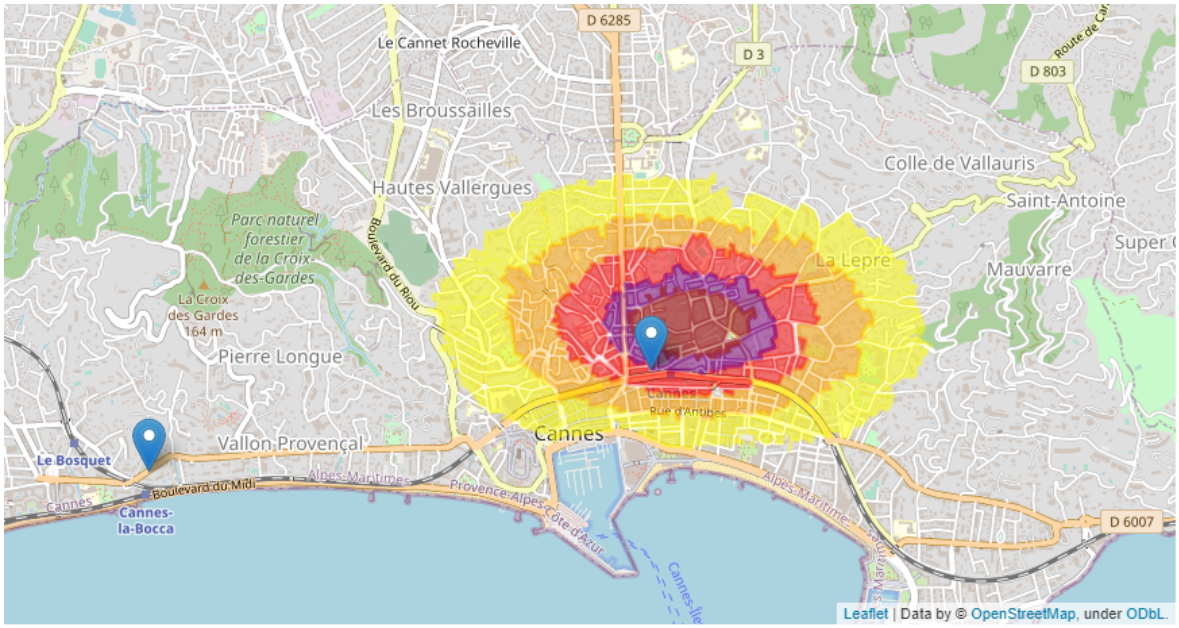

Figure 4.12: Near the train station, Cannes.

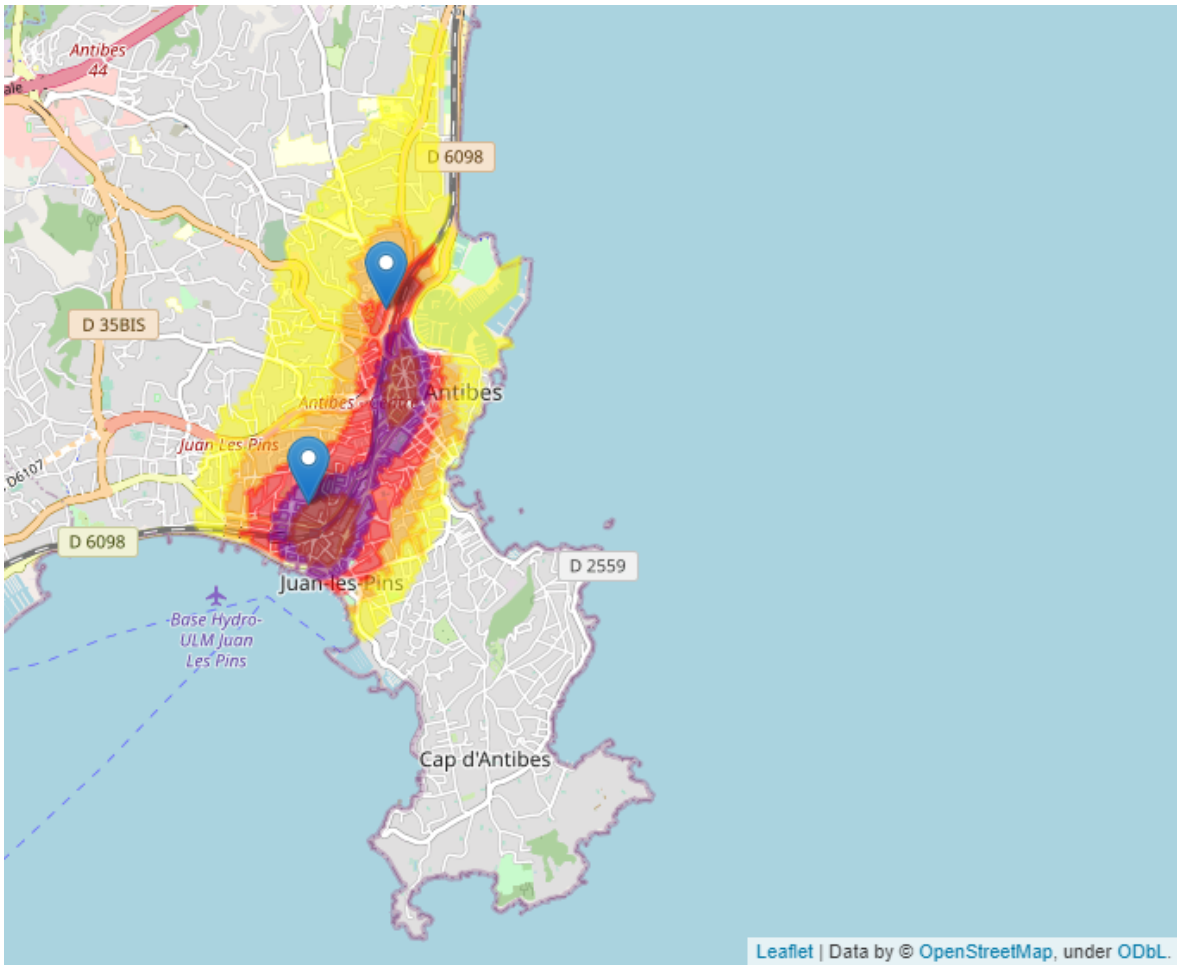

Figure 4.13: Near the train station, Antibes.

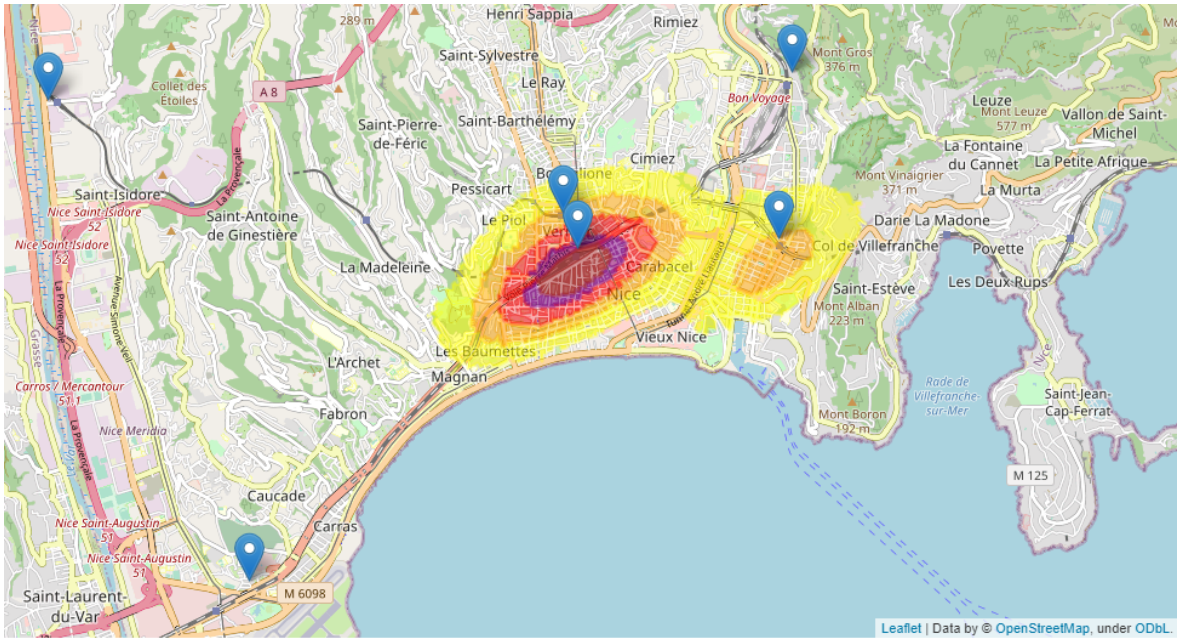

Figure 4.14: Near the train station, Nice.

5 Information Fusion

The estimation and representation of vague spatial location extracted in the real estate advertisements allow us to approximately geolocate each real estate property. For instance, the following example, in Figure 4.15, highlights several locations that give a clue to estimate the position. Although the first place refers to the city which is not a new fact, the three other pieces of information help to delineate the area where the real estate property should be located. In this example, we could expect that the property is very close to the sea since the three locations are nearby.

The different pieces of information can be seen as multiple data sources that have to be aggregated to provide a useful and unique piece of information. The process of information fusion consists of combining several information from multiple sources to create a more complete picture of a given phenomenon. This process is divided into four steps: modelization, estimation, combination and fusion. The first stage aims at formalizing the representation of the information. We chose to use fuzzy set theory as presented in Section 4 in order to represent the imprecision and vagueness of the information. This approach is more suitable than the probabilistic one because the probability approach could only model the uncertainty of the information. For each source of information S_j and decision d_i to take, the information M_i^j is represented as follows:

$$M_i^j(x) = \mu_i^j(x), \tag{4.5}$$

where μ_i^j is the membership function computed for each information in Section 3.

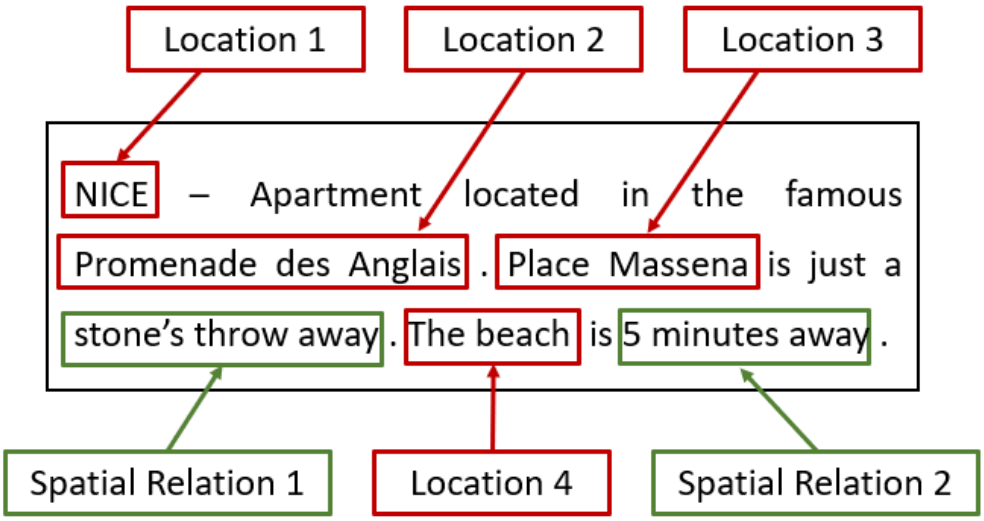

Figure 4.15: Example of spatial information in a Real Estate advertisement.

The combination stage intends to gather all the information, which is the most important part of the information fusion. Fuzzy set theory offers a large choice of operators [Zimmermann, 2011] that could be used to fuse the information. We chose to compute the ordered weighted averaging (OWA) [Yager, 1988] membership function which is a compromise between the most restrictive operation *min* (i.e., intersection) and the less restrictive operation *max* (i.e., union). The OWA operator isdefined as follows:

Definition 5.1.

$$\mu_{OWA}(x) = \sum_j w_j \mu_j(x)$$

where $\sum_j w_j = 1$.
If the OWA-Operator is the arithmetic mean then $\forall j, w_j = \frac{1}{n}$, where n is the number of information items.

The final stage relies on defining a criterion to make a decision on the approximate location of the real estate property. In our work, we aggregated all the spatial location found in a text, and computed the OWA operator for each parcel. Therefore, each parcel has a certain value describing the membership of a real estate property to it. To approximate the location, we created 5 different α-cuts (i.e., $\alpha \in [0.2, 0.4, 0.6, 0.8, 1]$) that give the degree of membership of the real estate property to each area. Figure 4.16 represents an application of the pipeline of information fusion. In this example,

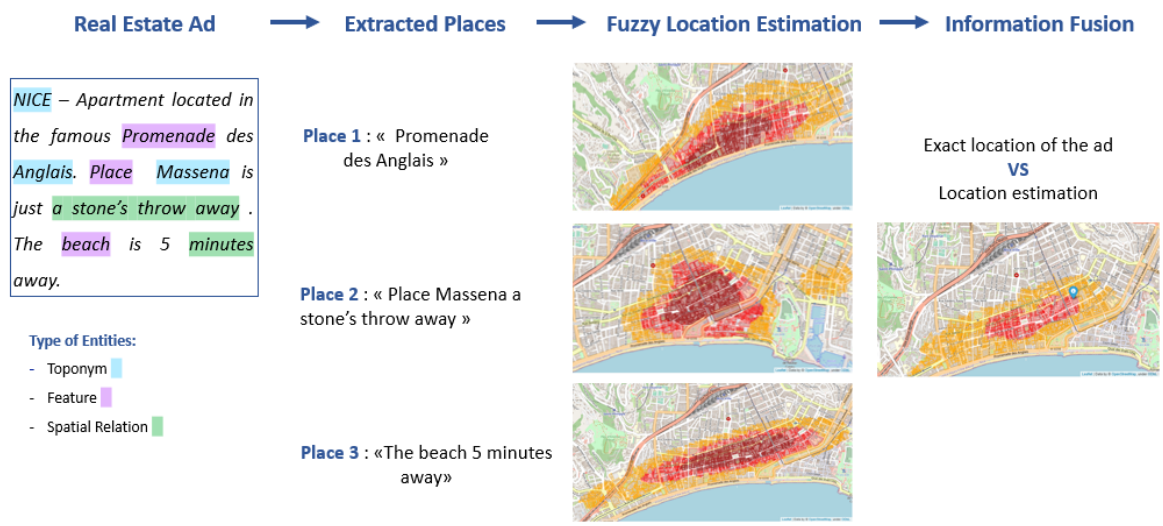

Figure 4.16: Example of information fusion to retrieve the location of a real property.

we extracted and estimated 3 places. Then, we combined the value of each parcel for each information to get an approximate position of the real estate property. Finally, we can notice that the exact location (i.e., blue icon) is within a parcel with a high degree of membership.

6 Evaluation

The evaluation process aims at assessing the performance of the pipeline by applying it to new advertisements located in Nice, Cannes and Antibes. In this section, we first describe the experimental setup and metrics to evaluate our method. Then, we present and discuss the results.

6.1 Experimental setup

We evaluated the pipeline by applying it to a new dataset. First, we extracted new advertisements with reliable coordinates that have not been used in the process of learning the area and located in Nice, Cannes and Antibes. Then, we applied the whole pipeline to each advertisement: (1) Information Extraction, (2) Retrieving area of each places and (3) Combination of each area. Finally, we evaluated the approximate location found by the model by comparing it to the exact coordinates of the advertisement.

To the best of our knowledge, there is no standard evaluation to measure the quality of a fuzzy and imprecise location. In information retrieval and, in particular toponym resolution, the Precision (P), Recall (R) and F-Score (F) are widely used to evaluate

City	Number of Ads
Nice	11,916
Cannes	2,269
Antibes	1,902

Table 4.1: Number of evaluated advertisements for each city.

methods. As we do have coordinates for each ad, we suggest to use these three metrics with some adaptations to our problem as proposed by Leidner [Leidner, 2008]. We define the following notation:

- L_I^α : number of ads located within a given α-cut;
- L_O^α : number of ads located outside a given α-cut;
- L_U^α : number of ads for which the model did not find an area for a given α-cut.

Therefore, we can compute Precision, Recall and F1-Score as follows :

$$\begin{aligned} P &= \frac{L_I^\alpha}{L_I^\alpha + L_O^\alpha} \\ R &= \frac{L_I^\alpha}{L_I^\alpha + L_O^\alpha + L_U^\alpha} \\ F1 &= 2 \times \frac{P \times R}{P + R} \end{aligned}$$

Precision refers to the ratio of the number of ads correctly found within the area of a given α-cut among the number of ads for which the model found an area for the given α-cut. A high precision means that when the model finds an area for a given α-cut, then the ads are within the area. Recall is the ratio of the number of ads correctly found within the area of a given α-cut among the total number of ads. For this task, the recall has the same definition as accuracy. Recall gives information about the capability of the model to find a zone for a given α-cut. Indeed, a high precision and a low recall mean that the model is good at finding ads within a fuzzy location, but fails to resolve many areas for a given α-cut. F1-Score summarizes precision and recall in one metric by computing their harmonic mean.

A limitation of these metrics is that we only differentiate if an advertisement is inside or outside a given alpha-cut. This binary distinction does not take the area of the zone into account. A fuzzy area that equals to the entire city would not be penalized, whereas it is not precise at all. On the other hand, a small fuzzy area where the advertisement is not within but at a very close distance would be penalized. Thus, we propose to also use a continuous metric called Root Mean Squared Distance (RMSD) defined by Leidner [Leidner, 2008]. RMSD is derived from the Root Mean

Squared Error, which is frequently used to compare predicted and observed values. Here, RMSD is the root of the arithmetic mean of the squared distance, in meters, between the coordinates of the advertisment p_i and the centroid c_i of a given alpha-cut:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Delta(p_i, c_i)^2}.$$

Nevertheless, RMSD only uses advertisements with an area found for a given alpha-cut, and does not evaluate the performance of the model to find a fuzzy representation. Thus, a good performance of our model should be a good compromise between a high F1-score and a low RMSD.

6.2 Results and Discussion

In this section, we discuss the results of the experiment by applying the method to the three biggest cities in the French Riviera: Nice, Cannes and Antibes. We computed the evaluation metrics for several values of α as summarized in Table 4.2. We noticed that the method reaches high precision for the three cities, in particular for a low α. On the other hand, the recall is always lower which highlights that the method does not always find an area for a given α-cut (i.e., low recall), but is very good at giving a correct area (i.e., high precision). Moreover, the greater the α the lower the performance which could be explained by the size of the area. Indeed, the area for α equals to 0 is very large (i.e., around 30% of the total of the city) which leads to a high precision. Nevertheless, for α equals to 0.4, the area is pretty small for each city (i.e., less than 10% of the total of the city) and the precision is higher than 50%. Regarding the RMSD, the distance between the location of an advertisement and the centroid of the area is between 1.5 and 2.5 kilometers. The RMSD shows that the advertisements are not so far from the centroid despite a low precision for higher α. Finally, the evaluation shows that the method is robust since it has similar results for the three cities. The city of Nice has slightly better results regarding the precision, recall and F1-score which could be explained by the larger area. However, the RMSD for Antibes is lower which shows that the computed areas are more precise and accurate.

We also compared the OWA operator to the minimum and maximum operators as defined in Chapter 2 by applying them to the city of Antibes. Table 4.3 shows that the OWA operator is a good compromise between the minimum and maximum. Indeed, the minimum operator returns very small areas which leads to very low precision and recall despite a better RSMD. On the other hand, the maximum operator has better precision and recall for each α, but the areas are very large and the RMSD is greater. Therefore, the OWA operator performs well enough to retrieve an accurate fuzzy area since the F1-score is high enough for certain α and the RMSD is around 1.6 kilometers.

α	Metrics	Nice	Cannes	Antibes
0	P	0.90	0.87	0.86
	R	0.67	0.65	0.65
	F1-score	0.77	0.74	0.74
	RMSD	2421	1915	1751
	Area (km^2)	20.2	7.9	11.2
0.2	P	0.82	0.76	0.73
	R	0.61	0.57	0.55
	F1-score	0.70	0.65	0.63
	RMSD	2267	1895	1644
	Area (km^2)	9.2	4.6	6.2
0.4	P	0.59	0.54	0.49
	R	0.44	0.41	0.37
	F1-score	0.50	0.47	0.42
	RMSD	2197	1858	1655
	Area (km^2)	2.9	1.9	2.5
0.6	P	0.34	0.34	0.28
	R	0.23	0.24	0.20
	F1-score	0.27	0.28	0.23
	RMSD	2089	1810	1697
	Area (km^2)	0.89	0.71	0.90
0.8	P	0.21	0.19	0.14
	R	0.07	0.09	0.07
	F1-score	0.11	0.12	0.09
	RMSD	2051	1720	1781
	Area (km^2)	0.40	0.28	0.37
1	P	0.13	0.11	0.08
	R	0.02	0.03	0.02
	F1-score	0.04	0.05	0.03
	RMSD	2207	1784	1908
	Area (km^2)	0.25	0.17	0.19

Table 4.2: Results of the evaluation for Nice, Cannes and Antibes.

α	Metrics	Minimum (Intersection)	OWA	Maximum (Union)
0	P	0.59	0.86	0.86
	R	0.43	0.65	0.64
	F1-score	0.5	0.74	0.73
	RMSD	1543	1751	1751
	Area (km^2)	3.6	11.2	11.2
0.2	P	0.59	0.73	0.86
	R	0.43	0.55	0.64
	F1-score	0.5	0.63	0.73
	RMSD	1543	1644	1751
	Area (km^2)	3.6	6.2	11.2
0.4	P	0.36	0.49	0.70
	R	0.24	0.37	0.52
	F1-score	0.29	0.42	0.60
	RMSD	1610	1655	1706
	Area (km^2)	1.7	2.5	6.5
0.6	P	0.24	0.28	0.52
	R	0.11	0.20	0.39
	F1-score	0.15	0.23	0.45
	RMSD	1784	1697	1727
	Area (km^2)	0.87	0.90	3.6
0.8	P	0.13	0.14	0.34
	R	0.04	0.07	0.26
	F1-score	0.06	0.09	0.3
	RMSD	1879	1781	1740
	Area (km^2)	0.42	0.37	1.6
1	P	0.08	0.08	0.21
	R	0.02	0.02	0.16
	F1-score	0.03	0.03	0.18
	RMSD	1908	1908	1762
	Area (km^2)	0.19	0.19	0.74

Table 4.3: Comparison of minimum, OWA and maximum operators.

7 Summary and Perspectives

This chapter describes the geocoding of each advertisement from vague spatial information extracted in Chapter 3. We first showed that the spatial objects described in the texts are imprecise because of vernacular places and spatial relations. We studied the completeness of two famous gazetteers (OpenStreetMap and GeoNames) regarding the toponyms extracted in real estate advertisements, and noticed that a large number of toponyms are not easily retrieved in gazetteers. Particularly, GeoNames has a very small coverage, compared to OpenStreetMap, due to its worldwide target audience. Moreover, we compared the use of the preposition *near* in real estate advertisements according to the type of objects and the territory. We showed that modelling a spatial relation to geocode a place highly depends on the context. Then, we described the method to approximate the location of each advertisement. First, we learnt a density estimation for each places extracted in the advertisements to delimit their boundaries, and we transformed the density into a fuzzy set to represent the imprecision on a map and to easily combine the vague information. We presented several examples to show the proficiency of the method to estimate boundaries. The last part focused on the combination of the information to retrieve the approximate location of each advertisement, which can be seen as an aggregation of multiple data sources. We described the chosen modelization and the operator used to aggregate the different sources. Finally, we defined the metrics used to measure the quality of the fuzzy locations and we evaluated this method on the three biggest cities of the French Riviera: Nice, Cannes and Antibes. We showed that the OWA operator is the more suitable for our work and the method is robust since it has similar results for the three cities.

Regarding the limitations and further improvements, we first identified the size of the city. Indeed, we only applied the Information Fusion process on three big cities. This process requires a sufficient amount of data in order to learn an estimation of the boundaries of each place whereas small villages lack data. Moreover, the size of a village is sometimes equal to the size of a neighborhood in a big city. Additionally, it often does not exist Place Of Interest (POI) in villages to delimit the location, which makes it difficult to precisely locate a property except if the street is mentioned. As a perspective, we could use an hybrid approach by geocoding some places using gazetteers if we lack data, while gazetteers suffer from incompleteness. Also, it would be interesting to compare boundaries from the gazetteers and the ones from our estimation. This comparison could either evaluate our estimation or combine the different boundaries of a place in order to get a more reliable representation.

As another future work, we could evaluate the quality of the estimation of the boundaries by conducting a user experiment. In this work, we only assessed the quality of the estimation by discussing the boundaries of well-known places with experts of

SepteoProptech and a PhD student in geography. A thorough evaluation could expand our analysis to potential experts and users (e.g., real estate agents or geographers) and, validate the boundaries by showing them several maps from a sample of places. Then, we could ask users to give a score to the boundaries of each place. This evaluation could allow us to distinguish places that are well represented from those that are too far from the reality. For instance, very frequent POIs (e.g., *Promenade des Anglais*) are very exaggerated and, their limits could be very extended compared to the official ones. On the other hand, places with few data might have narrower boundaries.

Finally, in this work, we did not take the mode of transportation into account despite we extracted it. It would be easy to implement it by adding the mode of transportation into the definition of a place. For instance, '*Promenade des Anglais 5 minutes away walking*' is different from '*Promenade des Anglais 5 minutes away driving*' and, we could create two different places with two different boundaries. Adding this information will help to get a more precise location of the real estate property.

Chapter 5

SURE-KG: A Knowledge Graph to Represent Real Estate and Uncertain Spatial Data from Advertisements

Objectives

This chapter aims at describing SURE-KG, a new knowledge graph built from a real dataset at the core of the industrial application of Septeo-Proptech. First, we define a new ontology to represent Real Estate and uncertain spatial information given several motivating scenarios. Then, we give an overview of the pipeline set up to process the initial corpus and generate the RDF dataset. We also detail the characteristics of the knowledge graph and services made available to exploit it. Finally, we illustrate and discuss potential applications and use cases.

Contents

1 Introduction

In Chapters 3 and 4, we proposed an automatic workflow to extract information from texts and we addressed the problem of geocoding vague spatial descriptions. The produced knowledge is valuable for the company SepteoProptech but also for broader communities, such as the GIS community, which could study it. However, this knowledge is unstructured, which makes it difficult to exploit and share. Thus, representing knowledge is crucial to effectively solve complex tasks, share, and discover new knowledge. A knowledge graph is one of the possible structured representations and has several benefits, such as a more flexible manner to design and maintain data, reasoning with ontologies or discovery of hidden patterns. Furthermore, the Semantic Web promotes the publication and linking of data on the Web in order to be reused by users for wider applications.

This chapter presents SURE-KG, a new knowledge graph built from a real dataset at the core of the industrial application of SepteoProptech. The company aims at using real estate advertisements to give more insights to real estate agents wishing to sell a property (e.g., by comparing a property for sale to similar ones) and analyze the real estate market. However, these data are not structured, and the uncertainty and vagueness used to describe the location and environment of a real estate property need a suitable representation to be exploited, in particular to reason over them. Therefore, the objective is to design an ontology and to build a knowledge graph to represent the extracted information and facilitate their interoperability. To the best our knowledge, this knowledge graph is the first one to represent, query and reason over uncertain and vague spatial data. Particularly, the contributions of this chapter may be summarized as follows:

- we define a new ontology to represent real estate and uncertain spatial information;
- we build an extraction pipeline to process real estate advertisements;
- we generate a RDF dataset and publish it according to the standards and best practices of the linked open data.

The remainder of this chapter is structured as follows. Section 2 details the motivating scenarios and choices made to design the ontology. Section 3 explains the extraction pipeline set up to process the initial corpus and generate the RDF dataset, and describes its characteristics and services made available to exploit it. Finally, Section 4 presents and discusses statistics and illustrates potential applications. Section 5 summarises and concludes this chapter.

2 SURE Ontology

An ontology is a formal representation of the meaning of concepts and their properties within a domain. Its goal is to facilitate data organization, integration and interoperability in a knowledge graph. In the context of Real Estate, the information extracted from the advertisements is sometimes uncertain and vague and needs a suitable representation.

In this section, we present the SURE[1] (Spatial Uncertainty and Real Estate) ontology, which has been developed to formalize and represent a real estate property, its attributes and its location as well as uncertain spatial entities and their boundaries. We describe the domain-specific classes and properties identified by several motivating scenarios, as well as classes extracted from the advertisements. The ontology follows the standards and best practices of the linked open data, and the prefix *sure:* is used to refer to it. Figure 5.1 shows a representation of a real estate advertisement using this ontology.

2.1 Motivating Scenarios

The development of an ontology is often motivated by scenarios that describe real world applications [Uschold, 1996]. A motivating scenario is a story problem which could not be solved by existing ontologies and helps to formalize the proposed ontology. It is composed of a title, a formal description of the problem and several examples of the use case. Given the scenarios, competency questions are defined and expressed in natural language to specify the requirements that the ontology should meet. The competency questions are used to evaluate the ontology. In our work, we adopted a user-centered approach to design this project. The industrial partner (Septeo Proptech) and geographers were closely involved to help us to identify motivating scenarios, competency questions and potential users. The following scenarios name (*Name*), describe (*Desc.*), illustrate (*Ex. 1, 2*) and deduce competency questions (*CQ*).

Scenario 1:

- **Name**: Real Estate Property Search
- **Desc.**: A buyer is looking for a property in a particular city and neighborhood, and precise characteristics. To take a decision, the buyer needs to know the characteristics, the price and the location of the property.
- **Ex. 1**: Mathilde wants to buy a property in Cannes, France. She would like a 1-bedroom flat with a sea view and close to the Croisette. The price should not

[1] http://ns.inria.fr/sure

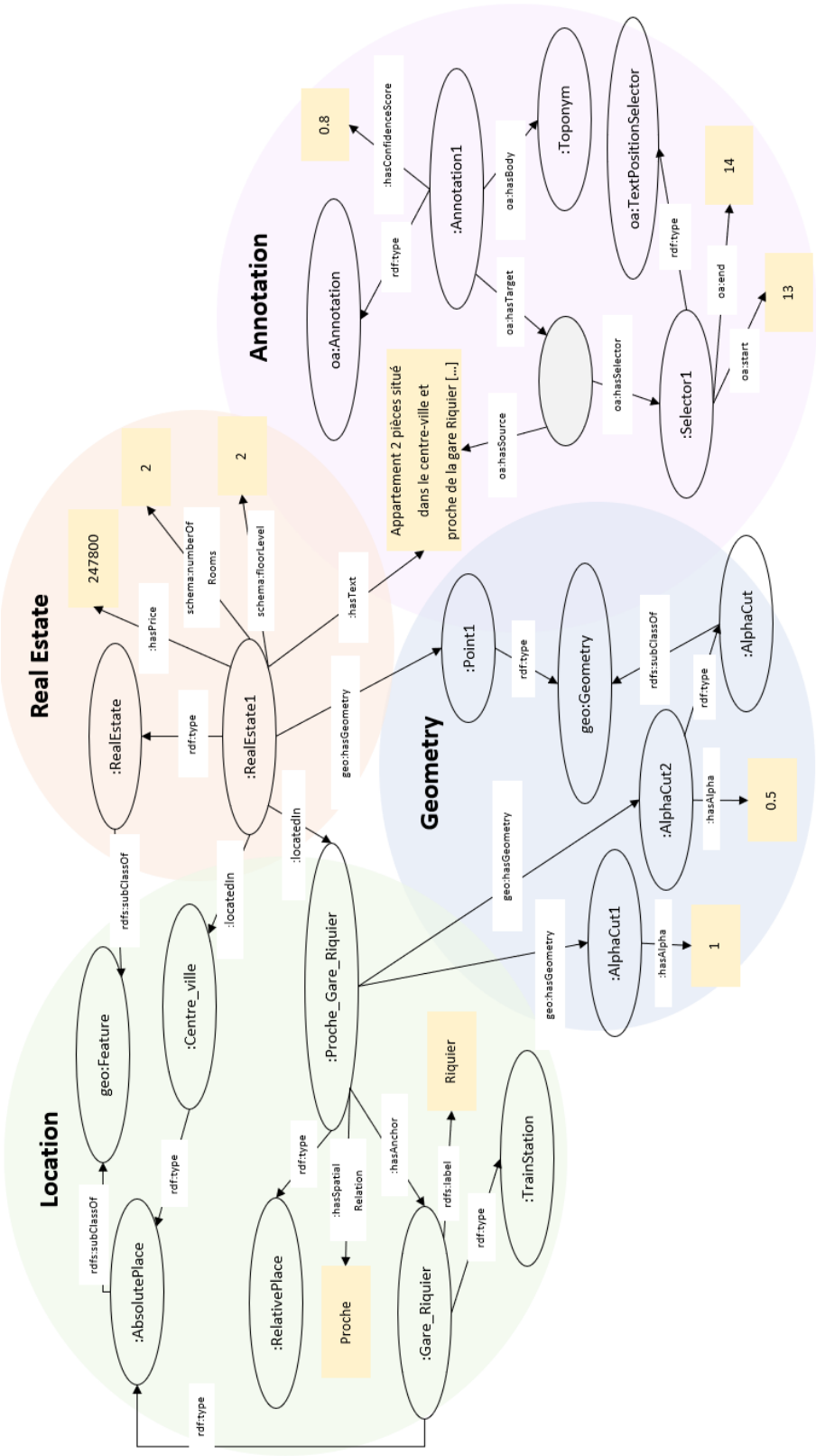

Figure 5.1: Example of the RDF graph of a real estate advertisement using the SURE ontology.

exceed 300,000 €.

- **Ex. 2**: Paul moves to Nice for a new job. He does not know the city and would like a flat in a quiet area, but also close to the public transports and shops.
- **CQ** :
  - What are the characteristics of the property ?
  - In which city is the property located ?
  - In which neighborhood is the property located ?
  - What are the nearby amenities and points of interest (POI) ?
  - What is the type of environment (e.g., quiet, noisy) ?

Scenario 2:

- **Name**: Real Estate Market Analysis
- **Desc.**: A real estate agent might need to understand the market and know the sold/for sale properties to align the price of the property he has to sell. He needs to know the similar properties, the mean price, the number of sells in the neighborhood, etc.
- **Ex. 1**: Denis is a real estate agent and is selling a villa located in Cap d'Antibes, France. However, he is not used to selling properties in this neighborhood. He would like to know what are the average price in this area and the similar properties already sold or for sale before publishing his advertisement.
- **Ex. 2**: David is a real estate developer and is looking for the best neighborhood to build his next building. He needs to know for each neighborhood the price per square meter and the nearby amenities (e.g., schools, transports, etc.)
- **CQ** :
  - What are the other properties located in the same area ?
  - Are they similar to the one for sale ?
  - What is their average price ?
  - What are the amenities often mentioned in this area ?
  - What are the other places mentioned in this neighborhood ?
  - What is the number of sales per year in this neighborhood ?
  - What is the price per square meter in this neighborhood ?

Scenario 3:

- **Name**: Urban Analysis
- **Desc.**: The real estate agents have a good knowledge of the territory and give its description through the advertisements that could be used by geographers or urban planners. They could study the social representations to understand which part of a territory better suits to one type of population than another. Urban planners could also analyze how real estate agents mention the amenities (e.g., transports, schools, shops, etc.) to highlight a lack of services in a neighborhood.
- **Ex. 1**: Alicia is a geographer and would like to study the social representation of the city in order to understand which part of the city is more suitable for a certain type of population (e.g., family, seniors, etc.).
- **Ex. 2**: Clément works for the City Hall of Nice as an urban planner, and would like to better study the access to transports and amenities (e.g., a neighborhood with a lack of transport) to give advice about the future tramway line.
- **CQ** :
  - How do the real estate agents describe the neighborhood (e.g., residential, quiet, etc.) ?
  - What are the amenities mentioned in this neighborhood ?
  - What are the amenities NOT mentioned in this neighborhood ?
  - Are some places always/never mentioned together ?
  - In which part of the city is the tramway often mentioned ?
  - Are the advertisements dealing with the tramway located far from the city center ?
  - Is a place within another place ?

Scenario 4:

- **Name**: Use of Annotated Textual Data
- **Desc.**: The entities automatically retrieved in the texts are a great source of annotated data to help NLP researchers to test models for French language and work with a confidence score.
- **Ex. 1**: Julien is a PhD student in NLP and would like to test his new model to extract spatial entities in French.

- **Ex. 2**: Fabrice would like to add the confidence scores in his search engine about the advertisement to filter the results with highest confidence score.
- **CQ** :
  - What are the advertisements with extracted toponyms having a confidence score higher than 0.8 ?
  - What are the advertisements having all the extracted entities with a confidence score higher than 0.5 ?

2.2 Ontological Formalization

The motivating scenarios and competency questions helped us to formalize and define the classes and relations found in our ontology. As discussed below, we identified three main components of this ontology: the Real Estate, the definition and formalization of a place, and the representation of the uncertain location. Listing 1 gives an example of our proposed representation.

Real Estate:

The Real Estate is the accommodation described in an advertisement through the text and metadata. Thanks to the first two motivating scenarios, we identified three kinds of information to describe a real estate :

1. Type (house, apartment, etc.);
2. Features (price, floor size, floor level, number of rooms, etc.) ;
3. Location (coordinates, city, neighborhood, proximity to the amenities, etc.).

To formalize the real estate information, we used the *GeoSPARQL* and *schema.org* vocabularies. First, the classes *schema:Apartment* and *schema:House* define the type of accommodation. We do not need other accommodations since we only have extracted advertisements describing houses and apartments. Although the properties underlying these classes describe several features (*schema:floorLevel*, *schema:numberOfRooms*, etc.), we created other properties such as *sure:hasPrice*. Furthermore, the real estate is also a spatial object represented as a point (latitude/longitude) or located with respect to other places (i.e., qualitative spatial relations). Thus, we have created the class *sure:RealEstate* that is a sub-class of *geo:Feature*. An instance of *sure:RealEstate* might have a geometry (*sf:Point*) if the coordinates are given, or be linked to other places (*sure:locatedIn*).

2.2.1 Place:

A place is a spatial object that is often described by its name and its feature (e.g., Nice, Masséna Square, etc.). For instance, digital gazetteers mostly refer to a place thanks to its names and its attributes. However, a significant number of places are vernacular and might consist of a spatial relation (e.g., "city center", "the old downtown", "West of Nice", "Nearby the Promenade des Anglais"). In Lesbegueries et al. [Lesbegueries, 2006] and, Syed et al. [Syed, 2022] define two types of places : absolute vs relative places. An absolute place is a named place (e.g., Promenade des Anglais) while a relative place is a place that needs a linguistic or spatial reasoning processes (e.g, "West of Nice", "Downtown Nice"). We proposed to follow this definition by creating two classes, *sure:AbsolutePlace* and *sure:RelativePlace*, to represent a place in the ontology. The first one represents all places (named or not) where the real estate is located *inside* while the latter only describes the place compound of proximity-related relations (e.g., "nearby", "5 kilometers", "10 minutes", etc.). We added two properties to this class to define the spatial relation and its object: *sure:hasSpatialRelation* and *sure:hasAnchor*. We used *rdfs:label* to refer to the name of a place and *rdf:type* to specify its class (e.g., "neighborhood", "street", "school", etc.). In the literature, many vocabularies have been developed to represent geographic features. *GeoNames*[2] uses the *SKOS* concepts to describe upper-level classes. *GeoLinkedData*[3] defines three ontologies depending on the use (administrative, transport, hydrography) and uses existing vocabularies. Finally, the National Institute of Geographic and Forest Information (*IGN*[4]) developed its own ontology using the *BDTOPO* dataset to describe topographic and administrative entities (buildings, road network, green area, etc.). Since our application focuses on a representation of a place according to its use and its perception, while the existing vocabularies describe the spatial entity according to their nature and their topography, their use is limited. Thus, we defined two upper-classes, *sure:LocativeArea* and *sure:Amenity*, to distinguish the amenities from the place of living. Then, we chose to extract and generate classes from the ads despite they might be noisy.

2.2.2 Uncertain Location:

A spatial object is often linked to a geometry to represent its boundary. However, the places extracted in the ads are described according to the view of the real estate agent. Thus, the agent might not have the same limits as the official ones, or exaggerate them in order to sell [McKenzie, 2017b]. The location is vague and cannot be represented as a single point or polygon. A classic way to overcome the vagueness is the use of

[2]http://geonames.org/ontology/ontology_v3.0.rdf
[3]http://geo.linkeddata.es
[4]http://data.ign.fr/def/topo/20190212.htm

fuzzy set theory [Schneider, 1999; Bunel, 2018], which we used to approximate a vague place, as explained in Chapter 4.

In fuzzy set theory, a fuzzy subset A of a set E is defined by a function called membership function μ_A. The function gives the degree of membership to the set A for each element x of E. The degree is often ranged between 0 and 1. If $\mu_A(x) = 1$, then x completely belongs to A while if $\mu_A(x) = 0$, then x does not belong to A. We applied this theory to the places to capture the uncertainty of their location by computing the membership degree for each point in the space. Also, we could retrieve crisp sets using alpha-cuts. An alpha-cut $\tilde{A}_\alpha$ is a crisp subset where each element has a membership degree greater than α.

$$\tilde{A}_\alpha = \Big\{x \in A; \mu_{\tilde{A}}(x) \geq \alpha\Big\}.$$

The core and the support are specific α-cuts where α is respectively equal to 1 and 0 :

$$\begin{aligned} cor(A) &= \Big\{x \in A; \mu_A(x) = 1\Big\}, \\ supp(A) &= \Big\{x \in A; \mu_A(x) > 0\Big\}. \end{aligned}$$

In the ontology, we have represented the geometry of a place as a collection of alpha-cuts. We defined *sure:AlphaCut* as a subclass of *geo:Geometry*, and its property *sure:hasAlpha* to set the membership degree. Then, *GeoSPARQL* allows to associate a collection of geometries to the same object. Thus, a place could have several alpha-cuts in order to represent as reliably as possible its uncertain boundaries.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@prefix : <http://ns.inria.fr/sure#> .

############## Classes and Properties
:RealEstate rdfs:subClassOf geo:Feature.
:Amenity rdfs:subClassOf geo:Feature.
:TrainStation rdfs:subClassOf :Amenity.
:AbsolutePlace rdfs:subClassOf geo:Feature.
:RelativePlace rdfs:subClassOf geo:Feature.
:AlphaCut rdfs:subClassOf geo:Geometry.

:hasAlpha a rdfs:Property ;
    rdfs:domain :AlphaCut ;
    rdfs:range xsd:double .

:hasAnchor a rdfs:Property ;
    rdfs:domain :RelativePlace ;
    rdfs:range geo:Feature.

:hasSpatialRelation a rdfs:Property ;
    rdfs:domain :RelativePlace ;
    rdfs:range xsd:string.

############## Instances
:RealEstate1 a schema:House, :RealEstate ;
    schema:floorSize "200"^^xsd:double;
    schema:hasPrice "975000"^^xsd:double;
    schema:numberOfRooms "6"^^xsd:double;
    sure:locatedIn :Proche_Gare_Riquier ;
    :hasDescription :Text1 .

:Gare_Riquier a :TrainStation, :Place;
  rdfs:label "riquier"@fr.

:Proche_Gare_Riquier a :RelativePlace;
   :hasAnchor :Gare_Riquier;
   :hasSpatialRelation "proche";
   geo:hasGeometry :AlphaCut1, :AlphaCut2.

:AlphaCut1 a :AlphaCut ;:hasAlpha "0.5"^^xsd:double;
     geo:asWKT "MULTIPOLYGON (((43.6957 7.280889,..., 43.69578 7.280882)))"^^geo:wktLiteral.

:AlphaCut2 a :AlphaCut ;:hasAlpha "0.8"^^xsd:double;
     geo:asWKT "MULTIPOLYGON (((43.695 7.2808,..., 43.6955 7.280892)))"^^geo:wktLiteral.
```

Listing 1: Example of the RDF representation of real estate information and vague places.

3 SURE-KG

The ontological formalization provides a structured representation of the information extracted and processed from the advertisements. A knowledge graph is a collection of entities where the relationships between them are modeled by nodes and edges that could be defined by the ontology. In our work, we created SURE-KG, a knowledge graph built from a real dataset at the core of the industrial application of Septeo Proptech[5]. In this section, we present the generation pipeline involving two main steps described in Chapter 3 and Chapter 4 applied to a real dataset. Finally, the resulting RDF dataset and its reusability are described and discussed.

3.1 Dataset

The initial dataset[6] is a corpus gathering real estate advertisements written in French and located in the French Riviera from various online advertisers and provided by SepteoProptech, our industrial partner. Two advertisers (i.e., Bien ici[7] and Leboncoin [8]) provide more than 80% of the advertisements. The first one is a professional advertiser where only real estate agent can publish ads. On the other hand, the second advertiser is a collaborative platform that puts individuals in France in touch with each other when they want to buy or sell. Both individuals and professionals could publish their ads. Hence, the dataset gathers advertisements written by professionals as well as private individuals, which might differ in terms of natural language and structure.

Furthermore, a real estate advertisement is mainly composed of a text describing the property and its location and is surrounded by pictures and metadata such as the price, the floor area, the city and the coordinates (latitude/longitude). In our work, we focused on the text and metadata to extract Real Estate attributes and location information (proximity to amenities, neighborhood, scenic view, etc.). Moreover, we only processed the advertisements promoting houses or apartments for sale, and did not apply our pipeline to rental ads or other types of property (e.g., parking). The dataset is mainly composed of apartments, that make up for more than 80% of the advertisements.

Last but not least, the coordinates linked to each advertisement are not always reliable. Indeed, most of the real estate agents hide the exact position of the property in order to keep the sales agreement exclusive and avoid business competition. For instance, the coordinates often fall in the middle of the city or a neighborhood instead of at the building. Hence, we applied a rule of thumb to keep the coordinates that

[5] https://septeo-proptech.fr/
[6] http://github.com/Wimmics/sure/tree/main/dataset
[7] https://www.bienici.com/
[8] https://www.leboncoin.fr/

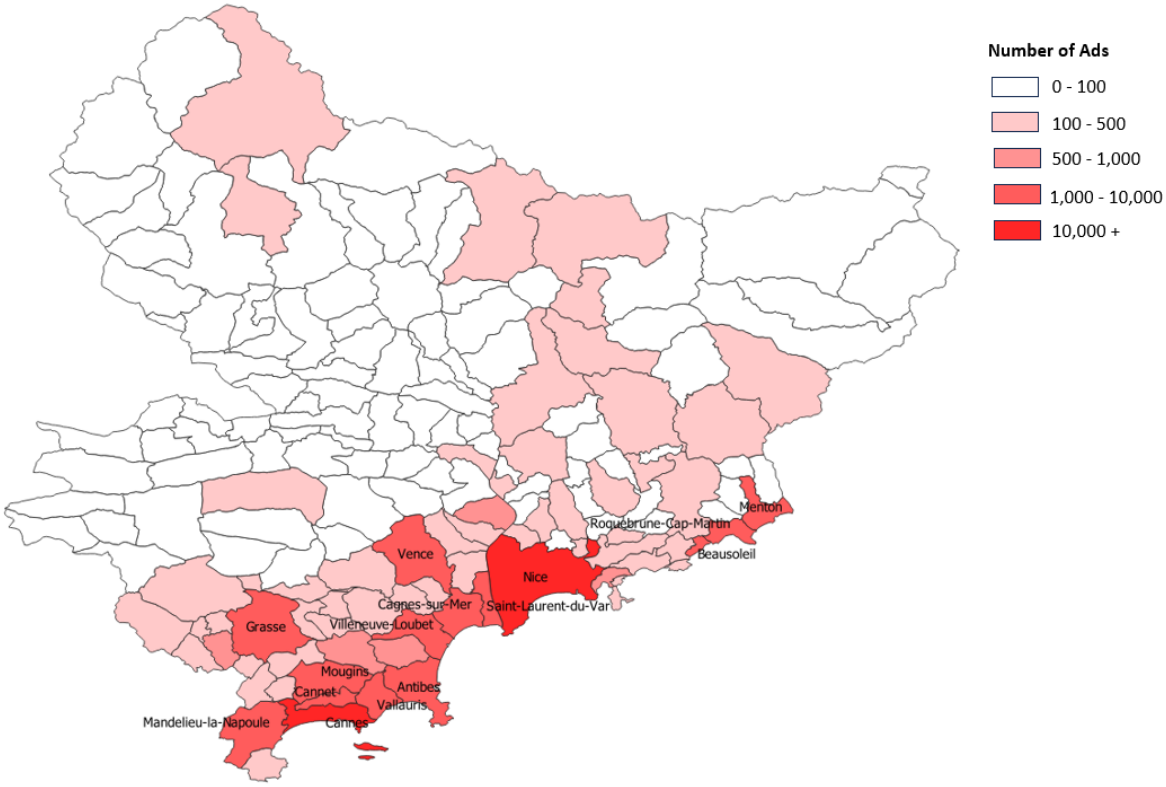

Figure 5.2: Number of ads in the dataset by city

seem reliable: the coordinates should not appear more than 20 times for each city. The remaining advertisements only constitute 10% of the data stored by SepteoProptech.

Overall, we processed 100,000+ ads located in the Alpes-Maritimes, France. Figure 5.2 shows the number of advertisements by city, while Figure 5.3 represents the population by city for the Alpes-Maritimes, France. The ads density is similar to the population density since the number of ads is very high along the coast and declines in the countryside.

3.2 Generation Pipeline

To generate SURE-KG, we applied a pipeline to the dataset, which involves two main steps: (1) process each document of the corpus to extract information and, in particular, spatial information, and (2) estimate the uncertain boundary of each place extracted. Finally, the output of both treatments is translated into a unified and consistent RDF dataset.

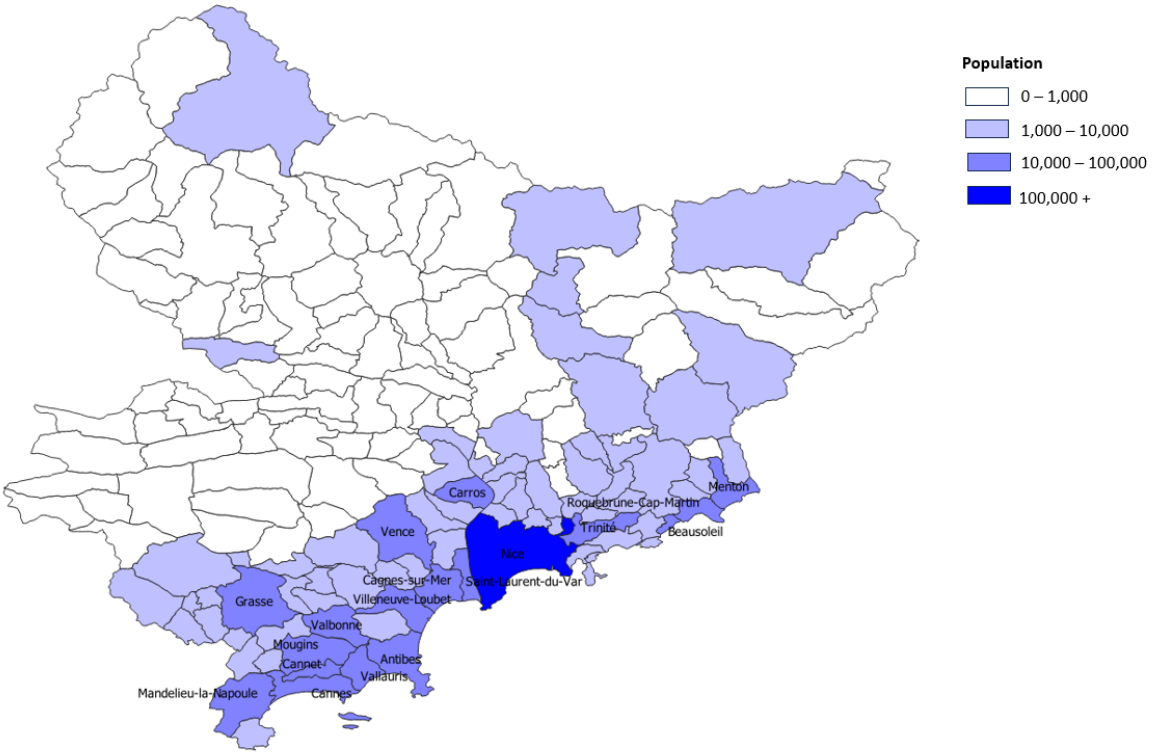

Figure 5.3: Population by city in the Alpes-Maritimes, France

3.2.1 Spatial Information Extraction:

As presented in Chapter 3, we specifically designed a model to extract spatial information in real estate advertisements that we applied to our dataset. This method is a two-stage workflow involving Named Entity Recognition and Relationship Extraction. The Named Entity Recognition model architecture is a *BiLSTM+CRF model* combined with a text embedding. The Relationship Extraction is based on Dependency Parsing methods. Both have been trained from scratch and evaluated on a corpus of French Real Estate advertisements written in French and located in the French Riviera. The NER method detects 4 type of entities to better capture the spatial information in a Real Estate advertisement:

- **Toponym**: entity referring to a place's proper name which is the easiest way to describe a place;
- **Feature**: entity representing the type of a place (e.g., natural features, constructions and subdivisions of land).
- **Spatial Relation entity**: a spatial relation describes the location of an object by specifying its direction with respect to a reference object whose location is known (e.g., "nearby", "5 minutes away");

- **Mode of transportation**: entity referring to the travel mode between two places (e.g., "walk")

Moreover, the model gives a structured knowledge by reconstructing relations such as the spatial relation (i.e., relations between a Spatial Relation entity and a Toponym or Feature) or the nature of a place (i.e., relation between a feature and a Toponym).

After applying the method to our dataset, we post-processed the output to clean data from misspelling, plural and abbreviations. We replaced the well-known abbreviations by its correct terms (e.g., "min" for "minutes", "m" for "meters"). We also applied the Jaro-Winkler distance to retrieve very similar terms and correct misspeling (threshold 0.9 and 0.95 for Toponym because there are more possibilities). Lastly, we retrieved the most obvious hyponyms for the Feature class by applying a simple term inclusion heuristic. Table 5.1 shows general statistics about the spatial information extraction and post-processing output. We can see that the post-processing part significantly reduced the number of unique entities and increased the number of times each entity is mentioned in the advertisements.

Nb of ads processed	102,335
Nb of ads with spatial information extracted ≥ 1	80,200
Median of nb of spatial information extracted per ad	3
Maximum of nb of spatial information extracted per ad	53

Type of entity	Nb of entities before post-processing	Nb of entities after post-processing	Median of the count of each entity before post-processing	Median of the count of each entity after post-processing
Feature	4,212	486	2	35
Toponym	11,084	5,501	2	2
Spatial Relation	491	50	2	113
Mode of Transportation	14	10	8	4

Table 5.1: Statistics about (1) the spatial information extraction and (2) post processing

3.2.2 Uncertain Location Estimation

The second stage of the generation pipeline consists in the estimation of the boundaries of each place. Since we have extracted a significant number of vernacular places that could not be retrieved in official database, and the real estate agents might exaggerate the limit of a place, we decided to create our own knowledge using the advertisements and their reliable coordinates. As stated in Chapter 4, we used the Kernel Density Estimation method, which is a non-parametric estimation method that infers the shape of a variable from a sample, and gives a probability (density) for each point of the support. For each extracted place, we selected all the ads mentioning it, removed outliers and estimated the footprint based on the coordinates. In order to get a reliable

estimation, we only applied the method for each place with a minimum of 10 ads mentioning it. Finally, the method returns a density, which can be easily transformed into a membership function of a fuzzy set.

3.2.3 RDF Generation

The final stage of the pipeline is the translation of both treatments into an RDF model. We created a script[9], available on our Github repository, using the python library RDFLib[10]. We defined a template of triples about the real estate and spatial information. We also linked the cities mentionned in the metadata of the real estate ads to Geonames using the library Geocoder[11] and Geonames' attributes (e.g., feature class). Among 281 unique cities, we found 182 Geonames related entities that we linked with the predicate *owl:sameAs*. Finally, we added the named entity annotations given by the NER model to the knowledge graph to propose a reusable dataset. We chose the Web Annotation Data Model[12] vocabulary to annotate the text as presented in Listing 2). The annotation points to the annotated ad, the text position (the target), and the named entity category (the body). We also give a confidence score of the extraction (*sure:confidence*) provided by the Named Entity Recognition model.

3.3 Knowledge graph and resulting linked dataset

The resulting SURE-KG dataset is an RDF graph that provides an RDF representation of Real Estate ads and the spatial information automatically derived from the textual data and metadata. It contains more than 7M triples, 100K Real Estate ads and 6K places. We paid a thorough attention to adopt the FAIR principles [Wilkinson, 2016] to ensure reproducibility, allow other researchers to compare our results and methods, and explore our dataset.

Dataset Description. In line with best practices [Lóscio, 2018], the dataset comes with a thorough self-description, comprising licensing, authorship and provenance information, used vocabularies, interlinking and access information, described with Dublin Core Metadata Information, DCAT and VOID.

Dataset Accessibility. The dataset is made available by means of a DOI identified RDF dump downloadable from Zenodo, and a public SPARQL endpoint. A Github repository provides a comprehensive documentation, source codes and query templates. The ontology has been published following the standards and best practices of the linked open data. This information is summarized in Table 5.2.

[9] https://github.com/Wimmics/sure/tree/main/src/GraphGeneration
[10] http://rdflib.readthedocs.io
[11] http://geocoder.readthedocs.io/
[12] https://www.w3.org/TR/annotation-model/

```
@prefix dc: <http://purl.org/dc/elements/1.1> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <https://ns.inria.fr/sure#> .

############# Instances
:RealEstate1 :hasDescription :Text1

:Text1 dc:description "Biot : dans domaine ferme a proximite du village [...].";
   dc:language "fr" .

:Entity0 a oa:Annotation ;
   oa:hasBody :Toponym ;
   oa:hasTarget [oa:hasSource sure:Text1],
       [a oa:TextPositionSelector ;
          oa:end "0"^^xsd:double ;
          oa:start "0"^^xsd:double] ;
   oa:motivatedBy oa:classifying ;
   :confidence "0.99"^^xsd:double.

:Entity1 a oa:Annotation ;
   oa:hasBody :SpatialRelation ;
   oa:hasTarget [oa:hasSource sure:Text1],
       [a oa:TextPositionSelector ;
          oa:end "6"^^xsd:double ;
          oa:start "6"^^xsd:double ];
   oa:motivatedBy oa:classifying ;
   :confidence "0.99"^^xsd:double.
```

Listing 2: Example of RDF representation of annotations of Toponym and Spatial Relation.

Reproducibility In compliance with the open science principles, all the scripts and files involved in the pipeline are provided in the project's Github repository under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, so that anyone may rerun the whole processing pipeline.

Dataset Licensing The SURE-KG RDF Knowledge Graph is published under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[13]. In particular, this license allows anyone to use the dataset for a non-commercial purpose since the data come from an industrial project.

Sustainability Plan. In the short term, we plan to apply the pipeline to all the regions of France. This will be the opportunity to assess the quality of the method on other data and areas. In the middle and long term, we intend to improve the pipeline (e.g., linking spatial information to other gazetteers) and to fit it to other language (e.g., English). Furthermore, we have deployed a server to host the SPARQL endpoint that benefits from a high availability infrastructure and 24/7 support.

[13] https://creativecommons.org/licenses/by-nc-sa/4.0/

Dataset DOI	10.5281/zenodo.7885757
Downloadable RDF dump	https://doi.org/10.5281/zenodo.7885757
Public SPARQL endpoint	http://erebe-vm2.i3s.unice.fr:5000/sparql/
Source Code and Documentation	http://github.com/Wimmics/sure
URIs Namespace	http://ns.inria.fr/sure#
Dataset URI	http://ns.inria.fr/sure/data/
Citation	Lucie CADOREL, Fabien GANDON, Andrea G. B. TETTAMANZI, 2023. SURE-KG dataset. https://doi.org/10.5281/zenodo.7885757

Table 5.2: Dataset availability.

3.4 Potential Impact and Reusability

Altough the SURE-KG dataset is firstly designed for the real estate domain and our industrial partner Septeo Proptech, we highlight its potential impact and reusability beyond these two aspects.

Current Use. The processing pipeline and the dataset are used in the context of the industrial need of the company to retrieve the location of each advertisement, as described in Chapter 4, in order to better analyse the market. Moreover, Scenario 1 is one of the use cases that the company faces on to provide a tool to the real estate agents. On the other hand, ongoing research works are conducted by geographers from the University Côte d'Azur about the social representation of the city of Nice according to the real estate agents [Blanchi, 2022].

Interest of communities in using the dataset. Beyond the real estate domain, the dataset could be used by the Semantic Web community in works and experimentation with uncertain and spatial data. For instance, GeoSPARQL implements topological relations (e.g., union, intersection, overlaps, etc.) for crisp geometries, but it could be extended to fuzzy geometries [Schockaert, 2011]. Furthermore, the GIS community could be interested in this dataset to enrich gazetteers with vernacular and cognitive places, or to study qualitative spatial relations, such as *near*, according to their context [Aflaki, 2022].

Potential for reuse. To the best of our knowledge, the SURE-KG dataset is the first one integrating uncertain and vague spatial data in a knowledge graph, which could serve for benchmarking spatial algorithms. We provide a documentation on the Github repository including a demonstration notebook to help users to query the graph. For a potential wider application, the processing pipeline do not require any adaptation for the extraction of spatial information from French texts, but might need

small adjustments to extract metadata (e.g., price, floor size, etc.). For the adaptation to the other languages, a more substantial work may be required to train a model on the data.

4 Statistics and Usage of the Dataset

To assess the quality and usability of the dataset, we analyse and discuss some statistics describing the classes and properties. We also study different use cases highlighted by the motivating scenarios described in Section 2.1 and several examples of queries.

4.1 Statistics

The SURE-KG dataset comprises 7,497,370 triples, describing 102,335 real estate properties and 6,033 places. A real estate property is, on average, linked to 3 places extracted from its advertisement. The maximum number of places linked to a single real estate property is 34, while 29,011 real estate properties are not linked to any places. The type of place linked to the real estate property is either *sure:AbsolutePlace* or *sure:RelativePlace*, and both types are almost equally represented: respectively 56,212 real estate properties have at least a link to an instance of *sure:AbsolutePlace*, and 55,629 to an instance of *sure:RelativePlace*. The difference between *sure:AbsolutePlace* and *sure:RelativePlace* is the presence of a proximity-related relation in the latter. The number of instances of *sure:AbsolutePlace* is slightly higher because an instance of *sure:RelativePlace* is always composed of a spatial relation and an instance of *sure:AbsolutePlace*, which is automatically created if it does not already exist. Moreover, the instances of *sure:AbsolutePlace* are either an instance of *sure:LocativeArea* or *sure:Amenity*. Table 5.3 shows that only a few instances belong to *sure:Amenity*. Indeed, the subclasses of *sure:Amenity* and *sure:LocativeArea* have been automatically generated from the extracted spatial features. However, these classes are very noisy and often misclassified as subclass of *sure:LocativeArea*. We retrieved 30 subclasses of *sure:Amenity* while 456 classes have been classified as a subclass of *sure:LocativeArea*. For instance, features such as high school (*Lycée*), bakery (*Boulangerie*), or restaurant (*Restaurant*) should be amenities but they have been classified as subclass of *sure:LocativeArea* in our dataset. A postprocessing step should be added to clean these errors in the ontology. Last but not least, we recorded 20,530 instances of *sure:AlphaCut* which represent the boundaries of each place. A place has a collection of 5 different alpha-cuts to capture its uncertain boundaries. However, among the 6,033 places, we managed to compute and record a reliable estimation of the geometry for 4,106 places. The places without geometry mainly belong to the class *sure:AbsolutePlace*. Indeed, an instance of *sure:AbsolutePlace* is always created when an instance of *sure:RelativePlace*

is added and if its anchor does not already exist. Nevertheless, the geometry for these instances are not computed since any advertisements mention them as an instance of *sure:AbsolutePlace*. These places are often points of interest (e.g., train station, beach, university) and, their geometry may be found in gazetteers.

Class URI	nb of instances
http://ns.inria.fr/sure#RealEstate	102,335
http://ns.inria.fr/sure#AbsolutePlace	3,760
http://ns.inria.fr/sure#RelativePlace	2,273
http://ns.inria.fr/sure#AlphaCut	20,530
http://ns.inria.fr/sure#LocativeArea	1,904
http://ns.inria.fr/sure#Amenity	77
http://ns.inria.fr/sure#Quartier	193

Property URI	nb of instances
http://www.w3.org/2000/01/rdf-schema#label	1,736
http://www.opengis.net/ont/geosparql#hasGeometry	122,865
http://ns.inria.fr/sure#hasAnchor	2,273
http://ns.inria.fr/sure#hasSpatialRelation	2,273
http://ns.inria.fr/sure#hasDescription	81,911
http://ns.inria.fr/sure#locatedIn	330,654

Table 5.3: Selected statistics on typical properties and classes.

4.2 Usage of the Dataset and Examples of Queries

The dataset should be able to answer the different competency questions defined by the motivating scenarios described in Section 2.1. For each scenario, we queried the dataset to verify each competency question. Furthermore, we implemented some potential applications to demonstrate the usability of the dataset.

4.3 Query the Competency Questions

Scenario 1: Real Estate Property Search

In this scenario, we would like to help a buyer to find a property according to predefined criteria (e.g., price, location, number of rooms, etc.). We identified 5 competency questions of which 4 could be easily answered with a SPARQL query. The competency question '*What are the nearby amenities and places of interest (POI) ?*' can be partially answered with a SPARQL query since we extracted the amenities and spatial relations in each advertisement. On the other hand, GeoSPARQL functions could also be used to compute the distance between a real estate property and a particular point

of interest. Finally, external resources (e.g. coordinates of schools, restaurants, etc.) could be added to fully answer this competency question.

Scenario 2 : Real Estate Market Analysis

The market analysis scenario is a very important application because it could help the real estate professionals in their business. It comprises 7 competency questions. Only one competency question ('*Are they similar to the one for sale ?*') requires an additional AI method to find a proper solution (i.e., a method to compute similarity between real estate properties).

Scenario 3 : Urban Analysis

This scenario, used by geographers or urban planners, focuses on the analysis of the territory through the knowledge of the real estate agents. Among the seven competency questions, three of them require further analysis and two of them need the use of GeoSPARQL functions (e.g., distance, within). Nevertheless, the user does not need external information to answer the questions.

Scenario 4: Use of Annotated Textual Data

The last scenario proposes to use the automatically annotated textual data as a training dataset to test models for French language and spatial entities. Simple SPARQL queries easily answer the competency questions.

4.4 Illustration of Potential Applications

Application 1: Find your home !

This application showcases the use of a the dataset to retrieve the best real estate properties according to a potential buyer's criteria. Figure 5.4 shows an example of a buyer who is looking for a 2-room apartment in the Vieux Nice district, with a budget limit of 500,000 €. This search is easily answered by a SPARQL query which retrieves all the real estate properties corresponding to the criteria. The map shows the location of all the results. To dig deeper this example, a potential improvement would be a real estate search engine.

Application 2: Inspect your neighborhood 🔎

The second application aims at providing a deeper knowledge about the market to the real estate agents. To illustrate a possible market analysis, we computed the price per square meter for each neighborhood in the city of Nice, France (see Appendix D for Cannes and Antibes). The name of the neighborhoods might not be exactly the same as the official ones since we used the one defined by the advertisements. Figure

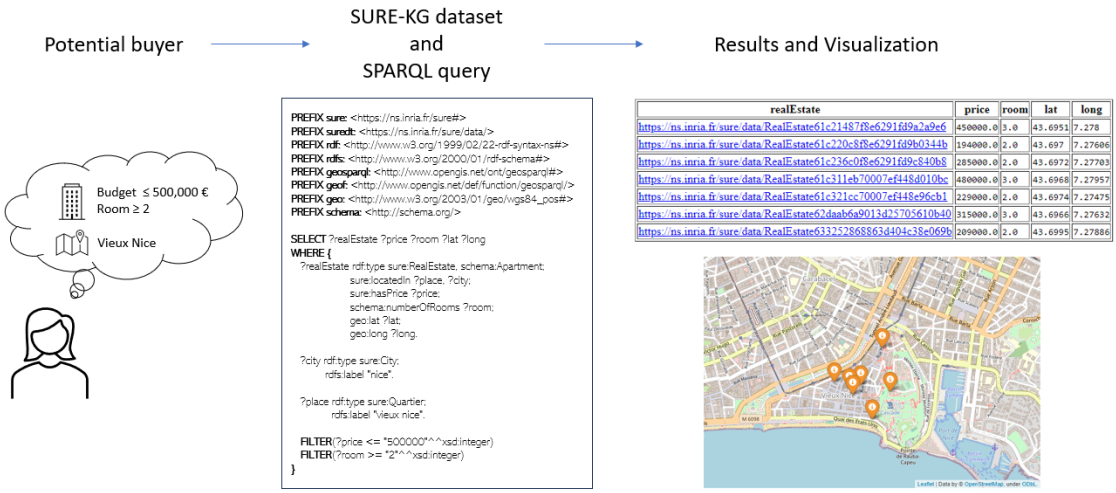

Figure 5.4: Example of the SPARQL query of a potential buyer looking for an apartment in the Vieux Nice district, and its results.

5.5 shows the SPARQL query to compute the price per square meter and the centroid of the 5 most expensive (in red) and the 5 cheapest (in blue) neighborhoods. We can notice that the most expensive neighborhoods are very close to the city center and the sea, while the cheapest are far from the centre, on the hills or near the airport. Moreover, Table 5.4 shows that there is a very important gap between the most expensive neighborhood (Carré d'Or) and the cheapest one (Ariane). Then, we studied the most frequent attributes used by the real estate agents to describe these two groups of neighborhoods. Figures 5.6 and 5.7 represent the two word clouds of the attributes. First, the word cloud of the most expensive neighborhoods contains more words. Also, the vocabulary used by the agents deals with luxury (e.g., *prévilegié* — privileged, *prestigeux* — prestigious), reputation (e.g., renommé — renowned, *réputé* — reputed) and trend (e.g., *branché* — trendy, *prisé* — popular, *convoité* — sought after). On the other hand, the second word cloud only comprises 5 words and evokes the history of the neighborhoods (e.g., *historique* — historical, *nouveau* — new) and the atmosphere (e.g., *dynamique* — vibrant, *vivant* — animate , *résidentiel* — residential). Finally, we studied the most frequent spatial features mentioned with the neighborhoods. Figure 5.8 shows that the agents mainly mention the downtown (*Centre Ville*), the bay (*Baie)* and the public transports (*Transports en Commun*) with the 5 most expensive neighborhoods. Indeed, we previously identified these neighborhoods as close to the city centre and the sea, and these two elements are promoted by the agents. Regarding the cheapest neighborhoods, the spotlighted features refer to services: convenience stores (*Commerces de proximité*), supermarket (*Supermarché*), drugstore (*Pharmacie*), tramway (*Ligne de Tramway*). As these neighborhoods are far from the centre, the

```
PREFIX sure: <https://ns.inria.fr/sure#>
PREFIX suredt: <https://ns.inria.fr/sure/data/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX schema: <http://schema.org/>

SELECT ?place ROUND(AVG(?priceSize)) AS ?priceSqm  WHERE {
    ?realEstate rdf:type sure:RealEstate;
                sure:hasPrice ?price;
                schema:floorSize ?floorSize;
                sure:locatedIn ?city. ?place.

    ?city rdf:type sure:City;
          rdfs:label "nice".

    ?place rdf:type sure:AbsolutePlace, sure:Quartier;
           rdfs:label ?label.

    BIND(if(?floorSize >"0"^^xsd:integer, ?price/?floorSize, "0"^^xsd:integer) AS ?priceSize )
    FILTER(?priceSize > "0"^^xsd:integer)
}
GROUP BY ?place
ORDER BY DESC(?priceSqm)
```

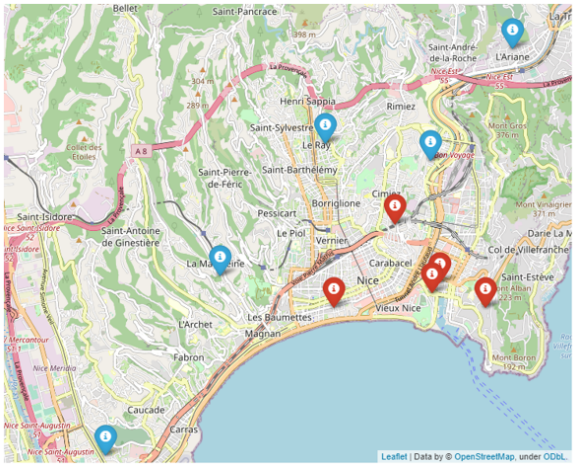

Figure 5.5: SPARQL query to compute price per square meter and the centroid of the 5 most expensive (in red) and the 5 cheapest (in blue) neighborhoods.

real estate agents advance the proximity to the amenities. In a nutshell, the dataset and the SPARQL queries allow to study a neighborhood in different ways (i.e., price, attributes, amenities, etc.) which could be used by the real estate professionals to promote it.

Name	**Price per Square Meter (€)**
Carré d'Or	8,327
Mont Boron	7,983
Antiquaires	7,747
Port	7,444
Cimiez	7,278
Madeleine	3,921
Ray	3,844
Pasteur	3,767
Saint-Agustin	3,315
Ariane	1,986

Table 5.4: Price per Square Meter of the 5 most expensive and the 5 cheapest neighborhoods in Nice, France.

Figure 5.6: Word cloud of the attributes of the 5 most expensive neighborhoods.

Figure 5.7: Word cloud of the attributes of the 5 cheapest neighborhoods.

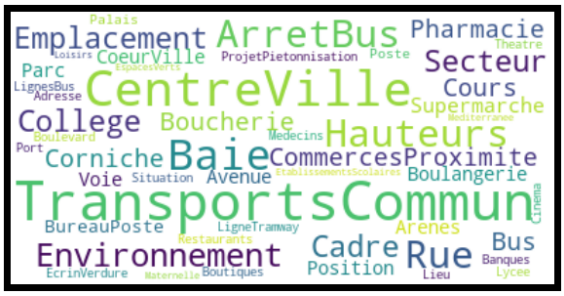

Figure 5.8: Word cloud of the features mentioned with the 5 most expensive neighborhoods.

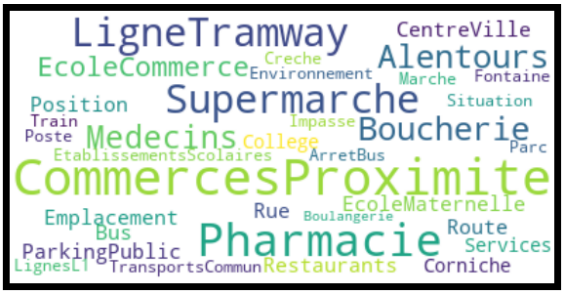

Figure 5.9: Word cloud of the features features mentioned with the 5 cheapest neighborhoods.

5 Summary and Perspectives

In this chapter, we focused on the last contribution of this thesis: the representation, storage and reasoning over the produced knowledge. We presented the SURE-KG knowledge graph built from a real dataset at the core of the industrial application of SepteoProptech and containing 7,497,370 triples. We designed an ontology to represent real estate and uncertain spatial data which follows the standards and best practices of the linked open data. The design choices have been motivated and formalized thanks to several scenarios and competency questions. This ontology allows to define a real estate object, an imprecise place and the uncertain boundaries. We also developed a pipeline to turn the raw data to a RDF knowledge graph, and made several services available to exploit SURE-KG. Finally, we assessed the quality and usability of the dataset by analysing statistics, querying competency questions and presenting several use cases. The competency questions have been widely answered by only querying the dataset despite some of them require additional AI methods. Regarding the use cases, they have been derived from the motivating scenarios and have shown the potential applications and analysis in the real estate domain. Particularly, we proposed a study of the real estate market in Nice by computing the average price per square meter in

each neighborhood thanks to a SPARQL query. We also analysed the most frequent attributes and features mentioned in the five most expensive and the five cheapest neighborhoods. In a nutshell, we showed that the dataset and the SPARQL queries allow to study a neighborhood in different ways (i.e., price, attributes, amenities, etc.).

Several directions could be considered to expand this work. First, we only have applied our pipeline to advertisements located in the Alpes-Maritimes. We would like to apply it to all the regions of France. It will give us the opportunity to evaluate the method's ability to adapt to new data. Moreover, covering the whole country will provide a good knowledge of the Real Estate market for the real estate agents. Nevertheless, it also raises the questions of the completeness and the generality of the SURE ontology designed for our application. Indeed, we automatically generated classes from the advertisements to describe the type of place (e.g., Train Station, School, Beach) but it only covers the geographic type used in the French Riviera. For instance, some features exist in specific cities or regions, such as the Réseau Express Régional (RER) in Paris (i.e., a hybrid train/underground serving Paris and its suburbs). Moreover, the definition of a term might be different from a city or region to another. For instance, the term *Tramway* in Antibes corresponds to a *Bus* having its own lane, which is totally different from the general definition. Therefore, the question of a global or local enrichment of the ontology has to be studied. A more global ontology would help to extend our work to different territories but it would also remove the specificity of each region and thus, it would be less accurate. Furthermore, another limitation of the ontology is the errors introduced by the automatic extraction of the classes. We showed in Section 4 that some of these classes have been classified as subclass of *sure:LocativeArea* instead of *sure:Amenity*. These errors must be corrected in order to get a more reliable and usable ontology.

Lastly, one of the main goals of the Semantic Web is to interrelate datasets on the Web. In this work, we only linked cities to the ones described in GeoNames. As a perspective, matching the place-names extracted from the advertisements to the same places stored in other datasets (e.g., GeoNames, DBpedia) can enrich our KG as well as the other gazetteers. It can also be used used to compare official boundaries to the cognitive ones estimated with our approach. Furthermore, some entities in our KG refer to the same spatial object but are not linked. Therefore, a matching must be performed to reduce the number of entities referring to the same object.

Chapter 6

Conclusion

Contents

1 Summary of Contributions

The main challenge of this thesis was to automatically retrieve the imprecise location of the real estate advertisements from the texts. This thesis involves connection between three main fields of research: Natural Language Processing, Geographic Information Science and Semantic Web. Particularly, this interdisciplinary work gave the opportunity to collaborate with geographers from the research group ESPACE and to develop tools that could be exploited by different communities and domains. Moreover, this work has been conducted in collaboration with an industrial partner, SepteoProptech, which allowed to strengthen the links between all the actors of the innovation and to improve the communication between researchers and engineers from different backgrounds. In this thesis, we faced several challenges to automatically geocode the location of each real estate advertisement, and we divided the problem into three tasks: (1) information extraction, (2) geocoding places and real estate advertisements and (3) building a knowledge graph to store and share the data.

In Chapter 3, we described the first contribution of this thesis, i.e. a method to automatically extract and represent the spatial information from the real estate advertisement which describes the location of the property. We first analyzed the language used in the real estate advertisements such as the typical vocabulary, errors or the structure since it is different from traditional text in NLP (e.g., tweets, scientific papers, news). We detailed five challenging specificities, with few examples, that have to be tackled before performing a textual analysis. We also identified the different ways to

describe a place in an advertisement to show that standard NLP models might not be suitable for this task. Then, we presented our method based on a Named Entity Recognition to detect and classify entities in text into predefined categories. We described the guidelines used to annotate spatial information in the advertisements and to create a dataset to train our model. The model is based on a BiLSTM-CRF architecture combined with three text representations to capture the language specificities: a basic word embedding architecture, a context-based and character-level language model and a transformer language model. As the output of a NER model is not structured, we also proposed to extract relations between entities. We identified four types of relationship that have been retrieved thanks to the analysis of the grammatical structure. Finally, we performed an evaluation of the pipeline and we achieved good results for the NER model as well as the RE approach.

The second contribution of this thesis, described in Chapter 4, is the geocoding of each advertisement from the vague spatial information extracted in Chapter 3. We first showed that the spatial objects described in the texts are imprecise because of vernacular places and spatial relations. We studied the completeness of two famous gazetteers (OpenStreetMap and GeoNames) regarding the toponyms extracted in the real estate advertisements, and noticed that a large number of toponyms are not easily retrieved in the gazetteers. Particularly, GeoNames has a very small coverage, compared to OpenStreetMap, due to its worldwide target audience. Moreover, we compared the use of the preposition *near* in the real estate advertisements according to the type of objects and the territory, and showed that to model a spatial relation to geocode a place, it also depends on the context. Then, we described the method to approximate the location of each advertisement. First, we learnt a density estimation for each places extracted in the advertisements to delimit their boundaries, and we transformed the density into a fuzzy set to represent the imprecision on a map and to easily combine the vague information. We presented several examples to show the proficiency of the method to estimate boundaries. The last part focused on the combination of the information to retrieve the approximate location of each advertisement, which can be seen as an aggregation of multiple data sources. We described the chosen modelization and the operator used to aggregate the different sources. Finally, we defined the metrics used to measure the quality of the fuzzy locations and we evaluated this method on the three biggest cities of the French Riviera: Nice, Cannes and Antibes. We showed that the OWA operator is the more suitable for our work and the method is robust since it has similar results for the three cities.

In Chapter 5, we focused on the last contribution of this thesis: the representation, storage and reasoning over the produced knowledge. We presented the SURE-KG knowledge graph built from a real dataset at the core of the industrial application of SepteoProptech and containing 7,497,370 triples. We designed an ontology to represent

real estate and uncertain spatial data which follows the standards and best practices of the linked open data. The design choices have been motivated and formalized thanks to several scenarios and competency questions. This ontology allows to define a real estate object, an imprecise place and the uncertain boundaries. We also developed a pipeline to turn the raw data to a RDF knowledge graph, and made several services available to exploit SURE-KG. Finally, we assessed the quality and usability of the dataset by analysing statistics, querying the competency questions and presenting several uses cases. The competency questions have been widely answered by only querying the dataset despite some of them require additional AI methods. Regarding the use cases, they have been derived from the motivating scenarios and have shown the potential applications and analysis in the real estate domain. Particularly, we proposed a study of the real estate market in Nice by computing the mean price per square meter in each neighborhood thanks to a SPARQL query. We also analysed the most frequent attributes and features mentioned in the five most expensive and the five cheapest neighborhoods. In a nutshell, we showed that the dataset and the SPARQL queries allow to study a neighborhood in different ways (i.e., price, attributes, amenities, etc.).

2 Future Research Perspectives

This section details the research possibilities deriving from the contributions proposed in this thesis. Although we have shown the feasibility of our proposal on real data, there are still some limits and some possible improvements for further works. In each chapter, we presented *minor* improvements to develop a better version of our approach in short-term. Hereafter, we present and discuss more general perspectives for extending this thesis, that require a more substantial amount of work.

Information Extraction and Other Types of Text and Language

A first perspective is the extension of our first contribution to other types of text and language. In this work, we only focused on the real estate advertisements written in French, while the geographic and spatial information might be found in other unstructured data, such as travel blogs, social media or historical documents. Although the real estate professionals use a particular vocabulary, humans similarly describe a place, as explained in Chapter 2. For instance, they often use a toponym (e.g., the name of a city, neighborhood or POI) or a spatial relation to locate an object by specifying its direction with respect to a reference object whose location is known. As a first extension, we could apply our Information Extraction workflow to similar texts (e.g., holiday rentals such as Airbnb descriptions) before deploying to very different types of text, such as itineraries [Moncla, 2014]. Applying our approach to other type of texts

would allows us to evaluate the robustness of the model. Also, we could fine-tune our model with few labeled data from other types of text in order to get better results.

Furthermore, we developed a model for the French language while the NLP resources are limited. On the contrary, the English language have been largely studied and many Geoparsers have been developed. As a future work, we could design an English version of the workflow in order to compare our Geoparser to others. However, we need to create a new labeled dataset and train our model with new text representations adapted to English. Hence, the deployment to a new language is difficult and time-consuming. Multilingual models have been developed to bridge language gaps and allow to analyze data from multiple language sources. In particular, Massively Multilingual Language Models (MMLMs) have shown their ability to to generalize across languages, even for languages with limited training data [Wu, 2019]. Thus, a research possibility to improve our contribution is to use a MMLM and data from different language sources to create a model adapted to several languages. Nevertheless, there are still some challenges with these models, such as difficulties with low-resource languages or the presence of language biases in the training data, that are still ongoing research works.

Uncertainty and Ranking Information

In Chapter 4, we proposed a method to estimate the boundaries of imprecise spatial information. However, we did not take into account the uncertainty which can also arise from the advertisements. Uncertainty results from ignorance and describes the degree of knowledge required to decide if a statement is true or false. In this work, we assumed that the information extracted from an advertisement is reliable and always true. However, uncertainty can arise because of several reasons: the real estate agents can lie about the location by mentioning well-reputed places, or our model extracting information can miss some toponyms, features or spatial relations. Thus, the question of how to quantify and represent the uncertainty remains open. Moreover, we noticed that some places are often mentioned in the advertisements, such as famous POIs, but they are not precise enough to locate a property. For instance, retrieving the name of the street in the text is more interesting than retrieving the name of the most famous POI of the city. Hence, we should rank the information according to their scarcity among the corpus in order to give more weight to precise information. Also, giving a weight to each information allow to choose a method or another: if the information is very precise (e.g., a street name) we could use a Geocoder, while if the information is vague we could use our approach described in Chapter 4. Finally, as a future work, we could imagine learning a weight for each information with a neural network by minimizing the distance between the coordinates of the advertisement and the estimated boundaries in order to rank the information and quantify the uncertainty.

Modelling Spatial Prepositions

In this thesis, we showed that modelling qualitative spatial relations is difficult and depends on the context, the spatial object and the territory. We proposed an approach to take the spatial relations into account to estimate boundaries of vague spatial description, which could be applied to any context with a sufficient amount of data. However, our method defines a place as a combination of a toponym and/or a feature, and/or a spatial relation, and estimates the boundaries of the combination. Therefore, the boundaries highly depend on the data (i.e., the real estate advertisements) and they might differ according to the application. Also, we did not design the spatial relations (e.g., *near*) as a quantitative model (e.g., acceptance model) to avoid creating a lot of models for each spatial preposition and context, but we can not estimate a place if we lack data. For instance, if we want to know where is *near* a certain place-name, we need to have enough advertisements mentioning the spatial relation with this place-name. Both approaches have limitations and modelling a spatial preposition remains a challenging research question. As a perspective, we could conduct a detailed study about the spatial prepositions, as presented in Chapter 4, and extend this work to other spatial prepositions according to the size of the spatial object, the territory and the type of text, in order to better understand how spatial relations are used by humans.

Spatial Objects Changing Over Time

Our last perspective deals with the temporality of the spatial objects and its representation in the Knowledge Graph. In this thesis, we treated real estate advertisements published between 2020 and 2022, and we extracted the limits of places according to their description over this period. However, a spatial object can change over time, in particular its name or its boundaries. For instance, our model detected the amenity *future tramway line* in NIce, which is now available. The entity corresponding to the new line is the same as the entity describing *future tramway line* but its name has evolved. On the other hand, a city or a neighborhood can keep the same name over years but having its limits reduced or expanded. Therefore, as a future work, it is necessary to find a representation of the different versions of the spatial objects according to their temporality, in order to apply the pipeline to new advertisements over the time. A first exploration of several ontologies developed to tackle the problem of information, in particular territories, changing over time in a KG [Kauppinen, 2007; Welty, 2006; Claramunt, 1995; Charles, 2023], may help to address this limitation.

Appendix A

List of Publications

The following papers were published as part of this thesis:

- Lucie Cadorel et al. “Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text”. *K-CAP 2021 - International Conference on Knowledge Capture*. Virtual Event USA, United States: ACM, 2021, pp. 41–48
- Lucie Cadorel et al. “Fuzzy representation of vague spatial descriptions in real estate advertisements”. *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, LocalRec 2022*. Seattle Washington, United States: ACM, 2022, pp. 1–4
- Lucie Cadorel et al. “Towards a representation of uncertain geospatial information in knowledge graphs”. *GKG 2022 - Proceedings of the 1th ACM SIGSPATIAL International Workshop on Geospatial Knowledge Graphs*. Ed. by ACM. Seattle, Washington, United States, 2022, pp. 1–2
- Lucie Cadorel et al. “Connaissances géospatiales dans les annonces immobilières : détection et extraction d’information spatiale à partir du texte”. *IC 2022 - Journées francophones d’Ingénierie des Connaissances (dans le cadre de PFIA 2022)*. Ed. by Fatiha Saïs. IC, Journées francophones d’Ingénierie des Connaissances, PFIA 2022. AfIA. Saint-Étienne, France: AfIA, 2022, pp. 20–21
- Lucie Cadorel et al. “Graphes de Connaissances et Ontologie pour la Représentation de Données Immobilières Issues d’Annonces en Texte Libre”. *IC 2023 - 34es Journées francophones d’Ingénierie des Connaissances @ Plate-Forme Intelligence Artificielle (PFIA 2023)*. IC2023 : 34es Journées francophones d’Ingénierie des Connaissances. Strasbourg, France, 2023

- Lucie Cadorel et al. “A Comparative Study of Text Representations for French Real-Estate Classified Advertisements Information Extraction”. *Proceedings of 1st Workshop on AI-driven heterogeneous data management: Completing, merging, handling inconsistencies and query-answering, co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*. Rhodes, Greece, 2023

- Alicia Blanchi et al. “Studying urban space from textual data: Toward a methodological protocol to extract geographic knowledge from real estate ads.” *Computational Science and Its Applications – ICCSA 2022 Workshops. Proceedings Part II*. ed. by O. Gervasi et al. Vol. 13378. Lecture Notes in Computer Science. Springer, 2022, pp. 520–537

- Lucie Cadorel et al. “Mining RDF Data of COVID-19 Scientific Literature for Interesting Association Rules”. *WI-IAT’20 - IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT’20), 14-17 December 2020, a fully virtual conference. Melbourne, Australia, 2020

- Aline Menin et al. “ARViz: Interactive Visualization of Association Rules for RDF Data Exploration”. *IV 2021 - 25th International Conference Information Visualisation*. Vol. 25. 2021 25th International Conference Information Visualisation (IV). Melbourne / Virtual, Australia, 2021, pp. 13–20

Appendix B

Text Annotation Tool: Doccano

This appendix shows some examples of the text annotation tools, called *doccano*, used in Chapter 3 to label our dataset.

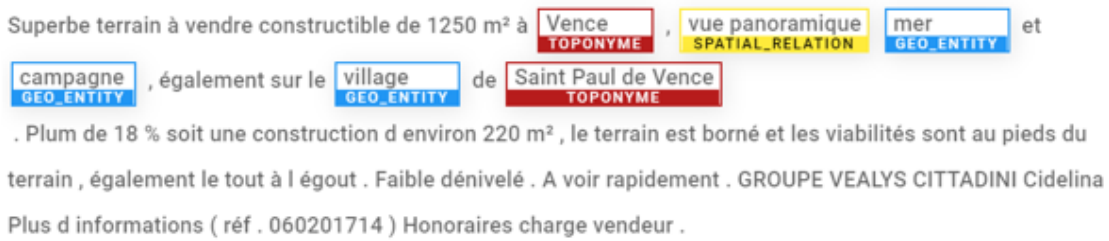

Figure B.1: Example of Text Annotation with *doccano*

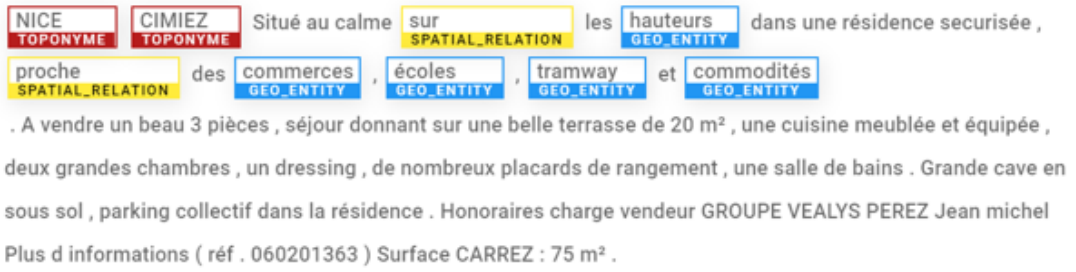

Figure B.2: Example of Text Annotation with *doccano*

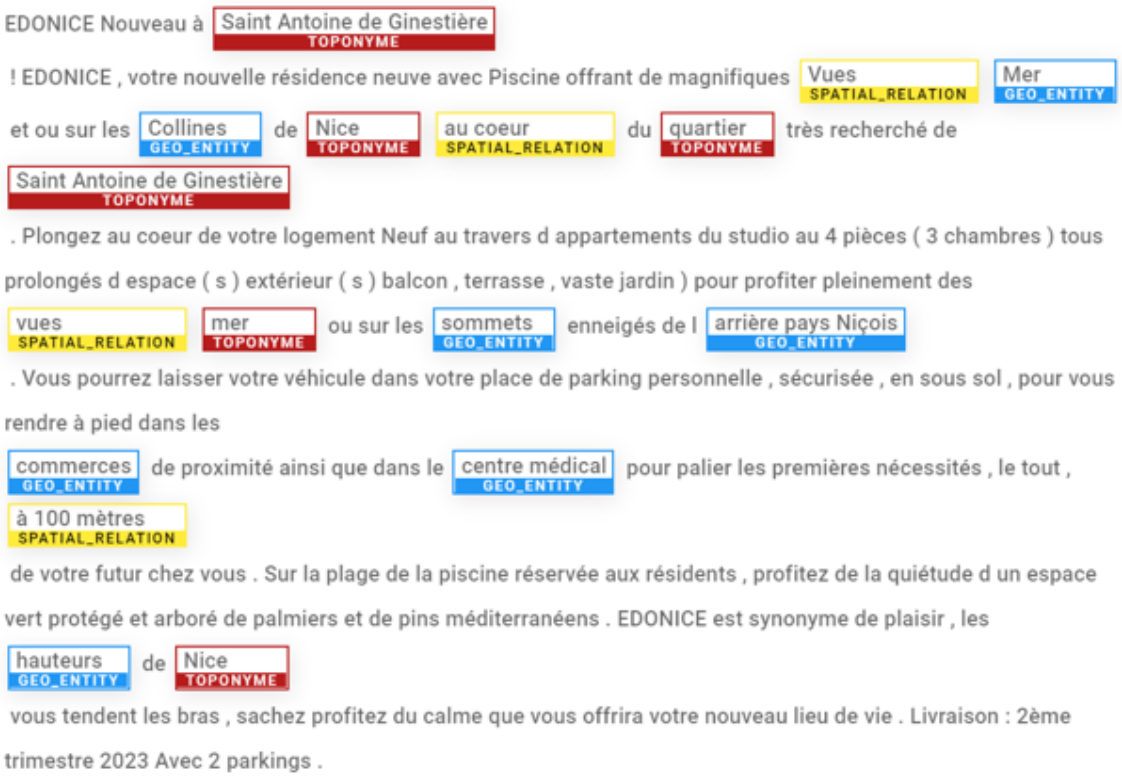

Figure B.3: Example of Text Annotation with *doccano*

Figure B.4: Example of Text Annotation with *doccano*

Appendix C

Fuzzy Representation of Places

In this appendix, we present the fuzzy representation of different places or neighborhoods in Nice, Cannes and Antibes. Figures C.1, C.2, C.3, C.4, C.5, C.6 show famous places and neighborhoods, such as the old town in Antibes (i.e., Vieux Antibes) and Cannes (i.e., Suquet). Figure C.7 is a representation of '*Access to the A8 highway*' and shows the two main highway entrances in Nice. Figures C.8, C.9, C.10 shows the estimation of the three tramway lines in Nice. Regarding Figure C.11, which is the map of the tramway lines given by the city of Nice, the estimations are quite accurate. Finally, Figures C.12, C.13, C.14 represent the vague location '*Near the beach*' in the three cities. As expected, the three representations are very close to sea, although the boundaries are very large.

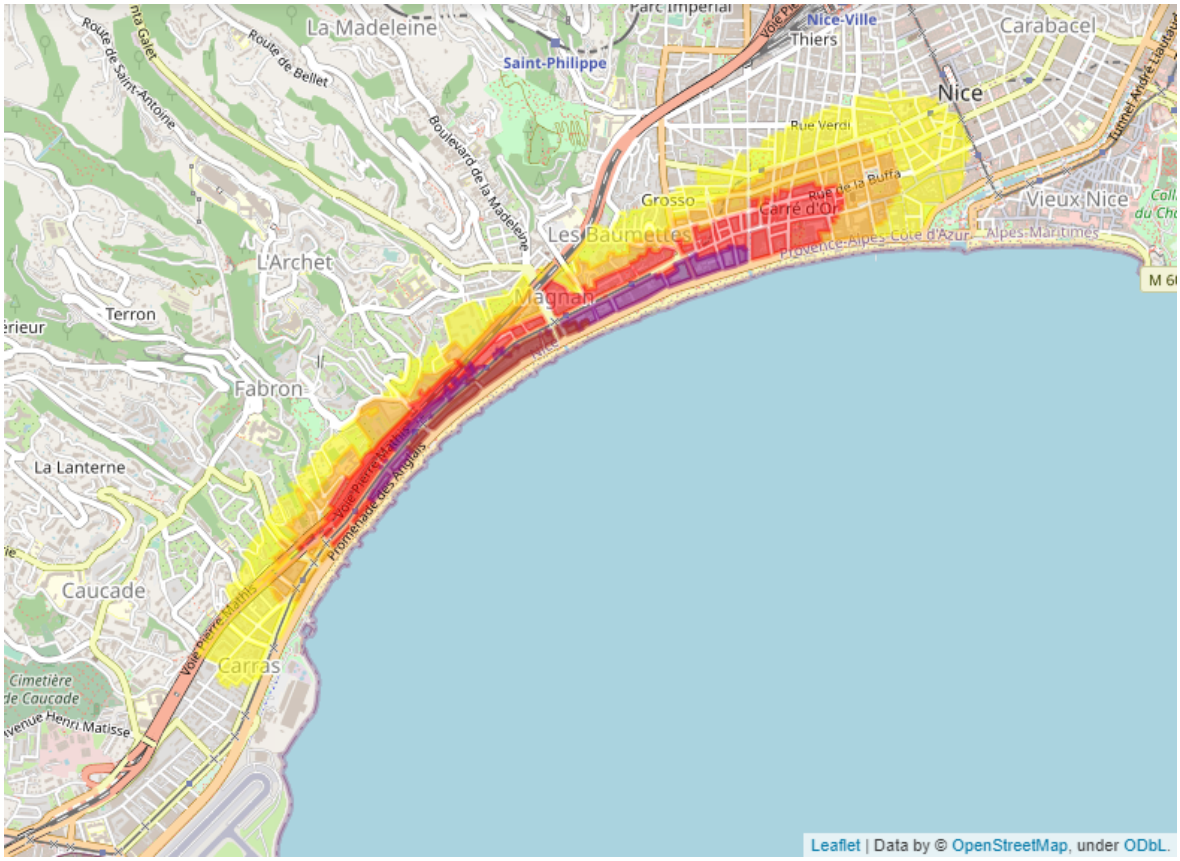

Figure C.1: Promenade des Anglais, Nice.

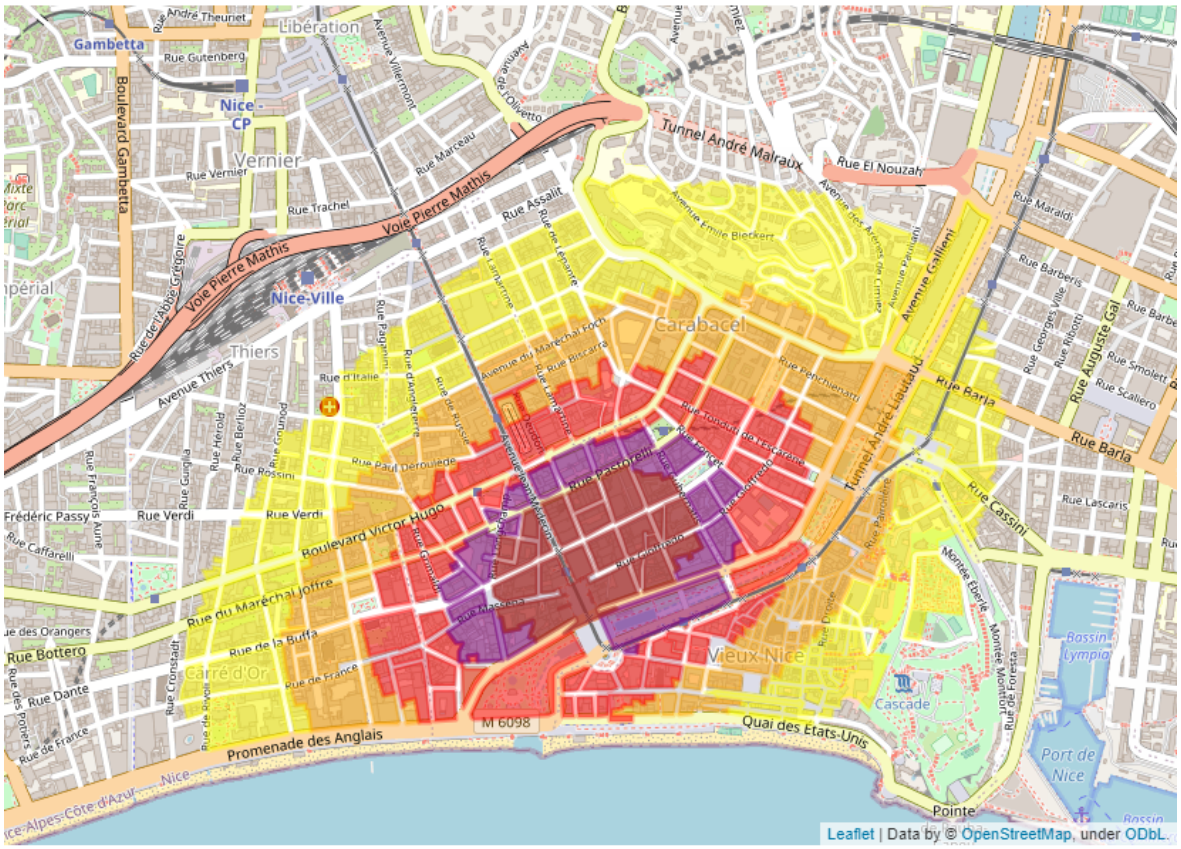

Figure C.2: Place Masséna, Nice.

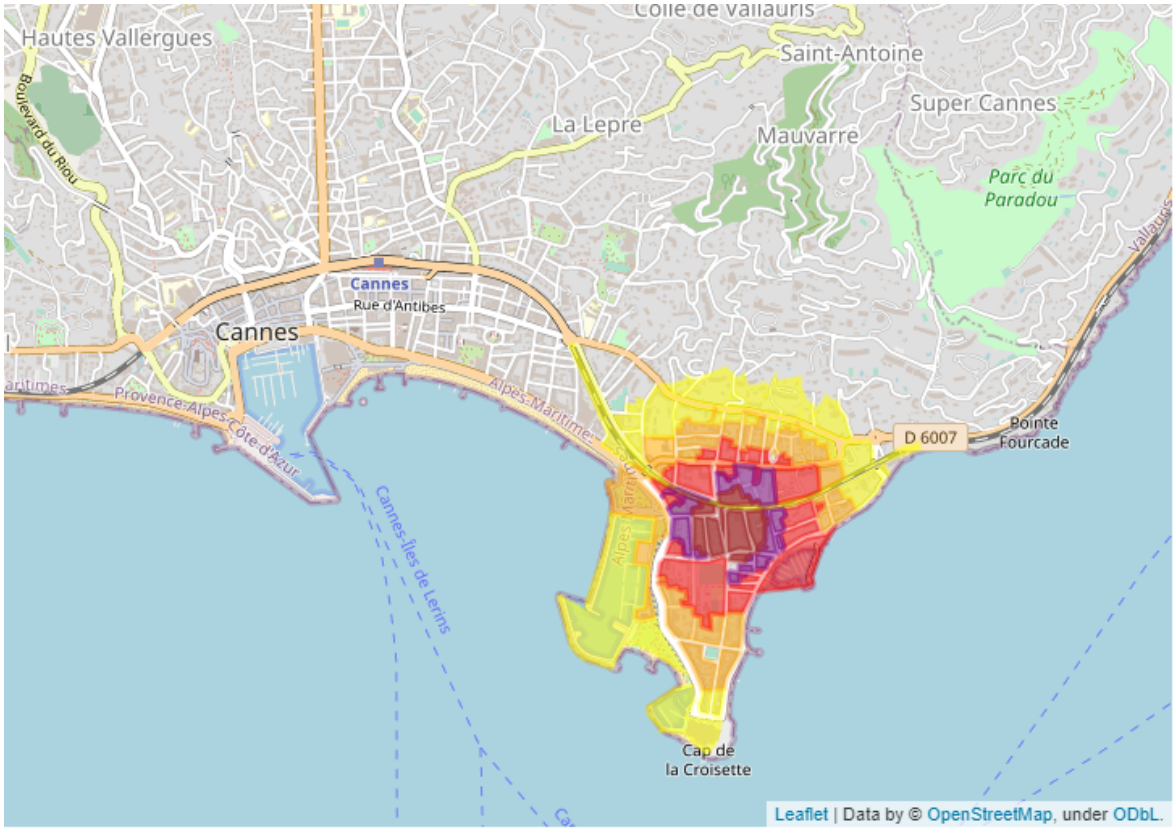

Figure C.3: Palm Beach, Cannes.

Figure C.4: Suquet, Cannes.

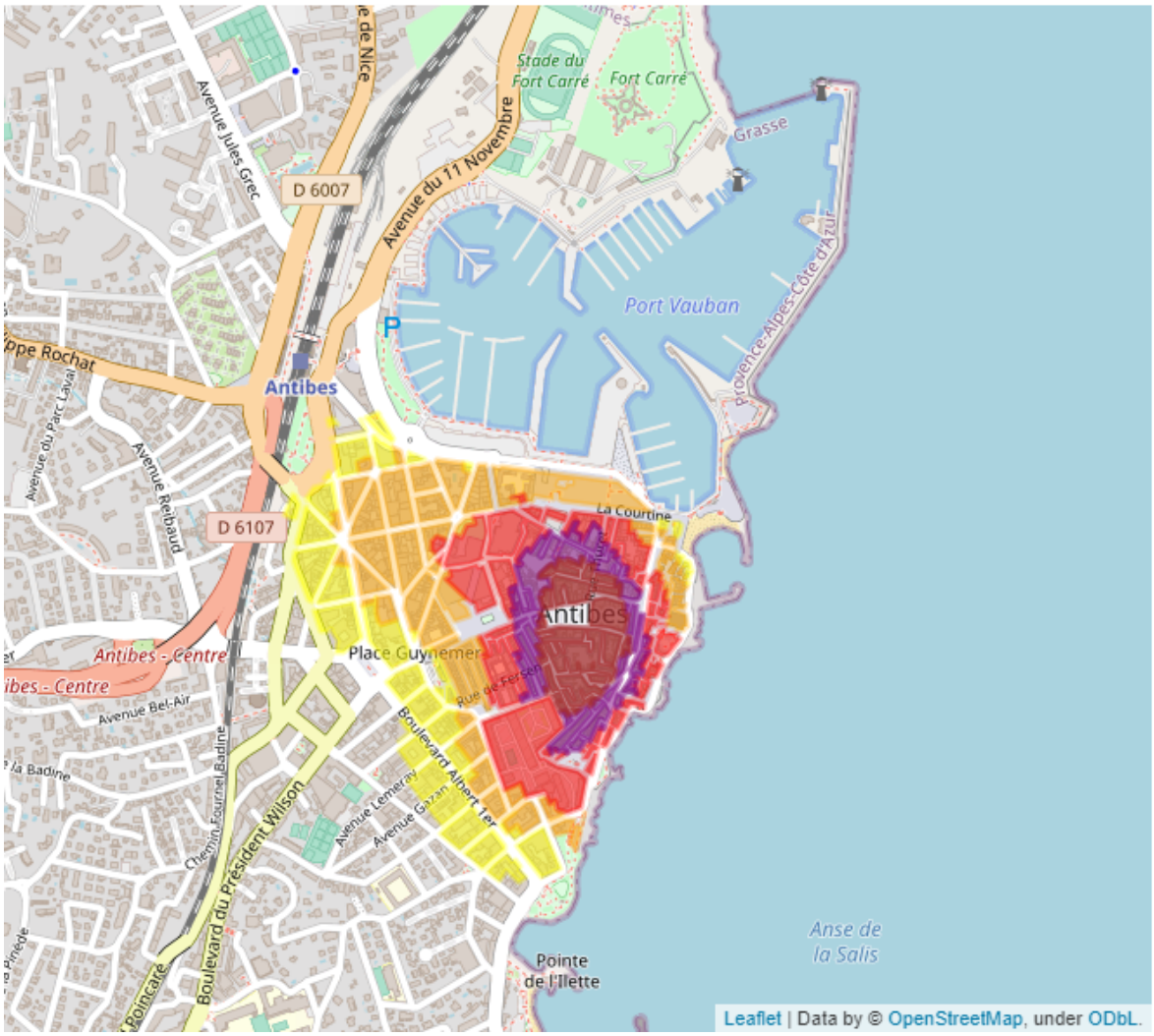

Figure C.5: Vieille Ville, Antibes.

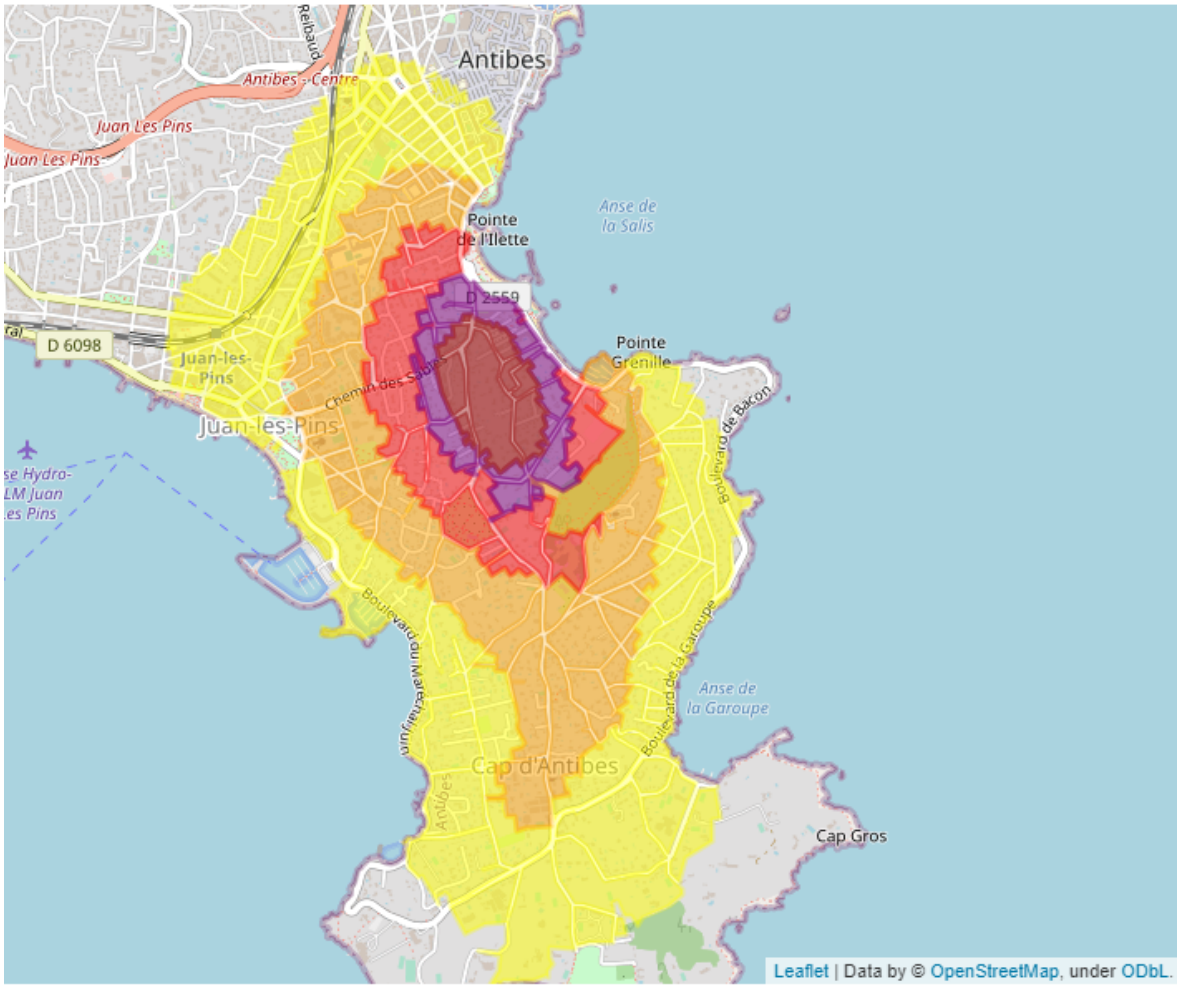

Figure C.6: Cap d'Antibes, Antibes.

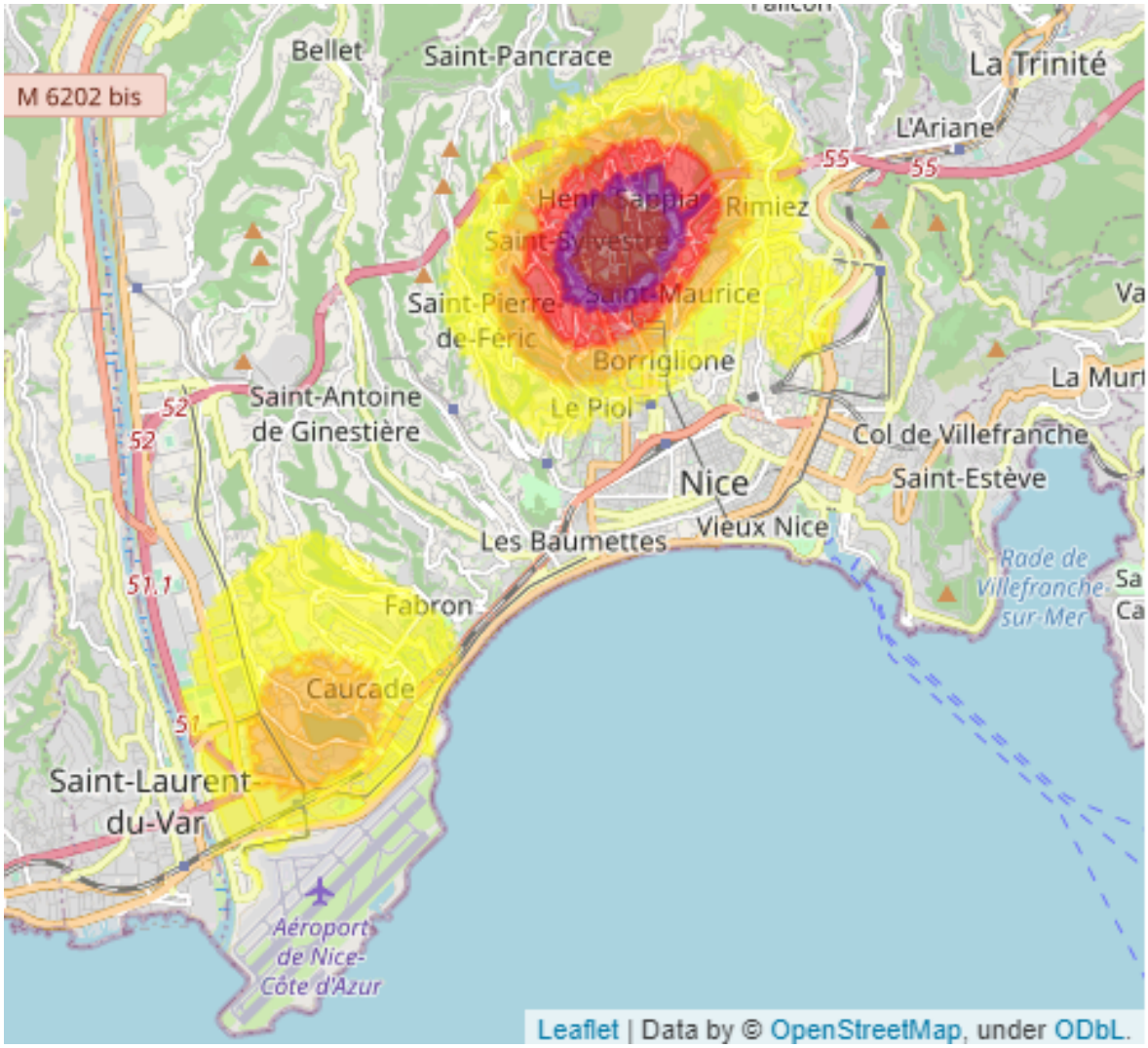

Figure C.7: Access to the A8 Highway, Nice.

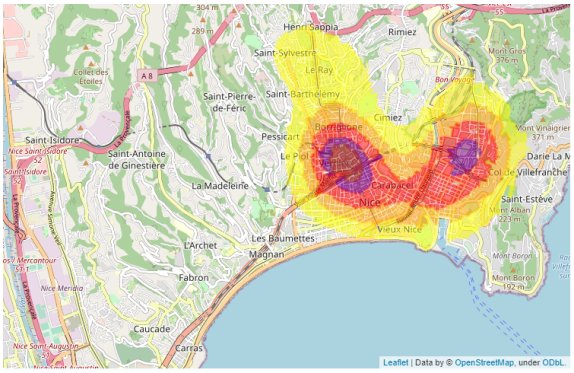

Figure C.8: Tramway Line 1, Nice.

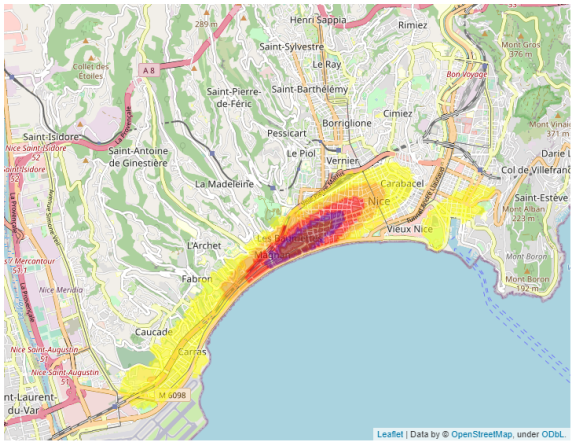

Figure C.9: Tramway Line 2, Nice

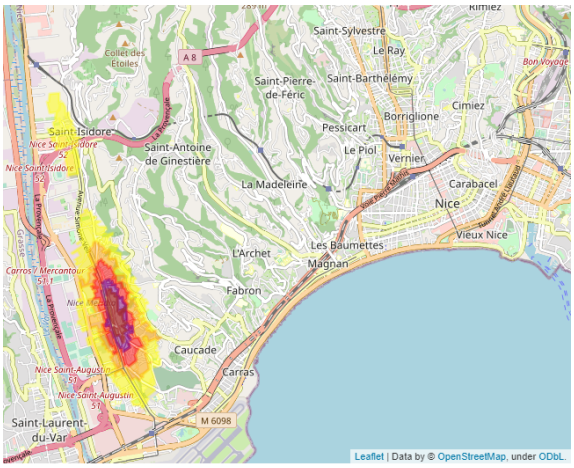

Figure C.10: Tramway Line 3, Nice

Figure C.11: Map of the tramway lines in Nice (https://projets-transports.nicecotedazur.org/)

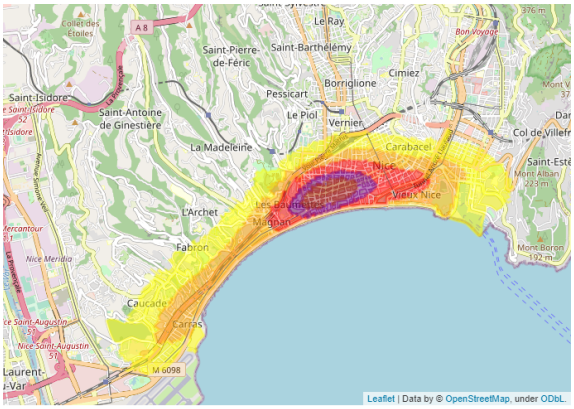

Figure C.12: Near the beach, Nice.

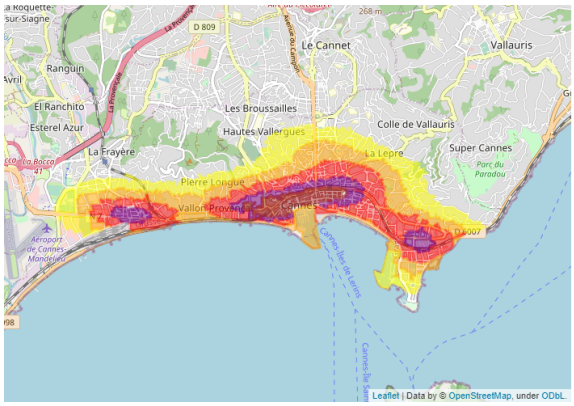

Figure C.13: Near the beach, Cannes

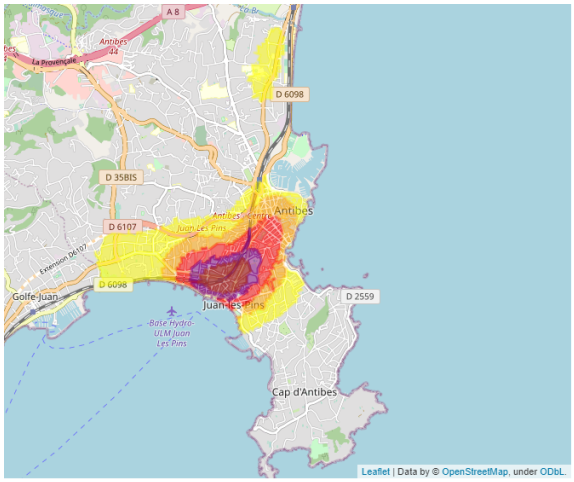

Figure C.14: Near the Beach, Antibes

Appendix D

Real Estate Market Analysis

This appendix shows the computed prices per square meter for the five most expensive and five cheapest neighborhoods in the cities of Cannes and Antibes, as well as word clouds of the most frequent attributes and features, as presented in Chapter 5 for the city of Nice. We can notice that the word clouds are similar to those found for the city of Nice. For instance, the most expensive neighborhoods are often mentioned with the downtown and the sea, while the cheapest ones are mentioned with features referring to services. Moreover, the most expensive neighborhoods have more attributes than the cheapest ones.

Name	**Price per Square Meter (€)**
Basse Californie	11,060
Palm Beach	9,519
Banane	9,032
Californie	8,541
Montfleury	7,377
Anglais	5,009
Carnot	4,788
Brousailles	4,704
République	4,495
Gallieni	4,437

Table D.1: Price per Square Meter of the 5 most expensive and the 5 cheapest neighborhoods in Cannes.

Name	Price per Square Meter (€)
Ilette	8,957
Safranier	8,207
Salis	7,790
Rostagne	6,266
Puy	6,095
Badine	5,337
Fontmerle	5,306
Fontonne	4,538
Croix Rouge	4,339
Semboules	3,926

Table D.2: Price per Square Meter of the 5 most expensive and the 5 cheapest neighborhoods in Antibes.

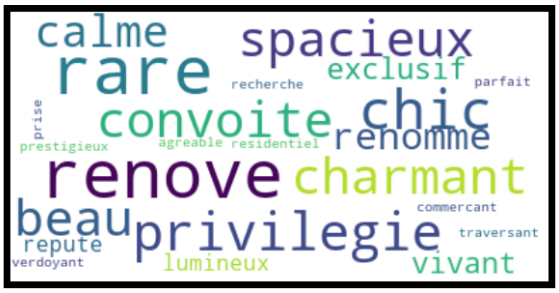

Figure D.1: Word cloud of the attributes of the 5 most expensive neighborhoods in Cannes.

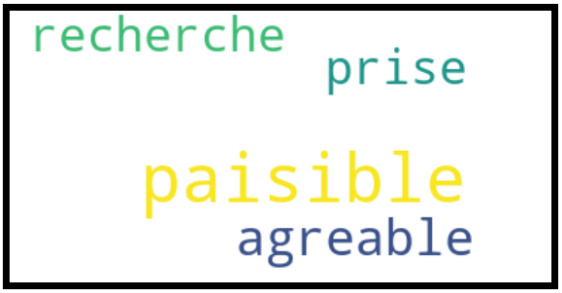

Figure D.2: Word cloud of the attributes of the 5 cheapest neighborhoods in Cannes.

Figure D.3: Word cloud of the features mentioned with the 5 most expensive neighborhoods in Cannes.

Figure D.4: Word cloud of the features features mentioned with the 5 cheapest neighborhoods in Cannes.

Figure D.5: Word cloud of the attributes of the 5 most expensive neighborhoods in Antibes.

Figure D.6: Word cloud of the attributes of the 5 cheapest neighborhoods in Antibes.

Figure D.7: Word cloud of the features mentioned with the 5 most expensive neighborhoods in Antibes.

Figure D.8: Word cloud of the features features mentioned with the 5 cheapest neighborhoods in Antibes.

Bibliography

[Abdallah, 2017] Sherief Abdallah and Deena Abu Khashan. "Using text mining to analyze real estate classifieds". *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2*. Springer. 2017, pp. 193–202 (cit. on p. 13).

[Abeillé, 2003] Anne Abeillé, Lionel Clément, and François Toussenel. "Building a treebank for French". *Treebanks: Building and using parsed corpora* (2003), pp. 165–187 (cit. on p. 18).

[Adams, 2012] Benjamin Adams and Krzysztof Janowicz. "On the geo-indicativeness of non-georeferenced text". *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 6. 1. 2012, pp. 375–378 (cit. on pp. 2, 24, 41, 44).

[Aflaki, 2022] Niloofar Aflaki, Kristin Stock, Christopher B Jones, Hans Guesgen, Jeremy Morley, and Yukio Fukuzawa. "What Do You Mean You're in Trafalgar Square? Comparing Distance Thresholds for Geospatial Prepositions". *15th International Conference on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2022 (cit. on pp. 2, 31, 70, 107).

[Akbik, 2018] Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual string embeddings for sequence labeling". *Proceedings of the 27th international conference on computational linguistics*. 2018, pp. 1638–1649 (cit. on pp. 23, 50).

[Alex, 2019] Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander. "Geoparsing historical and contemporary literary text set in the City of Edinburgh". *Lang. Resour. Evaluation* 53.4 (2019), pp. 651–675 (cit. on pp. 24, 25).

[Altman, 1994] David Altman. "Fuzzy set theoretic approaches for handling imprecision in spatial analysis". *International journal of geographical information systems* 8.3 (1994), pp. 271–289 (cit. on p. 31).

[Aone, 1998] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. "SRA: Description of the IE2 system used for MUC-7". *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998 (cit. on p. 20).

[Atemezing, 2012] Ghislain Auguste Atemezing and Raphaël Troncy. "Comparing Vocabularies for Representing Geographical Features and Their Geometry." 2012 (cit. on p. 36).

[Auer, 2009] Sören Auer, Jens Lehmann, and Sebastian Hellmann. "Linkedgeodata: Adding a spatial dimension to the web of data". *The Semantic Web-ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings 8*. Springer. 2009, pp. 731–746 (cit. on p. 36).

[Ballatore, 2016] Andrea Ballatore. "Prolegomena for an ontology of place". *Advancing geographic information science* (2016), pp. 91–103 (cit. on p. 36).

[Barrière, 2019] Valentin Barrière and Amaury Fouret. "May I Check Again? - A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts". *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa 2019, Turku, Finland, September 30 - October 2, 2019*. Ed. by Mareike Hartmann and Barbara Plank. Linköping University Electronic Press, 2019, pp. 327–332 (cit. on pp. 18, 50).

[Battle, 2012] Robert Battle and Dave Kolas. "Enabling the geospatial Semantic Web with Parliament and GeoSPARQL". *Semantic Web* 3 (2012), pp. 355–370 (cit. on p. 37).

[Baum, 2017] Andrew Baum. "PropTech 3.0: the future of real estate". 2017 (cit. on p. 2).

[Bekoulis, 2018] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. "An attentive neural architecture for joint segmentation and parsing and its application to real estate ads". *Expert Syst. Appl.* 102 (2018), pp. 100–112 (cit. on pp. 14, 18).

[Benefield, 2011] Justin D Benefield, Christopher L Cain, and Ken H Johnson. "On the relationship between property price, time-on-market, and photo depictions in a multiple listing service". *The Journal of Real Estate Finance and Economics* 43 (2011), pp. 401–422 (cit. on p. 13).

[Bennett, 2007] Brandon Bennett and Pragya Agarwal. "Semantic categories underlying the meaning of 'place'". *Spatial Information Theory: 8th International Conference, COSIT 2007, Melbourne, Australiia, September 19-23, 2007. Proceedings 8*. Springer. 2007, pp. 78–95 (cit. on p. 15).

[Berners-Lee, 2006] Tim Berners-Lee. "Linked data-design issues". *http://www. w3. org/DesignIssues/LinkedData. html* (2006) (cit. on p. 34).

[Bikel, 1997] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. "Nymble: a High-Performance Learning Name-finder". *Fifth Conference on Applied Natural Language Processing*. Washington, DC, USA: Association for Computational Linguistics, 1997, pp. 194–201 (cit. on p. 20).

[Bittner, 2002] Thomas Bittner and John G Stell. "Vagueness and rough location". *Geoinformatica* 6 (2002), pp. 99–121 (cit. on p. 31).

[Blanchi, 2022] Alicia Blanchi, Giovanni Fusco, Karine Emsellem, and Lucie Cadorel. "Studying urban space from textual data: Toward a meth-odological protocol to extract geographic knowledge from real estate ads." *Computational Science and Its Applications – ICCSA 2022 Workshops. Proceedings Part II*. Ed. by O. Gervasi, B. Murgante, S. Misra, A.M.A.C. Rocha, and C. Garau. Vol. 13378. Lecture Notes in Computer Science. Springer, 2022, pp. 520–537 (cit. on pp. 107, 121).

[Bosvieux, 2018] Jean Bosvieux. *L'immobilier, poids lourd de l'économie*. Vol. 49. 1. 2018, pp. 10–14 (cit. on p. 2).

[Brickley, 2014] Dan Brickley, Ramanathan V Guha, and Brian McBride. "RDF Schema 1.1". *W3C recommendation* 25 (2014), pp. 2004–2014 (cit. on p. 32).

[Brown, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. "Language models are few-shot learners". *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 23).

[Bunel, 2018] Mattia Bunel, Ana-Maria Olteanu-Raimond, and Cécile Duchêne. "Référencement spatial indirect: Modélisation à base de relations et d'objets spatiaux vagues". *Sageo 2018*. 2018 (cit. on p. 99).

[Bunescu, 2005] Razvan Bunescu and Raymond Mooney. "A Shortest Path Dependency Kernel for Relation Extraction". *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 724–731 (cit. on pp. 23, 53, 59).

[Buscaldi, 2011] Davide Buscaldi. "Approaches to disambiguating toponyms". *Sigspatial Special* 3.2 (2011), pp. 16–19 (cit. on p. 25).

[Buscaldi, 2008] Davide Buscaldi and Paulo Rosso. "A conceptual density-based approach for the disambiguation of toponyms". *International Journal of Geographical Information Science* 22.3 (2008), pp. 301–313 (cit. on pp. 25, 69).

[Cadorel, 2021] Lucie Cadorel, Alicia Blanchi, and Andrea G. B. Tettamanzi. "Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text". *K-CAP 2021 - International Conference on Knowledge Capture*. Virtual Event USA, United States: ACM, 2021, pp. 41–48 (cit. on p. 120).

[Cadorel, 2022a] Lucie Cadorel, Alicia Blanchi, and Andrea G. B. Tettamanzi. "Connaissances géospatiales dans les annonces immobilières : détection et extraction d'information spatiale à partir du texte". *IC 2022 - Journées francophones d'Ingénierie des Connaissances (dans le cadre de PFIA 2022)*. Ed. by Fatiha Saïs. IC, Journées francophones d'Ingénierie des Connaissances, PFIA 2022. AfIA. Saint-Étienne, France: AfIA, 2022, pp. 20–21 (cit. on p. 120).

[Cadorel, 2022b] Lucie Cadorel, Denis Overal, and Andrea G. B. Tettamanzi. "Fuzzy representation of vague spatial descriptions in real estate advertisements". *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, LocalRec 2022*. Seattle Washington, United States: ACM, 2022, pp. 1–4 (cit. on p. 120).

[Cadorel, 2020] Lucie Cadorel and Andrea G. B. Tettamanzi. "Mining RDF Data of COVID-19 Scientific Literature for Interesting Association Rules". *WI-IAT'20 - IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20), 14-17 December 2020, a fully virtual conference. Melbourne, Australia, 2020 (cit. on p. 121).

[Cadorel, 2023a] Lucie Cadorel and Andrea G. B. Tettamanzi. "A Comparative Study of Text Representations for French Real-Estate Classified Advertisements Information Extraction". *Proceedings of 1st Workshop on AI-driven heterogeneous data management: Completing, merging, handling inconsistencies and query-answering, co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*. Rhodes, Greece, 2023 (cit. on p. 121).

[Cadorel, 2022c] Lucie Cadorel, Andrea G. B. Tettamanzi, and Fabien Gandon. "Towards a representation of uncertain geospatial information in knowledge graphs". *GKG 2022 - Proceedings of the 1th ACM SIGSPATIAL International Workshop on Geospatial Knowledge Graphs*. Ed. by ACM. Seattle, Washington, United States, 2022, pp. 1–2 (cit. on p. 120).

[Cadorel, 2023b] Lucie Cadorel, Andrea G. B. Tettamanzi, and Fabien Gandon. "Graphes de Connaissances et Ontologie pour la Représentation de Données Immobilières Issues d'Annonces en Texte Libre". *IC 2023 - 34es Journées francophones d'Ingénierie des Connaissances @ Plate-Forme Intelligence Artificielle (PFIA 2023)*. IC2023 : 34es Journées francophones d'Ingénierie des Connaissances. Strasbourg, France, 2023 (cit. on p. 120).

[Carlson, 2005] Laura A Carlson and Eric S Covey. "How far is near? Inferring distance from spatial descriptions". *Language and Cognitive Processes* 20.5 (2005), pp. 617–631 (cit. on pp. 2, 30, 70).

[Carlson-Radvansky, 1999] Laura A Carlson-Radvansky, Eric S Covey, and Kathleen M Lattanzi. ""What" effects on "where": Functional influences on spatial relations". *Psychological Science* 10.6 (1999), pp. 516–521 (cit. on pp. 16, 17).

[Carrillo, 2008] Paul E Carrillo. "Information and real estate transactions: the effects of pictures and virtual tours on home sales". *Department of Economics, The George Washington University, Washington, DC. Luettu* 14 (2008), p. 2016 (cit. on p. 13).

[Charles, 2023] William Charles, Nathalie Aussenac-Gilles, and Nathalie Hernandez. "HHT: An Approach for Representing Temporally-Evolving Historical Territories". *European Semantic Web Conference*. Springer. 2023, pp. 419–435 (cit. on p. 119).

[Claramunt, 1995] Christophe Claramunt and Marius Thériault. "Managing time in GIS an event-oriented approach". *Recent Advances in Temporal Databases: Proceedings of the International Workshop on Temporal Databases, Zurich, Switzerland, 17–18 September 1995*. Springer. 1995, pp. 23–42 (cit. on p. 119).

[Clementini, 2019] Eliseo Clementini. "A conceptual framework for modelling spatial relations". *Information Technology and Control* 48.1 (2019), pp. 5–17 (cit. on p. 16).

[Cohn, 2020] Anthony G Cohn and Nicholas Mark Gotts. "The 'egg-yolk'representation of regions with indeterminate boundaries". *Geographic objects with indeterminate boundaries*. CRC Press, 2020, pp. 171–187 (cit. on pp. 2, 31).

[Collins, 1999] Michael Collins and Yoram Singer. "Unsupervised models for named entity classification". *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*. 1999 (cit. on p. 20).

[Copara, 2020] Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. "Contextualized French Language Models for Biomedical Named Entity Recognition". *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*. Nancy, France: ATALA et AFCP, 2020, pp. 36–48 (cit. on p. 18).

[Coventry, 2004] Kenny R Coventry and Simon C Garrod. *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press, 2004 (cit. on p. 15).

[Culotta, 2004] Aron Culotta and Jeffrey Sorensen. "Dependency tree kernels for relation extraction". *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*. 2004, pp. 423–429 (cit. on p. 23).

[Cyganiak, 2014] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. "The RDF data cube vocabulary". *W3C recommendation* 16 (2014) (cit. on p. 32).

[DeLozier, 2015] Grant DeLozier, Jason Baldridge, and Loretta London. "Gazetteer-independent toponym resolution using geographic word profiles". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015 (cit. on p. 25).

[Derungs, 2016] Curdin Derungs and Ross S Purves. "Mining nearness relations from an n-grams Web corpus in geographical space". *Spatial Cognition & Computation* 16.4 (2016), pp. 301–322 (cit. on p. 31).

[Devlin, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 23).

[Dittrich, 2015] André Dittrich, Maria Vasardani, Stephan Winter, Timothy Baldwin, and Fei Liu. "A classification schema for fast disambiguation of spatial prepositions". *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming*. 2015, pp. 78–86 (cit. on p. 16).

[Dozat, 2017] Timothy Dozat and Christopher D. Manning. "Deep Biaffine Attention for Neural Dependency Parsing". *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017 (cit. on p. 55).

[Dsouza, 2021] Alishiba Dsouza, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, and Elena Demidova. "Worldkg: A world-scale geographic knowledge graph". *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 4475–4484 (cit. on p. 36).

[Dubois, 1985] Didier Dubois and Henri Prade. "A review of fuzzy set aggregation connectives". *Information Sciences* 36.1 (1985), pp. 85–121 (cit. on p. 28).

[Dubois, 2012] Didier Dubois and Henri Prade. *Fundamentals of fuzzy sets*. Vol. 7. Springer Science & Business Media, 2012 (cit. on p. 26).

[Dürst, 2005] Martin Dürst and Michel Suignard. *Internationalized resource identifiers (IRIs)*. Tech. rep. 2005 (cit. on p. 32).

[Egenhofer, 2005] Max J Egenhofer. "Reasoning about binary topological relations". *Advances in Spatial Databases: 2nd Symposium, SSD'91 Zurich, Switzerland, August 28–30, 1991 Proceedings*. Springer. 2005, pp. 141–160 (cit. on p. 16).

[Ehrlinger, 2016] Lisa Ehrlinger and Wolfram Wöß. "Towards a definition of knowledge graphs." *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4 (2016), p. 2 (cit. on p. 32).

[Erwig, 1997] Martin Erwig and Markus Schneider. "Vague regions". *Advances in Spatial Databases: 5th International Symposium, SSD'97 Berlin, Germany, July 15–18, 1997 Proceedings 5*. Springer. 1997, pp. 298–320 (cit. on p. 31).

[Fisher, 2000] Peter Fisher. "Sorites paradox and vague geographies". *Fuzzy sets and systems* 113.1 (2000), pp. 7–18 (cit. on p. 29).

[Fisher, 1991] Peter F Fisher and Thomas M Orf. "An investigation of the meaning of near and close on a university campus". *Computers, Environment and Urban Systems* 15.1-2 (1991), pp. 23–35 (cit. on p. 30).

[Gaio, 2017] Mauro Gaio and Ludovic Moncla. "Extended Named Entity Recognition Using Finite-State Transducers: An Application To Place Names". *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*. Nice, France, 2017 (cit. on p. 24).

[Goodchild, 2000] Michael F Goodchild. “GIS and transportation: status and challenges”. *GeoInformatica* 4 (2000), pp. 127–139 (cit. on pp. 2, 31).

[Goodchild, 2007] Michael F Goodchild. “Citizens as sensors: the world of volunteered geography”. *GeoJournal* 69 (2007), pp. 211–221 (cit. on p. 69).

[Goodchild, 1998] Michael F Goodchild, Daniel R Montello, Peter Fohl, and Jon Gottsegen. “Fuzzy spatial queries in digital spatial data libraries”. *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36228)*. Vol. 1. IEEE. 1998, pp. 205–210 (cit. on pp. 2, 31).

[Grace, 2021] Rob Grace. “Toponym usage in social media in emergencies”. *International Journal of Disaster Risk Reduction* 52 (2021), p. 101923 (cit. on pp. 14, 18, 24, 44).

[Grishman, 1996] Ralph Grishman and Beth M Sundheim. “Message understanding conference-6: A brief history”. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 1996 (cit. on p. 18).

[Habibi, 2017] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. “Deep learning with word embeddings improves biomedical named entity recognition”. *Bioinformatics* 33.14 (2017), pp. i37–i48 (cit. on p. 20).

[Harris, 2013] Steve Harris, Andy Seaborne, and Eric Prud’hommeaux. “SPARQL 1.1 query language”. *W3C recommendation* 21.10 (2013), p. 778 (cit. on p. 33).

[Harrison, 2022] Zachary Harrison and Anish Khazane. “Taxonomic Recommendations of Real Estate Properties with Textual Attribute Information”. *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 479–481 (cit. on p. 14).

[Hasegawa, 2004] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. “Discovering relations among named entities from large corpora”. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 2004, pp. 415–422 (cit. on p. 24).

[Hassan, 2014] Mohsen Hassan, Adrien Coulet, and Yannick Toussaint. “Learning subgraph patterns from text for extracting disease–symptom relationships”. *1st International Workshop on Interactions between Data Mining and Natural Language Processing*. Vol. 1202. ceur-ws. 2014 (cit. on p. 24).

[Herskovits, 1985] Annette Herskovits. “Semantics and pragmatics of locative expressions”. *Cognitive science* 9.3 (1985), pp. 341–378 (cit. on pp. 2, 70).

[Herskovits, 1986] Annette Herskovits. *Language and spatial cognition*. Cambridge university press Cambridge, 1986 (cit. on p. 15).

[Hitzler, 2009] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, Sebastian Rudolph, et al. “OWL 2 web ontology language primer”. *W3C recommendation* 27.1 (2009), p. 123 (cit. on p. 32).

[Hochreiter, 1997] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 20).

[Hogan, 2021] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, et al. "Knowledge graphs". *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–37 (cit. on pp. 2, 32).

[Hu, 2019] Yingjie Hu, Huina Mao, and Grant McKenzie. "A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements". *International Journal of Geographical Information Science* 33.4 (2019), pp. 714–738. eprint: https://doi.org/10.1080/13658816.2018.1458986 (cit. on pp. 14, 18, 24, 73).

[Hu, 2021] Yingjie Hu and Jimin Wang. "How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey". *11th International Conference on Geographic Information Science, GIScience 2021, September 27-30, 2021, Poznań, Poland - Part I*. Ed. by Krzysztof Janowicz and Judith Anne Verstegen. Vol. 177. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021, 6:1–6:16 (cit. on pp. 2, 14, 18, 24, 41, 44).

[Huang, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". *arXiv preprint arXiv:1508.01991* (2015) (cit. on p. 20).

[Humphreys, 1998] Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Christian Huyck, Brian Mitchell, Hamish Cunningham, et al. "University of Sheffield: Description of the LaSIE-II system as used for MUC-7". *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998 (cit. on p. 20).

[Jabbari, 2020] Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. "A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News". *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 2293–2299 (cit. on p. 18).

[Janowicz, 2022] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, et al. "Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence". *AI Magazine* 43.1 (2022), pp. 30–39 (cit. on p. 36).

[Ji, 2021] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. "A survey on knowledge graphs: Representation, acquisition, and applications". *IEEE transactions on neural networks and learning systems* 33.2 (2021), pp. 494–514 (cit. on p. 32).

[Jones, 2008] Christopher B Jones, Ross S Purves, Paul D Clough, and Hideo Joho. "Modelling vague places with knowledge from the Web". *International Journal of Geographical Information Science* 22.10 (2008), pp. 1045–1065 (cit. on pp. 14, 25, 30, 73).

[Kambhatla, 2004] Nanda Kambhatla. “Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction”. *Proceedings of the ACL interactive poster and demonstration sessions*. 2004, pp. 178–181 (cit. on p. 23).

[Karalis, 2019] Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis. “Extending the YAGO2 knowledge graph with precise geospatial knowledge”. *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer. 2019, pp. 181–197 (cit. on p. 36).

[Kauppinen, 2007] Tomi Kauppinen and Eero Hyvönen. “Modeling and reasoning about changes in ontology time series”. *Ontologies: A handbook of principles, concepts and applications in information systems*. Springer, 2007, pp. 319–338 (cit. on p. 119).

[Keßler, 2009] Carsten Keßler, Krzysztof Janowicz, and Mohamed Bishr. “An agenda for the next generation gazetteer: Geographic information contribution and retrieval”. *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems*. 2009, pp. 91–100 (cit. on p. 36).

[Kim, 2000] Ji-Hwan Kim and Philip C Woodland. “A rule-based named entity recognition system for speech input”. *Sixth International Conference on Spoken Language Processing*. 2000 (cit. on p. 20).

[Klement, 2004] Erich Peter Klement, Radko Mesiar, and Endre Pap. “Triangular norms. Position paper I: basic analytical and algebraic properties”. *Fuzzy sets and systems* 143.1 (2004), pp. 5–26 (cit. on p. 28).

[Kolbe, 2021] Jens Kolbe, Rainer Schulz, Martin Wersing, and Axel Werwatz. “Real estate listings and their usefulness for hedonic regressions”. *Empirical Economics* (2021), pp. 1–31 (cit. on p. 13).

[Krupka, 1998] George R. Krupka and Kevin Hausman. “IsoQuest Inc.: Description of the NetOwl^{TM} Extractor System as Used for MUC-7”. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. 1998 (cit. on p. 20).

[Kulik, 2001] Lars Kulik. “A geometric theory of vague boundaries based on supervaluation”. *International conference on spatial information theory*. Springer. 2001, pp. 44–59 (cit. on p. 76).

[Kuru, 2016] Onur Kuru, Ozan Arkan Can, and Deniz Yuret. “Charner: Character-level named entity recognition”. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 911–921 (cit. on p. 22).

[Laddada, 2020] Wissame Laddada, Fabien Duchateau, Franck Favetta, and Ludovic Moncla. “Ontology-based approach for neighborhood and real estate recommendations”. *Proceedings of the 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising*. 2020, pp. 1–10 (cit. on p. 37).

[Lafferty, 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data" (2001) (cit. on p. 21).

[Lample, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. "Neural architectures for named entity recognition". *arXiv preprint arXiv:1603.01360* (2016) (cit. on pp. 20, 21, 50).

[Landau, 1993] Barbara Landau and Ray Jackendoff. "Whence and whither in spatial language and spatial cognition?" *Behavioral and brain sciences* 16.2 (1993), pp. 255–265 (cit. on p. 16).

[Le, 2019] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, et al. "Flaubert: Unsupervised language model pretraining for french". *arXiv preprint arXiv:1912.05372* (2019) (cit. on p. 18).

[Lehmann, 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, et al. "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia". *Semantic web* 6.2 (2015), pp. 167–195 (cit. on p. 34).

[Leidner, 2008] Jochen L Leidner. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers, 2008 (cit. on pp. 24, 69, 84).

[Leidner, 2011] Jochen L Leidner and Michael D Lieberman. "Detecting geographical references in the form of place names and associated spatial natural language". *Sigspatial Special* 3.2 (2011), pp. 5–11 (cit. on p. 24).

[Lesbegueries, 2006] Julien Lesbegueries, Christian Sallaberry, and Mauro Gaio. "Associating spatial patterns to text-units for summarizing geographic information". *ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop*. 2006, pp. 40–43 (cit. on pp. 15, 98).

[Li, 2002] Huifeng Li, Rohini K Srihari, Cheng Niu, and Wei Li. "Location normalization for information extraction". *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002 (cit. on p. 24).

[Lieberman, 2011] Michael D Lieberman and Hanan Samet. "Multifaceted toponym recognition for streaming news". *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 2011, pp. 843–852 (cit. on p. 24).

[Lieberman, 2012] Michael D Lieberman and Hanan Samet. "Adaptive context features for toponym resolution in streaming news". *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 2012, pp. 731–740 (cit. on pp. 25, 69).

[Lieberman, 2010] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. "Geotagging with local lexicons to build indexes for textually-specified spatial data". *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. IEEE. 2010, pp. 201–212 (cit. on p. 50).

[Lóscio, 2018] Bernadette Farias Lóscio, Caroline Burle, Newton Calegari, et al. *Data on the web best practices. W3C recommendation (2017)*. 2018 (cit. on p. 105).

[M Hall, 2011] Mark M Hall, Philip D Smart, and Christopher B Jones. "Interpreting spatial language in image captions". *Cognitive processing* 12 (2011), pp. 67–94 (cit. on p. 70).

[Martin, 2019] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, et al. "CamemBERT: a tasty French language model". *arXiv preprint arXiv:1911.03894* (2019) (cit. on pp. 18, 50).

[McKenzie, 2017a] Grant McKenzie and Benjamin Adams. "Juxtaposing thematic regions derived from spatial and platial user-generated content". *13th international conference on spatial information theory (COSIT 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017 (cit. on p. 73).

[McKenzie, 2017b] Grant McKenzie and Yingjie Hu. "The "Nearby" exaggeration in real estate". *Proceedings of the Cognitive Scales of Spatial Information Workshop (CoSSI 2017), L'Aquila, Italy*. 2017, pp. 4–8 (cit. on pp. 6, 17, 98).

[Medad, 2020] Amine Medad, Mauro Gaio, Ludovic Moncla, Sébastien Mustière, and Yannick Le Nir. "Comparing supervised learning algorithms for spatial nominal entity recognition". *AGILE: GIScience Series* 1 (2020), p. 15 (cit. on p. 44).

[Menin, 2021] Aline Menin, Lucie Cadorel, Andrea G. B. Tettamanzi, Alain Giboin, Fabien Gandon, and Marco Winckler. "ARViz: Interactive Visualization of Association Rules for RDF Data Exploration". *IV 2021 - 25th International Conference Information Visualisation*. Vol. 25. 2021 25th International Conference Information Visualisation (IV). Melbourne / Virtual, Australia, 2021, pp. 13–20 (cit. on p. 121).

[Mikolov, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems* 26 (2013) (cit. on pp. 22, 50).

[Miles, 2009] Alistair Miles and Sean Bechhofer. "SKOS simple knowledge organization system reference. W3C Recommendation, 2009". *URL: https://www.w3. org/TR/skos-reference* (2009) (cit. on p. 36).

[Moncla, 2017] Ludovic Moncla, Mauro Gaio, Thierry Joliveau, and Yves-François Le Lay. "Automated geoparsing of paris street names in 19th century novels". *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. 2017, pp. 1–8 (cit. on pp. 14, 24, 44, 50).

[Moncla, 2014] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. "Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus". *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*. 2014, pp. 183–192 (cit. on pp. 18, 117).

[Montello, 2003] Daniel R Montello, Michael F Goodchild, Jonathon Gottsegen, and Peter Fohl. "Where's downtown?: Behavioral methods for determining referents of vague spatial queries". *Spatial Cognition & Computation* 3.2-3 (2003), pp. 185–204 (cit. on pp. 1, 30, 78).

[Moratz, 2006] Reinhard Moratz and Thora Tenbrink. "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations". *Spatial cognition and computation* 6.1 (2006), pp. 63–107 (cit. on p. 70).

[Nadeau, 2007] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification". *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26 (cit. on p. 18).

[Nadeau, 2006] David Nadeau, Peter D Turney, and Stan Matwin. "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity". *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*. Springer. 2006, pp. 266–277 (cit. on p. 20).

[Nakayama, 2018] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. *doccano: Text Annotation Tool for Human*. 2018 (cit. on p. 45).

[Nguyen, 2015] Thien Huu Nguyen and Ralph Grishman. "Relation Extraction: Perspective from Convolutional Neural Networks". *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 39–48 (cit. on p. 24).

[Nivre, 2016] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, et al. "Universal dependencies v1: A multilingual treebank collection". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 1659–1666 (cit. on pp. 23, 54).

[Nivre, 2020] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, et al. "Universal Dependencies v2: An evergrowing multilingual treebank collection". *arXiv preprint arXiv:2004.10643* (2020) (cit. on p. 56).

[Nowak, 2017] Adam Nowak and Patrick Smith. "Textual analysis in real estate". *Journal of Applied Econometrics* 32.4 (2017), pp. 896–918 (cit. on p. 13).

[Ortiz Suárez, 2020] Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. "Establishing a New State-of-the-Art for French Named Entity Recognition". *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 4631–4638 (cit. on p. 18).

[Overell, 2008] Simon Overell and Stefan Rüger. "Using co-occurrence models for place-name disambiguation". *International Journal of Geographical Information Science* 22.3 (2008), pp. 265–287 (cit. on pp. 25, 69).

[Paasch, 2005] Jesper M Paasch. "Legal Cadastral Domain Model: An Object-orientated Approach". *Nordic Journal of Surveying and Real Estate Research* 2.1 (2005), pp. 117–136 (cit. on p. 37).

[Paulheim, 2017] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". *Semantic web* 8.3 (2017), pp. 489–508 (cit. on p. 32).

[Platonov, 2018] Georgiy Platonov and Lenhart Schubert. "Computational models for spatial prepositions". *Proceedings of the First International Workshop on Spatial Language Understanding*. 2018, pp. 21–30 (cit. on p. 70).

[Purves, 2018] Ross S Purves, Paul Clough, Christopher B Jones, Mark H Hall, Vanessa Murdock, et al. "Geographic information retrieval: Progress and challenges in spatial search of text". *Foundations and Trends® in Information Retrieval* 12.2-3 (2018), pp. 164–318 (cit. on p. 15).

[Qi, 2019] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. "Universal dependency parsing from scratch". *arXiv preprint arXiv:1901.10457* (2019) (cit. on p. 55).

[Qi, 2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020 (cit. on p. 55).

[Randell, 1992] David A Randell, Zhan Cui, and Anthony G Cohn. "A spatial logic based on regions and connection." *KR* 92 (1992), pp. 165–176 (cit. on p. 16).

[Sainsbury, 2009] R. M. Sainsbury. "Vagueness: the paradox of the heap". *Paradoxes*. 3rd ed. Cambridge University Press, 2009, pp. 40–68 (cit. on p. 29).

[Sarawagi, 2008] Sunita Sarawagi et al. "Information extraction". *Foundations and Trends® in Databases* 1.3 (2008), pp. 261–377 (cit. on p. 18).

[Schneider, 1999] Markus Schneider. "Uncertainty management for spatial datain databases: Fuzzy spatial data types". *Advances in Spatial Databases: 6th International Symposium, SSD'99 Hong Kong, China, July 20—23, 1999 Proceedings 6*. Springer. 1999, pp. 330–351 (cit. on p. 99).

[Schneider, 2001] Markus Schneider. "A design of topological predicates for complex crisp and fuzzy regions". *Conceptual Modeling—ER 2001: 20th International Conference on Conceptual Modeling Yokohama, Japan, November 27–30, 2001 Proceedings 20*. Springer. 2001, pp. 103–116 (cit. on p. 31).

[Schockaert, 2008] Steven Schockaert, Martine De Cock, and Etienne E Kerre. "Location approximation for local search services using natural language hints". *International Journal of Geographical Information Science* 22.3 (2008), pp. 315–336 (cit. on p. 70).

[Schockaert, 2011] Steven Schockaert, Martine De Cock, and Etienne E Kerre. *Reasoning about fuzzy temporal and spatial information from the web*. Vol. 3. World Scientific, 2011 (cit. on pp. 31, 107).

[Schreiber, 2000] August Th Schreiber, Guus Schreiber, Hans Akkermans, Anjo Anjewierden, Nigel Shadbolt, Robert de Hoog, et al. *Knowledge engineering and management: the CommonKADS methodology*. MIT press, 2000 (cit. on p. 31).

[Schutz, 2005] Alexander Schutz and Paul Buitelaar. "Relext: A tool for relation extraction from text in ontology extension". *The Semantic Web–ISWC 2005: 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005. Proceedings 4*. Springer. 2005, pp. 593–606 (cit. on p. 24).

[Sekine, 2004] Satoshi Sekine and Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." *LREC*. Lisbon, Portugal. 2004, pp. 1977–1980 (cit. on p. 18).

[Shi, 2017] Ling Shi, Nikolay Nikolov, Dina Sukhobokb, Tatiana Tarasovac, and Dumitru Roman. "The prodatamarket ontology for publishing and integrating crossdomain real property data". *journal Territorio Italia. Land Administration, Cadastre and Real Estate* 2 (2017) (cit. on p. 37).

[Shi, 2018] Ling Shi and Dumitru Roman. "Ontologies for the real property domain". *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. 2018, pp. 1–8 (cit. on p. 37).

[Singhal, 2012] Amit Singhal. "Introducing the knowledge graph: things, not strings". *Official google blog* 5.16 (2012), p. 3 (cit. on p. 31).

[Siniak, 2020] Nikolai Siniak, Tom Kauko, Sergey Shavrov, and Ninoslav Marina. "The impact of proptech on real estate industry growth". *IOP Conference Series: Materials Science and Engineering* 869.6 (2020), p. 062041 (cit. on p. 2).

[Skoumas, 2016] Georgios Skoumas, Dieter Pfoser, Anastasios Kyrillidis, and Timos Sellis. "Location estimation using crowdsourced spatial relations". *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 2.2 (2016), pp. 1–23 (cit. on p. 70).

[Sladić, 2013] Dubravka Sladić, Miro Govedarica, D Pržulj, Aleksandra Radulović, and Dušan Jovanović. "Ontology for real estate cadastre". *Survey Review* 45.332 (2013), pp. 357–371 (cit. on p. 37).

[Smets, 1997] Philippe Smets. "Imperfect information: Imprecision-uncertainty". *Uncertainty management in information systems. from needs to solutions* (1997), pp. 225–254 (cit. on p. 25).

[Starr, 2021] Christopher W Starr, Jesse Saginor, and Elaine Worzala. "The rise of PropTech: Emerging industrial technologies and their impact on real estate". *Journal of Property Investment & Finance* 39.2 (2021), pp. 157–169 (cit. on p. 2).

[Stock, 2022] Kristin Stock, Christopher B Jones, Shaun Russell, Mansi Radke, Prarthana Das, and Niloofar Aflaki. "Detecting geospatial location descriptions in natural language text". *International Journal of Geographical Information Science* 36.3 (2022), pp. 547–584 (cit. on pp. 16, 17).

[Stock, 2018] Kristin Stock and Javid Yousaf. "Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data". *International Journal of Geographical Information Science* 32.6 (2018), pp. 1087–1116 (cit. on pp. 2, 70).

[Stokes, 2008] Nicola Stokes, Yi Li, Alistair Moffat, and Jiawen Rong. "An empirical study of the effects of NLP components on Geographic IR performance". *International Journal of Geographical Information Science* 22.3 (2008), pp. 247–264 (cit. on p. 24).

[Stubkjaer, 2017] Erik Stubkjaer. *The ontology and modelling of real estate transactions*. Routledge, 2017 (cit. on p. 37).

[Suchanek, 2007] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge". *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 697–706 (cit. on p. 34).

[Syed, 2022] Mehtab Alam Syed, Elena Arsevska, Mathieu Roche, and Maguelonne Teisseire. "GeoXTag: Relative spatial information extraction and tagging of unstructured text". *AGILE: GIScience Series* 3 (2022), p. 16 (cit. on pp. 24, 98).

[Szarvas, 2006] György Szarvas, Richárd Farkas, and András Kocsor. "A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms". *Discovery Science: 9th International Conference, DS 2006, Barcelona, Spain, October 7-10, 2006. Proceedings 9*. Springer. 2006, pp. 267–278 (cit. on p. 20).

[Takeuchi, 2002] Koichi Takeuchi and Nigel Collier. "Use of support vector machines in extended named entity recognition". *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002 (cit. on p. 20).

[Talmy, 2000] Leonard Talmy. *Toward a cognitive semantics*. Vol. 2. MIT press, 2000 (cit. on p. 15).

[Tran, 2017] Quan Tran, Andrew MacKinlay, and Antonio Jimeno Yepes. "Named entity recognition with stack residual lstm and trainable bias decoding". *arXiv preprint arXiv:1706.07598* (2017) (cit. on p. 22).

[Tyler, 2003] Andrea Tyler and Vyvyan Evans. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press, 2003 (cit. on pp. 2, 70).

[Uschold, 1996] Mike Uschold and Michael Gruninger. "Ontologies: Principles, methods and applications". *The knowledge engineering review* 11.2 (1996), pp. 93–136 (cit. on p. 93).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) (cit. on p. 51).

[Vrandečić, 2014] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". *Communications of the ACM* 57.10 (2014), pp. 78–85 (cit. on p. 34).

[Wang, 2019] Jimin Wang and Yingjie Hu. "Are We There yet? Evaluating State-of-the-Art Neural Network Based Geoparsers Using EUPEG as a Benchmarking Platform". *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*. GeoHumanities '19. Chicago, Illinois: Association for Computing Machinery, 2019 (cit. on p. 25).

[Welty, 2006] Chris Welty, Richard Fikes, and Selene Makarios. "A reusable ontology for fluents in OWL". *FOIS*. Vol. 150. 2006, pp. 226–236 (cit. on p. 119).

[Wilkinson, 2016] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3.1 (2016), pp. 1–9 (cit. on p. 105).

[Worboys, 2001] Michael F Worboys. "Nearness relations in environmental space". *International Journal of Geographical Information Science* 15.7 (2001), pp. 633–651 (cit. on p. 30).

[Wu, 2019] Shijie Wu and Mark Dredze. "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT". *arXiv preprint arXiv:1904.09077* (2019) (cit. on p. 118).

[Xu, 2018] Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. "A bidirectional LSTM and conditional random fields approach to medical named entity recognition". *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*. Springer. 2018, pp. 355–365 (cit. on p. 20).

[Yager, 1988] Ronald R Yager. "On ordered weighted averaging aggregation operators in multicriteria decisionmaking". *IEEE Transactions on systems, Man, and Cybernetics* 18.1 (1988), pp. 183–190 (cit. on pp. 28, 29, 82).

[Yang, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "Xlnet: Generalized autoregressive pretraining for language understanding". *Advances in neural information processing systems* 32 (2019) (cit. on p. 23).

[Yao, 2015] Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Waqas Anwar. "Biomedical named entity recognition based on deep neutral network". *Int. J. Hybrid Inf. Technol* 8.8 (2015), pp. 279–288 (cit. on p. 22).

[Yu, 2021] Wei Yu, Zhongming Ma, Gautam Pant, and Jing Hu. "The effect of virtual tours on house price and time on market". *Journal of Real Estate Literature* 28.2 (2021), pp. 133–149 (cit. on p. 13).

[Zadeh, 1965] L Zadeh. "Fuzzy sets". *Inform Control* 8 (1965), pp. 338–353 (cit. on p. 26).

[Zhai, 2017] Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. "Neural models for sequence chunking". *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017 (cit. on p. 22).

[Zhang, 2020] MeiShan Zhang. "A survey of syntactic-semantic parsing based on constituent and dependency structures". *Science China Technological Sciences* 63.10 (2020), pp. 1898–1920 (cit. on p. 55).

[Zhou, 2007] Guodong Zhou, Min Zhang, DongHong Ji, and Qiaoming Zhu. "Tree kernel-based relation extraction with context-sensitive structured parse tree information". *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 2007, pp. 728–736 (cit. on p. 23).

[Zhou, 2017] Peng Zhou, Suncong Zheng, Jiaming Xu, Zhenyu Qi, Hongyun Bao, and Bo Xu. "Joint extraction of multiple relations and entities by using a hybrid neural network". *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 16*. Springer. 2017, pp. 135–146 (cit. on p. 22).

[Zimmermann, 2011] Hans-Jürgen Zimmermann. *Fuzzy set theory—and its applications*. Springer Science & Business Media, 2011 (cit. on pp. 28, 82).