Fuzzy representation of vague spatial descriptions in real estate advertisements

Lucie Cadorel, Denis Overal, Andrea G. B. Tettamanzi

Fuzzy Representation of Vague Spatial Descriptions in Real Estate Advertisements

Lucie Cadorel
lucie.cadorel@inria.fr
Université Côte d'Azur, Inria, CNRS, I3S, KCityLabs
Sophia-Antipolis, France

Denis Overal
denis.overal@kcitylabs.fr
KCityLabs
Sophia-Antipolis, France

Andrea G. B. Tettamanzi
andrea.tettamanzi@univ-cotedazur.fr
Université Côte d'Azur, Inria, CNRS, I3S
Sophia-Antipolis, France

ABSTRACT

Geocoding a spatial description is challenging since vernacular place names and vague spatial expressions give uncertainty and ambiguity to the description. Usually, digital gazetteers are used to match geospatial objects to their boundaries. However, gazetteers do not contain all places. Therefore, a number of studies have proposed to enrich gazetteers by estimating and representing the vernacular places. Nevertheless, only a few approaches have taken into account vague spatial expressions such as "nearby", and have represented geospatial objects as sharp boundaries. In this work, we present an automatic workflow to retrieve a location approximation of vague spatial description. We propose a model to estimate a fuzzy representation of each mentioned geospatial information and spatial expressions. Then, we perform information fusion to find a location approximation of a property. Lastly, we demonstrate our proposed method by applying it to the case of French Real Estate advertisements with two real-world datasets in Nice and Paris. Real Estate advertisements allow us to deal with uncertain geospatial objects since a vague and exaggerated property location's description is usually provided. Our results show that our proposed method is promising and able to correctly approximate a location from uncertain spatial descriptions.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Geographic information systems**.

KEYWORDS

Uncertainty, Fuzzy sets, Spatial relationships, Natural Language, Geocoding

ACM Reference Format:
Lucie Cadorel, Denis Overal, and Andrea G. B. Tettamanzi. 2022. Fuzzy Representation of Vague Spatial Descriptions in Real Estate Advertisements . In *The 6th ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (LocalRec '22) (LocalRec '22), November 1, 2022, Seattle, WA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557992.3565994

LocalRec '22, November 1, 2022, Seattle, WA, USA

1 INTRODUCTION

Geographic coordinate systems, such as the World Geodetic System (WGS84) used by GPS, quantitatively and precisely locate a place. However, a lot of unstructured (textual) data using natural language, such as travel blogs, social media in emergencies or Real Estate advertisements, qualitatively refer to locations. Humans often use spatial expressions with toponyms (e.g., "West of Nice, France") which give vagueness and ambiguity to the description. Sometimes, place types are even preferred to toponyms to locate a place (e.g., "Near the beach"). Therefore, uncertainty arises in the description of locations and poses a challenge to geocode places.

Representing spatial descriptions and in particular spatial expressions, with exact geometries and sharp boundaries might not be suitable to capture uncertainty. Some solutions represent toponyms with quantitative spatial expression as polygons (e.g., isochron). Nevertheless, spatial expressions are not always quantitative (e.g., "nearby", "north of", "not far from") and presume fuzzy and imprecise boundaries.

Fuzzy set theory uses membership functions to define the degree to which each point in the space belongs to a fuzzy set. This approach could deal with uncertain descriptions and represent geospatial objects with fuzzy boundaries. Several studies have proposed to use fuzzy sets to model natural language location descriptions ([2, 4, 5, 9, 10]) but, to the best of our knowledge, evaluations have never been carried out on a large scale.

In this paper, we propose to automatically retrieve a location approximation of a property based on Real Estate descriptions written in French using fuzzy set theory. Indeed, Real Estate professionals do not often give the exact position of a property, and use vague spatial expression to locate it [7]. Also, location is one of the most valuable factors of purchasing. Therefore, Real Estate agents exaggerate boundaries of place names that are popular and well-reputed to promote a property [8]. Thus, Real Estate ads are a great source of data to deal with uncertain geospatial objects. Furthermore, many applications can result from the extracted fuzzy representation. First, understanding Real Estate language and positioning online ads are fundamental to evaluate the Real Estate market and to obtain an in-depth knowledge of the territory. It could also help professionals to compare their properties for sale to similar ones. Lastly, fuzzy representations of geospatial objects might enrich Geographic Information Systems since Real Estate professionals often mention non-official/local place-names in their advertisements.

In this work, we present a method to represent a location approximation of natural language description from Real Estate advertisements. We first locate and create a fuzzy representation of

each mentioned geospatial information item. Then, we perform information fusion to find the (fuzzy) location approximation of the property description. We evaluate the model on two real-world datasets in Paris and Nice, France.

The rest of the paper is organized as follows: Section 2 presents the dataset and the methodological details of our model; Section 3 presents and discusses the results of experiments in Paris and Nice; Finally, Section 4 draws some conclusions and outlines directions for further research.

2 PROPOSED METHOD

2.1 Dataset

As a case study, we focused on two cities in France: Paris and Nice. We first collected Real Estate advertisements from various online advertisers written in French. We extracted the text describing the property and its location, and the coordinates (latitude/longitude) given in the metadata. Then, we selected advertisements with precise coordinates. Indeed, we computed frequency of each pairs of latitude and longitude and, we found out coordinates with a high frequency that correspond to the center of the city or the neighbourhood. Thus, we kept pairs with a low frequency (e.g., less than 15 times in the dataset) in order to be at building or street level. The number of samples for each city is reported in Table 1. Then, we applied a geospatial information extraction workflow, described in [3] and designed for Real Estate advertisements, to retrieve entities (Toponym, Place Type, Spatiotemporal expression, and Mode of Transportation) and relationships. This method is a two-stage pipeline involving Named Entity Recognition and Relationship Extraction. The Named Entity Recognition model architecture is a BiLSTM+CRF combined with a text embedding, whereas the Relationship Extraction is based on Dependency Parsing.

City	Number of ads
Nice	1593
Paris	2384
Total	3977

Table 1: Number of ads by cities

2.2 Model

We propose to automatically represent a location approximation of a property from spatial descriptions found in Real Estate advertisements. Our method mainly consists of representing the footprint of each spatial information item extracted from the text and performing information fusion to find a location approximation of the property at hand.

2.2.1 Positioning Uncertain Geospatial Information. As we extracted place names using our workflow designed for Real Estate advertisements, we had to deal with vernacular and local place names. Also, professionals give a spatial position using vague spatiotemporal expressions (e.g., "nearby", "close to", "not far away from", etc.). A place name depends on its name but also on its type and spatial relationship. For instance, "Champs-Elysées" is different from "Avenue des Champs-Elysées", which in turn is different from "Nearby Avenue des Champs-Elysées". The first one could refer to a neighborhood or the avenue, the second explicitly refers to the avenue whereas the last one gives an uncertain spatial location around the avenue. Therefore, we defined a place name ([1]) as follows :

- Toponym,
- or Place Type + Toponym,
- or Spatial Relationship + Toponym,
- or Spatial Relatiosnhip + Place Type + Topoynm.

Then, we estimated place names by using Kernel Density Estimation, 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. In our study, we chose Gaussian kernels to approximate the boundary of a geospatial object, mainly because they are well-supported by existing libraries and Gaussian membership functions are a popular choice for fuzzy sets. For each place name, we selected all geotagged ads mentioning it, removed outliers and estimated its footprint based on the advertisements' coordinates.

2.2.2 Information Fusion. The second step of our model is about combining all geospatial information to retrieve a location approximation of the property. We proposed to represent the kernel density estimation as a fuzzy set to deal with the uncertainty of the data and use fuzzy operators. In fuzzy sets theory, elements of a set have degrees of membership, generally in the interval [0,1]. A major advantage of fuzzy set theory is that it returns an approximation of the location instead of a sharp area that should be the location. A fuzzy set A is characterized by its membership function μ_A , which describes the degree of membership of a point in the space to a fuzzy geographic set. Another advantage is that we can retrieve crisp sets from the membership function thanks to α-cuts, defined as follows:

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

An α-cut of a set A, $\tilde{A}_\alpha$, is a crisp set where all the points belonging to the set have a degree of membership greater than or equal to α. We easily transformed Gaussian kernels into a fuzzy set by taking the density function (normalized between 0 and 1) as a membership function. Then, we computed ordered weighted averaging (OWA) [11] membership functions defined as follows :

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

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.

Figure 1 gives an example of the method applied on a Real Estate advertisement in Paris. The Real Estate professional mentions three identified place names. Our method estimates the kernel density of the three place names, and then computes a membership function based on the fusion of the three relevant fuzzy sets with the arithmetic mean operator. We can see that the red icon, which is the exact location of the property, belongs to the fusion of the estimated footprints of these three information items with a high degree of membership.

3 EXPERIMENTS AND RESULTS

In this section, we first introduce the metrics to evaluate our method. Then, we present and discuss the performance of the method. All the results are based on the two datasets presented in 2.1 and split in 10 folds to carry out a cross-validation.

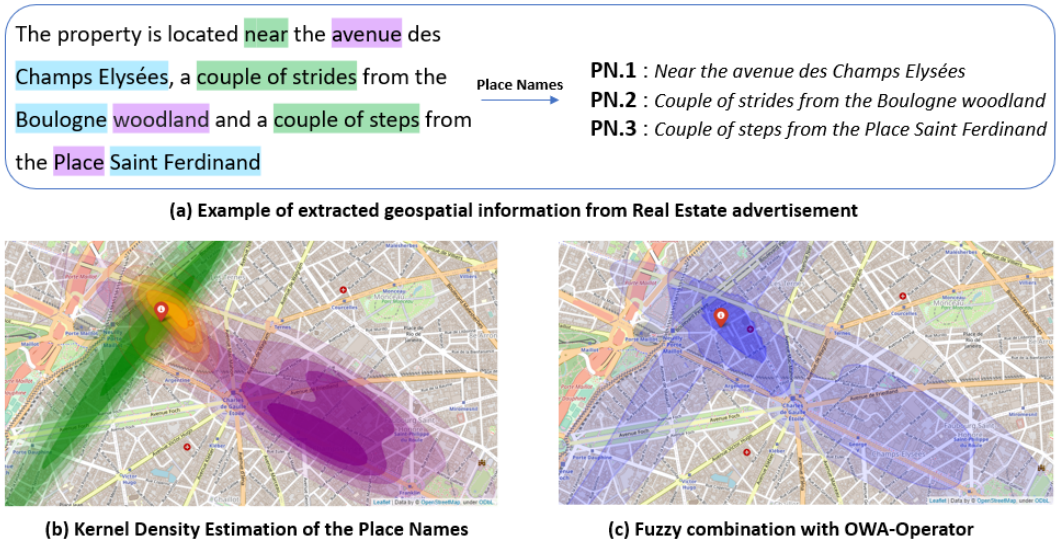

Figure 1: Example of location approximation of a Real Estate advertisement in Paris

3.1 Evaluation Metrics

To the best of our knowledge, there is no standard evaluation metric 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 methods. As we do have coordinates for each ad, we suggest to use those three metrics with some adaptations to our problem. We define the following notation:

- L_C^α : number of ad locations inside the zone found by the system corresponding to a given α-cut
- L_I^α : number of ad locations outside the zone found by the system corresponding to a given α-cut
- L_U^α : number of ad locations that the system did not find a zone for a given α-cut

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

$$P = \frac{L_C^\alpha}{L_C^\alpha + L_I^\alpha}$$
$$R = \frac{L_C^\alpha}{L_C^\alpha + L_I^\alpha + L_U^\alpha}$$
$$F1 = 2 \times \frac{P \times R}{P + R}$$

Precision refers to the ratio of the number of ad locations correctly found inside the zone of a given α-cut among the number of ad locations that the system found a zone for the α-cut. A high precision means that when the system find a zone for the α-cut, then the ad coordinates are inside the area.

Recall is the ratio of the number of ad locations correctly found inside the zone of a given α-cut among the total number of ad locations. For this task, the recall has the same definition as accuracy. Recall gives information about the capability of the system to find a zone for a certain α-cut. Indeed, a high precision and a low recall mean that the system is good at finding ad coordinates inside fuzzy location, but fails to resolve many areas for a given *alpha*-cut.

F1-Score summarizes precision and recall in one metric by computing their harmonic mean.

A limitation of those metrics is that we only differentiate if an ad location is inside or outside the α-cut. This binary distinction does not take into account the area of the zone. A fuzzy location that equals to the entire city would not be penalized, whereas it is not precise at all. On the other hand, a small fuzzy location where the ad coordinates are not within but at a very close distance would be penalized with those metrics. Thus, we propose to also use a continuous metric called Root Mean Squared Distance (RMSD) defined by Leidner [6]. 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 ad coordinates p_i and the most representative point on surface[1] c_i of the α-cut:

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

Nevertheless, RMSD only uses ad locations with a zone found for the α-cut, and does not evaluate the performance of the system to find a fuzzy representation. Thus, we should find a good compromise between a high F1-score and a low RMSD.

3.2 Discussion

We first compared the metrics evolution with different α-cuts based on the dataset in Nice. We computed the evaluation metrics for 9 values of α from 0.1 to 0.9 by cross-validation. We did not calculate $\alpha = 1$ because the fuzzy representation is a single point and the ad coordinates would practically never match. Figure 2 summarizes the mean of the evaluation metrics based on the results of the 10 folds. Regarding precision, recall and F1-score, the method reaches a high precision for the three operators whereas recall remains lower. These results highlight that the method does not always find an area for a given *alpha*-cut (i.e., low recall), but is very good at giving a correct area (i.e., high precision). We also notice that the higher the α, the lower the performance. For precision, recall and F1-Score this could be explained by a smaller area (e.g., $\alpha = 0.9$ yield a very small area and very low performance). On the other hand, RMSD decreases and then surprisingly increases for $\alpha \geq 0.6$. We could have expected a better RMSD for small areas. It means that the smaller the area, the less representative the location approximation. Lastly, it is difficult to choose a good α since the performance of F1-score decreases and that of RMSD increases with α.

[1] https://postgis.net/docs/ST_PointOnSurface.html

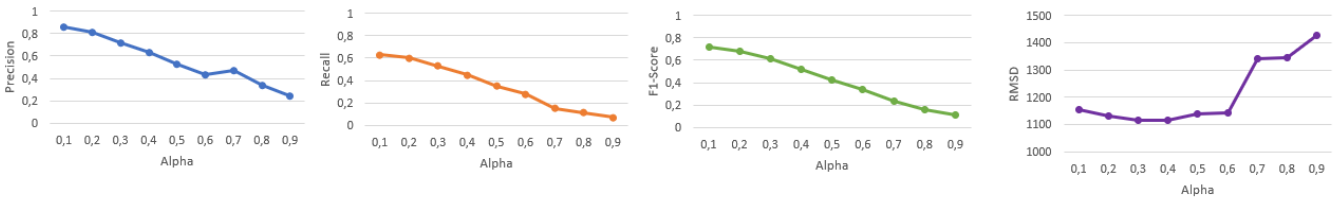

Figure 2: Metrics evolution according to different α-cuts

Secondly, we applied the method on the two datasets. We chose to extract the α-cut with $\alpha = 0.2$ since the previous evaluation showed that precision is above 0.8 and RMSD is smaller than for $\alpha = 0.1$. In Table 2, we noticed that the method reaches high precision, particularly for Nice. As we trained the geospatial information extraction model, described in [3], on advertisements located in the French Riviera, it is not surprising that our method obtains better results for Nice. Nevertheless, the results shows that the method is able to generalize to Paris. Regarding the RMSD and area, the method achieved very good results for Paris. Indeed, we noticed that more points of interests (POI) such as subways, monuments or museums, are mentioned by real estate professionals in Paris. This gives a big help to our method in order to delineate a zone. On the other hand, smaller cities, such as Nice, possess fewer POIs and real estate professionals often mention the same one (e.g., "Promenade des Anglais" in Nice) despite the property is pretty far from it. Our method is obviously more accurate with more specific geospatial information.

In a nutshell, this evaluation shows promising results since the method is able to correctly locate uncertain spatial descriptions in two different cities (i.e., high precision). A drawback of this method is the low recall which means that we do not always find boundaries for a number of geospatial objects. Nevertheless, one could easily boost the results simply by collecting more data. Another limitation of this study is the choice of $\alpha = 0.2$ based on the first evaluation in Nice. This value for α might not be equally suitable for every city and method. Finally, we evaluated the method on two big cities, and a great challenge could be to apply it on a rural area.

Metrics	Nice	Paris
P	**0.81**	**0.75**
R	**0.60**	**0.52**
F1-score	**0.68**	**0.62**
RMSD	**1131**	**677**
Area (km^2)	**5.3**	**1.3**

Table 2: Model performance evaluation results

4 CONCLUSION

In this paper, we have presented a method to automatically retrieve a location approximation of a property from its vague spatial description in Real Estate advertisements. In order to deal with uncertainty, we have proposed to use fuzzy set theory to represent place names, and combine several geospatial information items. The method returns an approximate area of the vague description. Moreover, our method estimates spatial footprints with a Kernel Density Estimation, based on the coordinates of advertisements, that models the real estate professionals' exaggeration of using place names. This method also helps to approximate vernacular place names that are not found in official gazetteers, and could be used to enrich the latter. Several directions could be considered to expand this work. First, it would be desirable to define a metric or criterion capable of summing up in one figure the quality of a (fuzzy) location approximation and help to choose an optimal α. Furthermore, we have not yet treated unnamed entities (e.g., "close to the beach", "nearby the university"), whereas they are detected by our geospatial information extraction model. After delineating a zone from place names, those terms can restrain the area if we can match them to footprints (e.g., if there is only one university in the city, we can easily match it). Also, we would like to add the mode of transportation to better model spatial expressions (e.g., 5 minutes by car is different from by walk).

5 ACKNOWLEDGMENTS

This work has been partially supported by the French government, through the 3IA Côte d'Azur "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

REFERENCES

[1] Brandon Bennett and Pragya Agarwal. 2007. Semantic Categories Underlying the Meaning of 'Place'. In *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4736)*, Stephan Winter, Matt Duckham, Lars Kulik, and Benjamin Kuipers (Eds.). Springer, 78–95. https://doi.org/10.1007/978-3-540-74788-8_6

[2] H. S. Al-Olimat et al. 2019. Towards Geocoding Spatial Expressions (Vision Paper). In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, IL, USA) *(SIGSPATIAL '19)*. Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10.1145/3347146.3359356

[3] L. Cadorel et al. 2021. Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text. In *Proceedings of the 11th on Knowledge Capture Conference* (Virtual Event, USA) *(K-CAP '21)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/3460210.3493547

[4] Peter Fisher, Jo Wood, and Tao Cheng. 2004. Where Is Helvellyn? Fuzziness of Multi-Scale Landscape Morphometry. *Transactions of the Institute of British Geographers* 29, 1 (2004), 106–128. http://www.jstor.org/stable/3804433

[5] Christopher B. Jones, Ross S. Purves, Paul D. Clough, and Hideo Joho. 2008. Modelling vague places with knowledge from the Web. *Int. J. Geogr. Inf. Sci.* 22, 10 (2008), 1045–1065. https://doi.org/10.1080/13658810701850547

[6] Jochen L. Leidner. 2007. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum* 41, 2 (2007), 124–126. https://doi.org/10.1145/1328964.1328989

[7] G. McKenzie and Y. Hu. 2017. The " Nearby " Exaggeration in Real Estate A Position Paper.

[8] D. Medway and G. Warnaby. 2014. What's in a Name? Place Branding and Toponymic Commodification. *Environment and Planning A: Economy and Space* 46, 1 (2014), 153–167. https://doi.org/10.1068/a45571 arXiv:https://doi.org/10.1068/a45571

[9] Steven Schockaert. 2011. Vague regions in Geographic Information Retrieval. *ACM SIGSPATIAL Special* 3, 2 (2011), 24–28. https://doi.org/10.1145/2047296.2047302

[10] J. Xu and X. Pan. 2020. A Fuzzy Spatial Region Extraction Model for Object's Vague Location Description from Observer Perspective. *ISPRS International Journal of Geo-Information* 9, 12 (Nov 2020), 703. https://doi.org/10.3390/ijgi9120703

[11] R. R. Yager. 1993. Families of OWA operators. *Fuzzy Sets and Systems* 59, 2 (1993), 125–148. https://doi.org/10.1016/0165-0114(93)90194-M