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

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

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

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

Alicia Blanchi
alicia.blanchi@kcitylabs.fr
Université Côte d'Azur, ESPACE, CNRS, KCityLabs
Sophia-Antipolis, France

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

ABSTRACT

Information of the geographical and spatial type is found in numerous text documents and constitutes a very challenging target for extraction. Geoparsing applications have been developed to extract geographic terms. However, off-the-shelf Named Entity Recognition (NER) models are mainly designed for Toponym recognition and are very sensitive to language specificity. In this paper, we propose a workflow to first extract geographic and spatial entities based on a *BiLSTM-CRF* architecture with a concatenation of several text representations. We also propose a Relation Extraction module, particularly aimed at spatial relationships extraction, to build a structured Geospatial knowledge base. We demonstrate our pipeline by applying it to the case of French housing advertisements, which generally provide information about a property's location and neighbourhood. Our results show that the workflow tackles French language and the variability and irregularity of housing advertisements, generalizes Geoparsing to all geographic and spatial terms, and successfully retrieves most of the relationships between entities from the text.

CCS CONCEPTS

• **Information systems → Specialized information retrieval**;
• **Computing methodologies → Machine learning**.

KEYWORDS

Information Extraction, Geographical Knowledge, Text Mining, Named Entity Recognition, Relation Extraction

ACM Reference Format:
Lucie Cadorel, Alicia Blanchi, and Andrea G. B. Tettamanzi. 2021. Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text. In *Proceedings of the 11th Knowledge Capture Conference (K-CAP '21), December 2–3, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3460210.3493547

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 environments, or the location of events (e.g., natural disasters). To capture and extract spatial information from text, Geoparsing applications have been widely developed and are mainly focused on Place-name or Toponym extraction. Indeed, names are often used by people to refer to places and they can be linked to existing Digital Gazetteers. However, Gazetteers mostly record official Place names, whereas text documents may contain non-official, local Place names. Also, spatial descriptions might use geographic entities (i.e., features like station, beach, city center, ...) to mention places. For example, in housing advertisements, geographic entities give more information about a neighbourhood and its facilities.

Another challenge of extracting spatial information is to create a structured knowledge base. Relationships, and more particularly spatial relationships, between extracted entities provide a structured representation of Geospatial information.

Applications and studies from various domains (geography, urban planning, real estate intelligence, ...) could result from the extracted knowledge. For example, spatial information could help to locate an entity—and, in particular, a property—since real estate advertisers do not always provide the exact location. Spatial and urban analyses may also be carried out about perception of residential spaces, valuable neighborhood factors or influential zone of a toponym.

In this work, we propose an automatic workflow for extracting Geospatial knowledge from text written in French. The workflow is divided in two main stages. First, we extend Named Entity Recognition to Spatial entities designed for French housing advertisements. Then, we develop a Relationship Extraction model to locate Spatial entities. The contributions of this paper may be summarized as follows:

- We define new Spatial entity categories to describe locations.
- We propose a Named Entity Recognition model to extract Toponyms and our new Spatial entities from French housing advertisements.
- We develop a Relationship Extraction model to create a structured Geospatial knowledge base.

The rest of the paper is organized as follows: Section 2 reviews related work on Geographic Named Entity Recognition and Relationship Extraction; Section 3 presents the methodological details of our automatic workflow for extracting Geospatial Knowledge; Section 4 presents and discusses the results of experiments based on French housing advertisements dataset; Finally, Section 5 draws some conclusions and outlines directions for further research.

K-CAP '21, December 2–3, 2021, Virtual Event, USA.

2 RELATED WORK

2.1 Geographic Named Entity Recognition

Geoparsing is the task to detect geographic terms from text and has been widely used in various types of texts such as travel blogs [1], social media in emergencies [7, 11], housing advertisements [9], or fictional novels [21]. This method is often a subtask of Named Entity Recognition (NER) applied to geographic entities. In [15] and [30], the authors reviewed traditional to state-of-the-art NER approaches. The traditional approach is based on linguistic rules and Knowledge systems and often uses Gazetteers. This approach is frequently used for Geoparsing [16, 21] but gives limited results and depends on the completeness of the rules and Gazetteers. In recent years, deep neural network models have been developed and achieved very good results on NER taks and particularly for Toponym Recognition [28, 29].

Most of the above-mentioned works only detect Toponyms in the English language. However, our research focused on all geographical and spatial terms in the French language, for which limited linguistic resources are available. In [19], the authors have defined and extracted French Spatial Nominal Entities that refer to physical objects in a spatial context (i.e., a geographic term that is not linked to a Toponym), which is similar to our proposal. Regarding the French language, [4] compared different deep neural network models to retrieve Named Entities applied to French Legal texts. They demonstrated that BiLSTM-CRF combined with text representations gives the best results for this task.

2.2 Relationship Extraction

Relationship extraction (RE) aims at extracting semantic relationships from a text, usually between two or more entities of a certain type. A large number of works use features of dependency-parsed sentences as extraction patterns or Machine Learning input [3, 6, 26]. Several methods have been developed from Dependency Graph such as shortest path dependency or frequent sub-graph. In [5], the authors have proven that the shortest path dependency kernel is very efficient in extracting binary relationships. [12] has extended this method to n-ary relationships. Another approach [8] is to find frequent sub-graph from the Dependency Graph to discover patterns. In a recent work [10], a Dependency Graph has also been used in a neural network. The idea is to embed several adjacency matrices to get a better dependency representation. Thus, a Graph Convolutional Network is applied to the Dependency Graph and then the new embedding feeds a classification model. The use of deep Learning methods is more and more frequent in the field of RE, as summarized in [22], and achieves state-of-the-art results. Nevertheless, neural networks need huge annotated corpora for training, and might not be effective for relationships occurring in just a few examples.

3 METHODOLOGY

3.1 Overview

In this paper, we propose a two-stage method to automatically retrieve spatial information in housing advertisements, mostly located in the French Riviera, and extract relationships to create a structured knowledge base.

The two main stages of the proposed method are:

- Perform Spatial Named Entity Recognition;
- Retrieve relationships between Spatial entities.

3.2 Spatial Information in Housing Advertisements

Housing advertisements have several advantages to extract spatial information as they generally provide information about a property's environment and location. Indeed, location is one of the most important factors in purchasing decisions. Thus, advertisers give a description of the neighbourhood and its facilities to promote the location to a future resident. Also, housing advertisements are a fairly exhaustive and updated source of data since many geographic areas are found and they are posted recurrently. Finally, advertisers often provide an online version that facilitates data collection (e.g., by crawling housing websites).

Thus, before retrieving information, we identified 4 spatial entity categories in housing advertisements that describe the neighbourhood of a property and spatial relationships:

- **Geographic entity**: entity representing natural features, constructions and subdivisions of land which is located on or near the surface of the earth.
- **Toponym**: entity referring to proper names of places (also known as place name or geographic name)
- **Spatio-Temporal entity**: entity localizing geographic entity or Toponym
- **Mode of transportation**: entity describing the travel mode between two places

The first two categories, Toponym and geographic entity, explicitly define the environment at different levels. Entity can refer to places in the residential area of a property, its neighbourhood or even the city. Toponym, or place name, is the most classic way to characterize a place and is found in many Named Entity Recognition frameworks. Also, Gazetteers, such as GeoNames, already record millions of place names around the world. On the other hand, a geographic entity also describes a place but is frequently associated with a Toponym (e.g., **University** of Nice, Nice-Riquier **station**, Masséna **square**, etc.). Thus, even if geographic entities are often mentioned in the literature, this category is never extracted. However, a geographic entity itself still gives information about a neighbourhood and its facilities. For example, we can extract the number of schools in an area or the proximity with public transport. Also, we can infer the exact place by cross-analysis with other databases and the spatial context.

After referencing places with Toponyms and geographic entities, we want to locate them to better understand their spatial context. Spatial relationships play a key role to link all those entities and a property in the space. However, a spatial relationship might be expressed in different ways (e.g., temporal or kilometer distance). Thus, we propose to create a Spatio-Temporal entity, which describes the spatial relationship between two places or a place and a property. It can be an exact or vague distance (e.g., near, 10 kilometers away, 5 minutes, etc.). Lastly, the Mode of transportation entity specifies the Spatio-Temporal entity (e.g., 5 minutes walking is different from 5 minutes driving) and can reduce its vagueness.

Table 1 shows several examples of those 4 types of entities.

Category	Examples	Count
Geographic Entity	- This apartment reveals a magnificent view over the **sea**. - Excellent location, close to the **local shops**, **Place** Masséna 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
Spatio-Temporal Entity	- 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 20 minutes **driving**	160

Table 1: Examples of the Spatial entity categories

3.3 Named Entity Recognition

The first stage of our pipeline is the Named Entity Recognition module, which extracts all the entities mentioned above. The proposed model is based on a *BiLSTM+CRF* architecture [13], which has achieved very good results on NER tasks. We also 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 1).

The first text representation is a basic Word Embedding architecture [20] that we trained on our corpus of French housing advertisements. Secondly, we fine-tuned the two pre-trained French Flair Language Models [2] with our corpus, which corresponds to a *BiLSTM* Language Model. This Language Model has a context-based and character-level representation that 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 [17], a French Transformer Language Model. *Transformers* are faster and more efficient than *BiLSTM* or *CNN* architectures, since they use parallel attention layers [27]. 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 those 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.

Housing advertisements have language and stylistic specificity and variability, and an informal format that are captured with the above three text representations.

3.4 Relationship Extraction

The previous stage identifies spatial entities thanks to a NER model. However, information is totally unstructured, whereas there exist relationships between entities in the text. The last part of our method tries to reconstruct relationships between spatial entities and provide a structured representation of extracted information. We determined four types of relationships to get a structured knowledge base:

- **Attributes**: attributes of spatial entities such as adjectives, numerical, noun object (e.g., **5** minutes, **residential** neighbourhood, etc.);
- **Geographic Composition**: affiliation link between geographic entity and/or Toponym (e.g., Nice City Center, City of Cannes, Audiberti High School, etc.);
- **Spatial**: relationship between Spatio-Temporal entity and geographic entity, Toponym or the property itself;
- **Mode of transportation**: relationship between Spatio-Temporal entity and Mode of transportation.

We made several assumptions to perform the relation extraction. The first is that a relationship occurs only between two entities of the same sentence. Then, we assumed that a direct or indirect connection between the two entities always exists in the sentence, which can be captured by a dependency graph. Our last assumption is that words between the two entities also play an important role.

Thus, we highlighted two types of features to recognize the relationships:

- Shortest path dependency between entities;
- Sub-phrase between entities.

3.4.1 Shortest-Path Dependency. In a sentence, the syntactical structure based on grammar gives relationships between words as a tree. Two methods are frequently used: Constituency and Dependency parsing. On the one hand, Constituency parsing uses the formalism of context-free grammars and the sentence is divided in sub-phrases that belong to the same grammatical category. On the other hand, Dependency parsing directly assigns grammatical connections between words. A relation is binary and consists of a head word, a dependent word and a grammatical function. This approach is more suitable for analysing housing advertisements, since the order of words does not always follow classical grammar (e.g., housing 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 [24] is general enough to easily capture dependencies in housing advertisements and can be adapted to another language. As we did not have a labeled corpus available, we performed dependency parsing using Stanza [25] for French, which is based on the Universal Dependencies taxonomy and is already trained on a large corpus.

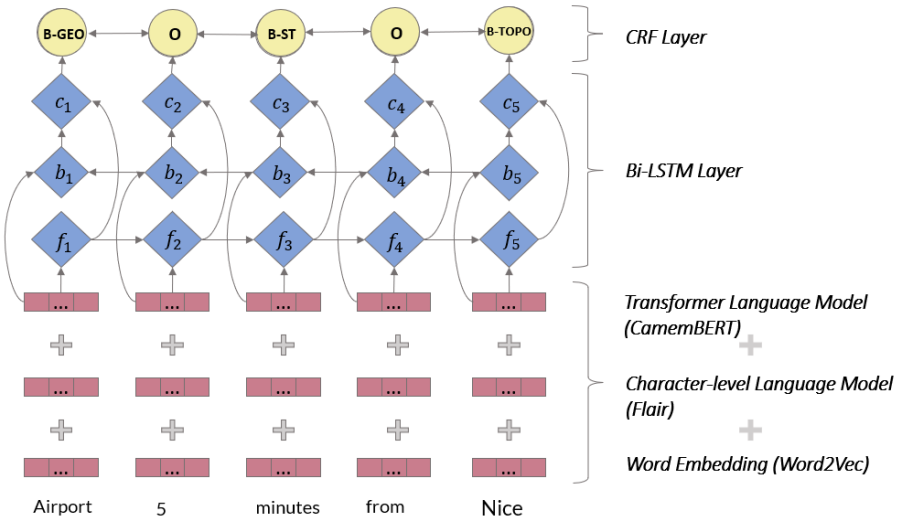

Figure 1: Overall NER architecture

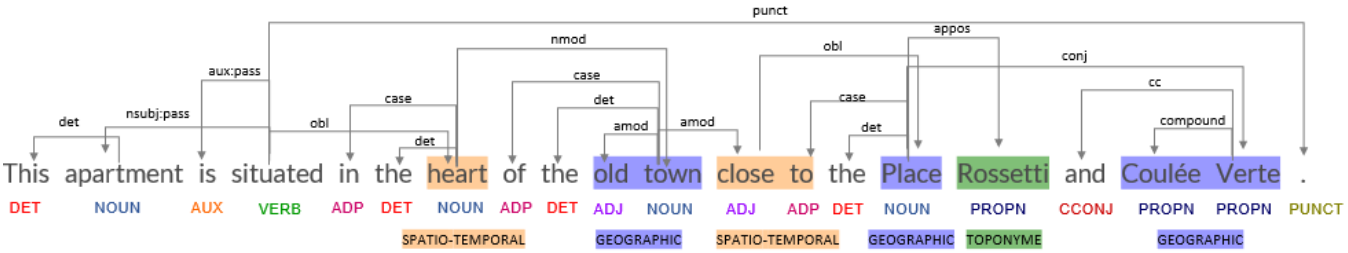

Figure 2: Example of tagged Sentence

Entities	Type of Relation	Shortest path dependency	Sub-phrase
(heart, old town)	Spatial	SPATIO-TEMPORAL ↓ nmod GEOGRAPHIC	(heart, of, the, old town)
(old town, close to)	Spatial	GEOGRAPHIC ↓ amod SPATIO-TEMPORAL	(old town, close to)
(close to, Place)	Spatial	SPATIO-TEMPORAL ↓ obl GEOGRAPHIC	(close to, the, Place)
(close to, Coulée Verte)	Spatial	SPATIO-TEMPORAL ↓ obl GEOGRAPHIC ↓ conj GEOGRAPHIC	(close to, the, Place, Rossetti, and, Coulée Verte)
(Place, Rossetti)	Geograpihc Composition	GEOGRAPHIC ↓ appos TOPONYM	(Place, Rossetti)

Table 2: Features of candidate relationships

However, this model was not very accurate, especially with the Part-of-Speech tagging module. Thus, we trained a Part-of-Speech tagger by first predicting labels with an existing tagger and then correcting predictions to get a labeled data set.

After carrying out dependency parsing, we got a dependency graph of each sentence and extracted the shortest directed path for each candidate relationship. The path is composed of nodes that represent entities or Part-of-Speech tags and edges corresponding to grammatical relations.

Figure 2 shows Spatial Named entities, Part-of-Speech tags and Dependency parsing in a sentence. Two type of relationships occur in this example: spatial and geographic composition. To find a spatial relationship, the shortest-path dependency is extracted between a Toponym or Geographic entity and a Spatio-Temporal entity. For geographic composition relationship, the aim is to find a path between two Geographic entities, two Toponyms or a Geographic entity and a Toponym. All the candidate relationships and their shortest-path dependencies are summarized in Table 2.

3.4.2 Classification. After extracting all the candidate relationships thanks to the shortest path, we built four classification models, one for each type of relation. Those models take as input the sub-phrase between two entities and their shortest-path dependency. We transformed both inputs into vectors with an embedding model. For the sub-phrase, we used the same embedding as for the Named Entity Recognition task.

Regarding the shortest-path dependency, we chose to model the path as a sequence of entities (Named Entities or Part-of-Speech tags) and grammatical relations. Thus, we embedded each sequence based on the Sqn2Vec method [23]. This method is unsupervised and designed for small-vocabulary sequential datasets. It combines two techniques: Sequential Pattern Mining and Neural Embedding Learning. The idea is to first extract all the Sequential Patterns from the dataset and then to learn an embedding for each sequence. The embedding is a concatenation of two representations. The former is only based on predicting which single symbols are in the sequence while the latter predicts to which Sequential Pattern the sequence belongs to. Both predictions are based on a Paragraph Vector-Distributed Bag-of-Words [14].

We could have used a graph embedding model [10], since dependency relationships are represented as a graph. Graph embedding is often based on the adjacency matrix, which embeds either nodes or edges. However, we noticed that the entities (i.e., nodes) and the grammatical relations (i.e., edges) are both important in the shortest-path dependency. Thus, we needed an embedding that takes both into account.

3.4.3 Output. The final step is focused on the representation of the extracted relationships in a structured knowledge base. Figure 3 shows a graph representation of the previous example. Nodes represent Spatial entities and edges the type of relationship. Regarding spatial relationships, Spatio-Temporal entities (*at the heart* and *close to*) should describe the relationship between two places or a place and a property. Thus, spatial relationship can be viewed as a ternary relationship with a place or a property as subject and/or object, and a Spatio-Temporal entity as attribute of the predicate. In this example, the subject is missing, as we have only detected a link between a Spatio-Temporal entity and one geographic entity. If (and only if) the subject is missing, we always assume that the property being advertised is the subject.

Finally, a post-processing step should be applied to get ternary and binary relationships in order to create a Geospatial Knowledge Graph, whose description is out of the scope of this work.

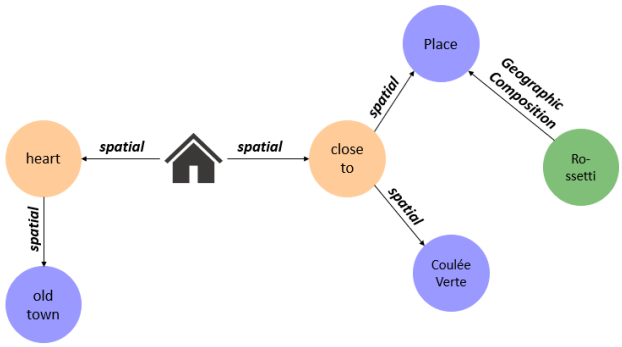

Figure 3: Graph representation of the output

4 RESULTS AND DISCUSSIONS

In this section, we apply the proposed two-stage workflow to a dataset of French housing advertisements in the French Riviera. We first describe the dataset and then present the results and the performance of the workflow.

4.1 Dataset

As a standard French dataset is not available to train and evaluate our method, we collected about 1,200 housing advertisements[1] written in French gathered from various online advertisers on the French Riviera. The average length of the ads is about 650 words after a preprocessing stage. Cleaning the text was necessary since the French house ads were full of noisy, repetitive words and abbreviations. We also removed new lines, URLs, special symbols and characters (e.g., &, #, *) using regular expressions. Then, we manually annotated the dataset following the BIESO tagging scheme. Finally, we split the dataset into 10 folds and evaluated our pipeline through cross-validation.

4.2 Performance Evaluation

In this subsection, we evaluate the performance of the two stages of our workflow. We start by comparing several architectures for Named Entity Recognition, trained on our annotated spatial entities. We chose the Spacy and Bi-LSTM-CRF architectures as they do not use the same type of neural networks (CNN and RNN respectively). We could not compare them with off-the-shelf models, as they do not extract the same Named Entities. Regarding Spacy, we fine-tuned a pre-trained model for French with our entities and trained another from scratch. For the second architecture, we tried different embeddings and combinations to handle French and the language specificity of housing advertisements. We fine-tuned the Flair embedding and trained a Word2Vec model.

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

	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: Performance of NER models.

To quantify the performance, we employed, as it is usual in tasks like these, Precision, defined as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{1}$$

Recall, defined as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}, \tag{2}$$

and F1-score, defined as

$$\text{F1-Score} = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}. \tag{3}$$

We also computed a Welch's t-test, with t defined as

$$t = \frac{\overline{X_i} - \overline{X_j}}{\sqrt{s^2_{\overline{X}_i} + s^2_{\overline{X}_j}}}, \tag{4}$$

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.

In Table 3, we present average Precision, Recall, F1-Score and their standard deviation, based on the 10-fold cross-validation. Table 4 shows the models that have weak or no evidence of a significantly different F1-Score.

First, we notice that *BiLSTM+CRF* architecture mostly outperforms Spacy architectures except with Word2Vec embedding at a 1% significance level. Spacy architecture gets slightly better results with a model trained from scratch. Regarding *BiLSTM+CRF* models with a single embedding, CamemBERT achieves very good performance. Flair gets a similar Recall to CamemBERT but the Precision is lower. We can underline that the combination of Flair and CamemBERT increases the performance of the 3 metrics. The combination of Flair and Word2Vec or CamemBERT and Word2Vec increases the Precision but decreases the Recall. Moreover, there is no evidence of statistically significant differences between those models. Finally, the combination of the three representation (Flair + CamemBERT + Word2Vec) achieves the best results while the Welch's t-test does not show difference with Flair + CamemBERT. Nevertheless, we chose to keep this model for our application.

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 4: Level significance > 1%

The second evaluation consists of estimating the performance of each relationship classification. We created four datasets with candidates for each type of relationships and tagged them with a binary label. We tried different classification models such as Naive Bayes, Support Vector Machine or XGBoost, and quantified their performance with Accuracy defined as

$$\text{Accuracy} = \frac{\#(Labeled \cap Recognized)}{\#Labeled}. \tag{5}$$

In Table 5, we present the best performance of each classification using an XGBoost model. All classifications get an accuracy above 95% and even 97% for three of them. However, those results might be biased because of the small dataset for some relationships and imbalanced classes. Especially, Mode of Transportation relationship has been evaluated on a very small dataset (216 rows) and totally imbalanced classes (209 label 1 and 7 label 0). The other three relationships are more accurate since their dataset contains more than 3,000 examples for Geographic Composition and Spatial and 6,000 examples for Attributes.

Relations	Accuracy
Attributes	97.1%
Geographic Composition	97.7%
Spatial	95.7%
Mode of Transportation	97.8%

Table 5: Performance of Relationship Extraction models

All in all, our proposed workflow performs well to retrieve Spatial entities and relationships. Nevertheless, we separately evaluated

each model which does not represent the global performance of the pipeline. Indeed, the Relationship Extraction stage depends on the good results of Named Entity Recognition model. Thus, the error is propagated across the model and should be quantified with a fully annotated dataset with both Spatial entities and relationships.

5 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a two-stage method to automatically retrieve Geospatial information from housing advertisements. A major contribution is the recognition of Geographic and Spatio-Temporal entity to better understand the spatial context. Also, the Named Entity Recognition model is fine-grained to discover new Toponyms that would have not been found with off-the-shelf NER models. Another contribution is the Relationship Extraction tool and in particular the Spatial relationship. We have proposed to build binary relationship between Spatio-Temporal entities and Toponyms, Geographic entities or the property to easily capture a ternary Spatial relationship in a post-processing part. Finally, our method is designed for French language and housing advertisements but can be adapted to other languages and types of text by annotating a corpus, changing the embedding and retraining the model. In addition, some domain- and task-specific heuristics, such as assuming that an implicit subject is the property being advertised, would have to be dispensed with or replaced when adapting our method to other tasks or domains.

Several directions could be considered to expand this work. First, Entity Recognition is only the first step of Information Extraction, and adding Entity Resolution would help to reduce ambiguity of Toponym and Geographic entity. A number of Toponym Resolution models already exist but may not be efficient for Geographic entity. Also, it would be interesting to link Toponym and Geographic entities to existing Knowledge bases (e.g GeoNames, DBpedia, Wikidata). Second, we have extracted Spatio-Temporal entities in order to build Spatial relationships; however the Real Estate market often exaggerates proximity-related terms and introduces uncertainty and vagueness in their advertisements [18]. A promising work would be to qualify and quantify the uncertainty and vagueness of those terms to improve the reliability of the location of the mentioned places. Third, the extracted information combined with their socio-spatial context may be analysed and studied within urban space researches (e.g., perception of residential spaces, elements of a neighbourhood, valuable neighborhood factor).

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] Benjamin Adams and Krzysztof Janowicz. 2012. On the Geo-Indicativeness of Non-Georeferenced Text. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci (Eds.). The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4629

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.

[3] Alan Akbik, Thilo Michael, and Christoph Boden. 2014. Exploratory Relation Extraction in Large Text Corpora. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, Jan Hajic and Junichi Tsujii (Eds.). ACL, 2087–2096. https://aclanthology.org/C14-1197/

[4] Valentin Barrière and Amaury Fouret. 2019. May I Check Again? - A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa 2019, Turku, Finland, September 30 - October 2, 2019*, Mareike Hartmann and Barbara Plank (Eds.). Linköping University Electronic Press, 327–332. https://aclweb.org/anthology/W19-6136/

[5] Razvan C. Bunescu and Raymond J. Mooney. 2005. A Shortest Path Dependency Kernel for Relation Extraction. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*. The Association for Computational Linguistics, 724–731. https://aclanthology.org/H05-1091/

[6] Aron Culotta and Jeffrey Sorensen. 2004. Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 423–429. https://doi.org/10.3115/1218955.1219009

[7] Rob Grace. 2021. Toponym usage in social media in emergencies. *International Journal of Disaster Risk Reduction* 52 (2021), 101923. https://doi.org/10.1016/j.ijdrr.2020.101923

[8] Mohsen Hassan, Adrien Coulet, and Yannick Toussaint. 2014. Learning Subgraph Patterns from text for Extracting Disease - Symptom Relationships. In *Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing co-located with The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, DMNLP@PKDD/ECML 2014, Nancy, France, September 15, 2014 (CEUR Workshop Proceedings, Vol. 1202)*, Peggy Cellier, Thierry Charnois, Andreas Hotho, Stan Matwin, Marie-Francine Moens, and Yannick Toussaint (Eds.). CEUR-WS.org, 81–96. http://ceur-ws.org/Vol-1202/paper6.pdf

[9] Yingjie Hu, Huina Mao, and Grant McKenzie. 2019. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *Int. J. Geogr. Inf. Sci.* 33, 4 (2019), 714–738. https://doi.org/10.1080/13658816.2018.1458986

[10] Yanfeng Hu, Hong Shen, Wuling Liu, Fei Min, Xue Qiao, and Kangrong Jin. 2021. A Graph Convolutional Network With Multiple Dependency Representations for Relation Extraction. *IEEE Access* 9 (2021), 81575–81587. https://doi.org/10.1109/ACCESS.2021.3086480

[11] Yingjie Hu and Jimin Wang. 2021. How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey. In *11th International Conference on Geographic Information Science, GIScience 2021, September 27-30, 2021, Poznań, Poland - Part I (LIPIcs, Vol. 177)*, Krzysztof Janowicz and Judith Anne Verstegen (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 6:1–6:16. https://doi.org/10.4230/LIPIcs.GIScience.2021.I.6

[12] Gitansh Khirbat, Jianzhong Qi, and Rui Zhang. 2016. N-ary Biographical Relation Extraction using Shortest Path Dependencies. In *Proceedings of the Australasian Language Technology Association Workshop 2016, Melbourne, Australia, December 5 - 7, 2016*, Trevor Cohn (Ed.). ACL, 74–83. https://aclanthology.org/U16-1008/

[13] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 260–270. https://doi.org/10.18653/v1/n16-1030

[14] Quoc V. Le and Tomás Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings, Vol. 32)*. JMLR.org, 1188–1196. http://proceedings.mlr.press/v32/le14.html

[15] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A Survey on Deep Learning for Named Entity Recognition. *CoRR* abs/1812.09449 (2018). arXiv:1812.09449 http://arxiv.org/abs/1812.09449

[16] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, Feifei Li, Mirella M. Moro, Shahram Ghandeharizadeh, Jayant R. Haritsa, Gerhard Weikum, Michael J. Carey, Fabio Casati, Edward Y. Chang, Ioana Manolescu, Sharad Mehrotra, Umeshwar Dayal, and Vassilis J. Tsotras (Eds.). IEEE Computer Society, 201–212. https://doi.org/10.1109/ICDE.2010.5447903

[17] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics. Association for Computational Linguistics, Online, 7203–7219. https://www.aclweb.org/anthology/2020.acl-main.645

[18] Grant McKenzie and Yingjie Hu. 2017. The "Nearby" exaggeration in real estate. In *Proceedings of the Workshop on Cognitive Scales of Spatial Information*.

[19] Amine Medad, Mauro Gaio, Ludovic Moncla, Sébastien Mustière, and Yannick Le Nir. 2020. Comparing supervised learning algorithms for Spatial Nominal Entity recognition. *AGILE: GIScience Series* 1 (2020), 15. https://doi.org/10.5194/agile-giss-1-15-2020

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[21] Ludovic Moncla, Mauro Gaio, Thierry Joliveau, and Yves-François Le Lay. 2017. Automated Geoparsing of Paris Street Names in 19th Century Novels. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, GeoHumanities@SIGSPATIAL 2017, Redondo Beach, CA, USA, November 7-10, 2017*. ACM, 1–8. https://doi.org/10.1145/3149858.3149859

[22] Tapas Nayak, Navonil Majumder, Pawan Goyal, and Soujanya Poria. 2021. Deep Neural Approaches to Relation Triplets Extraction: A Comprehensive Survey. *CoRR* abs/2103.16929 (2021). arXiv:2103.16929 https://arxiv.org/abs/2103.16929

[23] Dang Nguyen, Wei Luo, Tu Dinh Nguyen, Svetha Venkatesh, and Dinh Q. Phung. 2018. Sqn2Vec: Learning Sequence Representation via Sequential Patterns with a Gap Constraint. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11052)*, Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim (Eds.). Springer, 569–584. https://doi.org/10.1007/978-3-030-10928-8_34

[24] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 1659–1666. https://aclanthology.org/L16-1262

[25] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[26] Hans Uszkoreit. 2011. Learning Relation Extraction Grammars with Minimal Human Intervention: Strategy, Results, Insights and Plans. In *CICLing*.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[28] Jimin Wang, Yingjie Hu, and Kenneth Joseph. 2020. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Trans. GIS* 24, 3 (2020), 719–735. https://doi.org/10.1111/tgis.12627

[29] Xiaobin Wang, Chunping Ma, Chu Zheng, Huafei andLiu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM_NLP at SemEval-2018 Task 12: A Pipeline System for Toponym Resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 917–923. https://doi.org/10.18653/v1/S19-2156

[30] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 2145–2158. https://www.aclweb.org/anthology/C18-1182/