

GeospaCy: A tool for extraction and geographical referencing of spatial expressions in textual data

Mehtab Alam SYED

CIRAD, Montpellier, France
mehtab-alam.syed@cirad.fr

Elena ARSEVSKA

CIRAD, Montpellier, France
elena.arsevska@cirad.fr

Mathieu ROCHE

CIRAD, Montpellier, France
mathieu.roche@cirad.fr

Maguelonne TEISSEIRE

INRAE, Montpellier, France
maguelonne.teisseire@inrae.fr

Abstract

Spatial information in text enables to understand the geographical context and relationships within text for better decision-making across various domains such as disease surveillance, disaster management and other location-based services. Therefore, it is crucial to understand the precise geographical context for location-sensitive applications. In response to this necessity, we introduce the *GeospaCy* software tool, designed for the extraction and georeferencing of spatial information present in textual data. *GeospaCy* fulfills two primary objectives: 1) Geoparsing, which involves extracting spatial expressions, encompassing place names and associated spatial relations within the text data, and 2) Geocoding, which facilitates the assignment of geographical coordinates to the spatial expressions extracted during the Geoparsing task. Geoparsing is evaluated with a disease news article dataset consisting of event information, whereas a qualitative evaluation of geographical coordinates (polygons/geometries) of spatial expressions is performed by end-users for Geocoding task.

keywords: Geoparsing, Geocoding, Spatial Expressions, Natural Language Processing

1 Introduction

In recent years, spatial information recognition from textual data has gained more attention in the natural language processing (NLP) field. The importance and relevance of the work can be strengthened by highlighting the potential impact and benefits of accurate spatial information extraction in various domains, i.e., disaster management, disease surveillance. For instance, in disease surveillance, a disease outbreak in 'central Paris' is not the same as one in the 'southern part of Paris'. Moreover, the extraction and georeferencing of spatial information have significant implications in various domains, including healthcare,

financial markets, and e-learning (Hassani et al., 2020). Therefore, the extraction and interpretation of spatial information from textual data play a fundamental role in understanding geographical contexts.

The spatial information can be expressed in the textual documents in both simple and complex ways, depending on the syntax and semantic of expression. This geospatial information is available in the form of absolute spatial information (precise location names, e.g., Milan) and relative spatial information (spatial relations associated with the location name, e.g., North Milan). Both absolute and relative spatial information are essential in providing accurate context for locations in the text, ensuring precision in understanding and responding specific to geographical sensitive applications. While absolute spatial data offers concrete locations, relative spatial information provides contextual references that help to refine and to detail the specific area of interest, resulting into more accurate geographical reference in the text. Therefore, a possible research question is: "Can we develop an efficient and accurate algorithm for extracting spatial relations from textual data and transforming them into valid geospatial representations"?

Traditional methods of text mining often overlook important geographical details by ignoring the complex spatial information found within the text. The motivation behind the development of **GeospaCy** is to overcome this limitation and provide a robust tool specifically tailored to identify and georeference spatial expressions in textual data. The main purpose of *GeospaCy* software tool is to address the demand of precise geographical insights, which are essential for making informed decisions in various domains such as disease surveillance, disaster management, and other location-based services. *GeospaCy* performs two main tasks, i.e., 1) Geoparsing and 2) Geocoding. Geoparsing within the context of the *GeospaCy* tool involves

the identification and extraction of spatial expressions embedded within unstructured textual data. This task primarily revolves around recognizing spatial expressions such as place names, and spatial relations associated with the place names from the text. Geoparsing task provide a foundation for subsequent geocoding to understand the geographical context of the textual data. In contrast, Geocoding within the GeospaCy tool represents the process of assigning precise geographical coordinates to the spatial expressions identified during the geoparsing task. After the spatial information such as place names or locations have been extracted, geocoding works to convert these textual references into geographical coordinates.

The remainder of this article is structured as follows: Section 2 provides the related work associated to *GeospaCy*. Subsequently, Section 3 describes the software overview, Section 4 briefly details the methodology, Section 5.2 explains the real world use cases and Section 6 presents the conclusion.

2 Related Work

Different research studies have been carried out for both Geoparsing and Geocoding process. The details of the related work are discussed in subsequent sections.

2.1 Geoparsing

Numerous research studies have been carried out with diverse approaches enhancing Geoparsing that revolves around the extraction of spatial information from unstructured text. These Geoparsing approaches include i.e., rule-based approaches, machine learning, ontology-based reasoning, geographical databases and transformer-based language models (Kokla and Guillet, 2020; Alonso Casero, 2021). In a research study carried out, a rule-based named-entity recognition method was proposed to address specific cases involving spatial named entities in textual data. This approach was validated using historical corpora (McDonough et al., 2019). However, the proposed approach did not address the complex relationship that involves other linguistic features, i.e. part-of-speech (POS), dependency parsing, word vectors etc. In another research (Chen et al., 2017), a best-matched approach is proposed to extract geospatial relations that are referred to anchor places, gazetteered places, and non-gazetteered places.

However, it is not defined in the coordinate system to be represented in geographical systems. A further research proposed a voting approach (SPENS) to extract place names through five different system including Stanford NER, Polyglot NER, Edinburgh Geoparser, NER-Tagger, and spaCy (Won et al., 2018). Another research combine multiple features that capture the similarity between candidate disambiguations, the place references, and the context where the place references occurs, in order to disambiguate place among a set of places around the world (Santos et al., 2015). Furthermore, another research (Medad et al., 2020) proposed an approach that is the combination of transfer learning and supervised learning algorithm for the identification of spatial nominal entities. However, the scope of the work was limited to the spatial entities without proper nouns e.g. conferences, bridge at the west, summit, etc. Afterwards, another research (Wu et al., 2022) proposed deep learning models i.e., CasREL and PURE in order to extract geospatial relations in the text. The proposed models were validated with two main approaches, i.e., 1) spatial entities and relations were dealt separately and joint approach. The quantitative results demonstrated that pipeline approach performed better than joint approach using deep learning models. Another research (Zheng et al., 2022) proposed a knowledge-based system (GeoKG) that described geographic concepts, entities, and their relations in order to search through queries. The system is used for geological problem solution and their decision-making. However, the solution is only limited to the geological domain that contains information about geographical events, geographical relationships and concepts. Another research proposed an approach for extracting place names from tweets, named GazPNE2 by combining global gazetteers (i.e., OpenStreetMap and GeoNames) to train deep learning, and pretrained transformer models i.e. BERT (Hu et al., 2022). The extracted place names taken coarse (e.g., city) along with fine-grained (e.g., street and POI) levels and place names with abbreviations. Moreover, recent advancements have introduced the UniversalNER model with more entity types, demonstrating remarkable NER accuracy across various domains, including healthcare, biomedicine, and others (Zhou et al., 2023).

2.2 Geocoding

Diverse research studies have been carried out about geocoding methodologies with the primary

objective of transforming toponyms, which are place names or location references in text, into precise geographical coordinates (Gritta et al., 2018). Mostly, geocoding methods rely on address matching, where textual toponyms are compared to a database of known addresses to retrieve latitude and longitude information (Behr, 2010). In a research study carried out, an unsupervised geocoding algorithm is proposed by taking leverage of clustering techniques to disambiguate toponyms extracted from gazetteers and estimate the spatial footprints of fine-grain toponyms that are not present in gazetteers (Moncla, 2015). A further research proposed a system that extracts place names from text, resolves them to their correct entries in a gazetteer, and returns structured geographic information for the resolved place name (Halterman, 2017). The system can be used for various tasks including media monitoring, improved information extraction, document annotation, and geolocating text-derived events. Further research proposed a geotagging algorithm constructed a model in which they used DBpedia-based entity recognition for places disambiguation, and Geonames gazetteer and Google Geocoder API for resolution of geographical coordinates of locations (Middleton et al., 2018). One more research introduced a deep neural network that incorporates Long Short-Term Memory (LSTM) units (Fize et al., 2021). The approach was focused on modelling pairs of toponyms, where the first input toponym is geocoded based on the context provided by the second toponym. The approach effectively reduced contextual ambiguities and generates precise geographical coordinates as output. A further research proposed a representational framework that employed rules, semantic approximations, background knowledge, and fuzzy linguistic variables to geocode imprecise and ad-hoc location referents in terms of fuzzy spatial extents as opposite to atomic gazetteer toponyms (Al-Olimat et al., 2019). Additionally, geocoding services and APIs offered by technology companies and government agencies have become increasingly accessible, providing convenient and efficient solutions for geocoding tasks (Longley and Cheshire, 2017). However, there is no existing Geocoding service or method to convert the extracted toponyms associated with spatial relations into geographical coordinates.

3 GeospaCy Overview

GeospaCy have the capabilities to precisely extract and reference spatial information within unstructured textual data. The main purpose of this software tool is to provide a comprehensive understanding of geographical contexts within textual information. This software tool has a set of features and functionalities which are as follows:

3.1 Geoparsing

In GeospaCy, we extract three kinds of spatial entities, i.e., 1) Geopolitical entities (GPE) i.e., place names e.g., Paris, Lyon etc, 2) location entities (LOC) i.e., physical locations e.g. Alpe d'Huez and 3) spatial relation entity (RSE) i.e., spatial relations associated with place names e.g., nearby Paris, south of Montpellier etc. ‘GPE’ and ‘LOC’ are extracted through state-of-the-art NER approach, whereas ‘RSE’ extraction is the main contribution of this software tool. We further categorized RSE into four main categories i.e., Level-1, Level-2, Level-3 and Compound RSE. *Level-1 RSE* is cardinal/ordinal associated with place names, *Level-2 RSE* is spatial keywords (nearby, border, neighbourhood) associated with place names, *Level-3 RSE* is distance keywords (1 km radius, 2 miles) associated with place name and compound is the combination of Level-1, Level-2 and Level-3 combination.

3.2 Geocoding

The geocoding process, to identify the specific geographic locations of entities, is acquired using the Nominatim API (Clemens, 2015). The coordinates of the GPE and LOC are directly obtained through this API. Nevertheless, the coordinates of RSE are determined differently. An algorithm developed for establishing spatial relationships processes place name coordinates, retrieved from the Nominatim API, to compute the coordinates of RSE entities. The main contribution of geocoding process is the computation of geographical coordinates of RSE. The output coordinates after geocoding is visualized on OpenStreetMap (OpenStreetMap contributors, 2017) leaflet or downloaded as GeoJson format. Figure 1 shows the overview of the GeospaCy software tool.

4 Methodology & Implementation

Our methodology (Syed et al., 2023) is divided into two main phases: 1) Extraction phase (Geoparsing), 2) Geocoding phase respectively. The process

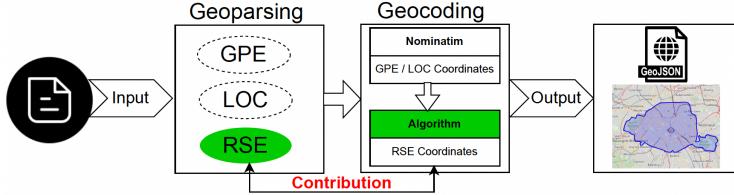


Figure 1: GeospaCy Overview

workflow of extraction of RSE and its georeferencing is shown in Figure 2. The details of the two phases of are explained in the subsequent sections.

4.1 Extraction Phase

In the first phase, RSE are extracted from the text data. We selected state-of-the-art natural language processing (NLP) library *spaCy* (Honnibal and Montani, 2017) for python. *spaCy* has a better performance for NER tasks as compared to other NLP libraries (Vajjala and Balasubramanian, 2022). The *spaCy* NER pipeline is customized for recognition of RSE label entity types. The steps for the customization of *spaCy* NER pipeline to recognize RSE are as follows:

Model Selection: *GeospaCy* offers three linguistic models, i.e., `en_core_web_sm`, `en_core_web_md`, `en_core_web_lg` and `en_core_web_trf`. The `en_core_web_trf` model is computationally expensive compared to smaller models, and requires significant computational resources to run. However, its high performance and accuracy make it a popular choice for a wide range of NLP applications. After the selection of a linguistic model, the environment is set up for the NLP NER task.

Apply NER: The next step in extraction phase is to apply NER on the textual data. We recognized spatial entities from the textual data with the labels ‘GPE’ e.g., Paris and ‘LOC’ e.g., Safari Desert respectively. To identify the RSE, we extract the clauses that contain the spatial entities in the text.

Clause Extraction: We split the sentence into clauses and save the clause that contains the spatial entity (GPE or LOC) and ignore the rest of the clauses in the sentence.

Spatial Relations Identification: We extract spatial relations from the candidate clauses. Candidate clauses are identified in the text document as the clauses that contain GPE/LOC. In order to extract RSE in the clauses, we defined regular expressions for Level-1, Level-2, Level-3 RSE. The regular ex-

pressions of these geospatial relations are defined using *Python* regex *re* with the help of *Python* library *quantities*. *quantities* library is used to get the different quantity units, its abbreviations and their interconversions. If spatial relations are identified in the clause that contained the GPE/LOC, then we adjust the span offset according to the spatial relation. The span offset is either adjusted from the end or in the start according to the occurrence of spatial relation relative to GPE/LOC.

RSE Spans Extraction: We further identify the spatial relations clauses and made the compatible span having in the NER linguistic pipeline.

RSE Injection: The GPE/LOC having spatial relation in the clause is replaced with RSE span in the ‘DOC’ (the element that contains linguistic feature information) element of *spaCy* NER pipeline. The label of the RSE are injected in the ‘DOC’ element as ‘RSE’.

4.2 Geocoding Phase

In this phase, the translation of geographical coordinates is derived either by slicing the polygon or by deriving using geospatial operations. The steps involved to extract the geographical coordinates of RSE are as follows:

Acquire coordinates from Nominatim: Nominatim API (Clemens, 2015) provides search by place name, feature description or free text search in OpenStreetMap (OpenStreetMap contributors, 2017) database and return its geographical coordinates based on search queries. The API provides the *GeoJSON* which contains the geometry along with their feature attributes. The coordinates of RSE are further determined from the place name (GPE/LOC) coordinates.

Derive/Slice RSE Coordinates: The next step is to derive the coordinates of the RSE. Slicing depends on the type of RSE. Level-1 (cardinal/ordinal) RSE coordinates are acquired by slicing the main geometry of the place into 9 RSE geometries.

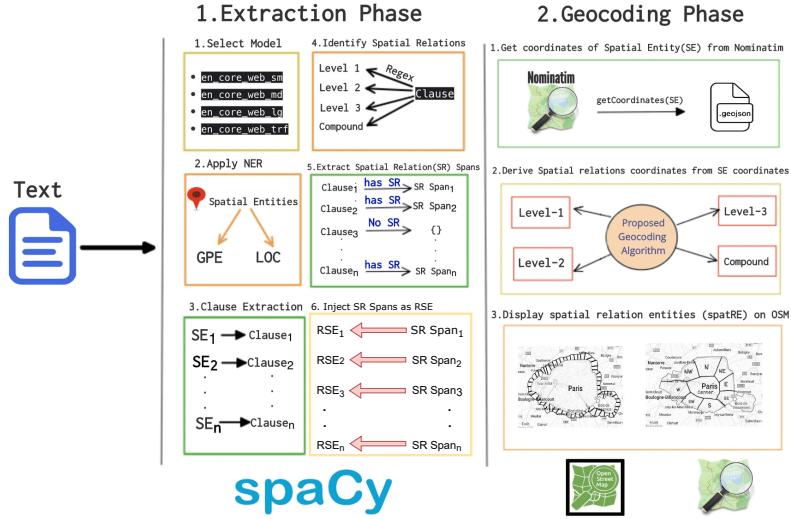


Figure 2: GeospaCy: RSE Extraction and Georeferencing Pipeline

For instance, The Level-1 slicing of Paris can be sliced into 9 geographical shapes: ‘Northern Paris’, ‘Southern Paris’, ‘Eastern Paris’ etc. In contrast to Level-1 RSE, Level-2, Level-3 and compound RSE are derived by applying spatial operations i.e. spatial joins, spatial unions, intersections by using *GeoPandas* (Jordahl et al., 2020) and *Shapely* (Gillies et al., 2007) Python libraries.

RSE output: The RSE output coordinates can be downloaded as GeoJson file or visualize on OpenStreetMap leaflet.

5 Experiments

GeospaCy results are evaluated for each phase of the software. These two main phases are: 1) Geoparsing i.e., the extraction of RSE from text and 2) Geocoding i.e., the computed geographical coordinates for RSE. The evaluation of these phases are as follows:

5.1 Extraction and Geocoding Evaluation

The *extraction phase* focused on extraction of RSE in unstructured text, which is then evaluated using a dataset related to disease surveillance. The dataset contains the news extracted by PADI-web¹, which is an event-based surveillance system related to animal health events. The dataset contains the news articles of different diseases i.e., 1) *Antimicrobial Resistance (AMR)* 2) *COVID-19*, 3) *Avian-Influenza*, 4) *Lyme* and 5) *Tick-borne Encephalitis*

(TBE) with manually annotated RSE. Precision, recall and F-Score are calculated for the RSE. The RSE recognition task have a precision of **0.9**, recall of **0.88** and F-Score of **0.88**. The detail of the evaluation are available in Table 1 of the Section A.

GeospaCy calculated the geographical coordinates of RSE. These coordinates were computed and evaluated for cities such as Paris, London, Milan, Madrid, Zagreb, Utrecht, Delft, Lyon, and Florence. For each city, 19 RSE shapes were assessed through qualitative evaluation by end-users, resulting in an average accuracy of **75%**. The detail of the evaluation are available in Table 2 of the Section A.

5.2 Use Cases

GeospaCy can be a useful tool in different use cases, including disease Surveillance, disaster Response management, environmental geographical analysis and other geographical sensitive applications etc. The detail of some use cases associated with disease surveillance are as follows:

Disease Surveillance: In the field of public health (PH) and animal health (AH), health professionals often deal with unstructured textual data from various sources, such as reports, articles, or social media, containing vital information about disease outbreaks, symptoms, and affected areas. *GeospaCy* can parse and extract spatial expressions from these texts, identifying affected regions, hotspots, and areas prone to outbreaks. An example of African Swine Fever (ASF) outbreak with RSE location by

¹<https://padi-web.cirad.fr/en/>

Enter the text to extract [Spatial Entities]

African Swine Fever (ASF) has been detected in a sample from 7 wild boar found just south-east of Fagersta

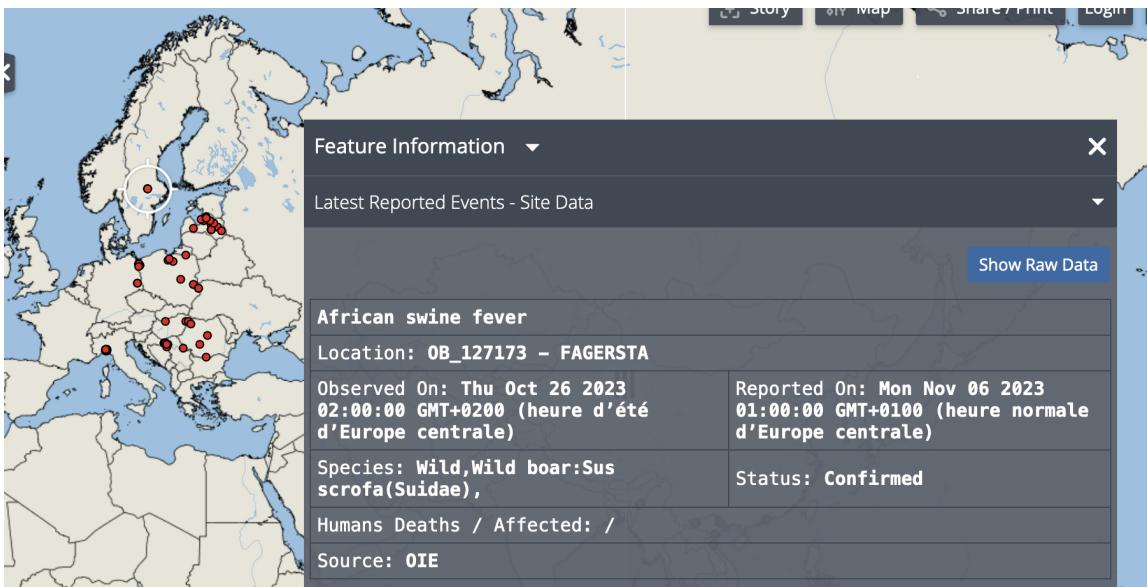
Extract

African Swine Fever (ASF) has been detected in a sample from 7 wild boar found just **south-east of** **Fagersta RSE**

(a) GeospaCy detected outbreak RSE in the text: **South-east of Fagersta**



(b) GeospaCy detected outbreak location coordinates of **South-east of Fagersta**



(c) Official Outbreak detected by Empress-i/OIE with location **Fagersta**

Figure 3: An African Swine Fever (ASF) outbreak in South-east of Fagersta, Sweden

GeospaCy and official outbreak by Empress-i are as follows:

ASF: African Swine Fever (ASF) has been detected in a sample from 7 wild boar found just south-east of Fagersta²

The provided text highlights about “an African Swine Fever (ASF) outbreak occurring in the south-east region of Fagersta, Sweden, with suspected involvement of wild boars on October 26, 2023”. The geographical identification of this outbreak was conducted using the GeospaCy tool in conjunction with the official source, Empress-i, as depicted in Figure 3. Figure 3a illustrates how GeospaCy extracted the location information, denoted as (*RSE: South-east of Fagersta*), from the text. Following this, Figure 3b demonstrates the subsequent step where GeospaCy computed the precise coordinates of the identified RSE, forming a polygon that ac-

curately corresponds to the south-eastern vicinity of Fagersta. For comparison, Figure 3c displays the official source location of the ASF outbreak, pinpointing it to the center of Fagersta. Notably, this example clarifies that GeospaCy provides a more granular and precise region of the outbreak compared to the location indicated by the official source. This indicates the tool potential in offering enhanced spatial precision in identifying outbreak locations.

6 Conclusion

GeospaCy focused on extracting spatial expressions such as GPE, LOC and the primary contribution of RSE extraction from text, subsequently translating the geographic coordinates of the identified RSE. We proposed a combination of NLP techniques to extract RSE from unstructured text. The results of the RSE extraction are evaluated with news article disease dataset having a precision of 0.9, recall of 0.88 and micro F-Score of 0.88. Sub-

²<https://www.pigprogress.net/health-nutrition/health/asf-sweden-first-outbreak-found-in-wild-boar/>

sequently, we conducted a qualitative assessment of RSE geographical coordinates (shapes) with an observed accuracy of 75%.

Short Video

The short video of the GeospaCy tool is available on YouTube for EACL 2024 demonstration on the following link : <https://youtu.be/sZb1aUkcRcs>.

Software Availability Statement

The code support the findings in this article are openly available in GitHub repositories dedicated to GeospaCy tool³.

Acknowledgements

GeospaCy tool is partially funded by EU grant 874850 MOOD and is catalogued as MOODXXX. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

- Hussein S. Al-Olimat, Valerie L. Shalin, Krishnaprasad Thirunarayan, and Joy Prakash Sain. 2019. *Towards geocoding spatial expressions (vision paper)*. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL ’19*, page 75–78, New York, NY, USA. Association for Computing Machinery.
- Álvaro Alonso Casero. 2021. *Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature*. Ph.D. thesis, ETSI_Informatica.
- Franz-Josef Behr. 2010. Geocoding: Fundamentals, techniques, commercial and open services. *AGSE 2010*, page 111.
- Hao Chen, Maria Vasardani, and Stephan Winter. 2017. Geo-referencing place from everyday natural language descriptions. *arXiv preprint arXiv:1710.03346*.
- Konstantin Clemens. 2015. *Geocoding with open-streetmap data*. *GEOProcessing 2015*, page 10.
- Jacques Fize, Ludovic Moncla, and Bruno Martins. 2021. Deep learning for toponym resolution: Geocoding based on pairs of toponyms. *ISPRS International Journal of Geo-Information*, 10(12):818.
- Sean Gillies et al. 2007. *Shapely: manipulation and analysis of geometric objects*.
- Milan Gritta, Mohammad Taher Pilehvar, Nut Lim-sopatham, and Nigel Collier. 2018. What’s missing in geographical parsing? *Language Resources and Evaluation*, 52:603–623.
- Andrew Halterman. 2017. Mordecai: Full text geoparsing and event geocoding. *J. Open Source Softw.*, 2(9):91.
- Hossein Hassani, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani, and Mohammad Reza Yeganegi. 2020. *Text mining in big data analytics*. *Big Data and Cognitive Computing*, 4(1):1.
- Matthew Honnibal and Ines Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Xuke Hu, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann. 2022. Gazpne2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal*, 9(17):16259–16271.
- Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D. Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, Sergio Rey, Christopher Ren, Dani Arribas-Bel, Leah Wasser, Levi John Wolf, Martin Journois, Joshua Wilson, Adam Greenhall, Chris Holdgraf, Filipe, and François Leblanc. 2020. *geopandas/geopandas: v0.8.1*.
- Margarita Kokla and Eric Guilbert. 2020. *A review of geospatial semantic information modeling and elicitation approaches*. *ISPRS International Journal of Geo-Information*, 9(3):146.
- Paul A Longley and James A Cheshire. 2017. Geographical information systems. In *The Routledge Handbook of Mapping and Cartography*, pages 251–258. Routledge.
- Katherine McDonough, Ludovic Moncla, and Matje van de Camp. 2019. Named entity recognition goes to old regime france: geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, 33(12):2498–2522.
- 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*, 1:1–18.
- Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)*, 36(4):1–27.

³<https://github.com/mehtab-alam/GeospaCy>

Ludovic Moncla. 2015. *Automatic reconstruction of itineraries from descriptive texts*. Ph.D. thesis, Pau.

OpenStreetMap contributors. 2017. *Planet dump retrieved from <https://planet.osm.org>* . <https://www.openstreetmap.org>.

João Santos, Ivo Anastácio, and Bruno Martins. 2015. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80:375–392.

Mehtab Alam Syed, Elena Arsevska, Mathieu Roche, and Maguelonne Teisseire. 2023. **Geospatre: extraction and geocoding of spatial relation entities in textual documents**. *Cartography and Geographic Information Science*, 0(0):1–17.

Sowmya Vajjala and Ramya Balasubramaniam. 2022. **What do we really know about state of the art ner?** In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 5983–5993. European Language Resources Association.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. ensemble named entity recognition (ner): evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5:2.

Kehan Wu, Xueying Zhang, Yulong Dang, and Peng Ye. 2022. **Deep learning models for spatial relation extraction in text**. *Geo-spatial Information Science*, 0(0):1–13.

Kun Zheng, Ming Hui Xie, Jin Biao Zhang, Juan Xie, and Shu Hao Xia. 2022. **A knowledge representation model based on the geographic spatiotemporal process**. *International Journal of Geographical Information Science*, 36(4):674–691.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Appendices

The appendix contains the software demonstration, some example use cases and evaluation of GeospaCy tool. The details of these are as follows:

A.1 Software Demonstration

Figure 4 shows the GeospaCy user interface for spatial entity extraction.

Figure 4: Geoparsing: **1:** selection of spaCy language model for NER (GPE and LOC) recognition, **2:** selection of spatial entity type user wish to extract from text, **3:** input text, **4:** text with highlighted selected spatial entity types and **5:** Table with list of spatial entities with option to view coordinates as GEOJson or on Map

Figure 5 shows the GeospaCy user interface for RSE geographical coordinates visualization.

Figure 5: Geocoding: **6:** view geographical coordinates as GEOJson or on Map, **7:** Compute level-1 corrdinates using midpoint(less area inside polygon) or midmidpoint (more area inside polygon), **8:** Export GEOJson on local device and **9:** This region shows the coordinates as Map or GEOJson

A.2 Disease Surveillance Use cases Scenarios

We discuss here some more disease outbreaks detected by PADI-web⁴ with improvement of precise location information using GeospaCy tool. The examples are as follows:

AI: The Czech Republic has found a second case of the bird flu virus, at a commercial poultry farm, an Agriculture Ministry spokesman said on Sunday. The spokesman said more details of the case, in a region east of Prague,.⁵



Figure 6: Region of outbreak: east of Prague (RSE)

AI: The outbreak recorded in poultry and captive birds near Melton, Mowbray, Leicestershire, and the outbreak in captive birds at a wetland centre near Stroud, Gloucestershire have both been confirmed as highly pathogenic.⁶

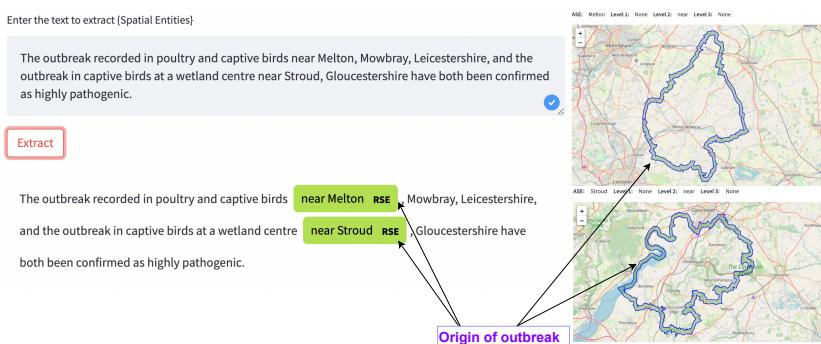


Figure 7: Region of outbreak: near Melton (RSE), near Stroud (RSE)

AI: France has detected a highly pathogenic strain of bird flu in a pet shop in the Yvelines region near Paris, days after an identical outbreak in one of Corsica's main cities.⁷



Figure 8: Region of outbreak: near Paris (RSE)

⁴<https://padi-web.cirad.fr/en/>

⁵<https://www.agriculture.com/markets/newswire/czech-republic-reports-second-bird-flu-case>

⁶<https://www.thepoultrysite.com/news/2020/11/2-bird-flu-clusters-in-the-uk-confirmed-as-highly-pathogenic>

⁷<https://www.reuters.com/article/us-health-birdflu-france-idUSKBN27Z35D>

AI: In less than a week after bird flu was detected in two poultry farms in Vengeri and west Kodiyathoor in Kozhikode district.⁸

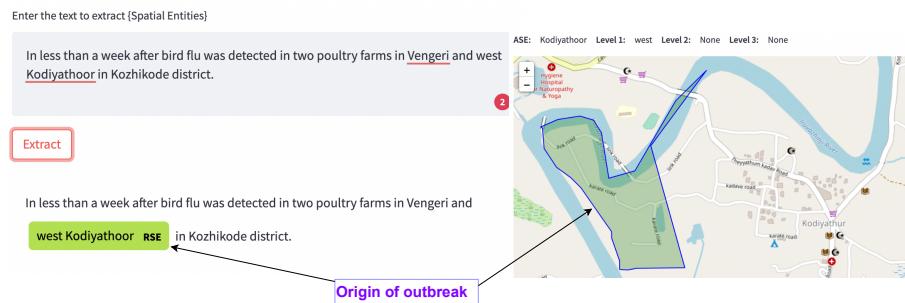


Figure 9: Region of outbreak: west Kodiyathoor (RSE)

AI: The Taipei Times reports that Taiwan has moved to block imports of live poultry after cases of highly pathogenic bird flu were detected at a farm near Cheshire, England.⁹

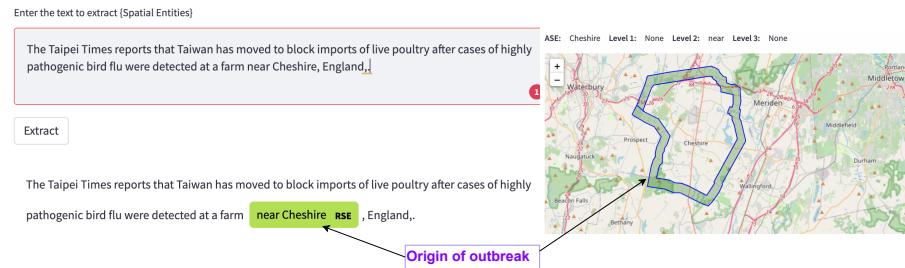


Figure 10: Region of outbreak: near Cheshire (RSE)

⁸<https://english.manoramaonline.com/news/kerala/2020/03/12/bird-flu-kozhikode-malappuram.html>

⁹<https://www.thepoultrysite.com/news/2020/11/south-korea-and-taiwan-ban-uk-poultry-imports-on-bird-flu-fears>

A.3 Evaluation: Extraction Phase

We evaluated RSE extraction through state-of-the-art evaluation protocol. For that, we had Named-entity recognition (NER) RSE annotated dataset to evaluate RSE extraction through standardized evaluation scores i.e., precision, recall and F-score. Table 1 shows the precision, recall and F-Score for RSE extraction task. For instance, in the first row of Table 1, the data reveals that 25 news articles were processed for AMR disease. The GeospaCy tool extracted 4 RSE, while 5 RSE were annotated. The evaluation results with precision, recall, and F-score of 1.0, 0.8, and 0.88, respectively. The overall score for all the RSE annotated disease dataset is calculated with precision of **0.9**, recall of **0.88** and F-Score of **0.88**.

Disease Name	No. of Articles	spatRE Extracted	spatRE Actual	Precision	Recall	F-Score
Antimicrobial resistance (AMR)	25	4	5	1	0.80	0.88
COVID-19	100	100	92	0.87	0.94	0.90
Avian-Influenza	150	57	68	0.87	0.83	0.84
Lyme	29	10	10	0.83	1	0.90
Tick-borne Encephalitis (TBE)	73	73	81	0.93	0.83	0.87
Average	377	244	256	0.9	0.88	0.88

Table 1: Extraction Phase Results (RSE Extraction)

A.4 Evaluation: Geocoding Phase

In order to evaluate the Geocoding phase, there is no state-of-the-art mechanism to evaluate the geographical coordinates of RSE. Therefore, we applied a qualitative assessment to evaluate the geometry of the geographical coordinates of RSE. The criteria of the geometry evaluation are 1) how well the geometry of the RSE geographical coordinates are represented, 2) how well the geometry shows the real geographical region of RSE. The geometries were being evaluated by geographical information system (GIS) end users. For instance, in Table 2, in first row the total score was computed for the 19 spatial relations associated with Paris, including N, S, E, and W, using evaluations provided by end users. The score 19 RSE of Paris evaluated by end user is 136 with the accuracy of 89.5%, with an average score of 3.6 out of 4. Overall, the accuracy of geometries of geographical coordinates of RSE computed by GeospaCy tool for London, Zagreb, Delft and Florence is better as compared to the other mentioned cities.

City	Obtained Score	Total Score	Accuracy(100%)	Mean(4)	Remarks
Paris	136	152	89.5	3.6	Excellent
London	142	152	93.4	3.7	Excellent
Milan	106	152	69.7	2.8	Good
Madrid	77	152	50.7	2	Weak
Zagreb	116	152	76.3	3.1	Excellent
Utrecht	105	152	69.1	2.8	Good
Delft	121	152	79.6	3.2	Excellent
Lyon	114	152	75	3	Good
Florence	477	608	78.5	3.1	Excellent

Table 2: Qualitative Evaluation of Spatial Relation by City