Geo-coding of textual input

Master internship

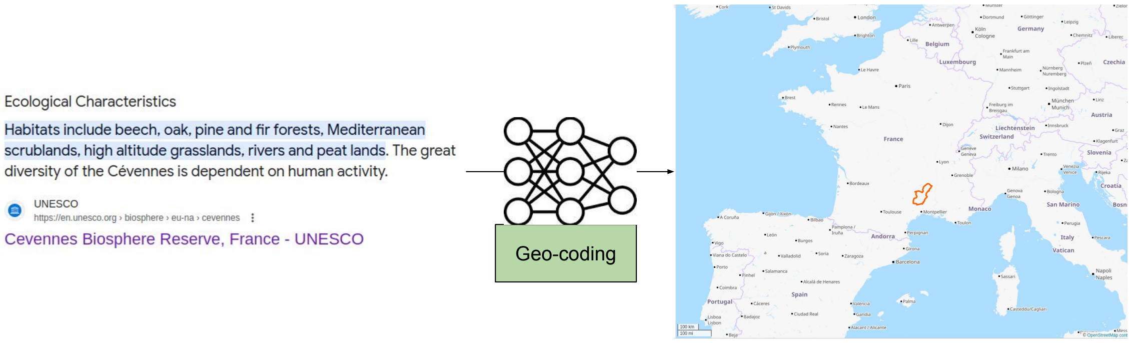
General information

- Duration: 6 months (standard stipend). To start between February and April 2024.
- Institutes: Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), team Systèmes Intelligents de Perception and team EVERGREEN (Inria, INRAE, Cirad)
- Location: 45 rue des Saints-Pères, 75006, **Paris** (LIPADE)
  or 500, rue Jean François Breton, 34090, **Montpellier** (EVERGREEN)
- Supervision: Sylvain Lobry, Camille Kurtz, Laurent Wendling (LIPADE), Diego Marcos (Inria), Dino Ienco (INRAE)
- **Application: Please apply on TOCOME. The position is open until filled, and full consideration will be given to application received before 15/12/2023.**

Proposed topic

Context

By using location on the Earth's surface as the common link between different modalities, a geo-spatial foundation model would be able to incorporate a variety of data sources, including remote sensing imagery, textual descriptions of places, and features in maps. Leveraging the large amounts of available unlabeled geo-spatial data from these different sources, the GEO-ReSeT [1] (Generalized Earth Observation with Remote Sensing and Text) ANR project has the objective to learn a better representation of any geo-spatial location and convey a semantic representation of the information. Such a foundation model has the potential to revolutionize Earth observation by allowing for few or zero-shot solutions to classical problems such as land-cover and land-use mapping, target detection, and visual question answering. It will also be useful for a wide range of applications with a geo-spatial component, including environmental monitoring, urban planning and agriculture. By leveraging several data modalities, this foundation model could provide a more comprehensive and accurate understanding of the Earth's surface, enabling more informed decisions and actions. This will be particularly valuable for new potential users in sectors such as journalism, social sciences or environmental monitoring, who may not have the resources or expertise to collect their own training datasets and develop their own methods, thus moving beyond open Earth observation data and democratizing the access to Earth observation information.



Ecological Characteristics

Habitats include beech, oak, pine and fir forests, Mediterranean scrublands, high altitude grasslands, rivers and peat lands. The great diversity of the Cévennes is dependent on human activity.

UNESCO
https://en.unesco.org › biosphere › eu-na › cevennes

Cevennes Biosphere Reserve, France - UNESCO

Figure 1: The main objective of this project is to design a geo-coding pipeline able to generate a geographic footprint of the area described in any piece of text with a geographic component.

[1] https://geo-reset.sylvainlobry.com/

Work to be done

The work to be conducted during the proposed M2 internship will contribute to the ambition of the GEO-ReSeT ANR project by linking textual descriptions of places, found online, to their approximate geo-location, a task known as geocoding [1]. This text-location link will then be used along the project in combination with other geospatial data modalities, such as those stemming from remote sensing sensors, in order to train multi-modal models that are aware about the way in which people describe locations. This will be done by first combining information stemming from different databases containing geographic named entities, such as Open Street Maps, Wikipedia and gazetteers, such that geographic points or polygons can be linked to each entity. In a second step, a pipeline will be developed to obtain the most likely geographic named entities that are referred to in any piece of text that describes a place. In order to avoid restricting us to cases where entities' names appear exactly as in the databases, we will leverage pre-trained Large Language Models (LLM) [2] to resolve ambiguities and gather evidence towards the most likely entities that are being described in the text.

In this work, our objective is to develop a pipeline to automatically link any piece of text describing a place with its most likely geographical footprint. The work to be performed in this internship will lead to the following three contributions:

- Contribution A: a first pipeline to allow querying a variety of databased that include geographic named entities, such that, given a name, a list of possible geographic footprints, either in the form of points or polygons, is obtained.
- Contribution B: a second pipeline in which LLM are used to determine if a piece of text does contain geographic information and proposes potential named entities that would be associated to it. These proposals will the be used to query in the pipeline developed in the previous contribution in order to obtain candidates for the geographic footprint that is relevant to the text. An additional module will be developed in which LLM will be once again used to determine which of the proposed footprints is the most likely.
- Contribution C: the developed pipelines will be used to build a large dataset of text and the corresponding geographic footprints. The candidate will propose a methodology for evaluating the obtained dataset.

Desired background

We are looking for a Master 2 student or final year of MSc, or engineering school in computer science. The ideal candidate should have knowledge in image processing, computer vision, natural language processing, geo-information sciences, Python programming and an interest in handling large amount of data, in particular remote sensing.

Bibliography

[1] Fernando Melo and Bruno Martins. "Automated geocoding of textual documents: A survey of current approaches". In: *Transactions in GIS* 21.1 (2017), pp. 3–38.

[2] Bonan Min et al. "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40.