2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

# Enriching Wikipedia Texts through Geographic Information Extraction

Laura Ventrice
*Dept. of Computer Science*
*University of Turin*
Turin, Italy
laura.ventrice@unito.it

Luigi Di Caro
*Dept. of Computer Science*
*University of Turin*
Turin, Italy
luigi.dicaro@unito.it

***Abstract*—Geographic Information Extraction (GIE) involves the extraction of geo-referenced information from a data collection through steps of geoparsing and geocoding. The former is a process that starts from a free textual description of locations with the goal of identifying an unambiguous location, such as specific geographic coordinates expressed as latitude-longitude. Differently, geocoding regards the easier task of translating an exact and well-formatted location such as postal addresses. This paper presents *MAWI*, i.e. a pipeline that starts from generic texts about cities that first extracts geographic information to automatically detect possible points of interest, then generates textual snippets from their contexts by means of Natural Language Processing (NLP) techniques. The adopted methodology involves several modules, ranging from publicly available geocoding systems to NLP libraries for Named Entity Recognition and text segmentation. The impact of the proposal includes multiple tasks and applications, e.g. *i)* the enrichment of public platforms of geographic data, *ii)* the detection of geographic scopes in textual documents, *iii)* a geo-centric exploration of locations in the tourism domain, and so forth. In this contribution, we present an experimentation of the system with 50 input Wikipedia pages referring different cities, first demonstrating its effectiveness with a running example, then evaluating its power to detect and structure a highly-significant amount of novel geo-referenced information with respect to what currently encoded in Wikipedia. Data and code are publicly available for future research at https://anonymous.4open.science/r/PointOfInterest-8D80/.**

***Index Terms*—geographic information extraction, named entity recognition, geoparsing, wikipedia enrichment**

## I. Introduction

A great amount of information can be extracted from textual or unstructured data which are currently available online through social networks, blogs, and online encyclopedias such as Wikipedia. In this work, we focus on investigating the possibility of detect, extract and structure pieces of textual information which can be mapped to physical places through the employment of Natural Language Processing techniques and existing resources.

This geo-referenced information can be used to generate spatial indexes to enable users to use geographic information in Web searches. In addition, there are services that specifically rely on these types of information, e.g. map-based systems such as Google/Apple Maps and social networks aiming to collect information that is scattered across the web, reducing the information overload by filtering through location-based principles and strategies [1].

For these purposes, geographic information often requires some manual intervention and annotation phase involving spatial references and coordinates. Therefore, the annotation process we are interested in can be divided into two phases: *i)* the first one extracts entities describing geo-referenced places (i.e., geoparsing), and *ii)* a second phase where entities are mapped with spatial references (i.e., geocoding).

The main objective of this short paper is to present an end-to-end pipeline, named *MAWI*, which parses textual documents to detect, extract and enrich geographic entities through NLP technologies and existing open-source resources. As experimental environment, we tested the method on Wikipedia articles about notable cities in two different languages (Italian and English). As a result, the system was able to automatically unveil approximately 50% of new high-quality geographical information from the unstructured text.

The paper is organized as follows: Section II briefly outlines the related work concerning the different parts of the developed modules; Section III illustrates an overview of the proposed system; in Section IV, the results obtained are presented; finally, Section V concludes the paper with future prospects.

## II. Related Work

Automatic extraction of locations from unstructured texts is a crucial task in Natural Language Processing and Information Retrieval. In recent years, there have been several works that have explored different approaches to tackle this problem. Here, we discuss some of the most recent and relevant studies in this area.

One of the primary methods to extract locations from texts is Named Entity Recognition (NER), which focuses on



*ASONAM '23*, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3630110

identifying and classifying named entities within a given text. Recently, the literature has been mainly occupied by BERT-based NER Models, such as [4]. BERT (Bidirectional Encoder Representations from Transformers) [5] is a powerful pre-trained language model that has significantly improved the performance of various NLP tasks, including NER. Geoparsing is a more specific task that involves extracting location information from text and resolving it into geographical coordinates [8], [11], [12]. A popular geoparsing tool, CLIFF-CLAVIN, was developed by [3] and uses a combination of NER techniques, gazetteers, and heuristics to extract and disambiguate location information from unstructured texts.

Generally speaking, to extract geographic information from textual resources, it is necessary to parse the text and identify geospatial entities, and annotate them with references, when available. Several studies have explored the value of geo-referenced data and its automatic extraction from natural language. Some of these studies are based on the premise that every document has a geographic scope, which can be leveraged to enhance information retrieval, as exemplified in [2], [7], with a focus on public and commercial interest, and the societal applications of this information. Other research has concentrated on developing automatic information extraction systems to populate resources, including the identification of relationships between them [6]. Other efforts focused on the extraction of locations within social messages [10]. Some research focused on the specific implementation of systems, e.g. GeoTxt [9], which included six different NER modules to index geographic information within unstructured texts.

In summary, recent research on automatic extraction of locations from unstructured texts has seen significant advancements in NER, geoparsing, and end-to-end systems. These approaches have leveraged deep learning techniques, unsupervised clustering, and knowledge graph-based methods to enhance the performance of location extraction tasks. To the best of our knowledge, our *MAWI* system represents the first attempt to create an end-to-end pipeline working on the Italian language, in addition to English. Furthermore, our experimental setting based on Wikipedia represents a novel, simple and effective methodology to concretely evaluate geographic information extraction systems.

## III. Architecture of the proposed system

This paper proposes the application of a Named Entity Recognition model to extract fine-grained entities of geographic significance, including historically, culturally, and naturally important locations. Specifically, we have gathered a specific set of document classes comprised of Wikipedia articles about world-renowned cities in both Italian and English. The rationale for selecting these documents stems from the clarity of their geographic scope, which facilitates the disambiguation of extracted entities during the geocoding process. The aim of this experiment is to assess the extent to which geo-referenced information, which is not already linked to other Wikipedia articles, can be automatically extracted. The architecture of *MAWI* is categorized based on the tasks it carries out, i.e. geoparsing and geocoding.

### A. Geoparsing

Initially, a preprocessing phase is executed wherein the text is purified to guarantee accurate decoding of the HTML source of Wikipedia articles. Particular attention is paid to sections containing bulleted lists, which are frequently abundant in location entities in this document type. To pursue the goal, each selected document is processed by a language-specific SpaCy[1] model used in the text, specifically transformer-based and fine-tuned for NER[2].

Afterwards, from the set of entities identified by the model, only the elements labeled as "*LOC*" or "*FAC*" are retained, representing respectively non-*GPE* locations, mountain ranges, bodies of water and in the second case airports, roads, buildings, bridges and so forth. Entities labeled as *GPE*, on the other hand, are used to identify geographic scopes, which for the chosen document types will be unambiguous but still need to be identified. In this scenario, the idea was to find the most frequently-mentioned *GPE* entity, which intuitively will be the city that is the subject of the article. The geopolitical entity is then searched within the OpenStreetMap geographic database[3] to obtain the coordinates of its location and additional information such as the state/region in which it is located. This information will be useful in the next phase.

Subsequently, the entities are scrutinized to eradicate all generic ones that would consequently provide no useful information. This is accomplished by utilizing Part-of-Speech tagging information. In case the entity consists of only one word tagged as *Noun*, it is discarded. To improve the accuracy of the extraction and subsequent search for locations, a further step was added to check that the entities are well-formed in terms of punctuation and POS useful for the identification, removing, for example, determinative articles and prepositions.

From the results produced by the pipeline, information about the segmentation of the text into sentences is also stored, so that textual snippets can be associated with the extract entities as a form of context enrichment.

### B. Geocoding

To link geographic coordinates with the data obtained in the previous step, Nominatim[4] was employed. Nominatim is a geocoding service that functions as a search engine for the OpenStreetMaps (OSM) geographic database.

To request coordinates, a query is executed wherein the name of the location to be searched is entered without any specific formatting. With Nominatim, structured queries can also be performed by inputting information such as streets, cities, and countries to expedite and refine the search. Moreover, the viewbox parameter allows for a preferred search area

[1]https://spacy.io

[2]A trained pipeline based on transformers directly from SpaCy is not available for the Italian language, so a model available on HuggingFace was used https://huggingface.co/bullmount/it_nerIta_trf

[3]https://www.openstreetmap.org

[4]https://nominatim.org

to be specified. These features presented were the key to make the system effective and efficient.

In general, two difficulties may arise during this procedure: *i)* the query can be ambiguous so that multiple results can be obtained with no unique coordinates pair, or, *ii)* due to the limitations of the service, no results are obtained even though the requested location exists.

To alleviate the initial issue, structured queries were made to the service employing the geographic scope of the article. As a result, the received outcomes are more closely associated with the entity referenced in the article, and the quantity of results is reduced. As previously stated, the corpus comprises articles with a clearly defined geographic scope, which simplifies this phase. The "*spatial minimality*" hypothesis [9] was adopted for this purpose. The principle behind this approach is that the bounding box of the geographic scope represents the region within which entities should be included.

At the end of geographic coordinate retrieval, the results are saved in a GeoJSON format, and each item is provided with entity name, location name, coordinates, and textual snippet.

## IV. Results

In this section, we first present a running example of the proposed system, then reporting the detailed results of the whole experimentation. Consider the following Wikipedia sentence included in the experimental set[5]:

> [Italian] "*Tra i monumenti di Torino più noti anche all'estero sono da citare l'ottocentesca Mole Antonelliana, simbolo incontrastato della città, che ospita il Museo nazionale del Cinema; il Palazzo Reale (antica dimora dei duchi ed in seguito dei re di Casa Savoia)*"
>
> [English] "*Among the most well-known monuments of Turin, also abroad, we must mention the 19th-century Mole Antonelliana, an undisputed symbol of the city, which houses the National Museum of Cinema; the Royal Palace (ancient residence of the dukes and later of the kings of the House of Savoy).*"

The recognized entities and the result of geocoding for the Italian case are reported in Figure 1.

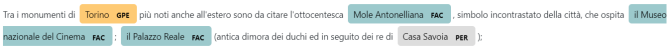


Fig. 1. Named Entity Recognition of the sample sentence using the SpaCy model for Italian.

From the sentence, we can note that the places "*Mole Antoneliana*", "*Museo nazionale del Cinema*" and "*Palazzo Reale*" are correctly recognized. Then the coordinates are searched in the geocoding stage, and the results of the extraction are displayed in a map as in Figure 2, with *geojson.io*[6].

[5]Extracted from the Italian Wikipedia article linked to the city of "*Torino*" (Turin).

[6]https://geojson.io/

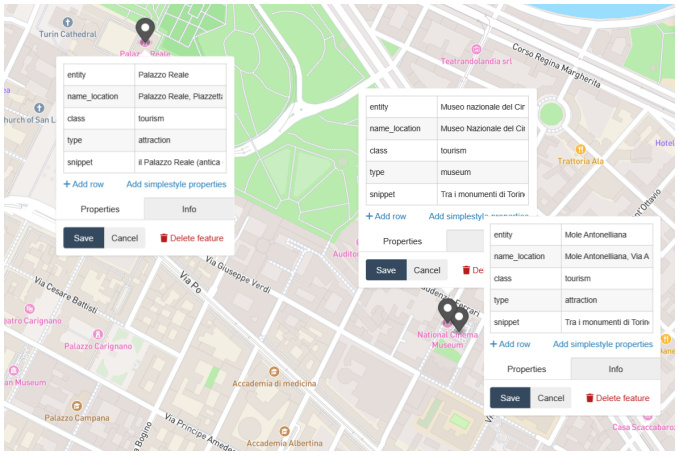


Fig. 2. Example of geoparsed entities in the Wikipedia page of the city of Turin.

In order to provide a thorough assessment of the system's performance from both quantitative and qualitative perspectives, this study presents system statistics alongside detailed information on the utilized NER type. Notably, the efficacy of transformer-based models for this particular task is emphasized, and a comparative analysis between transformer-based and non-transformer-based NER models is proposed. The models we used are executable through the SpaCy library, as shown in Table I with their specifications and sources.

The results presented below were obtained from a dataset comprising 50 articles each in Italian and English, denoted as DOCIT and DOCEN, respectively. These articles relate to notable cities across the globe. Tables II and III display findings pertaining to information extraction, including the average number of locations identified and the average proportion of entities not already featured in Wikipedia articles. The latter metric reflects the extent to which extraction contributes novel insights beyond preexisting links to Wikipedia. Additionally, the final value reports the ratio of entities detected via geoparsing to those identified through geocoding.

In regards to the Italian language, it can be observed that both kinds of model generate a similar ratio of new information compared to the Wikipedia links. However, there is a significant difference in the average number of extracted locations between the two models. The reason for this difference lies in their availabile labels. The non-transformer-based model lacks the labels "*FAC*" and "*GPE*", consequently, all location entities are labeled as "*LOC*". This results in the inclusion of all geopolitical entities, although with less accuracy for the proposed task. Despite this, the geocoding phase helps filter out some of the results, which is why the ratio remains relatively consistent across both models. An example of models differences is shown in Table IV for the input sentence "*This neighbourhood hosts the significant architecture of Santuario di Maria Ausiliatrice ("Maria Ausiliatrice Sanctuary") in the homonymous square and behind the church stands San Pietro in Vincoli old cemetery*".

Regarding the English language, both models used have

TABLE I
THE EMPLOYED MODELS WITHIN THE SPACY LIBRARY ALONG WITH THE PERFORMANCE SCORES REPORTED IN THE SOURCES.

| Model name | Language | Source | Transformer-based | Precision | Recall | F-Score |
|---|---|---|---|---|---|---|
| it_core_news_lg | Italian | SpaCy | no | 0.88 | 0.88 | 0.88 |
| en_core_web_lg | English | SpaCy | no | 0.85 | 0.86 | 0.85 |
| it_nerIta_trf | Italian | Hugging Face | yes | 0.92 | 0.91 | 0.92 |
| en_core_web_trf | English | SpaCy | yes | 0.90 | 0.90 | 0.90 |

TABLE II
FREQUENCY OF LOCATIONS AND RATIOS USING TRANSFORMER-BASED NER MODELS.

| Corpus | # docs | # locations | avg # locations | avg ratio | novelty ratio |
|---|---|---|---|---|---|
| DOCIT | 50 | 2808 | 56.16 | 0.54 | 0.41 |
| DOCEN | 50 | 3720 | 74.40 | 0.51 | 0.53 |

TABLE III
FREQUENCY OF LOCATIONS AND RATIOS USING BASIC (NON-TRANSFORMER) MODELS.

| Corpus | # docs | # locations | avg # locations | avg ratio | novelty ratio |
|---|---|---|---|---|---|
| DOCIT | 50 | 4915 | 98.30 | 0.53 | 0.44 |
| DOCEN | 50 | 1471 | 29.42 | 0.49 | 0.42 |

the same set of labels, resulting in a similar ratio of new entities. However, when the non-transformer-based NER is employed, the number of locations extracted is significantly lower compared to the transformer-based model. This can be also attributed to the differing levels of precision and recall between the two models.

The obtained results reveal that approximately 50% of the extracted information is not present in the corresponding Wikipedia links. This highlights the significant contributions of this system, which include utilizing a Transformer-based NER model and linking short text fragments to the extracted information.

TABLE IV
RESULTS OBTAINED WITH THE TRF-MODEL IN THE FIRST ROW, AND WITH THE BASE MODEL IN THE SECOND.

| GPE | LOC | FAC | Other |
|---|---|---|---|
| Vincoli | NA | Santuario di Maria Ausiliatrice, San Pietro | NA |
| San Pietro, Vincoli | NA | NA | Maria Ausiliatrice (PER), Maria Ausiliatrice Sanctuary (WORK_OF_ART) |

## V. CONCLUSIONS AND FUTURE WORK

This work shows that information extraction using Transformer-based models can contribute in entity identification of places of historical, cultural and naturalistic importance. The reported results also confirm that the extracted information can contribute in making Wikipedia articles more informative, particularly in reference to cities. An experimentation involving one hundred Wikipedia articles in Italian and in English demonstrated the validity of the approach.

In addition to the current study, our future endeavors include extending Information Extraction to encompass other document types that harbor distinct geographic scopes, as well as exploring the diverse properties and relationships that entities may reveal. The resulting insights may enhance and complement other geographic data, enabling more precise determination of a document's geographic scope. Furthermore, such knowledge may have practical applications in the tourism domain, such as facilitating visitors' access to relevant content associated with sites of interest, which is derived from extracted information.

## REFERENCES

[1] Antonini, A., Boella, G., Calafiore, A., Salaroglio, C., Sanasi, L. & Schifanella, C. First Life, the Neighborhood Social Network: A Collaborative Environment for Citizens. *Proceedings Of The 19th ACM Conference On Computer Supported Cooperative Work And Social Computing Companion*. pp. 1-4 (2016), https://doi.org/10.1145/2818052.2874310

[2] Pantaleo, G. & Nesi, P. Ge(o)Lo(cator): Geographic Information Extraction from Unstructured Text Data and Web Documents. *Proceedings - 9th International Workshop On Semantic And Social Media Adaptation And Personalization, SMAP 2014*. (2014,11)
[3] D'Ignazio, C., Bhargava, R., Zuckerman, E. & Beck, L. Cliff-clavin: Determining geographic focus for news articles. (NewsKDD: Data Science for News Publishing, at KDD 2014,2014)
[4] Hu, X., Zhou, Z., Sun, Y., Kersten, J., Klan, F., Fan, H. & Wiegmann, M. GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet Of Things Journal*. **9**, 16259-16271 (2022)
[5] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)
[6] Lima, E. & Junior, C. Geographic information extraction using natural language processing in Wikipedia texts. *Brazilian Symposium On GeoInformatics*. (2017)
[7] Andogah, G., Bouma, G. & Nerbonne, J. Every document has a geographical scope. *Data & Knowledge Engineering*. **s 81–82** pp. 1-20 (2012,11)
[8] Leidner, J. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding. *SIGIR Forum*. **41** pp. 124-126 (2007,1)
[9] Karimzadeh, M., Pezanowski, S., MacEachren, A. & Wallgrün, J. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions In GIS*. **23**, 118-136 (2019), https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12510
[10] Wang, J., Hu, Y. & Joseph, K. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions In GIS*. **24**, 719-735 (2020)
[11] Gritta, M., Pilehvar, M. & Collier, N. A Pragmatic Guide to Geoparsing Evaluation. (2019)
[12] Aldana-Bobadilla, E., Molina-Villegas, A., Lopez-Arevalo, I., Reyes-Palacios, S., Muñiz-Sanchez, V. & Arreola-Trapala, J. Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/18/3041