UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Geographical Named Entity Recognition from Travel Articles

Author:
Andrii MAISTRUK

Supervisor:
Vasyl MYLKO

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2021

Declaration of Authorship

I, Andrii MAISTRUK, declare that this thesis titled, "Geographical Named Entity Recognition from Travel Articles" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Like all great travellers, I have seen more than I remember and remember more than I have seen."

Benjamin Disraeli

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Geographical Named Entity Recognition from Travel Articles

by Andrii MAISTRUK

Abstract

Geographical Named Entity Recognition, also known as geoparsing, is the task of obtaining geographical coordinates from free-format text, arises in many real-world applications such as understanding location instructions in auto-response systems, determining a document's geographic scope, real-time social media geographical event analysis, and more. Geoparsing consists of two parts: toponym extraction from the text; toponym resolution, disambiguating and connecting toponyms to fully specified real-world locations, geographical coordinates.

In this work, I tackle the problem of geoparsing travel guide articles. For this reason, I developed the geoparsing system. As input data, National Geographic Travel articles describing road trips in the United States were used. The code is available on Github[23].

Acknowledgements

First of all, I am very grateful to Vasyl Mylko for spending a tremendous amount of time guiding me through all thesis research, helpful advice, and always giving valuable feedback. I want to thank Applied Sciences Faculty at the Ukrainian Catholic University for all the knowledge and experience I acquired during the four years.

Contents

List of Figures

List of Tables

Dedicated to my family

Chapter 1

Introduction

Geoparsing is the vital component of the Geographical Information Retrieval system, whose task is to extract and infer geographical coordinates from unstructured text that frequently has noisy and ambiguous information. For the information searching systems, geographical data represents an additional dimension by utilizing which they can provide query results based on a more profound understanding of web page content. In particular, knowing what region and places the web page text describes leads to better suggestions for queries like "California coast road trip" or "Chicago places to visit." The practical relevance of the problem also arises when people create travel advising tools. They inevitably go through reading travel tour articles and resolving all places mentioned in the article to their geographical coordinates. Automating such processes requires natural language understanding and an up-to-date worldwide geographical knowledge base.

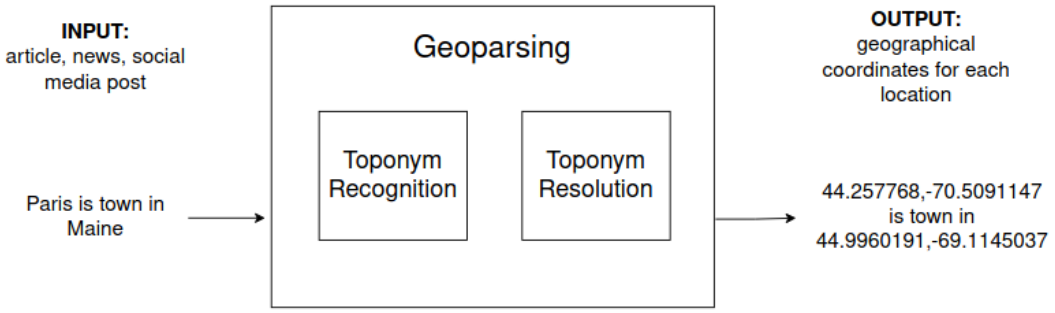

FIGURE 1.1: Geoparsing pipeline

The geoparsing typically consists of two components: toponym recognition and toponym resolution. The toponym recognition part is to extract the possible toponyms from the text. Due to the essential NER component for geotagging stage, investigation and comparison of different off-the-shelf NER taggers focusing on the toponym extraction capability are conducted in this work's scope. The task of toponym resolution is posed as follows, having extracted several toponyms in the text, for each, there might be more than one possible geolocation in the gazetteers, assign the correct geolocation meant in the text. The complexity also arises due to various definitions of the toponym from the most common (and ambiguous) dictionary definition as a place name to the United Nations Conference on the Standardization of Geographical Names definition as the general name for any place or geographical

entity. Other researchers extend this definition to include the names for topographical features. Therefore as toponyms that require coordinates in this thesis, we call countries, cities, villages, regions, neighborhoods, lakes, monuments, points of interest, outdoor areas, landmarks, and buildings (museum, restaurants, etc.).

Chapter 2

Related works

The first attempts at geocoding were made in 1994 by Woodruff, who introduced the first geoparsing prototype within GIPSY. [1]. The task of geoparsing remains an open problem to this date due to the complex interaction between spatial, temporal, and thematic sub-space within the text that needs to be addressed depending on the problem domain [2]. The first comprehensive survey and critical evaluation of state-of-the-art geoparsers with heterogeneous datasets were made by [3], concluding that featured geoparsing systems can only be used as additional input acknowledging the technology limits. Such a conclusion illustrates the complexity of coordinating the extraction of location entities and exact geographical resolution from unstructured text.

2.1 Toponym Recognition methods

The first task in geoparser is to determine which text tokens denote the names of places. This stage in the geoparsing pipeline is known as geotagging or toponym recognition. Toponym recognition can also be considered as a specialized form of Named Entity Recognition (NER) with the focus on recognizing named geographical entities[26]. For it, there is a requirement for methods that can discriminate toponyms from other recognized entities. The common geotagging approach is to employ gazetteer lookup, a simple string matching process with records from the external resource of place names and basic geographic information. A gazetteer is a geographical dictionary or geographical thesaurus. It typically contains information concerning the geographical makeup, social statistics, physical features of a country, region, continent, and alternative names. Gazetteers vary in their coverage of names, associated geographical information, and hierarchical structure. One of the classification categories for gazetteers is whether it has a toponym hierarchy or not. Gazetteer, which has a toponym hierarchy, is called ontological gazetteer[24]. An ontological gazetteer that has the correct hierarchy for all its entries is called a strict gazetteer. Geonames[15] is an ontological gazetteer, and it does not have a strict geo-ontology. For example, there are records of a village (administrative level 4) placed directly under a province-level entry (level 1), whereas it should be under sub-district (level 3). The data-driven method for toponym recognition is most widely used. Data-driven methods require an annotated corpus (often annotated using BIO scheme) for training to distinguish entity types such as Person (PER), Location (LOC), or Organization (ORG). The commonly used NER tools are AllenNLP[20], StanfordNER[21], Spacy[14], and Flair[19].

2.2 Toponym Pragmatics

As for inherent toponym ambiguity, the first systematization was made in [4]. The authors address semantic ambiguity of toponyms and do a deep dive into toponym pragmatics, i.e., a toponym from a linguistic point of view and the practical NLP implications. They propose toponym taxonomy, review and consolidate metrics for geocoding. The toponyms, according to the proposed taxonomy, are divided into two groups: literal and associative. Literal toponym types refer to places where something is happening or is physically located. Associative toponym types refer to or are used to modify non-locational concepts associated with locations rather than directly referring to their physical presence. For example, U.S. Supreme Court has the U.S. embedded in its name; however, the U.S. does not refer in this toponym to the country. As for geocoding metrics, they conclude that F-score is an inappropriate metric for the following reasons. Incompatibility of results for geoparsers built with different knowledge databases. The all-or-nothing approach makes errors as 5-10 km from gold coordinates equal to 500 km deviation. The underspecification of the difference between recall and precision for the geocoding, i.e., is a correctly geotagged toponym resolved to a more distant place than X km, a false positive or false negative. As for geoparsing evaluation, one of the recommended metrics is Accuracy@161km, which is a fast and intuitive way to inform of "correct" resolutions (error within 160 kilometers of gold coordinates), ignoring the rest of the error distribution. For geoparsers benchmarking, they removed from the used dataset the most difficult toponyms to geographically resolve, such as buildings, venues, streets, and demonyms and homonyms. The result of such removal is that 94–95% of the 1547 correctly recognized toponyms were resolved to within 161 km. In this work, the removed toponym types are evaluated for a complete picture of current geoparsing limitations for real-world scenarios, such as geocoding travel articles.

2.3 Toponym Resolution methods

The summary of the toponym disambiguation methods was done in [5]. There are three approaches: map-based, knowledge-based, and data-driven or supervised. Map-based methods use an explicit representation of toponyms on a map, such as geographical area or coordinates, for instance, to calculate the average distance from specific context toponyms to ambiguous toponym referents. These methods, however very sensitive to the context, the places that are known to be far away from each other can be geocoded in the same region. In case Paris, in France, is mentioned in the text describing places to see in Maine U.S. state, the Paris would be resolved to Paris, Maine rather than to Paris, France. Knowledge-based methods utilize knowledge sources such as ontologies, DBpedia, or gazetteers to find any clues on the most probable referent. The examples of such clues are the place's population, more populated places are more likely to be mentioned; hierarchical information from the gazetteer is used for containment relationship, as the places mentioned in the same context tend to be in one region. The data-driven or supervised methods rely on machine learning techniques. The machine learning-based approaches are not typical for toponym resolution, due to the lack of open geographically tagged data. The advantage of supervised methods is that they can exploit non-geographical context. For example, if a person or organization is based at a place, their presence in the context of toponym may give an essential clue for disambiguation. The implemented

in this work toponym resolution approach represents a hybrid of map-based and knowledge-based methods.

2.4 Resolution scope of geoparsers

The geoparsing tools fall into three categories: toponym level resolution scope, document level resolution scope, event level resolution scope[6]. Toponym-level resolution scope type geoparser for input text will output the list of extracted toponyms with the assigned coordinates. Toponym-level is the prevalent type of geoparsers created. Edinburgh Geoparser [7], CLAVIN[8] are examples of toponym-level geoparser. Edinburgh Geoparser consists of a rule-based toponym recognition system and heuristics georesolver scoring toponym referents based on feature type, population, contextual information, locality parameter from the user, clustering. CLAVIN (Cartographic Location And Vicinity INdexer) works by using AllenNLP for extracting place names and heuristics-based combinatorial optimization for the toponym resolution stage. The document-level resolution scope geoparser's goal is to identify the geographical focus of the document. The examples are CLIFF [9] and Newstand[10]. CLIFF is built on top of the CLAVIN architecture; a multi-stage heuristic disambiguation pipeline replaced the toponym resolution method to remedy disambiguation errors found after evaluation on their internal dataset. Newstand is news articles geoparser that for geotagging used NE tagger of the Ling Pipe toolkit[25] and for toponym resolution used multiple heuristics filters. The event-level resolution scope geoparser is set to detect the event(s) mentioned and resolve the location or geographical scope of those events. This type of geoparsers is the most recent development and is constrained by event models and ontology. The examples are Mordecai[11] and Profile[12]. Mordecai uses Spacy for toponym extraction and custom Geonames gazetteer setup; for the toponym resolution stage, they trained two neural networks on the private dataset to infer the correct country and correct gazetteer entries for each placename. Mordecai explicitly defines event notion and links the (possibly several) event(s) with its locations. Profile for toponym recognition used StanfordNER, and for detection sentences that contain focus and non-focus locations, they trained an SVM classifier. The presented in this work geoparser is of toponym-level scope type.

Chapter 3

Data

As data, we decided to use National Geographic Travel articles for the U.S., with the use case in mind to create itineraries based on travel stories autonomous or at least faster. The source's credibility is supported by the survey that concluded that National Geographic is one of the most trusted U.S. brands [13]. The data used for NER open-source tools comparison and evaluation of proposed toponym resolution method's accuracy is ten road trip articles by area ranging from one to several states. For the articles, category classification and sentiment analysis on a scale from -1.0 to 1.0 were conducted using Google Cloud Natural Language API. Nine out of ten were written neutrally, except for one written in a positive manner describing the best places in Silicon Valley. The total number of unique toponyms in the articles is 382. The length of articles varies from 696 to 2340 words. The categories to which articles were assigned are: travel, tourist destinations, science, restaurants, hobbies and leisure, regional parks, and gardens.

TABLE 3.1: Page Link

Article ID	Article URL
1	https://www.nationalgeographic.com/travel/article/route-66
2	https://www.nationalgeographic.com/travel/article/must-see-stops-destinations-californoia-route-1
3	https://www.nationalgeographic.com/travel/article/virginia-tennessee-kentucky-national-parks
4	https://www.nationalgeographic.com/travel/article/national-parks-california
5	https://www.nationalgeographic.com/travel/article/dark-sky-road-trip-in-the-southwest
6	https://www.nationalgeographic.com/travel/article/camping-road-trip-through-magical-caverns
7	https://www.nationalgeographic.com/travel/article/places-to-go-coastal-road-trip
8	https://www.nationalgeographic.com/travel/article/hudson-valley-new-york-road-trip
9	https://www.nationalgeographic.com/travel/article/explorers-guide-7
10	https://www.nationalgeographic.com/travel/article/geek-retreat-the-best-of-silicon-valley

TABLE 3.2: Data Statistics

Article ID	Sentiment	Category	Word number	Toponym number
1	0.0	/Travel	1850	61
2	0.1	/Travel/Tourist Destinations	1175	34
3	0.1	/Travel/Tourist Destinations/Regional Parks and Gardens	2340	48
4	0.1	/Travel/Tourist Destinations	1594	60
5	−0.1	/Travel/Tourist Destinations /Science	1385	38
6	0.0	/Travel/Tourist Destinations /Science/Earth Sciences	1191	20
7	0.1	/Hobbies and Leisure	1189	25
8	0.1	/Travel/Tourist Destinations	1350	52
9	0.2	/Travel /Food and Drink/Restaurants	578	22
10	0.5	/Food and Drink/Beverages/Alcoholic Beverages	696	22

Chapter 4

Proposed Method

4.1 System Architecture

The developed geoparsing system consists of three parts: extraction of possible place names, lookup in gazetteers, and the disambiguation of several toponym geolocations. For toponym extraction, Spacy[14] library is used after recall evaluation of several open-source NER tools on the National Geographic Travel articles. As geolocations source, three gazetteers are used Geonames[15], Gisgraphy[16], and DBpedia[17]. For toponym resolution, a hybrid of map-based and knowledge-based methods is employed.

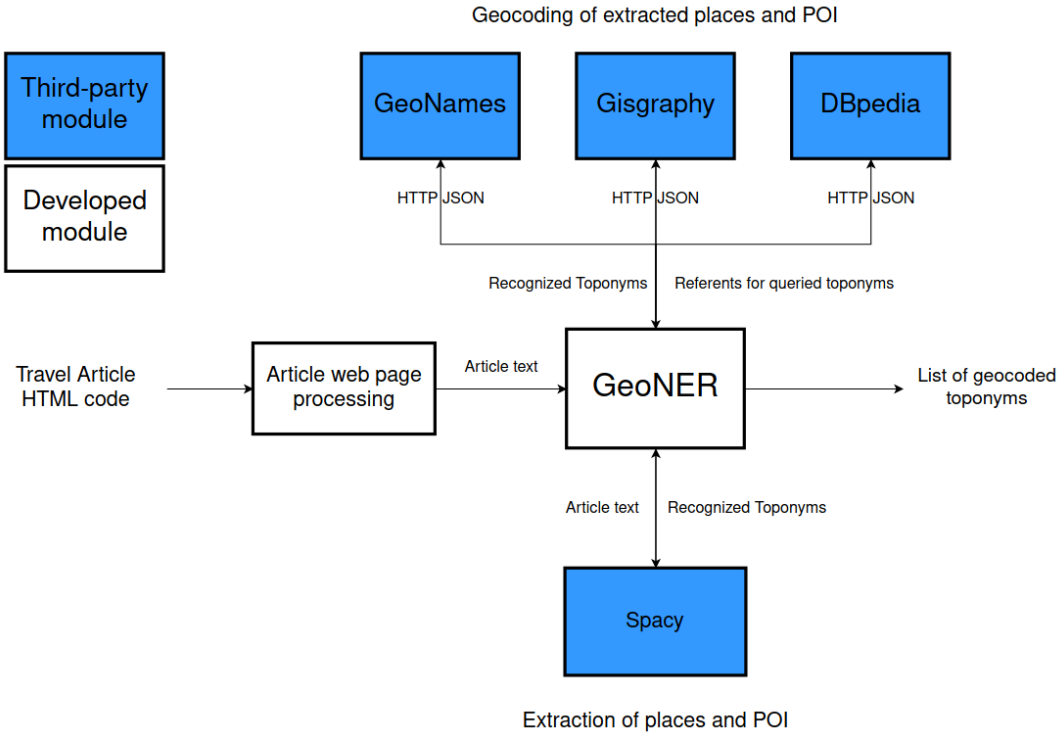

FIGURE 4.1: High-level system architecture

4.2 Article web page processing

As article text extraction constitutes an initial step of the pipeline, handling different page layouts arose. I decided to develop the parser for one of the most common

National Geographic travel story layouts in order to focus on other components of the system.

4.3 Toponym recognition

For extraction of possible toponyms, I considered several open-source tools such as Flair[19], Spacy, AllenNLP[20], StanfordNER[21], and NLTK[22]. The evaluation was conducted with a focus on toponym extraction capability. For each article, the set of target toponyms was created. For each tool, an extractor wrapper was written with a tool-specific list of entity types taken into account for accuracy evaluation. Spacy achieved the best result and was used in the developed geoparsing pipeline.

TABLE 4.1: NER open-source tools accuracy evaluation

Tool Name	Model Name	Accuracy (%)
NLTK	`ne_chunk`	51.52
Spacy	`en_core_web_trf`	95.76
Flair	`ner-english`	89.72
AllenNLP	`tagging-fine-grained-transformer-crf-tagger`	80.18
StanfordNER	`english.conll.4class.distsim.crf`	70.34

4.4 Gazetteer lookup

As a geolocations knowledge base, three gazetteers are used: Geonames, Gisgraphy, and DBpedia. In total, they cover more than 500 million addresses, places, and POI. Due to high RAM and storage requirements for locally hosting all three gazetteers, more than 500 GB for storage, I decided to query all of them using public API. The increased lookup time was an acceptable trade-off for the exemption from satisfying high computational requirements for local setup.

For gazetteer lookup, two modes of querying are used: exact match with the extracted toponym from the text and second more general one when the gazetteer, in case toponym consists of two or more words, is allowed to consider not all the words specified as required. The gazetteer lookup is done in the following sequence. The first stage is to make an exact lookup of an extracted toponym from the text against DBpedia, Geonames, and Gisgraphy. The motivation for querying all three used gazetteers is that they might have different referents for the same place name. This way, they complement each other, and the toponym resolution stage has a broader view of the text's possible geographical region. For determining the similarity between the obtained result and the queried toponym, the Ratcliff-Obershelp algorithm is used through Python `difflib` library. The second stage happens only when the exact match failed; this typically happens when the author either shortens the toponym or, in case a toponym name consists of several words, the author skips some of them. For the general match lookup, different API search parameters are applied for Geonames and Gisgraphy. DBpedia is not used for general lookup due to the enormous list of results; sorting its relevance would require incorporating DBpedia ontologies and increasing time for gazetteer lookup. The obtained lookup results are unified to the same view. They all share the following attributes: place name, latitude, longitude, country, feature code (e.g., park, city, village, restaurant, mountain), first-level administrative division. The unified results are then passed down the geoparsing pipeline to the toponym resolution module.

4.5 Toponym resolution

Toponym resolution is built to resolve referential ambiguities of the recognized toponyms. For example, given the toponyms in a document Paris, Sea Bags, Maine, which location of Paris is the correct referent? Is it (a) Paris, Maine, United States; (b) Paris, France; or one of many other possible candidates for Paris worldwide? To answer this question, I employed a set of toponym resolution heuristics. These heuristics represent toponym resolution insights embedded into the geoparsing system as simple rules and simplifying assumptions. The heuristics used are population heuristics, one sense per discourse heuristics, geographic proximity heuristics, and containment relationship heuristics. The population heuristics prefers higher population referent to lower population referent candidates. The one sense per discourse heuristics assigns only one interpretation across several instances of the same toponym mentioned in the article. The geographic proximity heuristics uses the map information to disambiguate toponyms by minimizing the geographic distance to unambiguous ones. The relationship heuristics uses hierarchical knowledge from gazetteer looking for containment relationships, e.g., Clinton and Oklahoma occurring in the same paragraph or as the bigram Clinton, Oklahoma. The heuristics are used in the following order: the first is containment relationship heuristics; for each ambiguous toponym, it looks for toponym type neighbors in the text and uses hierarchical knowledge to check containment relationship. For example, if in the text Needles is mentioned after Arizona and before California and among the referents, there is one in the state of California, we would disambiguate Needles to the city in California. The next step would be resolving to the regions of the specific toponyms. For example, if unambiguous toponyms are in three regions, we would assign ambiguous toponyms to one of those regions if a referent exists. The next step is to try relationship heuristics again, as the toponyms resolved at the previous step might allow containment relationship check to be applied for other unambiguous toponyms. Next, the population heuristics is used for toponyms in the regions not explicitly mentioned. For the last step, geographic proximity heuristics is used. The number of unambiguous toponyms is already big enough to form region clusters; for referent to be assigned to the toponym, it must be within 150 kilometers to at least 3 unambiguous toponyms. This way of assigning to the clusters also checks for referents that match the extracted toponym, but the gazetteers do not know the place in the region mentioned in the article.

Chapter 5

Results

5.1 Evaluation

The developed geoparsing system was evaluated on ten National Geographic Travel articles as the primary metric Accuracy@161km was used. Accuracy@Xkm is the percentage of toponyms resolved within Xkm of their gold coordinates. The choice of 161 km instead of 5 or 20 is motivated to account for coordinate records from different gazetteers that differ but still lie within the same topographical feature area. The results table presents the percent of toponyms correctly geocoded within 161 km, the percent of toponyms not geocoded due to the gazetteers not knowing about such a place in the mentioned region, and the percent of toponyms not recognized by Spacy. From the obtained results, we can see that the knowledge base highly limits fine-grained geoparsing. The worst results are for articles sixth and ninth; the gazetteers do not know more than a third of the toponyms mentioned. The other limitation of the geoparsing system is the toponym recognition component. For the articles, Spacy failed for non-English toponyms or combinations of words never seen to denote toponyms. The average geocoding accuracy@161km of the developed system is 73 percent.

TABLE 5.1: Geoparsing results

Article ID	Accuracy@161km	Not found in gazetteers	Not recognized by Spacy	Toponym number
1	80.33	8.19	1.60	61
2	88.23	8.82	2.94	34
3	83.33	12.50	0.00	48
4	81.66	8.33	3.30	60
5	68.42	23.68	2.63	38
6	55.00	35.00	5.00	20
7	72.00	20.00	4.00	25
8	80.76	9.61	1.92	52
9	54.54	40.90	4.54	22
10	68.18	9.09	9.00	22

5.2 Visualization

For geoparsing results visualization, unambiguously geocoded toponyms for the top 5 articles by acurracy@160km are plotted on the U.S. map.

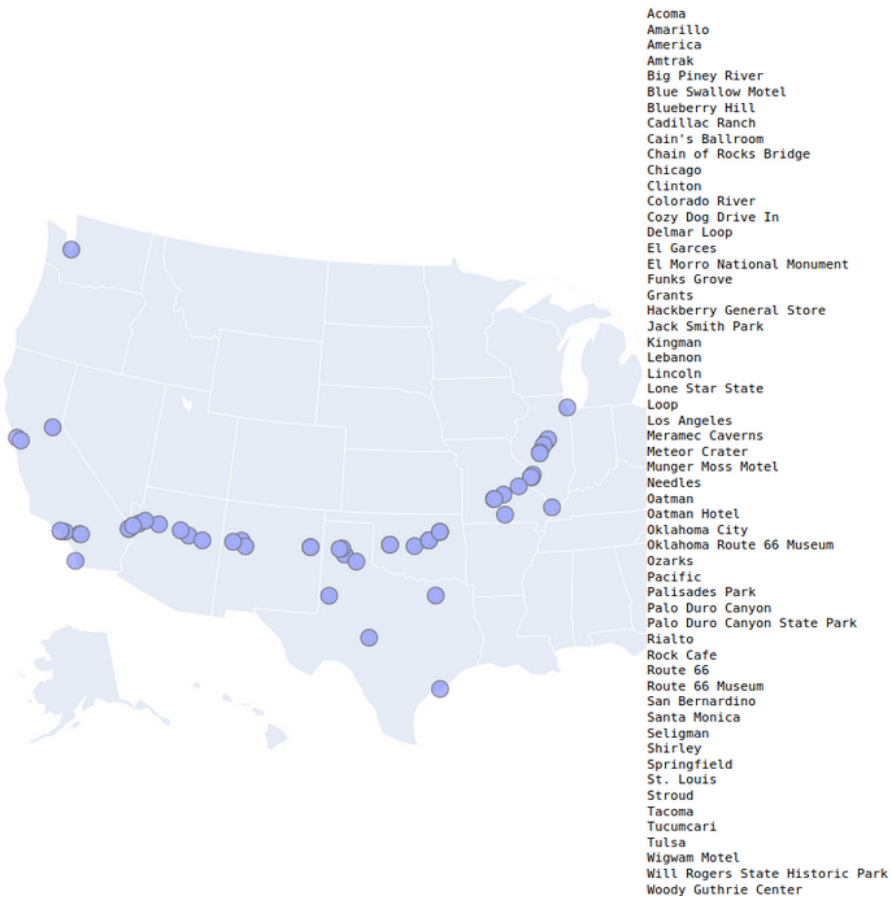

FIGURE 5.1: Route 66 road trip. Article ID: 1

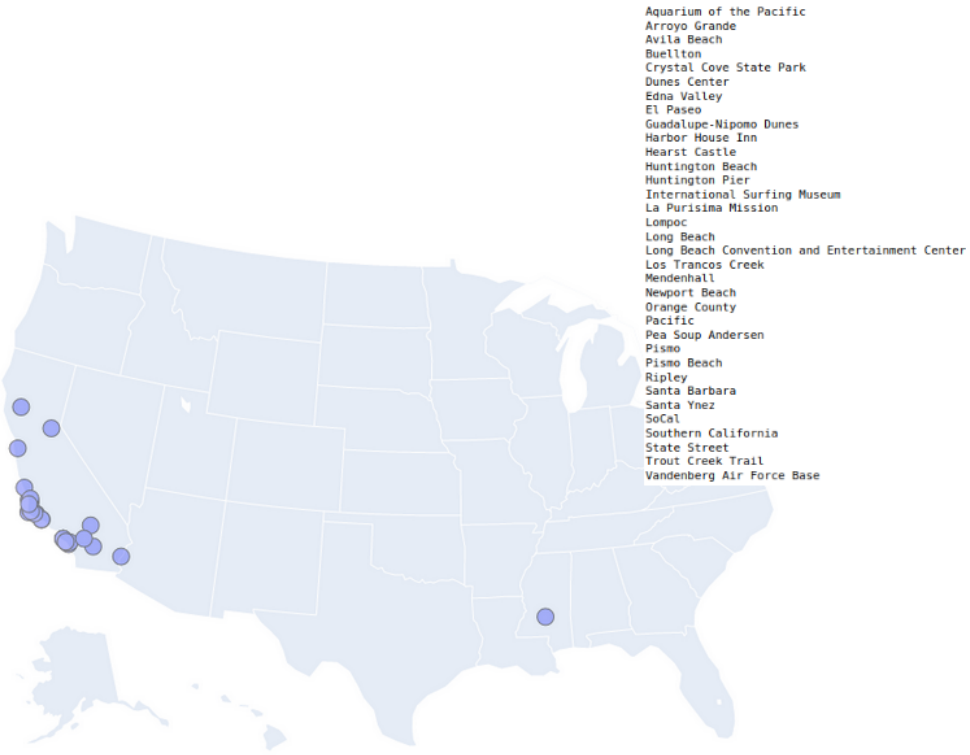

FIGURE 5.2: California route 1 road trip. Article ID: 3

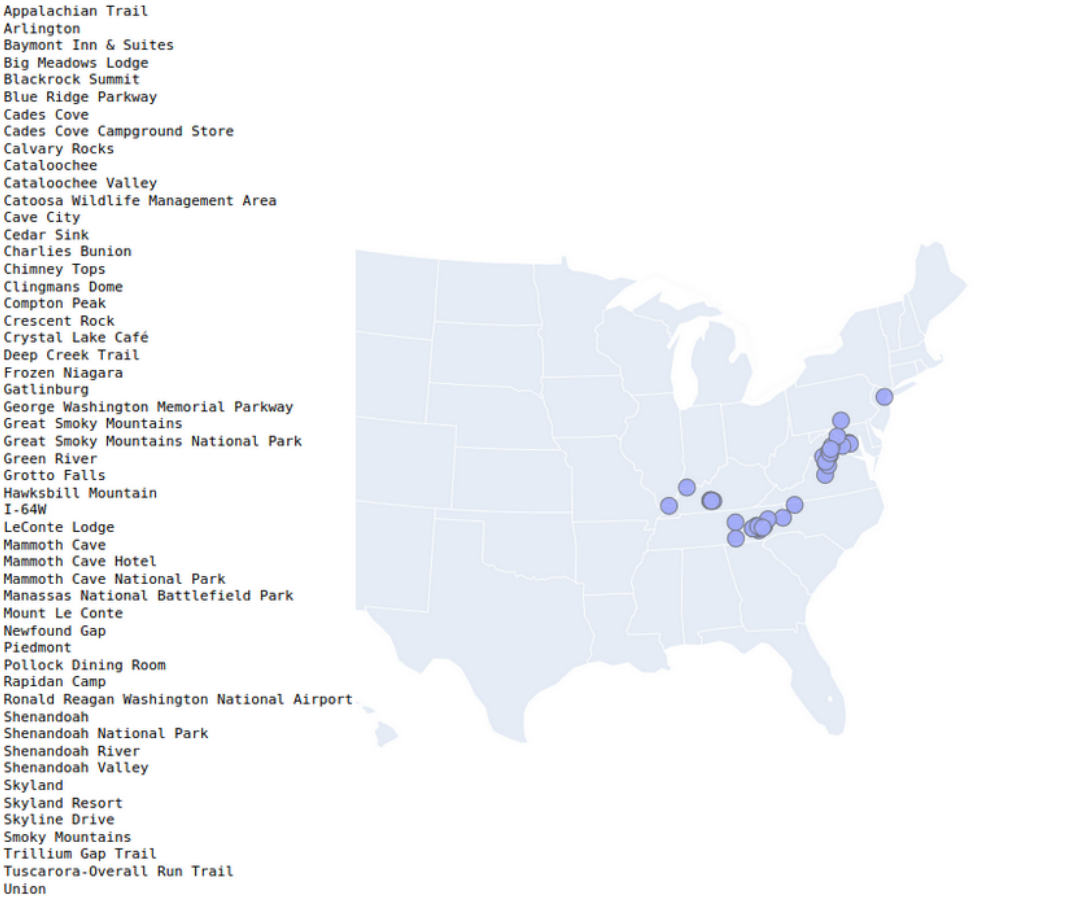

FIGURE 5.3: National Parks Road Trip: East Coast. Article ID: 3

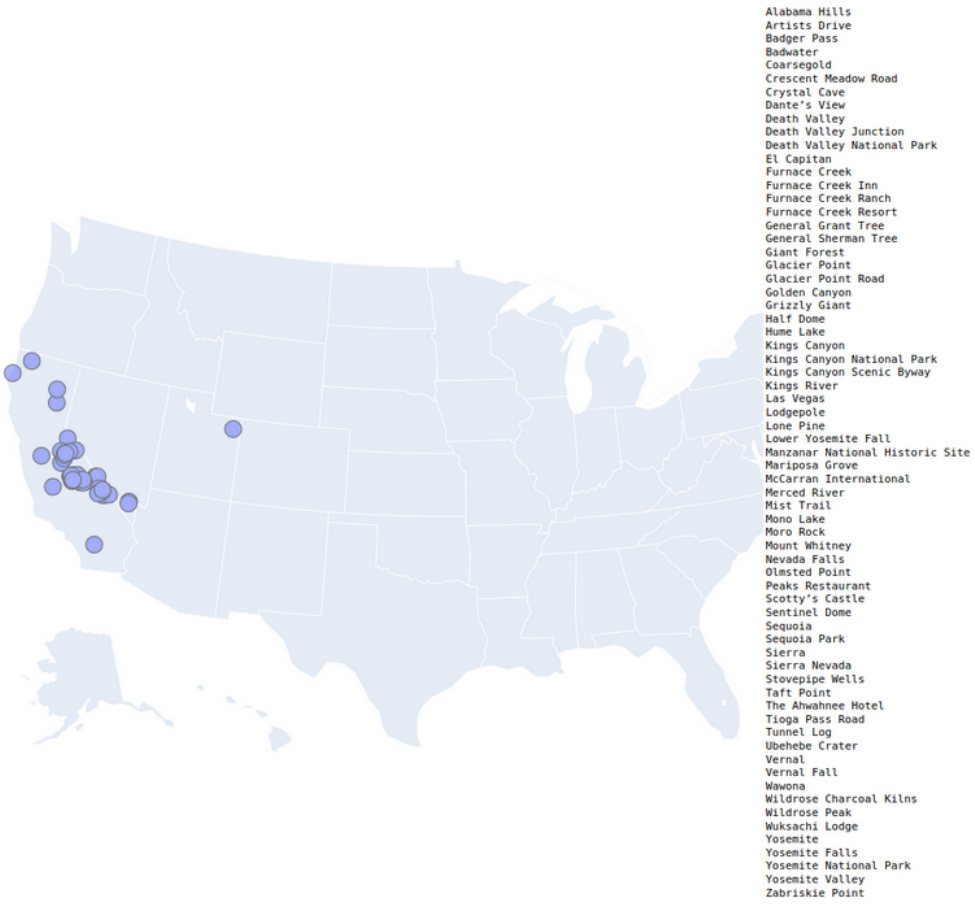

FIGURE 5.4: National Parks Road Trip: California. Article ID: 4

Albany
Bear Mountain
Bear Mountain Inn
Bear Mountain State Park
Beekman Arms
Clermont State Historic Site
Continental Army
Culinary Institute of America
Edward Hopper House Art Center
Eleanor Roosevelt National Historic Site
Goshen
Goshen Historic Track
Greenwich Village
Gunks
Harness Racing Museum
Home of Franklin D. Roosevelt National Historic Site
Hudson
Hudson River
Hudson River Maritime Museum
Hudson Valley
Hyde Park
Kingston
Kingston-Rhinecliff Bridge
Kykuit
Lyndhurst
Mohonk
Mohonk Lake
Mohonk Mountain House
Monroe
Montgomery Place
NY 9A
National Register of Historic Places
New Paltz
Newburgh
Nyack
Old Dutch Church
Old Kingston
Old Rhinebeck Aerodrome
Philipsburg Manor
Rhinebeck
Rondout
Shawangunks Ridge
Sleepy Hollow
Sleepy Hollow Cemetery
Springwood
Staatsburgh State Historic Site
Stockade District
Stony Point
Storm King Mountain
Tappan Zee Bridge
Tarrytown
The Visitor Center
Trolley Museum of New York
U.S. 6
United States Military Academy
Val-Kill
Vanderbilt Mansion National Historic Site
Visitor Center
West Point

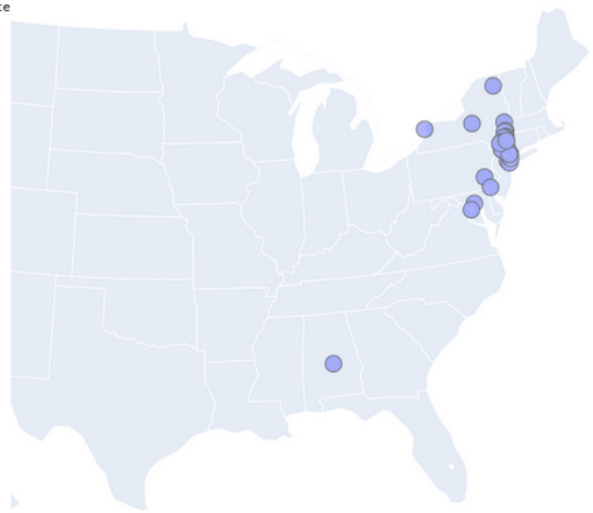

FIGURE 5.5: Road Trip: Hudson Valley, New York. Article ID: 8

Chapter 6

Conclusion

Toponym recognition and resolution are both active research topics and have been around for more than a decade. The approach described in this thesis deals with fine-grained geoparsing of travel articles. Fine-grained includes geocoding the mentioned street names, buildings, restaurants, venues, monuments, landmarks, parks, museums, shops, and other points of interest. From the results, we can see the existing limitations, such as the number of records in the geographical knowledge base and toponym extraction capability of modern NER tools. Nevertheless, the developed system achieved 73% toponym geoparsing accuracy@161km for free-format English travel texts. That signifies the approach's ability to be used as the automation tool for travel itinerary creation, which was the primary use case of the system.

Bibliography

[1] Woodruff, A.G. (GIPSY) Georeferenced Information Processing System. J. Am. Soc. Inf. Sci.1994, 45, 1–44.

[2] Bo, A.; Peng, S.; Xinming, T.; Alimu, N. Spatio-temporal visualization system of news events based on GIS. In Proceedings of the IEEE 3rd International Conference on Communication Software and Networks, Xi'an, China, 27–29 May 2011; pp. 448–451, doi:10.1109/iccsn.2011.6014089.

[3] Gritta, M., Pilehvar, M.T., Limsopatham, N. et al. What's missing in geographical parsing?. Lang Resources Evaluation 52, 603–623 (2018). https://doi.org/10.1007/s10579-017-9385-8

[4] Gritta, M., Pilehvar, M.T. Collier, N. A pragmatic guide to geoparsing evaluation. Lang Resources Evaluation 54, 683–712 (2020). https://doi.org/10.1007/s10579-019-09475-3

[5] Buscaldi, Davide. (2011). Approaches to Disambiguating Toponyms. SIGSPATIAL Special. 3. 16-19. 10.1145/2047296.2047300.

[6] Dewandaru, Agung Widyantoro, Dwi Akbar, Saiful. (2020). Event Geoparser with Pseudo-Location Entity Identification and Numerical Argument Extraction Implementation and Evaluation in Indonesian News Domain. International Journal of Geo-Information. 9. 712. 10.3390/ijgi9120712.

[7] Grover, C.; Tobin, R.; Byrne, K.; Woollard, M.; Reid, J.; Dunn, S.; Ball, J. Use of the Edinburgh geoparser for georeferencing digitized historical collections. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 2010, 368, 3875– 3889, doi:10.1098/rsta.2010.0149.

[8] B.;Technologies, CLAVIN. Available online: https://github.com/Novetta/CLAVIN

[9] D'Ignazio, C.; Bhargava, R.; Zuckerman, E.; Beck, L.; CLIFF-CLAVIN: Determining Geographic Focus for News Articles, Proc. NewsKDD Data Sci. News Publ., 2014.

[10] Teitler, B.E.; Lieberman, M.D.; Panozzo, D.; Sankaranarayanan, J.; Samet, H.; Sperling, J. NewsStand. Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems GIS '08 2008, 2008, 1,, doi:10.1145/1463434.1463458.

[11] Halterman, A., Massachusetts Institute of Technology Political Science Department Linking Events and Locations in Political Text Andrew Halterman, Massachusetts Institute of Technology, 2018.

[12] Imani, M.B.; Chandra, S.; Ma, S.; Khan, L.; Thuraisingham, B. Focus location extraction from political news reports with bias correction. 2017 IEEE International Conference on Big Data (Big Data) 2017, 1956–1964, doi:10.1109/bigdata.2017.8258141.

[13] Morning Consult, 2020. "The 25 Most Trusted Brands in America."January 13, 2019 [Survey Report] Retrieved from https://morningconsult.com/wp-content/uploads/2020/01/Morning-Consult-Reveals-The-Most-Trusted-Brands-of-2020.pdf

[14] Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, Adriane. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

[15] Geonames.org. "Geonames". 2021. Available online: https://geonames.org.

[16] Gisgraphy.com. "Gisgraphy". 2021. Available online: https://gisgraphy.com

[17] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia.. Semantic Web, 6, 167-195.

[18] Akbik, R. (2018). Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1638–1649). Association for Computational Linguistics.

[19] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 54–59).

[20] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.H., Peters, M.E., Schmitz, M., Zettlemoyer, L. (2017). A Deep Semantic Natural Language Processing Platform.

[21] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/ manning/papers/gibbscrf3.pdf

[22] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[23] Andrii Maistruk, GeoNER, (2021). Github repository, https://github.com/kwh44/geoner

[24] Jochen L. Leidner and M. D. Lieberman, "Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language," SIGSPATIAL Spec., vol. 3, no. 2, pp. 5–11, 2011.

[25] B. Baldwinan B. Carpenter. Lingpipe. Available online: http://alias-i.com/lingpipe

[26] Leidner, J.L.Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names, The University of Edinburgh, 2008.