Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 37–45
Marseille, 24 June 2022

Enhancing Geocoding of Adjectival Toponyms with Heuristics

Breno Alef Dourado Sá, Ticiana L. Coelho da Silva, José Antônio Fernandes de Macêdo
Department of Computer Science
Federal University of Ceará, Fortaleza, Brazil
{brenoalef, ticianalc, jose.macedo}@insightlab.ufc.br

Abstract

Unstructured text documents such as news and blogs often present references to places. Those references, called toponyms, can be used in various applications like disaster warning and touristic planning. However, obtaining the correct coordinates for toponyms, called geocoding, is not easy since it's common for places to have the same name as other locations. The process becomes even more challenging when toponyms appear in adjectival form, as they are different from the place's actual name. This paper addresses the geocoding task and aims to improve, through a heuristic approach, the process for adjectival toponyms. So first, a baseline geocoder is defined through experimenting with a set of heuristics. After that, the baseline is enhanced by adding a normalization step to map adjectival toponyms to their noun form at the beginning of the geocoding process. The results show improved performance for the enhanced geocoder compared to the baseline and other geocoders.

Keywords: geocoding, toponyms, adjectival toponyms

1. Introduction

In everyday life, people often use place names to give directions, inform the location of events, and provide spatial information based on the shared knowledge of said names (Vasardani et al., 2013). These references to places, also called toponyms, are often present in documents with geographic content such as news, blogs, and even posts on social media. This geographic information can be used in many applications, such as disaster warning (Wu and Cui, 2018), emergency response (Singh et al., 2019), monitoring of epidemics (Lampos and Cristianini, 2012), crime prevention (Vomfell et al., 2018), news aggregation (Abdelkader et al., 2015), touristic planning (Colladon et al., 2019), among others.

The usage of geographic information embedded in unstructured text requires a process of toponyms extraction and resolution called geoparsing. Geoparsing comprises two steps: geotagging and geocoding.

Geotagging is a particular case of Named Entity Recognition (NER), a Natural Language Processing (NLP) task, which identifies named-entity mentions in texts and classifies them into predefined categories Person, Location, and Organization. For the task of geotagging, only entities corresponding to locations are relevant.

Geocoding is a process of disambiguating, and linking toponyms to geographic coordinates (Gritta et al., 2018b). This is not a trivial task, as it is common to see different locations sharing the same name around the world, for instance, *Springfield, Oregon*, and *Springfield, Queensland*. Moreover, toponyms sometimes appear in adjectival form, e.g., "*Spanish* sausages sales top €2M."

A geocoding technique can be defined as a model G_c such that for a given text T, $G_c(< t_1, t_2, \ldots, t_n >) =< p_1, p_2, \ldots, p_n >$, where t_i, $1 \leq i \leq n$, is a toponym extracted from T and p_i is its corresponding (latitude, longitude) tuple. The latitude and longitude are usually obtained from a gazetteer, a geographic dictionary containing place names and their coordinates.

The geographic information obtained using a geocoder can be used to automatically collect event information from news articles, which researchers may use to observe and extract information on politically relevant events as they occur (Lee et al., 2019). SPERG (Gunasekaran et al., 2018) is one of these initiatives. SPERG focuses primarily on archived newspaper reports on political events and aims to parse the exact event location with high accuracy of every place mentioned in a report. Political scientists require information from these reports for various study purposes, including the impact, attendee profile, and event location.

Another geocoding application relevant and related to political themes is built-in epidemiological early warning systems. First, epidemiological data typically requires time to be available due to time-consuming laboratory tests. Due to its prevalence, social media data, such as Twitter and Facebook, have been used for epidemiological studies on different infectious diseases such as Influenza (Allen et al., 2016), Dengue (Albinati et al., 2017), and COVID-19 (Jiang et al., 2021), among others. By geocoding such text data, the authorities can plan and act appropriately on effective interventions to control infectious diseases, reducing mortality and morbidity in human populations. Another application geared through the use of geocoding information for early conflict warning is ICEWS (O'brien, 2010).

Applications that use the geographic information of unstructured texts need a geocoder capable of assigning the best coordinates for the locations referenced in the text. This task can be a challenge when dealing with toponyms in adjectival form. For instance, consider-

ing the Geonames[1] gazetteer and the text "The *French* President and his foreign minister have been promoting a new course," the expected output of geocoding for the toponym is the tuple (lat=46, long=2), corresponding to the Republic of France. A simple lookup in the gazetteer is not enough to geocode correctly, as "French" is not the country's name, and other places are called in the same manner.

Figure 1 illustrates the problem with a map. Denoted by red markers are several locations named "French" in the gazetteer around the world, and indicated by a blue marker is the Republic of France. Although there is a possible location for the toponym inside France, it is not the entry corresponding to the country. Incorrectly geocoding the toponym could cause an application to treat the text as about a location in the United States of America instead of the French Republic. Thus, it is necessary to treat this kind of toponym somehow.

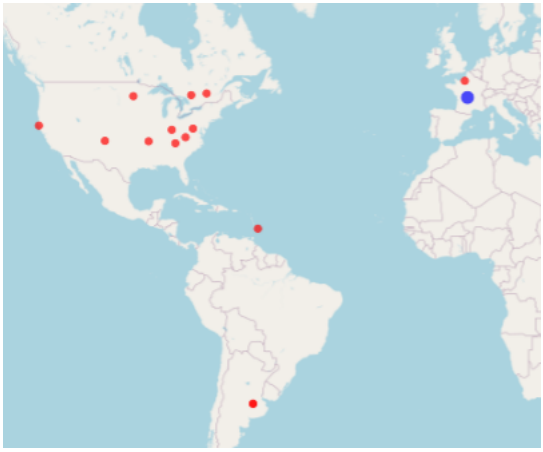

Figure 1: Possible locations for "French" in Geonames.

This paper addresses the geocoding task and aims to improve the process for adjectival toponyms, a type of toponym that other geocoders do not treat. Although adjectival toponyms have already been recognized and annotated in corpora(Kamalloo and Rafiei, 2018; Gritta et al., 2019), geocoders usually either ignore it like CLAVIN[2] or treat it as any other toponym in noun form like CamCoder (Gritta et al., 2018a).

The main contribution of this work is the proposal of a new heuristic to treat adjectival toponyms based on a dictionary of adjectival forms of places. A baseline geocoder is defined through experiments on a set of heuristics. It is further enhanced by adding a normalization step that maps adjectival toponyms to their noun form at the beginning of the process. The experiments confirm that the enhanced geocoder outperforms the baseline.

The structure of this paper is as follows. Section 2 gives a background of related works in the task of geocoding. Section 3 presents the data and methodology used in detail. Section 4 shows the results obtained and a comparison to other geocoders. Finally, Section 5 presents final thoughts and future work.

2. Related Work

Other works on geocoding have varied strategies depending on the focus of the application. Some geocoding plans assign a single location to an entire document, like the one proposed by (Rahimi et al., 2015) to geolocate Twitter users. The approach presented in this paper tries to assign a coordinate to every location referenced in a text. Current toponym resolution methods can be categorized as rule-based, statistical, and machine learning-based.

Several works propose rule-based approaches for geocoding tasks. (Rauch et al., 2003) and (Amitay et al., 2004) use population data as a disambiguation criterion. (Clough, 2005), on the other hand, prioritizes candidate locations with a higher administrative level. (Leidner, 2008) is one of the first comprehensive surveys on geocoding heuristics, addressing methods such as one sense per discourse and geometric minimality.

CLAVIN (Cartographic Location And Vicinity INdexer) is an open-source rule-based geocoder that gets candidates through Lucene[3] with score increments for some fields and values. It performs disambiguation by calculating a score for candidate combinations based on the commonality of countries and states. In other words, when there is more than one candidate for a location, priority is given to the candidate contained in the same administrative region as the precise locations. If the user specifies, CLAVIN also allows disambiguation based solely on population.

Approaches based on statistics seek to solve the problem through distribution models. This strategy is used in several works that focus on the geolocation of entire documents, as in (Butt and Hussain, 2013) and (Hulden et al., 2015), but it can also be applied to individual locations.

The *TopoCluster*, proposed by DeLozier et al. (2015) improves the work of Butt and Hussain (2013) and does the geocoding through pseudo-documents containing the toponym context, using windows of 15 words in each direction. Its resolution works by dividing the world into a grid with 0.5x0.5 degree cells and models the geographic distribution of context words over it. With its hot spots analysis, *TopoCluster* assigns toponyms to the most overlapping cells of the individual word distributions. In the same direction, there is an alternative version of *TopoCluster* called *TopoClusterGaz*. It uses a hybrid geographic dictionary of GeoNames and Natural Earth[4]. This solution searches on the gazetteer at the end of the process and assigns to the toponym the coordinates of the candidate closest to the predicted cell.

[1] https://www.GeoNames.org/
[2] https://github.com/Novetta/CLAVIN
[3] https://lucene.apache.org/
[4] https://www.naturalearthdata.com/

Strategies based on machine learning use trained models to predict the geographic coordinates for toponyms. Among current methods, the usage of bag-of-words representations combined with Support Vector Machines or Logistic Regression has achieved good results (Gritta et al., 2018b).

The CamCoder proposed by (Gritta et al., 2018a) divides the world into a grid. It uses a vector representation called MapVec to model the geographic distribution of the locations mentioned in the text. It uses a deep neural network to predict grid cells for toponyms. It then queries a Geonames database, choosing candidate places based on their population and distance to the predicted cell.

The strategy proposed in this paper also uses a Geonames based gazetteer and does the geocoding task using information such as population and alternate names. However, unlike the aforementioned works, the proposed geocoder in this work treats adjectival toponyms normalizing them to noun form at the beginning of the geocoding process. CLAVIN is the most similar geocoder to the one proposed in this work, but it filters out adjectival toponyms as it doesn't consider them references to places. TopoCluster addresses the same type of named entity but doesn't show effective results (DeLozier et al., 2015). CamCoder does not address adjectival toponyms.

3. Data and Methods

This section describes the methodology for the definition of the baseline geocoder and the enhanced version proposed in this paper.

3.1. Dataset and Metrics

This work uses the toponym taxonomy proposed by (Gritta et al., 2019), in which a toponym is classified based on the semantics of the noun phrase containing it and the context of the surrounding clause. For instance, in the phrase "A former *Russian* double agent was poisoned in the *English* city of *Salisbury*," there is an associative adjectival modifier ("*Russian*"), a literal adjectival modifier ("*English*"), and a literal toponym ("*Salisbury*").

Due to the taxonomy used, GeoWebNews, a dataset also proposed by (Gritta et al., 2019), is used in the experiments. The dataset comprises 200 news articles from globally distributed news sites collected during the first eight days of April 2018. Table 1 presents the GeoWebNews toponym classes according to the taxonomy.

In this work, only the 2401 toponyms annotated with latitude, longitude, and an entry in Geonames are considered. The reason for that is to avoid the types of toponyms as languages and homonyms, which do not have ground truth coordinates as they are not locations, and the most difficult toponyms like festival venues, which do not have an entry in the gazetteer and would require additional resources specific to the domain to be geocoded.

Class	Category	Type
Literal	Literal	Literal
Coercion	Literal	Coercion
Mixed	Literal	Mixed
Embedded_Literal	Literal	Embedded Literal
Literal_Modifier	Literal	Noun Modifier Adjectival Modifier
Demonym	Associative	Demonym
Language	Associative	Language
Metonymic	Associative	Metonymy
Non_Literal_Modifier	Associative	Noun Modifier Adjectival Modifier
Embedded_Non_Lit	Associative	Embedded Associative
Homonym	Associative	Homonym

Table 1: Taxonomy of GeoWebNews classes

The following metrics are used for performance evaluation:

- **Mean Error Distance (MED)**: the mean of great-circle distances[5], in kilometers, between annotated locations and geocoder output locations;
- **Accuracy@X (Acc@X)**: the percentage of toponyms geolocated within X kilometers of the annotated locations. The chosen distance is 161 km (100 miles), previously used in other works such as (DeLozier et al., 2015; Gritta et al., 2019; Wang and Hu, 2019). The reason for that is the possible differences between gazetteer and annotated coordinates;
- **Area Under the Curve (AUC)**: a metric for the overall deviation between geolocated toponyms and ground-truth coordinates. Its value is calculated through the trapezoidal rule[6] using Equation 1, where x denotes the distances, $dim(x)$ is the number of elements in x, and 20039 is the approximated value of half the Earth's circumference in kilometers. The highest possible error is when the output location is diametrically opposed to the expected coordinates on the planet's surface. The better the geocoding, the closer the AUC must be

[5] https://geopy.readthedocs.io/en/stable/#geopy.distance.great_circle

[6] https://docs.scipy.org/doc/numpy/reference/generated/numpy.trapz.html

to 0.

$$AUC = \frac{\int_0^{dim(x)} ln(x)\, dx}{dim(x) * ln(20039)} \quad (1)$$

3.2. Baseline Heuristic Geocoder

The Heuristic Geocoder (HG) used as baseline breaks the task into two steps in which different heuristics can be used. The first step is to obtain the candidates by querying the gazetteer, and the second is the disambiguation. In the end, the geocoder outputs a gazetteer entry for the input toponym.
The geocoder receives a list of toponyms as input and outputs coordinates according to the following parameters:

- **Obtaining Candidates** :
  - **Search Type**: the type of search used for the toponym. "Filter" indicates exact matching, and "Full-text" indicates loose matching;
  - **Ordering**: tells the geocoder if candidates should be ordered by score, feature class, or population;
- **Candidate Disambiguation**:
  - **Top-K Geometric Minimality**: tells the geocoder how many candidates should be considered for disambiguation. If $K > 1$, chooses the candidate closest to previously geocoded locations.

The gazetteer is searched using ElasticSearch[7], a Lucene interface(Divya and Goyal, 2013) that has already shown effective results in geocoding applications due to its dynamic ranking (Clemens, 2015). Before the geocoder usage, an ElasticSearch index is created and populated with Geonames data, including information such as name, alternate names, feature class, and population. To allow filtering and full-text searches on name fields, those are created as text and keyword fields.
The geocoding process is done using the following heuristics:

- **Exact matching (H1)**: consider a place a candidate only when one of its names is exactly equal to the queried text. That means the entry for the *United States of America* is considered a candidate for "*United States*" or "*USA*", which are alternate names in the gazetteer, but not for "*States of America*";
- **Loose matching (H2)**: consider a place a candidate if there is a partial match between one of its names and the queried text. That means querying "*States of America*" will return *USA*'s entry as a candidate;
- **Order candidates by Score (H3)**: ranks candidates based on ElasticSearch default score. The score depends on the place's name and queried text;
- **Order candidates by Feature Class (H4)**: ranks candidates based on their Geonames' feature class. This means the country *Angola* will take precedence over the city *Angola, Indiana*;
- **Order by population (H5)**: ranks candidates based on their population. In this case, the entry for the *Republic of Korea* will take precedence over the one for the *Democratic People's Republic of Korea*;
- **Geometric Minimality (H6)**: minimizes the average distance between all geocoded toponyms. This is done by choosing candidates based on their mean distance to previously geocoded toponyms, assuming places mentioned in a text are as close as possible.

The parameter values for HG used as baseline are defined by evaluating combinations on GeoWebNews and comparing their performances. The one with the best result is chosen and later enhanced to treat adjectival toponyms. Table 2 shows the values for each parameter.

Parameter	Values
Search Type	Filter, Full-text
Ordering	Score, Feature Class, Population
Top-K Geometric Minimality	$K \in \{1, 5, 10\}$

Table 2: HG parameters values

3.3. Enhanced Heuristic Geocoder

When trying to geocode toponyms like *"Australian"* or *"Finnish"* using simple Geonames lookups, even though these words are references to places and can be classified as adjectival modifiers, they are not the places' names. Therefore, such terms are not included as the official names or alternative names on Geonames entries, meaning such toponyms must be normalized before geocoding. That means the usage of a new heuristic:

- **Adjectival Toponym Normalization (H7)**: normalize adjectival toponyms to their noun form at the beginning of geocoding. In this case, instead of querying "*Dutch*", the geocoder will get candidates for "*Kingdom of the Netherlands*".

To do so, an ElasticSearch index is created to be used as the dictionary. The index is then populated with a list of country names, as they appear in Geonames and their

[7] https://www.elastic.co/

adjectival and demonymic forms. For instance, the entry corresponding to the *Kingdom of Denmark* includes the adjectival form "*Danish*", which describes something as being from the country, and the demonymic form "*Danes*", which refers to its people. In this work, the adjectives are taken from Wikipedia's list of nationalities [8].

Thus, the strategy for geocoding adjectival toponyms involves adding a step before obtaining candidates. The toponyms are consulted in the dictionary index and replaced by a normalized version. That means the toponym *Danish* is normalized to *Kingdom of Denmark* before querying the gazetteer. The Heuristic Geocoder enhanced with this strategy is hereafter referred to as HG+.

4. Experimental Results

This section describes the results obtained for the baseline and enhanced geocoders.

4.1. HG Results

Table 3 shows the evaluation metrics for the 5 best parameter combinations ranked by AUC. The best result for every metric was obtained by using a more strict search method and population as the ordering criterion, that is, the combination of H1 and H5. Hence, that was the combination of parameters chosen for HG as the baseline for later improvement.

Geocoder Parameters	MED	Acc@161	AUC
Filter (H1) Pop. (H5) Top-1	**1162.74 ±1612.78**	**0.7713 ±0.2519**	**0.2036 ±0.1801**
Full-text (H2) Pop. (H5) Top-1	1228.28 ±1583.49	0.7187 ±0.2494	0.2729 ±0.1904
Filter (H1) Pop. (H5) Top-5 (H6)	1530.77 ±2106.15	0.6369 ±0.3515	0.3088 ±0.1922
Filter (H1) Pop. (H5) Top-10 (H6)	1613.80 ±2248.68	0.6210 ±0.3601	0.3239 ±0.2826
Full-text (H2) Pop. (H5) Top-5 (H6)	1252.69 ±1742.95	0.6262 ±0.3173	0.3472 ±0.2321

Table 3: Best results for the Heuristic Geocoder parameter combinations

Figure 2 presents the results for HG divided by GeoWebNews classes. Each bar shows the distribution of geocoding outputs for the toponym class, given by the y-axis. The color red indicates toponyms for which no candidates were found in the gazetteer, blue denotes places geocoded to the expected coordinates, and purple indicates location references geocoded to coordinates more than 161 km away from the expected ones. For toponyms of the "Literal" class, direct references to physical locations (e.g. "Harvests in *Australia*"), the geocoder shows a high number of correct predictions. However, for the ones of the "Non_Literal_Modifier" class, toponyms that modify a non-locational concept associated with a location (e.g. "*British* voters"), there are many cases in which no candidates were found or the geocoded coordinates were too far away from the expected.

4.2. HG+ Results

After the baseline geocoder was defined as the combination of H1 and H5, also called HG; it was improved to process adjectival toponyms. Table 4 shows the results for the HG+ in comparison to HG. The enhanced geocoder obtained the best performance in every metric.

Geocoder	MED	Acc@161	AUC
HG (H1 & H5)	1162.74 ±1612.78	0.7713 ±0.2519	0.2036 ±0.1801
HG+ (H1, H5 & H7)	**729.97 ±1340.19**	**0.8188 ±0.2412**	**0.1618 ±0.1669**

Table 4: Results for HG and HG+ on GeoWebNews

Figure 3 presents the results for HG+ split by GeoWebNews classes. Compared to the baseline performance, the geocoding has been improved for most types, increasing toponyms correctly geocoded. Associative modifier toponyms, indicated by the "Non_Literal_Modifier" class, are the ones with the most noticeable improvement.

This difference in performance is due to the new heuristic of processing adjectival toponyms. For instance, considering the sentence "They were found in the southern *English* city of *Salisbury*," HG would assign the coordinates for the town of *English, Indiana* instead of the ones for *England* regarding the adjectival toponym. HG+ can deal with this toponym because of the normalization step added, which makes it search for candidates matching "England."

4.3. Comparison to Other Proposals

HG+ was also compared to other works. The comparisons were done using the ground-truth files[9] provided by the EUPEG (Wang and Hu, 2019). The tests were done on GeoWebNews and TR-News, a dataset proposed by Kamalloo and Rafiei (2018) containing 118 human-annotated news articles from global and local news sources. Both datasets were chosen due to

[8] https://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations

[9] https://github.com/geoai-lab/EUPEG/

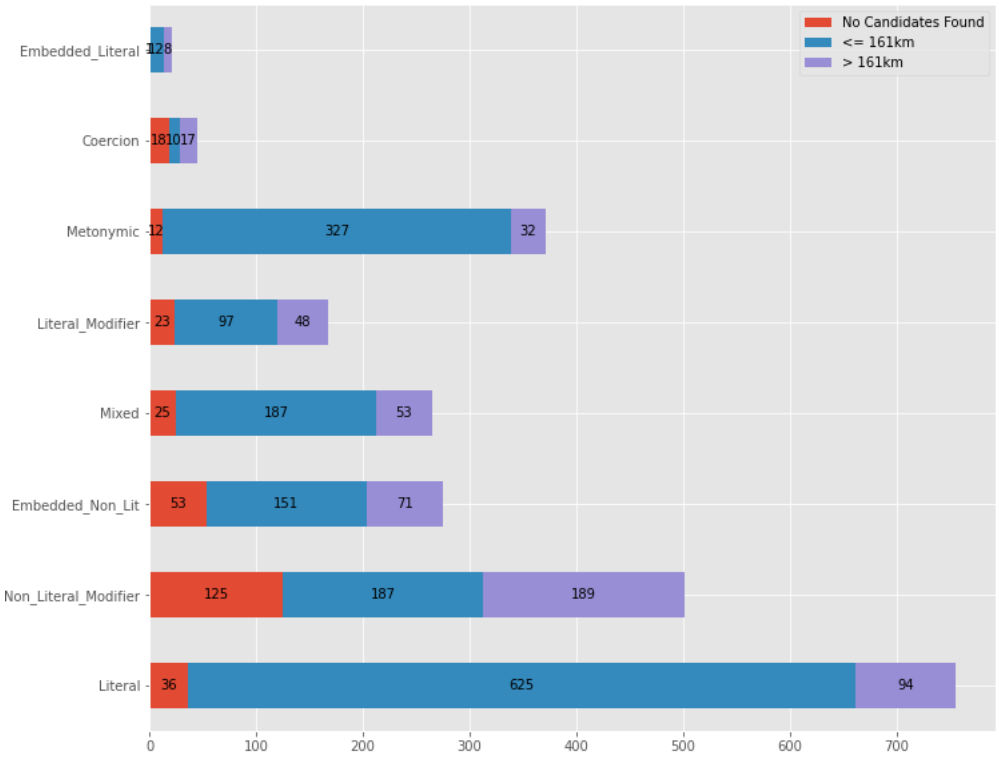

Figure 2: Results for HG on GeoWebNews by toponym type.

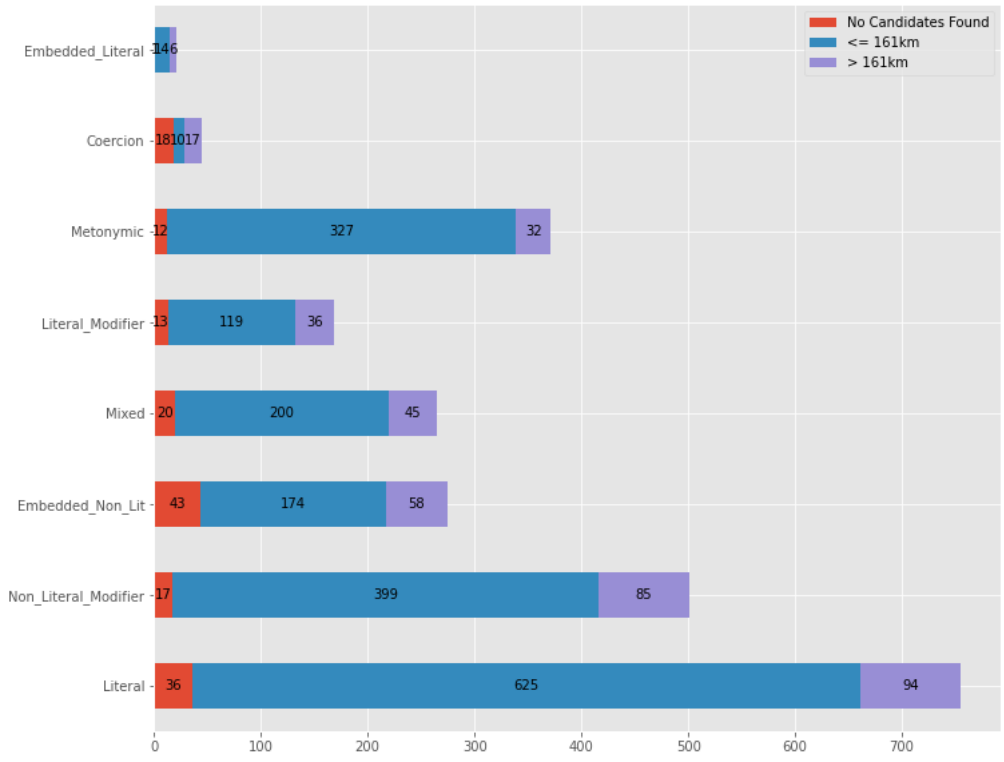

Figure 3: Results for HG+ on GeoWebNews by toponym type.

their coverage of adjectival toponyms[10]. Other well-known datasets, such as Geovirus (Gritta et al., 2018a), provide incomplete or no annotations of adjectival to-

[10] Approximately 14.9% of toponyms in GeoWebNews and 10.7% in TR-News are in adjectival form.

ponyms. Although LGL (Lieberman et al., 2010) also presents this type of toponym, the dataset was not used since its locations are highly region-specific, making them very difficult to disambiguate using the population heuristic with a global gazetteer.

For the CLAVIN geoparser, the REST version[11] was used in the comparison. Because CLAVIN doesn't allow isolated geocoding, HG+ was tested using the toponyms recognized by the geoparser, thus discarding any difference in performance caused by the geotagging process. Table 5 presents the results for CLAVIN and HG+ for GeoWebNews dataset, whilst Table 6 shows the results for TR-News. HG+ achieved the best result for all the evaluated metrics for both datasets.

Geocoder	MED	Acc@161	AUC
CLAVIN	790.18 ±1585.18	0.8100 ±0.3101	0.1268 ±0.2153
HG+ (H1, H5 & H7)	**392.85 ±1105.98**	**0.8432 ±0.2914**	**0.1086 ±0.1739**

Table 5: Comparison to CLAVIN geocoder on GeoWebNews

Geocoder	MED	Acc@161	AUC
CLAVIN	1570.10 ±2685.97	0.7687 ±0.3275	0.1889 ±0.2581
HG+ (H1, H5 & H7)	**1424.23 ±2644.26**	**0.7770 ±0.3346**	**0.1864 ±0.2592**

Table 6: Comparison to CLAVIN geocoder on TR-News

For the CamCoder geoparser, the code available on Github[12] was used. CamCoder allows the execution of its geocoder separately if provided with a formatted ground-truth file. Thus, the ground-truth files for GeoWebNews and TR-News, provided by EUPEG, were used to geocode annotated toponyms as they appear on each text.

Before the comparison, the CamCoder database was updated with the same Geonames dump used to populate the ElasticSearch index in which HG+ operates. CamCoder geocoding was then applied directly to the annotated toponyms, and the same was done for HG+. Table 7 presents the results for GeoWebNews, and Table 8 the results for TR-News. HG+ outperforms CamCoder on GeoWebNews for all three metrics. On TR-News the geocoder achieves better performance for Acc@161 and AUC.

When applied to both TR-News and GeoWebNews, HG+ showed a significant improvement for locations

[11] https://hub.docker.com/r/novetta/clavin-rest

[12] https://github.com/milangritta/Geocoding-with-Map-Vector

Geocoder	MED	Acc@161	AUC
CamCoder	1033.53 ±1527.37	0.7536 ±0.2703	0.2007 ±0.1893
HG+ (H1, H5 & H7)	**729.98 ±1340.19**	**0.8188 ±0.2412**	**0.1617 ±0.1670**

Table 7: Comparison to CamCoder geocoder on GeoWebNews

Geocoder	MED	Acc@161	AUC
CamCoder	**1112.25 ±1566.50**	0.7933 ±0.2429	0.1966 ±0.2005
HG+ (H1, H5 & H7)	1250.09 ±1597.01	**0.8034 ±0.2380**	**0.1956 ±0.2068**

Table 8: Comparison to CamCoder geocoder on TR-News

of the A-class (e.g., countries, mountains, and islands) and the T-class (e.g., mountains, capes, and islands) on GeoNames. For instance, it correctly geocodes the toponyms in "The chancellor of a *Spanish* university [...]," which CLAVIN ignores, and CamCoder wrongfully geocodes to *Spanish, Ontario*. However, as expected of a geocoder based on the population heuristic, locations such as buildings, airports, parks, villages, and sections of populated places are still a problem, especially on TR-News, due to ambiguities like in the case of *Heathrow*, the airport in *London, England*, and *Heathrow*, the suburban community in *Florida, United States*.

5. Discussion and Future Work

This paper proposed the usage of a country adjectives dictionary as a heuristic to improve the geocoding of adjectival toponyms. The proposed geocoder uses ElasticSearch to query a Geonames gazetteer and a dictionary of country adjectives and demonyms. To disambiguate candidates, it uses the population heuristic.

The experiments carried out showed that the processing of adjectival toponyms improved the geocoding performance compared to the baseline. When tested against other known geocoders, it also improved results in both GeoWebNews and TR-News datasets.

For future work, more experiments can be carried out using other datasets to verify differences in performance. The normalization of adjectival toponyms could be improved by adding more adjectives related to other administrative regions such as provinces and cities. Furthermore, processing embedded adjectival toponyms could also improve geocoding.

Acknowledgment

The research reported in this work was supported by the Cearence Foundation for Support of Research (FUNCAP) project "Big Data Platform to Accelerate the

Digital Transformation of Ceará State" under the number 04772551/2020 and "Citizen Platform" under the number 04772314/2020. The first author was partially funded by grant #133938/2019-0, Brazilian National Council for Scientific and Technological Development (CNPq).

6. Bibliographical References

Abdelkader, A., Hand, E., and Samet, H. (2015). Brands in newsstand: Spatio-temporal browsing of business news. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4.

Albinati, J., Meira Jr, W., Pappa, G. L., Teixeira, M., and Marques-Toledo, C. (2017). Enhancement of epidemiological models for dengue fever based on twitter data. In *Proceedings of the 2017 International Conference on Digital Health*, pages 109–118.

Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., and Gawron, J.-M. (2016). Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza. *PloS one*, 11(7):e0157734.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280.

Butt, M. and Hussain, S. (2013). Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Clemens, K. (2015). Geocoding with openstreetmap data. *GEOProcessing 2015*, page 10.

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 25–30.

Colladon, A. F., Guardabascio, B., and Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems*, 123:113075.

DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Divya, M. S. and Goyal, S. K. (2013). Elasticsearch: An advanced and quick search technique to handle voluminous data. *Compusoft*, 2(6):171.

Gritta, M., Pilehvar, M., and Collier, N. (2018a). Which melbourne? augmenting geocoding with maps.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018b). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

Gritta, M., Pilehvar, M. T., and Collier, N. (2019). A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, pages 1–30.

Gunasekaran, A. K., Imani, M. B., Khan, L., Grant, C., Brandt, P. T., and Holmes, J. S. (2018). Sperg: Scalable political event report geoparsing in big data. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 187–192. IEEE.

Hulden, M., Silfverberg, M., and Francom, J. (2015). Kernel density estimation for text-based geolocation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Jiang, Y., Huang, X., and Li, Z. (2021). Spatiotemporal patterns of human mobility and its association with land use types during covid-19 in new york city. *ISPRS International Journal of Geo-Information*, 10(5):344.

Kamalloo, E. and Rafiei, D. (2018). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1287–1296.

Lampos, V. and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–22.

Lee, S. J., Liu, H., and Ward, M. D. (2019). Lost in space: Geolocation in event data. *Political science research and methods*, 7(4):871–888.

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers.

Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, pages 201–212. IEEE.

O'brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International studies review*, 12(1):87–104.

Rahimi, A., Vu, D., Cohn, T., and Baldwin, T. (2015). Exploiting text and network context for geolocation of social media users. *arXiv preprint arXiv:1506.04803*.

Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 50–54.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., and Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, 283(1):737–757.

Vasardani, M., Winter, S., and Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532.

Vomfell, L., Härdle, W. K., and Lessmann, S. (2018). Improving crime count forecasts using twitter and taxi data. *Decision Support Systems*, 113:73–85.

Wang, J. and Hu, Y. (2019). Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419.

Wu, D. and Cui, Y. (2018). Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision support systems*, 111:48–59.