Geoparsing at Web-scale - Challenges and Opportunities

Sheikh Mastura Farzana[1,†], Tobias Hecking[1,†]

[1]German Aerospace Center (DLR), Institute for Software Technology, Linder Höhe, 51147 Cologne, Germany

Abstract

The increasing amount of web data being generated and stored along with geographic information is of great importance to enrich future search applications in science, news, economics, etc.. In addition to location information provided by users or content providers directly, the potential to extract geographic entities from unstructured web content and linking them to geographic coordinates at scale has not been fully exploited. This paper highlights the importance of geoparsing large web archives and associated challenges. Furthermore, this paper evaluates different existing methods with regard to accuracy and scalability to showcase future directions for improving their efficiency.

Keywords

Geoparsing, Geographic information retrieval, Web Data Analysis

1. Introduction

Linking web content with geographic coordinates associated with locations enables a wide range of applications. Examples are, among others, geographical information retrieval, improved situational awareness for crisis management, localised search of real-estate, attractions, or products, or supporting environmental studies.

While many web search applications rely on user reported locations or microformats[1] to link web resources and locations, extracting geographic information also from unstructured content can be of great added value for building a geo-enriched search index.

The process of extracting geographical information from textual data is known as geoparsing. Most general geoparsing techniques comprise two steps, namely geotagging and geocoding. Geotagging is the process of extracting mentions of geographical locations in texts using natural language processing techniques such as Named Entity Recognition (NER). NER techniques [1] identify mentions of entites in texts and associate them to categories (usually including location). Therefore, geotagging can be considered as a sub-task of NER. Geocoding links such place names with geographical coordinates, which typically requires disambiguation of named entities and entity linking to location gazetteers.

GeoExT 2023: First International Workshop on Geographic Information Extraction from Texts at ECIR 2023, April 2, 2023, Dublin, Ireland

*Corresponding author.

†These authors contributed equally.

Sheikh.Farzana@dlr.de (S. M. Farzana); Tobias.Hecking@dlr.de (T. Hecking)

0000-0003-4242-2458 (S. M. Farzana); 0000-0003-0833-7989 (T. Hecking)

[1]e.g. schema.org

While the geoparsing process in itself has several methodological challenges, for example, place name detection, toponym resolution and disambiguation [2], additional issues regarding scalability and data preprocessing arise when it comes to building a large web-index enriched with geo-references. For example, even a small index used for the ChatNoir [3] search engine exceeds the size of 115 Terabytes. Therefore, geoparsing at web-scale cannot only focus on accuracy but also needs to have high robustness and throughput.

This paper will first outline challenges and opportunities of geo-referencing web resources and derive requirements for geoparsing at web-scale. In the second part of the paper, we focus on current solutions for place name extraction from texts and present a first comparative study in terms of precision, recall, F-1 score and runtime.

2. Review on geoparsing methods

There exists a variety of technical approaches to geoparsing. These approaches can be categorised into four sections: rule based approaches, gazetteer matching, learning based models and hybrid approaches that combine 2 or 3 different techniques [4].

Rule-based approaches of geo-parsing mainly involves defining a set of grammar rules to extract location information from text data. These rules may include regular expressions, pattern matching, and lexical analysis. Rule-based approaches are often quick and efficient, but can be limited in their ability to handle complex or ambiguous text data. There have been many works such as [5] and [6] where grammar based simple rules have been used to extract location information. However, with the rise of learning based approaches and their accuracy in terms of location identification, pure rule based approaches are very rare.

Gazetteer matching geo-parsing mechanism involve comparing the location information in text data to a pre-existing database of locations such as GeoNames [2] and OpenStreetMap (OSM) [3], known as a gazetteer. This approach is often effective for accurately resolving location names, but can be limited in its ability to handle ambiguous or incomplete location information.

Learning-based approaches of geo-parsing involve using machine learning algorithms to identify and extract location information from text data. These algorithms are trained on a large corpus of text data and learn to identify location information based on patterns in the data. Learning-based approaches are often more flexible and able to handle complex or ambiguous text data, but can be more computationally intensive and require large training datasets. There are many examples available of such models based on deep learning, entropy based modeling, decision trees etc. [[7], [8]]

Hybrid approaches to geo-parsing are currently the most popular technique for geo-parsing. It involves combining elements of multiple approaches, such as combining rule-based and learning-based approaches, or combining gazetteer matching and rule-based approaches. Hybrid approaches can often provide a balance of accuracy and efficiency, while overcoming some of the limitations of individual approaches. A very successful example of such technique is the GazPNE2 model by Hu et. al [4]. However, similar to learning based models, hybrid approaches containing ML modules suffer from high computational requirements.

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

3. Challenges for building a geoparser for the web

Based on the literature review above we elicit requirements specific for web-scale geoparsers. To the best of our knowledge only a few of them are fulfilled in existing geo-parsing systems.

- **Location Disambiguation -** Location disambiguation also referred to as toponym resolution is one of the main components in any geoparser. Since there can be different places with the same name (e.g. several cities are named Santiago) disambiguation needs context information such as references to close-by locations or geographic distributions of words [9]. Others make additionally use of word embeddings that are known to capture geographic information as well [10]. While these approaches are generally more suitable for large-scale geoparsing disambiguation of toponyms with different names, for example, 'New York' and 'Big Apple' often rely on the use of large gazetteers such as GeoNames. Querying such large databases through APIs or directly can be time consuming, and thus, limits throughput. Consequently, gazetteer-free approaches are preferable in this regard.
- **Location inference from context -** In many descriptions it is observed that a place or an event is sometimes described without mentioning the location. This is especially the case in short notes in microblogs or discussion forums. For example, the sentence 'You should visit this small city on the banks of river Rhine, the former capital of West Germany.' refers to Bonn (Germany). This can be inferred from the given context information although the actual location was not explicitly mentioned. While this is an easy task for humans (at least locals), it can be a challenge for automated location inference. While geographic word distributions and language models can be an approach to this, to the best of our knowledge, inference of implicit location mentions is not explicitly targeted by available geoparsers. This, however, would greatly improve the capabilities of geographic information retrieval since search indexes can be enriched with considerably more geo-information.
- **Semantics of place mentions -** Different geographic web search applications need to focus on different types of place mentions in web data. For example, for local search of shops and businesses addresses of corresponding website providers are most important. When the focus is on situational awareness (e.g. in crisis management, traffic information, etc.) one is interested in event related location information (i.e. a web resource reports about a place), which also has a time component. Other applications, such as touristic information retrieval may focus on web resources that report factual knowledge about a place. In order to support such different types of search application a geographic search index should provide at least some minimal semantics on the context in which a place was mentioned. Related to this is also determining the focus location of a web page if multiple places are mentioned [11]. Thus, an ideal geo-parser for the web should extract such information from text and metadata along with the actual geo-coordinates.
- **Tool chains for web data processing -** Web data usually comes with a lot of boilerplate content, such as advertisements that sometimes contain place names and additional metadata not useful for the task. Furthermore, content classification is necessary to model contextual information. Extracting the relevant information from unstructured web content is an issue not only for geoparsing but also for other information extraction

tasks. While efficient tools for large-scale processing of crawled web data (e.g. stored in WARC[4] files) exist (e.g. [3], geoparsing is not yet an integral part). In this light, a design requirement of a web-scale geoparser is compatibility with such libraries for distributed processing of web data.

- **Scalability and Robustness -** For all the aforementioned points scalability is a serious challenge. At the moment, there is an essential trade-off between speed and accuracy in geoparsing [12]. A web-scale geoparser cannot rely on external API calls nor can it use complex models that cannot process a web document in milliseconds. As we will show later moderate accuracy comes with a massive compromise in terms of speed. For this reason, reducing the gap between accuracy and computing time is considered as one of the most important problems to solve.
- **Lack of annotated data -** For building improved geoparsing models for web data, it is necessary to build large annotated and multilingual web corpora for model building and evaluation. Most available datasets so far cover mostly English resources, are very domain specific, focus only on one type of content (e.g. social media), or have annotation schemes that do not fully fit the tasks described above. Especially a corpus for implicit location inference is missing. Possibilities to mitigate these issues are using available web resources that are annotated with microformats describing location information or using weak supervision techniques for information extraction [13].

4. Comparison of Place Name Extractors

In this section we report an initial comparison study of existing solutions for place name extraction from texts in terms of accuracy and computation time. It is worth to mention that place name extraction is only one but essential part of geoparsing that precedes disambiguation and coordinate association. Comparison of methods beyond place name extraction is left for future works.

4.1. Models

Hu et. al [12] did an evaluation of methods for place name extraction. From this we have selected some of the best performing models with regard to reported time consumption (SpaCy [5], StanfordNER [14], OpenNLP [6], Polyglot [15]) and in terms of accurracy (Flair [16], GazPNE2 [4]) for our experiments. Most of these models are general NER models and thus capable of extracting multiple types of entities. We have extracted only location entities (and geopolitical entities for SpaCy) for our study.

Moreover, as a representative of a very fast but possibly less accurate approach for place name extraction we have developed a set of regular expressions that match different variations of prepositions that precede capitalised words (e.g. 'going/went/gone to', 'in', 'arrive/arriving/arrived at', ...) or succeed possible locations (e.g. 'airport', 'station', ...) based on a prior statistical analysis.

[4]https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/
[5]https://spacy.io/
[6]https://opennlp.apache.org/

4.2. Datasets

- *Dataset* 1 : In order to evaluate time needed by geo-parsing models to process large articles, we downloaded 1000 articles related to populated places from Wikipedia [7]. We used the dataset without annotation hence we cannot calculate accuracy of location extraction by different models on it, but we have calculated possible time requirements to process such data. This dataset is important for our study as the content and length of the articles are similar to web articles thus reflects the time consumption of general web data compared to datasets containing short sentences.
- *Dataset* 2 : For evaluating the accuracy of the models we have used the dataset compiled by Al-Olimat et al[17] of tweets posted during three different floods. The texts are annotated with location information. We have only considered the location type *inLOC* (location the flood took place) and *inLOC* (other locations mentioned). The three datasets contain 4500 tweets combined. We have removed the '#'s and split word such as 'ChennaiFloods' into 'Chennai Floods' for better recognition rates.

4.3. Evaluation

For *Dataset* 1, since the articles are not annotated, we only calculated the time required by different geoparsers as time plays a huge factor in processing large scale data [8].

For *Dataset* 2 precision, recall and F1-score of extracted place names were computed in addition.

Figure 1 shows the time in seconds for processing the Wikipedia articles in *Dataset* 1. It can be seen that there are huge differences in computation time. While the regular expression baseline processes the 1000 documents in milliseconds Flair and GazPNE2 require much more time. This can be possibly attributed to the use of mixed approaches (rules, machine learning models, gazetteers) that are simultaneously applied.

Table 1 reports precision, F1-score and recall for the different methods. It is important to note that some methods cannot identify any place name in certain documents. In these cases it is not possible to calculate the precision and consequently also the F1-score. Hence, for every method the average precision and F1-score (denoted as $precision_{avg}$ and $F1 - score_{avg}$) was only computed on documents where at least one place name was extracted. The numbers in brackets in Table 1 indicate the fraction of documents for which place names could be actually extracted. From the table the trade-off between processing time and accuracy becomes very apparent. The regular expression baseline achieves considerable precision if some of the rules match. However, this is only the case for roughly a quarter of the tweets. The reason is that tweets are often written in a very shortened language. Named entities such as locations, do not always start with a capital letter and are often contained in hashtags that are not captured by linguistic patterns. This is also to some extent the case for SpaCy which is a more sophisticated approach but fails to capture locations that are not properly written in the tweets. It produces only a few false positives but many false negatives. Supporting the results reported in [4], GazPNE2

[7] https://www.wikipedia.org/

[8] Experiments were performed on a machine with Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GH RAM 32 GB

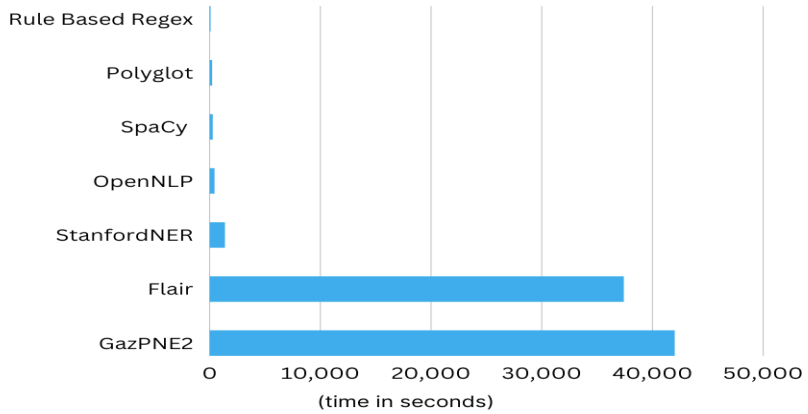

Figure 1: Time consumption of different geo-parsing models on *Dataset*[1],

appears to achieve the highest F-1 score and recall by still having good precision. However, this comes at the cost of the longest runtime.

Table 1
$Precision_{avg}$, $Recall_{avg}$, $F1-score_{avg}$ and Time consumption of different geo-parsing models

-	Rule Based Regex	Polyglot	SpaCy	StanfordNER	Flair	GazPNE2
$Precision_{avg}$	0.64 (0.23)	0.78 (0.75)	**0.84** (0.43)	0.52 (0.72)	0.54 (0.72)	0.64 (**0.88**)
$Recall_{avg}$	0.11	0.47	0.29	0.48	0.45	**0.59**
$F1-score_{avg}$	0.19	0.59	0.43	0.50	0.49	**0.61**
$Time(sec)$	**0.57**	26.17	35.16	5144.96	1859.32	10894.71

5. Discussion

The full potential of geotagging large amounts of web data has not yet been fully exploited. This concerns the linking of web content with geographical information such as earth observation data, environmental and climate studies, as well as economic applications. On the way to a full geo-enriched web search index research has to overcome several challenges, some of them outlined in this paper. Apart from improvements in the quality of geoparsers there is a high need for robust and parallelisable tools that are capable of extracting geo-references from web content at petabyte scale. As our preliminary evaluation showed, there is still a tradeoff between throughput and accuracy already at the geotagging stage of geoparsing, which calls for further research in the regard possibly using mixed methods. For example, simple models can be used for quick but low recall annotations of web resources with coordinates while for particular cases more sophisticated methods can do a refinement.

Acknowledgement

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://-doi.org/10.3030/101070014).

References

[1] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: a systematic review, Computer Science Review 29 (2018) 21–43.

[2] M. Gritta, M. T. Pilehvar, N. Limsopatham, N. Collier, What's missing in geographical parsing?, Language Resources and Evaluation 52 (2018) 603–623. URL: http://link.springer.com/10.1007/s10579-017-9385-8. doi:10.1007/s10579-017-9385-8.

[3] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic chatnoir: Search engine for the clueweb and the common crawl, in: Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40, Springer, 2018, pp. 820–824.

[4] X. Hu, Z. Zhou, Y. Sun, J. Kersten, F. Klan, H. Fan, M. Wiegmann, Gazpne2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models, IEEE Internet of Things Journal 9 (2022) 16259–16271.

[5] L. Zou, D. Liao, N. S. Lam, M. A. Meyer, N. G. Gharaibeh, H. Cai, B. Zhou, D. Li, Social media for emergency rescue: An analysis of rescue requests on twitter during hurricane harvey, International Journal of Disaster Risk Reduction 85 (2023) 103513.

[6] P. Giridhar, T. Abdelzaher, J. George, L. Kaplan, On quality of event localization from social network feeds, in: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), IEEE, 2015, pp. 75–80.

[7] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on gps data, in: Proceedings of the 10th international conference on Ubiquitous computing, 2008, pp. 312–321.

[8] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, D. S. Weld, Knowledge-based weak supervision for information extraction of overlapping relations, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 541–550.

[9] G. DeLozier, J. Baldridge, L. London, Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles (2015) 7.

[10] M. Gritta, M. T. Pilehvar, N. Collier, Which Melbourne? Augmenting Geocoding with Maps, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1285–1296. URL: http://aclweb.org/anthology/P18-1119. doi:10.18653/v1/P18-1119.

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

[12] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location reference recognition from texts: A survey and comparison, arXiv preprint arXiv:2207.01683 (2022).
[13] A. J. Ratner, S. H. Bach, H. R. Ehrenberg, C. Ré, Snorkel: Fast training set generation for information extraction, in: Proceedings of the 2017 ACM international conference on management of data, 2017, pp. 1683–1686.
[14] J. R. Finkel, T. Grenager, C. D. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05), 2005, pp. 363–370.
[15] R. Al-Rfou, V. Kulkarni, B. Perozzi, S. Skiena, Polyglot-ner: Massive multilingual named entity recognition, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 586–594.
[16] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.
[17] H. S. Al-Olimat, K. Thirunarayan, V. Shalin, A. Sheth, Location name extraction from targeted text streams using gazetteer-based statistical language models, arXiv preprint arXiv:1708.03105 (2017).