Towards a toolbox to map historical text collections

Adrien Barbaresi

Towards a toolbox to map historical text collections

Adrien Barbaresi
Austrian Academy of Sciences
Vienna, Austria
adrien.barbaresi@oeaw.ac.at

ABSTRACT

This article presents an effort to integrate spatial and textual data processing tools into a modular software package which features preprocessing, geocoding, disambiguation and visualization.

CCS CONCEPTS

• **Human-centered computing** → *Visualization toolkits*;

KEYWORDS

Information extraction, Geocoding, Maps, Open Source Software

ACM Reference Format:
Adrien Barbaresi. 2017. Towards a toolbox to map historical text collections. In *GIR'17: 11th Workshop on Geographic Information Retrieval, November 30-December 1, 2017, Heidelberg, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3155902.3155905

1 INTRODUCTION

Among the current trends in geographic information retrieval and geocoding [11], the extraction and normalization of named places, itineraries, spatial relations and locative expressions are particularly relevant to study text collections, as advocates of distant reading employ computational techniques to mine the texts for significant patterns and make statements about them [19]. From the point of view of computational linguistics, toponyms are often out-of-vocabulary tokens. As such, they are a potential error source and are supposed to be identified and dealt with. The processing chains usually stop at this point and do not provide geographical visualizations even if the place names can be linked to meta-information such as type and georeference. Besides, publicly available geocoding and cartographic software solutions are scarce and do not usually integrate linguistic information such as annotation layers. This article describes an effort to go more conveniently from texts to maps by integrating several key steps in a modular software package: data curation and preparation, processing of linguistic corpora, geocoding, and projection on maps. The toolkit is meant to be flexible in terms of formats and software environment; it currently focuses on the functions which form the bases of previous studies, with two main goals: provide an up-to-date common ground for hypothesis testing and visualization, since the solution used is not maintained anymore, and ensure replicability in an open science perspective.

GIR'17, November 30-December 1, 2017, Heidelberg, Germany

2 DESCRIPTION

2.1 Approach

Especially for historical corpora, researchers face a lack of general-purpose tooling to analyze geographic references in texts. In order to produce cartographic visualizations, the necessity to adapt to different contexts [3] and to complement existing resources with a precise historical gazetteer [7] has been highlighted, as combined approaches lead to more complete or finer knowledge bases [14]. Such databases of geographic locations featuring coordinates and relevant metadata exist, but their development is challenging [16] even for 20th century Europe [12]. Existing toolboxes, such as AATOS [17], mostly feature candidate extraction and ranking as well as entity linking. HeidelPlace [15] does implement a comparable series of operations but it is currently tied to a series of different engineering decisions (Java, PostgreSQL, Leaflet as output). My approach is more light-weight and extensible, with a similar scope as CORE [10] but with an overall greater focus on usability, text input, integration of registers, and export of maps.

The toolkit is currently developed with historical texts in mind within a generic, language-independent framework. It has been used so far to extract and disambiguate toponyms in different German corpora ranging from the 17th to the 20th century [4–6]. First, the toolkit can be used on raw or previously annotated text, and also in combination with various NLP solutions, e.g. Polyglot [2], as Python is currently the most used programming language in academia.[1] Second, it includes data helpers to bootstrap geographical data, as knowledge-based methods using fine-grained data improve the results [18]. Import filters for the generic gazetteer Geonames and structured data from Wikipedia and Wikidata are available with facilitated download and data cleaning. Third, an additional layer can be activated to bypass geocoding for selected locations.

2.2 From extraction to visualization

The toolkit features an extraction function which operates on token level, using either regular expressions and wildcards on raw text or linguistic annotation layers such as lemma and POS-tags. The extraction is performed by a sliding window which captures single tokens as well as multi-word expressions. This recognition phase is knowledge-based and grounds on gazetteers bootstrapped from existing geographical databases or provided by the user. It can be complemented by external cues like stoplists or linguistic information such as suffixes and derivation.

As it is often necessary to assign the right coordinates to a toponym among several possible ones, a disambiguation process is included. Pouliquen et al. [13] showed that an acceptable precision

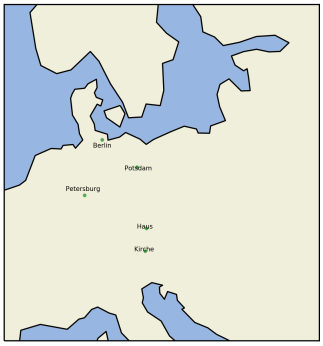
[1] https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

can be reached by applying a series of heuristics based on existing meta-information or various data generated on-the-fly. Accordingly, two different methods [8] have been implemented so far: map-based (geographically relevant contextual information) and knowledge-based (supplied or external meta-information). The information taken into account consists here in type and importance of the entries (as known from data extracted from Geonames or Wikipedia) as well as immediate context (e.g. the expected range and the last countries and locations seen), which can be controlled by user-defined parameters, most notably customized distance calculations, filter level or size of the search radius.

Finally, the toolbox integrates its own visualization component, as the future of the previously used solution (*TileMill*) is unclear. The visualization grounds on the Python module *matplotlib* and its extension *cartopy*[2], which allows for adaptability of projection and design. In this case, the toolkit is meant to leave it open to the user to refine the map, in a particular emphasis on the concept of visualization, most notably regarding the forms and colors used to convey meaning and further for clustering and labeling functions.

2.3 Example

A historical example can be used to demonstrate the impact of the settings on both form and content. The sentence to be analyzed is from the late 19th century and features a series of proper nouns.[3] Geonames is known to be prone to coverage and data quality issues [1]. Figure 1 displays an unfiltered view using raw text and Geonames as gazetteer, where only one point out of five is placed correctly while two other are wrongly considered to be place names, and one place name is missing. Figure 2 shows the impact of both filtering (knowledge-based and POS-based filtering both remove the false positives) and external resources (proper geocoding with a historical gazetteer) which combined lead to the expected result. Additionally, this example shows that quality control and text analysis benefit from the projection of the results on a map.



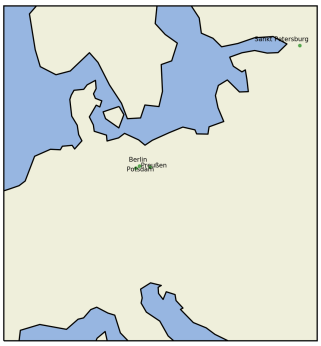


Figure 1: No filter, Geonames Figure 2: Cleaned, meta-infos

[2]http://scitools.org.uk/cartopy/

[3]taken from *Der Stechlin* by Theodor Fontane: *"Ich sage Ihnen, Hauptmann, das waren Preußens beste Tage, als da bei Potsdam herum die 'russische Kirche' und das 'russische Haus' gebaut wurden, und als es immer hin und her ging zwischen Berlin und Petersburg."*

3 CONCLUSION

The main contribution of the toolbox resides in the bundling of spatial and textual data processing in a light-weight system based on the Python technology stack, most notably preprocessing helpers for text and gazetteers, a disambiguation algorithm, and cartographic processing. Three common issues in geographic information extraction [9] are addressed: detecting geographical references, disambiguating place names, and developing effective user interfaces. The toolbox targets historical collections but is not limited to them. The streamlined process from text to map involves a series of decisions as well as a critical reading of the map. Furthermore, being able to go through all the process in one shot is well-suited to spot methodological or data-related problems. I provided an example of the issues raised by generic extraction and showed the benefits of integrated data cleaning and filtering. The toolkit is designed to be modular, it can be extended in its functionality or integrated into third-party software. Future work includes further geocoding techniques as well as benchmarking functions. A working alpha version is available along with the code under an open-source license.[4]

REFERENCES

[1] Dirk Ahlers. 2013. Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on GIR*. ACM, New York, 74–81.

[2] Rami Al-Rfou et al. 2015. Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 586–594.

[3] Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. 2015. Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing* 9, 1 (2015), 15–35.

[4] Adrien Barbaresi. 2016. Visualisierung von Ortsnamen im Deutschen Textarchiv. In *DHd 2016*. DH im deutschprachigen Raum e.V., Leipzig, 264–267.

[5] Adrien Barbaresi. 2017. Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel. In *Digital Humanities 2017*. ADHO, Montréal, 159–161.

[6] Adrien Barbaresi. 2018. A constellation and a rhizome: two studies on toponyms in literary texts. In *Visual Linguistics*, Bubenhofer Noah and Kupietz Marc (Eds.). Heidelberg University Publishing, Heidelberg. To appear.

[7] Lars Borin, Dana Dannélls, and Leif-Jöran Olsson. 2014. Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing* 29, 3 (2014), 400–404.

[8] Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *SIGSPATIAL Special* 3, 2 (2011), 16–19.

[9] Christopher B. Jones and Ross S. Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science* 22, 3 (2008), 219–228.

[10] Eetu Mäkelä, Thea Lindquist, and Eero Hyvönen. 2016. CORE – a Contextual Reader Based on Linked Data. In *Digital Humanities 2016*. ADHO, 267–269.

[11] Fernando Melo and Bruno Martins. 2017. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS* 21 (2017), 3–38.

[12] Paolo Plini, Sabina Di Franco, and Rosamaria Salvatori. 2016. One name one place? Dealing with toponyms in WWI. *GeoJournal* (2016), 1–13.

[13] Bruno Pouliquen et al. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC*. ELRA, Genoa, 53–58.

[14] Thomas Rebele et al. 2016. YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In *ISWC 2016: 15th International Semantic Web Conference*, Paul Groth et al. (Ed.). Springer, Cham, 177–185.

[15] Ludwig Richter, Johanna Geiß, Andreas Spitz, and Michael Gertz. 2017. HeidelPlace: An Extensible Framework for Geoparsing. In *Proceedings of EMNLP 2017: System Demonstrations*. ACL, Copenhagen, 85–90.

[16] Humphrey Southall, Ruth Mostern, and Merrick Lex Berman. 2011. On Historical Gazetteers. *International Journal of Humanities and Arts Computing* 5, 2 (2011), 127–145.

[17] Minna Tamper et al. 2017. AATOS – a Configurable Tool for Automatic Annotation. In *International Conference on Language, Data and Knowledge*. Springer, Cham, 276–289.

[18] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a Free Collaborative Knowledge Base. *Commun. ACM* 57, 10 (2014), 78–85.

[19] Clifford E. Wulfman. 2014. The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies* 5, 1 (2014), 94–109.

[4]https://github.com/adbar/geokelone