Exploiting multiple heterogeneous data sets for improving geotagging quality

Discussion paper

Laura Di Rocco[1], Roberto Marzocchi[2,3], Barbara Catania[1], Tiziano Cosso[3], and Giovanna Guerrini[1]

[1] DIBRIS, Università degli Studi di Genova
[2] DICCA, Università degli Studi di Genova
[3] Gter srl

Abstract. Geotagging is the process of associating with textual data items the geographic position they denote, usually in the form of geographical coordinates (latitude and longitude). Automatic geotagging is often trivial relying on one of the many available gazetteers, such as OpenStreetMap (OSM)[1]. However, such knowledge bases are not free of errors, and, while this simple match works for popular locations, automatic annotation of less relevant venues and events may be significantly inaccurate. The goal of this work is to increase geotagging quality (in terms of completeness and accuracy) by also identifying and jointly exploiting diverse data sources as gazetteers. This will also allow us to cope with ambiguity by additionally performing semantic queries in various open knowledge bases.

1 Introduction

The ability to geolocate a user (i.e., to assign a geographic position) enables a number of location-aware applications and services. Similarly, associating geolocation information to data enables managing and analyzing them on the basis of the geographic dimension. When no explicit georeferencing (e.g., in the form of location metadata coming from the application providing data) is available, geolocation can be implicit and can be inferred, with variable degree of confidence, by the data itself, which may contain, for example, names of entities with known spatial location. Georeferencing by place name is an informal and the most common form of georeferencing approach. We commonly use place names in conversations, correspondence, reporting, and documentation.

Data are said to be implicitly (or indirectly) georeferenced [6] when they are not associated with explicit geospatial references (such as positioning on maps or spatial coordinates), rather they are referenced by place names, geocodes, and addresses. The term geotagging highlights the fact that additional steps are required to identify the locations on maps.

[1] `openstreetmap.org`

More preciselly, geotagging is the computational process of transforming a textual location description[2] into a location on the Earth's surface (spatial representation in numerical coordinates). The geotagging process therefore produces a function that, given a toponym, returns the corresponding coordinates.

Automatically geotagging is usually trivial using one of the many available curated geographical knowledge bases. Dictionaries of placenames are called gazetteers [4]. Gazetteers contain descriptive information about named places, which can include their geographic locations, categories, etc. However, such gazetteers are not free of errors, and while automatic annotation is actually easy for popular locations, automatic annotation of less important venues and events may be significantly inaccurate. This low accuracy is not only introduced by human mistakes in the strings, but it is inherent in the data which could be ambiguous due to the multiple, context-dependent, interpretations toponym. For example, even for the case of very popular toponym as "Eiffel Tower", there is a replica in the city of Paris, Texas, USA, yielding a second match, with different geographical metadata.

The quality problems that may affect geotagging can be classified in three categories: (i) accuracy/correctness issues: datasets can contain human mistakes; (ii) completness issues: no metadata are available; (iii) consistency issues leading to ambiguity: when names of places are generic and there are no additional information to devise whether a name refers to a village or to a mountain and so on. These ambiguity problems are usually difficult to solve even with human intervention.

In this paper, we propose to cope with these issues by jointly exploiting heterogeneous datasets and relying on strong assumptions on the geographical domain, in terms of: (i) a specific geographical area, and (ii) a specific domain of interest the geonames refer to.

Diverse data sources are exploited in order to achieve a fast and accurate automatic geotagging, allowing to cope with both human induced errors, gazetteer incompleteness, and ambiguity. In particular, we combine the usage of a geotagging approach relying on a crowdsourced gazetteer containing toponyms related to the specific domain, with semantic searches on dataset, like DBpedia[3], in order to disambiguate toponyms and achieve a higher accuracy.

The advantage of using a combined approach is due to the specific characteristics of the different datasets. Crowdsourced gazetteers help us in addressing the problems related to the use of vernacular and very specific names. On the other hand, semantic sources help to exploit a semantic layer with the aim of achieving an higher precision and a lower ambiguity. As a specific use case, we will consider data from an outdoor tourism domain.

Related work. Geotagging consists of toponym recognition and toponym resolution. Toponym recognition assigns a possible geographical metadata (e.g., matching the toponym in a gazetteer) to geographical name. Toponym resolu-

[2] In the rest of the paper, we will use the words *toponym*, that means placename, and *geoname*, as synonyms, to refer to textual location descriptions.

[3] `dbpedia.org/`

tion, instead, eliminates the geo/non-geo ambiguity (e.g., Washington can be a city in the USA or the name of a person). There are different strategies to do this: (i) finding names in the text that exist in a gazetteer [1, 5]; (ii) using Name Entity Recognition techniques [8, 9]; (iii) using a geographic ontology (for understanding the context) [9, 8]. In general, a lot of studies in geotagging are related to the analysis of social media data [3, 2]. Moreover, implicit georeferencing information has been exploited, for instance, for localizing news on maps [10].

Outline of the paper. The remainder of the paper is structured as follows: Section 2 states the problem and presents the proposed solution, Section 3 illustrates the specific scenario, and Section 4 concludes by discussing future work.

2 Problem Statement and Proposed Approach

In this section, we first present the formalization of the problem, starting from a simple geotagging function and extending it to cope with issues that may impact geotagging quality. Then, we present a specific istantiation of the proposed approach.

Geotagging function. Given a set of input toponyms, belonging to some texts or to an input dataset, we want to define a geotagging function associating geographical metadata with these toponyms.

Let $C \subseteq \mathbb{R}^2$ be a set of geographical coordinates corresponding to locations in the physical world, and T be a set of toponyms, i.e., strings denoting locations. Let G be a gazetteer, i.e., a set of pairs of the form (t, c), where $t \in T$ is a toponym referring to a specific location $c \in C$ (coordinates).

The simplest geotagging function f_G relying on a gazetteer G is defined as: $f_G : T \longrightarrow C$, where $f_G(t) = c$ s.t. $\exists!(t, c) \in G$. We notice that f_G is partial since it is undefined for t if (i) $t \notin \pi_1(G)$ or (ii) several pairs in G for t are found.

Assuming that our dataset is split into specific known geographic bounding boxes define as $BB \subseteq C \times C$, e.g., country or geographical regions. We can now define the geotagging process related to a specific geographical area as a bounding box $b \in BB$ represented as a pair of coordinates. We define a function g relying on G that, given a specific toponym $t \in T$ and a specific $b \in BB$ returns its coordinates $c \in C$. More formally, $g_G : T \times BB \longrightarrow C$ where $g_G(t, b) = c$ if $(t, c) \in G$ and c is included in bounding box b. The g function is partial but the point (ii), explained above, is partial solved.

In Figure 1, we graphically depict the geotagging function for a specific bounding box.

Issues in geotagging and proposed approach. There are a number of issues in realizing the geotagging function g_G: (i) typos and alternatives names, (ii) G incompleteness, (iii) ambiguity. To avoid these issues, in the following we discuss how we extend the geotagging function g_G.

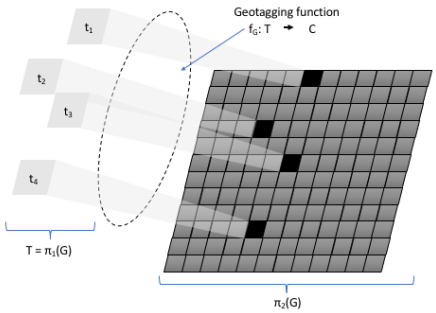

Fig. 1. The geotagging function g_G.

Naming differences and typos. To solve typos and alternative name problems that toponyms can have, we introduce a correction function γ. For example, we can have different toponyms like "Tour Eiffel", "Eiffel Tower" or "eifel tower" and we would be able to geotag correctly "Tour Eiffel" in all these cases. To do this, we define γ as follow:

$$\forall t \in T \ \gamma(t) = t \text{ if } \gamma(t) \subseteq \pi_1(G)$$

$$\gamma(t) = t' \text{ otherwise}$$

where $t' \in T$ is the toponym closest to t according to a string similarity function (and s.t. t and t' similarity according to such function is above a fixed threshold).

Therefore, we define function h_G as follows:

$$h_G(t, b) = g_G(\gamma(t), c)$$

Dataset incompleteness and ambiguity. The problem in the application of function h_G is that, in a real scenario, it is difficult to have a set G complete for a specific set of toponyms T. Since individual datasets are incomplete, we propose the combined use of different datasets for which an order has been specified. The use of a sequence of gazetteers works like a filter on T. Performing geotagging on different gazetteers, we can also disambiguate toponyms. Ambiguity arises when several alternative geographical coordinates are found for one toponym. The same name may indeed denote different geolocated objects.

Let $\mathcal{G} = G_1, ..., G_n$ be a sequence of gazetteers, we define a new function k on $\mathcal{G}$ as:

$$k_{\mathcal{G}}(t, b) = c \text{ if } h_{G_i}(t, b) = c \wedge \forall j < i.(t, c) \notin G_j$$

Note, therefore, to cope with ambiguity, we rely on two approaches:
- the usage of bounding box $b \in BB$ in order to reduce the geographical area in which we search a toponym (see g_G).
- the usage of multiple gazetteers G_i in order to understand exactly which toponym we found (see $k_{\mathcal{G}}$).

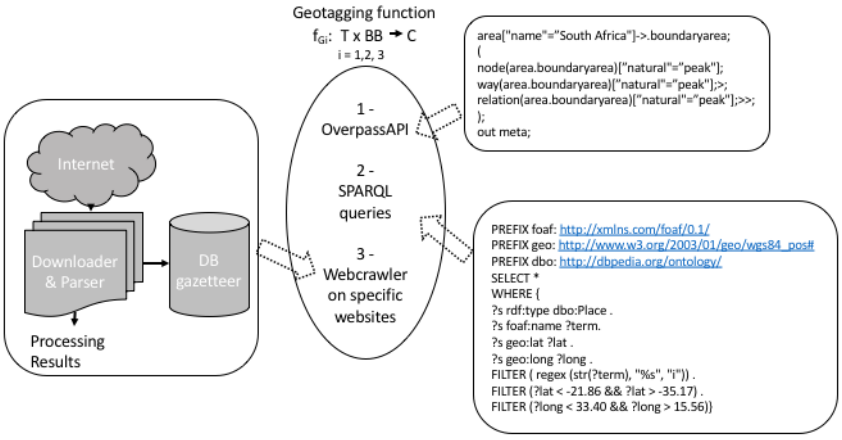

Fig. 2. Graphical representation of the geotagging function $k_{\mathcal{G}}$.

Instantiation of geotagging function $k_{\mathcal{G}}$ We describe a specific solution exploiting three different gazetteers. The described solution relies on the assumption that toponyms refers to a specific domain of interest. In details, as we can see in Figure 2, the istantiation of $\mathcal{G}$ is:
1. G_1 : OpenStreetMap(OSM)
2. G_2 : DBpedia.
3. G_3 : specific domain websites.

Our function $k_{\mathcal{G}}$ cycles on $\mathcal{G}$ in order. In the follow, we describe in detail G_1, G_2 and G_3. The way we query these gazetteers is the implementation of h_{G_i}. The bounding boxes provided as input to the functions for different G_i are the same.

OpenStreetMap We use OSM and OverpassAPI as a geotagger. The use of OSM helps us to extract vernacular toponyms. Moreover, knowing a specific tag of the toponym, we are able to assert the correctness of our preliminary solution. In this specific case our correction function γ is the Levenshtein distance[7]. After this process, for each toponym three different results may be obtained : (i) a single solution, (ii) more that one solutions, (iii) no solutions. As discussed before, toponyms in (i) are solved, while for the other ones, the new gazetteers will be used.

DBpedia We use DBpedia for data geotagging. We query DBpedia with a SPARQL endpoint (see the query in Figure 2). The use of semantic information helps us to disambiguate in case of multiple matches coming from the previous step. Moreover, in case of extraction of new toponyms, if we extract more than one solution,

we are able to choose the correct one. This is because we know the meaning of this object. With this gazetteer, the correction function γ in the filter option of our SPARQL query. We use regex function in order to find similar strings.

Web sites We query specific information from the web that we know being related to our specific domain. We create a web crawler on websites related to the specific domain and we extract from them the coordinates of our toponyms. This implementation is still in progress and we cannot provide more details about it.

3 An Application to the Outdoor Tourism Domain

As an application of the proposed approach, we consider a dataset containing toponyms related to the touristic outdoor domain. The dataset used in this application is a private dataset. We are, thus, not allowed to share the data.

The entire dataset contains more or less 50000 toponyms from all over the world. The toponyms are split in three different types of area: data from administrative areas (e.g., Ande), countries (e.g., South Africa) and geographical areas (e.g., Caucaso). Therefore, the possibility to split the dataset in subsets related to a specific geographic bounding box is due to information provided in the dataset itself (and provided by the data owner). These toponyms are not uniformly distributed over the world. In order to provide some information related to the sparsity of the dataset, in South Africa we have $\sim$300 toponyms, in the Ande area we have $\sim$3000 toponyms and in Causaso we have $\sim$100 toponyms.

The data are structured and each data item contains the following fields: toponym name, toponym name without typology, main locality (nullable field) and altitude (nullable field). These data have been collected over 30 years from different sources (e.g. books, journals, website, ecc.). These data are very noisy and imprecise but the expertise of the owner of the data can give us the assurance that the data are related to the touristic outdoor domain only.

Since we have toponyms from all over the world, our algorithm works on separate subsets of this domain. In case of administrative areas and countries, we have specific BBs retrieved from geographical gazetteers. The geographical area has a different problem: we must choose *a priori* the relative BB. In this case we can have two different problems that could increment errors retrieving toponyms: (i) we can take a BB that does not correctly cover the entire area, therefore loosing some toponyms, (ii) we can take a "big" BB and increment the ambiguity in retrieving toponyms.

Following the process explained in Section3, first of all we clean the OSM data by filtering out geographic information that are not useful in the outdoor touristic domain. Relying on this external knowledge, we perform a first level of geotagging. When we obtain only one solution, we can already assume that we found the correct geographical location. Unfortunately, we have a lot of toponyms without a location. On this new subset of data, we will use the second gazetteer. With a specific SPARQL query on DBpedia, we obtain new results without ambiguity. We can see in Figure 3 an example of ambiguity coming from the

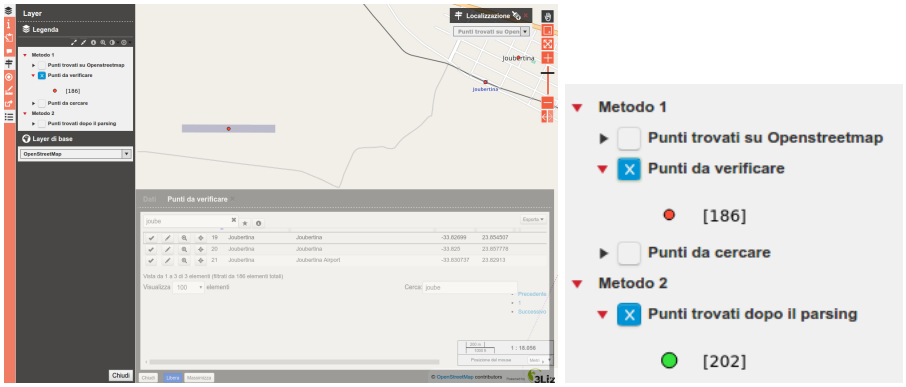

Fig. 3. Our user interface with an example of ambiguity.

first geotagging step with OSM. We provide two examples to better illustrate the method:
1- Toponym Cockscomb: It does not exist in OSM. We retrieve this information from DBpedia. This is an example of incompleteness that we solve using different gazetteers.
2- Toponym Joubertina: Three different points are retrieved in OSM. We found it in DBpedia as a village. This is an example of ambiguity. We solve it with the combined use of a correction function γ and multiple gazetteers.

Gazetteers, OSM and DBpedia, are not enough to geotag the entire set of toponyms. The third gazetteer will allow us to find the remaining location.

Using the geotagging function, we obtain a set of toponyms with the corresponding location metadata. As a real scenario, we do not want to compute error measures such as precision and recall, but we use a web visual interface (as the one shown in Figure 3) available on the web platform `www.gishosting.gter.it`, implemented using the QuantumGIS open source software, in order to get a feedback from the end user of the dataset.

We notice that we are able to show the user the possible location of a toponym. In this case, the end user can assess the correctness of geographical coordinates.

We report information related to the percentage of toponyms that we are able to retrieve. We cannot report all the results because this is work in progress and we have to perform the analysis on the whole world, but on the set of toponyms of South Africa, we geotag 50% of the toponyms with OSM and 20% with DBpedia solving some of the arising ambiguities.

4 Conclusions

In this paper we have proposed the joint use of different datasets for improving the quality of geotagging and we have applied the proposed approach in a specific scenario. The motivation of this work is to meet the needs of the data owner.

We introduced a geotagging function $k_{\mathcal{G}}$, taking as input a toponym and a specific geographic bounding box. The function, iterating over multiple gazetteers, returns the geographical metadata associated with a toponym. Our function does not perform a simple string matching rather it uses a specific correction function for each gazetteer. The consciousness of the specific geographical domain allows us to select and to decide how to use the relevant gazetteers.

This approach is still work in progress. Our future efforts will be the implementation of the web crawler to increase completeness and to complete the experimental evaluation.

References

[1] E. Amitay, E. Amitay, N. Har'El, N. Har'El, R. Sivan, R. Sivan, A. Soffer, and A. Soffer. Web-a-where: geotagging web content. *Proceedings of SIGIR '04 conference on Research and development in information retrieval*, pages 273–280, 2004.

[2] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768. ACM, 2010.

[3] J. Gelernter and N. Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.

[4] M. F. Goodchild and L. L. Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.

[5] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889, 2010.

[6] L. L. Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.

[7] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[8] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of spirit: A spatially aware search engine for information retrieval on the internet. *Int. J. Geogr. Inf. Sci.*, 21(7):717–745, Jan. 2007.

[9] N. Stokes, Y. Li, A. Moffat, and J. Rong. An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science*, 22(3):247–264, 2008.

[10] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM, 2008.