Georeferencing Wikipedia pages using language models from Flickr

Chris De Rouck[1], Olivier Van Laere[1], Steven Schockaert[2], and Bart Dhoedt[1]

[1] Department of Information Technology, IBBT, Ghent University, Belgium, {chris.derouck,olivier.vanlaere,bart.dhoedt}@ugent.be
[2] Department of Applied Mathematics and Computer Science, Ghent University, Belgium, steven.schockaert@ugent.be

Abstract. The task of assigning geographic coordinates to web resources has recently gained in popularity. In particular, several recent initiatives have focused on the use of language models for georeferencing Flickr photos, with promising results. Such techniques, however, require the availability of large numbers of spatially grounded training data. They are therefore not directly applicable for georeferencing other types of resources, such as Wikipedia pages. As an alternative, in this paper we explore the idea of using language models that are trained on Flickr photos for finding the coordinates of Wikipedia pages. Our experimental results show that the resulting method is able to outperform popular methods that are based on gazetteer look-up.

1 Introduction

The geographic scope of a web resource plays an increasingly important role for assessing its relevance in a given context, as can be witnessed by the popularity of location-based services on mobile devices. When uploading a photo to Flickr, for instance, users can explicitly add geographical coordinates to indicate where it has been taken. Similarly, when posting messages on Twitter, information may be added about the user's location at that time. Nonetheless, such coordinates are currently only available for a minority of all relevant web resources, and techniques are being studied to estimate geographic location in an automated way.

For example, several authors have applied language modeling techniques to find out where a photo was taken, by only looking at the tags that its owner has provided [9, 10]. The main idea is to train language models for different areas of the world, using the collection of already georeferenced Flickr photos, and to subsequently use these language models for determining in which area a given photo was most likely taken. In this way, implicit geographic information is automatically derived from Flickr tags, which is potentially much richer than the information that is found in traditional gazetteers. Indeed, the latter usually do not contain information about vernacular place names, lesser-known landmarks, or non-toponym words with a spatial dimension (e.g. names of events), among others. For the task of assigning coordinates to Flickr photos, this intuition

seems to be confirmed, as language modeling approaches have been found to outperform gazetteer based methods [5].

For other types of web resources, spatially grounded training data may not be (sufficiently) available to derive meaningful language models, in which case it seems that gazetteers would again be needed. However, as language models trained on Flickr data have already proven useful for georeferencing photos, we may wonder whether they could be useful for finding the coordinates of other web resources. In this paper, we test this hypothesis by considering the task of assigning geographical coordinates to Wikipedia pages, and show that language models from Flickr are indeed capable of outperforming popular methods for georeferencing web pages. The interest of our work is twofold. From a practical point of view, the proposed method paves the way for improving location-based services in which Wikipedia plays a central role. Second, our results add further support to the view that georeferenced Flickr photos can provide a valuable source of geographical information as such, which relates to a recent trend where traditional geographic data is more and more replaced or extended by user contributed data from Web 2.0 initiatives [2].

The paper is structured as follows. Section 2 briefly reviews the idea of georeferencing tagged resources using language models from Flickr. In Section 3 we then discuss how a similar idea could be applied to Wikipedia pages. Section 4 contains our experimental results, after which we discuss related work and conclude.

2 Language models from Flickr

In this section, we recall how georeferenced Flickr photos can be used to train language models, and how these language models subsequently allow to find the area that most likely covers the geographical scope of some resource. Throughout this section, we will assume that resources are described as sets of tags, while the next section will discuss how the problem of georeferencing Wikipedia pages can be cast into this setting.

As training data, we used a collection of around 8.5 million publicly available photos on Flickr with known coordinates. In addition to these coordinates, the associated metadata contains tags attributed to each photo, providing us with a textual description of their content, as well as an indication of the accuracy of the coordinates as a number between 1 (world-level) and 16 (street level). As in [10], we only retrieved photos with a recorded accuracy of at least 12 and we removed photos that did not contain any tags or whose coordinates were invalid. Also, following [9] photos from bulk uploads were removed. The resulting dataset contained slightly over 3.25 million photos. In a subsequent step, the training data was clustered into disjoint areas using the k-medoids algorithm with geodesic distance. Considering a varying number of clusters k, this resulted in different sets of areas $\mathcal{A}_k$. For each clustering, a vocabulary V_k was compiled, using χ^2 feature selection, as the union of the m most important tags (i.e. the tags with the highest χ^2 value) for each area.

The problem of georeferencing a resource x, in this setting, consists of selecting the area a from the set of areas $\mathcal{A}_k$ (for a specific clustering k) that is most likely to cover the geographic scope of the resource (e.g. the location of where the photo was taken, when georeferencing photos). Using a standard language modeling approach, this probability can be estimated as

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \tag{1}$$

where we identify the resource x with its set of tags. The prior probability $P(a)$ of area a can be estimated as the percentage of photos in the training data that belong to that area (i.e. a maximum likelihood estimation). To obtain a reliable estimate of $P(t|a)$ some form of smoothing is needed, to avoid a zero probability when encountering a tag t that does not occur with any of the photos in area a. In this paper, we use Jelinek-Mercer smoothing ($\lambda \in [0, 1]$)]:

$$P(t|a) = \lambda \cdot \frac{O_{ta}}{\sum_{a' \in \mathcal{A}_k} O_{ta'}} + (1 - \lambda) \cdot \frac{\sum_{a' \in \mathcal{A}_k} O_{ta'}}{\sum_{a' \in \mathcal{A}_k} O_{ta'} \sum_{t' \in V_k} O_{t'a'}}$$

O_{ta} is the number of occurrences of tag t in area a while V_k is the vocabulary, after feature selection. In the experiments, we used $\lambda = 0.7$ although we obtained good results for a wide range of values. The area that is most likely to contain resource x can then be found by maximizing the right-hand side of (1). To convert this area into a precise location, the area a can be represented as its medoid $med(a)$:

$$med(a) = \arg\min_{x \in a} \sum_{y \in a} d(x, y) \tag{2}$$

with $d(x, y)$ being the geodesic distance. Another alternative, which was proposed in [10] but which we do not consider in this paper, is to assign the location of the most similar photo from the training data which is known to be located in a.

3 Wikipedia pages

The idea of geographic scope can be interpreted in different ways for Wikipedia pages. A page about a person, for instance, might geographically be related to the places where this person has lived throughout his life, but perhaps also to those parts of the world which this person's work has influences (e.g. locations of buildings that were designed by some architect). In this paper, however, we exclusively deal with finding the coordinates of a Wikipedia page about a specific place, such as a landmark or a city. It is then natural to assume that the geographic scope of the page corresponds to a point.

While several Wikipedia pages already have geographic coordinates, it does not seem feasible to train area-specific language models from Wikipedia pages

with a known location, as we did in Section 2 for Flickr photos. The reason is that typically there is only one Wikipedia page about a given location, so either its location is already known or its location cannot be found by using other georeferenced pages. Moreover, due to the smaller number of georeferenced pages (compared to the millions of Flickr photos) and the large number of spatially irrelevant terms on a typical Wikipedia page, the process further complicates. One possibility to cope with these issues might be to explicitly look for toponyms in pages, and link these to gazetteer information. However, as we already have rich language models from Flickr, in this paper we pursue a different strategy, and investigate the possibility of using these models to find the locations of Wikipedia pages.

The first step consists of representing a Wikipedia page as a list of Flickr tags. This can be done by scanning the Wikipedia page and identifying occurrences of Flickr tags. As Flickr tags cannot contain spaces, however, it is important that concatenations of word sequences in Wikipedia pages are also considered. Moreover capitalization should be ignored. For example, an occurrence of "Eiffel tower" on a page is mapped to the Flickr tags "eiffeltower", "eiffel" and "tower".

Let us write $n(t, d)$ for the number of times tag t was thus found in the Wikipedia page d. We can then assign to d the area a which maximizes

$$P(a|d) \propto P(a) \cdot \prod_{t \in V_k} P(t|a)^{n(t,d)} \tag{3}$$

where V_k is defined as before and the probabilities $P(a)$ and $P(t|a)$ are estimated from our Flickr data, as explained in the previous section. Again (2) can be used to convert the area a to a precise location.

Some adaptations to this scheme are possible, where the scores $n(t, d)$ are defined in alternative ways. As Wikipedia pages often contain a lot of context information, which does not directly describe the location of the main subject, we propose two techniques for restricting which parts of an article are scanned. The first idea is to only look at tags that occur in section titles (identified using HTML tags of the form `<h1>`), in anchor text (`<a>`) or in emphasized regions (`<strong>` and `<b>`). This variant is referred to as *keywords* below. The second idea is to only look at the abstract of the Wikipedia page, which is defined as the part of the page before the first section heading. As this abstract is supposed to summarize its content, it is less likely to contain references to places that are outside the geographical scope of the page. This second variant is referred to as *abstract*. Note that in both variants, the value of $n(t, d)$ will be lower than when using the basic approach.

4 Experimental results

In our evaluation, we used the Geographic Coordinates dataset of DBPedia 3.6 to determine an initial set of georeferenced Wikipedia pages. To ensure that all articles refer to a specific location, we only retained those pages that are

Table 1. Comparison of the Flickr language models for different numbers of clusters k and Yahoo! Placemaker (P.M.). We report how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc. Accuracy refers to the percentage of test pages for which the language models identified the correct cluster.

k	1 km	5 km	10 km	50 km	100 km	Acc
50	20	156	262	745	1470	76.52
500	334	1060	1385	2993	4195	69.17
2500	736	1636	2139	4020	4995	57.98
5000	892	1857	2377	4194	5075	51.62
7500	1008	1996	2557	4396	5239	49.85
10000	1052	2086	2670	4471	5233	47.80
12500	1103	2131	2697	4528	5263	45.73
15000	1129	2154	2743	4551	5212	44.45
17500	1159	2213	2783	4578	5243	43.71
P.M.	313	1583	2395	4257	5056	–

mentioned as a "spot" in the GeoNames gazetteer. This resulted in a set of 7537 georeferenced Wikipedia pages, whose coordinates we used as our gold standard.

Using the techniques outlined in the previous section, for each page the most likely area from $\mathcal{A}_k$ is determined (for different values of k). To evaluate the performance of our method, we calculate the accuracy, defined as the percentage of the test pages that were classified in the correct area, i.e. the area actually containing the location of page d. In addition, we also look at how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc.

Our main interest is in comparing the methods proposed in Section 3 with the performance of Yahoo! Placemaker, a freely available popular webservice capable of geoparsing entire documents and webpages. Provided with free-form text, Placemaker identifies places mentioned in text, disambiguates those places and returns the corresponding locations. It is important to note that this approach uses external geographical knowledge such as gazetteers and other undocumented sources of information. In contrast, our approach uses only the knowledge derived from the tags of georeferenced Flickr photos.

In a first experiment, we compare the results of language models trained at different resolutions, i.e. different numbers of clusters k. Table 1 shows the results for k varying from 50 to 17 500, where we consider the basic variant in which the entire Wikipedia page is scanned for tag occurrences. There is a trade-off to be found, where finer-grained areas lead to more precise locations, provided that the correct area is found, while coarse-grained areas lead to a higher accuracy and to an increased likelihood that the found location is within a certain broad radius. In [10], it was found that the optimal number of clusters for georeferencing Flickr photos was somewhere between 2500 and 7500, with the optimum being higher for photos with more informative tags. In contrast, the results from Table 1 reveal that in the case of Wikipedia pages, its is beneficial

Table 2. Analysis of the effect of restricting the regions of a Wikipedia article that are scanned for tag occurrences (considering $k = 17500$ clusters). We report how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc.

k	1 km	5 km	10 km	50 km	100 km
article	1159	2213	2783	4578	5243
abstract	1194	2163	2707	4419	5051
keywords	1200	2361	3018	5052	5778

to further increase the number of clusters. This finding seems to be related to the intuition that Wikipedia pages contain more informative descriptions than Flickr photos. Comparing our results with Placemaker, we find a substantial improvement in all categories, which is most pronounced in the 1km range, where the number of correct locations for our language modeling approach is 3 to 4 times higher than for the Placemaker.

In a second experiment, we analyzed the effect of only looking at certain regions of a Wikipedia page, as discussed in Section 3. As the results in Table 2 show, when using the abstract, comparable results are obtained, which is interesting as this method only uses a small portion of the page. When only looking at the emphasized words (method *keywords*), the results are even considerably better. Especially for the 50km and 100km categories, the improvement is substantial. This seems to confirm the intuition that tag occurrences in section titles, anchor text and emphasized words are more likely to be spatially relevant.

5 Related work

Techniques to (automatically) determine the geographical scope of web resources commonly use resources such as gazetteers (geographical indexes), and tables with locations corresponding to IP addresses, zipcodes or telephone prefix codes. These resources are often handcrafted, which is time-consuming and expensive, although this results in accurate geographical information. Unfortunately, many of these sources are not freely available and their coverage varies highly from country to country. If sufficiently accurate resources are available, one of the main problems in georeferencing web pages is dealing with the high ambiguity of toponyms [6]. For example, when an occurrence of *Paris* is encountered, one first needs to disambiguate between a person and a place, and in the case it refers to a place, between different locations with that name (e.g. Paris, France and Paris, Texas).

It is only recently that alternative ways have been proposed to georeference resources. In [8], names of places are extracted from Flickr tags on a subset of around 50000 photos. Also, as studied by L. Hollenstein in [4], collaborative tagging-based systems are also useful to acquire information about the location of vernacular places names. In [1], methods based on Wordnet and Naive Bayes classification are compared for the automatic detection of toponyms within articles. To the best of our knowledge, however, approaches for georeferencing

Wikipedia pages, or webpages in general, without using a gazetteer or other forms of structured geographic knowledge have not yet been proposed in the literature.

An interesting line of work aims at automatically completing the infobox of a Wikipedia page by analyzing the content of that page [11]. This work is related to ours in the sense that semantic information about Wikipedia pages is made explicit. Such a strategy can be used to improve semantic knowledge bases, such as YAGO2 [3], which now contains over 10 million entities derived from Wikipedia, WordNet and GeoNames. Similarly, in [7], a gazetteer was constructed based on geotagged Wikipedia pages. In particular, relations between pages are extracted from available geographical information (e.g. *New York* is part of the *United States*). Increasing the number of georeferenced articles may thus lead to better informed gazetteers.

6 Conclusions

In this paper, we investigated the possibility of using language models trained on georeferenced Flickr photos for finding the coordinates of Wikipedia pages. Our experiments show that for Wikipedia pages about specific locations, the proposed approach can substantially outperform Yahoo! Placemaker, a popular approach for finding the geographic scope of a webpage. This is remarkable as the Placemaker crucially depends on gazetteers and other forms of structured geographic knowledge, and is moreover based on advanced techniques for dealing with issues such as ambiguity. Our method, on the other hand, only uses information that was obtained from freely available, user-contributed data, in the form of georeferenced Flickr photos, and uses standard language modeling techniques.

These results suggest that the implicit spatial information that arises from the tagging behavior of users may have a stronger role to play in the field of geographic information retrieval, which is currently still dominated by gazetteer-based approaches. Moreover, as the number of georeferenced Flickr photos is constantly increasing, the spatial models that could be derived are constantly improving. Further work is needed to compare the information contained implicitly in such language models with the explicit information contained in gazetteers.

References

1. D. Buscaldi and P. Rosso. A comparison of methods for the automatic identification of locations in wikipedia. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, pages 89–92, 2007.
2. M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.
3. J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference on World Wide Web*, pages 229–232, 2011.

4. L. Hollenstein. Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich, 2008.
5. M. Larson, M. Soleymani, and P. Serdyukov. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.
6. J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* PhD thesis, University of Edinburgh, 2007.
7. A. Popescu and G. Grefenstette. Spatiotemporal mapping of Wikipedia concepts. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 129–138, 2010.
8. T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web*, 3(1):1–30, 2009.
9. P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 484–491, 2009.
10. O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.
11. F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pages 41–50, 2007.