Expanding the Utility of Geospatial Knowledge Bases by Linking Concepts to WikiText and to Polygonal Boundaries

Bruno Martins
University of Lisbon
IST and INESC-ID
bruno.g.martins@ist.utl.pt

Francisco J.
López-Pellicer
Universidad de Zaragoza
fjlopez@unizar.es

Dirk Ahlers
NTNU – Norwegian University
of Science and Technology
dirk.ahlers@idi.ntnu.no

ABSTRACT

This vision paper argues that a geospatial knowledge base combining textual descriptions for concepts such as places, together with place types, semantic relations between concepts and, most importantly, polygonal geometries associated to the geospatial concepts, constitutes a valuable resource for researchers working on the computational modeling of spatial language. We describe a simple procedure for producing one such resource from existing open datasets, and discuss possible ways for moving beyond the current state-of-the-art within the general area of geospatial text mining, through studies supported by one such knowledge base.

CCS Concepts

•**Information systems** → *Information integration; Spatial-temporal systems; Information extraction;*

Keywords

Sources of Geospatial Knowledge, Polygonal Geometries as Geospatial Footprints, Resources for Processing Geospatial Language

1. INTRODUCTION

Computational models for supporting the analysis of spatial language are important in applications related to the general areas of information retrieval, information extraction, and natural language processing. Much work has for instance been done on extracting and disambiguating geospatial entities [6], on extracting relations between these entities [2, 7, 4], or on inferring the geospatial foci of textual contents [5]. Much of this previous work has leveraged:

- small corpora where textual contents (e.g., noun phrases corresponding to toponyms) are annotated with geospatial coordinates (i.e., latitude and longitude), or with links to entries in gazetteers such as GeoNames, which associate named places to types and to the corresponding latitude and longitude coordinates;
- large resources built indirectly through community efforts, such as collections of geo-tagged Flickr photos or geo-referenced Wikipedia articles. There resources are, in turn, often linked to structured knowledge bases, such as DBpedia or YAGO, where the geospatial concepts are also associated to proper types and to the corresponding coordinates of latitude and longitude.

In previous studies, the use of geospatial information (e.g., for the validation of computation models that analyze natural language in order to predict geospatial properties) has been mostly limited to point geometries (i.e., geospatial coordinates of latitude and longitude). Considering polygonal geometries (i.e., multi-polygons describing the boundaries of places) could lead to more interesting and relevant results. For example, it would be more interesting to consider polygonal regions, instead of individual latitude and longitude coordinates, when assigning documents to their corresponding geospatial foci [5], given that some of the documents are likely to correspond to large regions such as entire countries or states.

Other studies have considered the development of knowledge bases (KBs) where entities and facts (i.e., relations between pairs of entities) are anchored in time and/or space (e.g., the latest version of YAGO [1], where both facts and entities can be associated to spatio-temporal footprints), as well as the extension of such KBs with information extracted from the Web or from large corpora. However, these efforts have again been limited to point-based geospatial information, severely restricting their practical application.

2. ADDING POLYGONAL FOOTPRINTS

We argue that large KBs combining textual descriptions for concepts such as places, together with proper place types, semantic relations between concepts, and polygonal geometries associated to the geospatial concepts, constitute a valuable resource for researchers working on the computational modeling of spatial language, supporting studies that can significantly advance the current state-of-the-art. We also argue that existing resources, available openly and under very permissive licensing schemes, can easily be combined to form such a KB. For demonstrating this claim, we followed the approach described next:

- We started with the contents from the QuattroShapes[1] global polygon gazetteer and from the Who's On First gazetteer[2], i.e. two authoritative sources of non-overlapping polygons associated to curated lists of places. These resources are based on information from the Natural Earth[3] public domain map, together with open data from multiple other sources. Computational geometry techniques (e.g., alpha-shapes[4]) were used to back-fill regions without complete open data, e.g. by leveraging FourSquare checkins and geo-tagged Flickr photos. Alignments with the GeoNames and Yahoo! GeoPlanet[5] gazetteers are also provided,

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837693

[1] http://quattroshapes.com

[2] http://mapzen.com/blog/who-s-on-first

[3] http://www.naturalearthdata.com

[4] http://code.flickr.net/2008/10/30/the-shape-of-alpha/

[5] http://developer.yahoo.com/geo/geoplanet/

Table 1: Statistical characterization for the resulting KB.

Property	Lang	Count
Concepts associated to polygons	–	371,346
+ also with textual descriptions (i.e., linked to entries in Wikipedia)	EN	100,784
	DE	48,451
	ES	60,903
	PT	62,977
+ also linked to DBpedia	-	100,757
+ also linked to YAGO	-	98,803
+ also linked to temporal information	-	14,917
Geospatial references in text descriptions	EN	9,802,181

allowing us to combine these data with other resources.

- We leveraged existing alignments for linking the polygonal geometries to the corresponding GeoNames[6] concepts, which provides structured descriptions for a very large set of toponyms. Geospatial concepts within GeoNames are typed according to well-defined classes, and are internally linked according to administrative and proximity relations. Concepts in this ontology are also associated to the corresponding entries in Wikipedia.
- We associated the GeoNames concepts to the textual descriptions, in multiple languages, that are provided in Wikipedia. This was again made by leveraging existing alignment information. Recent community efforts such as Wikivoyage[7] (i.e., travel guides in multiple languages) or the Simple English Wikipedia[8] are also associated to Wikipedia and, as such, we linked some of the GeoNames concepts to these other textual descriptions.
- Wikipedia descriptions can refer to concepts, by linking to other Wikipedia pages. This linkage structure has been used for supporting named entity recognition [3] and disambiguation models, and we therefore also kept in our KB the links between phrases in the Wikipedia descriptions for geospatial concepts (e.g., for supporting studies in toponym resolution [6]).
- We linked the GeoNames concepts to the corresponding entries in the DBpedia ontology[9], which in turn provides links to WordNet[10] and to YAGO [1]. Both DBpedia and YAGO classify and inter-relate different concepts (not just geospatial), in multiple languages, within a detailed ontology.

3. APPLICATIONS

Table 1 shows some statistics for the produced KB, highlighting the large number of concepts with textual descriptions and with polygonal information. Possible applications for this KB include:

- The polygonal geometries can support the training and evaluation of models for geocoding textual documents. Previous studies have proposed language modeling (LM) approaches [5], in which (i) the geographic space is first discretized into a set of non-overlapping cells, (ii) LMs are trained from Wikipedia documents whose coordinates lay within each cell, and (c) new documents are geocoded to the centroid coordinates of the most likely cell for their text, as inferred from the LMs. Alternatively, through our KB, we can consider approaches where documents are assigned to a polygonal region, e.g. computed from all the cells whose LM probabilities are above a given threshold.
- The KB can support the evaluation of document geocoding models over different types of text (e.g., general contents, descriptions from travel guides, or texts with a simple vocabulary and short sentences), over textual descriptions for different types of concepts, and for multiple languages.
- Previous studies on toponym resolution have also relied on point coordinates associated to noun phrases [6], but one could alternatively consider disambiguating toponyms, occurring over texts in multiple languages, into encompassing polygonal footprints.
- Formal models of spatial relations, together with the polygonal footprints in the KB, can be used to derive topological, directional, and distance relations between pairs of geo-referenced concepts. One can thus consider addressing tasks related to the analysis of how spatial relations are expressed in natural language [2, 7] (e.g., we can use textual descriptions for pairs of concepts, of different conceptual types, in order to predict the spatial relation that holds between them).
- The KB can support studies that jointly address the modeling of spatial and temporal aspects in natural language (e.g., joint document dating and geocoding), given that many concepts and/or facts in DBpedia/YAGO are also associated to a temporal validity or to a focus time. These data can perhaps also support experiments involving places whose boundaries change over time.

4. CONCLUSIONS

We discussed applications for a KB integrating polygonal geometries, typed geospatial concepts, and textual descriptions. We described a simple approach for producing one such KB through existing open resources, although there are many possible ways to extend the work described in this paper (e.g., by further integrating KB concepts with the FrameNet[11] dictionary of word senses, as well as with other resources related to computational semantics, thus advancing research in spatial semantics [4]). Ongoing work is focused on the integration of other resources (e.g., increase the detail for some regions, add more information regarding concepts such as rivers, increase the amount of temporal information, consider links to Wikipedia from resources such as WikiSource[12] or WikiNews[13]), as well as on producing a more detailed characterization of the resulting KB. Through this last aspect, we also hope to start addressing data quality issues in a principled manner.

5. REFERENCES

[1] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.*, 194, 2013.

[2] A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *ACM SIGSPATIAL Workshop on Computational Models of Place*, 2013.

[3] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.*, 194, 2013.

[4] J. Pustejovsky, P. Kordjamshidi, M.-F. Moens, A. Levine, S. Dworman, and Z. Yocum. SemEval-2015 Task 8: SpaceEval. In *International Workshop on Semantic Evaluation*, 2015.

[5] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Conference on Empirical Methods in Natural Language Processing*, 2012.

[6] M. Speriosu and J. Baldridge. Text-driven toponym resolution using indirect supervision. In *Annual Meeting of the Association for Computational Linguistics*, 2013.

[7] J. O. Wallgrün, A. Klippel, and T. Baldwin. Building a corpus of spatial relational expressions extracted from Web documents. In *ACM SIGSPATIAL Workshop on Geographic Information Retrieval*, 2014.

[6] http://www.geonames.org/ontology/

[7] http://en.wikivoyage.org

[8] http://simple.wikipedia.org/

[9] http://wiki.dbpedia.org/services-resources/ontology

[10] http://wordnet.princeton.edu

[11] http://framenet.icsi.berkeley.edu/

[12] http://en.wikisource.org/

[13] http://en.wikinews.org