Which Kobani? A Case Study on the Role of Spatial Statistics and Semantics for Coreference Resolution Across Gazetteers

Rui Zhu, Krzysztof Janowicz, Bo Yan, and Yingjie Hu

STKO Lab, Department of Geography, University of California Santa Barbara, Santa Barbara, California, USA

Abstract

Identifying the same places across different gazetteers is a key prerequisite for spatial data conflation and interlinkage. Conventional approaches mostly rely on combining spatial distance with string matching and structural similarity measures, while ignoring relations among places and the semantics of place types. In this work, we propose to use spatial statistics to mine semantic signatures for place types and use these signatures for coreference resolution, i.e., to determine whether records form different gazetteers refer to the same place. We implement 27 statistical features for computing these signatures and apply them to the type and entity levels to determine the corresponding places between two gazetteers, namely GeoNames and DBpedia. The city of Kobani, Syria, is used as a running example to demonstrate the feasibility of our approach. The experimental results show that the proposed signatures have the potential to improve the performance of coreference resolution.

Introduction and motivation

Coreference resolution across gazetteers is an important prerequisite for spatial **data conflation** and **interlinkage**.

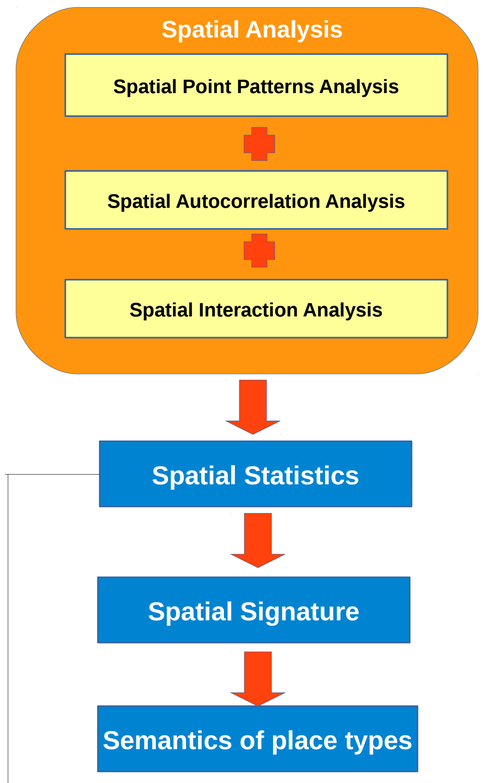
Conventional approaches and their limitations:

(1). Coordinate matching:
Centroids for all geographic features → difficult to select a place type agnostic distance threshold as initial search radius.

(2). String matching:
Same place but different names;
Different places but the same name.

(3). Feature type matching:
Incompatible typing schemata/ontologies.

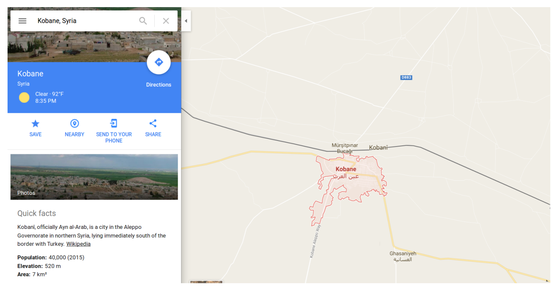
Proposed approach → Semantic matching:



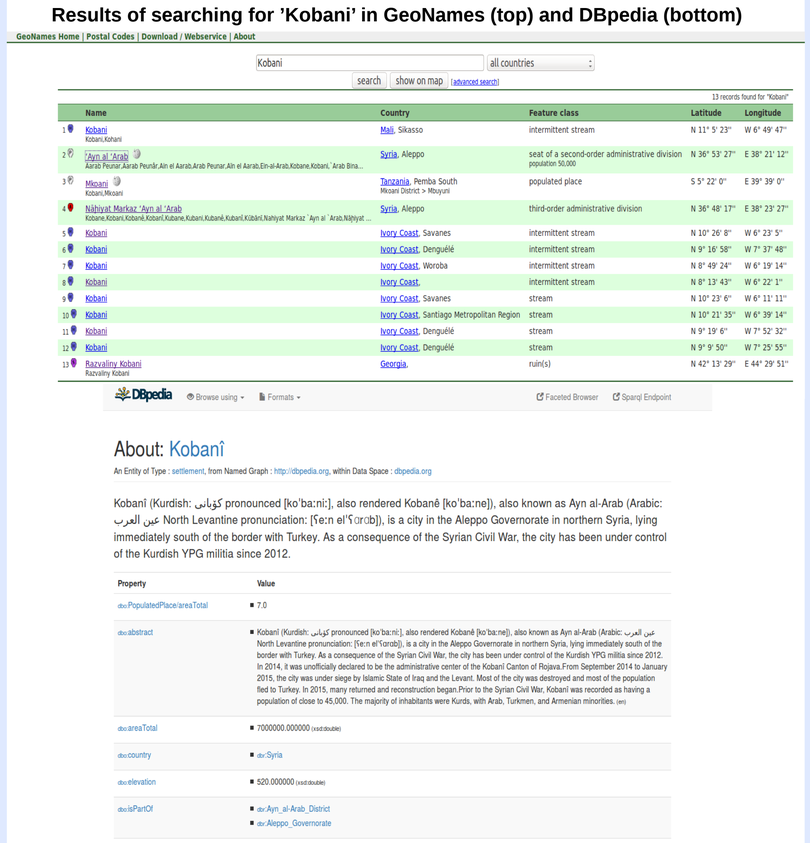
Case

To illustrate our method, Kobani, Syria is used as an example. Kobani is a city that lies near the border between Syria and Turkey. It is a typical example for the complexities arising when multiple parties such as the local population, news outlets around the world, government agencies from different states, and so forth, refer to a place by different names such as Aarab Peunar, Kubani, Kobane and ‘Ayn al‘ Arab, to name but a few.

Kobani, Syria, in Google Map



Results of searching for ’Kobani’ in GeoNames (top) and DBpedia (bottom)



Methodologies and Results

1. Place type signature as additional matching characteristics

We propose to use the mined place type signatures as an additional matching characteristic that communicates the semantics of place types beyond labels alone.

Step 1 - Select candidates: there are three place types associated with candidates for Kobani in GeoNames and one place type (populated place) in DBpedia.
Step 2 - Calculate dissimilarities: computing the Euclidean distance between these three GeoNames place signatures and the populated place signature from DBpedia.

Note: The place signature in this work is essentially the feature vector comprised of the 27 statistics listed in the Table. We use Euclidean distance and regard all statistics the same weight, but more sophisticated models are under investigation as well.

	Spatial Point Patterns	Spatial Autocorrelations	Spatial Interaction with Other Geographic Features	
Local	Intensity	Global Moran's I	**Internal**	Count of distinct nearest feature types
	Mean distance to nearest neighbor			
	Variance distance to nearest neighbor			Entropy of nearest feature types
	Kernel density (bandwidth)			
	Kernel density (range)	Semivariogram value (at first distance lag)	**External**	Population value (max)
	Ripley's K (range)			
	Ripley's K (mean deviation)			Population value (min)
	Standard deviational ellipse (rotation)	Semivariogram value (at median distance lag)		Population value (mean)
	Standard deviational ellipse (std dev along x-axis)			Population value (std dev)
	Standard deviational ellipse (std dev along y-axis)			Shortest distance to road (max)
Global	Intensity	Semivariogram value (at last distance lag)		Shortest distance to road (min)
	Kernel density (bandwidth)			Shortest distance to road (mean)
	Kernel density (range)			Shortest distance to road (std dev)

Results: dissimilarities between the populated place signature for DBpedia and three example place type signatures in GeoNames.

Dissimilarity (Euclidean distance)	DBpedia: *Populated Place*
GeoNames: *seat of second-order administrative division*	7.22
GeoNames: *stream/intermittent stream*	8.96
GeoNames: *populated place*	9.22

2. Place type signature of neighboring places

One drawback of the first method is that if multiple candidates shared the same place type, the signatures are incapable of providing any further distinctions. Therefore, we propose to include the signatures of neighboring places as well.

Step 1 - Query nearest neighbors: 9 nearest neighbors are queried for each candidate place and their place types are recorded.
Step 2 - Obtaining averaged neighboring signatures: the averaged signatures of these 9 place types are calculated for characterizing the neighborhood of the specific candidates.
Step 3 – Calculating dissimilarities: Euclidean distances are calculated between candidates in GeoNames and the one in DBpedia.
Note: the averaged neighboring signature is the averaged feature vector of the 27 statistics listed in the Table.

	’Ayn al’ Arab (GeoNames: seat of a second-order administrative division)	Kobani (GeoNames: stream)	Mkoani (GeoNames: populated place)	Kobani (DBpedia: populated place)
Feature types of 9 nearest neighbors	section of populated place	populated place	populated place	settlement
	office building	stream	populated place	settlement
	school	stream	populated place	village
	square	stream	third-order administrative division	settlement
	prison	stream	third-order administrative division	tunnel
	section of populated place	stream	third-order administrative division	settlement
	section of populated place	stream	populated place	village
	market	stream	populated place	dam
	square	populated place	populated place	populated place

Results: Dissimilarities between Kobani ’s neighboring signatures in DBpedia and the three example places’ neighboring signatures in GeoNames.

Dissimilarity (Euclidean distance)	DBpedia: *Populated Place*
’Ayn al’ Arab (GeoNames)	4.23
Kobani (GeoNames)	10.57
Mkoani (GeoNames)	6.98

Conclusion

In this work, we presented an initial case study that demonstrates how signatures mined from spatial statistics can reveal additional information about the semantics of place types on top of relying on type labels alone. Our work shows how spatial statistics and ontology engineering and alignment can go hand in hand to provide additional characteristics for tasks such as coreference resolution which play an increasingly important role as drivers of record linkage and conflation. In essence, we make use of the fact that different types of places can be told apart by the results of various spatial statistics performed over their instances, i.e., particular places. This, in turn, enables us to regard the resulting place type specific signatures as feature vectors and compute their dissimilarity using Euclidean distance (or other measures), thereby gaining an additional matcher on top of the string, spatial distance, and structural matchers used in the literature. Finally, we also go beyond existing work by taking neighboring places into account to improve the matching, instead of comparing 1:1 matches in isolation. In the future, we will apply the presented work to more (Linked Data) gazetteers and all their places.

References

1. V. Sehgal, L. Getoor, and P. D. Viechnicki. Entity resolution in geospatial data integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 83–90. ACM, 2006.
2. P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, *IEEE Transactions on*, 25(1):158–176, 2013.
3. R. Zhu, Y. Hu, K. Janowicz, and G. McKenzie. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*, 2016.

Contact: ruizhu@geog.ucsb.edu