

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



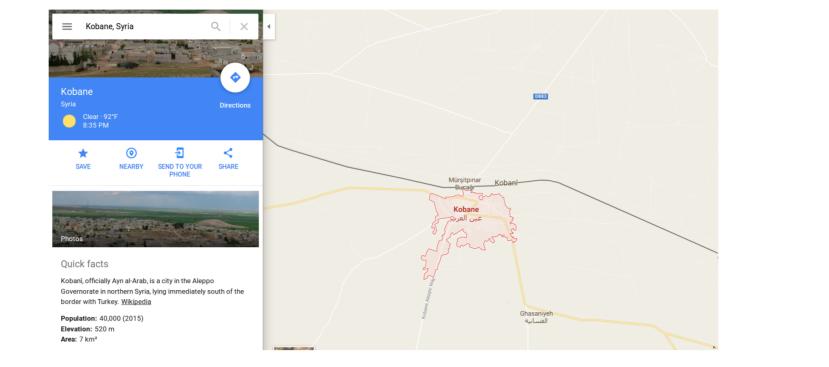
Abstract

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.

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

Dissimilarity (Euclidean distance)	DBpedia: <i>Populated</i> <i>Place</i>
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.

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 \rightarrow difficult to select a place type agnostic distance threshold as initial search radius.

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

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

Proposed approach \rightarrow Semantic matching:

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

	Kobani	all count	ries ‡		
	Sr	earch show on map [advanced search]			
				13 records	found for "Kobani"
	Name	Country	Feature class	Latitude	Longitude
1 🖤	<mark>Kobani</mark> Kobani, Kohani	<u>Mali</u> , Sikasso	intermittent stream	N 11° 5' 23''	W 6° 49' 47''
2 🕅	<u>'Ayn al 'Arab</u> 🧐 Aarab Peunar,Aarab Peunâr,Ain el Aarab,Arab Peunar,Aïn el Aarab,Ein-al-Arab,Kobane,Kobani,`Arab Bina	<u>Syria</u> , Aleppo	seat of a second-order administrative division population 50,000	N 36° 53' 27''	E 38° 21' 12"
з 🕅	<u>Mkoani</u> 🐌 Kobani, Mkoani	<u>Tanzania</u> , Pemba South Mkoani District > Mbuyuni	populated place	S 5° 22' 0''	E 39° 39' 0''
4 🥊	<u>Nāḥiyat Markaz 'Ayn al 'Arab</u> Kobane,Kobani,Kobanê,Kobani,Kubane,Kubani,Kubanê,Kubanî,Kübānī,Nahiyat Markaz `Ayn al `Arab,Nāḥiyal	<u>Syria</u> , Aleppo	third-order administrative division	N 36° 48' 17"	E 38° 23' 27''
5 🖲	Kobani	lvory Coast, Savanes	intermittent stream	N 10° 26' 8''	W 6° 23' 5''
6	Kobani	Ivory Coast, Denguélé	intermittent stream	N 9° 16' 58''	W 7° 37' 48''
7 🖲	Kobani	Ivory Coast, Woroba	intermittent stream	N 8° 49' 24''	W 6° 19' 14''
8 🖲	Kobani	Ivory Coast,	intermittent stream	N 8° 13' 43"	W 6° 22' 1"
9 🖲	Kobani	lvory Coast, Savanes	stream	N 10° 23' 6"	W 6° 11' 11''
10 🥊	Kobani	Ivory Coast, Santiago Metropolitan Region	stream	N 10° 21' 35"	W 6° 39' 14''
11 🖲	Kobani	lvory Coast, Denguélé	stream	N 9° 19' 6''	W 7° 52' 32''
12 🥊	Kobani	Ivory Coast, Denguélé	stream	N 9° 9' 50''	W 7° 25' 55''
13 🦻	<u>Razvaliny Kobani</u> Razvaliny Kobani	<u>Georgia</u> ,	ruin(s)	N 42° 13' 29"	E 44° 29' 51''
	Srowse using - Formats -		C Faceted Browser C Sparql Endpoint		

About: Koban

n Entity of Type : settlement, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Kobanî (Kurdish: كۆبانى pronounced [ko'baːniː], also rendered Kobanê [ko'baːne]), also known as Ayn al-Arab (Arabic: العرب North Levantine pronunciation: [ʕeːn el'ʕɑrɑb]), is a city in the Aleppo Governorate in northern Syria, lying عين العرب mmediately south of the border with Turkey. As a consequence of the Syrian Civil War, the city has been under contro of the Kurdish YPG militia since 2012

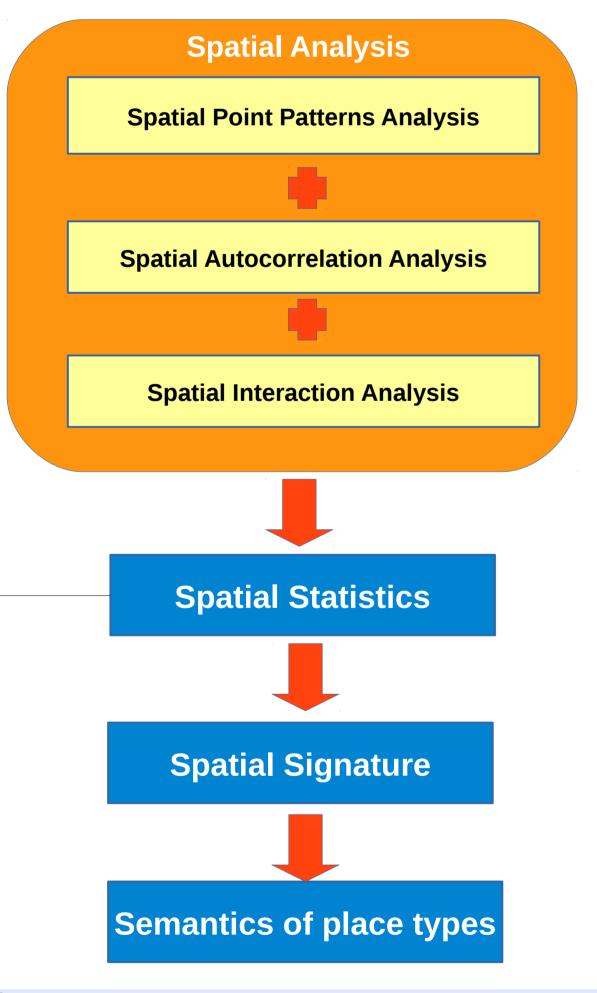
Property	Value			
dbo:PopulatedPlace/areaTotal	■ 7.0			
dbo:abstract	Kobanî (Kurdish: كوبانى pronounced [ko'ba:ni:], also rendered Kobanê [ko'ba:ne]), also known as Ayn al-Arab (Arabic: عين العرب North Levantine pronunciation: [Se:n el'Sorob]), is a city in the Aleppo Governorate in northern Syria, lying immediately south of the border with Turkey. As a consequence of the Syrian Civil War, the city has been under control of the Kurdish YPG militia since 2012 In 2014, it was unofficially declared to be the administrative center of the Kobanî Canton of Rojava.From September 2014 to Janua 2015, the city was under siege by Islamic State of Iraq and the Levant. Most of the city was destroyed and most of the population fled to Turkey. In 2015, many returned and reconstruction began.Prior to the Syrian Civil War, Kobanî was recorded as having a population of close to 45,000. The majority of inhabitants were Kurds, with Arab, Turkmen, and Armenian minorities. (en)			
dbo:areaTotal	700000.000000 (xsd:double)			
dbo:COUNTry	■ dbr:Syria			
dbo:elevation	520.000000 (xsd:double)			
dbo:isPartOf	 dbr:Ayn_al-Arab_District dbr:Aleppo_Governorate 			

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)
	section of populated place	populated place	populated place	settlement
	office building	stream	populated place	settlement
	school	stream	populated place	village
ature types	square	stream	third-order administrative division	settlement
9 nearest ighbors	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



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.

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

Dissimilarity (Euclidean distance)	DBpedia: <i>Populated</i> <i>Place</i>
'Ayn al' Arab (GeoNames)	4.23
Kobani (GeoNames)	10.57
Mkoani (GeoNames)	6.98

Conclusion

Fea

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 flation. In essence, we make use of the fact that different es of places can be told apart by the results of various al statistics performed over their instances, i.e., particular es. This, in turn, enables us to regard the resulting place specific signatures as feature vectors and compute their imilarity using Euclidean distance (or other measures), eby gaining an additional matcher on top of the string, al distance, and structural matchers used in the literature. ally, we also go beyond existing work by taking hboring places into account to improve the matching, ead of comparing 1:1 matches in isolation. In the future, will apply the presented work to more (Linked Data) etteers and all their places.

					IIICICa
	Spatial Point Patterns	Spatial Autocorrelations	Spatial Interaction with Of	ther Geographic Features	confla
ocal	Intensity Mean distance to nearest neighbor			Count of distinct nearest feature types	types spatia
	Variance distance to nearest neighbor	Global Moran's I	Internal	Entropy of nearest feature types	places type s
	Kernel density (bandwidth)				dissin
	Kernel density (range)	Semivariogram value (at first distance lag)		Population value (max)	thereb spatia
	Ripley's K (range)				Finall neighl
	Ripley's K (mean deviation)			Population value (min)	instea
	Standard deviational ellipse (rotation)	Semivariogram value (at median distance lag)		Population value (mean)	we wi gazett
	Standard deviational ellipse (std dev along x-axis)		External	Population value (std dev)	Referen 1. V. Seh
	Standard deviational ellipse (std dev along y-axis)			Shortest distance to road (max)	integratio Advances
obal	Intensity	Semivariogram value (at last distance lag)		Shortest distance to road (min)	2. P. Shva challenge
	Kernel density (bandwidth)			Shortest distance to road (mean)	176, 2013 3. R. Zhu
	Kernel density (range)			Shortest distance to road (std dev)	feature ty in GIS, 20
			Contacts minhas Orea		

ences

ehgal, L. Getoor, and P. D. Viechnicki. Entity resolution in geospatial data tion. In *Proceedings of the 14th annual ACM international symposium on* ces in geographic information systems, pages 83–90. ACM, 2006. vaiko and J. Euzenat. Ontology matching: state of the art and future ges. Knowledge and Data Engineering, *IEEE Transactions on*, 25(1):158– 13.

nu, Y. Hu, K. Janowicz, and G. McKenzie. Spatial signatures for geographic types: Examining gazetteer ontologies using spatial statistics. *Transactions* 2016.

Contact: ruizhu@geog.ucsb.edu

Loc

Glo