# 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, USA {ruizhu,jano,boyan,yingjiehu}@geog.ucsb.edu

#### 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, which are 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.

Keywords: Spatial statistics, coreference resolution, gazetteers, semantic signatures

## **1** Introduction and Motivation

Coreference resolution across gazetteers is an important prerequisite for spatial data conflation and interlinkage. Conventional approaches, such as coordinate matching, string matching, and feature type matching, often focus on the footprints, names, and types of places, as well as the combination of these three properties (Sehgal et al., 2006; Shvaiko and Euzenat, 2013). However, such approaches have their limitations. Today, most gazetteers still rely on centroids for representing geographic features (even for feature types such as counties, rivers, or oceans). These centroids differ significantly across datasets, often by more than 100km. Furthermore, it is difficult to select a place type agnostic distance threshold as initial search radius. Polygon and polyline based matching, e.g., using Hausdorff distance, comes with its own limitations, scale and the resulting generalization being key problems. For string matching, such as using Levenshtein distance, the same place may have substantially different toponyms (e.g., Ayn al-Arab in TGN and Kobani in DBpedia) while different places may share common names. In addition, simply relying on direct feature type matching is likely to fail since different gazetteers employ incompatible typing schemata/ontologies. In conjunction, these problems often lead to either false negative or false positive matches.

In previous work, we proposed using *spatial signatures*, which are derived from spatial statistics, to understand the semantics of places types bottom-up (Zhu et al., 2016). In this work, we apply these signatures to coreference resolution. The used spatial statistics are selected from three perspectives; a detailed list is shown in Table 1:

- **Spatial point pattern analysis**. Point coordinates are used to quantitatively measure the spatial point patterns of place types (such as *populated place*). Kernel density estimation, Ripley's K, and standard deviational ellipse analysis are conducted and corresponding statistics are obtained for representing the signatures. Furthermore, we computed these statistics from both local and global aspects.
- **Spatial autocorrelation analysis**. In order to capture the interaction between places, we converted the point patterns into raster maps where each pixel represents the intensity of points. Spatial correlation statistics, such as Moran's I and semivariograms, are subsequently used to improve the signatures.
- Spatial interactions with other geographic features. In contrast to the first two perspectives, this group of statistical features is derived by integrating other geographic features. These

geographic features are further separated into internal features, in which the target feature's neighbors are considered, and external features, in which external data sources such as populations and road networks are incorporated.

A feature type's statistics (e.g., *Mountain*) are calculated considering all the points across the continuous US that belong to said feature type. To tackle the scalability issue of conducting spatial point analysis, like the Ripley's K, a spatial sampling technique was introduced (Zhu et al., 2016). Furthermore, in order to analyze the spatial autocorrelation of one feature type's intensity, the points are converted to a raster map whose cell sizes are consistent across all feature types. These statistics together (see Table 1), work as the spatial signature (here feature vectors) for the target feature type. Note that these 27 statistics are regarded as equally weighted in our work but more sophisticated models can be investigated in the future work.

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

Table 1: A summary of the 27 statistical features used to derive place type signatures.

## 2 Method and Case Studies

We have computed a unique semantic signature for each place type defined in the used gazetteers based on the statistical features outlined above. Next, these signatures are used within two case studies to show how they can be applied to improve coreference resolution. In the first case, we will use the signatures as an additional matching score to complement spatial distance as well string and type label similarity used in the existing literature. In the second case we also consider the signatures of neighboring places.

To illustrate our method, Kobani, Syria, is used as an example. Kobani is a city 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. Different and overlapping sets of toponyms are stored in gazetteers making straightforwards string similarity matching challenging. To further complicate issues, different places may have similar or even the same names while being in relative vicinity, such as a river's centroid that shares its name with a populated place. As we outlined above, the introduction of feature types into the matching process is often of limited use as there is no commonly agreed geographic feature type ontology and the individual ontologies and vocabularies used by gazetteers are under-specific to a degree where one has to rely on the types labels (and therefore string matching).

Figure 1 illustrates search results for Kobani in two gazetteers, GeoNames and DBpedia Places. Since the records of DBpedia are extracted from Wikipedia articles, only significant places are included. Therefore, the result for Kobani in DBpedia leads users directly to the city Kobani in Syria. However, since GeoNames is a more comprehensive gazetteer that attempts to collect all geographic features in the world, the search result for Kobani includes multiple records. Intuitively, and compared to types such as *stream* and *ruins*, the *seat of second-order administrative division* feature types in GeoNames is more likely to correspond to the *populated place* feature type in DBpedia despite having a very low string similarity between their labels. This make the matching task easy for humans but difficult for an automated machine-based matching.

#### 2.1 Place Type Signatures as Additional Matching Characteristic

So far, we established the argument that toponym and centroid distance based matching on its own is not always sufficient and that place types (present in any modern digital gazettes) should be used in addition.

lome	Postal Codes   Download / Webservice   About						_										
	Kobani	all counti	ries :														
	sear	ch show on map [advanced search]															
_					13 reco	rds found for "Kobani"											
	Name	Country	Feature class		Latitude	Longitude											
	Kobani, Kohani	Mali, Sikasso	intermittent stream		N 11° 5' 23'	W 6° 49' 47"											
Ð	'Ayra al 'Arab 🥥 Aarab Peunar Jarab Peunêr Jin el Aarab,Arab Peunar Jin el Aarab,Ein-al-Arab,Kobane,Kobani, 'Arab Bina	Syria, Aleppo	seat of a second-orde population 50,000	er administrative divi	ision N 36° 53' 21	" E 38° 21' 12"											
3	Mkoani 🕲 Kobari, Mkoani	Tanzania, Pemba South Mkoani District > Mbuyuni	populated place		S 5° 22' 0"	E 39* 39' 0"											
1	Nähiyat. Markaz. "Ayrı al. "Arab Kobare, Kobari, Kobarê, Kobari, Kubare, Kubari, Kubarê, Kubarê, Kubarê, Nahiyat. Markaz. "Ayrı al. "Arab, Nêjiyat	Syria, Aleppo	third-order administra	ative division	N 36° 48' 17	" E 38° 23' 27"											
)	Kobani	Ivory Coast, Savanes	intermittent stream	alia Dispedia	Browse using +	Formats +				C Faceted Brow	C Faceted Browsh	C Faceted Browser	C Faceted Browser C	C Faceted Browser C Spare	C Faceted Browser C Sparol En	🕑 Faceted Browser 🛛 🖒 Sparol Endp	C Faceted Browser C Sparol Endpo
8	Kobani	Ivory Coast, Denguélé	intermittent stream														
	Kobani	Ivory Coast, Woroba	intermittent stream														
)	Kobani	Ivory Coast,	intermittent stream	About: K	lobanî												
	Kobani	Ivory Coast, Savanes	stream	An Entity of Type : se	tlement, from Named	Graph : http://dbpedia.	org, within Data Space : dbpedia	.org	.org	org	org	org	org	org	org	.org	org
8	Kobani	Ivory Coast, Santiago Metropolitan Region	stream														
8	Kobani	Ivory Coast, Denguélé	stream	Kobanî (Kurdish: كَوَبَاني pronounced [koˈbaːniː], also rendered Kobanê [koˈbaːne]), also known as Ayn al-Arat North Levantine pronunciation: [ˈseːn elˈʕɑrob]), is a city in the Aleppo Governorate in northern Syr			anê [ko'ba:ne]), also known as Ayn al-Arab (Ar	anê [ko'ba:ne]), also known as Ayn al-Arab (Arab	anê [ko'ba:ne]), also known as Ayn al-Arab (Arabic	anê [ko'ba:ne]), also known as Ayn al-Arab (Arabic:							
8	Kobani	Ivory Coast, Denguélé	stream				in the Aleppo Governorate in northern Syr	in the Aleppo Governorate in northern Syria, I	in the Aleppo Governorate in northern Syria, lyin	in the Aleppo Governorate in northern Syria, lying	in the Aleppo Governorate in northern Syria, lying						
8	<u>Razvaliny Kobani</u> Razvaliny Kobari	Georgia,	ruin(s)	immediately s of the Kurdish	outh of the bon YPG militia sir	der with Turkey ce 2012.	. As a consequence of th	e	e Syrian Civil War, th	e Syrian Civil War, the city has be	e Syrian Civil War, the city has bee	e Syrian Civil War, the city has been	e Syrian Civil War, the city has been und	e Syrian Civil War, the city has been under c	e Syrian Civil War, the city has been under con	e Syrian Civil War, the city has been under contro	e Syrian Civil War, the city has been under contro
				Property		Value											
				do PopulatedPlace/	areaTotal	<b>7.0</b>											
				do abstract		<ul> <li>Kobani (Kurdish: North Levantine proboder with Turkey, In 2014, it was unof 2015, the city was u fied to Turkey, In 20 population of close !</li> </ul>	igs pronounced [ko/bacni;], also rendem nunciation; (Focn el Sorrol)), is a cibj in t As a consequence of the Syrian Civil We icially declared to be the administrative nder siege by Islamic State of Iraq and 1 5. mary returned and reconstruction b to 45,000. The majority of inhabitants w	ed Kr he Al ar, the centr the L regan ere K	obané [ko'bacne]), also eppo Governorate in no oby has been under o er of the Kobani Canton evant. Most of the oby v Prior to the Syrian Civi urds, with Arab, Turkm	obané (ilo 'ba:ne)), also known as Ayn al-Ar eppo Governorate in northern Syria, lying i o sly has been under control of the Kurdali er of the Kobani Canton of Rojava. From Se evant. Most of the city was destroyed and r Prior to the Syrian Call War, Kobani was r- urds, with Arab, Turkmen, and Armenian m	bland [lio bacnel], also known as Ayn al-Arab eppo Governorate in northern Syria, lying im or of the bacen under control of the Kurdish 't er of the Kobani Canton of Rojava. From Sept evant. Most of the city was destroyed and m Prior to the Syrian Coliv War, Kobani was rec urds, with Arab, Turkmen, and Armenian mir	obané (iko'ba:ne)), alao known as Ayn al-Arab (A eppo Governorate in northern Syria, lying imme o sly has been under control of the Kurdiah YPC er of the Kobani Canton of Rojava. From Septem evant. Most of the city was destroyed and most Prior to the Syrian Call War, Kobani was record urds, with Arab, Turkmen, and Armenian minori	bbind [iko bai:ne]], also known as Ayn al-Arab (Arabic: eppo Governorate in northerm Syria, lying immediate)s or of has basen under control of the Kuridik-YPG million or of the Kobani Canton of Rojava. From September 20 vaunt. Most of the city was destroyed and most of the Prior to the Syrian Coll War, Kobani was recorded ar urds, with Arab, Turkmen, and Armenian minorities. (r	المرب (الم تحدموا), alio known as Ayn al-Arab (Arabic: مرب oppo Governate in northern Syria, jving immediately south och has been under control of the Kurdini YPG mitlika and er of the Kobani Canton of Rojava From September 2014 ko avaari. Moa of the city was destroyed and most of the popul Prior to the Syria (Nd Wir, Kobani was recorded as haive urds, with Arab, Turkmen, and Armenian minorities. (m)	مي العرب، إلى المنابع المرابع المرابع وجهد من العرب المرابع المرابع المرابع المرابع المرابع المرابع المرابع المرابع من المرابع المر مام المرابع المرابع الم مام موالع المام عمل عما	منى المرب ، علمه الم موجود المرب ، علمه المربع ا مربع المربع المربع المربع المربع المربع المربع المربع الم المربع المربع المربع المربع المربع المربع المربع المربع المربع المربع المربع ا المربع المربع الم	من العرب : المرب : bland flub barnell), also known as Ayn al-Arab (Arabic: مراجع معلم المرب : papo Governorate in northem by sin, slying Immoduleed south of the cost has been under cost oil of the National YPG mitta and 2012. Not of the cost of the cost and the south of the south of the sweat. Most of the coly was distributed and most of the poolutions Prote to the Systen CMW kr. Kolani was recorded as having a under, with Arab, Turkmen, and Armenian misorities. (wi
				do areaTotal		<ul> <li>7000000.000000 (x)</li> </ul>	d:double)										
				do country		e ex:Syria											
				do.elevation		<ul> <li>520.000000 (estatout</li> </ul>	ie .										
				de isPartOf		as:Ayn_al-Arab_Dis	trict										
						actience Governo	rate										

Figure 1: Results of searching for 'Kobani' in GeoNames (left) and DBpedia (right).

However, we also pointed out that given today's gazetteer ontologies/vocabularies the comparison of place types often boils down to string matching or simple structural measures as the used ontologies rely on a lightweight axiomatization (if any) and thus are not readily alignable. Here we propose to use the mined place type signatures as an additional matching characteristic that communicates the semantics of place types beyond labels alone. To give an intuitive example, we will make use of the fact that the spatial distribution of places of type *populated place* differs substantially from those computed for *river* but not from those computed for *seat of second-order administrative division* irrespective of the fact that those places are in different gazetteers and that the type labels show no similarity (Zhu et al., 2016).

As illustrated in Table 2, there are three place types associated with candidates for Kobani in GeoNames and one place type (*populated place*) in DBpedia. Computing the Euclidean distance between these three GeoNames place signatures (essentially feature vectors comprised of the 27 statistics listed in Table 1) and the *populated place* signature from DBpedia shows that *geonames: seat of second-order administrative division* and *dbpedia: populated place* should indeed be matched. This matching score would then be combined with the other classical matchers such as toponyms, spatial (centroid) distance, and so forth; thereby improving coreference resolution.

Table 2: Dissimilarities between the *populated place* signature for DBpedia and three example placetype signatures in GeoNames.

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

It should be noted that the place type *populated place* in GeoNames shares the same name with the DBpedia type. However, there is a substantial dissimilarity between their signatures. Such an observation indicates that despite a high string similarity, two place types might still have a different underlying semantics and we were able to show exactly this in previous work (Zhu et al., 2016). This further confirms that using string matching in isolation, either for the feature names or types, is not necessarily sufficient for coreference resolution.

### 2.2 Place Type Signatures of Neighboring Places

One drawback of the approach outlined thus far 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. To the best of our knowledge, neighboring places have not been proposed as part of any existing matching framework.

For our running example, we queried the 9 nearest neighbors for each candidate place and recorded their place types. This can be done directly using the precomputed *nearbyFeature* RDF predicate in GeoNames. Next, the averaged signatures (i.e., the averaged feature vector of the 27 statistics listed in Table 1) of these 9 place types are calculated for characterizing the neighborhood of the specific candidates. Lastly, candidates that have the smallest Euclidean distance in terms of their averaged neighboring signatures are regarded as corresponding places. Since neighbors differ from candidate to candidate, their averaged neighboring signatures also differ despite potentially having the same place type. For instance, one populated place of a given name will have places of different types (e.g., a river and an island) nearby while another place of the same type (and similar names) will have places of other types (e.g., a mountain) nearby.

We tested this neighborhood based approach using our Kobani running example. Table 3 lists the place types of nearest neighbors for three example candidates in GeoNames and the Kobani from DBpedia. As can be seen, 'Ayn al 'Arab in GeoNames and Kobani in DBpedia both have relatively more diverse neighbors compared to the other two candidates. Table 4 shows the dissimilarity values which lead to the same conclusion (and correct matching) made in the direct matching case above and further supports our proposed approach.

Table 3: List of place types of the nearest neighbors for example candidates in GeoNames and DBpedia.

	'Ayn al' Arab	Kobani	Mkoani	Kobani		
	(GeoNames: seat of a second-	(GeoNames:	(GeoNames:	(DBpedia:		
	order administrative division) s		populated place)	populated place)		
	section of populated place	populated place	populated place	settlement		
	office building	stream	populated place	settlement		
	school	stream populated place				
Feature types of 9 nearest	square	stream	third-order administrative division	settlement		
neighbors	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 p		populated place	populated place		

 Table 4: Dissimilarities between Kobani's neighboring signatures in DBpedia and the three example places' neighboring signatures in GeoNames.

Dissimilarity (Euclidean distance) with Kobani (DBpedia					
'Ayn al' Arab (GeoNames)	4.23				
Kobani (GeoNames)	10.57				
Mkoani (GeoNames)	6.98				

## 3 Conclusions and Future work

In this work, we presented an initial case study that demonstrates how semantic 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 geographic features.

## References

- 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.
- 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.
- 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.