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.

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

<table>
<tr><th colspan="2">Spatial Point Patterns</th><th>Spatial Autocorrelations</th><th colspan="2">Spatial Interactions with Other Geographic Features</th></tr>
<tr><td rowspan="11">Local</td><td>Intensity</td><td rowspan="4">Global Moran' I</td><td rowspan="4">Internal</td><td rowspan="2">Count of distinct nearest feature types</td></tr>
<tr><td>Mean distance to nearest neighbor</td></tr>
<tr><td>Variance distance to nearest neighbor</td><td rowspan="2">Entropy of nearest feature types</td></tr>
<tr><td>Kernel density (bandwidth)</td></tr>
<tr><td>Kernel density (range)</td><td rowspan="3">Semivariogram value (at first distance lag)</td><td rowspan="8">External</td><td rowspan="2">Population value (min)</td></tr>
<tr><td>Ripley's K (range)</td></tr>
<tr><td>Ripley's K (mean deviation)</td><td>Population value (max)</td></tr>
<tr><td>Standard deviation ellipse (rotation)</td><td rowspan="3">Semivariogram value (at median distance lag)</td><td>Population value (mean)</td></tr>
<tr><td>Standard deviation ellipse (std dev along x-axis)</td><td>Population value (std dev)</td></tr>
<tr><td>Standard deviation ellipse (std dev along y-axis)</td><td>Shortest distance to road (min)</td></tr>
<tr><td rowspan="3">Global</td><td>Intensity</td><td rowspan="3">Semivariogram value (at last distance lag)</td><td>Shortest distance to road (max)</td></tr>
<tr><td>Kernel density (bandwidth)</td><td>Shortest distance to road (mean)</td></tr>
<tr><td>Kernel density (range)</td><td>Shortest distance to road (std dev)</td></tr>
</table>

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.

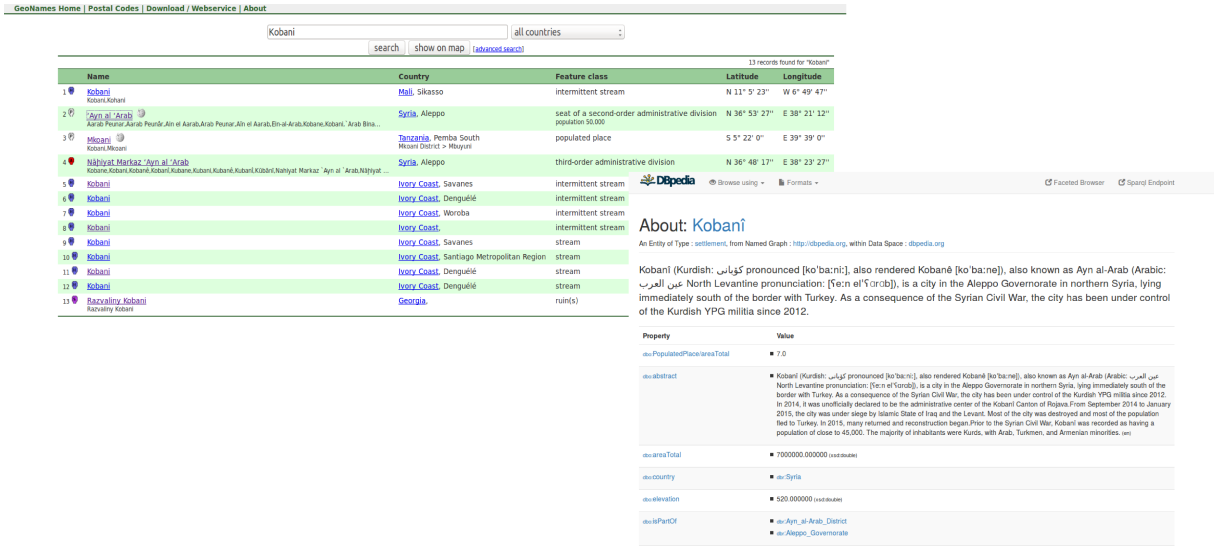

Figure 1: Results of searching for 'Kobani' in GeoNames (left) and DBpedia (right).

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 place type signatures in GeoNames.

Dissimilarity (Euclidean distance) with DBpedia: *Populated 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 (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

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.