A Dataset and Evaluation Framework for Complex Geographical Description Parsing

Egoitz Laparra
School of Information
University of Arizona
Tucson, AZ, USA
laparra@email.arizona.edu

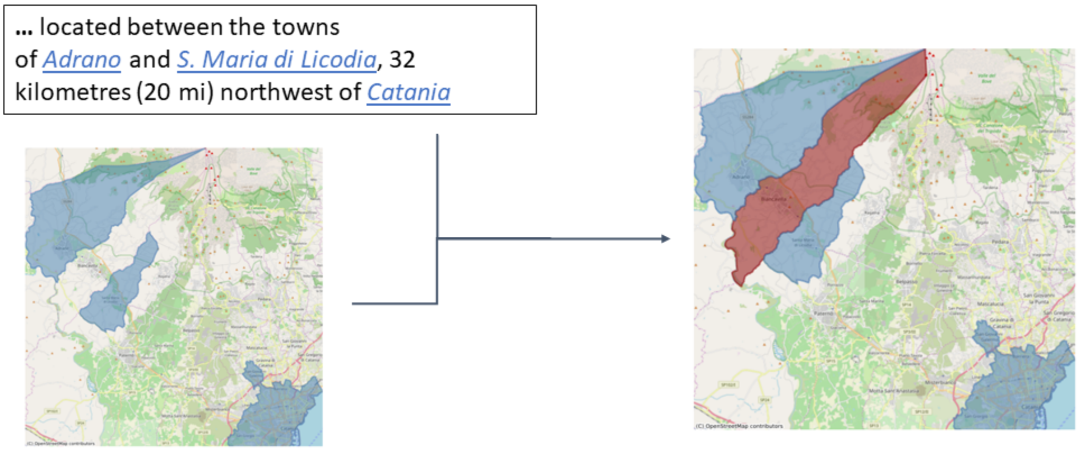
Steven Bethard
School of Information
University of Arizona
Tucson, AZ, USA
bethard@email.arizona.edu

Abstract

Much previous work on geoparsing has focused on identifying and resolving individual toponyms in text like *Adrano*, *S.Maria di Licodia* or *Catania*. However, geographical locations occur not only as individual toponyms, but also as compositions of reference geolocations joined and modified by connectives, e.g., *"…between the towns of Adrano and S.Maria di Licodia, 32 kilometres northwest of Catania"*. Ideally, a geoparser should be able to take such text, and the geographical shapes of the toponyms referenced within it, and parse these into a geographical shape, formed by a set of coordinates, that represents the location described. But creating a dataset for this complex geoparsing task is difficult and, if done manually, would require a huge amount of effort to annotate the geographical shapes of not only the geolocation described but also the reference toponyms. We present an approach that automates most of the process by combining Wikipedia and OpenStreetMap. As a result, we have gathered a collection of 360,187 uncurated complex geolocation descriptions, from which we have manually curated 1,000 examples intended to be used as a test set. To accompany the data, we define a new geoparsing evaluation framework along with a scoring methodology and a set of baselines.

1 Introduction

Geoparsing, or toponym resolution, is the task of identifying geographical entities, and attaching them to their corresponding reference in a coordinate system (Gritta et al., 2018b). In its traditional setting it



… located between the towns of Adrano and S. Maria di Licodia, 32 kilometres (20 mi) northwest of Catania

Figure 1: An illustrative example of complex geographical description parsing. Given a text describing a target geolocation the goal is to aproximate its geometry using as reference the geometries of the geolocations that appear in the description (*Adrano*, *S.Maria di Licodia* and *Catania*). In this example, the target geolocation is **Biancavilla**, but in general, there may not be a name for the target geolocation.

Proceedings of the 28th International Conference on Computational Linguistics, pages 936–948
Barcelona, Spain (Online), December 8-13, 2020

involves two main steps: parsing the text to recover geo-location mentions and linking those mentions to the entries of a toponym dictionary like GeoNames[1]. Thus, it can be seen as a special case of named entity recognition and disambiguation.

The task plays a key role in information extraction, since it allows events to be placed in an unequivocal location that could be plotted on a map. However, the setting described above restricts the task to geolocations with explicit names, whilst text in natural language can refer to geographical regions by complex descriptions that involve a set of different toponyms.

Take, for example, the following snippet: *"a town and comune in the Metropolitan City of Catania, Sicily, southern Italy… located between the towns of Adrano and S.Maria di Licodia, 32 kilometres (20 mi) northwest of Catania."*. Traditional geoparsing could return coordinates for the individual toponyms (*Adrano*, *S.Maria di Licodia*, etc.) but not the boundaries of the geolocation described by the entire phrase. Solving this problem requires understanding the meaning of linguistic connectives (e.g. *between*, *northwest*) and a geometric interpretation of how they combine the geographical entities. These complex descriptions are frequent in domains where there are no specific names for geographical regions of interest, for example in environmental impact statements describing the effects on the environment of activities such as constructing new mines or roads, or in food security outlooks describing the impact of localized agricultural production and other factors on humanitarian crises.

Though parsing these kinds of geographical descriptions and aproximating the corresponding geometries has been the goal of some previous work (Aflaki et al., 2018), research in this area has been constrained by the lack of a benchmark dataset. A naive approach to building such a dataset would require a huge amount of resources to find and collect complex geograhical descriptions, to annotate the reference toponyms and to edit the geometries that correspond to the target locations.

In this paper, we present a methodology to acquire such a dataset by combining Wikipedia and OpenStreetMap[2], making use of the links between those resources. We collect descriptions of geolocations from Wikipedia articles that are linked to an OpenStreetMap geometry, i.e. the set of coordinates that define the boundary of a geographical location. The text from the Wikipedia pages retrieved includes embedded links to other geolocations for which OpenStreetMap geometries can be obtained.

For example, the snippet presented above corresponds to the description of *Biancavilla* in Wikipedia. In our proposed task, we assume that the geometry of *Biancavilla* is unknown and must be recovered from its description using the rest of toponoym geometries (*Adrano*, *S. Maria di Licodia*, …) as reference, as shown in Figure 1.

As the content of both Wikipedia and OpenStreetMap has been manually created and reviewed by a broad commnunity of volunteers, the remaining steps needed for dataset construction are to collect the descriptions, filter non-compositional examples, and, in a some cases, complete missing reference geolocations. In summary, the main contributions of this paper are the following:

- We describe a methodology to adquire a dataset for complex geographical description by combining Wikipedia articles and OpenStreetMap geometries. As a result, we provide a new dataset that includes more than 300,000 uncurated descriptions to be used as a training set, and 1,000 manually curated descriptions to be used as test set.
- We complete an evaluation framework with a set of metrics to analyze the similarity between gold and predicted geometries
- We propose a set of baselines that can be used as reference in future approaches and highlight the need of proper language understanding of the descriptions.

The dataset and evaluation framework are available at `https://github.com/EgoLaparra/geocode-data` and `https://github.com/EgoLaparra/geocode` respectively.

[1] http://www.geonames.org

[2] https://www.openstreetmap.org

2 Task Description

Before going into further details about the building of our dataset, we first formally define the task of parsing complex geographical descriptions. Formally, a geoparser for such descriptions is a function $h : (T, L) \rightarrow G$, where:

- $G = \mathcal{P}(\{(lat, lon) : lat \in [-90, +90], lon \in [-180, +180]\})$ is the set of all sets of (latitude, longitude) points, i.e., the set of all possible geographical regions.
- $T = c_1 c_2 c_3 \ldots c_N$ is the text description of the geolocation, typically one or two brief paragraphs, where c_i is a character in the text.
- $L = \{(c_i, c_j, g_k) : 0 \leq i \leq N, 0 \leq j \leq N, g_k \in G\}$ is the set of other geolocations mentioned in T , where (c_i, c_j) are the character offsets of the mention, and g_k is the geometry of the mentioned geolocation.

Intuitively, the goal is to be able to map a natural language description and the geographical regions it mentions to a new geographical region that represents the complex geolocation described. All geographical regions are defined in terms of longitude and latitude points in the World Geodetic System (WGS 84) (Slater and Malys, 1998)[3].

This task is challenging because every linguistic composition of phrases must be matched to a geometric composition of geographic regions. Consider the first part of the description in Figure 1: *"between the towns of Adrano and S.Maria di Licodia"*. To parse this expression, one must first identify the phrases that represent spatial concepts (e.g. *between*, *Adrano*, *S.Maria di Licodia*). Phrases that represent known geographical regions must then be retrieved from an index of such geometries. For example, $\textsc{Geometry}(\textit{Adrano}) = \{(14.8258023, 37.6324199), (14.8275266, 37.6319645), \ldots\}$ in OpenStreetMap. Phrases that represent compositions of other geographical regions in the description must be mapped to formal geometric operations. For example, *between* should be mapped to:

$$\textsc{RegionBetween}(\textsc{Geometry}(\textit{Adrano}), \textsc{Geometry}(\textit{S.Maria di Licodia}))$$

where the REGIONBETWEEN function must be defined in terms of geometric operations over geographical regions. Note that there is no standard definition for REGIONBETWEEN, so one must be created as part of constructing a parser for complex geographic descriptions. This example illustrates that a successful geoparser will have to be able to map a wide variety of linguistic expressions to a wide variety of geographic regions and geometric operations.

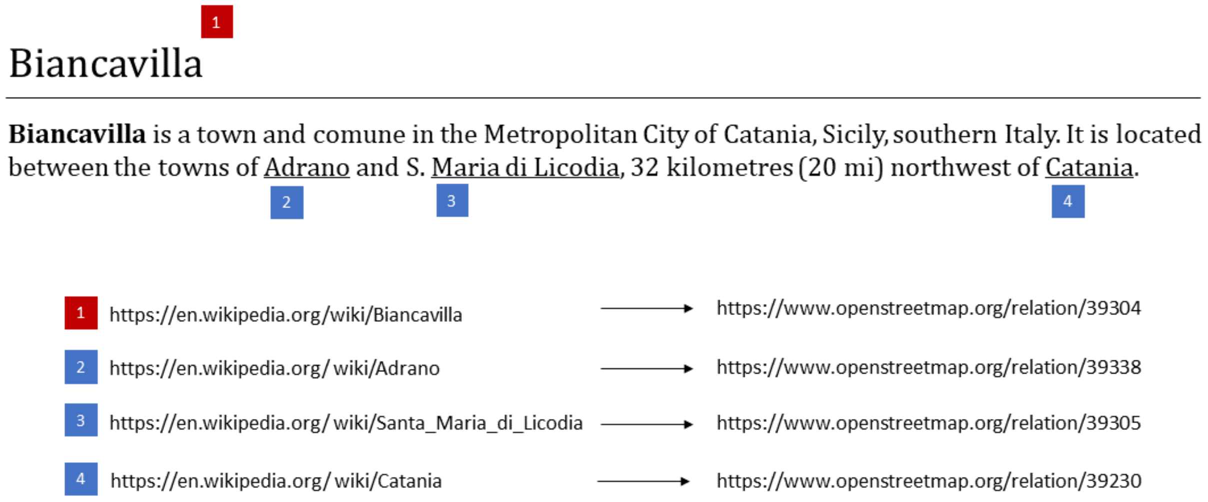
3 Related Work

The task proposed in this paper can be seen as a junction of two different NLP research lines.

On the one hand, toponym resolution has been widely studied, especially within the Entity Linking framework (Shen et al., 2015). But it has gained interest as an independent task in recent years. Its ultimate goal is to retrieve the coordinates for the geolocation mentioned in the text, for which gazetteers, such as GeoNames and Wikipedia, are frequently used as reference knowledge bases. The problem is commonly separated into two steps. First, location references are extracted from the text following a Named Entity Recognition strategy (Karagoz et al., 2016; Magge et al., 2018). Secondly, the references are disambiguated and linked to the reference knowledge base (Turton, 2008; Weissenbacher et al., 2015; Gritta et al., 2018a). As pointed out by Gritta et al. (2018b), the amount of annotated data for toponym resolution is not large, but a few datasets are available for different domains such as historical texts (DeLozier et al., 2016), social media (Wallgrün et al., 2014), scientific literature (Weissenbacher et al., 2019) and Wikipedia (Gritta et al., 2018b). Thanks to these works, it is possible to link locations mentioned in text to unambiguous geographic respresentations, in the form of coordinates. However, they do not address the compositional nature of the geographical descriptions.

On the other hand, although not necesarily related to toponyms, previous works have studied the structure of (geo)spatial natural language expressions. However, many of these works limited their

[3] Appendix A.1 includes examples of WGS84 points, e.g., (15.0915738 37.3582971).



1 Biancavilla

Biancavilla is a town and comune in the Metropolitan City of Catania, Sicily, southern Italy. It is located between the towns of Adrano and S. Maria di Licodia, 32 kilometres (20 mi) northwest of Catania.

1	https://en.wikipedia.org/wiki/Biancavilla	https://www.openstreetmap.org/relation/39304
2	https://en.wikipedia.org/ wiki/Adrano	https://www.openstreetmap.org/relation/39338
3	https://en.wikipedia.org/ wiki/Santa_Maria_di_Licodia	https://www.openstreetmap.org/relation/39305
4	https://en.wikipedia.org/ wiki/Catania	https://www.openstreetmap.org/relation/39230

Figure 2: On the top half, the first paragraph from the Wikipedia article of *Biancavilla* where 2, 3 and 4 are hyperlinks to other Wikipedia articles and will be used as reference geolocations. On the bottom half, the mappings between Wikipedia and OpenStreetMap for the target and reference geolocations in the paragraph. For simplicity, we only show the links of the second sentence but whole paragraph is included in the dataset.

attention to the role of specific elements, such as prepositions (Zwarts, 2005; Kracht, 2008; Radke et al., 2019), or particular grammatical structures (Stock and Yousaf, 2018). Some other works defined the task as a tagging problem, each specifying a different set of labels to tag the components of the text. These works were focused on the relations between the elements of the geospatial expressions in a syntactically inspired manner (Kordjamshidi et al., 2011; Mani et al., 2010; Kolomiyets et al., 2013; Pustejovsky et al., 2015). As pointed out by Aflaki et al. (2018), the manual annotation of compositional (geo)spatial expressions is very challenging and, in consequence, many of these schemas are limited to a reduced set of elements or define complex tag structures. Moreover, the goal of these works is to obtain a tag structure on top of the text but they do not provide an explicit way to intepret them to obtain a single geographical object, i.e. it is not possible to translate these structures into a set of coordinates.

Our work takes a different approach by not imposing any annotation structure on top of the text and providing, instead, the target and reference geometric shapes, in WGS84 coordinates, of the geolocation descriptions. By doing so, we aim to merge benefits from both research lines.

4 Dataset building

We constructed a dataset from Wikipedia, Wikidata and OpenStreetMap as follows:

1. Retrieved the articles from Wikipedia that have a link to OpenStreetMap. As shown in the top half of Figure 2, from every article we kept only the first one or two paragraphs if they satisfy the following criteria:
   - They contained a least two links to Wikipedia articles that are linked to OpenStreetMap (i.e., Wikipedia articles representing geolocations).
   - They contained at least one locative term (*north*, *close*, *between*, etc.). We manually assembled a dictionary of such terms from sources like WordNet and FrameNet. For the purposes of this filter, we did not consider as locative terms words that by themselves would yield only simple expressions, e.g., *in* as in *"Paris is in France"*.
2. Built a mapping dictionary between Wikipedia and OpenStreetMap (see bottom half of Figure 2). When the OpenStreetMap entity is linked to Wikidata or to a Wikipedia redirect we recover the corresponding Wikipedia page.

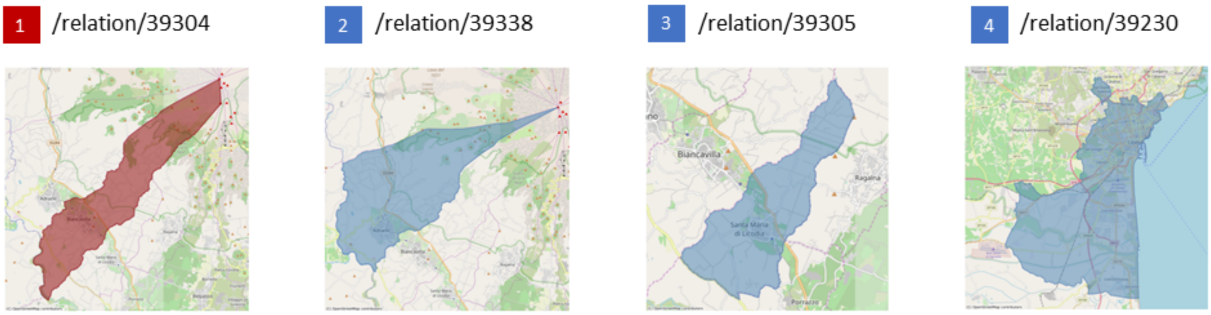


Figure 3: The geometries retrieved from OpenStreetMap for the geolocations in the example of Figure 2. Each geometry is defined by the set of coordinates that makes its boundary.

3. Gathered the geometries of entities from OpenStreetMap that have a link to Wikipedia or Wikidata. The OpenStreetMap geometries are defined by sets of coordinates and can be nodes (a single coordinate), ways (a collection of nodes forming a linestring or a polygon) and relations (a collection of any kind of entities). Figure 3 presents the plotted maps for some examples.

4.1 Analysis and curation

The result of the process described above is a corpus with 360,187 examples. Although it is derived from handcrafted resources, since our process is fully automatic, we performed a manual analysis to understand the quality of the acquired data. We sampled 4,000 descriptions and manually rated them with one of the following categories:

Complex/Complete (22% of descriptions): The description was a compositional description and each reference geolocation has a link to OpenStreetMap. E.g.: *"Fengnan District is a district of Tangshan, Hebei, China on the coast of the Bo Sea and bordering Tianjin to the west"*.

Complex/Incomplete (37%): The description is a compositional description but for some reference geolocations the link to OpenStreetMap is missing. For example, the link for *Weymouth Harbour* is missing in *"Hope Square is a historic square to the south of Weymouth Harbour in the seaside town of Weymouth, Dorset, southern England"*.

Simple/Complete (27%): The description is not compositional although each reference geolocation has a link to OpenStreetMap. E.g.: *"Waakirchen is a municipality in the district of Miesbach in Bavaria in Germany"*.

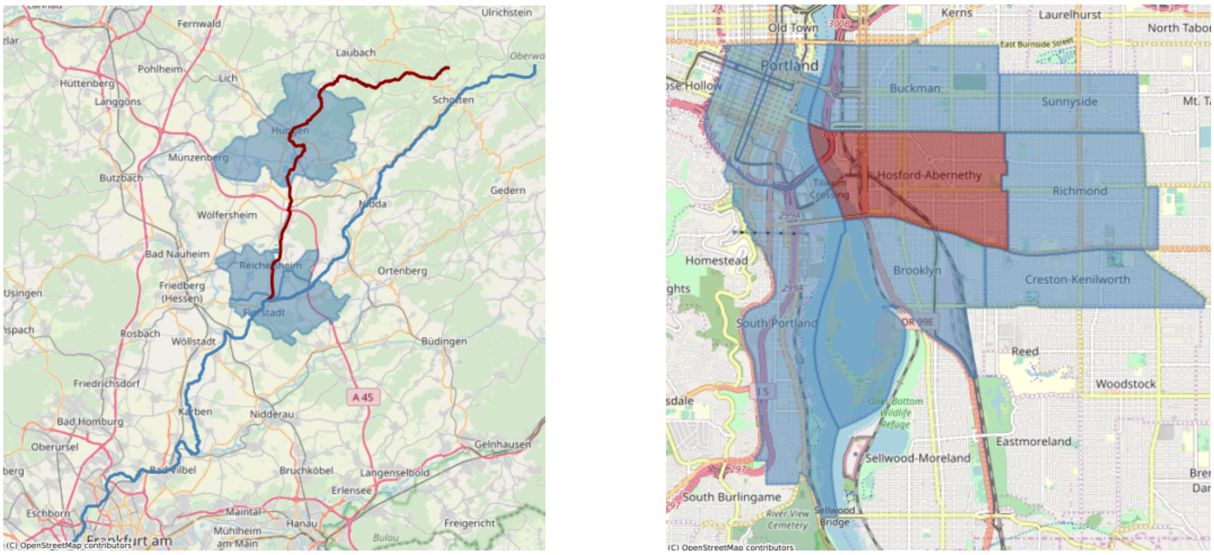
Simple/Incomplete (9%): The description is not compositional and for some reference geolocations the link to OpenStreetMap is missing. For example, the link for *San Gabriel Valley* is missing in *"North El Monte is a census-designated place in the San Gabriel Valley, in Los Angeles County, California, United States"*.

Not valid (5%): The text is not a geographical description. E.g.: *"Yoox Net-A-Porter Group is an Italian online fashion retailer created on 5 October 2015 after the merger between Yoox Group and The Net-A-Porter Group"*. This case was retrieved because it contains a locative term (*between*) and the *Groups* are linked to their headquarters.

From this analysis, it can be observed that only a small fraction of the examples do not correspond to geographical descriptions. It must be noted that, even for Simple or Incomplete categories, all the valid descriptions (95% of the total) are linked to a correct target geometry. However, the Complex/Complete category is the most interesting for our purposes.[4]

We also used this analysis to curate 1,000 examples for a test set. From the 4,000 descriptions sample, we kept only those that are Complex/Complete and a small portion of Complex/Incomplete that only had

[4] The categories we assigned here are intended only to assist in analyzing the quality of the extracted geographical descriptions, and to select items for the manually curated test set. We do not believe these categories are appropriate for a classification task.



(a) Horloff is a river of Hesse, Germany. It passes through Hungen and Reichelsheim, and flows into the Nidda in Florstadt.

(b) Hosford-Abernethy is a neighborhood in the inner southeast section of Portland, Oregon. It borders Buckman and Sunnyside on the north, Richmond on the east, Brooklyn and Creston-Kenilworth on the south, and (across the Willamette River) Downtown Portland and South Portland on the west.

Figure 4: Two examples of descriptions taken from out dataset, including the input text and the *reference* and **target** shapes: (a) example of a target linestring, (b) example of a target polygon. For clarity, the plots have been centered in the target entities and large scale geometries are not shown (e.g. Germany).

one unambiguous entity missing. For this latter category, we manually added the links for the missing entities. Although incorporating Simple and even Not valid descriptions could have resulted in a more realistic setting, we decided not to include them since the former are covered by traditional toponym resolution and the latter are not resolvable. To help us with both the analysis and the curation, we developed an annotation tool that is available at `https://github.com/EgoLaparra/geocode`.

5 Description of the resource

The set of descriptions obtained by the procedure described in Section 4 are stored in a collection of XML files. For every geolocation we include the cannonical Wikipedia name as well as its corresponding identifier and type of entity in OpenStreetMap. Along with the xml files, we built a PostgreSQL database to store the geographic objects collected from OpenStreetMap. Examples of the storage formats can be seen in Appendix A.1.

Examples from the data described are visualized in Figure 4. The maps have been plotted using the geometries retrieved from OpenStreetMap. All the linestrings and polygons correspond to the geolocations, both target and references, from the descriptions below each map.

Table 1 shows further details on the 360,187 gathered descriptions and the manually curated test set. The most remarkable differences between the uncurated and curated sets are the ratio of unique references and the number of references per description. These differences can be explained by the missing reference geolocations in the Complex/Incomplete descriptions and by the fact that Simple descriptions usually have fewer references that are also more frequently reused. For example, the uncurated set contains more than 20,000 French commune descriptions like: *“Le Mesnil-Esnard is a commune in the Seine-Maritime department in the Normandy region in northern France”*. These kinds of descriptions tend to share the same references (e.g. *Normandy*, *France*). These characteristics of the uncurated data should be taken into account when using it as a training set to avoid undesired biasses.

Finally, it should be noted that, although all the descriptions come from the same source, namely Wikipedia, the resource covers a wide variety of locations, inluding landforms, political entities, human

	Uncurated	Curated (test)
#descriptions	360,187	1,000
#refs	1,524,699	5,298
#unique refs	171,980	3,529
#refs per description	4.23	5.30

Table 1: Some statistics of the dataset.

settlements or artificial structures.

6 Metrics

As the main metric for models attempting this geoparsing task, we propose to measure the area of the predicted geometry that overlaps with the gold one. In other words, the metric calculates the area of the intersection between both geometries. This value can be obtained in precision and recall as follows:

$$P_i(S_i, O_i) = \frac{area(S_i \bigcap O_i)}{area(S_i)} \qquad R_i(S_i, O_i) = \frac{area(S_i \bigcap O_i)}{area(O_i)}$$

where $area(X_i)$ is the geometric area of the geolocation i from set X. Given the predicted (S) and gold (O) sets of geometries, we define the overall precision as the average of individual precisions in S. Similarly, overall recall is defined as the average of inidividual recalls in O:

$$P(S, O) = \frac{1}{|S|} \sum_{i \in |S|} P_i(S_i, O_i) \qquad R(S, O) = \frac{1}{|O|} \sum_{i \in |O|} R_i(S_i, O_i)$$

$F_1(S, O)$ can be calculated as the regular harmonic mean of $P(S, O)$ and $R(S, O)$.

This metric is run following two different criteria. First, we propose a **strict** evaluation by measuring the exact overlap between the original gold and predicted geometries. However, as geolocations can have very intrincate boundaries, this criterion becomes too demanding in many cases. Thus, we also propose a **relaxed** version where the metric calculates the overlap with the target oriented envelopes, i.e. the mimimum rotated rectangle enclosing each target geometry.

Note that the overlap based metric penalizes equally all predictions that do not intersect with a gold geometry. We propose an additional scoring method that gives some credit to predictions that are close to the gold geometries even if they do not overlap. For this, we extend the gold geometries by a scale factor of 2 and then calculate the precision of the predictions as explained previously. In other words, to calculate the precision P_{x2}, we double the size of gold geometries. Similarly, we scale the predicted geometry by a factor of 2 and calculate the recall R_{x2}. This two complementary metrics can be applied with both strict and relaxed criteria.

7 Baselines

In order to have a reference for future approaches, we propose four non-lingustic baselines that ignore the text description, and only execute straightforward combinations of the reference geolocations. Models which do any real language processing of the text descriptions should be able to exceed these baselines. We also propose a linguistic baseline that constructs a semantic parser from a hand-assembled synchronous grammar to parse the description of the target geometry and produce a composition of operations over the reference geometries. We expect machine learning models that combine both text and reference geometries will outperform this baseline.

7.1 Non-linguistic baselines

The non-linguistic baselines all consider only the set L of reference geometries for a particular target geolocation, ignoring the text T. The baselines are defined as follows:

	Strict					Relaxed					Coverage
baseline	P	R	F_1	P_{x2}	R_{x2}	P	R	F_1	P_{x2}	R_{x2}	%
Max reference	0.005	0.849	0.011	0.014	0.937	0.009	0.844	0.017	0.025	0.935	**100%**
Min reference	0.100	0.105	0.102	0.196	0.184	0.144	0.099	0.117	0.284	0.179	**100%**
Union	0.005	**0.928**	0.010	0.015	**0.954**	0.009	**0.927**	0.018	0.027	**0.955**	**100%**
Intersection	0.106	0.063	0.079	0.216	0.081	0.161	0.059	0.086	0.289	0.079	25.5%
Grammar	**0.172**	0.310	**0.221**	**0.272**	0.381	**0.213**	0.276	**0.240**	**0.358**	0.365	52.8%

Table 2: Performance of the five baselines using the scoring functions proposed. **Coverage** shows the percentage of cases where the baseline produces a prediction. The highest scores in each column are marked in bold.

- The **maximum** and **minimum** reference geometry in L according to the area sizes.

$$h_{\text{min}}(T, L) = \arg\min_{l \in L} area(l) \qquad\qquad h_{\text{max}}(T, L) = \arg\max_{l \in L} area(l)$$

- The **union** and **intersection** of all the reference geometries in L.

$$h_{\text{union}}(T, L) = \bigcup_{l \in L} l \qquad\qquad h_{\text{intersection}}(T, L) = \bigcap_{l \in L} l$$

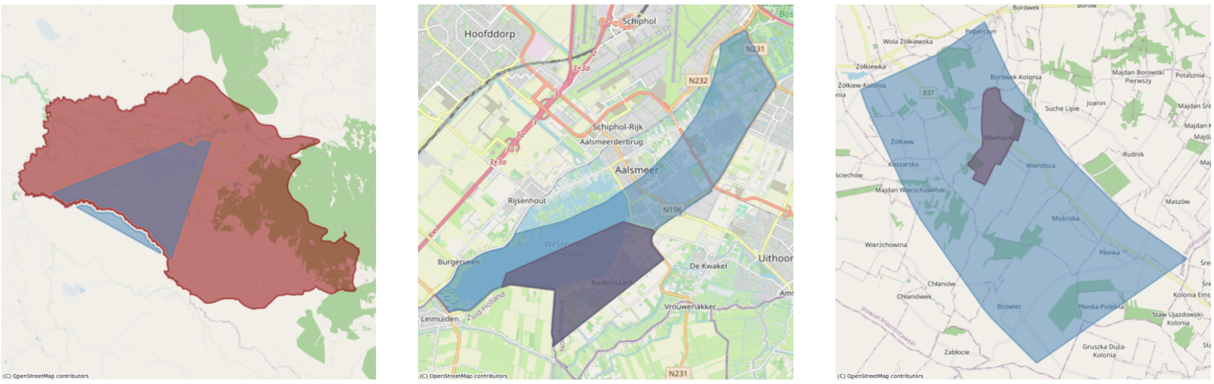
Table 2 shows the performance of the four baselines. The maximum reference and union baselines are overly inclusive, obtaining high recall in both strict and relaxed metrics, but with very low precisions. This reflects the large size of the predicted geometries that, in many cases, cover the target geolocation. The minimum reference baseline performs the best in terms of F_1, but has low recall. The intersection baseline seems to be the most precise, but has poor recall, since it only produces a prediction for a small fraction of the descriptions (only for 25.5% of the cases). In general, the poor performance of the four baselines show the need for natural language processing techniques that truly understand the textual descriptions.

7.2 Linguistic baseline

For the linguistic baseline, we have built a semantic parser based on a synchronous grammar using a dev set of ∼100 randomly sampled examples from the non-curated collection (ensuring that the test set was not used during the development of the grammar). Synchronous grammars allow the construction of two simultaneous trees, one in a source language and one in a target language. In our case, the source is the natural language text in English and the target is a formal grammar of geometry operators. Each of these operators defines a function to produce a geometry from others. For example, the BETWEEN operator takes two geolocations and calculates the region between them. We run our synchronous grammar with the extended CYK+ parsing algorithm described in Bethard (2013). The grammar contains 219 rules in total, 70 of which are lexical, e.g. $[\text{CARDINAL}] \rightarrow northwest \parallel NW$.

To illustrate our parsing process, we use as an example the following description: *"... located between the towns of Adrano and S.Maria di Licodia, 32 kilometres (20 mi) northwest of Catania"*. The text is first pre-processed, cleaning some unnecessary tokens and normalizing each geolocation to a *SHP Index* format: *"... between the towns of SHP 001 and SHP 002, 32 kilometres northwest of SHP 003"*. The *Index* is mapped to the geometry of its corresponding geolocation (e.g. 001 → *relation/39338*). Next, consider the following portion of our synchronous grammar with the source on the left of || and the target on the right:

$[\text{NIL}] \rightarrow towns \parallel$
$[\text{UNIT}] \rightarrow kilometres \parallel KM$
$[\text{CARDINAL}] \rightarrow northwest \parallel NW$
$[\text{LOCATION}] \rightarrow SHP\ [\text{INT}] \parallel \text{TOPONYM}([\text{INT}])$
$[\text{LOCATION}] \rightarrow between\ [\text{LOCATION}_1]\ and\ [\text{LOCATION}_2] \parallel \text{BETWEEN}([\text{LOCATION}_1], [\text{LOCATION}_2])$
$[\text{LOCATION}] \rightarrow [\text{INT}][\text{UNIT}][\text{CARDINAL}]\ of\ [\text{LOCATION}] \parallel \text{DISTANCE}([\text{LOCATION}], [\text{INT}], [\text{UNIT}], [\text{CARDINAL}])$
$[\text{LOCATION}] \rightarrow [\text{LOCATION}_1]\ ,\ [\text{LOCATION}_2] \parallel \text{INTERSECTION}([\text{LOCATION}_1], [\text{LOCATION}_2])$



(a) Gambela or Gambella. . . is one of the nine ethnic divisions (kililoch) of Ethiopia. . . The region is situated between the Baro and Akobo Rivers, with its western part including the Baro salient.
P: 0.914 R: 0.170.

(b) Kudelstaart. . . is a town in the Dutch province of North Holland. It is a part of the municipality of Aalsmeer, and lies about 10 km southeast of Hoofddorp.
P: 0.327 R: 1.000.

(c) Równianki. . . is a village in the administrative district of Gmina Rudnik, within Krasnystaw County, Lublin Voivodeship, in eastern Poland. It lies approximately 23 kilometres (14 mi) south-west of Krasnystaw and 47 km (29 mi) south-east of the regional capital Lublin.
P: 0.052 R: 1.000.

Figure 5: Three examples of predictions given by the grammar-based baseline. For comparison, we show the *prediction* and **target** geometries. We also include the **strict** precision (P) and recall (R) for each prediction.

The set of rules above produces the following operator composition:

$$\begin{aligned}&\textsc{Intersection}(\\&\quad\textsc{Between}(\textsc{Toponym}(001), \textsc{Toponym}(002)),\\&\quad\textsc{Distance}(\textsc{Toponym}(003), 32, KM, NW)))\end{aligned}$$

When the description is composed of multiple sentences, every sentence is parsed independently and then an INTERSECTION operation is performed. If the result of the INTERSECTION is empty, we return the UNION. We implement all the operators as deterministic functions, guided by their performance in the development set. For example, BETWEEN returns a geometry that corresponds to the area between the portions of the reference geometries that are closest to each other. DISTANCE guesses a geometry at the distance and direction defined by its arguments whose area is based on the reference geometry. See Appendix A.2 for a more detailed explanation.

The last row in Table 2 shows that this model more than doubles the F_1 of the non-linguistic baselines, despite being able to produce an intepretation for only 52.8% of the descriptions. If we consider only the cases where this model returns an interpretation, its strict recall rises from 31.0% to 58.7%, indicating that when the model is able to produce a prediction, it successfully retrieves much of the target geometry area.

The fact that the recall is higher than the precision reflects that the resulting geometries are still too large. This often results from incomplete interpretation; on average only 69.7% of a description's sentences are used when the grammar produces that description's interpretation. As a specific example, consider the description: *"Hornsby Heights is a suburb of northern Sydney. . . Hornsby Heights is located 27 kilometres north-west of the Sydney central business district, in the local government area of Hornsby Shire. . . "* When only the first sentence is parsed, the geometry returned covers the whole *"northern Sydney"*, an area that is aproximately 100 times larger than *Hornsby Heights*.

Besides the limitations of the grammar, the deterministic implemenation of our spatial operators introduces additional sources of error. For example, our BETWEEN operator can have flawed results when the reference geometries are not polygons but linestrings, as shown in Figure 5a where the returned geometry is calculated leaving out the portions of the *Baro* and *Akobo* rivers that are farthest from each other. The performance of DISTANCE operator may vary widely if the size of the reference and target geolocations differ, like in the examples Figures 5b and 5c, especially if the type of the geolocations do not match, as can be seen in Figure 5c where the reference geolocations *Krasnystaw* and *Lublin* are towns whilst the target geolocation is a village.

8 Conclusion

In this paper, we presented a new dataset for complex geographical description parsing. This dataset aims to push the research in this area that has been constrained by the lack of a benchmark dataset. We detailed the methodology to adquire such a dataset by combining Wikipedia and OpenStreetMap geometries, followed by a manual curation of 1,000 descriptions to be used as test set. We also described a novel scoring methodology that completes the evaluation framework. Finally, we proposed a set of baselines, including a grammar-based semantic parser, to be used as reference by future works. The results show that this a very challenging task, but we expect that the limitations of a rule-based system can be addressed by more sophisticated models trained in the uncurated portion of our dataset.

9 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1831551 and the Defense Advanced Research Projects Agency under the World Modelers program, Grant No. W911NF1810014. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

References

Niloofar Aflaki, Shaun Russell, and Kristin Stock. 2018. Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions (Short Paper). In *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:6, Dagstuhl, Germany. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Steven Bethard. 2013. A synchronous context free grammar for time normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 821–826, Seattle, Washington, USA, October. Association for Computational Linguistics.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a Novel Geolocation Corpus from Historical Texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198, Berlin, Germany, August. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018a. Which Melbourne? Augmenting Geocoding with Maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia, July. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018b. Whatś missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, June.

Pinar Karagoz, Halit Oguztuzun, Ruket Cakici, Ozer Ozdikis, Kezban Dilek Onal, and Meryem Sagcan. 2016. Extracting Location Information from Crowd-sourced Social Network Data. In Cristina Capineri, Muki Haklay, Haosheng Huang, Vyron Antoniou, Juhani Kettunen, Frank Ostermann, and Ross Purves, editors, *European Handbook of Crowdsourced Geographic Information*, pages 195–204. Ubiquity Press.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie-Francine Moens, and Steven Bethard. 2013. SemEval-2013 Task 3: Spatial Role Labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255–262, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36, December.

Marcus Kracht. 2008. The fine structure of spatial expressions. *Syntax and Semantics of Spatial P*, pages 35–62, May.

Arjun Magge, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics (Oxford, England)*, 34(13):i565–i573.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. SemEval-2015 Task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June. Association for Computational Linguistics.

Mansi Radke, Prarthana Das, Kristin Stock, and Christopher B. Jones. 2019. Detecting the Geospatialness of Prepositions from Natural Language Text (Short Paper). In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *14th International Conference on Spatial Information Theory (COSIT 2019)*, volume 142 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:8, Dagstuhl, Germany. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, February.

James A. Slater and Stephen Malys. 1998. Wgs 84 — past, present and future. In Fritz K. Brunner, editor, *Advances in Positioning and Reference Frames*, pages 1–7, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kristin Stock and Javid Yousaf. 2018. Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, 32(6):1087–1116, June.

Ian Turton. 2008. A System for the Automatic Comparison of Machine and Human Geocoded Documents. In *Proceedings of the 5th Workshop on Geographic Information Retrieval*, GIR '08, pages 23–24, New York, NY, USA. ACM. event-place: Napa Valley, California, USA.

Jan Oliver Wallgrün, Frank Hardisty, Alan Maceachren, Morteza Karimzadeh, Yiting Ju, and Scott Pezanowski. 2014. Construction and first analysis of a corpus for the evaluation and training of microblog/twitter geoparsers. In *Proceedings of the 8th Workshop on Geographic Information Retrieval, GIR 2014*, page a4. Association for Computing Machinery, Inc, November.

Davy Weissenbacher, Tasnia Tahsin, Rachel Beard, Mari Figaro, Robert Rivera, Matthew Scotch, and Graciela Gonzalez. 2015. Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics (Oxford, England)*, 31(12):i348–356, June.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 Task 12: Toponym Resolution in Scientific Papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 907–916, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Joost Zwarts. 2005. Prepositional Aspect and the Algebra of Paths. *Linguistics and Philosophy*, 28(6):739–779.

A Appendices

A.1 Examples of storage formats

The set of descriptions obtained by the procedure described in Section 4 are stored in a collection of XML files following the format of the example in Figure 6. For every target geolocation (<entity>) we include the cannonical Wikipedia name as well as its corresponding identifier and type of entity in OpenStreetMap. Each geolocation contains one or two paragraphs (<p>) with the text that describes it. The text is enriched with embedded <link> nodes corresponding to the reference geolocations that contains the same attibutes as the target geolocation <entity> node.

Along with the xml files, we built a PostgreSQL database to store the geographic objects collected from OpenStreetMap. We use the PostGIS plugin to gain support for geometric queries and operations. For example, the geometry of the geolocation *relation/39230*, shown in Figure 3, is stored as a MULTILINESTRING made of a set of coordinates like the following:[5]

[5] The coordinates are in (longitude latitude) format following the WGS 84 system.

```
MULTILINESTRING((15.0915738 37.3582971, 15.091534 37.3590011, 15.0914852 37.3594131,
     15.0913862 37.3625999, 15.0911766 37.3637703, 15.0912522 37.365941, 15.090798
    37.3687674, 15.090877 37.3725726, 15.0905403 37.3860332, 15.0905375 37.3861446,
    15.0904987 37.3878887, 15.0906826 37.3885978,...
```

```
<entity wikipedia="Biancavilla" osm="39304" type="relation">
  <p>
    Biancavilla is a town and comune in the <link osm="39181" type="relation"
    wikipedia="Metropolitan_City_of_Catania">Metropolitan City of Catania</link>,
    <link osm="39152" type="relation" wikipedia="Sicily">Sicily</link>, southern
    <link wikipedia="Italy" osm="365331" type="relation">Italy</link>. It is
    located between the towns of <link osm="39338" type="relation"
    wikipedia="Adrano">Adrano</link> and <link wikipedia="Santa_Maria_di_Licodia"
    osm="39305" type="relation">S. Maria di Licodia</link>, 32 kilometres (20 mi)
    northwest of <link wikipedia="Catania" osm="39230"
    type="relation">Catania</link>.
  </p>
</entity>
```

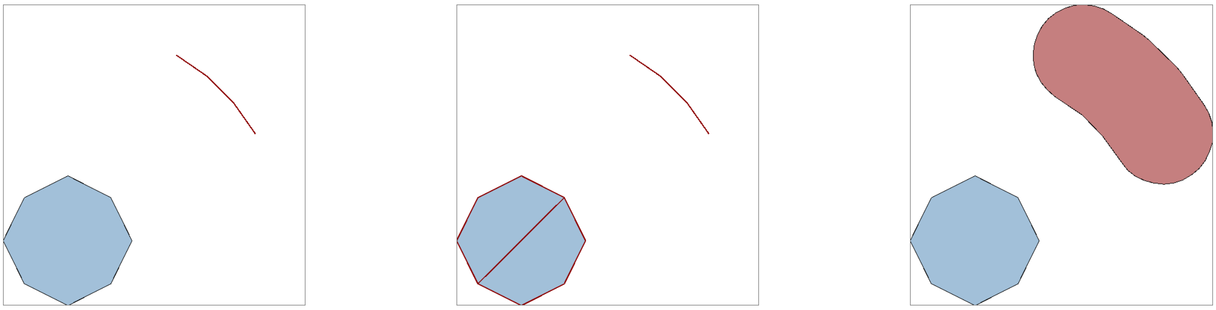
Figure 6: Example of the xml annotation of a geographical description in the dataset.

A.2 Examples of operator implementation

Algorithm 1 and Algorithm 2 contain the pseudo-code for the operators used as example in Section 7.2. Figure 7 and Figure 8 provide a visual description of how the operators are implemented. In both cases, **blue** is used to represent the reference geometries whilst **red** is used to show what the algorithms produce in each step as well as the final geometries.

Algorithm 1 DISTANCE operator

```
    Input: $Geometry, Integer, KM|MI, N|S|E|W|NE\ldots$
    Output: $Geometry$
1: function DISTANCE($ref\_geometry, distance, units, cardinal\_point$)
2:     $ref\_centroid$ = GETCENTROID($ref\_geometry$)
3:     $distant\_point$ = GETDISTANT($ref\_geometry, distance, units, cardinal\_point$)
4:     $distant\_arc$ = GETARC($ref\_centroid, distant\_point$)
5:     $max\_diagonal\_size$ = CALCULATEMAXDIAGONAL($ref\_geometry$)
6:     $distance\_geometry$ = BUFFERARC($distant\_arc, max\_diagonal\_size$)
7:     return $distance\_geometry$
8: end function
```



(a) Get an arc at *distance* from the centroid of the reference geometry and oriented acording to *cardinal_point*. Lines 2-4.

(b) Calulate the maximum diagonal of the reference geometry. Line 5.

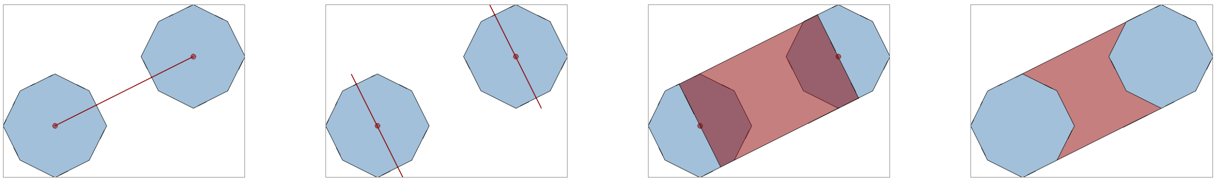
(c) Buffer the arc obtained previously so its width equals the maximum diagonal calculated in the previous step. Line 6.

Figure 7: Visual representation of Algorithm 1.

The BETWEEN operator takes two geometries as input. The DISTANCE operator takes a reference geometry, an integer with the distance value, a string representing the units used, “KM” for kilometers or “MI” for miles, and a string representing the cardinal point, e.g. “S” for south or “NE” for northeast.

Algorithm 2 BETWEEN operator

```
    Input: $Geometry, Geometry$
    Output: $Geometry$
 1: function BETWEEN($ref\_geometry\_1$, $ref\_geometry\_2$)
 2:     $ref\_centroid\_1 =$ GETCENTROID($ref\_geometry\_1$)
 3:     $ref\_centroid\_2 =$ GETCENTROID($ref\_geometry\_2$)
 4:     $shortest\_line =$ GETLINE($ref\_centroid\_1, ref\_centroid\_2$)
 5:     $orthogonal\_1 =$ GETORTHOGONAL($shortest\_line, ref\_centroid\_1$)
 6:     $intersections\_1 =$ GETINTERSECTIONPOINTS($ref\_geometry\_1, orthogonal\_1$)
 7:     $orthogonal\_2 =$ GETORTHOGONAL($shortest\_line, ref\_centroid\_2$)
 8:     $intersections\_2 =$ GETINTERSECTIONPOINTS($ref\_geometry\_2, orthogonal\_2$)
 9:     $between\_polygon =$ GETPOLYGON($intersections\_1 \cup intersections\_2$)
10:     $between\_geometry =$ DIFFERENCE($between\_polygon, ref\_geometry\_1 \cup ref\_geometry\_2$)
11:     return $between\_geometry$
12: end function
```



(a) Get the shortest line between the centroids of the reference geometries. Lines 2-4.

(b) Get the orthogonals of the line calculated in the previous steps that cross the centroids of the reference geometries. Lines 5-7.

(c) Get the points where the orthogonals intersect the reference geometries and build a polygon with them. Lines 8-9.

(d) The final geometry is calculated as the disjoint area of the polygon calculated previously and the reference geometries. Line 10.

Figure 8: Visual representation of Algorithm 2.