GeoDict: an integrated gazetteer

Jacques Fize
UMR TETIS, CIRAD
Montpellier, France
jacques.fize@cirad.fr

Gaurav Shrivastava
Birla Institute of Science and Technology
Pilani, India
gauravsh033@gmail.com

Abstract

Nowadays, spatial analysis in text is widely considered as important for both researchers and users. In certain fields such as epidemiology, the extraction of spatial information in text is crucial and both resources and methods are necessary. In most of spatial analysis process, gazetteer is a commonly used resource. A gazetteer is a data source where toponyms (place name) are associated with concepts and their geographic footprint. Unfortunately, most of publicly available gazetteer are incomplete due to their initial purpose. Hence, we propose Geodict, an integrated gazetteer that contains basic yet precise information (multilingual labels, administrative boundaries polygon, etc.) which can be customized. We show its utility when using it for geoparsing (extraction of spatial entities in text). Early evaluation on toponym resolution shows promising results.

1 Introduction

Nowadays, spatial analysis in text is widely considered as important for both researchers and users. For example, Google search engine is used 30 to 40%[1] of the time for spatial queries such as: *pizzeria in Pao Alto* or *Hotel near Coutances*, etc. In certain fields of research such as epidemiology, extracting information in text is crucial. In epidemiology, textual data represent 60% of the available information (Barboza, 2014). In particular, to study an epidemic spreading, different methods and techniques are necessary to extract spatial information in text.

Most of spatial analysis process are depending on geographical datasets such as gazetteers. A gazetteer is data source where toponyms (place names) are linked to concepts and their geographic footprint (Hill, 2000) (See Figure 1).

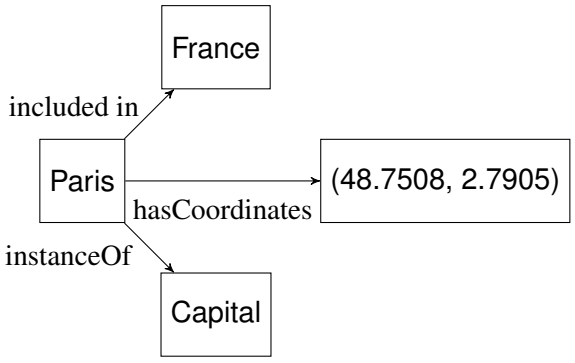

Figure 1: Example of information linked to Paris in a gazetteer

[1]Google Pinpoint 2012, London: https://www.youtube.com/watch?v=ucYiMBfyNfo

The extraction of spatial entities or *geoparsing* can be considered as one of the most important part in spatial analysis. Geoparsing is generally a two steps process:

(i) **Toponym identification**, *e.g. There is town near our house called Paris*

(ii) **Toponym resolution**, *e.g. Paris ⇔ Paris, France? Paris, Missouri? Paris, Illinois? ...*

Geoparsing is also known to be difficult due to text characteristics such as: context, language and text size. We can mention works on short text such as tweet or SMS for which the task is particularly challenging (Li and Sun, 2014; Zenasni et al., 2016).

Until now, most of publicly available gazetteers are incomplete because of their original usage. For example, Getty is destined to be used to catalog work of art. Thus complete data on administrative boundaries or precise coordinates are unnecessary. However, other users may have different usages and create new gazetteers by adding or restraining information to their needs. In this paper, we present Geodict, a customizable gazetteer that contains basic yet precise information (multilingual labels, administrative boundaries polygon, etc.) and its usage in geoparsing. Geodict is available **here**[2].

This paper is organized as follows. In Section 2, we review commonly used gazetteers and geoparsing methods. In Section 3, we present Geodict, its creation process and the associated features we defined for. Then in Section 4, a geoparsing use-case process using Geodict is presented. Finally, we conclude in Section 5.

2 Related Works

This section outlines related works on gazetteers and geoparsing.

2.1 Gazetteers

Geonames Geonames is a publicly available gazetteer. It contains more than 8 million entries linked to different information such as: *a unique ID, coordinates, used name, aliases, etc.* Each entry is classified by a tuple (class, code) *e.g (P, PPL)* → *Populated Place*.

Getty The Getty gazetteer or TGN (The Getty Thesaurus of Geographic Names) is part of datasets (AAT[3], ULAN[4]) used to improve the access of information about art, architecture and material culture. It is composed of approximately 1.3 million entries. Since Getty is destined for arts cataloging, data such as coordinates are less precise and not aimed for GIS[5](Geographic Information System). Interestingly, each entry has its label in different languages and sometimes the time period when it is used. Compared to Geonames, each entry may have coordinates of their administrative boundaries. However, the boundaries are only described by two points.

Others geographical resources like Geodict propose datasets built on linked open datasets. (Stadler et al., 2012) propose LinkedGeoData, a translation of OpenStreetMap to RDF model. However, it's hasn't been updated since 2015.

2.2 Geoparsing

Most of methods with good accuracy are rule-based. (Li et al., 2003; DeLozier et al., 2015) and (Clough et al., 2004) define a special gazetteer where each spatial entity is associated with a unique toponym based on different criteria (popularity, size, population, etc.). (Lieberman et al., 2010; Rauch et al., 2003;

[2] `http://dx.doi.org/10.18167/DVN1/MWQQOQ`

[3] The Art & Architecture Thesaurus

[4] Union List of Artist Names

[5] http://www.getty.edu/research/tools/vocabularies/tgn/about.html

Gazetteer	Nb. of SE[1]	A.B.[2]	Linked to	Customizable
Getty	1477816	✔[3]		
Geonames	11301264			
Geodict	4130301	✔	Geonames, OSM, Wikidata, Wikipedia	✔

[1] Spatial Entities
[2] Administrative Boundaries
[3] Two coordinates (rectangle boundaries)

Table 1: Comparison with other gazetteers

Li et al., 2003) or CLAVIN[6] use geographical scope defined by fixed spatial entities to disambiguate spatial entities. (Rauch et al., 2003; Clough et al., 2004) propose to use contextual information contained in words preceding (resp. following) a toponym.

Data-driven techniques adopt machine learning methods to disambiguate toponyms (Grossman and Frieder, 2004). The main issue of this method dwells within its training corpus which is not available in the community.

(Overell and Rger, 2008) propose to use co-occurrence models. Each document is associated with a list of words ordered by co-occurrences. Then, association rules can be extracted such as *Paris* $\rightarrow$ *France*.

3 GeoDict

A large number of geographical datasets and gazetteers store different pieces of information. Recently, data description strategies were harmonized. Hence datasets are strongly linked and follow similar representation formats (RDF model), it eases data aggregation from different datasets. To build Geodict, we chose to collect detailed representation for each attribute using different sources: Wikidata, Geonames, OpenStreetMap. Thanks to the policy within the Semantic Web (Berners-Lee et al., 2001), all mentioned data sources are easy to link. Ultimately, each entry in Geodict is associated with the attributes described in Table 2.

Wikidata. Wikidata is a publicly available and editable knowledge base. Entries in Wikidata are distinguished in two types: (i) items that represent all *things* in human knowledge *e.g. queen, Barack Obama, etc.*, (ii) properties that allow to represent information of items. Each item is described through *statements* which are composed of:

- a property, *e.g. country (P47)*
- one or multiple value(s), *e.g. France (Q142)*
- information reference/source, *e.g. https://en.wikipedia.org/wiki/Paris*

OpenStreetMaps. OpenStreetMap is free and editable map of the whole world. It was created to help people to access geographical data. OSM entries are divided in three types: *node, way, relation*. Each element is described with one or multiple *tags*. For example, Paris could be associated with tags like: *name=Paris; wikidata=Q90; alt_name=Lutèce*.

[6] `https://clavin.bericotechnologies.com/about-clavin/`
[7] P47: Share border with *e.g. France shares border with Belgium*
[8] P131: located in the administrative territorial entity *e.g. Paris is located in the adm. terr. entity Ile de France*
[9] P706: located on terrain feature *e.g. The Liberty Statue is located on terrain feature "Liberty Island"*

Field	Source	Example Value
Unique ID	Wikidata	*Q30: USA*
Labels	Wikidata	*fr: Cologne, de: Köln, etc.*
Administrative Boundaries	OpenStreetMap	*[[0,1],[1,0], . . .]*
Coordinates	Wikidata	*(48.7508,2.7905)*
Class(es)/Concept(s)	Geonames	*(P, PPL): populated place*
Spatial relationships (P47[7], P131[8], P706[9])	Wikidata	*See footnotes*

Table 2: Entry associated information

	Frequency
A (country, region, . . .)	281951
P (city, village,. . .)	856962
R (road, railroad, . . .)	292124
S (spot, building, farm, . . .)	642148
T (mountain, hill, rock, . . .)	1014332
U (undersea)	4317
V (forest, health, . . .)	10130
H (stream, lake, . . .)	976335
L (parks, area, . . .)	56943
With boundaries	172 645
Total	4 130 301

Table 3: Statistics on Geodict

3.1 Gazetteer creation

The creation process of Geodict is composed of 5 steps:

1. **Harvest basic information on Wikidata** (labels, coordinates, etc.). Since Wikidata is a general knowledge base, we only keep entries which one of the two following conditions:
   - Has a Geonames ID or a OpenStreetMapID (*resp. P1566* and *P402*)
   - Or has the property P706 or P131
2. **Associate one or multiple class(es) (city, canyon, etc.) for each entry**. We associate each available value contained in the property P31[10] (*e.g populated places*) to a Geonames class-code tuple (*e.g P, PPL*).
3. **Find the missing links**. All these data sources are strongly linked. However some links are missing and especially in OpenStreetMap entries. More precisely, some of the entries in OpenStreetMap don't have a Wikidata link but only a Wikipedia link. Fortunately, we know that each Wikipedia page is linked to a Wikidata entry (Vrandečić and Krötzsch, 2014) and each of these links are stored in Wikidata. Thus we search the missing links in OpenStreetMap entries by searching their Wikipedia link in Wikidata.
4. **Add user defined properties.** We associate user specified properties in Wikidata with each entity.
5. **Add the administrative boundaries.** Polygon coordinates representing administrative boundary(ies) are associated with their corresponding entry in the gazetteer.

Once the whole process is executed, a resulting gazetteer is created with 4,130,301 spatial entities divided in different Geonames class as illustrated in Table 3.

[10]P31: instance of *e.g. Barack Obama is an instance of [person, president, lawyer, etc.]*

3.2 Comparison with other gazetteers

We compare Geodict to other available gazetteers using three characteristics: (i) the number of spatial entities, (ii) linked datasets, (iii) if boundaries are available and (iv) if it is customizable. Table 1 sums up the characteristics for all gazetteers.

Obviously, Geodict isn't the most exhaustive because of specific constraints and the chosen pivot dataset (Wikidata). For example, by comparing Geodict with Geonames, we have less entries ($\approx$36% of Geonames). However, we remind that each spatial entity in Geodict is associated with complete information necessary to the geoparsing process. In order to fit different purposes, Geodict is customizable and linked to commonly used dataset such as Wikipedia. Future work will concentrate on different extraction processes to increase Geodict coverage.

3.3 Featured methods

To exploit the data in Geodict, basic methods were implement for spatial analysis.

Data access. We choose to store Geodict in an Elasticsearch (ES) instance for two reasons. First, running queries on Elasticsearch is really efficient. Second, ES is associated with various data types (nested object, geo-shape) and their related queries.

We implement simple functions such as:

- *ExistsInGazeteer*(toponym)
- *getEntityWithWikidataID*(WikiID)
- *getEntitiesWithLabel*(label,[lang])

Recently, the scientific community has taken an interest in spatial reasoning using GeoSPARQL with triple store (Anelli et al., 2016). Hence, we plan to propose Geodict using Linked Data suggested formats (JSON-LD, N-TRIPLES, etc.).

Adjacency Test. In order to detect two adjacent spatial entities, we use three methods:

- Using the separating axis theorem (SAT) on administrative boundaries convex hulls.
- Use Wikidata P47 (share borders with) properties.
- Use P131 (located in administrative territorial entity) and P706 (located on terrain feature). Two objects are considered adjacent if they belong to a common value inside those properties. For example, the Statue of Liberty and the Governors Island are adjacent since both of their P131 value are equal to Manhattan.

Customization Depending on different applications, users may need complimentary data. Since Wikidata is a general knowledge base, users are allowed to indicate relevant and complimentary properties to extract. However, Wikidata stored information can be incomplete. Fortunately, Geodict is stored in JSON format and stored entries are linked to common database such as Wikipedia. Thus, complementary information from other data sources can be easily merged with Geodict.

The source code of Geodict is available at `https://bitbucket.org/thedark10rd/geodict`.

4 A case study: using Geodict for geoparsing

In the previous section, we introduced Geodict, a gazetteer with basic yet precise information and customizable. In the following section, we present a usecase for geoparsing using Geodict.

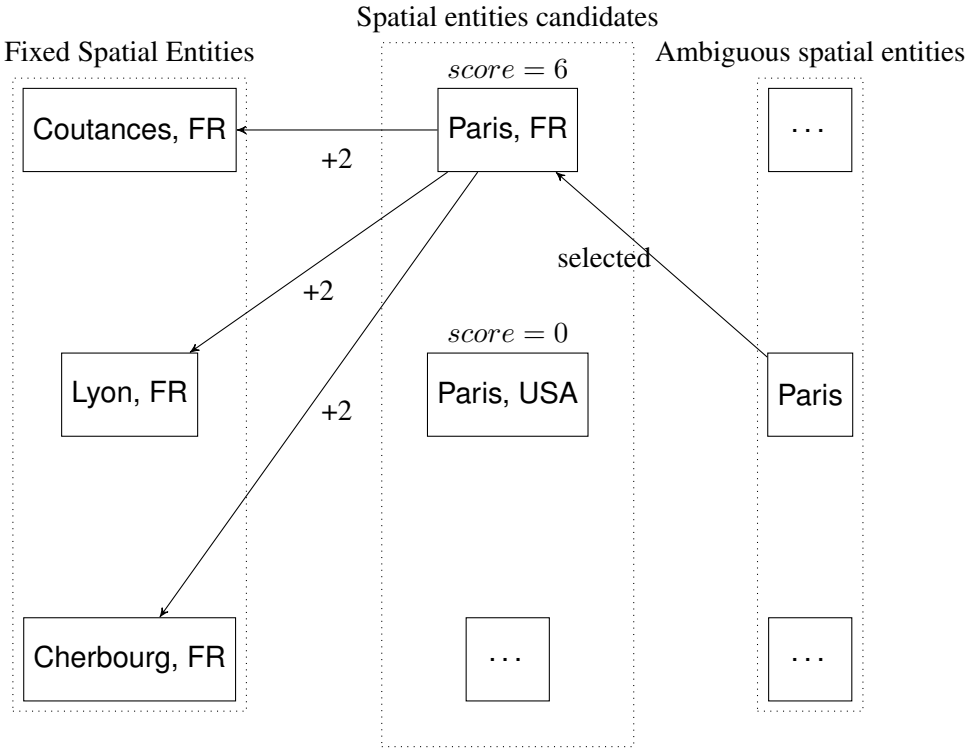

Figure 2: Example of toponym resolution with *Paris*

4.1 Toponym identification

To identify toponyms, we use a NER or Named Entity Recognizer. Various NERs have been proposed such as *StanfordNER* (Finkel et al., 2005), *NLTK*[11] and *Polyglot* (Al-Rfou et al., 2015). Since it supports 40 languages, we chose Polyglot. It increases our method coverage of available corpora.

Once the selected NER has returned detected named entities in a text, we only keep the locations. Then, each location is validated by checking their existence in the gazetteer.

4.2 Toponym Resolution

After identifying toponyms in text, we need to associate them with spatial entities. However, toponyms may be linked to different spatial entities *e.g. Paris, France* $\neq$ *Paris, Las Vegas*. To select which spatial entity is referred to a toponym in a text, we designed a disambiguation process divided in two parts.

First, we compute a score for each spatial entity candidate for a toponym. Second, we associate the toponym with the spatial entity having the maximum score. However, if the maximum score is not superior to a threshold (fixed to 4 in this use-case), we take the most frequently associated spatial entity for the corresponding toponym *e.g. Paris* $\rightarrow$ *Paris, France*. This process is illustrated in Figure 2.

Most Frequently Associated Spatial Entity If no spatial entity candidate is validated for a toponym, we associate the most frequently used one *e.g. Paris* $\rightarrow$ *Paris, France*. In order to do that, we need an "importance" value for each spatial entity. Every spatial entities stored in Geodict are indirectly linked to Wikipedia (using Wikidata). One way of computing popularity of webpage is to compute its page rank (Page et al., 1999). Hence, we decide to assign a page rank (PR) value computed on Wikipedia as proposed in (Thalhammer and Rettinger, 2016) to each spatial entity.

[11] `http://www.nltk.org/`

4.2.1 Score computation

In order to compute the score, we used 4 features associated with each spatial entity in Geodict:

- **P47** This property indicates which entities are adjacent to a corresponding entity. For example, Italy, Spain, U.K., Belgium, Gemany, etc. will be associated with France using P47. However, it does not give adjacency information between two adjacent entities at different (*e.g. country and city*)
- **P131** This property indicates in which administrative territorial entity is included a corresponding spatial entity *e.g. Paris is located in Ile de France*.
- **P706** This property indicates on which terrain feature is included a corresponding spatial entity *e.g. The Statue of Liberty is located on Liberty Island.*
- **Administrative boundaries** Polygon(s) describing administrative boundary(ies) of a spatial entity

For each spatial entity associated with a toponym, we search for existing relationships with the fixed spatial entities[12] in the text. Then each relationship is associated with a weight that denotes its importance. These relationships are using previously mentioned features and their weight are detailed in Table 4.

Relationship	**Weight**
Adjacency using Boundaries	2
P47 (Share Borders With)	3
Inclusion Score	See Paragraph 4.2.1

Table 4: Impact Weight of Properties on Disambiguations

Each weight is defined from different observations:

- In most cases, spatial relationships based on boundaries polygons are good indicators of the geographical context. However, in particular case, it can also bring confusion. For example, the boundaries between France and Surinam shown in Figure 3.
- As boundaries polygons, the property P47 contains relevant information to the geographical context. However, it contains simpler information (one scale adjacency) but less confusing. Hence, relationships found with P47 are more reliable than boundaries polygons.
- Finally, we considered spatial relationships found using P131 and P706 reliable since they contains precise information on the spatial entities in the spatial hierarchy ($Paris$ > *Ile de France* > $France > Europe > Earth$)

Inclusion Score To compute the inclusion score, we compare their inclusion chain made from P131 and P706. An inclusion chain is a list of spatial entities ordered by their inclusion. For example, the inclusion of Coutances and Caen in Figure 4 using P131.

Once inclusion chains using P131 and P706 of the two compared spatial entities are extracted, we compute the size of the intersection between them. For example, the size of the intersection between Caen and Coutances P131 inclusion chains is equal to 2. The inclusion score is defined as the sum of the intersections size of P131 and P706 inclusion chains. However, last spatial entities in inclusion chain are most likely to be equal. Therefore, we are summing the Fibonacci value of each intersection length value. It allows us to lower the impact of low score (resp. increase higher score).

[12] Spatial entities which does not share their toponym

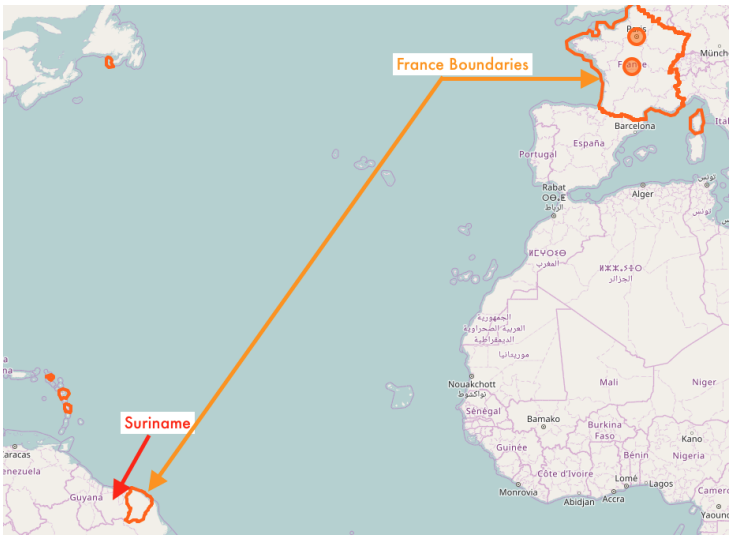

Figure 3: Adjacency Confusion

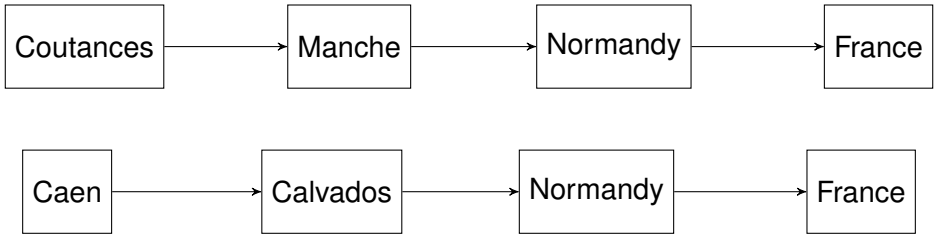

Figure 4: Inclusion chain of Coutances and Caen using P131

Ultimately, we define the inclusion score in Equation 1.

$$\begin{aligned} inclusion_{score}(se_1, se_2) = & fib(|inc(se_1, P_{706}) \cap inc(se_2, P_{706})|) \\ & + fib(|inc(se_1, P_{131}) \cap inc(se_2, P_{131})|) \end{aligned} \tag{1}$$

with:

- $fib(x)$, Fibonacci value of x
- $inc(se_i, P_y)$, inclusion chain of the spatial entity se_i using the property P_y

4.3 Disambiguation process evaluation

In this paper, we choose to focus on the toponym resolution process. In particular, for each document processed, we run our process on their list of annotated toponyms. Thus, the accuracy measure is the most adapted (Equation 2).

$$Accuracy(TP, SE) = \frac{\sum_{t \in TP, s \in SE} \delta(t, s)}{|TP|} \tag{2}$$

where:

- TP list of toponyms
- SE list of spatial entities associated to each toponym in TP
- $\delta(t, s)$ equal to 1 if the toponym t was correctly associated with s

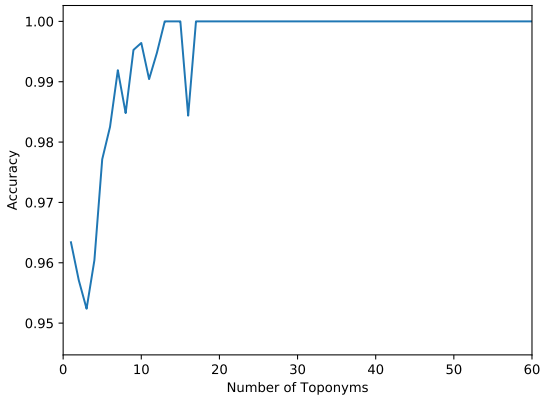

Figure 5: Accuracy evolution over different size of the list of toponyms to disambiguate

In order to evaluate our toponym resolution method, we built a corpus composed of 10000 random documents extracted from Wikipedia.

A Wikipedia article is written using a markup language composed of different tags. Among these tags, anchors allow to link different Wikipedia pages between them. Therefore, if a spatial entity exists in a Wikipedia article, it is referenced using an anchor. However, non-spatial entities can be referenced using these anchors. Thus, we use DBpedia to filter other named entities. In a nutshell, DBpedia is a knowledge base constructed on Wikipedia data including it's URI *e.g.* `http://fr.dbpedia.org/page/Louis_XIV` ⇔ `https://fr.wikipedia.org/wiki/Louis_XIV`. In DBpedia, each entity is associated with a main concept (Location, Person, etc.). Hence, it allows us to filter non spatial entity referenced in anchors for a Wikipedia article.

We obtain good performance with an average accuracy of 95.74% over the 10000 documents. In addition, we highlight our system efficiency over different size of toponym sets to disambiguate, as illustrated in Figure 5.

To strengthen the evaluation, our method could be compared to state-of-the-art methods on recognized corpora such as **TR-CoNLL** introduced in (Leidner, 2007), **LGL** in (Lieberman et al., 2010) or more recently **WarOfTheRebellion** by (DeLozier et al., 2016).

5 Conclusion

In this paper, we propose an integrated gazetteer Geodict that contains basic yet precise geographical information about places names. We conceived it to be multi-purpose by allowing users to customize its creation and link each spatial entity to commonly used datasets. Geodict was used for a geoparsing task and more precisely for toponym resolution. Based on a large corpus, we obtain good results and show the suitability of Geodict.

However, Geodict coverage must be improved by designing new extraction predicates over the different used sources. As for geoparsing, we consider improving our evaluation relevancy by comparing our method to state-of-art methods on referenced corpora.

References

Al-Rfou, R., V. Kulkarni, B. Perozzi, and S. Skiena (2015, April). Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015*.

Anelli, V. W., A. Cal'ı, T. Di Noia, M. Palmonari, and A. Ragone (2016). Exposing open street map in the linked data cloud. In *Trends in Applied Knowledge-Based Systems and Data Science - 29th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2016, Morioka, Japan, August 2-4, 2016, Proceedings*, pp. 344–355.

Barboza, P. (2014, December). *Evaluation of epidemiological intelligence systems applied to the early detection of infectious diseases worldwide.* Theses, Université Pierre et Marie Curie - Paris VI.

Berners-Lee, T., J. Hendler, O. Lassila, et al. (2001). The semantic web. *Scientific american 284*(5), 28–37.

Clough, P., M. Sanderson, and H. Joho (2004). Extraction of semantic annotations from textual web pages. *Deliverable D15 6201*.

DeLozier, G., J. Baldridge, and L. London (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *AAAI*, pp. 2382–2388.

DeLozier, G., B. Wing, J. Baldridge, and S. Nesbit (2016). Creating a novel geolocation corpus from historical texts. *LAW X*, 188.

Finkel, J. R., T. Grenager, and C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370. Association for Computational Linguistics.

Grossman, D. A. and O. Frieder (2004). *Information Retrieval - Algorithms and Heuristics, Second Edition*, Volume 15 of *The Kluwer International Series on Information Retrieval*. Kluwer.

Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pp. 280–290. Springer.

Leidner, J. L. (2007). Toponym resolution in text: annotation, evaluation and applications of spatial grounding. In *ACM SIGIR Forum*, Volume 41, pp. 124–126. ACM.

Li, C. and A. Sun (2014). Fine-grained location extraction from tweets with temporal awareness. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pp. 43–52.

Li, H., R. K. Srihari, C. Niu, and W. Li (2003). InfoXtract location normalization. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - 1*, 39–44.

Lieberman, M. D., H. Samet, and J. Sankaranarayanan (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pp. 201–212. IEEE.

Overell, S. and S. Rger (2008). Using cooccurrence models for placename disambiguation. *International Journal of Geographical Information Science 22*(3), 265–287.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Rauch, E., M. Bukatin, and K. Baker (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pp. 50–54. Association for Computational Linguistics.

Stadler, C., J. Lehmann, K. Hffner, and S. Auer (2012). Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal 3*(4), 333–354.

Thalhammer, A. and A. Rettinger (2016, October). PageRank on Wikipedia: Towards General Importance Scores for Entities. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers*, pp. 227–240. Cham: Springer International Publishing.

Vrandečić, D. and M. Krötzsch (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM 57*(10), 78–85.

Zenasni, S., E. Kergosien, M. Roche, and M. Teisseire (2016). Extracting new spatial entities and relations from short messages. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems, MEDES 2016, Biarritz, France, November 1-4, 2016*, pp. 189–196.