Kalpa Publications in Computing

Volume 13, 2019, Pages 39–46

Proceedings of the 1st International Conference on Geospatial Information Sciences

Georeference Assignment of Locations based on Context

Shanel Reyes-Palacios[1], Edwin Aldana-Bobadilla[2], Ivan Lopez-Arevalo[1], and Alejandro Molina-Villegas[3]

[1] Centro de Investigación y de Estudios Avanzados del IPN, Unidad Tamaulipas, México. contact: sreyes@tamps.cinvestav.mx

[2] CONACyT – Centro de Investigación y de Estudios Avanzados del IPN, Unidad Tamaulipas, México.

[3] CONACyT – Centro de Investigación en Ciencias de Información Geoespacial, Unidad Yucatán, México.

Abstract

Modern Text Mining techniques seek for extract information in useful formats such as georeferences in digital documents. Automatic recognition of location names in texts is usually solved through Named Entity Recognition (NER) systems. Most current NER are based on Machine Learning and have very high accuracy in detection of location entities in digital documents, especially if the texts are in English due to the lack of available annotated corpora in other languages. However, recent studies are dealing with the challenge of taking the output labels of a NER system and then gather, from a gazetteer, their exact unambiguous geographical coordinates. This is challenging mainly because toponyms use to be very ambiguous, so research in disambiguation methods is relevant. In this paper we describe some of the main ideas towards a method to associate locations with geographical data removing possible confusion between entities with the same name. So far, we have already accomplished Geographic NER and coordinates retrieval but the main research is still in course. We largely discuss about the state of the art around Geoparsing; we explain how our Geographic Entity Recognition module works and finally we describe the research proposal focusing in ambiguity detection.

1 Introduction

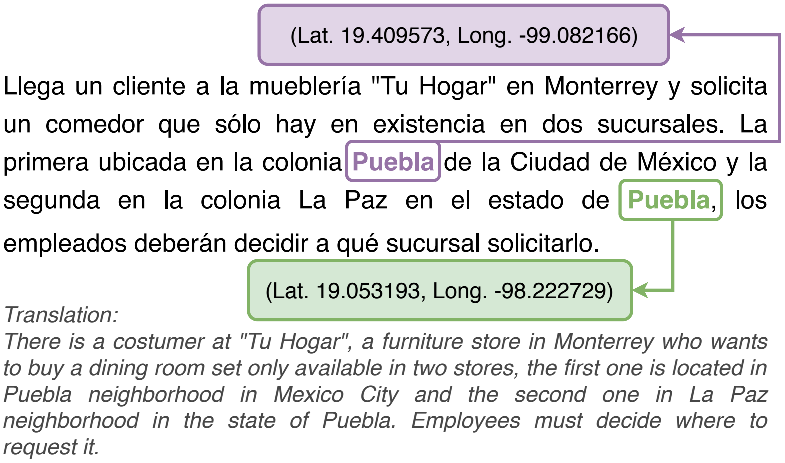
Nowadays, with the large number of digital texts, users request supplementary specialized information. More and more users ask for information related to geographical locations, to which a text refers. For example, what restaurants are closer to a public place described in a webpage? what hotels are near to a music festival announced on Facebook advertising? This paper proposes to enrich the information contained in digital texts adding a layer of georeferencing, for this purpose different areas were combined, such as Computational Linguistics, Text Mining and Geomatics.

Several methods from Information Extraction have been proposed in order to explore and analyze digital texts automatically; these methods used to simplify the information contained in a corpus through several transformations making the analysis simpler [1]. The Information Extraction process is based on Natural Language Processing (NLP) techniques. Within the

O. S. Siordia, J.L. Silván Cárdenas, A. Molina-Villegas, G. Hernandez, P. Lopez-Ramirez, R. Tapia-McClung, K. González Zuccolotto and M. Chirinos Colunga (eds.), iGISc 2019 (Kalpa Publications in Computing, vol. 13), pp. 39–46

NLP variety of techniques, Named Entity Recognition (NER) oversees charge of recognize from text such a Named Entities. A named entity is an element in the text composed of one or more words that has a meaning accepted by a community, it automatically identifies names of people, locations, organizations and other entities of interest [3]. NER has been successfully applied in several domains [12], it's main goal is simply to describe. NER systems have to recognize if a word within a text is an important entity for a specific domain or not. However, some related issues in the NER task are very challenging. One of these is the coarse-grained feature for the identification of Named Entities of location. That means that NER can not handle ambiguity of locations with the same label in a text.

For example, Figure 1 shows an extract of a document that refers to a furniture store with different store branches in Mexico, where two entities with the same name are mentioned (Puebla). In the first case (first mention related to the upper box in Figure 1), the text refers to the branch located in the Puebla neighborhood, in Mexico City. In the second case (the second mention related to the down box), the text refers to the state of Puebla in Mexico. Some NER tools like Stanford NER [2], spaCy (www.spacy.io) and Apache OpenNLP (https://opennlp.apache.org) will recognize these two entities as the same location name, so obtaining their georeferences would give us exactly the same coordinates for both. This is a problem known as named entity ambiguity.


(Lat. 19.409573, Long. -99.082166)
(Lat. 19.053193, Long. -98.222729)

Llega un cliente a la mueblería "Tu Hogar" en Monterrey y solicita un comedor que sólo hay en existencia en dos sucursales. La primera ubicada en la colonia **Puebla** de la Ciudad de México y la segunda en la colonia La Paz en el estado de **Puebla**, los empleados deberán decidir a qué sucursal solicitarlo.

Translation:
There is a costumer at "Tu Hogar", a furniture store in Monterrey who wants to buy a dining room set only available in two stores, the first one is located in Puebla neighborhood in Mexico City and the second one in La Paz neighborhood in the state of Puebla. Employees must decide where to request it.

Figure 1: Example of ambiguity in the recognition of two Named Entities in a text in Spanish.

To deal with the named entity ambiguity problem, some approaches have been proposed. In [8] and [15], authors propose to assign a location to the whole document (coarse grain allocation). This is done by considering the frequency of occurrences of locations mentioned in the text to assign it to a general location. In [18] authors assign one or more polygons that represent geographical locations, without obtaining the exact coordinates, but an approximate. Another important approaches, in English are [14], [17], [13] and [6].

It should be noted that in order to link the localities to geographical areas, the *georeferencing* process is always applied, which consists of query a geographic resource to obtain the geographical coordinates. A resource that allows obtaining the needed information to georeference is know as a digital *gazetteer*, which is a dictionary of geographical names that set names and coordinates [5].

2 Related Work

Rupp *et al.* [14] focus their work in the historical context. Each text was translated and transcribed from their original book to avoid inaccuracies that may occur when using OCR (Optical Character Recognition) and then a translator. Authors use a system called VARD, for spelling correction. They make use of a historical gazetteer from Cumbria in addition to a list of specific names from the Cumbria region and an ordinance survey. One of the limitations is that being historical texts, the names of current places may not match those they had in the past centuries, so the accuracy varies.

A different approach proposed by Tobin *et al.* [17] there were three historical collections digitized beforehand, so the problem of spelling variation is assumed to be solved. They used Information Extraction techniques to identify place names of places in the corpus. They rely on gazetteers to compare the results obtained with human annotations of the three collections. Similarly, the process is applied in the SpatialML corpus, which is a geo-annotated corpus of newspaper [7]. Two main parts are considered in this method, a geotagger and a georesolver. The first one processes input text and identifies the strings within it that denote place names. The second one takes the set of names of places recognized as input, searches them in one of the different gazetteers and determines for each name of places which of the possible referents is correct. In the case of ambiguity, classify the candidate entries in order of probability according to the context, if there are repeated places they are based on population or type of place to make a decision. A disadvantage is that it does not find place names if they are misspelling.

Another important work was proposed by Martins and Silva [8], they present a PageRank like application to assign documents with geographical scope. PageRank is a method based on graph centrality to rate web pages measuring human interest and the attention they are given [10]. This weighting is considered according to the frequency of occurrence of locations in the text. Another important feature to consider is the use of geographical references extracted from the text and a technique based on ontologies. This technique consists of two geographical ontologies that provide both vocabulary and relationships between geographical concepts. The first one is based on global multilingual and the second based only in the region of Portugal. One of the limitations is that the original PageRank algorithm gives the same weight to all edges (hyperlinks), which causes nodes with more links (links to other sites) that tend to obtain higher ranges, whether they are important for the problem. In addition, it does not consider the stage of Information Extraction, since it is assumed that the data must be previously preprocessed.

Likewise, Silva *et al.* [15] assigned geographic scopes to documents using a graphic classification algorithm like PageRank. It is focused on feature extraction, recognize and disambiguate geographical references. The method makes extensive use of an ontology of geographical concepts and includes an architecture system to extract geographic information from large collections of web documents. It should be noted that a geographical scope is specified as a relationship between an entity in the web domain (an HTML page or a website) and an entity in the geographic domain (such as an administrative location or region). The geographic scope of a web entity has the same footprint as the associated geographic entity. The scope assigned to a document is granted due to the frequency of occurrence of a term and also considering the similarity to other documents.

Radke *et al.*[13] proposed an algorithm for the geographical labeling of web documents considering all names of places together without disambiguate them individually. This approach requires a gazetteer and an NER tool for processing. This proposal is focused on a plain text. As a result, it reports latitude and longitude of the mentioned places. It can be applied to large sets of documents.

Woodruff *et al.* [18] developed an algorithm that automatically extracts the coordinates from the text in English, and the words and phrases that contain names of geographical places or their characteristics. They also proposed a prototype system called Gipsy (Georeferenced Information Processing System), which assigns to each text document indexes that consisting of one or more polygons that represent geographical locations, only in English.

Nes *et al.* [11] presented a system called GeLo, which extracts addresses and geographic coordinates of commercial companies, institutions and other organizations from their web domains. This extraction process is based on NLP techniques, specifically Part-Of-Speech-Tagging, pattern recognition and annotations. The tests carried out focused on web domains of organizations located in the region of Tuscany, in Italy. The platform was developed in Italian and English, and authors proposed an independent language option. The architecture is based on two modules, the first one is a tracking tool for indexing documents, and the second one is a linguistic analyzer that takes as input the documents and pages retrieved in the URLs of the web obtained by the previous module.

Inkpen *et al.* [6] developed an algorithm that extracts expressions composed of one or more words for each place name. They use a Conditional Random Fields classifier, which is based on an unguided graphical model that is used for unstructured predictions [4]. They focused on tweet location entities by defining disambiguation rules based on heuristics. The corpus contains tweets only in English basically in the states and provinces of the United States and Canada.

Although some works have been developed in multiple languages (Martins and Silva [8] and Silva *et al.* [15]), only one of them focuses on Mexican Spanish [9]. This is an important area of opportunity, since each language has specific language patterns that provide extra information in order to identify the name Georeferencing of entities. Notably, the work of Silva *et al.* [15] was tested for Spanish, specifically to locations from Spain. Martins and Silva's [8] work does not specify what languages it refers to, it is only mentioned that it is possible to use it in multiple languages. This is because it does not consider the context of the text, and just extracts each location separately for georeference it. The 2002 CoNLL edition [16] is also of interest, for the first time Spanish was considered as one of the languages to be processed. However, in CoNLL 2002, the data was collected by the Polytechnic University of Catalonia and the Autonomous University of Barcelona, and the annotation focused on documents from Spain, leaving aside any other variant, including the Mexican one. All the researches highlights the need of similar resources but for locations outside from Spain. In this sense, the main challenge is to develop and obtain a labeled quality corpus for such locations.

As well as we know, only two works take into account disambiguation: Silva *et al.* [15] and Inkpen *et al.* [6]. The first one recognizes if ambiguity is present, so it takes into account the elements related to involved entities, however it does not take into account variations in the names of places. The second one has a series of rules that allow disambiguation, but it focuses on states and provinces from United States and Canada, except smaller entities, such as neighborhoods or towns, or even variations in the names of places. The locations that can be found in a range of countries, states and municipalities, or even for more specific places, such as hospitals, restaurants or schools. As a reference, there is a description of the systems exposed in Table 1.

3 Research proposal

In general terms, Geoparsing involves a set of processes to transform an input text containing entities into a locations oriented data structure in which, such entities are georeferenced by

Method	**Input**	**Language**	**Tools**	**Disambiguation**
Molina-Villegas *et al.* 2019	Plain text	Spanish	Gazetteer	Not
Radke *et al.* 2018	URL	English	Gazetteer	Not
Inkpen *et al.* 2017	Tweets	English	Gazetteer	Yes
Nes *et al.* 2014	HTML	English and Italian	Ontology	Not
Rupp *et al.* 2013	HTML, plain text	English	Gazetteer	Not
Tobin *et al.* 2010	XML, plain text	English	Gazetteer	Not
Silva *et al.* 2006	Multiple formats	Multiple languages	Ontology	Yes
Martins and Silva 2005	URL	Multiple languages	Ontology	Yes
Woodruff *et al.* 1996	Plain text	English	Ontology	Not

Table 1: A categorization of methods of Geographic NER according to input type, language, external tools and disambiguation methods.

their most likely coordinates. These processes are illustrated in Figure 2 and described in [9].

3.1 Geographic Entity Recognition (GER)

The first important module is Geographic Entity Recognition (GER) which allows to recognize geographical entities within a text. Specifically, this module finds and labels all those words associated to locations. The system that we develop, GER, is based on a neural network model pre-trained with a corpus that includes Mexican locations called Corpus of Georeferenced Entities of Mexico (CEGEOMEX). The corpus has a total of 61,946 words distributed in 1,233 documents from important digital media in Mexico, it also contains a set of 5,870 geographical named entities have been manually annotated. This corpus can be accessed at `http://geoparsing.geoint.mx/mx/info/`.

In the short term, GeoparseMX researchers consider using this platform in several Big Data application scenarios with social impact. For example, to gather and standardize information from various sources and then process the data to relate it to other databases and maps. GeoparseMX possible applications are diverse: search for missing persons, clashes between armed groups, misogyny, biodiversity, culture, tourism for instance. GeoparseMX can be used as a web service at `http://geoparsing.geoint.mx/ws/` via HTTP POST request by sending the text in the body of the request as the *text* parameter.

3.2 Georeference Assignment

Depending on a set of location names extracted from a text document, the next challenge is to assign the most suitable coordinates to them. The most direct assignation is to associate a location name with the first option obtained from querying a digital gazeetter. However, the assignment can involve ambiguity. An example of these situations is illustrated in Figure

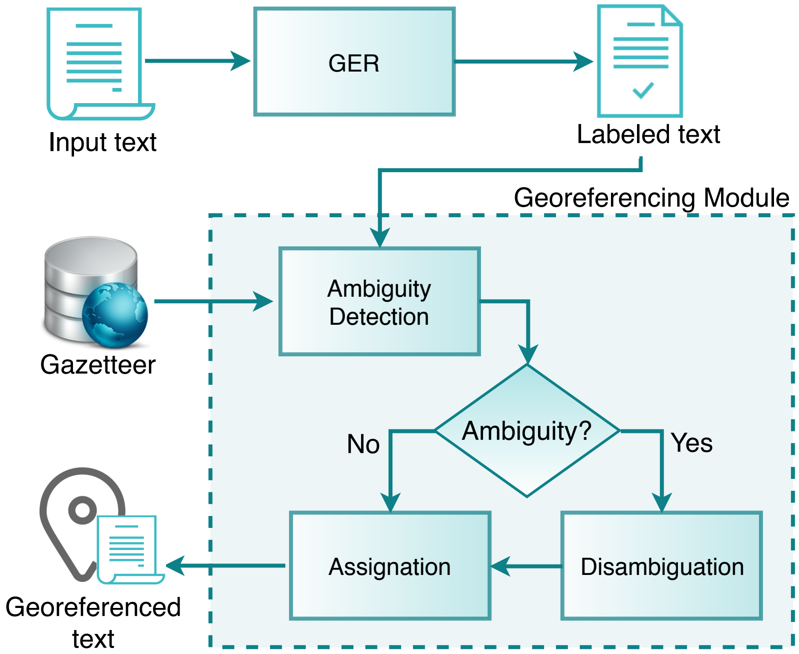


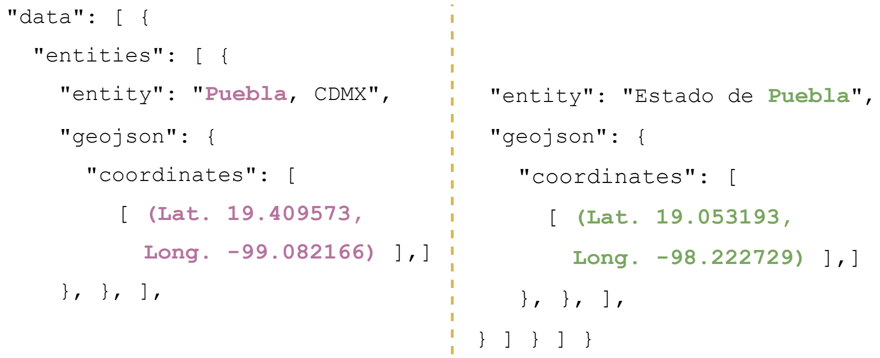
Figure 2: Research proposal

1 where the entity *Puebla* could refer either to a neighborhood located in Mexico City or a state in Mexico. This and other situations will be analyzed and solved through a process called *ambiguity detection*, which is the main subject of a thesis work in progress. As a main idea for disambiguation, we consider as the main input and starting point the results from a generic gazetteer. From this point, when a location entity has two or more possible assignments, we should determine which of the results is the most likely georeference. There are cases in which entities just do not appear in the gazetteer, we also consider this as an ambiguity. Under the presence of ambiguity, a sub-process of *disambiguation* will be applied. This process involves an inference engine that allows the evaluation of contextual ambiguities and solve them based on a set of previously defined rules. Finally, the geographical coordinates will be determined via *assignation*. In future work, we will prove that, our rule based proposal enhanced by cues in the text, allows us to make suitable choices about the coordinates to be assigned to an named entity.

3.3 Gazetteer

A Gazetteer is a geographical dictionary that contains structured information about places that have a particular geographic location. For this project, the gazetteer will consist of JSON files that will be generated manually. It will be composed of places in Mexico in Spanish, and will understand news from the main media in Mexico. In general, it will be tagged and reviewed manually, in it will be update or added new information once is generated.

The content that JSON generated by the system is shown in Figure 3, taking as example the one shown in Figure 1. This file is part of the gazetteer, that also shows the entity and the coordinates. In this case there is ambiguity, Puebla represents two different entities and the gazetteer will have information of each one, since the disambiguation module will know which one to refer to.



```
"data": [ {
  "entities": [ {
    "entity": "Puebla, CDMX",
    "geojson": {
      "coordinates": [
        [ (Lat. 19.409573,
          Long. -99.082166) ],]
    }, }, ],
```

```
"entity": "Estado de Puebla",
"geojson": {
  "coordinates": [
    [ (Lat. 19.053193,
      Long. -98.222729) ],]
  }, }, ],
} ] } ] }
```

Figure 3: JSON file generated for two entities with the same name ”Puebla”

4 Conclusions

We have presented the main ideas that are in progress in order to disambiguate geographic named entities. This is challenging mainly because toponyms are used to be ambiguous. We have described some of the main modules to develop a method to associate locations with geographical data removing ambiguity of entities with the same name. So far, we have already accomplished Geographic NER and georeferencing, although the main research around disambiguation is still in course. We discussed about the state of the art around Geoparsing and we have explained how our Geographic Entity Recognition module works. As future work, a method of disambiguation of locations based on dynamic rules will be proposed. At this point, the full process is in experimentation stage but some modules are in process by using a disambiguation manual.

References

[1] Jim Cowie and Wendy Lehnert. Information extraction. *Commun. Association for Computing Machinery*, 39(1):80 – 91, January 1996.

[2] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363 – 370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Venkat Gudivada. *Natural Language Core Tasks and Applications*, pages 403 – 428. North-Holland, 01 2018.

[4] Rahul Gupta. *Conditional Random Fields*, pages 146 – 146. Springer US, Boston, MA, 2014.

[5] Linda L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In Jose Borbinha and Thomas Baker, editors, *Research and Advanced Technology for Digital Libraries*, pages 280 – 290, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[6] Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237 – 253, Oct 2017.

[7] Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. Spatialml: Annotation scheme, corpora, and tools. In *LREC 2008*, 2008.

[8] Bruno Martins and Mario Silva. A graph-ranking algorithm for geo-referencing documents. In *Fifth IEEE International Conference on Data Mining*, pages 741 – 744, Nov 2005.

[9] Alejandro Molina-Villegas, Oscar S Siordia, Edwin Aldana-Bobadilla, César Aguilar Aguilar, and Olga Acosta. Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad. *Procesamiento del Lenguaje Natural*, 63:143 – 146, 2019.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[11] Gianni Pantaleo and Paolo Nesi. Ge(o)lo(cator): Geographic information extraction from unstructured text data and web documents. In *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 60 – 65, Nov 2014.

[12] Thierry Poibeau and Leila Kosseim. Proper name extraction from non-journalistic texts. *Language and Computers*, 37(1):144 – 157, 01 2001.

[13] Mansi Radke, Nitin Gautam, Akhil Tambi, Umesh Deshpande, and Zareen Syed. Geotagging text data on the web a geometrical approach. *IEEE Access*, 06:30086 – 30099, 06 2018.

[14] CJ Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, and Patricia Murrieta-Flores. Customising geoparsing and georeferencing for historical texts. In *Proceedings of the IEEE International Conference on Big Data, Big Data*, pages 59 – 62, 10 2013.

[15] Mario J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Cardoso Nuno. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378 – 399, 2006.

[16] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT - NAACL 2003 - Volume 4*, CONLL 03, pages 142 – 147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[17] Richard Tobin, Claire Grover, Kate Byrne, James Reid, and Jo Walsh. Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR 10, pages 1 – 8, New York, NY, USA, 2010. ACM.

[18] Allison Woodruff and Christian Plaunt. Gipsy: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45, 01 1996.