Geographic Information Extraction using Natural Language Processing in Wikipedia Texts

Edson B. de Lima[1], **Clodoveu Augusto Davis Jr.**[1]

[1]Departamento de Ciência da Computação – Universidade Federal do Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

edson@dcc.ufmg.br, clodoveu@dcc.ufmg.br

Abstract. *Geographic information extracted from texts is a valuable source of location data about documents, which can be used to improve information retrieval and document indexing. Linked Data and digital gazetteers provide a large amount of data that can support the recognition of places mentioned in text. Natural Language Processing techniques, which have evolved significantly over the last years, offer tools and resources to perform named entity recognition (NER), more specifically directed towards identifying place names and relationships between places and other entities. In this work, we demonstrate the use of NER from texts, as a way to detect relationships between places that can be used to enrich an ontological gazetteer. We use a collection of Wikipedia articles as a test dataset to demonstrate the validity of this idea. Results indicate that a significant volume of place/non-place and place-place relationships can be detected using the proposed techniques.*

1. Introduction

Currently, a relevant amount of information can be found in text or documents that are free of structure and widely available online, such as Wikipedia[1] articles and other forums or social networks. We are particularly interested in geographic information obtained from such textual sources, i.e., references to places embedded in natural language text, that can be used to characterize or to classify the documents. If the association between a document and a set of places can be correctly and reliably determined, spatial indexes on documents could be created, thereby enabling users to search by geographic location, keywords, or a combination of both. Furthermore, the co-occurrence of places and other entities in a document indicates relationships among them, which can be instrumental for ontological gazetteers and help in geographic information retrieval tasks [Moura and Davis Jr, 2013][Moura et al., 2017]. For example, the LinkedOntoGazzetter[2] (LoG) records geographic and semantic relationships between places and their various names, and between places and non-place entities, such as people and businesses.

Natural Language Processing (NLP) offers key resources for analyzing a document and extracting patterns that help identifying entities, including places, and establishing the relationships among entities as expressed in text. NLP techniques extract a potentially large set of features from sentences in the text. Selecting relevant features is difficult, since it involves a sequence of empiric tasks, based on linguistic intuition. Selected features are then used to feed a classifier, such as Support Vector Machine (SVM) [Hearst

[1]https://www.wikipedia.org/
[2]http://aqui.io/log/

et al., 1998], that determines a label for each word [Collobert et al., 2011]. Among these labels are indicators of entity names, and the type of entity is inferred by the structure of the sentence, using elements such as prepositions and the presence of other linguistic indicators of the nature of the entity.

The objective of this paper is to analyze information from Wikipedia documents extracted by NLP tasks and provide geographic characteristics obtained from linked data sources that relate to other challenges, such as place name disambiguation and geographic context resolution. The paper is organized as follows. Section 2 discusses related work and NLP-based feature selection from text. Sections 3 and 4 introduce the proposed approach and experimental results. Section 5 presents conclusions and future work.

2. Related work: geographic feature selection from text

Considering the usual contents of text documents, many location features are indicated by references to named entities, and by the relationships among them. Some references can be indirect, i.e., can be inferred from the contents of the text or from the generalization of the other references. For instance, a sentence such as "the earthquake struck Mexico City and regions of the Puebla and Morelos states" contains direct references to three entities (Mexico City and the two states) and an indirect reference to a fourth (the country of Mexico, which contains the three others). A Wikipedia page that refers to the event [3] contains many other geographic elements, such as the coordinates of the epicenter, the names of the tectonic plates involved, and references to several places affected by the disaster. It also contains names of related entities, such as the Mexican president or the local football championship, which reinforce the association of the text with the places. Automatically identifying such references to places is a complex task, for which many solutions have been proposed. Monteiro et al. [2016] provide a survey of current techniques for the recognition of the geographic context of documents.

Candidate names can be tested to verify if they correspond to place names. Gazetteers are ideally suited to this task, since they provide an efficient way to check if a candidate string corresponds or not to a known place name. In this work, we propose a approach that uses two NLP techniques to collect location attributes based on relationships and sentence structure features. The first technique is called Named Entity Recognition (NER) [Finkel et al., 2005]. NER receives a text as input and breaks it into sentences, using *sentence tokenization*. Each sentence goes through a similar procedure, this time to separate tokens using words and word groups, a process called *word tokenization*. A set of other methods analyzes the tokens to find patterns that confirm the characteristics detected for each word. The *Part of Speech (PoS)* procedure [Toutanova et al., 2003], for example, labels tokens based on the semantics of the words. A group of PoS-labeled words forms a *chunk*, that is used to establish a pattern, such as the indication of a named entity. Then, NER uses these chunks to identify entity types. The NER model used in this work identifies only three types of named entities: *person*, *organization* and *location*. Figure 1 illustrates NER applied to a sentence to obtain location entities.

The second NLP technique addresses relationship extraction, a task that is performed after the entity recognition subroutine. Each relationship between a location and another named entity is extracted from the text in the form of a triple, containing a subject

[3] https://en.wikipedia.org/wiki/2017_Central_Mexico_Earthquake

and an object that correspond to entity names, and a predicate that refers to the type of relationship inferred from the sentence. Figure 2 exhibits all existing relationships in a sentence given two named entities that were identified by the NER process. However, in this case the named entity types are not considered. Strings `New York City` and `United States` represent entities and each relationship is deduced. In the example, an entity type has been found in association to a relationship, but only the relationship's name is extracted from the sentence [Manning et al., 2014].

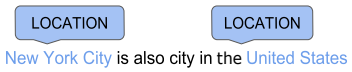

Figure 1. NER applied to a text sentence. Adapted from [Manning et al., 2014]

A similar procedure is implemented by Geo-NER, a system for detecting and recognizing geographic named entities [Perea-Ortega et al., 2009]. Geo-NER is based on a generic entity tagger, expanded with geographic resources generated from Wikipedia. Geo-NER uses GeoNames[4] as a gazetteer data source, and proposes some heuristics. It lacks, however, the possibility of considering geographic data from other sources to aid in the recognition of places from text. LoG, on the other hand, integrates data from GeoNames, FreeBase, DBPedia and OpenStreetMap that have been encoded as linked data. A richer source of place names such as LoG, which includes place/place and place/non-place relationships, should improve the recognition of geographic entities from text.

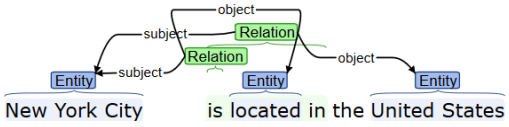

Figure 2. Relation Extraction using the CoreNLP toolkit [Manning et al., 2014]

3. Proposed approach

In this paper, we propose applying NER and relationship extraction to identify the type of place according to GeoNames feature classes and feature codes, in order to assess the possibility of obtaining rich sets of triples with which to enhance LoG. For that purpose, we collected three document classes, composed by Wikipedia articles that are listed in three different categories. The first document class (DOC1) contains 399 articles related to the most populous cities and states in the USA[5]. The second document class (DOC2) is composed of 110 articles that describe types of social networking tools[6]. And the third document class (DOC3) contains articles about online chat tools[7]. It is important to keep in mind that the algorithm does not need to obtain all named entities. The aim of this experiment is to verify how many relationships involving place entities can be extracted from document classes, thereby indicating a future strategy for using text sources to enrich LoG. We use three document classes with goal to analyze those that contains more

[4]http://www.geonames.org/
[5]https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
[6]https://en.wikipedia.org/wiki/List_of_social_networking_websites
[7]https://en.wikipedia.org/wiki/List_of_chat_websites

geographic enrichment through entities that can reveal features for that purpose. Furthermore, recognized places are classified using GeoNames feature classes and feature codes, in order to assess the most frequent types of places that appear in the relationships.

Named entities and relationships are obtained using Open Information Extraction (OpenIE) and NER annotations. OpenIE is a part of Stanford's CoreNLP [Manning et al., 2014] toolkit, which provides a set of NLP functions for text processing and parallel pipeline annotations. According to the authors, OpenIE is useful for relationship extraction tasks when there is limited or no training data, and when speed is essential. Since we used no training data for Wikipedia articles, and observing that the articles are relatively large, NER and OpenIE are suitable to recognize place entities that participate in a relationship. Even all OpenIE toolkit, extract relationship with named entities is not possible with CoreNLP modules. Then, we purpose a algorithm that combine the two tasks and extract relationship with the named entities related to location names.

After relationship extraction with place entities, in the next steps we analyze location features using feature codes and feature classes supplied by LoG's API [Moura et al., 2017], in turn obtained from GeoNames. GeoNames categorizes geographic features into nine classes, which are subdivided into more than 645 subcategories, identified by feature codes [Perea-Ortega et al., 2009]. Tables 1 and 2 show some of the main GeoNames feature classes and codes that were used in this work to classify each group of documents. For the experiments, we chose only four feature classes and their corresponding feature codes, because most of the place names are classified according to these types. Then, if a place name refers to feature class A, its feature code must be some of the ADM codes. On the other hand, the Populated Place subclasses (PPL and PPLA) are related to feature class P, while feature classes L and H refer to place names associated to Area features (parks, reserves, economic regions, etc.) and Hydrographic features (river, lake, sea, etc.) respectively.

Table 1. Feature Class

API	Feature Class	Description
1	A	Administrative Boundary
2	P	Populated Place
3	L	Area
4	H	Hydrographic

Table 2. Feature Code

API	Feature Code	Description
1	ADM1	First Adm. Division
2	ADM2	Second Adm. Division
3	PPL	Populated Place Code
4	PPLA	Seat of First Adm. Div

4. Experimental Results

An example of place name recognition and relationship extraction is presented next. From the sentences *"Chicago is located in northeastern Illinois."* and *"Chicago is the home of former president Barack Obama."*, the following triples indicating places and relationships were extracted:

⟨ `Chicago:LOCATION, related_to, Illinois:LOCATION` ⟩

⟨ `Chicago:LOCATION, related_to, Barack Obama:PERSON` ⟩

The locations recognized using NER in the sentences are then checked against LoG by the algorithm. If multiple places exist under the same name, a disambiguation step should follow. For disambiguation, the set of triples obtained in the document can be used to decide on a single place to correspond to the place names that have been identified. Then, the GeoNames Feature Class and Feature Code can be determined. So far, however,

we have not implemented this step. Furthermore, the type of relationship can be classified using other NLP techniques, such as Stanford's Relation Extractor.

Considering the three document classes, results indicate that DOC1 is the group of documents from which more relationships involving locations could be extracted, proportionally to the number of documents (over 28 triples per document) (Table 3). DOC2 and DOC3 achieved a lower proportion of relationships involving location entities (6 and 0.6 triples per document, respectively). However, the triples found in the process involve place names, identified as such using NER. These place names can be ambiguous, i.e., the may correspond to more than one actual place. LoG has a function by which all places that correspond to a given name are retrieved. We call such places *candidate places*, pending disambiguation. Table 4 shows the number of places involved for each dataset. Notice that the number of candidate places of the P feature class is much larger than the number in other classes. Similarly, Table 5 exhibits geographic characteristics subclasses that represent places and also the number of P feature code is larger than the value in the last classes. Finally, Table 6 compares the first two document classes considering the same number of location triples, and shows that DOC1 contains more geographic feature classes per candidate place than DOC2. Thus, there is an important disambiguation challenge in the actual integration of the relationships to LoG.

Table 3. Location triples

ID	Context	Docs	Triples	LOC Triples	Triples/Doc	LOC Triples/Doc	LOC Triples (%)
DOC1	USA Cities	395	15,861	11,375	30	28	71.72
DOC2	Social networks	110	1,009	623	9	6	61.74
DOC3	Chat websites	32	74	19	2.3	0.6	25.68

Table 4. GeoNames feature classes of candidate place names

ID	A	P	L	H	TOTAL	A Ratio	P Ratio	L Ratio	H Ratio
DOC1	21,148	235,600	26,436	13,500	296,684	5.11	56.95	6.39	3.26
DOC2	672	7,430	469	428	8,999	4.61	50.98	3.22	2.94
DOC3	24	272	16	3	315	1.96	22.17	1.30	0.24

Table 5. GeoNames feature codes of place names from Wikipedia documents

ID	ADM1	ADM2	PPL	PPLA	Total	ADM1 %	ADM2 %	PPL %	PPLA %
DOC1	2,431	3,193	211,969	1,340	218,933	0.80	1.05	69.44	0.44
DOC2	93	98	6,830	26	7,047	0.81	0.86	59.84	0.44
DOC3	3	7	249	4	263	0.29	0.68	24.31	0.39

Table 6. GeoNames feature classes, normalized number of triples

ID	A	P	L	H	TOTAL	Triples	LOC Triples	LOC Triples (%)
DOC1	1,747	20,730	1,647	1,876	26,000	1,009	698	69.18
DOC2	672	7,430	469	428	8,999	1,009	623	61.74

Therefore, in these experiments, documents with a clear geographic context are likely to contain more place names or relationships to locations. Notice that the feature class ratios are calculated from the percentage of extracted triples, in such a way that only location-related entities are considered.

5. Conclusions and future work

This work has shown that extracting information on the geographic context of documents using NLP can help in the identification of place and location properties. Some of these properties refer to entity recognition and relationship extraction between place names

and other entities. The LinkedOntoGazetter has provided support to analyze geographic properties of places, with access to GeoNames feature classes and linked data that are related to location entity names. Results confirm that location entities are more common in the context of articles that are related to populated places and administrative divisions, in comparison to other document classes.

As future contributions, we propose evaluating the information extraction with a disambiguation process, finding place properties of named entities according to the relationships between non-place entities and place names. Location triples can be used to enrich LoG and other linked data sources, by providing relevant connections between entities, obtained from natural language text. Therefore, we also plan to investigate how relationships involving places and other entities, as observed in text, can be helpful in place name disambiguation and other geographic information retrieval tasks.

Acknowledgements

The authors wish to thank CNPq, CAPES and FAPEMIG, Brazilian agencies in charge of fostering research and development.

References

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Association for Computational Linguistics*, pages 363–370.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA*, pages 55–60.

Monteiro, B. R., Davis Jr., C. A., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.

Moura, T. H. V. M. and Davis Jr, C. A. (2013). Linked geospatial data: desafios e oportunidades de pesquisa. In *Proceedings of the XIV Brazilian Symposium on GeoInformatics*, pages 13–18.

Moura, T. H. V. M., Davis Jr., C. A., and Fonseca, F. T. (2017). Reference data enhancement for geographic information retrieval using linked data. *Transactions in GIS*, 21(4):683–700.

Perea-Ortega, J. M., Santiago, F. M., Ráez, A. M., and López, L. A. U. (2009). Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. *Procesamiento del Lenguaje Natural*, 43:33–40.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the NAACL on Human Language Technology*, volume 1, pages 173–180.