Kalpa Publications in Computing

Volume 13, 2019, Pages 47–55

Proceedings of the 1st International Conference on Geospatial Information Sciences

A Vector Semantics Approach to the Geoparsing Disambiguation Task for Texts in Spanish

Filomeno Alcántara[1], Alejandro Molina[2], and Victor Muñiz[3]

[1] Centro de Investigacion en Matemáticas, Monterrey, Nuevo León, México
adler@cimat.mx http://www.cimat.mx

[2] CONACYT – Centro de Investigación en Ciencias de Información Geoespacial, Mérida, Yucatán, México
amolina@centrogeo.edu.mx http://mid.geoint.mx

[3] Centro de Investigación en Matemáticas, Monterrey, Nuevo León, México
victor_m@cimat.mx http://www.cimat.mx

Abstract

Nowadays, online news sources generate continuous streams of information that includes references to real locations. Linking these locations to coordinates in a map usually requires two steps involving the named entity: *extraction* and *disambiguation*. In past years, efforts have been devoted mainly to the first task. Approaches to location disambiguation include *knowledge-based*, *map-based* and *data-driven* methods. In this paper, we present a work in progress for location disambiguation in news documents that uses a vector-semantic representation learned from information sources that include events and geographic descriptions, in order to obtain a ranking for the possible locations. We will describe the proposed method and the results obtained so far, as well as ideas for future work.

1 Introduction

Nowadays, online news sources generate continuous streams of information with global coverage that directly impact the daily activities of people in a wide variety of areas. Although social media can also generate real-time news content, newspapers remain a source of information that incorporates a refined version of relevant events. In recent years, topic analysis has been used on these news to determine the types of news delivered by newspapers to their readers. The use of news analytics to perform tasks such as automatic location disambiguation and risk maps is interesting, especially given the increasing popularity of Natural Language Processing (NLP). In this work, we use the term "location" to mean a spatial mention, i.e., a place name (e.g. Mexico City).

Problems that involve mapping information geospatially from any kind of free-speech text require some sort of location extraction solution. This is usually made through some type of geoparsing, geocoding or geotagging.

Geocoding, geoparsing and geotagging are types of information extraction, which is itself a subset of information retrieval. Geocoding is the act of transforming a well-formed textual

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. 47–55

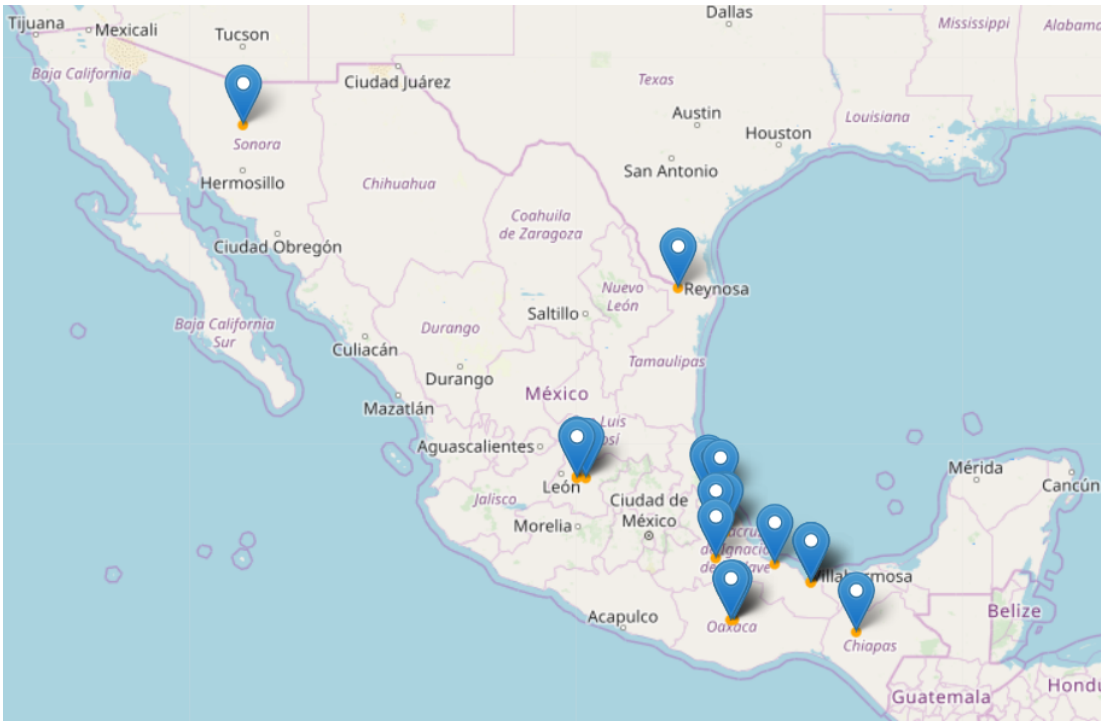

Figure 1: Locations with the name Guanajuato.

representation of an address into a valid spatial representation, such as a spatial coordinate or a specific map reference. Geoparsing does the same for unstructured text and involves location extraction and location disambiguation prior to the final geocoding. Geotagging assigns spatial coordinates to media content items, typically by building statistical models that, given a piece of text, can provide an estimate of the most likely location to which the text refers [4].

An extracted location can be ambiguous, meaning it may not refer to a unique location, for instance, figure 1 show all possible options for a location named Guanajuato, which include a state, a city and streets, among others locations with different administrative boundaries.

There are several approaches to deal with location disambiguation that can be grouped into three main categories, as described in [3]:

- *map-based*: methods that use an explicit representation of locations on a map, for instance, to calculate the average distance of unambiguous locations from options;
- *knowledge-based*: methods that exploit external knowledge sources such as gazetteers, Wikipedia or ontologies to find disambiguation clues;
- *data-driven* or *supervised*: methods based on machine learning techniques.

Nominatim is an open source geocoding tool that uses a *knowledge-based* method to deal with location disambiguation. The search algorithm in Nominatim includes searches for each valid combination of the tokens of a user query. After all searches are done, Nominatim returns a collection of places sorted (in local installations) by an importance index assigned by Nominatim

. This index is is calculated either from the tagging (i.e. town, city, country) or preferably from the linked Wikipedia article and has proven to be very effective in the disambiguation task. We will use this tool to obtain a list of possible options for a location.

This work explore the relationship between the semantic and geographic space, and how it can be used in the task of location disambiguation.

2 Data

The data used for this work consists of three corpora formed from different documents. The first one is a corpus provided by CentroGeo [6] with 5870 locations distributed in 1233 news documents from the main digital media of Mexico. Figure 2 shows an example of a news document that contains two different locations: Monterrey and Nuevo León. The locations were provided by CentroGeo based on OpenNLP's Named Entity Recognition (NER) module [1].

Figure 2: Example of News document with two locations: Monterrey and Nuevo León.

The second corpus consists of 12453 Wikipedia articles related to the Geography of Mexico: states, cities, municipalities, etc. These articles were obtained through the Wikipedia's API and were subsequently preprocessed in order to extract and clean their plain texts (no tables, URLs, pictures, infoboxes, references, html code, among other things). These documents were used to enrich the semantic space with more extensive descriptions of the regions of Mexico.

For the third corpus, we selected 20 news documents from the news corpus from which we obtained a total of 950 *pseudo documents* obtained as follows:

- Extract the locations.
- Get possible options (up to 20) for each location through Nominatim's API [7]. In this step, we also get additional information: latitude, longitude, administrative level, alternative names, among other things that are useful in the process.
- Get streets and amenities around each option in a $200m$ radius using OverPass API.
- Concatenate the results in the previous step to get a pseudo document for each option.

The objective of the pseudo documents is to describe the geographical surroundings of the locations mentioned in the news documents.

The purpose of the three corpora is to generate vector semantic models that help us to exploit the context of the news in order to disambiguate locations, as is explained in the next section.

3 Methodology

As we mentioned before, the main idea of this proposal is to explore the relationship between the geographic space (the distribution in a map) and the semantic space representation (the meaning according to the context) of the locations in a news document.

To obtain the vector semantic space, we used embeddings, which maps words from a vocabulary into dense vectors with real entries in a high-dimensional (generally, 100 to 300) space. The basic idea was first introduced in [2] as a means to beat the curse of dimensionality.

word2vec and doc2vec are among the most popular methods to make embeddings ([5]). The idea of word2vec is to map a word into a vector that represents the word's local context. The word representations include semantic and syntactic information obtained from the context words. Word2vec is used to build vector representations of words and phrases, and cannot be used to obtain representations of documents. Words put together have a logical structure based on syntax rules. In contrast, documents do not follow a logical ordering. The purpose of doc2vec is to address this matter. Word embeddings cannot be used to find similarities between documents that contain the corresponding words. The common approach is to average the word embeddings to make document embeddings; however, the semantic information of the word is usually lost. Doc2vec allow us to obtain both paragraph embeddings and word embeddings, by adding the documents IDs into the training data.

We considered these methods because they are able to elicit some complex connections of the texts as input for disambiguation, through the use of the semantic space learned from the combination of the corpora described in Section 2.

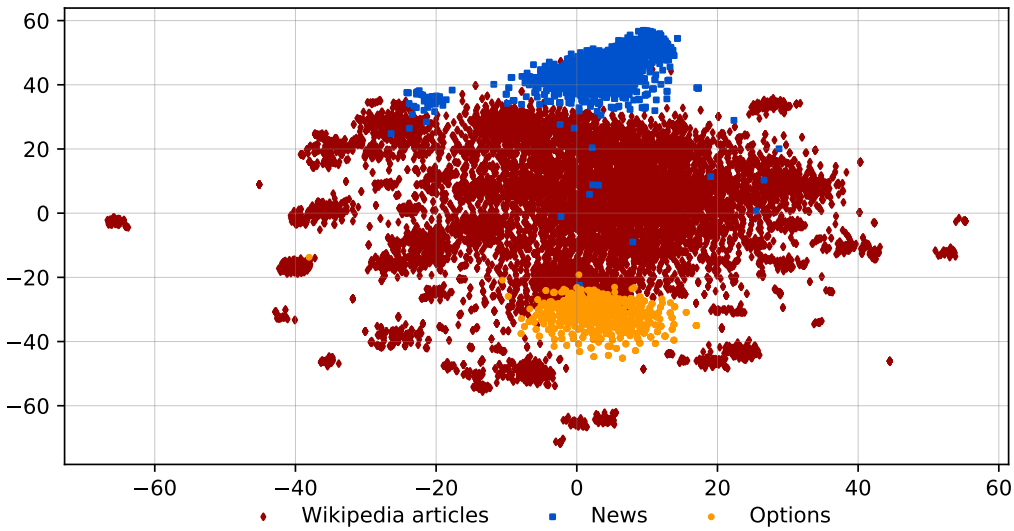

Figure 3: t-SNE representation of the three types of documents.

We used two models. The first model was built by using all documents (news, Wikipedia articles and pseudodocuments) in order to explore patterns and relations between the three types of documents. Figure 3 shows the t-SNE [8] representation of all 14636 documents. As expected, due to the nature of each type of document, news are more similar to news, and options to options. This fact can be verified by using the cosine distance (the standard metric for the similarity between document embeddings). In addition, we could see some clusters in the Wikipedia articles that correspond to articles on the same category. For instance, cloud-centered in coordinates (−40,−16) contains annexes of monuments throughout the country. Also, we could find cases in which news are close to the Wikipedia articles of the locations mentioned in them and cases where the options, the news documents and the Wikipedia articles of those locations are close.

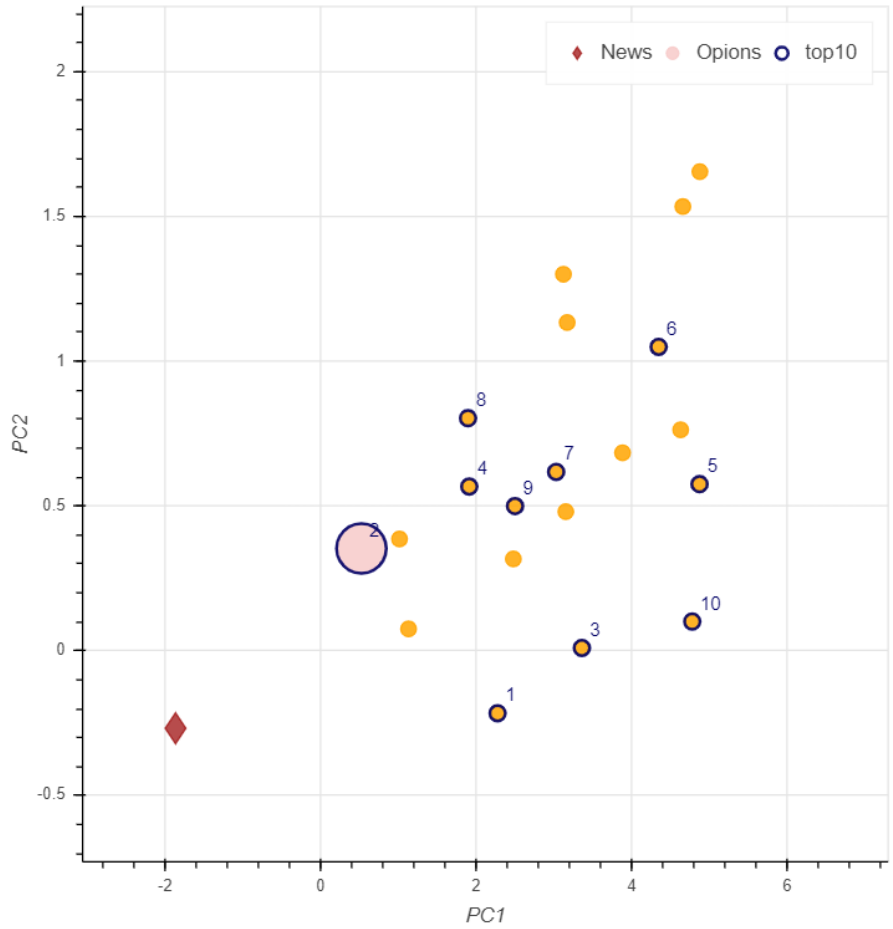

Figure 4: First two components of PCA obtained from the semantic model learned from news documents and Wikipedia articles. This model was used for the ranking procedure. The size and color of points representing location options, corresponds to their administrative level. In this example, the named location is Nuevo León (from news document showed in figure 2).

For the disambiguation procedure, we built another doc2vec model using the news documents and the Wikipedia articles. With this model, we obtained a semantic space in which we can project the options for locations and make the disambiguation. This way we can use the model for location disambiguation in documents that are not in the training corpus. Once we trained the model we proceed to do the disambiguation in the following way:

- We take one location from one of the 20 selected news documents.
- We extract the vector of the news document from the doc2vec model.
- We infer a vector for each one of the possible options for the location selected in the first step.
- We compute the distance (Euclidean and cosine distances) between the news document vector and each one of the vectors of the options.

- We sort the results in increasing order so that we get a ranking based in documents similarity.

In figure 4 we show the first two components of PCA for the ranking configuration for the location "Nuevo León" in the news document of the figure 2. As mentioned in the methodology section, the point labeled as 1 correspond to the closest possible option (in the semantic space).

4 Results

The rankings produced by Nominatim were compared to those of our proposed methods through the distance (in kilometers) between each option and the real locationHowever, this distance is not an absolute metric. In order to show the results, we selected some cases of interest that are presented in Table 1[1].

Table 1: Comparison of the rankings.

News ID	Location	True admin level	Admin level	Distance	Nominatim	Ranking Cosine	Ranking Norm
18	3	6	15	0	2	2	6
			6	7.36	1	1	1
			15	69.5	11	8	11
24	1	6	15	0	12	14	6
			15	22.01	15	11	16
			6	47.51	1	3	10
26	2	4	4	0	1	6	1
			15	69.72	6	8	13
			15	675.46	12	19	18
33	1	6	6	0	2	2	1
			15	58.62	1	4	5
			15	184.19	5	5	3
39	8	15	15	0	1	2	1
			15	0.77	2	6	6
			15	21.32	3	5	3
110	2	6	15	0	2	2	6
			6	7.36	1	3	1
			15	69.5	11	11	11
146	2	15	15	14.3	2	1	2
			15	37.82	8	6	11
			15	42.08	3	4	3
146	8	15	15	70.91	17	1	3
			15	102.48	19	13	12
			15	229.54	18	10	9
231	2	4	4	0	1	1	1
			6	29.33	3	10	8
			15	31.3	2	16	3
582	6	4	4	0	1	1	1
			15	35.82	8	16	8
			15	38.4	9	19	15

[1] We have 950 results in total

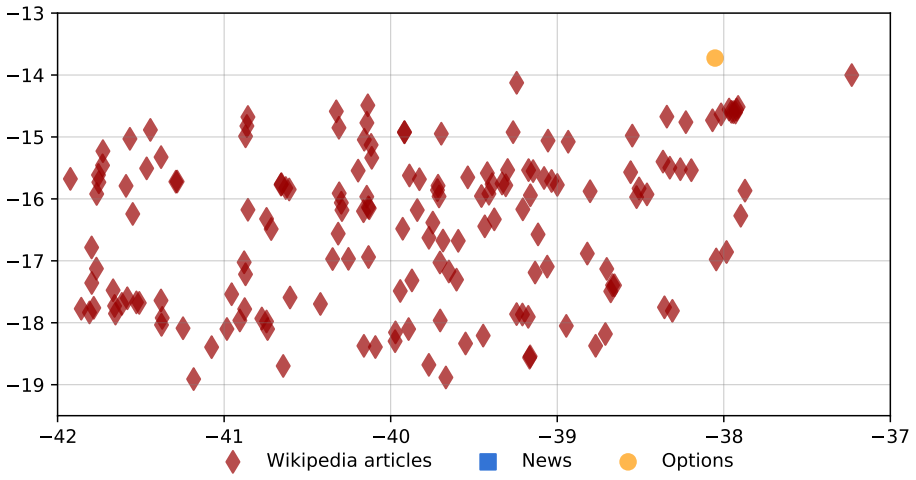

Figure 5: t-SNE projection of news document, Wikipedia articles and location options obtained from the embeddings. In this case Wikipedia articles are mostly from annexes that mention points of interest in the municipalities. These documents are composed by a series of short sentences. The same as the option.

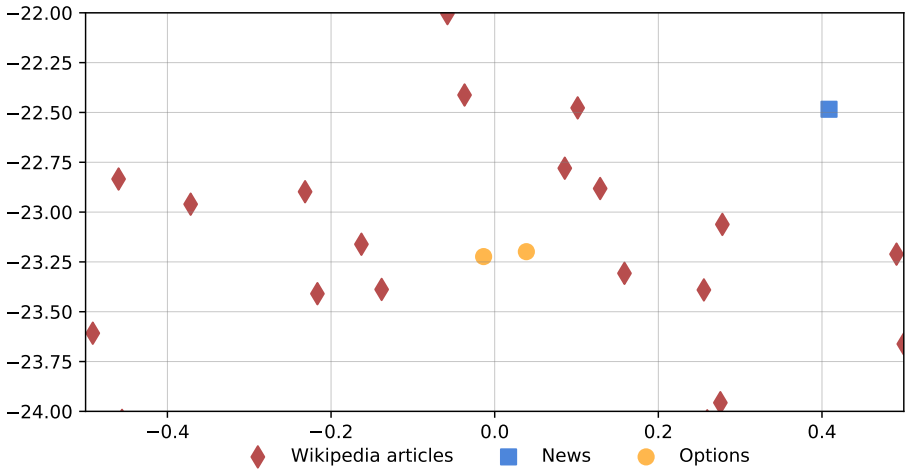

Figure 6: t-SNE projection of news document, Wikipedia articles and location options obtained from the embeddings showing the linkage among the three types of documents. One of the locations in the news document is Puebla. The options are: Puebla (the correct option) and Zoyatepec (a locality in Tecali, a municipality in the State of Puebla). Also, there are articles of several locations related to Puebla: Atoluca, Valle de Puebla-Tlaxcala, Llanos de San Juan, Casa del Traductor (BUAP) and some municipalities from Tlaxcala, among other articles.

Each entry of the table has three rows corresponding to the three nearer options for a location sorted by distance. As an example, consider the three rows of the first entry: they correspond to options for location 3 of the news document with ID-18, whose true administrative level is 6. The administrative levels of each option are 15, 6 and 15, respectively. These options are ranked by Nominatim as 2, 1 and 11. The last two columns of the table contain ranking scores obtained by using the cosine distance and the euclidean distance.

As we can see in table 1, except for news document 146, the nearest option appears in first or second place in the Nominatim ranking. The behavior of our proposed ranking is similar, but there are cases in which the nearer options are outside the top 5 of the rankings. For location 8 of news ID-146 our rankings perform better. However, since the administrative level 15 corresponds to streets, buildings and neighborhoods, a distance of $70km$ is quite big, i.e., even though an option can be the closest one, it may not be correct. Figure 5 shows relations between news documents and some Wikipedia articles. Figure 6 presents a news document, two options related to locations named in that news document and some Wikipedia articles related to those same locations. The structure of both figures could mean that a better version of the documents will be useful for the tasks we want to perform.

5 Discussion

The results indicate that there exist a relationship between the semantic and geographic spaces and that this relationship can be exploited to disambiguate locations in a news document. We found that there are cases in which news documents are related to the Wikipedia articles of the locations mentioned in them, and that most of the time distance to real location increases considerably when we consider rankings with more than 5 results.

In addition, we found that the general structure of the document plays an important role in the procedure. The Wikipedia articles for locations usually contain information about different topics such as culture, economy, education, government, history and so on. This means that we have a large document, so the comparison with the other types of documents is affected. Also, we have to consider that a pseudodocument is not a literal descriptions of the surroundings of a possible location. This could be the reason why the options and the pseudodocuments are not close in the semantic space.

6 Conclusions and future work

As we could see from the results, our proposal for disambiguation based on vector-semantic representation of the locations named in the news dataset, showed a good performance compared with the options given by Nominatim; we were able to model the relationships between geographic locations and their semantic representations in texts. Generally, the nearest option appears in the top 5 of our rankings.

Distance to real location is a good measure to compare the options given by the two methods, but can not be used as a definitive metric in this case, because our ranking is a reordering of the one given by Nominatim and both are based on different methods (Nominatim considers many more features in order to calculate its ranking). However, we think that our proposal can be used to incorporate relevant information to get a more robust disambiguation method.

There are several tasks to take into account for future work. Among them are:

- Include a better version of the Wikipedia articles (for instance, just the introductory part) so that the similarity between all three types of documents can be used more fruitfully.

- Improve the construction of the pseudodocuments for the possible options by adding simple words or phrases in order to get basic descriptions of the surroundings of a location.
- Use a bigger news documents corpus to have a better mapping of the semantic space of the Mexican territory.

7 Acknowledgments

This research was supported by The National GeoIntelligence Laboratory (GeoInt) `http://www.geoint.mx`.

References

[1] Apache Software Foundation. Opennlp.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[3] Davide Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19, July 2011.

[4] Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst.*, 36(4):40:1–40:27, June 2018.

[5] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[6] 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.

[7] OpenStreetMaps contributors. Nominatim . `https://nominatim.openstreetmap.org`, 2017.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.