Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: http://ees.elsevier.com

Geographic Named Entity Recognition and Disambiguation in Mexican News using word embeddings

Alejandro Molina-Villegas [a,*], Victor Muñiz-Sanchez [b], Jean Arreola-Trapala [b], Filomeno Alcántara [b]

[a] Conacyt - Centro de Investigación en Ciencias de Información Geoespacial, Yucatan Technological Science Park, Merida 97302, Mexico
[b] Centro de Investigación en Matemáticas, Monterrey Technological Research and Innovation Park, Monterrey 66628, Mexico

ARTICLE INFO

Article history:
Received 4 August 2020
Received in revised form 10 December 2020
Accepted 3 March 2021
Available online xxx

Keywords
Geographic Named Entity Recognition
Geographic Named Entity Disambiguation
Geoparsing

ABSTRACT

In recent years, dense word embeddings for text representation have been widely used since they can model complex semantic and morphological characteristics of language, such as meaning in specific contexts and applications. Contrary to sparse representations, such as one-hot encoding or frequencies, word embeddings provide computational advantages and improvements on the results in many natural language processing tasks, similar to the automatic extraction of geospatial information. Computer systems capable of discovering geographic information from natural language involve a complex process called geoparsing. In this work, we explore the use of word embeddings for two NLP tasks: Geographic Named Entity Recognition and Geographic Entity Disambiguation, both as an effort to develop the first Mexican Geoparser. Our study shows that relationships between geographic and semantic spaces arise when we apply word embedding models over a corpus of documents in Mexican Spanish. Our models achieved high accuracy for geographic named entity recognition in Spanish.

1. Introduction

The ability to integrate geolocation into voice assistants is creating interest in academic research and commercial investment. Some cutting-edge navigation systems include disruptive technologies that allow users to "tell" their cars where to go. Besides, there are also many interesting applications related to detecting specific locations in textual data, such as tracking information about clandestine graves or disappeared people in Mexico.

The cornerstone of the abovementioned applications is called geoparsing. Geoparsing is a sophisticated natural language processing (NLP) task for georeferencing entities naturally mentioned in free text (written or obtained through automatic transcription). Note that there is an essential difference between geoparsing and geocoding. In geoparsing, the input does not include clues about where the places mentioned in the input are located. In geocoding, a valid textual representation of a location (an address) is the input. It follows that a geocoder must simply find the coordinates of the input address in a gazetteer. Thus, what makes geoparsing so challenging is that it deals with raw natural language information. We present the first advances towards the creation of the first geoparsing system for Mexican Spanish. To achieve this aim, the first subprocess of our approach is to recognize where the locations are mentioned, which is known in NLP as Named Entity Recognition (NER). Once the geographic entities are detected, we use a geocoder to obtain the coordinates. Finally, a disambiguation stage is needed because of the high degree of ambiguity of toponyms that could correspond to many different coordinates, for example, Paris (France), Paris (Arkansas), and Paris (a street in Mexico City).

Throughout this article, we describe the algorithms, methods, and tools related to Geographic Named Entity Recognition and disambiguation of data in Mexican Spanish. We will explain the relationship between geographic space and semantic space and how to model such relationships to detect places immersed in free text. We also describe our first approach for entity disambiguation by data enrichment using Wikipedia.

2. Related work

2.1. Named Entity Recognition

Named Entity Recognition (NER) refers to the automatic detection and classification of entity names in domain-specific documents. This research topic has become very relevant since high-performance NER systems usually precede other complex NLP tasks, including information extraction, knowledge base population, named entity linking, machine translation, word sense disambiguation, and notably, geoparsing (Aguirre et al., 2015). Therefore, NER is considered the cornerstone for some ambitious projects, which is why it has been an active re-

* Corresponding author at: CentroGeo at the Yucatan Technological Science Park, Sierra Papacal-Chuburna Port Highway, Km 5 Sierra Papacal, Merida, YUC 97302, Mexico.
E-mail addresses: amolina@centrogeo.edu.mx (A. Molina-Villegas); victor_m@cimat.mx (V. Muñiz-Sanchez); jean.arreola@cimat.mx (J. Arreola-Trapala); filomeno.alcantara@cimat.mx (F. Alcántara)

https://doi.org/10.1016/j.eswa.2021.114855
0957-4174/

search area for some years and has been recently applied in many fields, such as medicine (Tanabe, Xie, Thom, Matten, & Wilbur, 2005), chemistry (Rocktäschel, Weidlich, & Leser, 2012), history (Smith & Crane, 2001) and geology (Sobhana, Mitra, & Ghosh, 2010).

Even though some of the first NER methods relied on lexical rules (Sekine & Nobata, 2004), modern NER is based on mathematical methods because they are more robust and generalize better among different languages. Three of the most successful NER frameworks are maximum entropy (MaxEnt) models, conditional random fields (CRFs), and, more recently, deep neural networks (DNNs).

In the MaxEnt framework, the posterior probability of NER labels for a sequence of words is modeled by the maximum entropy proposed by Berger, Pietra, and Pietra (1996). In this framework, there is a set of arbitrary feature functions that must be weighted using a single parameter. The feature functions could consider lexical aspects such as word-counts, capitalization, prefixes and suffixes, dictionary-based features, among other language-dependent characteristics.[1]

A CRF is a general stochastic model commonly used to label and segment sequential data; it provides a general framework to build sequence models for NER or any other task. A sequence observed during the training stage is the sequence of tokens that fit a sentence or a document, and the sequence of states corresponds to the entity labels provided during this stage (Finkel, Grenager, & Manning, 2005). Since languages differ from the conventions they use for named entities, the features have to be specific to each language. Therefore, a restriction of this approach is that its effectiveness is limited and may vary from language to language[2].

In the DNN approach, a Neural Network is used for both, features-learning and entity classification. The words from a sentence are tokenized, and then they are broken up into features and aggregated into a representative vector. This vector is then fed into a Convolutional Neural Network, which makes a classification based on the weight assigned to each feature within the text (Serrà & Karatzoglou, 2017). The training stage requires a lot of data manually labeled for NER[3].

2.2. Geographic Named Entity Recognition and Spanish

Despite the high performance of current NER systems, the research on Geographic Named Entity Recognition is still a very active area for many reasons. In the first place, Geographic Named Entity Recognition (GNER) aims not only to distinguish entities within the text but also tries to assign them with their explicit georeference (e.g., lat/long), which is more challenging than only detect and classify entities. Besides, the problem of toponym ambiguity is as common as it is complex. Indeed, there are different types of ambiguities (Smith & Mann, 2003; Wacholder, Ravin, & Choi, 1997; Moncla, Renteria-Agualimpia, Nogueras-Iso, & Gaio, 2014): the same name is used for several places (referent ambiguity); the same place has several names (reference ambiguity); the place name can be used in a non-geographical context, e.g. is the word *google* referring to the place or the company? (referent class ambiguity); the words constituting the place name are ambiguous, e.g. is the word *Lake* part of the toponym *Lake Grattaleu* or not? (structural ambiguity) or the place name is not found in gazetteers (unreferenced toponyms ambiguity). For GNER, some approaches have to consider the specific language patterns that could provide extra information to identify toponyms in a specific region. For example, Mexican Spanish still preserves ancient location names like *Xochimilco* or *Tulum*. In consequence, applying models generated from texts collected from Spain (Tjong Kim Sang & De Meulder, 2003) will not scale well for all the other Spanish variants of Latin America. Surprisingly, we have found that only some GNER approaches have been tested for documents in Spanish.

Silva, Martins, Chaves, Afonso, and Nuno (2006), presented a method tested on English, Spanish, Portuguese, and German. The main concept of their proposal is the use of what they called *the geographic scope* of web pages. The geographic scope specifies the relationship between an entity on the web and an entity on the geographic domain (such as an administrative location or a 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 by considering the similarity to other documents. The work was focused on feature extraction, recognition, and disambiguation of 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. Gelernter and Zhang (2013) presented a geoparser for Spanish translations from English. The proposed method uses four parsing steps: a lexico-semantic Named Location Parser, a rule-based building parser, a rule-based street parser, and a trained Named Entity Parser. Authors developed a parser for both languages and the NER module was trained using the GeoNames gazetteer and the CRF algorithm. The method was evaluated through four thousand tweets and the resulting model was able to recognize location words from Twitter by recognizing streets and buildings names. Moncla et al. (2014) proposed a processing pipeline to support geoparsing and geocoding of documents in French, Spanish, and Italian. The authors used a hiking corpus since this type of documents is rich in toponyms, displacements descriptions, spatial relations, and other useful features. The approach consists in two main parts: geoparsing based on syntactical-semantic combined patterns (in a cascade of transducers); and a disambiguation method based on clustering of spatial density. First, the geoparser module extracts toponyms and spatial relations. Then, for the disambiguation stage (toponym resolution) each entity extracted from the textual descriptions is searched in a gazetteer. During this last process, a lot of ambiguities may have arisen. Finally, the DBSCAN algorithm is used to detect the outliers; which in their GNER proposal context means a point that does not belong to the hiking trail cluster. There is a lack of GNER results for non-English documents and particularly in Spanish. In the case of Mexican Spanish, this is due to the unavailability of data ready-to-use for this task. Indeed, we have found only one project that focuses on Mexican Spanish as the target language (Aldana-Bobadilla et al., 2020).

3. Materials and methods

3.1. Labeled data from Mexican News and other resources

We used three different corpora to tackle three specific objectives.

The first objective was to produce our own set of word embeddings (Section 3.2); for this goal, we used the corpus C1 composed of 165354 news documents from the main digital media in Mexico.

The second objective was to generate a GNER model for Mexican Spanish trained on our set of word embeddings (Section 3.3); for this task, we used the embeddings created from C1 and the second corpus C2 composed by 1233 news documents manually labeled with geographic named entities tags. For example *"…those affected by the earthquake in <loc> Mexico City <loc>…"* makes an explicit distinction of a geographic entity in a specific context. A total of 5870 geographic entities in context were found and labeled for the C2 corpus.

The third objective was to propose a disambiguation method based on semantic space enrichment (Section 3.4); for this task, we used the third corpus C3 with 12453 Wikipedia articles related to the geography of Mexico: states, cities, departments, provinces, capitals, counties, dis-

[1] The Apache OpenNLP library provides a NER model based on MaxEnt.
[2] The Stanford CoreNLP offers NER software based on CRFs.
[3] The Spacy package provides a NER module based on CNN.

tricts, municipalities, villages, towns, valleys, lakes, rivers, counties, among other types. The Wikipedia articles were used to enrich the semantic space by adding more extensive descriptions of the regions of Mexico. The following sections detail how models were created from C1, C2, and C3.

3.2. Word embeddings creation from Mexican News

Word embeddings are dense vector representations of words which can be extended to documents of any length. From a machine learning point of view, word embeddings are useful because they are more capable of representing characteristics of texts, – such as semantic relationships –, than those obtained with word frequencies due to the reduction in the number of parameters to learn. There are different ways to obtain word embeddings, but we can group them in task-specific and domain-specific. The first group corresponds to representations of specific words to solve a learning problem such as classification. Embeddings of arbitrary length are weights in a real vector space which minimize a misclassification cost function, and they are treated as model parameters that are learned together with the other parameters of the classifier, generally, a neural network (Goldberg, 2015). The other group, the domain-specific word embeddings are focused on representing the semantic characteristics of texts according to the use of their words on specific domains, such as a special slang, or a discipline (medicine, laws, chemistry, biology, etc). These embeddings are learned from a language model based on a corpus related to the domain. There are different approaches to obtain domain-specific word embeddings most of them based on the neural network language model (NNLM) (Bengio, Ducharme, Vincent, & Janvin, 2003); those are *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013), *Glove* (Pennington, Socher, & Manning, 2014), *fastText* (Bojanowski, Grave, Joulin, & Mikolov, 2017), and contextual embeddings based on deep learning architectures like *ELMo* (Peters et al., 2018) or *BERT* (Devlin, Chang, Lee, & Toutanova, 2019).

Although there are pre-trained word embeddings for most of the models, for many languages, they were learned from big general-purpose corpora, and in our case of study, after extensive experimentation, we realized that we could not obtain the semantic-geographic relationship of our interest. For this work, the particular interest of using domain-specific word embeddings for GNER is the fact that the semantic-geographic relationship can be modeled in the embeddings space. For example, according to the official documentation of *word2vec*,[4] once the embeddings are trained from an appropriate corpus, if the generated model is used to query the word "*France*", the most similar words are "*Spain*", "*Belgium*", "*Netherlands*", "*Italy*", and so on. Furthermore, vector operations such as $Paris - France + Italy$ result in a vector that is very close to *Rome*.

Since this property is valid for other geographic entities, such as regions, cities, and even streets we have created our word embeddings set from the Mexican News corpus C1 in order to use it for Geographic Entity Recognition and Disambiguation in this language. When choosing the model for word embeddings, we have taken into account computational aspects and the capacity to handle vocabulary (OOV), which is particularly important for the GNER task. The best performance in computational aspects is achieved with *word2vec* since it simplifies the NNLM architecture eliminating the hidden non-linear layer and using a log-linear binary classifier with negative sampling in the skip-gram model, instead of the original softmax defined over the whole vocabulary. As for the OOV words, *fastText* and *ELMo* models are capable of generating embeddings for these words because they can operate at the character level (*ELMo*) and the sub-word level (*fastText*), although the computational cost increases and a huge corpus for training is needed. For this reasons, we decided to use *word2vec* with a Context Encoder (Horn, 2017) in order to deal with OOV words, in addition to obtaining additional advantages explained later in this section. Using the Mexican News corpus C1 described in Section 3.1, we have created three different word embeddings models: a Continuous Bag of Words model (CBoW) (Mikolov et al., 2013), a Context Encoder with Global information (ConEc Global), and a Context Encoder with Global and Local information (ConEc G&L) (Horn, 2017).

The CBoW model can be interpreted as a neural network that predicts the similarities of a word to other words. During training, for each occurrence i of a word w in the texts, a binary vector $x_{w_i} \in R^N$, which has 1 at the positions of the context words of w and 0 elsewhere, is used as input to the network and multiplied by a set of weights W_0 to arrive at an embedding $y_{w_i} \in R^d$ (the summed rows of W_0 correspond to the context words). This embedding is then multiplied by another set of weights W_1, which corresponds to the full matrix of word embeddings Y, to produce the output of the network: a vector $s_{w_i} \in R^N$ containing the approximated similarities of the word w to all other words. The training error is then computed by comparing a subset of the output to a binary target vector $t_{w_i} \in R^{k+1}$, which serves as an approximation of the true similarities s_w when considering only a small number of random words.

The ConEc training is very similar to the CBoW training however, the important difference is the computation of a word's embedding after the training is completed. In the case of the CBoW model, the word embedding is simply the row of the tuned W_0 matrix. With ConEc, the final vector representation is obtained by multiplying W_0 by the mean context vector x_w of the word. We have used two types of contextualization: global and global & local.

In the ConEc Global model, the global vector is obtained via the mean of all the binary context vectors x_{w_i} corresponding to the M_w occurrences of w in the training corpus according to Eq. (1).

$$x_{w_{global}} = \frac{1}{M_w}\sum_{i=1}^{M_w} x_{w_i} \tag{1}$$

In the ConEc G&L model, the local context vector is computed according to Eq. (2).

$$x_{w_{local}} = \frac{1}{m_w}\sum_{i=1}^{m_w} x_{w_i} \tag{2}$$

where m_w corresponds to the occurrence of w in a single document. The final embedding of a word is obtained according to Eq. (3).

$$y_w = (\alpha \cdot x_{w_{global}} + (1-\alpha)\cdot x_{w_{local}})^T W_0 \tag{3}$$

with $\alpha \in [0, 1]$. Note that the choice of the parameter α determines how much attention is placed on the word's local context, which helps to distinguish between multiple meanings of the word. The three-word embedding models were trained with a window of 5 words to obtain vectors of dimension 100. The results for the three different word embeddings models are presented in Section 4.1.

3.3. Geographic Named Entity Recognition

The main idea for the GNER model is to use the word embeddings models described in Section 3.2 combined with the labeled data (corpus C2 in Section 3.1) to train a Neural Network Classifier.

The embeddings to train the GNER classifier were obtained in the following way: all the words in the vocabulary of the labeled corpus C2 received the embeddings of a model trained from the unlabeled corpus C1. If a word of C2 does not appear in the C1 corpus, then it preserves the embedding from the C2 embeddings model. In this way, we ensure

[4] https://code.google.com/archive/p/word2vec/.

that all words have an embedding. Three models with different values of the context parameter α ($\alpha \in [0, 1]$) were trained to decide which local context level use. The metric to choose the best model was the F-score, obtaining the best context level when $\alpha = 0.5$. In addition to the embeddings, the inputs were enriched with lexical and syntactic characteristics. They included lexical boolean features such as *starts-with-a-capital-letter* or *has-an-internal-period*, which is a well know strategy to improve NER (Cucerzan & Yarowsky, 1999). Other lexical features were also included as binary variables. For instance, for each token, we checked several lexical properties such as if its individual characters are numeric, the number of characters, if the token is a stopword, and its part-of-speech tag.

Although some NER methods are based on deep encoder-decoder architectures, it is well known that they require a huge amount of labeled data to generate accurate models. In our case, this is a drawback, since our labeled corpus is not large enough for using such deep models. We have used Cross-Validation to select the appropriate complexity of the neural network which achieves a better performance. In our case, after testing a wide range of architectures, the best model for the GNER classifier was a neural network with one hidden layer and a sigmoid activation function with weight decay. At the end of the GNER classifier training, the last layer determines, whether a given token is a geographic named entity or not. The results for the GNER model are presented in Section 4.1.

3.4. Geographic Entity Disambiguation

The main idea of the geographic entity disambiguation approach is to enrich the knowledge about ambiguous geographic entities by creating what we called *pseudo documents* and then compare the original document to the *pseudo documents* in a semantic space. The *pseudo documents* of an ambiguous entity are created by collecting information from each of the possible entity candidates; that is, those entities that have the same name but different coordinates (referent ambiguity).

As an example, consider the extract of a news document, where we emphasize the ambiguous geographic entity "*Nuevo León*". A translation of the text reads thus:*The director of Preventive Programs said that due to the violent events in Nuevo León, an awareness program will be implemented to prevent the use of weapons…*

La directora de Programas Preventivos dijo que debido a los hechos violentos de **Nuevo León** se implementará un programa de concientización para prevenir el uso de armas…

First, the ambiguous entity is searched on OpenStreetMap Nominatim[5] which response could have up to 20 entity candidates with the same name. The entity candidates for "*Nuevo León*" are shown in Fig. 1. As we can see, the candidates are spread all over the country.

Then, the points of interest (streets, parks, banks, hotels, restaurants, etc.), of each entity candidate are retrieved, considering a radius of 200 meters, using the Overpass API.[6] All the collected information is combined in a single bag-of-words to represent the *pseudo documents* of each entity candidate. For illustration purposes, we show this procedure for three entity candidates in different states: (1, *"Nuevo León, Mexicali, Baja California, 21705, México"*), (2, *"Nuevo León, Guadalupe, Nuevo León, México"*) and (3, *"Nuevo León, Aguacatal, Xalapa, Veracruz de Ignacio de la Llave, 91130, México"*), which are mapped in Figs. 2–4, respectively and whose corresponding pseudo-documents are:

```
Nuevo León (1):[‘‘Nuevo León", ‘‘Carretera Mexicali-Estacion Coahuila", ‘‘Carretera Mexicali-Algodones", ‘‘Carretera Algodones"]

Nuevo León (2): [‘‘Paras", ‘‘Jose Maria Morelos", ‘‘Doctor Arroyo", ‘‘Maria Guadalupe", ‘‘Abasolo", ‘‘Mina", ‘‘San Victoria", ‘‘Aramberri", ‘‘Nuevo leon", ‘‘Doctor Gonzalez", ‘‘San Pedro", ‘‘Lampazos", ‘‘Nuevo León", ‘‘Maria", ‘‘Dr. Coss", ‘‘Concepcion", ‘‘Peral", ‘‘Bufalo"]

Nuevo León (3): [‘‘Queretaro", ‘‘Nuevo Leon", ‘‘Privada de Jorullo", ‘‘So- conusco", ‘‘Pestalozzi", ‘‘Centro Estatal de Cancerologia Miguel Do- rantes Meza", ‘‘Aguascalientes", ‘‘1a Privada Nuevo León", ‘‘Tacos Fide", ‘‘Salón de Eventos Los Anturios", ‘‘SEDEMA Direccion General de Desarrollo Forestal", ‘‘Michoacán", ‘‘2a Privada Nuevo León", ‘‘Pri- vada Guadalajara", ‘‘Avenida Diamante", ‘‘Escuela Primaria 16 de Sep- tiembre", ‘‘Jorullo", ‘‘Aguacatal"]
```

In the disambiguation process, we compare the original document vector to the *pseudo document* vectors in a semantic space which has been created by using the same embeddings models described in Section 3.2. The differences are that the corpus C1 and C3 were combined, to create the semantic space, and that *doc2vec* (Le & Mikolov, 2014) (an extension of *word2vec* to documents) were used to generate documents embeddings.

Using the cosine and the euclidean distances, we established a ranking where the *pseudo documents* are sorted according to their distances to the original document. This is illustrated in Fig. 5 where the embeddings are projected in the first two principal components (PCA) of the semantic space.

The rhombus represents the original document while the circles represent the *pseudo documents* of the candidate entities. It is worthwhile to mention that the rankings are computed by using the complete dimension of the embeddings, and the illustration of Fig. 5 is just a low-dimensional representation. The results of an evaluation with about 950 *pseudo documents* are described in Section 4.2.

4. Results and discussion

4.1. Results on Geographic Named Entity Recognition

The results of three different word embedding models are presented in Table 1. The best performance for GNER is obtained by using the ConEc Global & Local encoder. However, the three word embedding models (CBoW, ConEc Global and ConEc G&L) trained with the Mexican News corpus C1 have achieved to model correctly the relationship between the semantic space and the geographic space in the specific context of News in this language. As an example of the generated relationships between the semantic space and the geographic space, in Fig. 6, we present a PCA projection for the 10 most similar embeddings of the entity *"Mérida"* (a Mexican tourist destination) using the ConEc G&L model. We can see among the closest entities other Mexican tourist destinations like *"Cancun"*, *"Tulum"*, and *"León"*. However, other semantically related terms appear, such as *"capital yucateca"* (yucatecan capital) or *"destino turístico"* (tourist destination), which are, in the context of News, synonyms of *"Mérida"*. Moreover, this property is valid for other geographic entity types like states, departments, provinces, capitals, counties, districts, municipalities, villages, towns, valleys, avenues, roads, streets, among others.

An additional advantage of the ConEc models is that together, they allow to obtain representative embeddings even for OOV words. Since an OOV word does not have a global context (as it never occurred in the training corpus), its embedding is computed solely based on the local context (*i.e.*, by setting $\alpha = 0$).

To compare the performance of our GNER model with other NER software such as Spacy and OpenNLP, we collected 300 news documents from a digital news newspaper called *El Gráfico* which was not used to produce the embeddings. Then, the news documents were manually labeled with location tags and used for the evaluation. We com-

[5] Nominatim is a search engine to search coordinates by name: https://nominatim.org/.

[6] https://wiki.openstreetmap.org/wiki/Overpass_API.

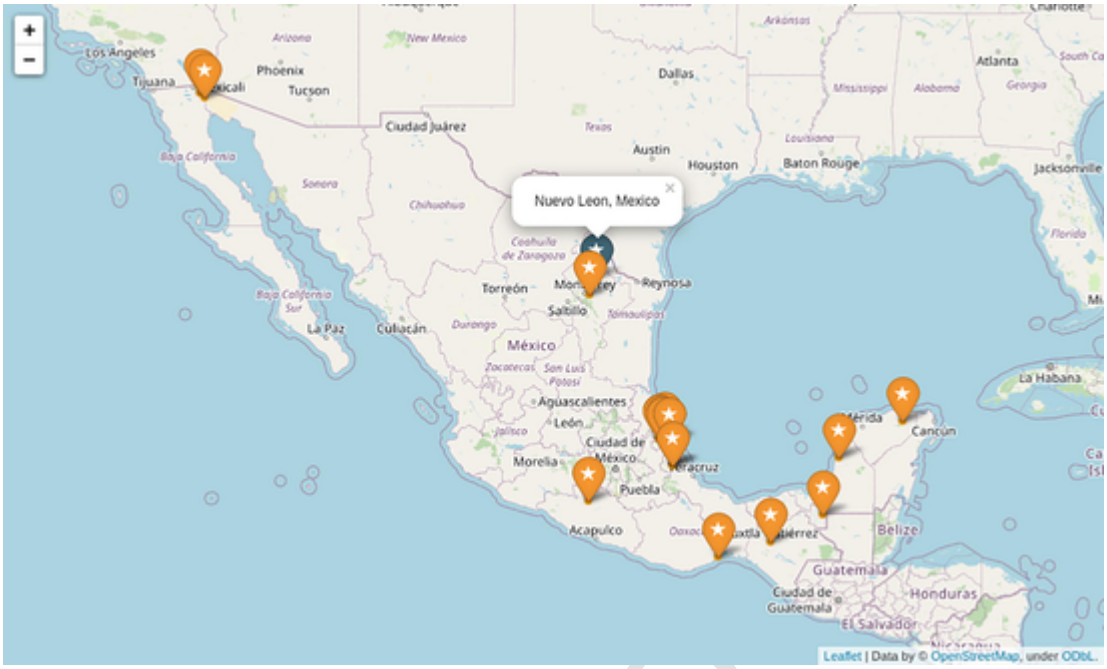

Fig. 1. Geographic Entity Candidates for the toponym Nuevo León. In blue, we show the correct location according to the textual context. The other candidates with the same name are presented in orange. All the entities coordinates were retrieved using Nominatim.

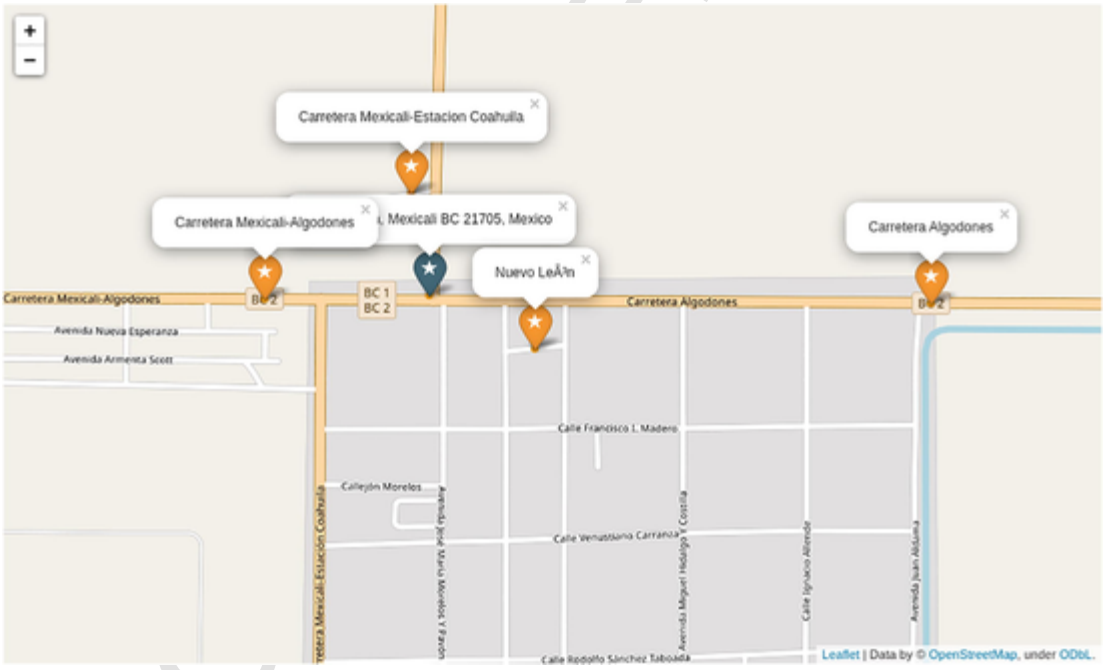

Fig. 2. Points of interests for the entity candidate 2 (*"Nuevo León, Mexicali, Baja California, 21705, México"*) obtained with the Overpass API. With blue color we show the entity candidate and with orange, the points of interests located in an area with a radius of 200 meters.

pared the out-of-the-box NER models from Spacy and OpenNLP (base) and we also create an OpenNLP model trained with the C2 corpus (trained). One important detail is that it is not possible to directly compare the resulting entities because each model has a different way to segment tokens and entities. To solve this problem, the corpus was segmented by unigrams so that words that conform an entity will be taken as entities by themselves. For example, for an entity such as *Estado de México* (State of Mexico), each word spanning the entity (*Estado*, *de* and *México*) should be classified as a location. In doing so, the comparison between models consisted of a token-by-token evaluation. The results are shown in Table 2.

Regarding F-measure, the proposed GNER model outperformed other state-of-the-art NER software. This result indicates that both precision and recall, are well balanced in the GNER model for Mexican Spanish. In contrast, we observe that the OpenNLP base obtained the best precision score while its recall was very low, i.e, is very unbalanced. Regarding to the accuracy, all models performed well. In practical applications, we would possibly be using a combination of models.

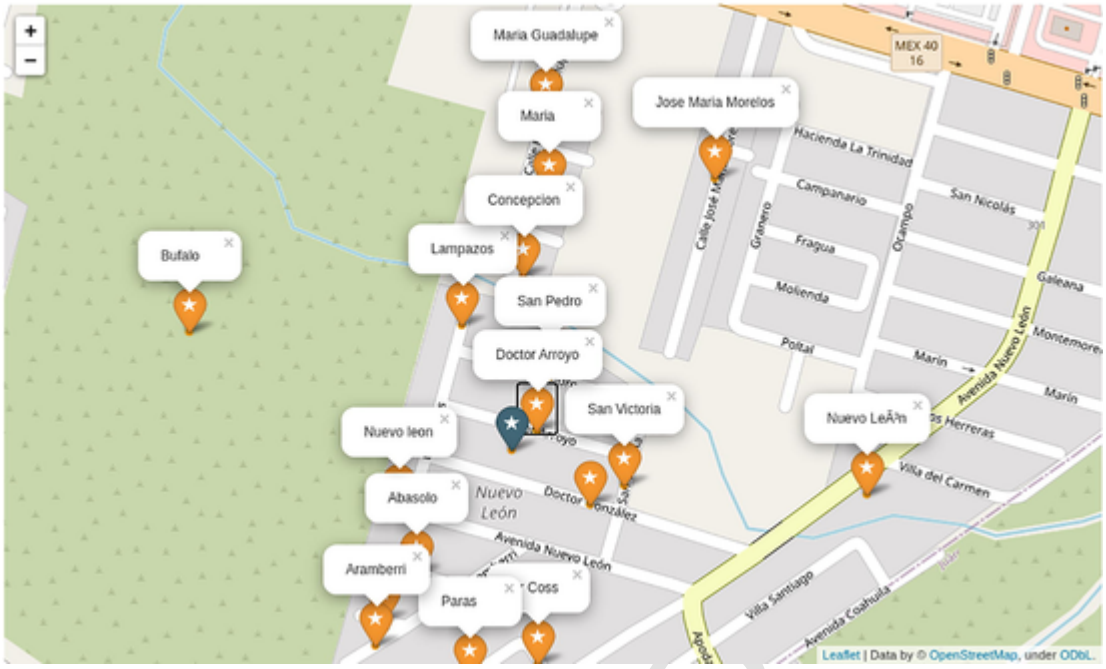

Fig. 3. Points of interests for the entity candidate 6 (*"Nuevo León, Guadalupe, Nuevo León, México"*) obtained with the Overpass API. With blue color we show the entity candidate and with orange, the points of interests located in an area with a radius of 200 meters.

Fig. 4. Points of interests for the entity candidate 12 (*"Nuevo León, Aguacatal, Xalapa, Veracruz de Ignacio de la Llave, 91130, México"*) obtained with the Overpass API. With blue color we show the entity candidate and with orange, the points of interests around it within an area with a radius of 200 meters.

4.2. Results on Geographic Entity Disambiguation

To evaluate the disambiguation proposal described in Section 3.4, we have considered 950 *pseudo-documents* to asses how many times the correct entity was placed on the first place, or the first three places in the ranking.

For the ranking assessment, we have used the Nominatim importance score[7] as the baseline to compare our disambiguation approach against. It is worth to note that Nominatim's ranking method is a high quality baseline since the heuristics used to score entities include lexical similarity (between the query and OSM data) and a location importance estimation called the *importance score* according to its prominence in Wikipedia.

[7] https://nominatim.org/.

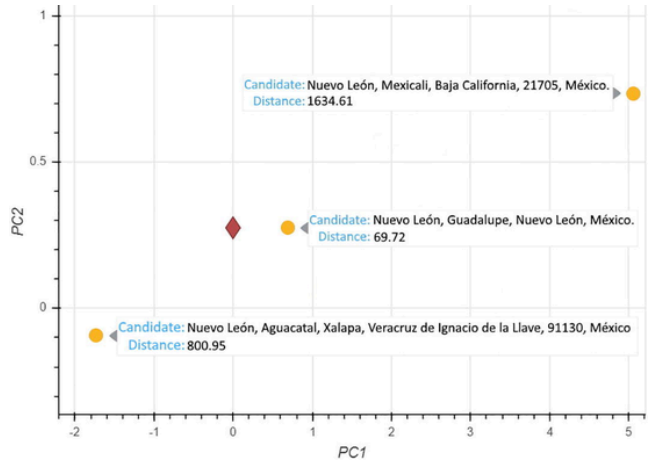

Fig. 5. First two components of PCA obtained from the semantic model learned from news documents and Wikipedia articles. The rhombus represents the original document while the circles represent the *pseudo documents* of the candidate entities.

Table 1
Results of three different word embeddings encoders for Geographic Named Entity Recognition in Mexican Spanish.

	Accuracy	Precision	Recall	F-measure
CBoW	0.9454	0.3953	0.07545	0.5348
ConEc Global	0.9633	0.7085	0.5663	0.8071
ConEc Global & Local	0.9626	0.7055	0.5761	0.8093

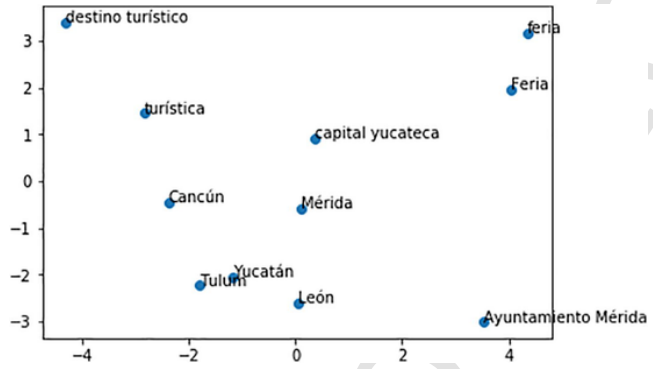

Fig. 6. PCA projection of the 10 most similar word embeddings to "Mérida" using a context encoder with global and local information trained with 165354 news documents from the main digital media in Mexico.

Table 2
Comparison of Geographic Named Entities Recognition models for Mexican Spanish.

	Accuracy	Precision	Recall	F-measure
Proposed GNER model	0.8394	0.5387	0.6759	0.6995
Spacy	0.8086	0.4608	0.4454	0.4529
OpenNLP trained	0.8461	0.6322	0.3332	0.4364
OpenNLP base	0.8523	0.9358	0.1863	0.3108

In Table 3 the results of the evaluation are presented. The first column corresponds to the proportion of ambiguous geographic entities correctly assigned to the first option in the ranking while the second column corresponds to the proportion of entities assigned among the first three options in the ranking. The first row in Table 3 are the score for the Nominatim importance score while the second and third rows

Table 3
Comparison of Geographic Named Entities Disambiguation for Mexican Spanish based on pseudo-documents and word embeddings.

	Correct disambiguation in ranking 1	Correct disambiguation in ranking 1–3
Nominatim	68.65%	88.05%
Proposal with Cosine	53.73%	82.08%
Proposal with Euclidean	74.62%	89.55%

correspond to the score of our approach, described in Section 3.4, using two different metrics to calculate the distance between *pseudo-documents* and original documents. The best results in both categories were for the disambiguation proposal using the euclidean distance. However, there are some aspects about the evaluation that must be discussed here. In the first place we have used only 950 *pseudo-documents* because the evaluation needed the manual creation of a golden standard which required time and a careful data preparation. As far as we know, there is no publicly available data for this task. Another aspect is that, at first, we have considered to report the distance, in kilometers, from the coordinates determined by the proposal to the actual points but the interpretation of this result could be confusing. In general, is quite challenging to define, what the correct location of a place is since geographic entities could have a variety of shapes and sizes: points, lines, multi-lines, polygons, among other shapes cannot be treated with the same criteria for evaluation. For example, observe the case where the entity is a big city and its canonical coordinates are in the geometric center of it. Then, our method assigns a point inside the city but not in the center, say at 2 km far from the center. We will probably say that this is a good result. Now, consider that the entity is a little park with 50×50 m^2 of surface and again our method assigns a point at 2 km far from the center of the park. We will probably say that the prediction is far from the actual coordinates. We concluded that in future research, we have to acknowledge these and other aspects to conduct a robust experimental design considering entity types and distances (in km) between entities.

5. Conclusions

In this paper, we aimed to explore semantic relationships of words and documents in Mexican Spanish in order to find useful patterns that could be used in the geoparsing task. The results indicate that the relationship between the geographic space and the semantic space can be exploited to recognize and disambiguate locations using embeddings from a news corpus. In light of the intrinsic properties of word embeddings, geographic entities can be accurately recognized in news documents in Spanish, for which we have compiled an annotated corpus. The resulting geo-entity recognition accuracy in our experiments is as good as the state-of-the-art software.

According to disambiguation, the general structure of the document plays an important role in the procedure and that enriching the embeddings model with Wikipedia articles is worth. The Wikipedia articles of locations usually contain information about different topics such as culture, economy, education, government, and history. We can take advantage of all this information to enrich the semantic space of word embedding models. While the proposed pseudo documents are not literal descriptions of the surroundings of a location, we are able to semantically relate such pseudo documents with geographic entities. We believe that the semantic relationships could be improved with a larger corpus and with a better description of the surroundings of the geographic entities of interest so we will increase our labeled corpus. In future research, we will also conduct a robust experimental design considering entity types and distances between entities.

CRediT authorship contribution statement

Alejandro Molina-Villegas: Conceptualization, Methodology, Formal analysis, Supervision, Resources. **Victor Muñiz-Sanchez:** Methodology, Formal analysis, Supervision, Resources. **Jean Arreola-Trapala:** Formal analysis, Software, Visualization. **Filomeno Alcántara:** Formal analysis, Software, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by CONACyT-CB-15-252996.

References

Aguirre, E., Alegria, I., Artetxe, M., Aranberri, N., Barrena, A., Branco, A., Popel, M., Burchardt, A., Labaka, G., Osenova, P., Sarasola, K., & Silva, J. (2015). Report on the state of the art of named entity and word sense disambiguation. Technical Report 4, Faculdade de Ciências da Universidade de Lisboa on behalf of QTLeap, Lisboa.

Y., Bengio, R., Ducharme, P., Vincent, & C., Janvin (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155.

A.L., Berger, V.J.D., Pietra, & S.A.D., Pietra (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 39–71.

P., Bojanowski, E., Grave, A., Joulin, & T., Mikolov (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.

S., Cucerzan, & D., Yarowsky (1999). Language independent named entity recognition combining morphological and contextual evidence. In Conference on empirical methods in natural language processing and very large corpora.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota. Association for Computational Linguistics.

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

J., Gelernter, & W., Zhang (2013). Cross-lingual geo-parsing for non-structured data. Proceedings of the 7th workshop on geographic information retrieval (pp. 64–71).

Goldberg, Y. (2015). A primer on neural network models for natural language processing. CoRR, abs/1510.00726.

Horn, F. (2017). *Context encoders as a simple but powerful extension of word2vec*. arXiv preprint arXiv:1706.02496.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Xing, E.P., & Jebara, T., (Eds.), *Proceedings of the 31st international conference on machine learning, volume 32 of Proceedings of machine learning research* (pp. 1188–1196). Beijing, China. PMLR.

T., Mikolov, K., Chen, G., Corrado, & J., Dean (2013). Efficient estimation of word representations in vector space. In Y., Bengio, & Y., LeCun (Eds.), 1st International conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings.

E., Aldana-Bobadilla, A., Molina-Villegas, I., Lopez-Arevalo, S., Reyes-Palacios, V., Muñiz-Sanchez, & J., Arreola-Trapala (2020). Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. Remote Sensing, 12(18), 3041. doi:doi.org/10.3390/rs12183041.

L., Moncla, W., Renteria-Agualimpia, J., Nogueras-Iso, & M., Gaio (2014). Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 183–192).

J., Pennington, R., Socher, & C.D., Manning (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. arXiv preprint arXiv:1802.05365.

T., Rocktäschel, M., Weidlich, & U., Leser (2012). Chemspot: A hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633–1640.

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*. Lisbon, Portugal.

J., Serrà, & A., Karatzoglou (2017). Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks. Proceedings of the eleventh ACM conference on recommender systems. New York, NY, USA: Association for Computing Machinery.

M.J., Silva, B., Martins, M., Chaves, A.P., Afonso, & C., Nuno (2006). Adding geographic scopes to web resources. Computers, Environment and Urban Systems, 30(4), 378–399.

D.A., Smith, & G., Crane (2001). Disambiguating geographic names in a historical digital library. Research and advanced technology for digital libraries (pp. 127–136). Springer.

Smith, D. A., & Mann, G. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references* (pp. 45–49).

N., Sobhana, P., Mitra, & S., Ghosh (2010). Conditional random field based named entity recognition in geological text. International Journal of Computer Applications, 1(3), 143–147.

L., Tanabe, N., Xie, L.H., Thom, W., Matten, & W.J., Wilbur (2005). Genetag: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics, 6(1), 1.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). 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 – Vol. 4, CONLL 03* (pp. 142–147). Stroudsburg, PA, USA: Association for Computational Linguistics.

N., Wacholder, Y., Ravin, & M., Choi (1997). Disambiguation of proper names in text. Fifth conference on applied natural language processing (pp. 202–208).