A systematic approach to literature analysis: traveling through stories

Ryusei Uenishi, Claudio Ortega, Ángel Pérez Martinez, Michelle Rodríguez-Serra and Paula Elías

Universidad del Pacífico

Abstract

Travel literature has captured humanity's imagination ever since the emergence of famous works such as *The Wonders of The World* by Marco Polo and *The Journal of Christopher Columbus*. Authors in this genre must process large and diverse volumes of data (visual, sensory, and written) obtained on their trips, before synthesizing it humanly in such a way as to move and communicate personally with the reader, without losing the factual nature of the story. This is the ultimate goal of the natural language processing (NLP) field: to process and generate human–machine interaction as naturally as possible. Hence, this article's purpose is to analyze and describe a nonfictional literary text, which is a type of documentary text that contains objective, qualitative, and quantitative information based on evidence. In this analysis, traditional methods will not be used. Instead, it will leverage NLP techniques to process and extract relevant information from the text. This literary analysis is a new kind of approach that encourages further discussions about the methodologies currently used. The proposed methodology enables exploratory analysis of both individual and unstructured corpus databases while also allowing geospatial data to complement the textual analysis by connecting the people in the text with real places.

Correspondence: Ryusei Uenishi, Jr. Gral. Luis Sánchez Cerro 2141, Jesús María, Lima, Perú.
E-mail: rm.uenishik@alum.up.edu.pe, c.ortegaariza@up.edu.pe

1. Introduction

Travel literature has captured humanity's imagination for years. Books such as *The Wonders of The World* by Marco Polo, *The Journal of* Christopher *Columbus*, and *The First Part of the Chronicle of Peru* by Pedro Cieza de León, among others, have been analyzed from various perspectives to understand the richness of their content. Authors of travel narratives must process the diverse and voluminous data (visual, sensory, and written) they obtain on their trips, and then synthesize it humanly without losing the factual nature of the story. This descriptive ability is one of the characteristics of this particular literary genre. Authors achieve such a level of complexity that they manage to move and communicate personally with the reader (Pérez Martínez, 2013). As discussed in the literature, the ultimate goal of natural language processing (NLP) is to process and generate human–machine interaction as naturally as possible (Manning and Schutze, 1999). Hence, in this article, we present travel narratives as an ideal use case for NLP, as well as a use case test. We suggest a tool for the literary analysis of Spanish-language travel narratives using NLP techniques and, primarily, a named entity recognition (NER) model; this tool enables both exploratory macro and micro analysis of unstructured corpus databases. In the case of the macro analysis, a map is shown to identify the studied area and make

doi:10.1093/llc/fqab081

connections with current places in the zone. Also, the user can select the top *n* most frequent part-of-speech (POS) sequence patterns, these patterns allow to understand the author's particular phrases compositions. In the other case, the microanalysis is applied to a user-selected book passage, on which you can visualize the model outputs (identified person and locations). This kind of tools can be helpful to urban planning since it allows knowing traditions, places, and monuments that are disappeared from cities. Authors like Lewis Munford (Mumford, 1938) invite the reader to think about cities further than the automotive paradigm. In order to do that it is necessary to remember the past, and history and literature are interesting tools. The so-called pre-statistics era lacks quantitative information that can be completed thanks to this type of text and analysis. The importance of this type of knowledge has been pointed out by several classical authors such as Ezra (1952), Pirenne (1969), or Sjoberg (1960).

In recent years, there have been numerous advances in models that allow the automation of tasks such as text classification, text generation, and machine translation (MT) (Otter *et al.*, 2020). These are just some of the potential practical applications from the field of NLP, which has found a successful formula in the use of deep neural networks (Belinkov and Glass, 2019). Building on the rediscovery of artificial neural networks—specifically of the recursive type—NLP researchers have continuously achieved results that advance the state of the art. Some widely used model architectures are recurrent neural networks, long short-term memory networks (LSTMs), gated recurrent unit networks (Sutskever et al., 2014), and, lately, transformers (Devlin *et al.*, 2019). These models have applications in the core areas of NLP, such as language modeling, morphology, syntax, and semantics (Otter *et al.*, 2020). The progress in the field is mostly related to improvements in algorithms, neural network architectures, and large annotated training data sets (Goodfellow *et al.*, 2016). Despite this, it is necessary to continue examining and validating the evaluation metrics of NLP models, not only using data-based indicators but also complementing the analysis with real use cases (Reiter, 2018). In this context, Digital Humanities emerges as a particularly interesting field for NLP applications, as one that, unlike Computational Linguistics, has not yet been subject to NLP techniques based on neural networks. A new field of study that seeks to analyze writing style computationally is Stylometry. Its main use is authorship attribution (Eder *et al.*, 2016). Nonetheless, current NLP techniques for literary analysis, in both near and far reading, have a much greater scope than just the identification of authorship. Therefore, it is necessary to illustrate the different forms of collaboration that could be pursued to generate greater scientific knowledge in both the NLP and Digital Humanities. However, it is not enough to apply these techniques to the available digitized texts. Travel narrative resides in the unexplored intersection between the two areas. These texts describe real-life travels in a factual way, in relation to the events experienced by the author. As such, they resemble what is known as history or documentary (Alburquerque García, 2011a,b). Therefore, to give a demonstration of the scope and potential of NLP, we present a use case based on the Pedro Benvenutto Murrieta book *Quince Plazuelas, Una Alameda y Un Callejón. Lima en los Años de 1884 a 1887. Fragmentos de Una Reconstrucción Basada en la Tradición Oral*, in which we verify the usefulness of the main NLP tools, as well as some of its applications in automating and supporting the literary analysis of travel narratives.

The proposed methodology is interdisciplinary. From the travel literature area, the theoretical platform proposed by Luis Alburquerque has been used, where he shares a series of recommendations regarding the travel narratives analysis (Alburquerque García, 2011a,b). The analyzed text (Benvenutto, 2003) has no records in the Peruvian bibliography and was not classified. At the end of the 19th century, the author gathered testimonies from the city's inhabitants. The text could be a chronicle or a popular traditions compilation, but it also fits as a travel narrative. Other authors who have collected the city traditions are Ricardo Palma, Acisclo Villarán, or Ismael Portal, yet Benvenutto provides rigorous documentation, interviews, and descriptions. In addition to these findings, the author also looked for information in ancient texts, engravings, and guides. He is not only a memorialist as José Jiménez Borja has called him (Borja, 1978), but an author who looks at modern urbanism with concern. Thus, he tried to send a message to future generations so they not forget about a human-scaled city. By suiting within the travel narrative paradigm, the original spirit of his work can be rescued, while its objective tendency can be turned into data and digital quantifications. Many other texts that have been listed on literary shelves were not able to be interpreted in all their richness. As such the modeling process is very suitable for this type of work. The proposed modeling process is intended to accommodate each author's own style, which is especially useful in the case of travel literature. Specifically, the results and interpretation of the model take into account the author's idiolect, uses of language, idioms, and even redundancies, as part of the search for the details and data that the text can offer us.

The remainder of the article is organized as follows: In Section 2, we present a literature review related to the use of computational techniques for text analysis, focusing on literary analysis. In Section 3, we describe each stage of the proposed methodology and its scope. In Section 4, we review the results obtained in the use case. Finally, we present our conclusions and discuss some technical and practical implications of this study.

2. Literature Review

This section is divided into a review of three groups of studies: (1) studies that have applied computational NLP techniques to analyze documents, (2) studies that have specifically applied NER, and (3) studies related to urban location mapping. Unfortunately, to the best of our knowledge, yet there are no academic studies that have applied NLP techniques to the analysis of travel narrative literature.

2.1 Natural language processing

Some articles on NLP are strongly based on text mining and feature engineering (that is, preprocessing). One example is Vani and Gupta (2017), who aimed to explore the power of syntax-based linguistic features extracted using shallow NLP techniques for plagiarism detection. Their results showed that the use of linguistic features empowers the classification of complicated cases of plagiarism. In another study, Ashraf *et al.* (2016) used stylometry to detect author traits (gender and age) for cross-genre author profiles. The results showed that a combination of different types of stylometric features, including lexical, syntactic, vocabulary richness, and character-based features, is helpful in identifying the age and gender of an author from his/her written text. Finally, Koto and Adriani (2015) proposed POS sequences as a feature for analyzing patterns or word combinations of tweets in two domains of sentiment analysis: subjectivity and polarity (Koto and Adriani, 2015). Using POS sequences to classify feelings tested in datasets and results, they obtained an increase in the certainty of each set, in comparison to the classification without the POS sequence.

Recent research on deep learning techniques to process raw text data (that is, without ad-hoc preprocessing) has aimed to learn internal representations by leveraging vast quantities of textual inputs. Sutskever *et al.* (2014) presented a general end-to-end approach to sequence learning that makes the fewest possible assumptions on the sequence structure. They used a multilayered LSTM method to outline the input sequence to a vector of fixed dimensionality, and later they used another deep LSTM to calculate the target sequence from the vector. The results showed that if there is a sufficient volume of training data, the LSTM-based approach to MT should perform well on other sequence learning problems. In addition, Wang *et al.* (2017) presented gated self-matching networks for question-answering based on a reading comprehension style. They showed that with an increase in length, the behavior remained stable. The clear fluctuation in the longer passages and questions was primarily because the proportion was too small. The authors stated that their model is broadly skeptical of long passages, and that it focuses on the important part of the passage. Finally, Devlin *et al.* (2018) improved the fine tuning-based approaches by proposing bidirectional encoder representations from transformers, which use masked language models to facilitate pre-trained deep bidirectional representations. It is important to mention that most of these large pre-trained models are open-source and available to the community through frameworks such as Spacy (Honnibal *et al.*, 2020) or Hugging Face Transformers (Wolf *et al.*, 2020).

2.2 Named entity recognition

Some studies in the field of NER have focused on the identification of entities in historical texts. For example, Borin *et al.* (2007) proposed a rule-based NER system for Swedish literary texts from the 19th century. Their system obtained an F1 score of 92.8%, but with errors in the identification of entities with complex structures. In addition, Iglesias Moreno *et al.* (2014) proposed an NER model based on the Freeling tool for Spanish texts from the Middle Ages. They obtained good results in identifying entities with simple structures, but not for entities with complex structures. Another notable study is Won *et al.* (2018), who evaluated NER tools for the extraction of geographical information from historical texts. The data they used for their research are the Mary Hamilton Papers and the Samuel Hartlib Papers. The results show that the best results were achieved by the Polyglot model in the case of the Hamilton dataset (with an F1 score of 61.1%), and the Stanford NER model in the case of the Hartlib dataset (with an F1 score of 70.8%). The low results are due to the spelling differences between the entities in the two datasets. Furthermore, studies on Spanish-language NER in recent years have been related to biomedical texts (Akhtyamova, 2020; Rivera-Zavala and Martineza, 2020). For example, Cotik *et al.* (2018) presented an NER method for poorly resourced languages. The authors applied this method to radiology reports and obtained better results than the state of the art, with an F1 score of 66.21%. Díez Platas *et al.* (2020) presented a tool to recognize entities in Spanish texts dating from the 12th to the 15th centuries, as well as new entities for medieval texts. The authors designed a tool using language characteristics and identification rules given the difficulty in identifying entities with complex structures and a lack of normalization in the expressions. Their model obtained an overall F1 score of 77% across all entities and can be applied to texts that have not been preprocessed. Turning back to travel literature, each text has its own style. The knowledge obtained about these styles through an NER model would allow analysis of the information they offer.

2.3 Urban location mapping

In recent years, frequently social networks have been used as data sources for applied NLP studies. Furthermore, sentiment mapping has been one of the most used techniques in studies regarding green spaces impact in cities (Lim *et al.*, 2019; Plunz *et al.*, 2019), citizens' happiness mapping (Mitchell *et al.*, 2013; Alharbi *et al.*, 2018), disaster resilience (Han and Wang, 2019; Yao and Wang, 2020), and human-aware smart cities (De Oliveira and Painho, 2021). Also, some studies provide methodologies and use cases for mapping places using historical travel texts, travelogues, or travel literature as data sources (Rupp *et al.* 2013; Afinoguénova *et al.*, 2020). These methodologies, as in the present work, use geocoding services. In addition, NLP techniques have also been used to amplify the people's voices in the urban planning process through digital platforms for citizen engagement where they receive opinions and comments to be processed by intelligent systems (Deitz *et al.*, 2018; Sayah and Schnable, 2019).

Several authors have acknowledged the importance that heritage and historic urban areas have in order to improve future urban planning. Abbot and Adler (1989) argue that there existed a need for including historical records within the planning practice. Stanley *et al.* (2012) stated that the flow of information and material among people interacting in open spaces represents a fundamental dimension of cultural, political, and economic life from early civilizations to the present. They proposed a typology that consists of seven categories. Specifically, Plazas are one of the open urban space categories that are poorly documented. In this article, use case test, we intent to extract information from Lima's Plazas in a semi-automated way. Finally, Giannopoulou *et al.* (2014) proposed a specialized Geographic Information System to model the urban process and its impact on heritage regions. The results show four distinctive urban environments within the Old Town of Canthi, an urban civilization from the 19th century in Northern Greece. In Section 3, the methodology proposal defines a standardized way to extract historical information from travel literature texts leveraging NLP and mapping techniques.

3. Methodology

The proposed methodology fills a gap in the available tooling and techniques used to analyze travel stories. To the best of our knowledge, there is currently no travel story exploration tool where you can interact with the texts and increase the reading experience by enriching the information using both NLP techniques and geographic data. As such, this article works

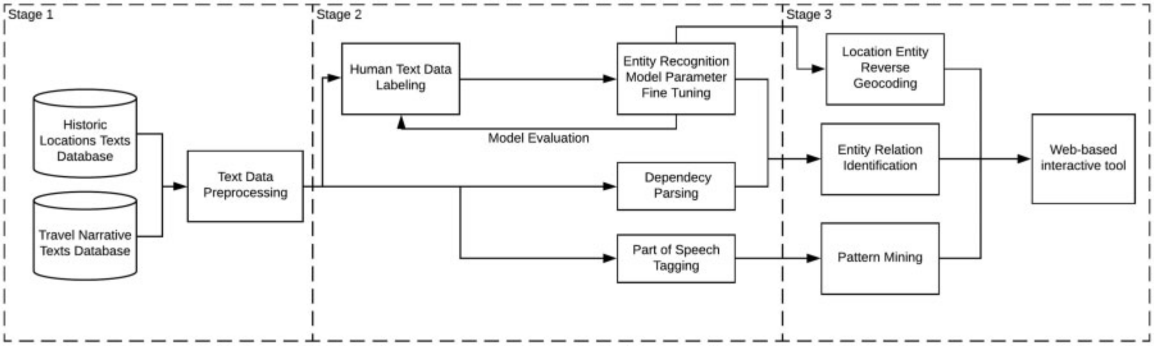

Stage 1
Stage 2
Stage 3
Historic Locations Texts Database
Travel Narrative Texts Database
Text Data Preprocessing
Human Text Data Labeling
Entity Recognition Model Parameter Fine Tuning
Model Evaluation
Dependecy Parsing
Part of Speech Tagging
Location Entity Reverse Geocoding
Entity Relation Identification
Pattern Mining
Web-based interactive tool

Fig. 1 Proposed methodology diagram divided by stages (columns) and tasks (boxes and arrows)

toward setting a standard framework for the macro and micro analysis of literary texts of travel stories that allow holistically understanding the different aspects of the text: location, characteristics of the descriptions, important characters, among others. In Fig. 1, we present a visual representation of each stage of the proposed methodology in detail. Stage 1 consists of the data acquisition and annotation processes. Stage 2 entails the fine tuning and evaluation of the pre-trained Spanish entity recognition model. Finally, Stage 3 is the description of the results, presented in the form of ranking, maps, and interactive visualizations.

In Stage 1, the data collection process can be carried out using one or many Spanish digitized books from the travel narrative literary genre as a primary source. The plain text format is sufficient. Then, according to the place and time of the main text source, the complementary data are identified and collected. Some examples of complementary data are historical dictionaries, curated and historical place references, among others. These complementary data allow one to find the current address of the place mentioned in the texts or books. Usually, in the NLP field the first step is to divide the analyzed text into units called tokens which normally consist of words, punctuation, or numbers. This step is also known as tokenization (Manning and Schutze, 1999). Thereafter, the data pre-processing step consists of two main tasks. First, sentence tokenization is applied to the corpus. Then, each sentence is labeled with its parent chapter.

The process of automatically finding entities in the text, such as people, locations, or organizations, through a statistical model is called NER. In Stage 2, the clean text data is human-annotated to improve the performance of the NER task, especially for the person (PER) and location (LOC) entities. This process requires a person to read the text and indicate, for each sentence, the start and end character position of one or more identified entities. Subsequently, the Spanish pre-trained NER model is fine-tuned using the labeled dataset. Hence, the dataset is divided into three samples: training, validation, and testing. The training and validation sets are used to adjust the model parameters to minimize the NER error by preventing overfitting of the model to the training set. Then, the model is evaluated using the test set, where the goal is to improve performance in comparison to the raw pre-trained model. Finally, the fine-tuned model is used to process the complete corpus; the results obtained consist of NER tags. Also, a Spanish-language trained model of a POS and dependency parser is applied to obtain the tags for the corpus data for use in the next steps.

The first step of Stage 3 consists of the location entity reverse geocoding, performed in a semi-automated fashion. First, with the locations entities already identified in the text, a dictionary of the locations and their current address is manually built using the complementary data. Variations of names in locations such as ‘Plazuela de Micheo’, ‘Micheo’, and ‘La Micheo’ are considered as one. Second, an online geocoding API is used to obtain the location latitude and longitude values for each location entity in the dictionary. Then, the resulting spatial data is used to build a location map for the final web application. In turn, the goal of the entity relation identification task is to automatize the identification of the location–person entity relationships described in the travel narrative text. For this purpose, using dependency parsing, two main identification rules are taken into account: (1) the entities must have a subject–object relation and (2) the entities must be less than a *d* word token distance, here *d* is an user-defined parameter, because the writing styles of different authors have a strong relation with the text distance that a subject–object relation can span.

To analyze the text structure, we used POS tagging which is the process of labeling each word or token with its corresponding tags such as noun, verb, adjective, preposition, or others (Manning and Schutze, 1999). Then, the pattern mining process is performed using the PrefixSpan algorithm (Han *et al.*, 2001) with the goal of automatically extracting the POS tag sequence patterns present in the input data. Besides, POS tags of the input data are encoded into numerical labels and a minimal support parameter is defined to obtain the top *n* most frequent sequence patterns. With the results obtained from the algorithm, the most frequent patterns are shown in the application. This pattern can be used in literary analysis to compare the author's narrative styles.

Finally, an interactive web-based tool is built. This tool has two sections: the first section contains a map generated using each georeferenced location and its related person entities, and the POS tag patterns ranking ordered by frequency. The second section allows the user to select a specific passage of the travel narrative text and analyze it using the trained model outputs. The NER tags are added to the selected text, and the dependency parsing and POS tags are shown using a nested-tree structure. As such, this tool allows the user to macro-analyze the complete travel narrative text input or focus on one specific passage. Furthermore, NLP techniques are essential for geographic information retrieval. On the degree of effectiveness of these techniques, we have the work of (Stokes et al., 2008). One of the ideas is the comparison between human effectiveness and that of NLP (Florian et al., 2003). The novelty here would be the use of old topomines within a story from the past. The peculiarity of geographical names can be cultural or chronological. The first has been dealt with in works such as those of (Hu et al., 2019) but it is necessary to delve deeper into the differences between place names over time. We believe that our work can be a contribution in this regard.

4. Use case test1

In this section, we present the use case results[1]. As noted earlier, the travel literature book we analyze is *Quince Plazuelas, Una Alameda y Un Callejón. Lima en los Años de 1884 a 1887. Fragmentos de Una Reconstrucción Basada en la Tradición Oral* de Pedro Benvenutto. The book presents stories related to certain places in the city of Lima, Peru, between 1884 and 1887, and describes experiences about the city's different squares. Furthermore, the book was in PDF format so had to be converted to plain text. Then, in the preprocessing stage, we obtained 2,049 sentences after the sentence tokenization process. In the next stage, we labeled 1,655 entities in the text, which resulted in 667 locations and 988 person entities. The labeling consists of manually annotating the type and position of the entities within the text. Then, using the annotated data, we performed the fine-tuning of the Spacy Spanish-language NER pretrained model. This model is a multi-task convolutional neural network (CNN) of the Spanish language, trained using the UD Spanish Ancora and WikiNER news-related corpus. The model performance scores reported by Spacy are presented in Table 1 which consists of two parts. On the one hand, Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS), and POS tagging accuracy (TAG) metrics help evaluate the model syntax accuracy. Regarding dependency parsing, LAS considers the percentage of words with both the correct syntactic

Table 1. Reported Spacy pre-trained Spanish multi-task CNN model syntax and NER accuracy metrics

Syntax accuracy (%)		**NER accuracy** (%)	
LAS	88.67	NER F	89.84
UAS	91.6	NER P	89.96
TAG	97.54	NER R	89.71

Source: Spacy models Spanish accuracy evaluation.

head (i.e. word relations) and the correct label (subject, object, punctuation, and others) (Nivre and Fang, 2017). UAS is equal to the percentage of words that get the right head. At last, TAG is the percentage of word tokens with the correct label (noun, verb, and others). On the other hand, the F-score (NER F), Precision (NER P), and Recall (NER R) metrics help evaluate NER accuracy. As such, Precision is the model's ability to avoid false positives outputs. Also, Recall is the percentage of words assigned to the correct entity label. Finally, F-score is the harmonic mean of the two previous metrics.

A machine learning algorithm is defined as an algorithm capable of performing better at a task with respect to the amount of data (experience) it has seen (Goodfellow *et al.*, 2016). In this case, the algorithm used is a multi-task CNN. Here, we are trying to improve the named entities recognition such as person and location, and the data are Benvenutto's book, the type of entities within the text, and its location. Our model acquires experience through a process called training. It consists of an iterative gradient-based error minimization algorithm that uses the input data to generate a model output and compare it to the expected output and correspondingly adjust the model internal parameters. This comparison is called validation if it is done during the training phase or evaluation if it is done after the training using an 'out-of-sample' dataset, also known as test dataset. Also, model performance on validation and evaluation sets should be close to the training set to avoid underfitting (i.e. training error higher than validation or test error) and overfitting (i.e. training error lower validation or test error), which, in any case, indicates poor model generalization (Goodfellow *et al.*, 2016). Given this, first, we separated the data into training, validation, and testing with the proportions of 70, 10, and 20%, respectively. Then, we deactivated the POS and dependency parser parts for NER model training. Later, we tested different combinations of hyperparameters, which are user-determined values outside the learning algorithm, for example, the number of *epochs* (i.e. one epoch is equivalent to a complete cycle through the input data). The best-performing model had 350 epochs, an incremental batch size in the range of 4–32, and a dropout rate of 0.2. The dropout rate determines the percentage of neurons that are randomly eliminated from the model in the training phase to improve the model generalization. The model training and validation loss curves are presented in Fig. 2. It can be observed that the training loss curve is above the validation loss curve (which means that overfitting is avoided) and that both loss curves converge before fifty epochs. We also evaluated the accuracy of the pre-trained and the fine-tuned models in order to determine the gained performances. The results are shown in Table 2, where it can be observed that fine-tuning helped to improve the Precision (NER P) and F-score (NER F) metrics in the validation and test sets.

After fine-tuning the NER model, we present an example of the model's application; Fig. 3 shows a sentence from the book in which the identified entities are highlighted (places in orange and people in gray). In this case, the identification of entities allows one to analyze the relationships between people and places. In the example, two places and one person are displayed. The person has a relationship with the 'Alameda de los Descalzos' due to having brought some sculptures there. However, the person has no relationship with the second place mentioned; in the next step we address this challenge.

Figure 4 presents the dependency parsing and POS tags within a tree structure in which the leaves and their color represent the words and their POS tags, respectively. The branches represent the dependency relations. This example shows that the verb in which the sentence is rooted can be identified, the subject is 'el tranvía', the object is 'nos', and the circumstantial complement is made up of the root 'mismita'. These graphs can help researchers to understand the syntactic structure of a sentence visually, but more importantly the tree structures can be leveraged to automatically extract meaningful location–person relationships. Moreover, these structures can aid automatic identification of author styling using methods such as pattern mining.

Accordingly, we processed the words and their POS tags using the PrefixSpan algorithm (Han *et al.*, 2001) to extract the most frequent POS tag sequence. We configured the algorithm with a minimum support of 0.45 since this number is small enough to capture a large set of patterns but not too small to compromise and increase of runtime (Han *et al.*, 2001). Table 3 shows the top five most common sequence patterns obtained from POS tags, and their highlighted examples. As mentioned, these patterns

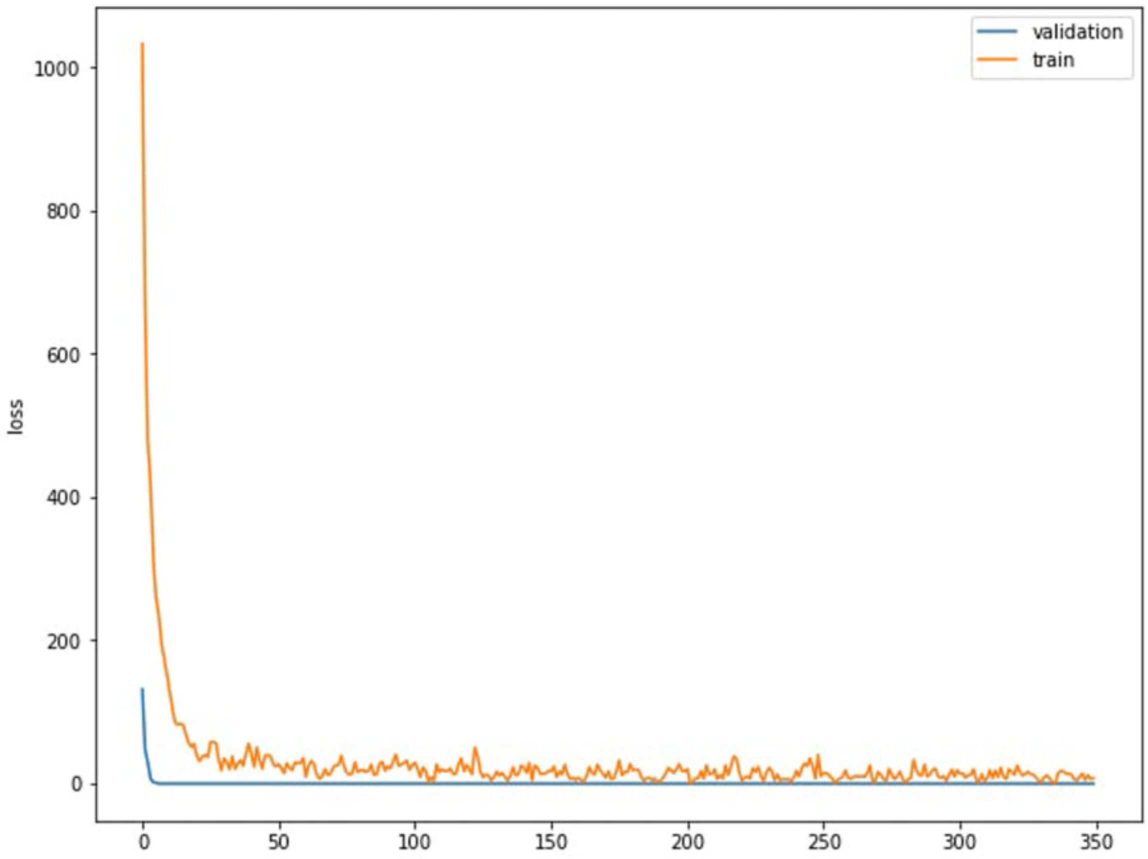


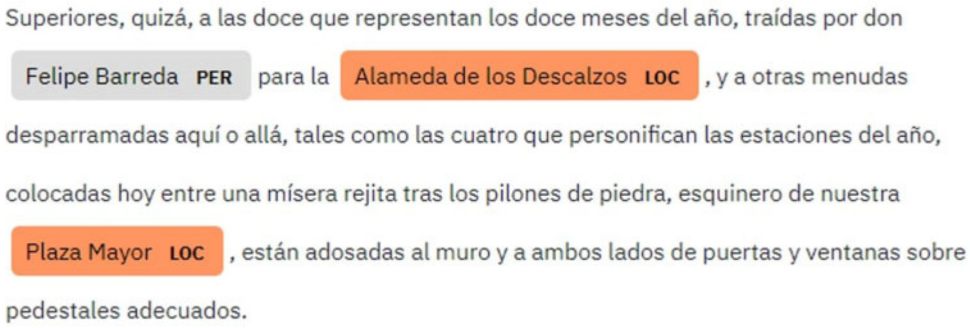
Fig. 2 The model training and validation loss curves. Loss value is presented in the Y-axis and the epoch number in the X-axis

Table 2. Comparison of the NER accuracy metrics for model applied the analyzed text before and after fine-tuning. Best values between pre-trained and fine-tuned test results are in bold.

Metric (%)	Pre-trained model	Fine-tuned model		
		Training	Validation	Testing
NER P	67.87	91.02	88.65	**87.99**
NER R	**90.44**	92.51	89.47	89.06
NER F	77.55	91.76	89.06	**88.52**

represent the syntactical features of the author's writing style. Based on the patterns found, the sequences are part of a sentence, and these structures are more frequent because the sentences displayed tend to be long. In addition, the patterns give an idea about the author's writing style, which is descriptive, in that the author gives many details about the story. The patterns also show the number of subordinate clauses found in the text.

We performed the location entity reverse geocoding process to obtain the geographical location and a dictionary. Then, to be able to apply the entity relation identification process, we selected a maximum word distance between entities. Since 60% of the text sentences have a maximum length of forty-five words, the maximum word distance parameter chosen was fifteen (see Fig. 5) to identify entity relations within the same context and avoid mismatches with non-related entities. Future studies could propose a metric to determine the optimal maximum word distance using a supervised dataset where the relationships



Superiores, quizá, a las doce que representan los doce meses del año, traídas por don Felipe Barreda PER para la Alameda de los Descalzos LOC , y a otras menudas desparramadas aquí o allá, tales como las cuatro que personifican las estaciones del año, colocadas hoy entre una mísera rejita tras los pilones de piedra, esquinero de nuestra Plaza Mayor LOC , están adosadas al muro y a ambos lados de puertas y ventanas sobre pedestales adecuados.

Fig. 3 Example paragraph, extracted from the analyzed text, annotated with the named entities found by the model. PER refers to person entity and LOC refers to location entity

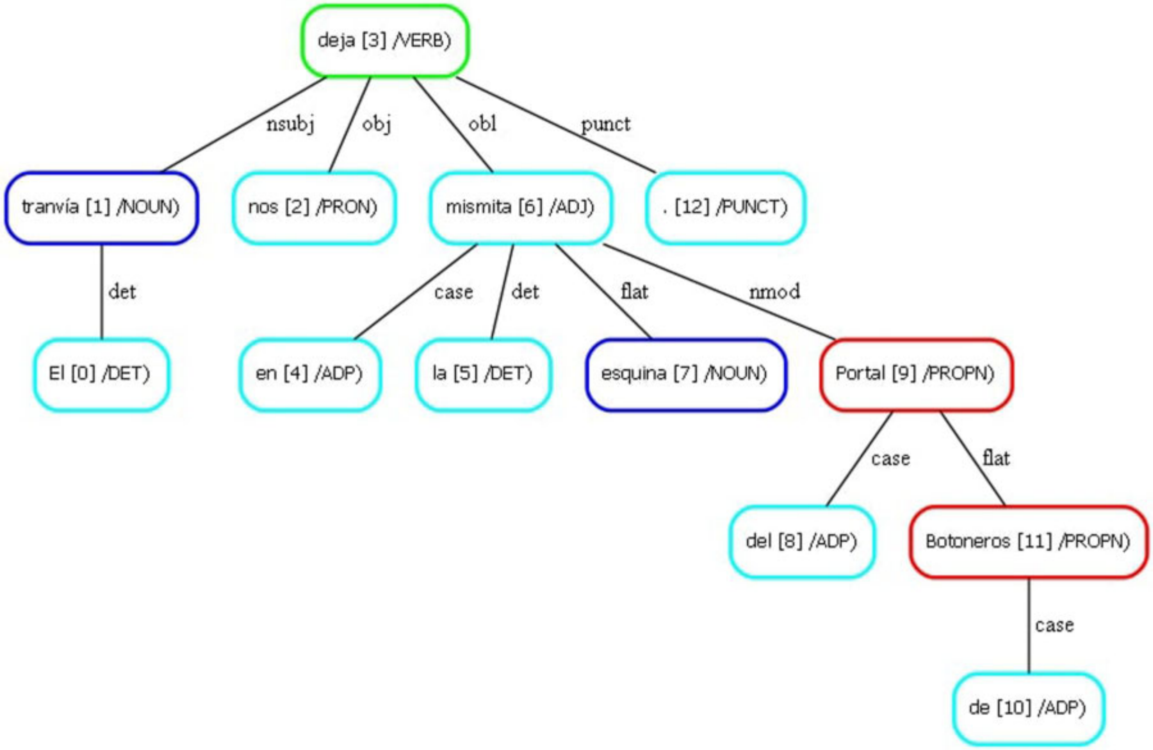


Fig. 4 Tree structure showing the model dependency parsing and POS tagging results using as input an example sentence extracted from the analyzed text

are known *a priori*. Furthermore, a machine learning model to automate this task could be proposed.

It is important to note that entities with compound names were counted as a single word when calculating the distance, to avoid overestimation of entity distance. Also, to identify entities relationships the corresponding dependency parsing labels had to be apposition (aposs), nominal subject (nsubj), or object (obj) in the case of PER entities, and nominal modifier (nmod), nsubj, or obj in the case of LOC entities.

Table 3. Five most frequent patterns present in the analyzed text using the PrefixSpan algorithm and an example sentence that contains the corresponding pattern (highlighted)

Frequent patterns	Example
ADP, DET, NOUN, ADP, DET, and NOUN	'A las ocho de la noche, un corneta de Santa Catalina sale a la puerta del fuerte y llama a las tropas a pasar lista'.
DET, NOUN, ADP, DET, NOUN, and ADP	'El río Huática sigue su curso a lo largo de la plazuela, los viejos y ruinosos balcones de la casa de Gómez Sánchez dan sobre él'.
VERB, DET, NOUN, ADP, DET, and NOUN	'El paseante limita su visita hasta la capilla de San Francisco de Paula el viejo, tal vez hasta la portada de Guia, y de estos lugares se regresan al carro que ha de volverlo al centro de la ciudad'.
DET, NOUN, ADJ, ADP, DET, and NOUN	'La vida nocturna de la plazuela queda descrita con el café Maximiliano en donde se resume toda ella'.
NOUN, ADP, DET, NOUN, ADP, and NOUN	'Los dueños convidan a los visitantes vasos de chicha pintorescamente llamados orines del Niño y estos agradecidos por la atención depositan una moneda en el platillo puesto con ese fin sobre una columnita de madera'.

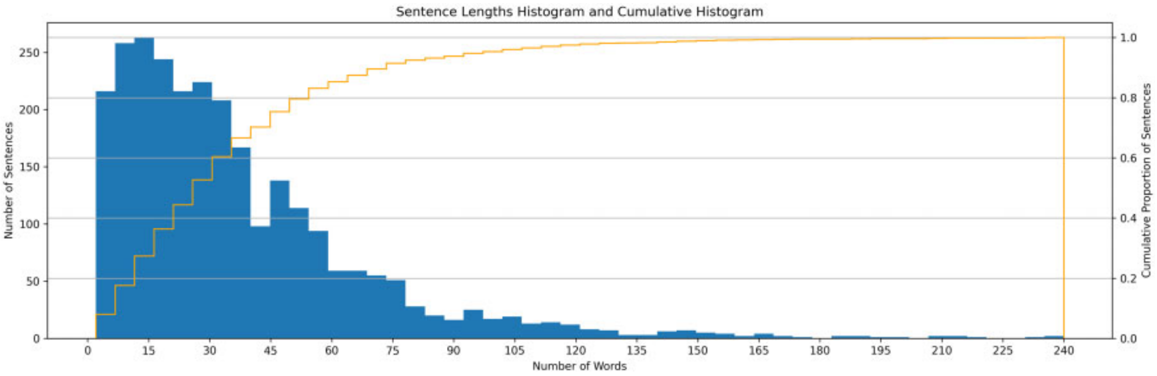


Fig. 5 Maximum Word distance selection using Sentences Word length histograms

Table 4. Four examples of the identified location–person entities relations

Location	People	Current location	Total people
Lima	José Santos, Don Carlos, el vecino Alty, Gaillour, Justiniano Alvarez, José Antonio, José, Pepito Gálvez	Lima	8
Casa de Boza	Don Carlos Paz Soldán, Pedro Paz Soldán, Manuel González de la Rosa, Mariano Felipe	Jirón de la Unión Cuadra 8	4
Plazoleta de la Merced	el sueco Young, Malmborg	Plazoleta de la Merced	2
Plazuela de la Micheo	Chávez, Manuel Gálvez	Jirón de la Unión Cuadra 10	2

Table 4 shows four examples of the obtained entity relations, as well as the locations, its related people, the total number of people related to it, and their current location. In total, we found 190 entity relationships. The entity with the highest number of relationships is Lima. The extracted relationships help us to gather information faster and also to discover the people that are related to those places. This can give us a clearer understanding of which people are related to each place. In these examples, spaces such as 'Casa de

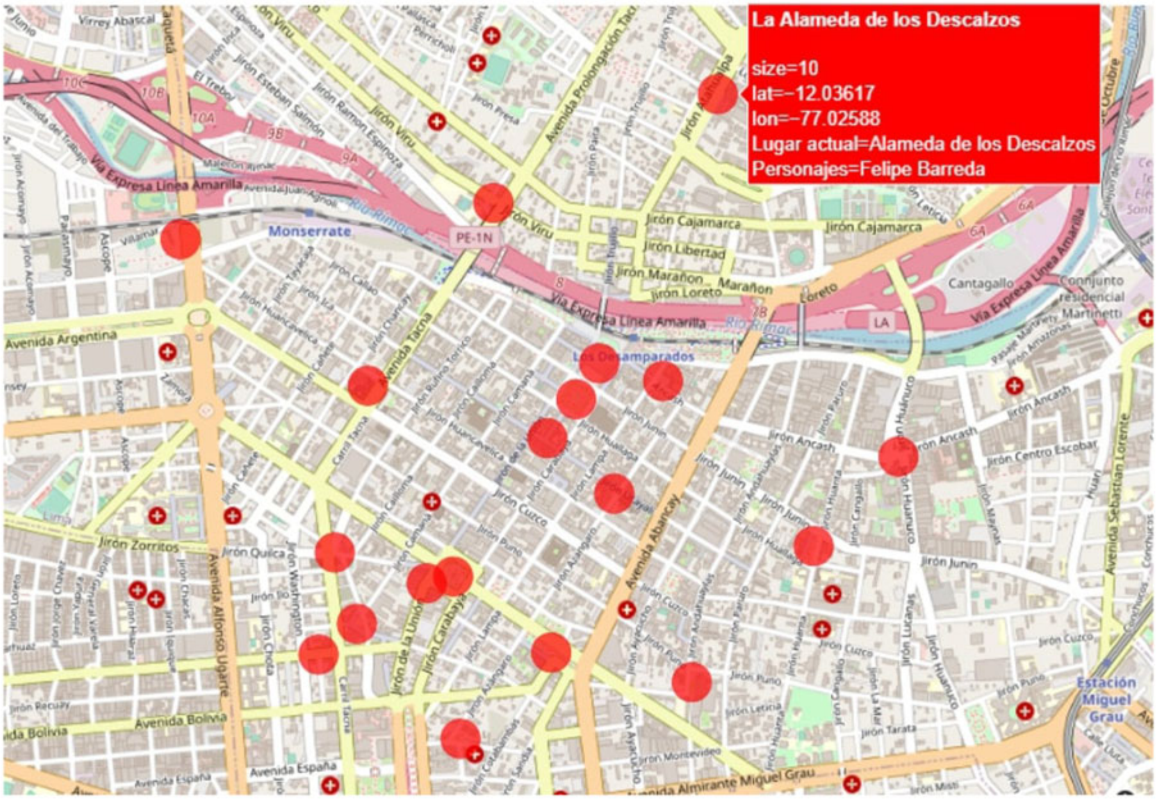


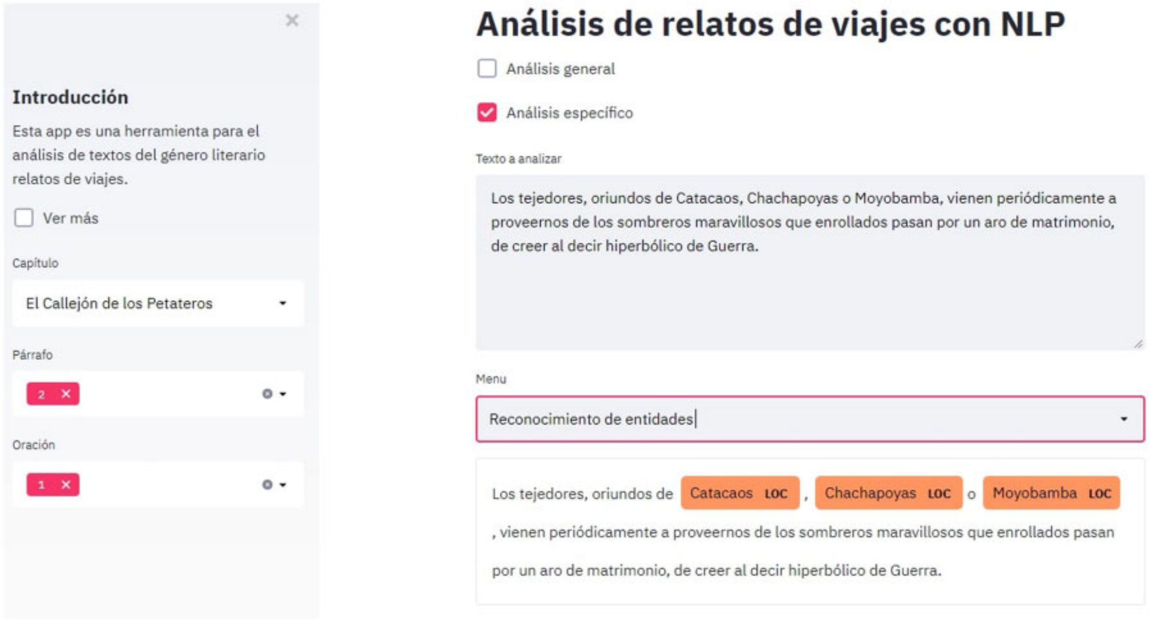
Fig. 6 The location entities were geolocated and visualized using an Entity Relation Map where each location contains its current name, original name, and related person entities

Boza' are related to famous or relevant figures of the time such as the poet Pedro Paz Soldán. In addition, the current address information gives us a deeper understanding of how the city has evolved over the centuries, and illustrates the difference between past and present buildings. For example, 'Casa de Boza' is located on Boza Street and corresponds to what is now block 8 of Union Street. To make the information more interactive, we constructed a map containing information such as related people, current address, name, and geographical points. Figure 6 presents a demo of the map that shows some spaces and the description of one of them.

Finally, we developed an interactive web application to deliver the macro-analysis (most frequent POS tags sequences patterns and a location–person map) and the micro-analysis (NER and the dependency and POS tree model outputs visualizations) in a user-friendly way to a nontechnical audience, such as Digital Humanities professionals or urbanists (see Fig. 7). The web application uses a static data file that contains every sentence tagged with its corresponding chapter and paragraph number. The user interacts with the application via a form where he can choose the text to analyze, then the NER model is used to identify PER or LOC entities and the text syntaxis tree is also shown to verify relations between entities. Also, the macro-analysis is done using as input data all the available texts. Here, we present an entity relation map and a ranking of the most frequent POS tags sequences patterns.[2]

5. Conclusions

In conclusion, NLP techniques have great value in advancing the fields of literary analysis, both close reading (i.e. specific passages in a single text) and macro-analysis (patterns of text databases). Thus far,



Introducción

Esta app es una herramienta para el análisis de textos del género literario relatos de viajes.

☐ Ver más

Capítulo: El Callejón de los Petateros

Párrafo: 2

Oración: 1

Análisis de relatos de viajes con NLP

☐ Análisis general

☑ Análisis específico

Texto a analizar

Los tejedores, oriundos de Catacaos, Chachapoyas o Moyobamba, vienen periódicamente a proveernos de los sombreros maravillosos que enrollados pasan por un aro de matrimonio, de creer al decir hiperbólico de Guerra.

Menu: Reconocimiento de entidades

Los tejedores, oriundos de Catacaos LOC , Chachapoyas LOC o Moyobamba LOC , vienen periódicamente a proveernos de los sombreros maravillosos que enrollados pasan por un aro de matrimonio, de creer al decir hiperbólico de Guerra.

Fig. 7 Screen capture of the Web Application developed as part of this research to allow non-technical digital humanities professional to explore the model results

the scope of NLP techniques has been limited to a direct or literal interpretation of the text, so it would be interesting to find ways to understand ‘artistic license’. This has implications for the organization of both semantics (metaphor) and information (stochasticity). In addition, it is important to stress that travel narratives are more accurate compared to other categories of literature, which usually have some fiction. An important issue to discuss is how much it affects the model generalization the fact that the language changes over time. As such, there is a need to determine whether a model pre-trained using another type of text could yield better results with the literature. One example might be a model pre-trained with historical texts. The application of NLP models can also include sentiment analysis, which in this case it offers information about the emotions or opinions in the text (Stone *et al.*, 1966). Despite this, this article analysis has only been descriptive, but it can be related to the language affective dimension. For example, to understand the reactions of the traveler to places or monuments it would be enough to add the conceptual variables that allow it. These variables would also help the assessment of the writer’s style. One of the tasks in the literary field is the relationship tensions between the objective and the subjective (Pang and Lee, 2008), and the description and its qualities. The dilemmas between narration and description are classic. As such, ancient authors such as Quintilian have worked on this topic (Kennedy, 1972). Thus, it is interesting trying to incorporate this knowledge into the world of computational analysis.

Furthermore, in the case of travel narratives, complementing textual analysis with geospatial data is very useful in reflecting the events experienced by the author (and the people) in real places. This would allow readers to connect the descriptions they are reading with real places. Percy Adams has noted that geography’s debt to travel literature is enormous (Adams, 1983); accordingly, another possible application of travel narratives is to facilitate urban planning through the recognition of textual objects in order to reproduce physical objects that no longer exist in a certain place (such as churches or squares that have been destroyed or modified). Thus, a virtual, literary reconstruction of a city can be carried out in certain spaces. Stories present the original or ancient configuration of cities in a literary way, allowing urban planners to make decisions based on historical knowledge. But narrations are not only descriptions as they also incorporate the author's perspective. Artificial intelligence has the potential to order these stories, differentiate the people and places, and even catalog the feelings that spaces evoke in the author. This would be very useful for those engineers and architects who want to approach the history of the city from a more human perspective, and therefore learn from past mistakes and successes of urban configuration.

Finally, future studies could explore the usage of unsupervised neural network models to perform the task of text summarization, the inclusion of other entities in the existing model, and the proposal and evaluation of an automatic geocoding process.

Notes

1. All the code use in this section is available in this repository.
2. Available at https://share.streamlit.io/ingenieriaup/traveling-through-stories-app/main.

References

Abbott C. and Adler, S. (1989). Historical analysis as a planning tool. *Journal of the American Planning Association*, **55**(4), 467–73.

Afinoguénova, E., Appel, S., Ballard, A., and McGowan, M. (2020). Letters from Spain in a Space-time Box: Historical GIS with timestamped itineraries for understanding the chronotopes of nineteenth-century travel writing. *International Journal of Humanities and Arts Computing*, **14**(1–2), 119–33.

Adams, P. G. (1983). *Travel Literature and the Evolution of the Novel.* Lexington, KY: University Press of Kentucky.

Akhtyamova, L. (2020). *Named Entity Recognition in Spanish Biomedical Literature: Short Review and BERT Model, 2020 26th Conference of Open Innovations Association (FRUCT)*, pp. 1–7, Yaroslavl, Russia, IEEE, 20–24 April 2020.

Alburquerque García, L. (2011a). *El "Relato de Viajes": Hitos y Formas en la Evolución del Género.* Revista de Literatura, 73, España: Consejo Superior de Investigaciones Científicas, pp. 15–34.

Alburquerque García, L. (2011b). Relatos y literatura de viajes en el ámbito hispánico: poética e historia. *Revista de Literatura*, **29**: 503–24.

Alharbi, A. A., Alotebii, H. A., and AlMansour, A. A. (2018). Towards measuring happiness in Saudi Arabia based on tweets: a research proposal. In *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–4. IEEE, Riyadh, Saudi Arabia, 4–6 April 2018.

Ashraf, S., Iqbal, H. R., and Nawab, R. M. A. (2016). Cross-genre author profile prediction using stylometry-based approach. *Working Notes of Conference and Labs of the Evaluation forum,* Portugal, Évora, CEUR Workshop Proceedings, pp. 992–9.

Benvenutto, P. (2003). Quince plazuelas, una alameda y un callejón. Lima en los años de 1884 a 1887. *Fragmentos de Una Reconstrucción Basada en la Tradición Oral.* Lima, Peru: Universidad del Pacífico.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: a survey, *Transactions of the Association for Computational Linguistics*, **7**: 49–72.

Borja, J. J. (1978). Nostalgia de Pedro Manuel Benvenutto Murrieta (1913–1978). *Boletín de la Academia Peruana de la Lengua*, **13**(13), 9. Retrieved from http://revistas.apl..pe/index.php/boletinapl/article/view/322.

Borin, L., Kokkinakis, D. and Olsson, L. J. (2007). Naming the past: named entity and animacy recognition in 19th century Swedish literature. *In Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007),* pp. 1–8.

Cotik, V., Rodríguez, H., and Vivaldi, J. (2018). Spanish named entity recognition in the biomedical domain, *Annual International Symposium on Information Management and Big Data.* Berlin/Heidelberg, Germany: Springer, pp. 233–48.

De Oliveira, T. H. M., and Painho, M. (2021). Open geospatial data contribution towards sentiment analysis within the human dimension of smart cities. *Open Source Geospatial Science for Urban Studies.* Cham Switzerland: Springer, pp. 75–95.

Deitz, M., Notley, T., Catanzaro, M., Third, A., and Sandbach, K. (2018). Emotion mapping: using participatory media to support young people's participation in urban design. *Emotion, Space and Society*, **28**: 9–17.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.

Díez Platas, M. L., Ros Muñoz, S., González-Blanco, E., Ruiz Fabo, P., and Alvarez Mellado, E. (2020). Medieval Spanish (12^{th}–15^{th} centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*, *72*(2), 224–38.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, **8**(1): 107–21.

Ezra, R. (1952). *Human Communities: The City and Human Ecology*. Glencoe, IL: Free Press.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T., 2003, Named entity recognition through classifier combination. In *Proceedings of the CoNLL-2003*, pp. 168–71, Edmonton, Canada, 31 May–1 June 2003.

Giannopoulou, M., Vavatsikos, A. P., Lykostratis, K., and Roukouni, A. (2014). Using GIS to record and analyse historical urban areas. *TeMA-Journal of Land Use, Mobility and Environment INPUT 2014, Eighth International Conference INPUT - Naples, 4-6 June 2014,* Italy. Naples: DICEA University of Naples "Federico II", pp. 487–97. 10.6092/1970-9870/2525.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning* (Vol. **1**, No. 2). Cambridge: MIT Press.

Han, J., Pei, J., Mortazavi-Asl, B. et al. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering, Heidelberg, Germany, 4/2/01 - 4/6/01* pp. 215–224, IEEE, Washington, DC.

Han, X. and Wang, J. (2019). Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in china. *ISPRS International Journal of Geo-Information*, **8**(4): 185.

Honnibal, M., Montani, I., van Landeghem, S., and Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. 10.5281/zenodo.1212303.

Hu, Y., Mao, H., and McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, **33**(4): 714–38.

Kennedy, G. A. (1972). *The Art of Rhetoric in the Roman World 300 B.C.–A.D. 300*. Princeton, NJ: Princeton University Press.

Koto, F. and Adriani, M. (2015). The use of POS sequence for analyzing sentence pattern in twitter sentiment analysis, *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, Gwangju, South Korea, IEEE, 24–27 March 2015, pp. 547–51.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One*, **8**(5): e64417.

Moreno, M. E. I., Aguilar-Amat, P. A. and Sánchez-Cuadrado, S. (2014). Primera aproximación para la extracción automática de entidades nombradas en corpus de documentos medievales castellanos, *Humanidades Digitales: Desafíos, Logros y Perspectivas de Futuro*, España: Universidade da Coruña, SIELAE, pp. 229–38.

Mumford, L. (1938). *The Culture of the Cities*. New York, NY: Harcourt Brace.

Nivre, J. and Fang, C. T. (2017). Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 22 May, Gothenburg, Sweden pp. 86–95.

Lim, K. H., Lee, K. E., Kendal, D. et al. (2019). Understanding sentiments and activities in green spaces using a social data–driven approach. *Smart Cities: Issues and Challenges*. Amsterdam, Netherlands: Elsevier, pp. 77–107.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*. New York, NY: IEE.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1–2), 2008, 1–135.

Pérez Martínez, A. (2013). Perspectivas (orteguianas) del paisaje en Quijote, *Anales Cervantinos*, **45**: 45–56.

Pirenne, H. (1969). *Medieval Cities: Their Origins and the Revival of Trade*. Princeton, NJ: Princeton University Press.

Plunz, R. A., Zhou, Y., Vintimilla, M. I. C. et al. (2019). Twitter sentiment in New York City parks as measure of well-being. *Landscape and Urban Planning*, **189**: 235–46.

Reiter, E. (2018). A structured review of the validity of bleu, *Computational Linguistics*, **44**(3): 393–401.

Rivera-Zavalaa, R. and Martineza, P. (2020). Deep neural model with contextualized-word embeddings for named entity recognition in spanish clin ical text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, Málaga, Spain, September 23th, 2020, *CEUR Workshop Proceedings*, pp. 385–95.

Rupp, C. J., Rayson, P., Baron, A. et al. (2013). Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data*, 6–9 Oct. 2013, Silicon Valley, CA, USA, pp. 59–62, IEEE.

Sayah, I., and Schnabel, M. A. (2019). *Amplifying Citizens' Voices in Smart Cities-An Application of Social Media Sentiment Analysis in Urban Sciences.* Intelligent & Informed - Proceedings of the 24th CAADRIA Conference - Volume 2, Victoria University of Wellington, Wellington, New Zealand, 15–18 April 2019, pp. 773–82.

Sjoberg, G. (1960). *The Pre-Industrial City*. New York, NY: The Free Press.

Stanley, B. W., Stark, B. L., Johnston, K. L., and Smith, M. E. (2012). Urban open spaces in historical perspective: a transdisciplinary typology and analysis. *Urban Geography*, **33**(8): 1089–1117

Stokes, N., Li, Y., Moffat, A., and Rong, J. (2008). An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science*, **22**(3): 247–64.

Stone, P. J., Dexter, D., and Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. pp. 3104–12. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Vani, K. and Gupta, D. (2017). Text plagiarism classification using syntax based linguistic features, *Expert Systems with Applications*, **88**: 448–64.

Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers), July 2017, Vancouver, Canada, pp. 189–98.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. 38. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Won, M., Murrieta-Flores, P., and Martins, B. (2018). Ensemble Named Entity Recognition (NER): evaluating NER tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, **5**: 2.

Yao, F. and Wang, Y. (2020). Towards resilient and smart cities: a real-time urban analytical and geo-visual system for social media streaming data. *Sustainable Cities and Society*, **63**: 102448.