PD-10: Natural Language Processing in GIScience Applications

Abstract

Natural Language Processing (NLP) has experienced explosive growth in recent years. While the field has been around for decades, recent advances in NLP techniques as well as advanced computational resources have re-engaged academics, industry, and the general public. The field of Geographic Information Science has played a small but important role in the growth of this domain. Combining NLP techniques with existing geographic methodologies and knowledge has contributed substantially to many geospatial applications currently in use today. In this entry, we provide an overview of current application areas for natural language processing in GIScience. We provide some examples and discuss some of the challenges in this area.

Keywords

natural language processing, text analytics, toponym disambiguation, topic modeling, question answering

1 Definitions

- Gazetteer: A dictionary or index of geographical names.
- n-gram: A sequence of n tokens, where n is a number. N-grams typically range between 1 (uni-gram) and three (tri-gram).
- Token: The building blocks of natural language. Small units of text that (e.g., characters, words, combinations of words) that have been split from a larger document or corpus.
- Toponym: A place name. Often derived from a topographic feature.

2 Natural Language Processing and GIScience

Natural Language Processing (NLP) is as an interdisciplinary research area that draws from the fields of linguistics, computational sciences, and many other related disciplines including geography and geographic information science (GIScience) that develop methods to analyze human language data. While the field includes a wide variety of topics it is primarily concerned with applying computational techniques to analyze human language in a variety of forms. In recent years, the field has focused on the extraction of patterns and meaning from large volumes of natural language data such as text and speech audio. Today, the field is moving towards "understanding" concepts and themes presented in natural language with the goal of answering questions and informing decision making.

Historically, the domain of natural language processing has focused on the extraction of structured content from unstructured text. Early *Symbolic* NLP approaches involved interpreting text and speech through a series of user-defined rules. In the 1980s and 1990s various statistical inference techniques were devised for identifying and applying these rules to natural language. More recently, the domain has seen a shift towards the use of machine learning, including deep learning, *Neural*, methods. These recent approaches do not take a rule-based approach but rather aim to understand natural language through statistical methods which can identify linguistic properties of words, sentences, or documents.

Though NLP does not fall solely within the discipline of Geography, a lot of human language is situated in geographic space and time and might make reference to inherently geospatial themes such as culture. Natural language varies by region meaning that GIScientists are well situated to process, identify, and contextualize patterns in language. Within the field of GIScience, NLP has been used to better understand a wide variety of geographic phenomena through identification of places, events, and activities as well as the extraction of linguistic patterns related to these entities. NLP techniques offer insight into geographic phenomenon that may not be accessible through traditional spatial and temporal analysis.

GIScientists are also able to leverage much of their existing expertise when processing natural language. Knowledge of spatial relationships, regional hierarchies and geographic laws & theories when combined with many leading NLP approaches result in cutting edge applications, many of which are actively used today. In the section to follow, a number of different NLP techniques are discussed with a specific focus on applications within the field of GIScience. The intent is to demonstrate how natural language processing is being used within GIScience applications today and discuss some of the challenges moving forward.

3 Applications of Natural Language Processing in GIScience

A number of natural language processing applications exist within GIScience. This section summarizes a small, but key set of application areas that have emerged in recent years.

3.1 Toponym disambiguation

Important locations on the Earth are usually given labels or *toponyms* to allow them to serve in a common reference system. When someone makes a reference to Montréal, Canada, for example, there is shared understanding of where this place is located on the Earth as well as what type of place it is, namely a city. Toponym disambiguation is the process of (a) identifying Montréal as a location, and (b) differentiating it from any other location labeled as Montréal.

To discuss toponym disambiguation in more detail, we must first take a large step back and discuss some of the building blocks necessary for many natural language processing tasks. The first step involves deconstructing natural language to a format that enables computational analysis, through a method known as tokenziation. Tokenization is the process of breaking down natural language into smaller lexical units which are referred to as *tokens*. Depending on the task, these units range from individual characters, to words (or sequences of words known as n-grams), sentences, paragraphs, or documents. The process of tokenization is easier for some languages than others. For instance, romance languages often delimit words with spaces whereas some Asian languages, such as Chinese, do not mark word boundaries with space delimiters making the process more complex [28].

In many languages, people use different inflection forms of words. For instance, *democratic*, *democracy*, *democracies*, and *democratization* all reference similar concepts, but for grammatical reasons the different words exist. For many applications these different concept references can be considered the same, thus it is advantageous to reduce them to a single token. Stemming is a simple solution to this problem that typically involves dropping the end of words such as derivational affixes, to reduce them to only those characters that the words have in common. For instance, a stemming approach to the above terms might be *Democra*. Lemmatization is a more complex approach that aims to identify the root term of the series of similar words. Often this root word is a term that represents a base concept rather than a sequence of common characters. For instance, a lemmatization of the example above might be *Democracy*. Lemmatization and stemming are often done as a first, data cleaning step along with tokenization.

Given these tokens, we come back to our objective of identifying and labeling these tokens. To achieve this, we use a technique known as *Named Entity Recognition (NER)*. NER is the process of labeling and categorizing lexical units extracted from unstructured natural language. This is typically an automated process of comparing tokenized entities found in unstructured text to an existing structured dictionary or determining the category of an entity based on the context in which the token exists. Pre-defined categories are often entities such as people, places, organizations, currencies, etc. This is not a trivial process as natural language can be quite complex and there is often a large amount of ambiguity in the meaning of words. Consider, for example, the sentence below.

I watched the Chicago Bulls game last night.

In this example, the term *Bulls* is ambiguous on its own as it is most often used to reference male cattle. It is only through analysis of contextual information that one is able to determine that *Bulls* in this instance refers to the Chicago-based professional basketball team. A state-of-the art NER application, such as Apache OpenNLP, would annotate each of the n-gram tokens in the example text with Chicago being labeled as a city in the United States, and the *Chicago Bulls* being labeled as a professional sports team. Today, many leading NER systems provide close to human-level performance in annotating unstructured text.

Even in the simple example above, the importance of geography is apparent. The region in which cattle are found, the city of Chicago, and dominance of basketball in discourse all relate to geography, and geographic knowledge can be leveraged in processing and labeling this information. NER is an important methodology to GIScientists as it is used in the first task of toponym disambiguation, which is that task of identifying and labeling a token as a geographic entity. Toponym disambiguation is typically accomplished through a look-up/matching process involving a geographic dictionary or what is often referred to as a *digital gazetteer* [13]. For lesser known or local toponyms, identification based on geographic context may be used. For instance, Hu et al. [16] use a geospatial clustering approach and contextual information from surrounding words to learn and train a machine learning model to identify toponyms based on unique spatial and linguistic patterns.

Once a token is identified as a toponym, the next challenge is differentiating it from other toponyms. The nature of human language and culture is that locations are often assigned the same label. For instance, there are at least 88 different locations in the United States with the name *Washington*, including cities, monuments, and a federal district. Identifying *which* Washington is the second task in toponym disambiguation. This is often a challenging task and involves examination of the contextual information and descriptive terms through which the toponym is referenced. In the Chicago Bulls example above, we can probabilistically identify *Chicago* as a large city in north-eastern Illinois, USA in a number of ways. First, Chicago, Illinois has the largest population of any known Chicago, and is therefore more likely to be mentioned in text. Second, an NER would likely identify the Bulls basketball team as an entity with a *home town* that also linked to the Chicago in Illinois. Leading research in this area has used a range of approaches that rely on existing geographic methods and spatial knowledge including graph-based approaches to linking toponyms [8], topic modeling for disambiguation [18], and co-occurrence models [24]. NER in general, and toponym disambiguation, more specifically, are central to foundational aspects of GIScience such as geocoding [11] and geographic information retrieval [17].

3.2 Spatial relationships in text

Aside from extracting geographic entities from natural language, researchers and industry professsionals are also very interested in understanding the relationships between geographic (and non-geographic) entities. Natural language data provides a rich source of relationship information as contributors of text often describe these relationships with rich detail. For instance if a body of text discusses the migratory patterns of people between two cities, this information could be extracted and represented as a geospatial flow between two network nodes in a GIS application. NLP extraction methods could also be used to identify mode of travel and quantify number of migrants.

As with toponym disambiguation, identifying and extracting relationships within unstructured natural language can be difficult. It requires us to determine which descriptors are applied to which words and which actions involve which actors. In the field of NLP, this process is called *coreference resolution.* Coreference resolution is the process of identifying which sub-components of a sentence or document, refer to which other sub-components, or tokens. In natural language, we often refer to specific entities or concepts through a variety of different terms and determine which entity is associated with which idea can be difficult for humans, let along computational model. Take the following example.

> *Seattle gets more days of rain than New York City, but it receives less total rainfall per year.*

In this case, we have two proper noun city names, Seattle and New York City as well as some facts about these cities. A coreference resolution task arises in the use of the pronoun, *it*. Within the context of this statement, *it* either refers to Seattle or New York City, and determining the correct referent is important when assigning information to a location. This may be a trivial task for a human to resolve, but the ambiguity of human language can often be difficult to represent computationally.

There are many ways to resolve ambiguity of coreferences within natural language and from a geospatial approach, we can leverage existing geographic knowledge. Early work in this discipline involved developing methods that applied a set of grammatical rules to natural language. This often meant the development of *parse trees* which aimed to represent dependency between tokens. Over the past couple of decades, techniques have been developed that take a probabilistic approach to identifying relationships through the construction of constituency parsing trees. While not all relationships are spatial, identifying relationships between entities can sometimes involve a spatial component, be it explicitly spatial (e.g., The museum in Montréal), or through regional or cultural context (e.g., The woman used the Algonquian word for fish). For example, Vasardani et al. [26] extracted mental representations of urban environments for use in emergency situations from verbal descriptions of places. Spatial hierarchies have also been extracted from user-generated text for use in qualitative spatial reasoning applications [29]. These, and many other processes demonstrate that spatial relationships can identified and extracted from unstructured linguistic content.

Having a background in GIScience also means that we are not solely reliant on the information extracted from natural language. We can use NLP techniques in conjunction with our existing geospatial expertise [22]. For example, Tobler's First Law of Geography can be applied in many cases to leverage the similarity of features in close proximity. Geographical theories such as *Central Place Theory* can be used to explain the relationships between nearby settlements, and gravity models can be employed to identify transfer and flow of entities described in text.

3.3 Discovering thematic patterns

Another approach to natural language processing is less concerned with labeling tokens and identifying individual toponyms in text and more interested in the broader themes or topics represented in natural language. The idea in this thematic approach to language is to extract groupings of terms that represent a set of topics on which a document can be characterized. This is important for representing ideas in documents as a whole as well as comparing themes across lexical units. The GIScience community has leveraged this approach to identify thematic patterns within geographic space and observe changes in patterns over time. One approach to this problem which has seen extensive use in the field of GIScience aims to extract themes or topics from corpora through an unsupervised probabilistic approach, called *Topic Modeling* that identifies the co-occurrence of tokens within documents. For example, applications of this technique have been used in clustering social media posts [14], location recommendation services [15], and ad hoc thematic search engines [3]. For instance, the Pteraform interactive search platform [1] shown in Figure 1 is built on top of geographically tagged Wikipedia data, and demonstrates how a topic modeling approach can be used to geographical depict themes over space and time. Notably, these approaches tend to ignore the sequence of tokens in a document or corpora and instead take what is commonly referred to as a *bag-of-words* approach.

Characterizing natural language text by themes is a form of classification, and there are also other ways we can classify a text. Sentiment analysis is the process of identifying and examining affective states within text and usually includes characterizing the emotions and attitudes towards a theme or topic. Techniques for identifying and extracting sentiment range from examining the *polarity* of individual tokens, to the *emotional state* of a document or grouping of tokens. Sentiment analysis is a notoriously challenging field of study as it involves analysis of subjective information and inference of intention by the

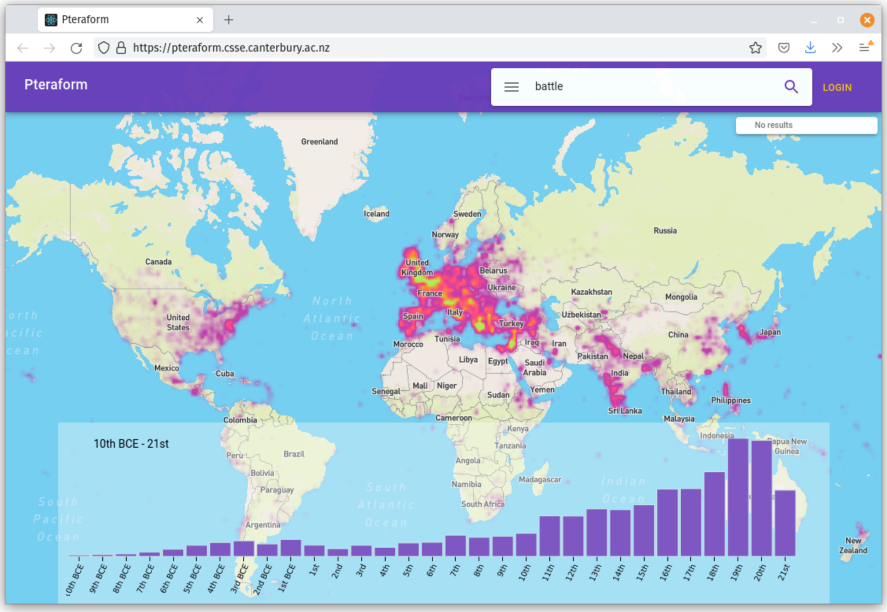

Figure 1: The Pteraform application showing spatial and temporal thematic (keyword: battle) trends over time.

language contributor. Applications of sentiment analysis in GIScience have included classification of parks through visitor contributions [20], understanding disaster response [4], and a plethora of research on attitudes towards travel destinations and places of interest [7, 19, 5].

3.4 Question Answering and Natural Language Generation

While humans can understand a sentence and the relationship between words through reading textual content or verbal communication, computers work in the realm of numerical values. Recent advances in NLP have moved towards not only representing words as numbers, but also the relationships between words. This allows analysts to perform mathematical and logical operations to compare terms, extract complex concepts, and better understand the ideas presented in natural language. This most often involve assigning a real-value representation to a sequence of terms and representing each unit as a numerical vector. Neural network-based methods such as word2vec or doc2vec are typically used to convert natural language to a series of numerical word vectors or matrices. The goal of this approach is to develop *word embeddings*. These encode the meaning of words, sentences, and concepts such that words that are closer in meaning are also closer in real-value vector space. Essentially, this involves embedding a multi-dimensional concept into a continuous lower-dimensional vector space. These word embeddings serve as the base unit on which many modern classification and predictive NLP tasks, including those in the geospatial field, are performed and often is a key pre-processing step for these other tasks.

Other techniques such as recurrent and convolutional neural networks have been applied to NER tasks with the goal of identifying geographic locations and places. Adams and McKenzie [2] used a character-level convolutional neural network to georeference noisy textual content and Cardoso et al. [6] used a variation on recurrent neural network for toponym resolution in text. Rather than applying rule-based approaches to identifying the features, deep learning methods use a representative classification approach to identifying latent features in natural language. These models thrive on large training datasets and the availability of rich and robust training data on which a model can be trained is critical. Transformer models such as Bidirectional Encoder Representations from Transformers (BERT) published by Google, have recently emerged. In this case, a learning model is pre-trained on an exceptionally large, generic dataset and then fine tuned for a specific task or application area. These *attention*-mechanized transformer models [27] have been shown to improve the accuracy and relevancy of many NLP-based applications, such as language translation and document search. These types of models are also being used for geospatial applications such as address validation [30], and identifying the locations of criminal organizations [23].

Question answering is a sub field within natural language processing, information retrieval, and artificial intelligence, in which a natural language questions, typically posed by a human are interpreted by a machine and appropriate responses are generated. In essence, this a fundamental test for many natural language processing techniques in that responding to a question requires comprehension of the concepts presented in the question itself. This approach involves a high level of automated reasoning. The field of *geographic* question answering has recently emerged with the goal of identifying and understanding the relationship between geographic features, places, and people through the use of many deep learning approaches. The nuances of geospatial concepts in natural language is unique and designing a system that can interpret and understand these concepts and relationships can be challenging. Take for example the question below.

> *How many people live in the capital of the third largest country on earth?*

Not only does the question above require entities to be extracted and labeled through an NER task or thematically encoded through a neural network, but it also requires leveraging existing geospatial knowledge such as administrative boundary hierarchies. For instance a capital is a city, a city exists within state, and a state with country. The term *largest* is ambigous here as well as it is unclear if this is in reference to population volume or physical area. Finally, *third*, it requires a system to know the populations or areas of all countries, rank them, and extract the third largest. While natural language processing techniques are increasingly able to learn many of these concepts, understanding the relationships and answering the question also involves accessing knowledge graphs, geographic databases, and range of other technologies. This area is proving to be a burgeoning subfield of GIScience. Scheider et al. [25] discuss the challenges associated with building a question-based geographic information system and how existing spatial techniques and technologies can be used within such a service. Mai et al. [21] demonstrate possibilities and limitations of geographic question answering through the use of geospatially enabled knowledge graph embeddings.

The complement to question answering is *natural language generation* (NLG). This approach aims to generate natural language text or speech based on semantically encoded concepts. In many ways, the second part of question answering demands generating natural language based on the interpreted understanding of the original question. Applied work in this field has predominantly focused on automating reports and responses to questions. Within the geographical sciences we see NLG techniques being applied to generating weather reports [12], descriptions of places and remotely sensed imagery [10], and the broader focus on chatbots and automated assistants capable of responding to basic questions.

4 Challenges

A number of challenges exist within the domain of natural language processing and many of them are uniquely spatial. Many of these were mentioned in the previous sections, but here the challenges are outlined in further detail.

Using NLP to interpret fine-grained spatial relationships in text is an active area of research. While many current NLP approaches are able to identify concepts, ideas, and relationships within natural language, surprisingly few of them explicitly model spatial relationships. Concepts such as spatial autocorrelation are fundamental to GIScience, yet very few approaches incorporate this idea in the process of understanding natural language.

Spatial cognition is a branch of cognitive psychology that studies the ways in which people use spatial information to gain knowledge, self locate, and wayfind. This field is closely linked with natural language processing in that understanding human-contributed natural language necessitates an understanding of how humans conceptualize space and communicate those concepts in language [9]. This presents a unique challenge, as how humans conceptualize and communicate spatial concepts is not fully understood, therefore making it difficult to train a computational model to represent spatial information in a similar way.

While substantial advances have been made in toponym disambiguation and co-reference resolution within NLP research, it still remains as a challenge. Given that places are labeled by humans, they tend to change over time, or have multiple, often localized, names. Humans reference places in different ways and the ability to identify a single place based on various colloquial references to the location remains a challenge.

Lastly, the automated generation of spatially-aware narratives is a challenge area that will likely see advances in the coming years. This will involve the integration of NLP more substantially in location-based systems such as tourism applications and will leverages geographic knowledge graphs and existing gazetteers.

5 Learning Objectives

The objective of this chapter is to

- Explain how natural language processing is being used in geographic information science applications.
- Differentiate between some of the key uses of natural language processing in geography and GIScience.
- Identify how *spatial is special* in the context of natural language processing.
- Identify challenges and future directions for applications of NLP in GIScience.

6 Instructional Assessment Questions

1. What does the field of geography bring to the discussion of natural language processing?
2. What are the two components necessary for toponym disambiguation?
3. How is geographic question answering different than traditional question answering?
4. What is the difference between stemming and lemmatization?

7 Additional Resources

- Apache OpenNLP https://opennlp.apache.org/index.html
- Stanford Natural Language Processing Toolkit https://nlp.stanford.edu/
- Python Natural Language Toolkit module https://www.nltk.org/
- R GeoParser package https://rdrr.io/cran/geoparser/
- An Extensible and Unified Platform for Evaluating Geoparsers https://geoai.geog.buffalo.edu/EUPEG/
- Creating the Corpus (Spatial Language) https://geospatiallanguage.massey.ac.nz/creatingthecorpus.htm
- EarthLings (Computational Linguistic Atlas) http://www.earthlings.io/

References

[1] Benjamin Adams. Chronotopic information interaction: integrating temporal and spatial structure for historical indexing and interactive search. *Digital Scholarship in the Humanities*, 2020.

[2] Benjamin Adams and Grant McKenzie. Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. *Transactions in GIS*, 22(2):394–408, 2018.

[3] Benjamin Adams, Grant McKenzie, and Mark Gahegan. Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th international conference on world wide web*, pages 12–22, 2015.

[4] Abdullah Alfarrarjeh, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. Geo-spatial multimedia sentiment analysis in disasters. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 193–202. IEEE, 2017.

[5] Andrea Ballatore and Benjamin Adams. Extracting place emotions from travel blogs. In *Proceedings of AGILE*, volume 2015, pages 1–5, 2015.

[6] Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. Using recurrent neural networks for toponym resolution in text. In *EPIA Conference on Artificial Intelligence*, pages 769–780. Springer, 2019.

[7] Mario Cataldi, Andrea Ballatore, Ilaria Tiddi, and Marie-Aude Aufaure. Good location, terrible food: detecting feature sentiment in user-generated reviews. *Social Network Analysis and Mining*, 3(4):1149–1163, 2013.

[8] Hao Chen, Stephan Winter, and Maria Vasardani. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 2018(17):31–62, 2018.

[9] Song Gao, Krzysztof Janowicz, Daniel R Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31(6):1245–1271, 2017.

[10] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[11] Daniel W Goldberg, John P Wilson, and Craig A Knoblock. From text to geographic coordinates: the current state of geocoding. *URISA journal*, 19(1):33–46, 2007.

[12] Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[13] Linda L Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pages 280–290. Springer, 2000.

[14] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88, 2010.

[15] Bo Hu and Martin Ester. Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 25–32, 2013.

[16] Yingjie Hu, Huina Mao, and Grant McKenzie. 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):714–738, 2019.

[17] Christopher B Jones and Ross S Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008.

[18] Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *European Knowledge Acquisition Workshop*, pages 353–367. Springer, 2016.

[19] Hanhoon Kang, Seong Joon Yoo, and Dongil Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010, 2012.

[20] Anna Kovacs-Györi, Alina Ristea, Ronald Kolcsar, Bernd Resch, Alessandro Crivellari, and Thomas Blaschke. Beyond spatial proximity—classifying parks and their visitors in london based on spatiotemporal and sentiment analysis of twitter data. *ISPRS International Journal of Geo-Information*, 7(9):378, 2018.

[21] Gengchen Mai, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, and Ni Lao. Se-kge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24(3):623–655, 2020.

[22] Grant McKenzie and Benjamin Adams. Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content. In Eliseo Clementini, Maureen Donnelly, May Yuan, Christian Kray, Paolo Fogliaroni, and Andrea Ballatore, editors, *13th International Conference on Spatial Information Theory (COSIT 2017)*, volume 86 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[23] Javier Osorio and Alejandro Beltran. Enhancing the detection of criminal organizations in mexico using ml and nlp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[24] Simon Overell and Stefan Rüger. Using co-occurrence models for place-name disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287, 2008.

[25] Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. Geo-analytical question-answering with gis. *International Journal of Digital Earth*, 14(1):1–14, 2021.

[26] Maria Vasardani, Sabine Timpf, Stephan Winter, and Martin Tomko. From descriptions to depictions: A conceptual framework. In *International Conference on Spatial Information Theory*, pages 299–319. Springer, 2013.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[28] Jonathan J Webster and Chunyu Kit. Tokenization as the initial phase in nlp. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, 1992.

[29] Xiaoyu Wu, Jianying Wang, Li Shi, Yong Gao, and Yu Liu. A fuzzy formal concept analysis-based approach to uncovering spatial hierarchies among vague places extracted from user-generated data. *International Journal of Geographical Information Science*, 33(5):991–1016, 2019.

[30] Liuchang Xu, Zhenhong Du, Ruichen Mao, Feng Zhang, and Renyi Liu. Gsam: A deep neural network model for extracting computational representations of chinese addresses fused with geospatial feature. *Computers, Environment and Urban Systems*, 81:101473, 2020.