[DC-02-037] Texts

Abstract

The integration of Geographic Information Retrieval (GIR) with advancements in Natural Language Processing (NLP) and Large Language Models (LLM) has revolutionized the utilization of unstructured text as a data source for Geographic Information Systems (GIS). Historically, unstructured text, unlike structured text such as XML documents or SQL queries, was predominantly leveraged by search engines and within the broader field of Information Science. However, the ubiquity of user-generated content on social media, combined with accessible online news outlet APIs, has prompted the integration of textual data in GIS applications. The fundamental shift in NLP technologies, particularly the advent of LLMs like GPT models and the evolution of text recognition algorithms, has enhanced the reliability of place name recognition, a subset of Named Entity Recognition (NER). These technologies enable the effective extraction of geographic references from vast quantities of textual data, offering substantial potential for enriching GIS databases. The primary challenges in this field include resolving place name ambiguities and vagueness, and adapting to the dynamic nature of geographic names and boundaries. Despite these challenges, GIR promises to unlock powerful new dimensions of spatial analysis and decision-making by integrating textual and geographic data.

Keywords: Geographic Information Retrieval (GIR), natural language processing, toponym disambiguation, toponyms

Author & citation

Karimzadeh, M. (2024). Texts. The Geographic Information Science & Technology Body of Knowledge (2024 Edition). John P. Wilson (Ed.). DOI: 10.22224/gistbok/2024.1.8

Explanation

1. Background
2. Methods
3. The Role of Foundational Models
4. Applications
5. Conclusion

1. Background

Up to early 2010s, unstructured text (i.e., free-form text, as opposed to structured text such as XML documents or SQL queries) was not commonly known as a major data source for GI Systems, and was mostly the focus of search engines and the broader field of Information Science. Nevertheless, the ubiquity of user-generated social media content, and online news outlet APIs as well as advances in Natural Language Processing (NLP), Machine Learning (ML), and Large Language Models (LLM) brought the wealth of information embedded in unstructured text to the attention of researchers and practitioners in GIS.

According to some analyst estimates, 80% of all data are unstructured, with the majority in

text. However, researchers have empirically confirmed that of all data stored as text, between 60% to 80% of all textual documents includes a geospatial reference, depending on how one defines the notion of 'geographic' (Hahmann et al., 2013). The salient point is that significant portions of all business data are textual, and a large majority of that data includes geographic references, making it a potential data source for GI Systems and Science. This includes both the spatial, and aspatial (attribute) aspects of data capture.

Textual data from sources like social media, news, and historical records offers qualitative insights that structured data like Census or GPS do not directly capture. It provides context to spatial patterns, revealing public perceptions, cultural significance, and human narratives tied to locations. These insights are essential for applications such as urban planning and disaster response, where understanding human experiences and historical contexts enhances the depth of GIS analyses.

2. Methods

Methods for capturing geospatial data, spatial or aspatial, from more conventional sources such as satellite or airborne observations or field surveys are comparatively mature—albeit ever progressing—in methodology. However, methods for harnessing and extracting geospatial information and knowledge from textual resources are younger, and undergoing consistent progress and paradigm shifts with more impressive results.

Geographic Information Retrieval (GIR) broadly refers to the process of searching and retrieving geographic information from textual documents, in order to develop spatially-aware search systems and support user's geographical information needs (Purves et al., 2018). It involves techniques and tools that allow users to find relevant information based on location-based queries. Unlike traditional information retrieval, which primarily focuses on text and content relevance, GIR integrates spatial queries with textual queries to provide more contextually and geographically relevant results. The field combines elements from information retrieval (IR), computational linguistics, and GIS to enhance the effectiveness of location-based searching.

GIR systems can leverage both explicit geographic expressions (such as place names) and implicit references (such as descriptions associated with particular locations) within documents. The primary challenge in GIR is understanding the spatial aspects of terms within the text and how these relate to user queries and real-world geography. This involves semantic interpretation, spatial indexing, and query processing to ensure that the geographical dimensions of the data are correctly understood and efficiently retrieved. The GIR-specific processes related to extracting geospatial information from text, however, can be generalized into two steps: Toponym Recognition and Toponym Resolution (Karimzadeh et al., 2019).

2.1 Toponym Recognition

Prompted by advances in NLP in the early 2010s, algorithmic approaches started to be more reliable in recognizing place names in text, such as recognizing 'Colorado' in the sentence "Colorado is home to 58 peaks above 14,000 feet of elevation". Called Place Name Recognition, this task is a special case of 'Named Entity Recognition' (NER), as subject of active research in NLP. Early approaches relied more on rule- and grammar-based algorithms, and phased out by the high performance of ML approaches, especially deep learning (DL) algorithms specialized for sequence data and natural language,

including Recurrent Neural Networks (RNN) and Transformers (Berragan et al., 2023). As opposed to rule-based methods that rely on the programming of language grammar, DL-based methods are trained on large bodies of text where considerable portions of place names (and other names) are manually 'annotated' by humans, allowing the models to learn from annotations. This, however, is also a limitation: creating large annotated datasets for a specialized task such as toponym recognition is expensive and time consuming. As we explain below, Large Language Models have changed the paradigm in this space. Nevertheless, given that NRE has been the focus of the larger field of NLP, and thanks to larger amounts of annotated training data enabled by the larger community of research and practice, NER and its subset of Place Name Recognition are relatively reliable methods of extracting place names. However, indirect references to place names, such as landscape descriptions require more creative approaches to address specific challanges (Koblet et al., 2020). It is worth noting that place is not a direct synonym for toponym. Rather, place name is synonymous with toponym, and places have names as a one property. It is possible—and an active area of research—to locate texts using their properties (e.g. locale and sense of place), without recourse to place names.

Once a place name is recognized, it needs to be geolocated to a place. For this, a gazetteer is needed.

2.2 Gazetteers

A gazetteer is a database of geographic names (toponyms) along with a unique identifier and corresponding coordinates, at a minimum. Gazetteers play a fundamental role in GIR tasks by serving as reference sources for place names and their locations. For instance, in the context of a query such as "schools within Boulder County close to Lafayette", "Boulder" and "Lafayette" need to be resolved to their corresponding gazetteer entry with

Gazetteers can vary widely in their geographic coverage, ranging from global to local scales. Some are comprehensive and cover the entire globe, while others might focus on specific countries or regions (Acheson et al., 2017). The level of detail in a gazetteer can also vary. Some provide detailed descriptions, including alternative names, alternative spellings, names in other languages, historical names, population sizes, and additional metadata such as the type of location (e.g., city, river, mountain). Gazetteers are regularly updated to reflect changes in toponyms and administrative boundaries. Some gazetteers include semantic information about places, such as the relationships between different locations (e.g., hierarchical relationships like a city being part of a state).

GeoNames is one of the most well-known, open, free and widely-used global gazetteers. GeoNames aggregates geographic data from various sources, providing over 25 million geographical names, consisting of over 12 million unique features whereof 4.8 million are populated places and 16 million are alternate names. GeoNames consistently integrates geographical data such as names of places in various languages, elevation, population and others from various sources, including official sources as well as crowd-sourced information. Users can manually edit, correct and add new names using a wiki interface. On the other side of the spectrum, is a gazetteer like the Geographic Names Information System (GNIS), which is the United State's Federal and national standard for geographic names, and serves as the official repository of geographic names data for the Government. As an official source and controlled by the U.S. Board on Geographic Names (BGN), the GNIS contains

information about physical geographic features of many types in the United States, associated areas, and Antarctica. GNIS is one of the sources of GeoNames gazetteer.

2.3 Toponym Resolution

Toponym resolution, the core geographic component of GIR, involves identifying and disambiguating place names (toponyms) mentioned in text, assigning them to an entry in the gazetteer. This process is essential for mapping text, as well as any task relying on geospatial references in documents. For instance, in the previous example “Colorado is home to 58 peaks above 14,000 feet of elevation”, resolving ‘Colorado” to the toponym with GeoNames ID 5417618 along with its geographic coordinates is the job of toponym resolution.

Despite its core role for unlocking text potential in GIS, toponym resolution is challenging. The primary issue is ambiguity, which arises when the same name refers to multiple geographic locations (e.g., Paris, Texas vs. Paris, France) or when different names refer to the same place. A text might refer to "Washington" the State, "Washington" the city (or even “Washington”, a person), and without leveraging contextual clues, systems can fail to properly disambiguate place names. In fact, Washington is one of the most common place names in the U.S., with 91 places or natural features only in the United States having Washington as part of their names! Context plays a significant role in correctly interpreting toponyms.

Another challenge is the dynamic nature of place names and geographic boundaries. Historical texts may refer to places that no longer exist in official databases, or whose names have changed, requiring systems to have access to historical geographic data as well as current data (Bol, 2013). Moreover, the informal or non-standard language often used in social media posts and similar content can further complicate the identification and resolution of toponyms (Snyder et al., 2019).

In addition to the ambiguity of place names, the vague boundaries of some places make toponym resolution and downstream tasks even more challenging. For instance, references to the “west coast”, even if resolved to the intended country or region of intent, still may refer to a vague, and at times, subjective region. Wallgrün et al. (2018) identified many special cases and challenges of toponym resolution, including references with uncertain semantics, vaguely qualified place names, and grouping errors. These special cases can act as a guide and test cases for implementing and testing GIR systems.

To address these challenges, toponym resolution systems must develop NLP and ML techniques that can interpret the context and semantics of text accompanying geographic references. In the case of social media posts, leveraging the network of related users and posts can provide further context. Toponym resolution methods will also need comprehensive gazetteers that include not only current but historical and alternative place names.

It's important to note that the effectiveness of NLP and GIR technologies can vary significantly across different languages. While many of the examples and technologies discussed here predominantly pertain to English, the challenges of geoparsing, GIR, and geo-information extraction from text are not language-neutral. For instance, languages with rich morphological variations or those with less digital resources pose unique challenges in entity recognition and toponym resolution. This underscores the need for developing

adaptable methods that can cater to the linguistic diversity inherent in global geographic data encoded in text.

3. The Role of Foundational Models

The advent of Large Language Models (LLMs) and the shift from traditional NLP methods to using advanced foundational models like GPT (Generative Pre-trained Transformer) models (Alec et al., 2019) represent a significant paradigm shift in GIR. This transition follows a broader trend within Artificial Intelligence (AI) and machine learning towards leveraging large-scale, pre-trained models capable of generalizing across a wide range of tasks, including complex NLP challenges, such a geoparsing or toponym resolution.

GPT models are a type of LLM-based on the Transformer architecture (Vaswani et al., 2017). The transformer architecture revolutionized NLP by moving away from sequential data processing (like in traditional RNNs) to a model that processes all words or tokens in parallel. This architecture significantly improves the efficiency of training models on large datasets and handling long-range dependencies in text. The resulting scalability enabled training models on very large corpora of text and learning more complex patterns in text.

The pre-trained aspect of GPT refers to the initial training phase, where the model is trained on an extremely large corpus of text data for predicting the next word in sequences of text, effectively (and surprisingly) learning the probability distribution of word occurrence in a language. With the large size of the training data and model complexity, this phase results in learning a variety of language patterns and structures without needing task-specific labels. The models learn to predict the next word in a sentence, which implicitly requires understanding syntax, grammar, semantics, and at least a semblance of world knowledge. The pre-training phase results in a model that has a broad understanding of language, which can then be fine-tuned to specific tasks with much smaller datasets.

After the pre-training, GPT models can be fine-tuned on smaller, task-specific datasets. Fine-tuning involves continuing the training process but with a dataset that is labeled for specific tasks like sentiment analysis, question answering, or, in the case of GIR, geographic entity recognition and disambiguation. More importantly, GPT models, with their deeper contextual understanding, can determine the meaning of ambiguous place names based on the context. For instance, Hu et al. (2023) showed how a GPT-based model can extract location descriptions from disaster-related social media messages, while fine-tuned using only 22 training examples. GPT models also have high potential for semantic parsing of geographic queries, as well as for responding to complex geographic queries, transforming natural language questions into actionable search parameters in GIR and GIS systems.

While LLMs offer powerful capabilities, they also come with challenges such as the need for substantial computational resources for (initial) training and inference, potential biases in the training data and pre-trained models including geographic biases, and the requirement for fine-tuning to adapt to specific domains like geography. LLMs like GPT show promise in GIR, but their scalability and efficiency in operational settings are still under examination. These models require substantial computational resources, with a notable trade-off between size and performance, limiting their practical use. Continued evaluation and optimization are essential to balance effectiveness with resource demands.

4. Applications

Geospatial information extracted from text can enhance and complement existing capabilities of GIS. Once text is transformed into geospatial data types, GIS can be used to conduct analysis and build models, and subsequently create new knowledge and understanding in various application domains:

- Urban Planning: Analyze social media and news articles to gauge public opinion on urban development projects, helping planners align initiatives with community needs.
- Disaster Response: Leverage real-time textual data from social media for rapid assessment of affected areas and to coordinate emergency services more effectively (Snyder et al., 2019).
- Cultural Studies: Utilize historical documents and contemporary narratives to map cultural heritage sites and understand changes in cultural landscapes over time.
- Environmental Monitoring: Use text from news reports, blogs, or social media to identify and track public concerns and reports about environmental issues like pollution or deforestation.
- Public Health: Analyze community health forums and public health updates to map the spread of diseases and public sentiment towards health interventions and policies, and likely adherence rates to non-pharmaceutical interventions.

Beyond individual applications, mapping the geographic references in documents, patterns and trends that are not otherwise obvious from text alone can be uncovered. Additionally, toponym resolution can enhance the user experience in navigation services, location-based services, and social media platforms by providing more contextually relevant content based on geographic references.

5. Conclusion

Extraction of (geospatial) information from text has paved the way for transformative advancements in data capture, analysis, and decision-making across various sectors using GIS. GIS platforms can incorporate a wealth of information that enriches spatial analysis and enhances our understanding of the world. As technology continues to evolve, the potential for further innovation at the intersection of text and spatial data remains limitless.

GIR and information extraction from text can open up new frontiers in data integration and knowledge discovery. For instance, combining the strong predictive power of traditional machine learning on earth observations with domain expertise has been a focus of research and remains challenging. Experts often describe their observations and knowledge in textual formats, detailing phenomena like climate dynamics, weather patterns, and forest fire dynamics. By integrating scientist expertise captured from text with mapping technology, we are one stop closer to further substantially improve inference and knowledge discovery.

References

Acheson, E., Sabbata, S. De, & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. Computers, Environment and Urban Systems, 64(Supplement C), 309–320.

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. International Journal of Geographical Information Science. 37(4), 747–766.

Bol, P. K. (2013). On the Cyberinfrastructure for GIS-Enabled Historiography: Space-Time Integration in Geography and GIScience. Annals of the Association of American Geographers. 103(5), 1087–1092.

Hahmann, S., & Burghardt, D. (2013). How much information is geospatially referenced? Networks and cognition. International Journal of Geographical Information Science, 27(6), 1171–1189.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., ... Joseph, K. (2023). Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. International Journal of Geographical Information Science. 37(11), 2289–2318.

Karimzadeh, M., Pezanowski, S., Wallgrün, J. O., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. Transactions in GIS, 23(1), 118–136.

Koblet, O., & Purves, R. S. (2020). From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally. Landscape and Urban Planning. Vol 197.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends in Information Retrieval. 12(2-3):164-318.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019, February 19). Language Models are Unsupervised Multitask Learners. Enhanced Reader, Open AI Blog.

Snyder, L. S., Lin, Y., Karimzadeh, M., Goldwasser, D., & Ebert, D. S. (2019). Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness. IEEE Transactions on Visualization and Computer Graphics, 1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pages 5998– 6008.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: building a corpus to test and train microblog geoparsers. International Journal of Geographical Information Science, 32(1), 1–29.