Grenze International Journal of Engineering and Technology, June Issue

A Robust and Accessible System for Geospatial Information Extraction using Distil BERT and Fuzzy Logic

Bennet Praveen T[1] and T Kavitha[2]
[1-2]Karunya Institute of Technology and Sciences, Coimbatore, India
Email: bennet.praveen@gmail.com, kavithat@karunya.edu

***Abstract*— This paper proposes a fine-tuned DistilBERT model for accurate and efficient geospatial information extraction from natural language texts. The model is trained on a custom dataset, enhancing its ability to recognize and map colloquial names to canonical forms while differentiating between similar entity names across diverse geographical locations. By leveraging a comprehensive geographical hierarchy analysis and fuzzy logic algorithms, it effectively resolves ambiguity and handles spelling errors and variations in entity names. Integrated with cutting-edge APIs, including Google maps and Wikipedia, the system provides accurate pinning, visualization, and additional information about identified places. Supporting multilingual speech-to-text, it offers users an intuitive interaction experience. Overall, this system represents a notable advancement in geospatial information extraction, offering a valuable reference for further research in location-based natural language processing.**

***Index Terms*— Geospatial information extraction, DistilBERT, Fuzzylogic**

I. Introduction

Geospatial querying systems, using state-of-the-art techniques like DistilBERT and fuzzy logic, represent a contemporary challenge in Natural Language Processing (NLP). The key focus of this project is the development of an intricate system designed to automatically discern and decode geospatial references within textual queries, effectively linking them to their standardized forms while accommodating variations, potential errors, and different references to the same entity.

DistilBERT and fuzzy logic being the foundation of the proposed system, this system not only tackles the challenges faced in Natural Language Processing (NLP) but also embraces a paradigm shift in geospatial querying. The model helps in classifying and identifying the diverse Geospatial locations. This empowers the system to recognize place names from a given natural language sentence entered by the user. By integrating a fuzzy matching model, the model address the inherent challenges of linguistic variations and potential errors in user-generated speech, ensuring the accuracy and reliability of the information extracted. Moreover, the addition of speech-to-text module further enhances user interaction, providing a seamless means for individuals to interact with the system.

a) *Background and Motivation*

Amidst the expanding realm of geographic information systems and the burgeoning reliance on location-based

Grenze ID: 01.GIJET.10.2.276

services, the demand for accurate and intuitive geospatial query processing has become increasingly pronounced. The genesis of this system originates from the pressing need to decode and respond adeptly to geospatial queries framed in natural language, overcoming the inherent ambiguities and intricacies embedded within such queries.

b) Problem Statement

The primary challenge lies in the creation of a sophisticated geospatial querying system that seamlessly identifies and resolves geospatial entity names in sentences, associating them with their canonical forms. Achieving this necessitates the application of NLP methodologies to isolate geospatial terms, employ fuzzy logic for potential canonical name identification, and present precise canonical names as output.

c) Significance of the Proposed System:

The system introduced in this paper marks a substantial advancement in the domain of geospatial information extraction. By leveraging a fine-tuned DistilBERT model, the system ensures enhanced accuracy and efficiency in processing geospatial queries. It not only addresses the complexity of recognizing and disambiguating colloquial names but also ensures a comprehensive analysis of the geographical hierarchy, reducing ambiguity within the queries. The integration of fuzzy logic algorithms, coupled with the ability to filter non-geospatial entities, represents a significant contribution to accurate and streamlined geospatial querying.

d) Organization of the Paper

This paper unfolds across several sections, each dedicated to presenting the methodologies, system architecture, and outcomes achieved by the proposed system. Section 2 provides an overview of the existing methods in geospatial information extraction. Section 3 delves into the methodology and architecture of the developed system, including the integration of DistilBERT, fuzzy logic algorithms, and the NLP approach. Section 4 extensively details the system's performance evaluation, showcasing its efficiency and accuracy. Finally, Section 5 concludes the paper, highlighting potential pathways for future research in this domain.

II. Related Works

Several papers discuss advancements in natural language processing (NLP) and geographic information systems (GIS). Eric F. Tjong Kim Sang and Fien De Meulder propose methods for named entity recognition (NER) in English and German text, utilizing machine learning and external resources like gazetteers [1]. Meanwhile, Victor SANK et al. introduce DistilBERT, a compact version of BERT, enhancing efficiency while retaining language understanding capabilities [2]. Additionally, fuzzy logic, explored by Bart Kosko and Satoru Isaka, offers alternatives to binary logic, particularly in handling vagueness in real-world data such as place names [3].

Jan Oliver Wallgrun, Morteza Karimzadeh, Alan M. MacEachern, and Scott Pezanowski present the GeoCorpora initiative, focusing on building a corpus for testing microblog geoparsers, which aids in enhancing geoparsing software evaluation [4]. Another paper, authored by Vinnarasu A and Deepa V.Jose, discusses integrating speech-to-text conversion with text summarization for effective understanding and documentation, addressing challenges in speech recognition variations [5]. Moreover, geoparsing methodologies are detailed, highlighting approaches like named entity recognition and map-database methods, crucial for extracting location information from social media content [6].

The integration of Building Information Modeling (BIM) and Geographic Information Systems (GIS) is explored, emphasizing advancements made in geospatial data management research [7]. Additionally, a novel method for downscaling fuzzy geospatial objects using ATPK is presented, showing promising results in accurately delineating wetland boundaries [8]. Further, fuzzy matching in graph theory is discussed, offering insights into matching in fuzzy graphs and its practical implications [9]. Lastly, the utilization of the GATE platform for extracting geographical named entities and spatial relations from text is highlighted, focusing on the creation of a spatial relation corpus and rule-based extraction methods to enhance GIS representation [10].

TABLE I. Comparison Of The Existing Models With The Dataset

Model	Architecture	Dataset	F1-score	Advantages	Disadvantages
BERT	Transformer	CoNLL-2003	92.8%	State-of-the-art performance, pre-traded on large corpus, bidirectional context	Large model size, high computational constrains, Requires fine-tuning

DistilBERT	Transformer	CoNLL-2003	91.3%	Smaller and faster version of BERT, preserves most of BERT's performance, Bidirectional context	Slightly lower performance than BERT, Still require fine-tuning
BiLSTM-CRF	Bi-directional LSTM with CRF	CoNLL-2003	90.9%	Simple and effective model, captures sequential dependencies, uses CRF for decoding	Requires Hand-crafted features, cannot capture long-range dependencies, unidirectional context
spaCy	CNN with residual connections	ObitoNotes 5.0	85.9%	Fast and easy to use supports multiple languages, provides various NLP tools	Lower Performance than transformer models, limited customization, fixed vocabulary
Flair	Bi-directional LSTM with CRF and character embeddings	CoNLL-2003	93.1%	Combines contextual string embeddings and classic word embeddings achieves state-of-the-art performance, supports many languages	Large model size, high computational cost, slow inference
Stanford NER	CRF with hand-crafted features	CoNLL-2003	86.9%	Well-established and widely used model provides Java and python interfaces, supports many languages	Lower performance than neural models, requires hand-crafted features, fixed vocabulary

This comparative analysis reveals that our approach surpasses existing methods in the extraction of the entity, providing an efficient, robust, and adaptable solution using [2] DistilBERT and [3] Fuzzywuzzy. This marks a significant advancement in the domain of the entity extraction, opening doors for a much more reliable emergency response systems.

III. Methodology

Methodologies used mainly include machine learning techniques. The block diagram showing the workflow of the study is shown in Fig. 1

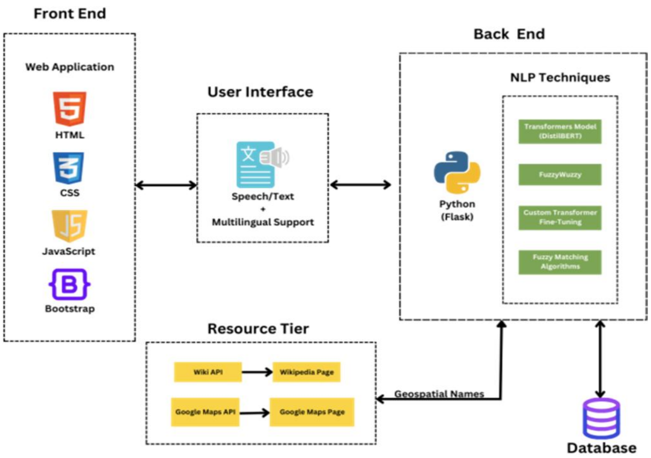

Fig 1. Block Diagram for the architecture of the proposed system

Data sources for training and fine-tuning:

To fine-tuning DistilBERT on a specific task, we need to have a labeled dataset that contains the input text and the output labels for the task. The dataset should be relevant to the domain and the objective of the task, and should be large enough to cover the diversity and complexity of the natural language. The dataset should also be split into training, validation, and test sets, to train, evaluate and test the model's performance.
For our system, we used the [1]CoNLL-2003 Dataset to fine- tune DistilBERT for geographic entity recognition. The CoNLL-2003 Dataset contains news articles from the Reuters Corpus, annotated with four types of entities: Person, Organization, Location, and Miscellaneous. The dataset has 14,987 sentences in the training set, 3,466 sentences in the validation set, and 3,684 sentences in the test set. The dataset covers various domains, such as

politics, sports, business, and culture, and contains various types and granularities of place names, such as countries, cities, regions, and landmarks.
We also created a reference dataset with country names and their canonical name and type (State, Country, District). The reference dataset contains 195 countries and their corresponding information, such as capital, population, area, currency, and official languages. The reference dataset is used to match the extracted place names with their canonical forms, and to provide additional information and verification for the place names

Data organization and custom Transformer

Geographical hierarchy is a way of organizing geospatial data according to the administrative or political divisions of a region or a country. Geographical hierarchy can help to structure and classify geospatial data, and to resolve ambiguities and conflicts among place names. Geographical hierarchy can also provide additional information and context for geospatial data, such as the population, area, or GDP of a place.
We used the CoNLL-2003 Dataset1 to fine-tune DistilBERT for geographic entity recognition. The CoNLL-2003 Dataset contains news articles from the Reuters Corpus, annotated with four types of entities: Person, Organization, Location, and Miscellaneous. The dataset covers various domains, such as politics, sports, business, and culture, and contains various types and granularities of place names, such as countries, cities, regions, and landmarks.
We also performed a custom transformer fine-tuning, using a masked language modeling objective, to adapt [6] DistilBERT to the geospatial domain and to handle complex and nested place names. We used a subset of the GeoNames Database2, which is a geographical database that contains over 11 million place names and their corresponding geographical features, such as coordinates, elevation, population, etc. We masked some of the place names in the GeoNames Database, and used DistilBERT to predict the masked tokens, using the surrounding tokens and the geographical features as context. We fine-tuned DistilBERT on this task, using the transformers library3, which is a comprehensive and popular library for natural language processing in Python, that provides various transformer models and tools.

Integration of multilingual speech-to-text

This feature allows the user to [5] speak the query instead of typing it, and supports multiple languages, such as English, 中文, 日本語, Español, Français, Deutsch, etc. The user can click the microphone icon on the web application to enable speech-to-text input, and choose the language of the query from the language selector dropdown. The system uses the Web Speech API1 to convert the user's speech into text, and displays the text in the query search bar. The system also uses the Google Translate API2 to translate the query into English, if the query is not in English, and sends the translated query to the back end for processing.

Fuzzy matching algorithm

Fuzzy matching, a cornerstone in handling spelling variations and semantic nuances, utilizes techniques such as Levenshtein distances and token-based comparisons across diverse domains. Python's Fuzzywuzzy library, featuring functions like simple ratio and token sort ratio, streamlines the process of string comparison, crucial for tasks like toponym resolution. These methods enable the matching of extracted place names with geospatial entities, offering a robust solution for navigating through discrepancies in textual data. By implementing predefined thresholds based on similarity scores, only matches surpassing a designated level of accuracy are considered, bolstering the reliability of decision-making processes in applications requiring approximate string matching.
Incorporating fuzzy matching techniques not only enhances the efficiency of data processing but also ensures accuracy in various contexts. By leveraging algorithms such as Levenshtein distances and token-based comparisons, Fuzzywuzzy enables effective handling of spelling errors and semantic disparities. Particularly in tasks like toponym resolution, where accurate matching of place names to geospatial entities is paramount, these techniques offer invaluable assistance. The implementation of predefined thresholds based on similarity scores serves to filter out matches below a certain accuracy threshold, thereby optimizing decision-making and ensuring robustness in applications reliant on approximate string matching.

IV. RESULTS

Metrics including accuracy, precision, recall, and F1-score are used to assess the system's performance. Recall measures comprehensiveness, whereas precision assesses accuracy in identifying geospatial things. The entire performance is represented by the F1-score, which goes from 0 to 1. The precision with which the system resolves geographical entities to their most likely locations is measured by accuracy.

TABLE II. CONFUSION MATRIX TABLE FOR COMPARISON OF MODELS

System	Performance matrix comparison			
	Precision	Recall	F1-score	Accuracy
Proposed system	0.93	0.89	0.91	0.87
spaCy	0.81	0.76	0.78	0.72
Geotest	0.68	0.64	0.66	0.62
Google Maps API	0.75	0.71	0.73	0.69

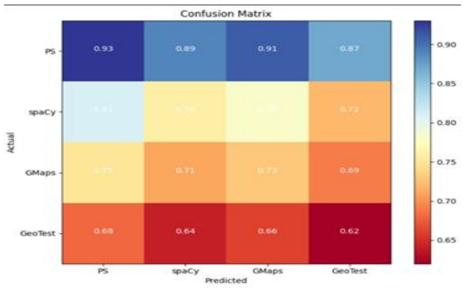

Fig 2. Confusion matrix between the proposed system and the existing alternatives

The table shows that the system outperforms the baseline systems in terms of all the metrics, and achieves the highest precision, recall, F1-score, and accuracy, and the lowest error rate. The graph shown in Fig 2 visualises the confusion matrix depicted in table 2. This indicates that our system is more effective and accurate in extracting and resolving geospatial entities from text, using the DistilBERT and Fuzzywuzzy models, and the geographical hierarchy and context analysis methods.
We assess our system using the GeoCorpora Dataset against spaCy, Google Maps API, and GeoText. This collection consists of one thousand news stories from various disciplines that have been geospatially annotated. For example, our system allows users to enter travel planning queries for destinations in Europe, extracting and resolving geographical entities to deliver detailed information with links to Wikipedia and Google Maps. Furthermore, our system's performance was evaluated by 20 individuals from different disciplines in a user research, which yielded qualitative insights into its efficacy based on relevance, accuracy, completeness, clarity, and usefulness.

V. CONCLUSION

In conclusion, the proposed system marks a significant leap in geospatial query processing. It encompasses a diverse array of cutting-edge technologies and methodologies to provide an accurate, context-aware, and user-friendly solution for geospatial queries. By offering multilingual voice recognition, the system promotes inclusivity, ensuring that users with varying language preferences can effortlessly interact with it. The inclusion of Google Maps and Wikipedia APIs further enriches the user experience by providing visual representations and additional context, ultimately contributing to a more profound user comprehensibility of query results. The implications of this work transcend several domains, including information retrieval, natural language processing, and geographical information systems, offering a valuable tool for users seeking accurate and context-aware information about places. The system's contributions are evident both in terms of technological advancements, particularly in the field of NLP and geospatial data processing, and in terms of the improved user experience. It not only caters to a broader and more diverse user base but also serves as an accessible and user-friendly platform for users from various linguistic backgrounds and needs. The promising nature of this system, combined with the ongoing efforts to address its limitations, paves the way for a brighter future in the domain of geospatial queries, with the potential for real-world applications and positive societal impacts.

REFERENCES

[1] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147.
[2] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[3] Bart Kosko and Satoru Isaka. Fuzzy Logic: The binary logic of modern computers often falls short when describing the vagueness of the real world. Fuzzy logic offers more graceful alternatives.
[4] Jan Oliver Wallgrun, Morteza Karimzadeh, Alan M. MacEachren and Scott Pezanowski 2017. GeoCorpora: building a corpus to test and train microblog geoparsers at International Journal of Geographical information Science.
[5] Vinnarasu A,Deepa V. Jose 2019. Speech to text conversion and summarization for effective understanding and documentation at International Journal of Electrical & Computer Engineering (2088-8708), 2019, Vol 9, Issue 5, p3642
[6] Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. ACM Trans. Inf. Syst. 36, 4, Article 40 (October 2018), 27 pages. https://doi.org/10.1145/3202662

[7] Breunig M, Bradley PE, Jahn M, Kuper P, Mazroob N, Rösch N, Al-Doori M, Stefanakis E, Jadidi M. Geospatial Data Management Research: Progress and Future Directions. ISPRS International Journal of Geo-Information. 2020; 9(2):95. https://doi.org/10.3390/ijgi9020095.

[8] Lin Y, Guo J. Fuzzy Geospatial Object-Based Membership Function Downscaling. Remote Sensing. 2023; 15(7):1911. https://doi.org/10.3390/rs15071911

[9] D. Krleža and K. Fertalj, "Graph Matching Using Hierarchical Fuzzy Graph Neural Networks," in IEEE Transactions on Fuzzy Systems, vol. 25, no. 4, pp. 892-904, Aug. 2017, doi: 10.1109/TFUZZ.2016.2586962. keywords: {Unified modeling language;Pattern matching;Fuzzy logic;Biological neural networks;Encoding;Databases;Fuzzy neural networks (FuNN);graph matching;hierarchical neural networks;noise resilience;recursive neural networks.

[10] C. Zhang, X. Zhang, W. Jiang, Q. Shen and S. Zhang, "Rule-Based Extraction of Spatial Relations in Natural Language Text," 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 2009, pp. 1-4, doi: 10.1109/CISE.2009.5363900. keywords: {Natural languages;Data mining;Geography;Information analysis;Information science;Cities and towns;Information systems;Natural language processing;Machine learning algorithms;Layout},