공학석사학위논문

GeoQA 시스템 성능 향상을 위한 장소 관련 질문 분류

Classification of Place-related Questions for Enhanced GeoQA system

2023년 8월

서울대학교 대학원

건설환경공학부

이 서 형

GeoQA 시스템 성능 향상을 위한 장소 관련 질문 분류

Classification of Place-related Questions for Enhanced GeoQA System

지도 교수 유 기 윤

이 논문을 공학석사 학위논문으로 제출함

2023년 06월

서울대학교 대학원

건설환경공학부

이 서 형

이서형의 공학 석사 학위논문을 인준함

2023 년 06월

위 원 장 ____________________ (인)

부위원장 ____________________ (인)

위 원 ____________________ (인)

Abstract

Classification of Place-related Questions for Enhanced GeoQA system

Seohyung Lee
Department of Civil and Environmental Engineering
The Graduate School
Seoul National University

Question classification (QC) plays a crucial role in delivering accurate and relevant answers to users by identifying the topic of a question. In Question Answering (QA) systems, accurate classification of questions is essential for providing precise responses that effectively address users' inquiries and ensure the retrieval of relevant answers from the information retrieval process.

While significant progress has been made in improving QC across various domains by categorizing questions into coarse and fine clas

ses, the field of geography presents unique challenges when it comes to answering questions related to geographic entities or concepts that require spatial operations. Place-related questions encompass a diverse range of topics, making their classification particularly challenging. Although efforts have been made in the geographical domain to determine the geographical nature of questions, analyze structural patterns in geographical questions for query generation, and classify geographic questions based on latent topics, accurately identifying specific topics within place-related questions remains a gap in current research.

The distinct characteristics of the geography field, such as the multitude of possible topics and the interplay between geographic and non-geographic elements in questions, contribute to the complexity. This study focuses on place-related questions to address these challenges and bridge the existing gap in current studies.

Therefore, the objective of this research is to develop an approach that accurately classifies specific topics within place-related questions, aiming to enhance the effectiveness of question classification in Geographic QA systems (GeoQA). The findings from this study will provide valuable insights for improving question answering system

s in the context of geography.

This thesis presents a methodology for classifying place-related questions in the domain of geography using predefined fine-grained topics. The goal is to accurately identify the specific topic of interest within place-related questions, which is crucial for providing relevant and informative responses to natural language queries about geographic locations. Accurately identifying the specific topic becomes essential in delivering precise information. For instance, consider the question "Henderson, TN zone." It is important to classify this question under the "Locator" topic, as it remains unclear whether the user is referring to the time zone or the hardiness zone of Henderson, TN.

By utilizing predefined fine-grained topics, this question classification model for place-related questions enhances the classification process and enables the system to capture the nuanced aspects of place-related questions. The findings from this research will contribute to the development of more accurate and effective question answering systems in the field of geography, facilitating better information retrieval.

To accomplish this objective, a total of 3,025 randomly selected place-related questions from the MS MARCO dataset were labeled according to 42 fine-grained topics. Subsequently, a BERT model was fine-tuned to classify place-related questions based on the relevant topics. The proposed QC model achieved an overall training accuracy of 87.9% and a test accuracy of 86.6%, demonstrating its effectiveness in classifying place-related questions. The proposed multi-class question classification model presented in this thesis provides an approach to question classification for place-related questions focusing on relevant topics. This work can enhance current GeoQA system by performing better retrieval of information from the external knowledge base.

Keywords: GeoQA, GeoQA Dataset, Close-domain QA system, Question Classification, Multi-class Question Classification

Student Number: 2021-22606

Table of Contents

List of Table

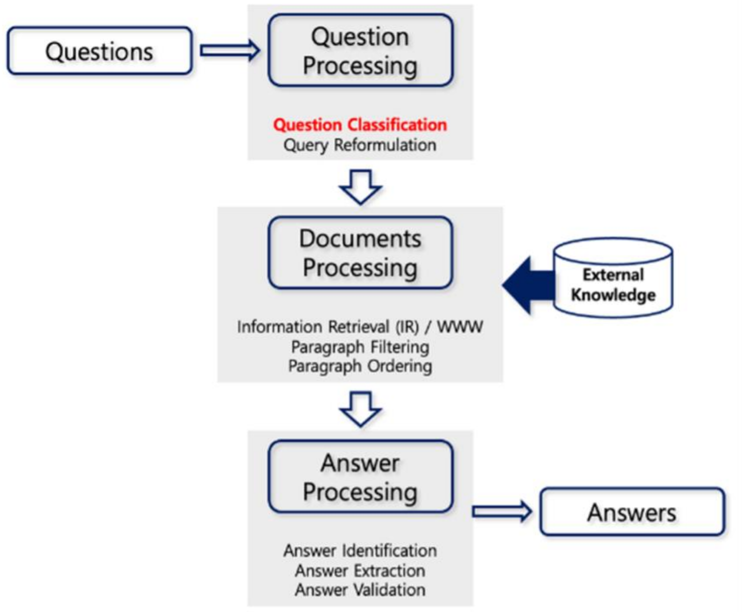
List of Figure

1. Introduction

1.1 Research Background and Purpose

The rapid advancement of computer technology, coupled with the widespread use of the internet and the significant growth of electronic information on the web, has raised concerns about efficiently and accurately retrieving the desired content from a vast amount of information (Wang & Qu, 2017). One of the most widely used applications for retrieving information is question answering (QA), which is an important task of natural language processing. The primary goal of the QA system is to provide an accurate response, rather than a collection of documents, to the user's arbitrary question expressed in natural language.

Question Answering (QA) systems can have different architectures, but they generally follow a common framework to answer questions ([Figure 1-1]). This framework involves several steps: determining the expected answer type through question classification, retrieving candidate answers from an external knowledge, identifying the correct answer within the retrieved information based on the expected answer type(s), and presenting the best answer to the user (Allam & Haggag, 2012).

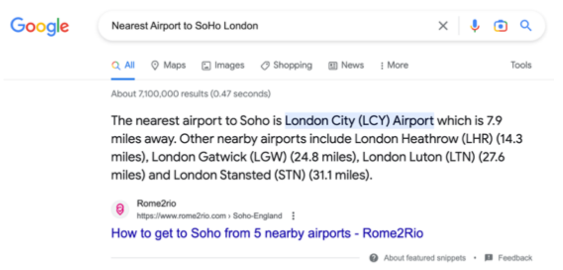
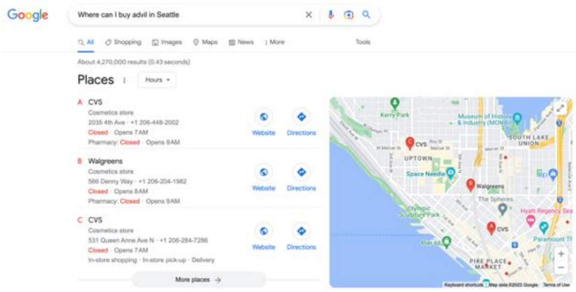


[Figure 1-1] General Architecture of Question Answering (QA) Systems (Allam & Haggag, 2012)

Aforementioned QA system aims to provide an accurate response to the user's question, based on prior knowledge of the expected answer type (Mallikarjuna & Sivanesan, 2022). To accomplish this, the question classification (QC) module plays a crucial role in providing relevant information, making it an essential component of the QA system architecture (Mohasseb et al., 2018). The objective of the QC procedure is to recognize the category of the question asked, which allows other modules to locate and verify an answer that is pertinent to the user's query. This is significant because determining the type of question can limit what is deemed as relevant information and can provide the answer (Allam & Haggag, 2012).

Before initiating this thesis, it is important to provide an overview of QA systems. QA systems can be categorized based on the type of information they utilize to answer questions. One such category is Text-based Question Answering (TextQA) which relies on unstructured text data to answer questions. TextQA uses machine learning algorithms to identify and extract relevant information from the text, providing an answer to the question asked. Another QA system is Knowledge Base Question Answering (KBQA), which focuses on structured knowledge bases, such as databases or ontologies, to answer questions. The knowledge base contains organized and structured information that can be easily queried by semantic parsing techniques used in KBQA systems. To summarize, TextQA systems are generally more flexible and can answer a broader range of questions, while KBQA systems are more specialized and provide more accurate and structured answers to specific types of questions. Research on GeoQA are mainly relying on knowledge base with the recent advances in pre-training language models, substantial improvements are in progress in KBQA models (Yang et al., 2023). Researchers have also developed a HybridQA approach that combines both structured and unstructured information (Oguz et al., 2022). This approach unifies the benefits of TextQA and KBQA to provide an enhanced question-answering experience. [Table 1-1] depicts the results of place-related questions that were searched on Google's open-domain search engine in May 2023.

[Table 1-1] Types of QA systems by Information Retrieval (IR) Source for Geographical Question

Question Input	Information Retrieval (IR) Source	Answers from Open-Domain Search Engine (Google)
Nearest Airport to SoHo London	TextQA	
Where can I buy advil in Seattle	KBQA	

Yet, it is still important to note that preliminary research on GeoQA relies on Knowledge base (Yang et al., 2023). QA systems can be broadly classified into two categories: open-domain QA systems and close-domain QA systems. Open-domain QA systems like ChatGPT and Siri are designed to answer a wide range of questions on various topics, while close-domain QA systems are intended to answ er questions with high accuracy in a specific field or domain (i.e., biomedical domain and geographical domain). The quantity of geographic information is increasing at an annual growth rate of approximately 20%, and such information makes up a significant portion of data on the internet (Lee & Kang, 2015). Despite the significant advancements in open-domain question answering, QA systems still face challenges when it comes to answering questions related to geographic entities or concepts that require spatial operations (Mai et al., 2021). To overcome these issues, researchers on GeoQA are building a knowledge-based QA system that can provide accurate answers from the structured information.

Over time, extensive research has been conducted by scholars in the field of Geographical Question Answering (GeoQA) (Hamzei et al., 2020; Hamzei, 2021; Hamzei et al., 2022; Mishra et al., 2010; Punjani et al., 2018; Xu et al., 2020). Previous research on GeoQA can be categorized into explaining the challenges of GeoQA system (Mai et al., 2021), constructing GeoQA datasets (Punjani et al., 2018; Xu et al., 2020), analyzing GeoQA datasets (Hamzei et al., 2020), explaining the challenges of Geo-analytical QA system (Scheider et al., 2021), categorizing geographic questions and answer pairs (Hamz ei et al., 2020; Mai et al., 2021; Xu et al., 2022), proposing models for transforming concepts (Xu et al., 2022), building overall pipeline for GeoQA system (Punjani et al., 2018), or introducing methods to translate place-related questions into GeoSPARQL queries (Hamzei, 2021; Hamzei et al., 2022)

Several studies mentioned above have classified geographic questions. For example, Hamzei et al. (2020) used a classification method based on a semantic encoding approach which recognizes the functional roles of the concepts and generates queries for a GeoQA engine or analyzes question types (e.g., factoid or analytical). However, the main purpose of classification of these questions using semantic encoding was to build a template for query generation, not to present the main topic for each question.

Accurately classifying questions into topics is a crucial element in the design and development of effective GeoQA systems, preceding any information retrieval (IR) process [Figure 1-1]. The focus of this study is to classify questions based on predefined topics, aiming to refine the scope of potential answers and enhance the precision of information retrieval. By achieving accurate question classification, the performance and user experience of GeoQA systems can be significantly improved. Hence, question classification holds significant relevance in advancing GeoQA systems.

To achieve this objective, this study focuses on the development of the question classification (QC) module for the GeoQA system. The QC module aims to accurately classify place-related questions by inferring their topics from the questions asked. The proposed QC module utilizes a machine learning method to identify the topic of the questions.

1.2 Related Work

This section aims to emphasize the significance of Question Classification (QC) within GeoQA and address research gaps through a comprehensive review of the field. While specific QC studies in GeoQA systems are limited, it is important to recognize the importance of QC in the broader domain of QA systems. GeoQA research covers key aspects such as dataset construction, query generation using SPARQL/Geo-SPARQL, and data analysis. Constructing appropriate datasets is crucial for training and evaluating QC models in GeoQA systems. Additionally, SPARQL/Geo-SPARQL query generation techniques enable retrieval of relevant information from geospatial databases or knowledge graphs. Data analysis techniques aid in understanding patterns and characteristics of geographic data, enhancing GeoQA system performance. Research gaps and motivations for further investigations in GeoQA are aimed to be identified through an examination of existing literature in these areas ([Table 1-2]).

[Table 1-2] GeoQA Research Scope Overview

Area	Paper	Description
Dataset Construction	Punjani et al. (2018)	Generated GeoQuestions201, which is created by Junior level students enrolled in AI course
	Xu et al. (2020)	Generated GeoAnQu, which is a dataset mainly for geo-analytic questions
Data Analysis	Hamzei et al. (2020)	Investigate the relationships between geographic questions and answers by categorization by alphanumeric representation for their semantic encoding
	Xu et al. (2020)	After generating GeoAnQu, they compare three major QA corpora (MS MARCO, Geo-Questions201, GeoAnQu) used in GeoQA research
Query (SPARQL/Geo-SPARQL) Generation	Hamzei (2021)	Translated geographic ques-tions to GeoSPARQL with the intermediate logical represent-tation form classified by semantic encoding using MS MARCO corpus
	Hamzei et al. (2022)	Improves existing methods in translating geographic questions to structured queries by BERT embedding to perform ontology mapping using GeoQuestions-201

This research focuses on the critical stage of question classification within the overall QA system and its significance in accurately c lassifying place-related questions into topics. However, the field of GeoQA lacks extensive exploration of the question classification (QC) module, creating a research gap. To address this gap and explore the potential of QC in GeoQA, ongoing studies in GeoQA are examined, and how the QC module is being studied in other QA domains is explored.

Existing literature in GeoQA reveals various research areas, including dataset construction, SPARQL/Geo-SPARQL query generation, and data analysis. These areas play crucial roles in training and evaluating QC models for GeoQA systems. Dataset construction is vital for obtaining suitable data to develop and test QC models tailored to the geography domain. SPARQL/Geo-SPARQL query generation techniques enable retrieval of relevant information from geospatial databases or knowledge graphs. Data analysis techniques contribute to understanding geographic data patterns, enhancing GeoQA system performance.

While direct research on the QC module in GeoQA systems is limited, broader QA studies have explored and developed QC models. Examining these studies provides valuable insights into designing and implementing QC modules for GeoQA systems.

To bridge the gap between GeoQA and existing QC research, this thesis introduces a fine-tuning approach using Bidirectional Encoder Representation from Transformer (BERT) (Devlin et al., 2019). BERT, pretrained on the Masked Language Model Task and Next Sentence Prediction Task with a diverse corpus from various domains incorporates a unique Masked Language Model approach (Sun et al., 2020). This approach allows BERT to predict randomly masked or substituted words, differentiating it from traditional bidirectional language models. By leveraging BERT's state-of-the-art capabilities through fine-tuning, this study aims to address the challenges of place-related question classification in GeoQA systems.

1.2.1 Overview of GeoQA

GeoQA plays a pivotal role in delivering precise and pertinent answers to queries specific to the geographic domain, underscoring its crucial significance in the realm of question answering systems. To provide a comprehensive understanding of the GeoQA field, it is important to clarify the two types of geographic questions: factoid (referred to as Geo-factoid) and analytical (referred to as Geo-analytic). These question types, although not directly utilized in the question classification module in this thesis, are commonly discussed in the literature, and contribute to the broader understanding of geographic question types in the field of GeoQA.

Factoid questions primarily aim to retrieve specific location-related information, as exemplified by questions such as "What is the capital of France?" or "What is the population of Tokyo?" On the other hand, more extensive analysis is required for analytical questions beyond straightforward location-based facts. Complex inquiries such as "What is the impact of climate change on crop yields in Southeast Asia?" or "What is the best location for a new retail store based on demographic data?" are involved in these questions. By discussing these question types, a framework for categorizing and understandi ng different types of questions commonly encountered in the geographic domain is provided.

Although the question classification module from this thesis does not explicitly focus on factoid or analytical questions, the explanation of these question types contributes to the broader landscape of GeoQA research. It allows for a better appreciation of the nuances and complexities involved in classifying and answering geographic questions. Moreover, the knowledge of these question types can inform future research directions and help researchers explore more targeted approaches for addressing factoid and analytical questions within the GeoQA framework (Xu et al., 2020).

Moreover, it is important to contextualize GeoQA within the broader field of question answering (QA) in natural language processing. QA systems aim to extract precise answers when users pose questions. While open-domain QA systems like ChatGPT can handle questions from various domains, close-domain QA systems specialize in specific areas such as medicine, cooking, or automobiles (Khilji et al., 2020). GeoQA falls into the category of close-domain QA systems, focusing specifically on answering geographic questions. It is designed and tailored to handle inquiries related to geography, includi ng topics such as locations, distances, maps, and other spatial information. Understanding the distinction between open−domain and close−domain QA systems provides valuable context for appreciating the unique challenges and opportunities in developing GeoQA systems.

1.2.2 GeoQA Datasets: Review and Analysis

In this section, the review and analysis of the existing datasets commonly employed in GeoQA research are delved into, highlighting the crucial role of dataset analysis in understanding the characteris tics and patterns of the questions encountered in the GeoQA domain.

GeoQA researchers have access to a range of datasets, both ge neral and specific to GeoQA. To qualify as a GeoQA dataset, the que stions must involve geographic entities, concepts, or relations. Thes e datasets serve as valuable resources for training and evaluating G eoQA systems, enabling researchers to improve the performance an d accuracy of their models.

Among the various datasets utilized in the GeoQA field, three pr ominent ones are MS MARCO, GeoQuestions201, and GeoAnQu. Eac h dataset has its unique attributes and focuses, encompassing divers e question types and complexities. [Table 1−3] provides a compreh ensive overview of the differences between these datasets, highligh ting their distinctive characteristics and features.

[Table 1-3] Datasets Typically Selected on GeoQA Research

Dataset	Size	Source	Focus	Category
MS MARCO	1,010,916 anonymized questions with human generated answer	Bing	General QA	(1) Description (2) Numeric (3) Entity (4) Location (5) Person
GeoQuestions 201	201 questions	Students enrolled in an AI course	GeoQA	Geo-factoid
GeoAnQu	429 questions	(1) 100 scientific articles (2) text-books on GIScience and GIS	GeoQA	Geo-analytical

Through the analysis of these datasets, valuable insights are gained into the nature of geographic questions and the challenges they pose. This analysis plays a crucial role in the development and refinement of the QC module, as common patterns, topics, and variations within the GeoQA datasets can be identified. The knowledge acquired from dataset analysis can be leveraged to enhance the classification accuracy and effectiveness of the QC module, thereby improving the overall performance of GeoQA systems.

In the subsequent sections, a deeper exploration will be undertaken into the review and analysis of these datasets, with a focus on e xamining their contents, structure, and implications for the development of the QC module. By harnessing the power of dataset analysis, the aim is to advance the state of GeoQA research and contribute to the ongoing efforts in building more accurate and efficient question answering systems for geographic queries.

MS MARCO (Bajaj et al., 2018) is a large-scale dataset for reading comprehension and question answering that contains 1,010,916 passages and their corresponding questions and answers. While not designed specifically for GeoQA, researchers extract "LOC" annotated questions and answers to identify geographic questions. Hamzei et al. (2020) found that approximately 65% of the annotated questions were geographic in nature.

GeoQuestions201 (Punjani et al., 2018) is a GeoQA-specific dataset of 201 geo-referenced questions on various geographic topics, collected from junior-level students in an Artificial Intelligence course. The dataset covers cities, countries, landmarks, and other spatial concepts.

GeoAnQu (Xu et al., 2020) is another GeoQA-specific dataset that contains 429 questions related to geographic locations, landmarks, and spatial relations. Unlike GeoQuestions201, the questions in G eoAnQu are formulated based on readings of scientific textbooks on GIScience and GIS. The dataset also includes multiple candidate answers for each question, making it more robust and detailed.

These datasets have been analyzed in several papers from different perspectives. One such paper by Hamzei et al. (2020) aimed to investigate the relationships between geographic questions and answers by categorizing them. To analyze the content of place-related questions and answers, the authors extracted patterns of place-related information through semantic encoding. They extended the semantic encoding schema by Edwardes & Purves (2007) to relatively short place questions and answers from MS MARCO. The categorization of place-related questions and human-generated answers was analyzed through semantic encoding and contextual semantic embedding. The authors computed clusters of semantic representations, modeled the encodings of question-answer pairs as 1024-dimensional vectors, and calculated the values of the vectors based on the Jaro similarity between the semantic encodings and the selected encodings. They then applied both k-means clustering and ELMo representations to measure whether the semantic representations retained the contextual similarity of the sentences. Finally, they manually inter preted the most frequent encodings in each cluster to derive the categories of place−related questions and answers. In essence, the researchers successfully identified the top−five most common place−related semantics for each of the six primary elements in both the question−and−answer datasets ([Table 1−4]). They also scrutinized the prevalent encoding patterns present in both questions and answers, and evaluated the proportion of these patterns that were classified as non−spatial, spatial, non−geographical, or ambiguous based on the types of questions, explicit or implicit localization, non−geographical and unanswered for the types of answers.

[Table 1−4] Top Five Frequent Place related Semantics Extracted from the Dataset from Hamzei et al. (2020)

Type	in Questions	in Answers
Place name	California (1393) Texas (1391) Florida (1148) New York (895) Illinois (692)	United States (4845) California (1482) Texas (989) Florida (961) New York (894)
Activity	Buy (340) Go (296) Play (120) Build (88) See (86)	Go (64) Run (62) Leave (55) Build (53) Move (38)

Type	in Questions	in Answers
Spatial Relation	In (3916) Near (153) At (142) On (109) Between (38)	In (10851) On (379) At (362) Near (275) Between (251)
Type	County (11702) State (2291) City (1630) Zone (745) Region (653)	City (1714) State (1653) County (1438) Area (882) Region (758)
Situation	Find (1412) Live (746) Have (662) Grow (321) Originate (237)	Find (695) Have (405) Live (305) Include (231) Originate (125)
Quality	Largest (242) Biggest (106) Highest (97) Expensive (56) Beautiful	Largest (121) Census-designated (68) Metropolitan (54) Small (46) Coastal (36)

Xu et al. (2020) compared two major QA corpora used in GeoQA research and suggested a new dataset called GeoAnQu, primarily designed for geo-analytic questions. The authors compared MS MARCO, GeoQuestions201, and GeoAnQu datasets against each other, using semantic encodings and parsing methods to extract semantic information from the questions. They then compared the three corpora at the word, phrase, and sentence level, using word clouds and quantifying syntactic patterns.

The papers discussed above predominantly employ alphanumeric encodings to extract semantics from place-related questions and answers, primarily focusing on appearance percentage. While these methods facilitate the classification of frequently asked questions and types of place-related information, they do not provide a means to identify the main topic of the questions. Additionally, each dataset possesses unique features and focuses, allowing researchers to select an appropriate dataset based on factors such as size, question source, expertise level, and focus.

Among the available datasets, the MS MARCO dataset stands out as the only one obtained from a real-world question answering engine (Yang et al., 2023). This unique attribute renders it highly advantageous for identifying frequently asked topics in the questions. Consequently, this thesis leverages the "LOC" annotated questions from the MS MARCO dataset to construct the QC module of the GeoQA system. The utilization of this dataset aims to enhance the classification and understanding of geographic questions within the GeoQA framework.

1.2.3 Query Generation

Query generation is an essential module that follows the question classification (QC) process and aims to create queries for information retrieval (IR). In the context of GeoQA, two papers have proposed methodologies to enhance the accuracy of query generation for geographic questions.

[Table 1-5] Alphanumeric Encoding Schema by Semantic Encoding Method used in Hamzei (2020)

Semantic Type	Part-of-speech	Code	Semantic Type	Part-of-speech	Code
where	WH-word	1	Place name	noun	n
what	WH-word	2	Place type	noun	t
which	WH-word	3	Object	noun	o
when	WH-word	4	Quality	adjective	q
how	WH-word	5	Activity	verb	a
whom	WH-word	6	Situation, and event	verb	s
whose	WH-word	7	Spatial relationship	preposition	r
why	WH-word	8			
Examples of Alphanumeric Encoding Schema					**Pattern**
What is the snowiest place in Canada?					2qtrn
Where is Mount Rainier					1n

Hamzei et al. (2020) proposed a method that translates place-related questions and answers using semantic encodings. They identified primary elements such as place names, place types, activities, s ituations, qualitative spatial relationships, and qualities. The questions were differentiated based on WH−words and mapped to corresponding place semantics. The researchers used an alphanumeric encoding schema to analyze and categorize the GeoQA datasets according to their structural patterns ([Table 1−5]). Consequently, the alphanumeric encoding schema lacked the ability to disambiguate specific semantic types, resulting in potential encoding errors.

Punjani et al. (2018) introduced a QA engine for geospatial questions, GeoQA, which translates natural language questions into GeoSPARQL queries. They implemented two reusable QA components, Qanary and Frankenstein, to achieve this. The architecture of their GeoQA system included a dependency parse tree generator, concept identifier, instance identifier, geospatial relation identifier, and SPARQL/GeoSPARQL query generator and executor. [Table 1−6] showcases examples of question patterns in natural language and their corresponding SPARQL/GeoSPARQL query templates.

[Table 1-6] Examples of question patterns (C for "Concept", I for "Instance", and R for "Geospatial Relation") for each natural language questions and its corresponding SPARQL/GeoSPARQL query in Punjani et al. (2018)

Pattern	Example Questions	SPARQL/GeoSPARQL Templates
CRI	Which rivers cross Limerick?	SPARQL: select ?x where { ?x rdf:type _Concept. ?x _Relation _Instance. } GeoSPARQL v1: select ?x where { ?x rdf:type _Concept; geo:hasGeometry ?xGeom. ?xGeom geo:asWKT ?xWKT. _Instance geo:hasGeometry ?iGeom. ?iGeom geo:asWKT ?iWKT. FILTER(_Relation(?xWKT, ?iWKT)) } GeoSPARQL v2: select ?x where { ?x rdf:type _Concept; geo:hasGeometry ?xGeom. ?xGeom geo:asWKT ?xWKT. ?instance owl:sameAs _Instance; geo:hasGeometry ?iGeom. ?iGeom geo:asWKT ?iWKT. FILTER(_Relation(?xWKT, ?iWKT)) }

Pattern	Example Questions	SPARQL/GeoSPARQL Templates
CRIRI	Which churches are close to the Shannon in Limerick?	select ?x where { ?x rdf:type _Concept; geo:hasGeometry ?xGeom. ?xGeom geo:asWKT ?xWKT. _Instance1 geo:hasGeometry ?i1Geom. ?i1Geom geo:asWKT ?i1WKT. _Instance2 geo:hasGeometry ?i2Geom. ?i2Geom geo:asWKT ?i2WKT. FILTER(_Relation1(?xWKT, ?i1WKT) && _Relation2(?i1WKT, ?i2WKT)) }

CRC	Which restaurants are near hotels?	select ?x where {
		?x rdf:type _Concept1;
		geo:hasGeometry ?xGeom.
		?xGeom geo:asWKT ?xWKT.
		?y rdf:type _Concept2;
		geo:hasGeometry ?yGeom.
		?yGeom geo:asWKT ?yWKT.
		FILTER(_Relation(?xWKT, ?yWKT))
		}

IRI	Is Hampshire north of Berkshire?	ASK where {
		_Instance1 geo:hasGeometry ?iGeom1.
		?iGeom1 geo:asWKT ?iWKT1.
		_Instance2 geo:hasGeometry ?iGeom2.
		?iGeom2 geo:asWKT ?iWKT2.
		FILTER(_Relation(?iWKT1, ?iWKT2))
		}

In addition, Hamzei et al. (2022) proposed an approach to properly formulate the intermediate representation of natural language questions, capturing their semantics without considering the technical features of the destination query language. They aimed to leverage domain knowledge by translating place-related questions into GeoSPARQL queries. This involved two steps: encoding natural language place-related questions and object-based conceptualization of place for location, place (object), event, properties, and relations. The logical representation enabled the translation of place-related questions to GeoSPARQL queries through concept identification, ontology mapping, and dynamic query generation. The researchers utilized BERT embeddings for ontology mapping and evaluated their model's performance in mapping place types and properties. Their focus was primarily on classifying place-related questions based on structural patterns to create query templates, rather than classifying them according to the main topic of the question. Thus, it is important to acknowledge that the primary objectives of these papers differ from the goals of this thesis.

In summary, query generation plays a crucial role in the GeoQA system by creating queries for information retrieval from the underlying data sources. While existing papers have proposed methodologies to improve query generation for geographic questions, their primary focus differs from the objectives of this thesis. In light of the ex isting research, this thesis aims to contribute to the GeoQA field by developing a question classification (QC) module. Unlike the aforementioned papers, the purpose of the QC module is to classify questions into their main topics, thereby enhancing the performance of GeoQA systems during the information retrieval process. By accurately categorizing questions, the QC module helps the QA system find better answers and deliver more precise and targeted results. The development of the QC module for GeoQA systems is motivated by the need to enhance the search capabilities of GeoQA systems, enabling users to obtain more relevant and accurate information regarding geographic queries. By understanding the main topics of the questions, the QC module facilitates more effective retrieval of spatial and geographical data, thereby enhance the overall GeoQA systems.

1.2.4 Question Classification

The classification of questions based on their main topic plays a fundamental role in natural language processing question-answering systems. By identifying the main topic, the system can retrieve pertinent information from external resources, enabling it to provide precise and relevant answers. This is particularly crucial in search engines or personalized recommendation systems, where understanding the intent and preferences of the user plays a significant role (Zhou et al., 2012). Accurate classification of questions based on the main topic enables the system to filter out irrelevant information and focus on delivering tailored answers that address the core purpose of the question. This leads to increased user satisfaction and improved system performance. The objective of this thesis is to classify place-related questions into specific topics using question classification methodologies employed in other fields. This section provides an overview of question classification.

Question classification has been extensively studied in the natural language processing (NLP) field, particularly in open-domain question answering (QA) systems. In the related work of question classification, the classical approach to addressing supervised NLP tasks will be briefly discussed. In the classical approach, text classification model is done by two steps: feature extraction and classification. Term−Frequency – Inverse Document Frequency (TF−IDF) (Salton & Buckley, 1988) and word2Vec (Mikolov et al., 2013) were commonly used methods for the feature extraction. In depth, TF−IDF calculates the importance of a word in a document within a collection of documents by weight. Then, it assigns a higher weight to words that appear frequently in a document while appearing less frequently in the entire document collection (González−Carvajal & Garrido−Merchán, 2023). This way, TF−IDF captures the relative importance of words in a document. Word2Vec (Mikolov et al., 2013) is a well−known word embedding technique that maps words into dense vector representations in a continuous vector space. It captures the semantic meaning of words by considering the context in which they appear. The vectors can be used as features for text classification tasks. After extracting features using TF−IDF or word2Vec, various machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), or Random Forest are applied for the classification step.

Previous research on question classification (QC) has mainly focused on using supervised machine learning methods to build open−

domain question answering systems. Li and Roth (2002), developed a list of semantic features and used multi-class classification to categorize questions into six broad and 50 fine-grained classes.

Zhang and Lee (2003) compared the performance of five different machine learning algorithms, including Support Vector Machine (SVM), Nearest Neighbors (NN), Naïve Bayes (NB), Decision Tree (DT), and Sparse Network of Winnows (SNoW), using the TREC10 dataset. They trained each model on different training dataset sizes and found that SVM outperformed the other models in most of the experimental results for both the coarse-grained and fine-grained evaluation.

Zhou et al. (2012) addressed the limitations of QA systems, which struggle to keep up with the rapid increase of posted questions and lack effective ways to identify interesting questions. This gap between posted questions and potential answers can degrade a QA service's performance and reduce users' loyalty to the system. To address this, Zhou et al. (2012) proposed a question routing approach. They defined question routing as a classification task and developed various features to capture different aspects of questions, users, and their relations. They crawled datasets from popular Community Que stion Answering (CQA) services, such as Yahoo! Answers, in 2009 and used support vector machines (SVM) (Hearst et al., 1998), one of the widely used classification algorithms at that time. Their experiments showed that question classification can improve QA systems by providing directions for user knowledge modeling and personalized question routing.

Despite the success of traditional algorithms in question classification, pre-trained algorithms such as BERT, which employs a deep learning architecture and pre-training on a large corpus of text data, have outperformed them. Cortes et al. (2020) evaluated and compared the performance of several machine learning and deep learning models for question classification (QC) in question answering (QA) systems. They aimed to assert the importance of accurate question classification in improving the overall performance of natural language processing (NLP) systems. The authors evaluated several traditional machine learning models, such as Support Vector Machines (SVM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Bidirectional Encoder Representation from Transformer (BERT), etc. For their experiment, they used the UIUC (Li & Roth, 2002)and DISEQuA (Magnini et al., 2004) datasets. They analyzed the accuracy of each model by different levels of dependency, distinct languages, and different dataset sizes of the training set. From their analysis of various test performances, they mention the fact that methods relying on pre-trained language models, such as BERT, reach the best performance among various models. This is true for both the performance of the methods in distinct languages in both of their datasets and performance in different sizes of the training set. The advantage of pre-trained models is that they can learn relevant features in each layer of the model, including grammar and entity type of each word, without using external tools to represent them.

Finally, research on question classification for place-related questions was proposed by Yang et al. (2023). In their study, topic modeling was utilized to analyze geographic questions by examining their semantic similarity, using the MS MARCO dataset. They successfully clustered questions by their latent topic, but their unsupervised learning method may not accurately classify questions into main topic by cluster.

The studies mentioned above shed light on the importance of accurate question classification in enhancing natural language processing systems. However, previous works on question classification hav e certain limitations that need to be addressed. These limitations primarily arise from the use of traditional algorithms, which have been outperformed by pre-trained algorithms like BERT. Pre-trained algorithms, such as BERT, leverage deep learning architectures and pre-training on large text corpora to achieve superior performance in question classification.

Cortes et al. (2020) found that methods relying on pre-trained language models, such as BERT, consistently outperformed traditional machine learning models like Support Vector Machines (SVM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). The advantage of pre-trained models lies in their ability to learn relevant features at different layers, including grammar and entity types, without the need for external tools.

While these studies have made significant contributions to question classification in general, there is a need to specifically address question classification for place-related questions. One recent study by Yang et al. (2023) utilized topic modeling to analyze geographic questions using the MS MARCO dataset. They successfully clustered questions based on their latent topics. However, their unsupervised learning approach may not accurately classify questions into the ma in topic based on the clusters alone.

Considering these limitations, an effort is made in this thesis to bridge the gaps in previous works on question classification by utilizing methodologies from broader NLP fields and employing them to classify place-related questions into specific topics inferred from each question. By leveraging the advancements in pre-trained algorithms and addressing the unique challenges posed by place-related questions, the aim is to enhance the accuracy and effectiveness of question classification in the field of GeoQA.

1.3 Research Scope and Method

The objective of this thesis is to create a question classification (QC) module for the GeoQA system that enables accurate categorization of place-related questions based on their main topics. To achieve this, a supervised learning approach is utilized, where a BERT fine-tuned model is trained on a GeoQA dataset to acquire knowledge of the underlying patterns and connections between the questions and predefined topics. The primary goal of the QC module is to improve the effectiveness of GeoQA systems, enabling them to deliver more precise and pertinent answers.

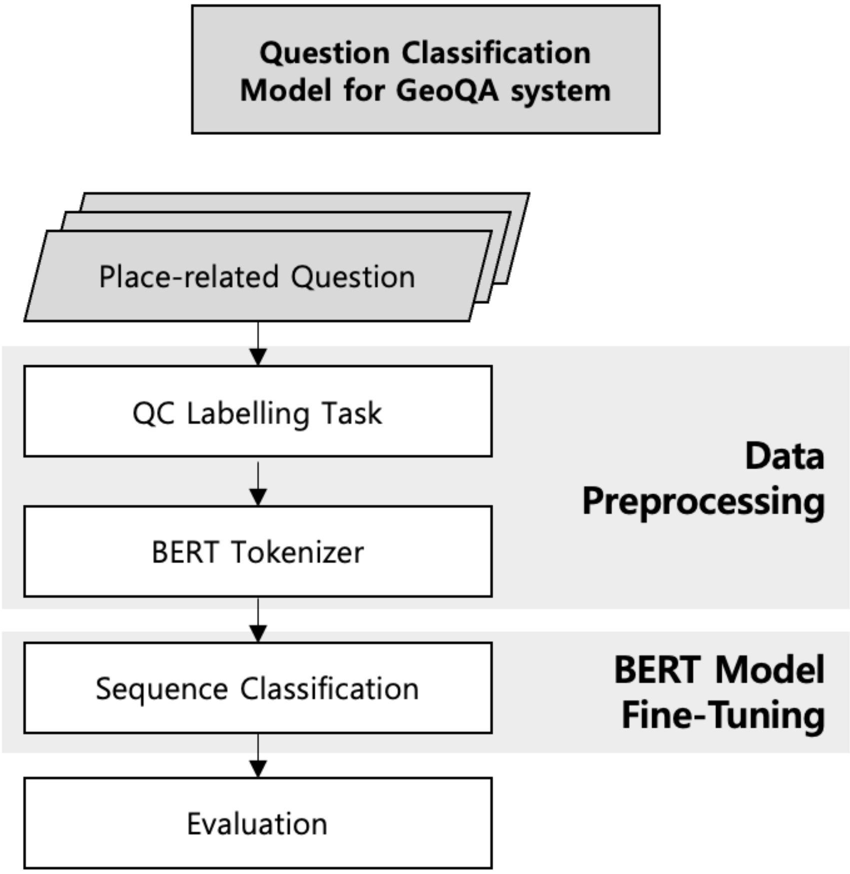
In this research, the focus is on addressing the multi-class classification problem in the GeoQA field. Each place-related question is assigned to a predefined topic that represents its main topic, based on a comprehensive analysis of frequently asked and prominent topics in such questions. The proposed QC module is developed and applied to a GeoQA dataset, enabling the accurate categorization of place-related questions according to their main topics.

The performance of the QC module is evaluated using various metrics, including accuracy, precision, and F1-score. The evaluation provides insights into the effectiveness of the proposed approach in classifying place−related questions. While the direct impact on the accuracy and relevance of answers provided by GeoQA systems is not examined in this research, the accurate categorization of questions based on their topics can lay the foundation for future improvements in the overall performance of GeoQA systems.

By utilizing a supervised learning approach and training a BERT fine−tuned model on a GeoQA dataset, the QC module aims to enhance the classification accuracy and enable a better understanding of the main topics of place−related questions. The results and insights gained from this research contribute to the ongoing efforts in developing intelligent GeoQA applications and advancing the field of question classification in the context of geographic queries.

2. Methodology

This chapter provides a detailed explanation of the multi−class question classification task for place−related questions, considering the relevant topics. To facilitate understanding, Figure 2−1 illustrates the flow of the methodology.



[Figure 2−1]Overview of the Question Classification (QC) Process by BERT Fine−Tuning

2.1 Dataset

2.1.1 Dataset Selection

For this experiment, the MS MARCO dataset (Bajaj et al., 2018) was selected for the question classification task of the GeoQA system. Among the three main QA corpora (MS MARCO, GeoQuestions201, and GeoAnQU) often used in GeoQA studies, the MS MARCO dataset is the only one that contains actual search questions that users have searched on Bing. Hence, it may be more representative of the information needs that users would typically ask in QA systems (Yang et al., 2023).

The MS MARCO (Microsoft MAchine Reading COmprehension) dataset was created by Bajaj et al. in 2018 as part of a large-scale effort by Microsoft to improve machine reading comprehension. The dataset contains a large number of question-and-answer pair from various sources, such as web pages and online forums, along with associated questions that are meant to test the ability of machine learning models to answer questions based on the content of the passages. The passages were selected to cover a wide range of topics and styles of writing, with the goal of creating a diverse dataset that wo uld be representative of real-world reading comprehension challenges.

This dataset is well-suited for training and evaluating machine learning models for question-answering tasks. It has been widely used in the research community and has become a benchmark for evaluating the performance of various machine learning models on reading comprehension tasks.

MS MARCO dataset is selected in this thesis for its size and diversity, which makes it a good candidate for training and evaluating machine learning models for QC tasks. The MS MARCO dataset contains over 1 million search queries with human-generated answers and is not specifically constructed for GeoQA purposes, meaning it includes all types of questions. This dataset is automatically annotated using machine learned classifier with five labels: (i) Numeric, (ii) Entity, (iii) Location, (iv) Person (v) Description (phrase). To ensure a representative sample for training and evaluation purposes, a carefully selected subset of 3,025 questions was chosen from the MS MARCO dataset. This subset was specifically derived from the "LOC" annotated questions, which are relevant to the topic of location in the context of GeoQA. The number of 3,025 questions was determine d to strike a balance between obtaining a sufficiently large dataset for robust model training and the practical constraints of manually labeling each question for topic classification. This sample size allows for comprehensive analysis and evaluation of the proposed question classification (QC) module while considering the available resources and time constraints.

2.1.2 Question Labelling

In this experiment, the topics for question classification were chosen based on the keywords and the type of information being asked in a sentence. However, it should be noted that the full version of the 'LOC(Location)' assigned questions in the MS MARCO dataset includes questions that are not necessarily related to geography. Therefore, Non-Geographic labels were assigned to sentences that ask for information about body parts or a place in a fictional world (e.g., "Where is Kidney located?") ([Table 2-1]). The remaining place-related questions are classified into more specific topics that reflect the nature of the questions.

[Table 2-1] Examples of Place relatedness for "LOC" annotated Questions from the MS MARCO Dataset

"LOC" annotated Questions from MS MARCO Dataset	**Place-relatedness**
where is calcium most commonly found in cells	X
where do i type the apple id verification code	X
where is scio ohio	O
where was the choice filmed	O

The extraction of the label, which represents the topic of each q uestion, was based on careful analysis and consideration of the content. The goal was to assign an accurate and informative topic that captures the essence of the question. To achieve this, the keywords and contextual information within the sentences were thoroughly examined. By identifying the key themes and subject matter of the questions, appropriate topics were assigned to provide a clear indication of the topic being addressed. This process ensured that the topics accurately reflected the underlying subject of each question, enabling effective classification and subsequent analysis.

After carefully reviewing previous studies in both open-domain QA and geographic-domain QA, a set of fine-grained categories was established to serve as the topics for accurately classifying place-related questions. The taxonomy used in open-domain QA, proposed by Li and Roth (2002), provided a foundation for capturing a broad range of topics related to various domains ([Table 2-2]). This taxonomy encompassed categories such as abbreviation, entity, animal, body, color, currency, and many others. Yet, there is no motivations of how specific categories were constructed (Sundblad, 2007).

[Table 2-2] Taxonomy defined in Li and Roth (2002) with 6-coarse categories and a total of 50 finer-categories

Class	Definition	Class	Definition
ABBREVIATIO	abbreviation	HUMAN	human beings

Class	Definition	Class	Definition
N			
abb	abbreviation	group	a group or organizati on of persons
exp	expression abbreviated	ind	an individual
ENTITY	entities	title	title of a person
animal	animals	descriptio n	description of a pers on
body	organs of body	LOCATIO N	locations
color	colors	city	cities
creative	inventions, books and oth er creative pieces	country	countries
currency	currency names	mountain	mountains
dis.med.	diseases and medicine	other	other locations
event	events	state	states
food	food	NUMERIC	numeric values
instrument	musical instrument	code	postcodes or other c odes
lang	languages	country	number of sth.
letter	letters like a−z	date	dates
other	other entities	distance	linear measures
plant	plants	money	prices
product	products	order	ranks
religion	religions	other	other numbers
sport	sports	period	the lasting time of st h.
substance	elements and substances	percent	fractions
symbol	symbols and signs	speed	speed
technique	techniques and methods	temp	temperature
term	equivalent terms	size	size, area and volum e
vehicle	vehicles	weight	weight
word	words with a special property		
DESCRIPTION	description and abstract c on−cepts		
definition	definition of sth.		

Class	Definition	Class	Definition
description	description of sth.		
manner	manner of an action		
reason	reasons		

In the context of geographic QA, Hamzei et al. (2020) introduced a schema that focused specifically on place-related questions ([Table 1-3]). Their semantic classifications included WH-word, place name, place type, object, quality, activity, situation and event, and spatial relationship. Building upon this foundation, Xu et al. (2020) further extended the semantic encoding by incorporating additional elements such as yes/no question verb, toponym, date, entity, place quality, entity quality, and others ([Table 2-3]).

[Table 2-3] Extended Semantic Encoding used in Xu et al. (2020)

Semantic Type	Part-of-speech	Code	Semantic Type	Part-of-speech	Code
where	WH-word	1	toponym	noun	n
what	WH-word	2	place type	noun	t
which	WH-word	3	date	noun	d
when	WH-word	4	entity	noun	e
how	WH-word	5	place quality	adjective	q
how+adj	WH-word	6	entity quality	adjective	p
why	WH-word	7	situation and event	verb	s

yes/no questions	verb	8	activity	verb	a
			spatial relationship	preposition	r

In order to capture the diverse range of topics in place−related question classification, a comprehensive set of 42 fine−grained categories was identified for this study based on the analysis of each question in the MS MARCO dataset. The selection of these categories was driven by the dataset's content, which predominantly encompasses both daily life−related words such as "address," "capital," and "terminal" (Xu et al., 2020), as well as specific geographic elements such as GPEs (geographic political entities) and habitats. By including these broader categories like GPEs and habitats alongside real−life inquiries, the topics used for classifying place−related questions cover a wide spectrum of user queries and ensure relevance to geographic queries. This comprehensive approach enables the effective categorization of diverse place−related questions, ultimately enhancing the performance and functionality of the question−answering (QA) system within the geographic domain.

[Table 2−4] Descriptions of Fine grained Topics used to Classify Place−related Questions

Topic(s)	**Description**
Abbreviation	Questions that ask for abbreviated state name or other abbreviation

Topic(s)	Description
Address	Questions that asks for 'address' of the place
Airlines	Questions that asks for airline destinations (ex. Where Does 'Airline Name' fly to?)
Airport	Questions that asks for location of the airport (ex. "Where is the nearest International Airport to NYC?")
Area Size	Questions that asks for the size of a place
Art/Music	Questions that are art/music related (ex. "Where is Mona Lisa?")
Association	Questions that asks for the location of the association
Building	Questions that asks for the location of certain building. (i.e., Rent, Mansion, Estate Complex, Medical Center, Store, Fort, Jail, Art Center, House, Coin−Machine, School, Resort/Hotel, Company, Hospital, Church, Meeting Location, Casino, Theatre, Notary, Bank, Restaurant, Stadium, etc.)
Climate	Questions that asks for weather/climate of a certain place (i.e., Sunniest City, Coldest Place, Rain the most, Temperate, Cold weather towns, Wettest Spot, etc)
Cost	Questions that are cost related (i.e., Expensive City to live in, Expensive Aiport, etc)
Crime	Questions that asks for crime rate of a certain place
Economic Geography	Questions that asks for Rust Belt, Fastest Growing Area etc.
Elevation	Questions that asks for elevation/highest point of a certain place
Entertainment	Questions that asks for a place of entertainment (i.e., Zoo, Fun Center, Campgrounds, Cove, Disney, Swimming Pool, Snorkeling, etc.)
Event	Questions that asks for certain event (i.e., Submarine Sink, Olympic Flame Lit, Marriage, Suicide, Super Bowls, Parade, Race, Festival, Terrorist Attack, Tournament, Fair, Hurriane, Tornadoes, etc.)
Fashion	Questions that fashion related (ex. "Which city is called the mecca of fashion?")
Film Location	Questions asking for place where certain movie/tv show was filmed

Topic(s)	Description
Geographic Feature	Questions that are related to geographic feature (i.e., Beach, Lake, Inside Passage, Glacier, Falls, River, Sea, Forest, Landscape, Vallet, Mammoth Cove, Volcano, Desert, Bull arm, Cliff, Desertification, Cliffs, Pipeline, Resorvior, etc.)
Geology	Questions that asks for the location of certain geology (i.e., Natural Bridge, Rock, Sinkhole, Minerals, Uplift, LimeStone, etc)
GPE	Questions that are related to Geo Political Entity (i.e., Country, State, County, District, City, Town, Island) (ex. "Largest City/Continent in the world")
Habitat	Questions that asks for where certain living things (i.e., animal/plant/fish) live, including biodiversity
History	Questions related to historical places (ex. Acient Greece live, Concentration Camp, Extermination Camp, War, Prince Henry Historical Movement, Colonization, Region Ruled by, Mayan Culture, Colony,)
Infrastructure	Questions that asks for the location of certain infrastructure. (i.e., Subway Station, Street, Station, Parking Area, Port, Dock, Dam, Road, Canal, Trail(Hiking), Park, Natural Reserve, Memorial Grove, etc)
Landmark(Tourist Attractions)	Questions that asks for certain landmark (ex. Golden Gate Bridge) or asks for recommendation of the most romantic place, popular city, most photographed places
Language	Questions that asks for where certain language is spoken (ex. "Where is Russian Spoken in the world?")
Locators	Questions that asks for locators (i.e., Country Code, Area Code, Zip Code, Airport Code, Telephone Area Code, Time Zone, Subduction Zone, Hardiness Zone, Convergence Zone, Growing Zone, Earthquake Zone, Other Zone, Hemisphere, Prime Meridian, Long-Latitude, Postcode, Horizon, etc.)
Life Expectancy	Questions that asks for the statistics of the life expectancy of the certain location
Military Field	Questions that asks for the location of the military field (i.e., afb, corps, fort sully, Camp (military Camp), Air Force/Navy Base etc.)

Topic(s)	Description
Non-geographic	Questions that are non-geographic questions (ex. "Where is the kidney located?")
Origination/Productio n	Questions that asks for where certain product was produced or a sks for the location of origination (i.e., Made, Originate From, Fou nd, Made, Quakers Start, etc.)
Place of Birth/Raised	Questions asking for where well-known figure was born or raise d.
Place of Death	Questions asking for where well-known figure died
Political/Legal Fact	Questions that asks for political or legal fact (i.e., Democracy, Vo te, Marijuana Legalized, Prostitution Legal, States Performing Fa milial Searching, First Independent Nation, Repressive Regimes, Banned, etc.)
Population	Questions asking for population of a certain place
Residence	Questions asking for the settlement or residence of well-known f igure
Retirement	Questions asking for places to retire (ex. "Where is the best plac e to retire in?")
Sport Team	Questions asking for where certain sport team is located (ex. Te xas A&M aggies located, What state do the kansas city royals pla y baseball, san antonio spurs play? etc.)
State Flag	Questions asking for the states that has certain state flag
State Motto	Questions asking for the state motto of a certain place (ex. "Whic h state, north to the future?")
Tax	Questions that are tax-related of a place (i.e., Form 1095, State Tax, Income Tax, Tax Table, Tax Interest, etc.)
Travel	Questions that are travel related (ex. "Best stay to stay in amster dam?")
Virus	Questions asking for certain virus (ex. "Regions where Hepatitis i s Common?", "Where was polio endemic")

To maintain focus on the specific geographic aspects, questions that primarily sought the location of geographic entities (GPEs) suc h as countries, counties, cities, and districts were not further catego rized. This decision was made considering the extensive analysis and classification of topics used to classify place−related questions in previous studies (see [Table 2−2] and [Table 2−3]). By employing this extensive range of detailed categories, the classification of place−related questions can be performed with higher levels of precision and granularity, leading to improved accuracy and customization of the question−answering (QA) system's responses to user queries within the geographic domain.

2.2 BERT for Question Classification

The QC module is designed to classify place−related questions into predefined topics that represent the main topic of the questions. After labeling the place−related questions with the corresponding main topic, the question classification process is performed.

2.2.1 Multi−Class Question Classification

In natural language processing (NLP), there are two types of question classification tasks. Multi−label question classification involves assigning multiple labels to a single question. In other words, each question can have multiple relevant classifications, and the goal is to predict all of them for a given question. For instance, consider the question "Where was Marilyn Monroe born?" The relevant labels could be "Celebrity", "Famous Figure", "Place of Birth", and so on. On the other hand, multi−class question classification involves assigning only one label to each question. The objective is to predict the single relevant classification for a given question. For example, given the question "What is the longest river in Africa?" the relevant label would be "river". In summary, the key distinction between multi−label question classification and multi−class question classification is the number of labels assigned to each question. Multi−label classification involves multiple labels per question, whereas multi−class classification involves only one label per question. The primary focus of this thesis is on multi−class QC, as the dataset used primarily featured one significant label for each question.

The proposed multi−class QC systems can be modelled with a set of questions asked on the QA systems and N number of pre−defined labels (l) used to classify each question [Equation (2−1)].

$$Q \in \{l_1, l_2, l_3, \dots\dots\dots., l_N\} \tag{2-1}$$

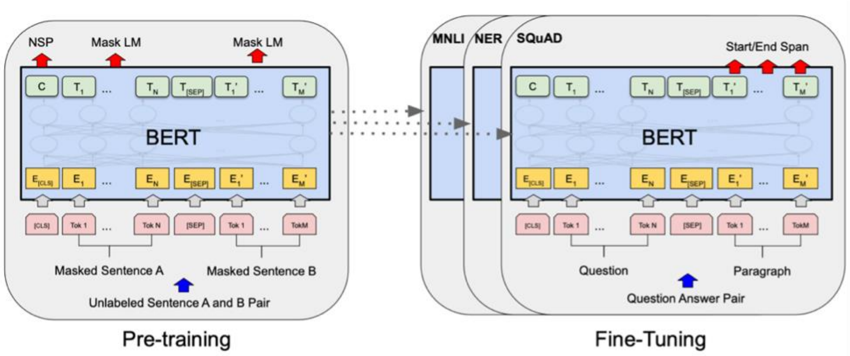
2.2.2 Model Selection

For the question classification task in this study, the primary model selected was BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). BERT has demonstrated strong performance on various natural language processing (NLP) tasks and achieved state-of-the-art results on benchmark datasets. The decision to choose BERT was based on several factors.

First, BERT is a pre-trained model that has learned contextualized word representations from a large amount of text data. Fine-tuning BERT has been shown to be effective for adapting the model to specific tasks, especially when working with smaller datasets like the randomly selected 3,025 questions from the MS MARCO dataset used in this study. Cortes et al. (2020) have highlighted the superiority of fine-tuning models on smaller datasets compared to other question classification methods for question answering systems.

Second, BERT is a transformer-based model, which allows it to capture long-range dependencies and contextual information effectively. This capability is crucial for the question classification task, where the meaning of a question can depend on its surrounding context. Moreover, unlike the traditional text classification models, BERT is an end−to−end model that does not separate feature extraction and classification task for text classification.

The BERT−base model consists of an encoder comprising 12 Transformer. Each transformer consists of 12 self−attention heads, and a hidden size of 768. When provided with a sequence input, BERT operates by producing a representation of the sequence. The sequence typically consists of one or two segments, with the first token always being [CLS], which contains a special classification embedding. Additionally, another special token, [SEP], is utilized to separate segments within the sequence.



[Figure 2−2] Overall pre training and fine tuning procedures for BERT (Devlin et al., 2019). [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token

For text classification tasks, BERT utilizes the final hidden state h of the first token, [CLS], as the representation for the entire sequ ence. A SoftMax classifier is incorporated at the apex of BERT to prognosticate the likelihood of a particular label c. The task-specific parameter matrix W is utilized, and in the process of fine-tuning, all the parameters from BERT, including W, are jointly optimized to maximize the log-probability of the correct label.

$$p(\hbar) = softmax(W\hbar) \qquad (2\text{-}2)$$

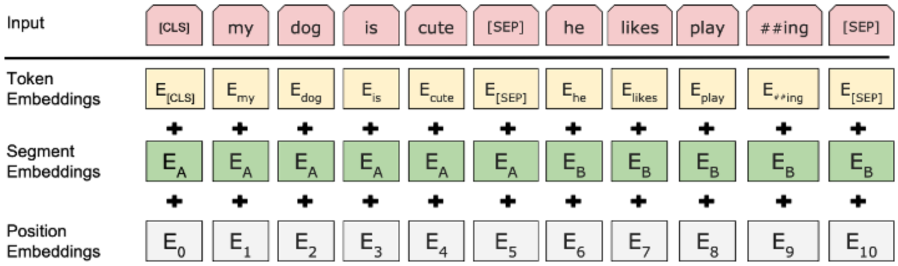
To facilitate the fine-tuning of the pre-trained BERT model, the Hugging Face Transformers library was employed, which provides an accessible interface for fine-tuning pre-trained models on various natural language processing (NLP) tasks. This library streamlined the implementation of the fine-tuning process and allowed for flexibility in customizing the training procedure.

The choice of the "bert-base-uncased" checkpoint as the starting point for fine-tuning was based on several key characteristics that made it highly suitable for the specific task at hand. Firstly, the pre-trained BERT model, upon which the "bert-base-uncased" checkpoint is based, has undergone extensive training on a large corpus of text data. This training process enables the model to capture and e ncode rich contextual information from diverse language patterns and structures. Furthermore, the "bert-base-uncased" checkpoint is composed of 12 transformer layers, each consisting of 768 hidden units, and incorporates 12 attention heads. These architectural specifications contribute to the model's ability to effectively learn and represent complex relationships and dependencies within the input data. The multiple transformer layers facilitate the extraction of hierarchical features and capture long-range dependencies, while the attention heads allow the model to focus on different parts of the input sequence simultaneously. The combination of these characteristics, including the pre-training on a large text corpus, the multi-layered transformer architecture, and the attention mechanism, makes the "bert-base-uncased" checkpoint well-suited for fine-tuning on the specific task of question classification in the GeoQA system. By leveraging these inherent strengths, the fine-tuned BERT model can effectively learn and adapt to the nuances and patterns present in the place-related questions, ultimately enhancing the accuracy and performance of the question classification module.

To fine-tune the BERT model suited for the QC module, the AdamW optimizer, a variant of the Adam optimizer that incorporates w eight decay, was employed. The hyperparameters of the optimizer were set as follows: a learning rate of 2e−5, epsilon of 1e−8, and weight decay of 0.01.

The fine−tuning process involved feeding the input data into the pre−trained BERT model and training it to predict the target variable, which pertained to question classification in this study. To prepare the input data for the model, the BERT tokenizer was utilized, which segmented the text into sub−words aligning with the BERT vocabulary. This tokenization process facilitated the model's ability to handle out−of−vocabulary words or those necessitating sub−word representation.

Subsequent to tokenization, the text was transformed into input features, encompassing input IDs, attention masks, and token type IDs. The input IDs represented the sequence of token indices, attention masks denoted the tokens requiring attention, and token type IDs distinguished between different segments within the input sequence. These features played a crucial role in ensuring the proper functioning of the BERT model.



[Figure 2-3] BERT input representation figure from Devlin et al. (2019). The input embeddings are the sum of the token embeddings

Once the input features were prepared, they were fed into the BERT model, and the model was trained to predict the target variable utilizing the cross-entropy loss function. During training, a batch size of 32 was employed, and the model underwent training for a total of 3 epochs. The selection of the number of epochs aimed to strike a balance between model performance and computational resources, and it was determined through an iterative experimentation process.

Throughout the training process, the performance of the model was monitored on a validation set, which constituted a randomly selected 10% portion of the training data. Early stopping based on the validation loss was implemented to prevent overfitting of the model. This approach facilitated the identification of the optimal model checkpoint exhibiting the most favorable performance on the validation se t.

In summary, the methodology encompassed the fine-tuning of a pre-trained BERT model employing established transformer fine-tuning procedures. The Hugging Face Transformers library and the PyTorch framework were utilized to facilitate the fine-tuning process. By adapting the model to the specific task through training on task-specific data, the aim was to leverage the contextual understanding of the pre-trained BERT model and achieve question classification task. The evaluation of the model's performance was conducted on a held-out test set, yielding insights into its effectiveness.

2.3 Evaluation Methodology

The evaluation of the fine−tuned BERT model's performance in question classification is a critical step to determine its capability in accurately categorizing questions. In this study, accuracy was selected as the primary evaluation metric for the model. Accuracy measures the proportion of questions that are correctly classified out of the total number of questions in the test set. It considers the True Positives (TP), which represent the number of correctly predicted positive instances (i.e., instances that are actually positive and are predicted as positive), as well as the True Negatives (TN), which denote the number of correctly predicted negative instances (i.e., instances that are actually negative and are predicted as negative). On the other hand, False Positives (FP) represent the number of negative instances incorrectly predicted as positive (i.e., instances that are actually negative but are predicted as positive), and False Negatives (FN) represent the number of positive instances incorrectly predicted as negative (i.e., instances that are actually positive but are predicted as negative).

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2-3)$$

The suitability of accuracy as a metric for the question classification task is demonstrated by Equation (2−3), as the objective is to maximize the correct classification of questions across all categories. The model's overall performance in terms of correctly classifying questions, considering both positive and negative instances, can be assessed by calculating the accuracy.

$$Precision = \frac{TP}{(TP + FP)} \tag{2-4}$$

Precision measures the accuracy of positive predictions made by the model. It quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive.

$$Recall = \frac{TP}{(TP + FN)} \tag{2-5}$$

Recall (also known as Sensitivity or True Positive Rate) measures the ability of the model to identify positive instances correctly. It represents the proportion of true positive instances identified by the model out of all actual positive instances.

The F1 Score, as shown in Equation (2−6), is a metric that combines precision and recall into a single value. It is the harmonic mea n of precision and recall, providing a balanced evaluation of the model's performance. The F1 Score considers both precision and recall, making it a suitable metric for tasks where false positives and false negatives have different implications. It balances the trade−off between precision and recall, offering a comprehensive evaluation of the model's performance.

$$F-1\,Score\;=\;2\cdot\frac{Precision\;\cdot Recall}{Precision+Recall} \tag{2-6}$$

The dataset was split into training, validation, and test sets using a stratified approach to ensure the maintenance of the distribution of categories in the original dataset across the splits. Eighty percent of the dataset was used for training, 10% for validation, and 10% for testing. The test set was kept separate and used only once to assess the final performance of the model. During training, the training loss, validation loss, f−1 score, and accuracy on the validation set were monitored to determine the stopping point of the training and to avoid overfitting. The model with the highest accuracy on the validation set was selected as the final model.

3. Experiment and Result

3.1 Experiment Environment

The experiment described in this thesis was conducted on a system with an Intel® Core™ i9-7900X CPU @ 3.30GHz and 64.0 GB memory, running the Windows 10 Education 64 Bit operating system. The experiment was performed in a controlled environment to ensure the consistency and accuracy of the results. To ensure the integrity of the results, the computer was optimized for performance by disabling unnecessary background processes and services. The system was also kept up to date with the latest updates and patches. The experiment was conducted using Python 3.8.16, Torch 1.9.0, and Transformers 4.28.1 to implement the full process of question classification for place-related questions.

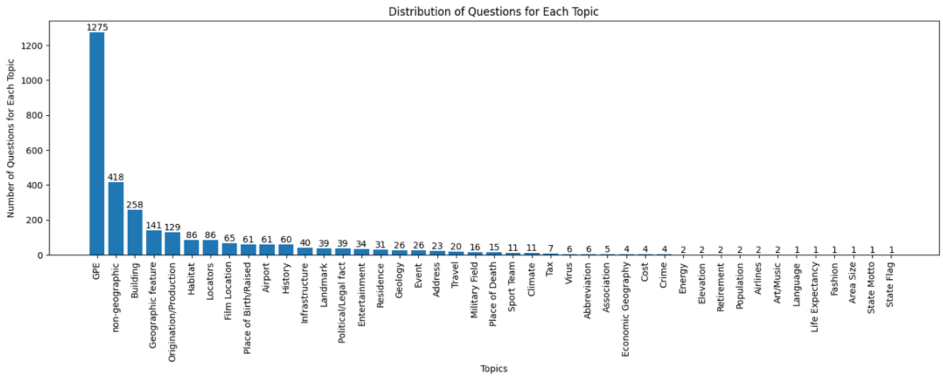
3.2 Question Labelling Result

Upon manually labeling randomly extracted place-related questions, it was evident that an unbalanced distribution existed among the topic labels. This observation can be attributed to the dominance of inquiries focused on the location of geopolitical entities (GPEs), including countries, states, counties, districts, cities, towns, and islands. In order to gain comprehensive insights into the various topics encompassing place-related questions, a meticulous labeling procedure was conducted to assign each question to one of the 42 fine-grained categories. This categorization allowed for a more nuanced analysis and understanding of the different topics. The distribution of the number of questions classified under each topic is presented in [Table 3-1] and [Figure 3-1].

[Table 3-1] Number of Questions Classified into Each Topic

Topic	Count	Topic	Count
GPE	1,275	Sport Team	11
non-geographic	418	Climate	11
Building	258	Tax	7
Geographic feature	141	Virus	6
Origination/Production	129	Abbreviation	6
Habitat	86	Association	5
Locators	86	Economic Geography	4
Film Location	65	Cost	4
Place of Birth/Raised	61	Crime	4

Topic	Count	Topic	Count
Airport	61	Energy	2
History	60	Elevation	2
Infrastructure	40	Retirement	2
Landmark (tourist attractions)	39	Population	2
Political/Legal fact	39	Airlines	2
Entertainment	34	Art/Music	2
Residence	31	Language	1
Geology	26	Life Expectancy	1
Event	26	Fashion	1
Address	23	Area Size	1
Travel	20	State Motto	1
Military Field	16	State Flag	1
Place of Death	15		



[Figure 3-1] Distribution of the Labelled Dataset

The distribution of question labels throughout 42 different topics was examined, revealing that questions frequently inquire about th e location of certain GPEs (geo−political entities), buildings, and geographic features. Additionally, queries in the MS MARCO dataset often pertain to the origin or production location of specific living things or products, as well as the habitats of various organisms. Furthermore, locators (such as country codes, area codes, airport codes, telephone area codes, time zones, etc.), film locations (where certain movies or shows were filmed), birthplaces of well−known figures (e.g., historical figures, celebrities), locations of airports, and sites of historical events are frequently sought by users.

To ensure an adequate number of training examples for each classification topic, a filtering process was implemented. Topics with less than 60 assigned questions were excluded from the analysis to ensure a reasonable distribution of data and to avoid potential biases caused by limited instances for certain topics. Although the specific threshold of 60 was chosen based on practical considerations, such as balancing dataset size and maintaining enough instances for reliable classification, it is worth noting that no statistically significant rationale or predefined criterion exists for this specific threshold selection. Furthermore, it was decided to proceed with the question classification, taking into account the practical considerations of dataset si ze and maintaining enough instances for reliable classification despite the presence of unbalanced data. By applying this filtering process, each remaining topic was assigned a unique numeric value, enabling further analysis and classification ([Table 3−2]).

[Table 3−2] Topics that have been assigned to more than 60 questions in each category

Topic	Topic No.	Question Count
GPE	0	1275
Non−geographic	2	418
Building	7	257
Geographic Feature	3	141
Origination/Production	6	129
Habitat	4	86
Locators	9	86
Film Location	5	65
Place of Birth/Raised	8	61
Airport	1	60

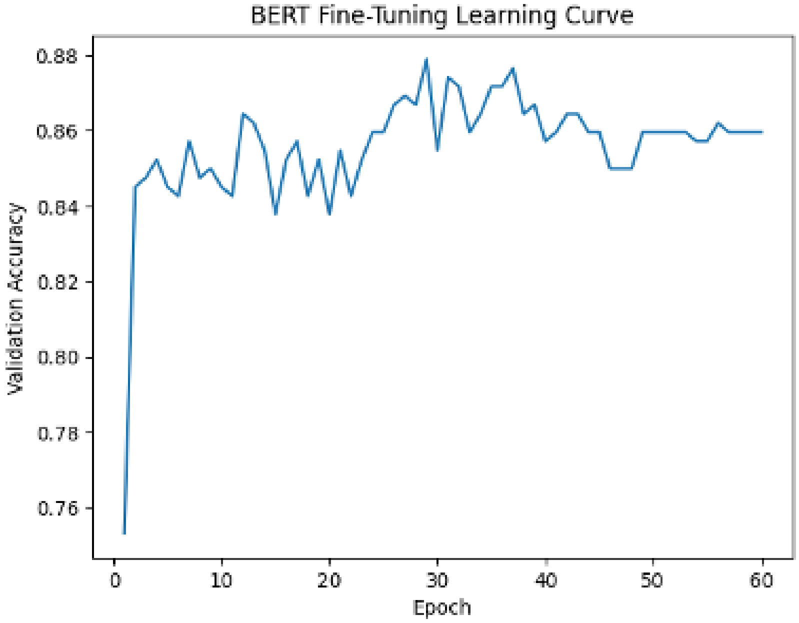
3.3 Evaluation

The proposed multi−class question classification model was evaluated on the MS MARCO dataset, which includes 10 topics (Airport, Building, Film Location, GPE, Geographic Feature, Habitat, Locators, Non−geographic, Origination/Production, Place of Birth/Raised) related to keywords of place−related questions that represent the main topic focus.

It is important to highlight that a recent study conducted by Yang et al. (2023) examined question classification for the MS MARCO place−related question dataset using an unsupervised learning approach with a T−5 model. In contrast, this study adopts a supervised learning approach by fine−tuning the BERT model.

This research makes a significant contribution to the field of natural language processing and showcases the ability of machine learning models to accurately classify questions according to their main topic, based on their content.

The learning curve was monitored during the validation step, spanning 60 epochs, and employing a learning rate of 1e−5. A validation accuracy of 87.9% was achieved by the proposed model on epoch 29, indicating the model's capacity to accurately classify questions based on their content and structure. The validation accuracy per epoch is illustrated in [Figure 3-2], revealing the model's commendable performance.

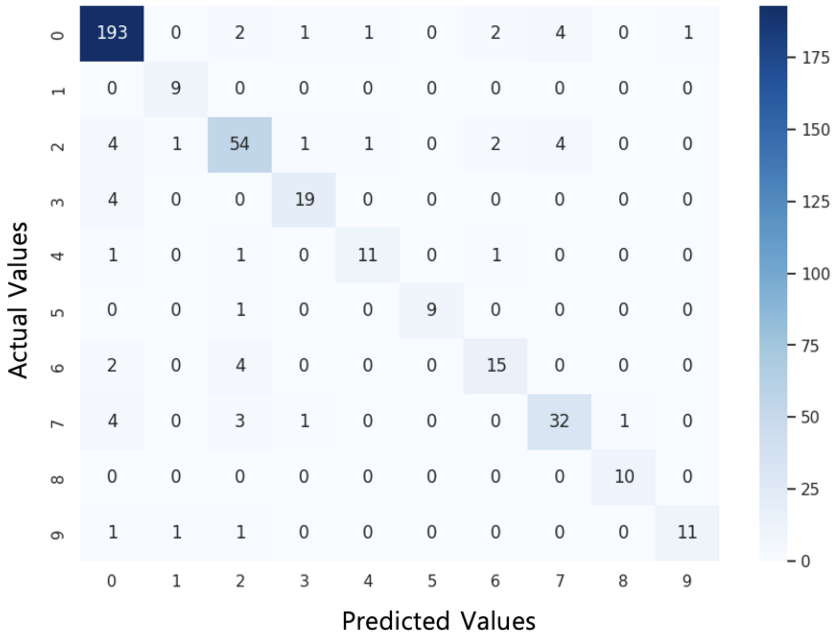


[Figure 3-2] Validation Accuracy over 60 epochs

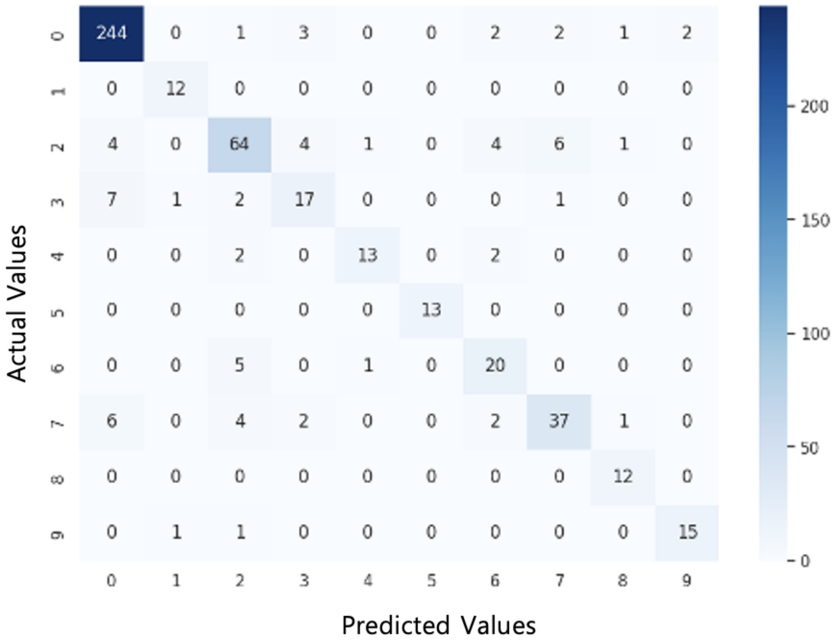
To assess the model's performance in greater detail, accuracy per class was checked for both the validation and test dataset ([Table 3-4]). Furthermore, confusion matrix was generated for both validation and test result ([Figure 3-3]). The confusion matrix provides insights into the model's ability to correctly classify each topic and identifies instances of confusion between similar topics. Overall, th e confusion matrix showcases the model's competence in accurately categorizing questions based on their content.

[Table 3-3] Accuracy per Topic on Validation and Test Dataset

Topic No.	Topic Label	Validation Dataset	Test Dataset
0	GPE	193/204 (94.61%.)	244/255 (95.69%)
1	Airport	9/9 (100%)	12/12 (100%)
2	Non-geographic	52/67 (77.61%)	64/84 (76.19%)
3	Geographic Feature	19/23 (82.61%)	17/28 (60.71%)
4	Habitat	11/14 (78.57%)	13/17 (76.47%)
5	Film Location	9/10 (90%)	13/13 (100%)
6	Origination/Production	15/21 (71.43%)	20/26 (76.92%)
7	Building	32/41 (78.05%)	37/52 (71.15%)
8	Place of Birth/Raised	10/10 (100%)	12/12 (100%)
9	Locators	11/14 (78.57%)	15/17 (88.24%)



(a) Confusion Matrix on Validation Dataset



(b) Confusion Matrix on Test Dataset

[Figure 3-3] Confusion Matrix on Both Validation and Test Dataset

Based on the observation of the confusion matrix, there were a significant number of topics that were misclassified as Topic No. 0 (GPE), Topic No. 2 (Non-Geographic), and Topic No. 7 (Building). This can be attributed to the limitations of the question classification dataset used in this experiment, where the quantities of examples for each topic were unevenly distributed. Uneven distribution of examples across topics can pose challenges for the model in accurately classifying questions. When certain topics have a relatively smaller number of examples compared to others, the model may struggle to learn and generalize patterns specific to those topics. Addressing the issue of uneven data distribution requires collecting or augmenting more examples for underrepresented topics. By ensuring a more balanced dataset, the model can be trained to better handle the challenges associated with misclassifications caused by imbalanced data distribution.

Overall, the question classification task for place-related questions achieved a high accuracy rate, with 87.9% on the validation set and 86.6% on the test set. However, the evaluation of the question classification results revealed instances of mislabeling, leading to inaccuracies in the predicted labels. These instances provide valuable insights into the factors contributing to mislabeling.

One prominent factor is the presence of multi-label situations, where a question encompasses multiple distinct topics. For example, in the text "what county is California doesn't run jails," the focus on searching for jails initially led to its label as "Building." However, the inclusion of the word "California" caused the model to misclassify it as "GPE." This example demonstrates the challenge of assigning a single label when multiple topics are involved, as the model struggles to differentiate between the emphasis on jails (Building) and the r eference to California (GPE).

Another factor is the difficulty in distinguishing between topics. In the text "what countries does the Mediterranean consist of," the original label assigned was "Geographic Feature," but it was misclassified as "GPE." This case highlights the model's struggle in accurately differentiating between the primary topic of interest, the "Mediterranean" as a geographic feature, and the mention of "countries" as geopolitical entities. The model's inability to discern between these concepts results in incorrect predictions.

Additionally, challenges arise in discerning geographic references from non-geographic ones. For instance, the text "greenwich japan limited" was initially labeled as "Non-geographic" but misclassified as "GPE." This example showcases the model's difficulty in distinguishing between geographic and non-geographic references, likely due to the presence of the location-related term "Japan."

Furthermore, the model's reliance on specific keywords presents another challenge to accurate classification. In the text "where can I get SAT past papers," the initial label assigned was "Non-geographic," but it was inaccurately predicted as "Building." This misclassification can be attributed to the model's association of the phrase "wh ere can I get papers" with physical structures, leading to an incorrect classification.

Examining these specific examples reveals that factors such as multi-label situations, difficulties in distinguishing between topics, challenges in discerning geographic references, and reliance on specific keywords have contributed to instances of mislabeling. Improving the model's sensitivity to context, enhancing differentiation between topics, developing a better understanding of geographic references, and reducing over-reliance on specific keywords are essential steps towards enhancing the accuracy and reliability of the question classification model for place-related questions.

4. Conclusion

In conclusion, this thesis has undertaken comprehensive research on the task of question classification for Geographic Question Answering (GeoQA) systems, with the primary objective of enhancing their performance. Previous research in the field of GeoQA mainly performed QC for the intention to analyze the place-related question datasets or to construct a query template for the IR process. Yet, this thesis introduces a novel approach by performing QC to classify pla ce-related questions based on their main topics, aiming to further enhance the GeoQA system by enabling accurate navigation for answers from external resources. Thus, a QC module specifically designed for place-related questions was constructed to bridge the existing gap in question classification for GeoQA systems.

To achieve this goal, a dataset consisting of 3,025 randomly selected place-related questions from the MS MARCO dataset with "LOC" annotations was manually labeled among 42 predefined topics. Subsequently, a BERT-based model was fine-tuned using this labeled dataset to develop a multi-class question classification module. The decision to employ multi-class classification instead of multi-label classification was based on the observation during the manual labeling procedure. It was found that the majority of the place-related questions had a single dominant topic, rather than multiple topics. In other words, most questions could be assigned to a primary category that represented their main focus. Consequently, it was determined that conducting multi-class classification would be appropriate for the question classification task. Multi-class classification involves assigning a single, exclusive label from a predefined set of categories to each instance. Given that most questions had a clear primary to pic, this approach was well-suited to capture the main focus of each question and provide a single, specific label.

By adopting multi-class classification, the model could effectively categorize each question into one of the predefined topics. This approach simplified the classification process and provided a straightforward categorization for each instance, reflecting the dominant topic of the respective questions. The utilization of BERT, a state-of-the-art language model, allowed for a more nuanced understanding of the context and semantics of place-related questions. The effectiveness of the model in accurately classifying place-related questions into relevant topic was demonstrated. The evaluation results revealed performance by achieving accuracy rates of 87.9% on the validation set and 86.6% on the test set. These outcomes highlight the model's capability to accurately classify place-related questions within the specific context of GeoQA systems.

The proposed multi-class question classification model makes a notable contribution to the field of natural language processing by offering a tailored approach specifically designed for the place-related questions acquired from the MS MARCO dataset. This model serves as a foundational component for future improvements in the over all GeoQA system. By accurately classifying place-related questions into relevant topics, the question classification (QC) module enables more precise and targeted retrieval of information. This advancement in question classification has the potential to enhance the overall performance and effectiveness of GeoQA systems, thereby improving the quality of answers provided to users in their place-related inquiries. By leveraging the predefined topic classification assigned to each question during the information retrieval (IR) process, the system can effectively navigate and retrieve relevant information, ensuring more accurate and meaningful responses.

While the developed model has demonstrated promising results, it is important to acknowledge the limitations of this research. Firstly, the generalizability of the model to other datasets or real-world applications may be limited, as it was primarily evaluated on the MS MARCO dataset. Future research should explore the performance of the model on diverse datasets and real-world scenarios to assess its generalizability and robustness.

Additionally, the predefined topic categories used for place-related questions in this study may not cover all possible topics and nuances, potentially limiting the model's ability to classify questions acc urately in certain situations. Future studies should consider expanding the set of topic categories or exploring alternative approaches, such as incorporating multi-label classification, to capture the complexity and variety of topics associated with place-related questions.

To further enhance the accuracy and reliability of question classification models for place-related questions and improve the performance of GeoQA systems, it is recommended that future research should focus on the accumulation of more datasets containing multi-label place-related questions. This would facilitate the exploration of multi-label classification approaches, enabling a more comprehensive analysis of the multiple topics associated with each question. By addressing these limitations and advancing the research in question classification for place-related questions, contributions can be made towards the development of more accurate and efficient GeoQA systems, wherein users are provided with precise and relevant answers to their place-related inquiries.

2. Appendix

Test Result			
Text	Predicted	Topic	Accuracy
what county is jefferson city mo located in	GPE	GPE	1
where does sugar bear live	Habitat	Habitat	1
where was sinner filmed	Film Location	Film Location	1
what county is monahans tx in	GPE	GPE	1
where is are the hamptons locate d	GPE	GPE	1
what county city of jacksonville, fl	GPE	GPE	1
where is osage county	GPE	GPE	1
where is great falls located	GPE	GPE	1
what county is whittier ca in?	GPE	GPE	1
where is the information about th e Valkyrie plot coming from	Non−geographic	Non−geographic	1
what county is california doesn't run jails	GPE	Building	0
where is charleston sc historic d istrict	GPE	GPE	1
where is darnu horn of oryx loca ted	Non−geographic	Non−geographic	1
what county is brownville maine in	GPE	GPE	1
what country is jerusalem	GPE	GPE	1
where is noxie, ok	GPE	GPE	1
what countries does the mediterr anean consist of	GPE	Geographic Feature	0
where is the nucleus	Non−geographic	Non−geographic	1
giant panda is situated in which c ountries	Habitat	Habitat	1
greenwich japan limited	GPE	Non−geographic	0
where can i get sat past papers	Building	Non−geographic	0
what building do the canadians pl ay in	Building	Building	1
what county dansville ny	GPE	GPE	1
what places can you send an ema il gift card from	Building	Non−geographic	0

Test Result			
Text	Predicted	Topic	Accuracy
where is bunker hill in relation to railroad hq.fallout	Non−geographic	Non−geographic	1
what county is mcdou	GPE	GPE	1
whats the country code for Ghana	Locators	Locators	1
what zone is iowa	Locators	Locators	1
what does it cost to fax at a fedex location	Non−geographic	Non−geographic	1
what airport do you use for banff, canada	Airport	Airport	1
where is weaver iowa	GPE	GPE	1
where is ashford university	Building	Building	1
where is the best place to stay if you want to see the amalfi Geographic Feature	Geographic Feature	Geographic Feature	1
where is cutaneous tissue	Non−geographic	Non−geographic	1
havasu falls, arizona, usa	GPE	Geographic Feature	0
rice where does it come from	Non−geographic	Origination/Production	0
what county is gladwin mi	GPE	GPE	1
where are the aurora teagarden movies filmed	Film Location	Film Location	1
where is el foster city	GPE	GPE	1
where is silsbee, tx	GPE	GPE	1
where is lexington square lowes foods	Building	Building	1
where is al pacino from	Place of Birth/Raised	Place of Birth/Raised	1
what county is kecksburg pa in	GPE	GPE	1
where is the pork loin end roast located on the hog	Non−geographic	Non−geographic	1
where are warn winches made	Origination/Production	Origination/Production	1
thorp wisconsin is in what county	GPE	GPE	1
where is dill found	Non−geographic	Habitat	0
where is vicarstown cork	GPE	GPE	1
where on the earth is the most amount of biodiversity	Non−geographic	Habitat	0

Test Result			
Text	Predicted	Topic	Accuracy
where are equilibrium receptors located quizlet	Non−geographic	Non−geographic	1
which dialing area is 0170	Locators	Locators	1
where do stones come from	Non−geographic	Origination/Production	0
where is williamston county sc?	GPE	GPE	1
in what county is chanhassen, mn?	GPE	GPE	1
what time zone is michigan in.	Locators	Locators	1
where does fertilization usually occur quizlet	Non−geographic	Non−geographic	1
what county is alpine ny in	GPE	GPE	1
where is ione washington	GPE	GPE	1
where is SALMONELLA TYPHI usually found	Non−geographic	Non−geographic	1
where are gogo squeeze apple made	Origination/Production	Origination/Production	1
what county is litchfield arizona?	GPE	GPE	1
where does honky come from?	Origination/Production	Non−geographic	0
what county is gordonville, texas in?	GPE	GPE	1
where is hartford wisconsin	GPE	GPE	1
what is the biggest state in the world?	GPE	GPE	1
where is curry alabama	GPE	GPE	1
where do peacocks live	Habitat	Habitat	1
what area code is 02 6393	Locators	Locators	1
what county is the city of dallas in	GPE	GPE	1
what county is mt. holly springs a	GPE	GPE	1
palace Resort/Hotels jamaica	Building	Building	1
where is clinton ms.	GPE	GPE	1
where can i finance a single wide mobile home	Building	Non−geographic	0
vacation rentals in maui	GPE	Building	0
where is helen ga	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where is cameroon located	GPE	GPE	1
what county is ness city in	GPE	GPE	1
where is the pound key	Non−geographic	Non−geographic	1
town of miller place	GPE	GPE	1
what county is wilmington nc	GPE	GPE	1
what county is savannah,ga in	GPE	GPE	1
where is hartford city indiana loc ated	GPE	GPE	1
what county is falls church va in	GPE	GPE	1
what county is murfreesboro ar i n	GPE	GPE	1
what county is hemet ca	GPE	GPE	1
what county is elizabeth iowa in	GPE	GPE	1
where is carlisle ar?	GPE	GPE	1
what county is dundee michigan i n	GPE	GPE	1
where was gilbert gil raised	Place of Birth/Raise d	Place of Birth/Raise d	1
what county is mount vernon mo	GPE	Geographic Feature	0
what state is the granite state	GPE	GPE	1
where is folded mountain brewer y	Building	Building	1
what county is richton, ms	GPE	GPE	1
what is the airport code for corn wall	Locators	Locators	1
what township is la grange il	GPE	GPE	1
what is the closest airport to he met ca	Airport	Airport	1
hotels in tybee island beach	Geographic Feature	Building	0
where is creede colorado located	GPE	GPE	1
what county is sandy hook ky in	GPE	GPE	1
what county is franklin massachu setts	GPE	GPE	1
what county am i in addison, tx	GPE	GPE	1
where does the drain tube attach to a dishwasher	Non−geographic	Non−geographic	1

Test Result			
Text	Predicted	Topic	Accuracy
where does most of the weather occur	Non−geographic	Non−geographic	1
ct600 where do I put business loans	Non−geographic	Non−geographic	1
what country is barbados in	GPE	GPE	1
where are the kenyon heights	GPE	Building	0
where are cinnamon buns from	Origination/Production	Origination/Production	1
where is erie pennsylvania	GPE	GPE	1
where is kiehl's products in new york?	Building	Building	1
where are stihl products manufactured	Origination/Production	Origination/Production	1
what's the largest city in the world?	GPE	GPE	1
where is mt elbert in colorado	Geographic Feature	Geographic Feature	1
what county is amsterdam ny in	GPE	GPE	1
what county is aspermont, tx	GPE	GPE	1
colorado theater massacre how did it happen	Non−geographic	Building	0
where is fenway park in boston	Geographic Feature	Building	0
where is ravenna nebraska	GPE	GPE	1
where does cork come from	Non−geographic	Non−geographic	1
what county is royal oak, mi	GPE	GPE	1
what city is callahan county in	GPE	GPE	1
where was chuck noland from	Place of Birth/Raised	Non−geographic	0
where is fermaca	Building	Building	1
where would be best place to fish for muskellunge	Habitat	Habitat	1
where was the abyss filmed	Film Location	Film Location	1
where is bernie sanders from	Place of Birth/Raised	Place of Birth/Raised	1
where do the detroit pistons play home games	Non−geographic	Building	0
which is small the smallest house in the world?	Building	Building	1
what county is kalispell mt in?	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where was i am malala born	Place of Birth/Raised	Place of Birth/Raised	1
where is leadville mt	GPE	GPE	1
where are rapid log data files located	Non-geographic	Non-geographic	1
what sass serves cgh hospital	Building	Non-geographic	0
where does the name tallant originate	Origination/Production	Non-geographic	0
where is dursley	GPE	GPE	1
where is the medallion in archons forge?	Non-geographic	Non-geographic	1
what county is conway arkansas in	GPE	GPE	1
where in the Bible does it talk about things done in secret	Non-geographic	Non-geographic	1
what county is dayton nevada	GPE	GPE	1
where should a tie fall	Non-geographic	Non-geographic	1
where is castle rock mn	GPE	GPE	1
where is the sabaru made	Origination/Production	Origination/Production	1
where is willard bay located	GPE	Geographic Feature	0
what time zone is coralville, ia	Locators	Locators	1
where to catch sf ferry	Building	Building	1
where is bull arm located in nl	Non-geographic	Geographic Feature	0
digital antenna location	Non-geographic	Non-geographic	1
where is bardstown	GPE	GPE	1
where is boonton nj	GPE	GPE	1
hotel rentals in maine	Building	Building	1
what county is palatine il in	GPE	GPE	1
what county is brightwood oregon	GPE	GPE	1
where is tecumseh ne	GPE	GPE	1
where is mittagong rsl	Non-geographic	Building	0
what county is crawfordville fl	GPE	GPE	1
where is culiacan mexico located	GPE	GPE	1
where is peru located	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
what county is pelham nh	GPE	GPE	1
where is omsk?	GPE	GPE	1
where is aiken	GPE	GPE	1
where is washington located in the america	GPE	GPE	1
location of the phalanges bones	Non-geographic	Non-geographic	1
where is scunthorpe on the map	GPE	GPE	1
where is the chez republic	GPE	GPE	1
where was the stemware marked with a z in a square made	Non-geographic	Non-geographic	1
most holy trinity parish, detroit	GPE	Building	0
where does the surname hatton come from	Non-geographic	Non-geographic	1
what county is greencastle pa	GPE	GPE	1
where is newport nc google earth	GPE	GPE	1
where is glucose stored	Non-geographic	Non-geographic	1
what county is fruitvale in	GPE	GPE	1
where are the deadmines in wow	Non-geographic	Non-geographic	1
where is the benton county detention center	Building	Building	1
what county is morrisonville wi in	GPE	GPE	1
where are great white found	Habitat	Habitat	1
where did grits originate from	Origination/Production	Origination/Production	1
where to fly into anaheim	Airport	Airport	1
where is chesterfield michigan	GPE	GPE	1
where is altuna in the united states	GPE	GPE	1
where does ozark leave off	Geographic Feature	Non-geographic	0
where is buhl?	GPE	GPE	1
which city is closer the thar desert- lahore	Geographic Feature	Geographic Feature	1
where is sullivan, wi	GPE	GPE	1
where can syphilis sores be located	Non-geographic	Non-geographic	1

Test Result			
Text	Predicted	Topic	Accuracy
where is atchison ks located	GPE	GPE	1
what is dfw airport	Airport	Airport	1
what region is us	GPE	GPE	1
what county is kilrush located	GPE	GPE	1
where are rainforests location	Geographic Feature	Geographic Feature	1
what is the nearest airport to pas adena	Airport	Airport	1
where was barcodes invented at	Non−geographic	Non−geographic	1
state of oregon hours required to maintain rn license	Non−geographic	Non−geographic	1
where do you see haploid and di ploid cells	Non−geographic	Non−geographic	1
what county is ocala	GPE	GPE	1
where is the scottish borders	GPE	GPE	1
where is longhorn steakhouse	Origination/Product ion	Building	0
where are traxxas products mad e	Origination/Product ion	Origination/Product ion	1
washingtonville what is the count y	GPE	GPE	1
where in the body does protein d igestion begin	Non−geographic	Non−geographic	1
columbia ms in what county	GPE	GPE	1
where is telephone area code 71 6	Locators	Locators	1
what district is mt vernon locate d	GPE	GPE	1
what county is orange park, fl in?	GPE	GPE	1
what region is sanford nc consid ered in nc	GPE	GPE	1
where is the forgotten river	Geographic Feature	Geographic Feature	1
where to find saguaro cactus tru nks	Habitat	Habitat	1
what city is the himalayas in	Geographic Feature	Geographic Feature	1
what county is rice, va	GPE	GPE	1
where is ooltewah tennessee	GPE	GPE	1
where can petroleum be found	Origination/Product ion	Origination/Product ion	1

Test Result			
Text	Predicted	Topic	Accuracy
where are the schedules saved in easy worship 6	Non-geographic	Non-geographic	1
where is bashur air field	Airport	Airport	1
where is the oxford institute for energy studies based	Building	Building	1
where does red eyed tree frog live	Habitat	Habitat	1
what county in aztec nm in	GPE	GPE	1
which miami beach pharmacist invented the first suntan cream	Building	Non-geographic	0
what county is delton mi	GPE	GPE	1
where is greek peak ny	GPE	Geographic Feature	0
where is columbus montana located	GPE	GPE	1
where is rolling hills estates ca vs san jose	GPE	GPE	1
where is rock river in iowa	Geographic Feature	Geographic Feature	1
what county is bladensburg oh	GPE	GPE	1
where is nickelback from	Origination/Production	GPE	0
where is gibtown	GPE	GPE	1
where is charles hill city map	GPE	GPE	1
where is 01268 postcode	Locators	Locators	1
where did denmark get education	Building	Building	1
where to find dividends rate on financial statements	Non-geographic	Non-geographic	1
where is tobago	GPE	GPE	1
where does angina radiate	Non-geographic	Non-geographic	1
where does the last name sabo come from	Non-geographic	Non-geographic	1
what district is the superdome in	Building	Building	1
where does sugar grow in australia	Origination/Production	Habitat	0
where is the grand canyon located latitude and longitude	Locators	Locators	1
where were bananas first domesticated	Origination/Production	Origination/Production	1
what county is caseyville il in	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where is watershed in mississipp i	Geographic Feature	Geographic Feature	1
where is albion prison	Building	Building	1
where is cayman islands	GPE	GPE	1
what airport is kul	Airport	Airport	1
where was kartini born	Place of Birth/Raise d	Place of Birth/Raise d	1
where is los angeles	GPE	GPE	1
where is centennial center	Building	Building	1
where can i go to get my body fa t precentage checked in oceansid e ca	Building	Building	1
where is morgantown	GPE	GPE	1
where can the goblin shark be fo und	Habitat	Habitat	1
what county is thomson ga locate d	GPE	GPE	1
where are salicylates absorbed	Non−geographic	Non−geographic	1
penn state pattee library hours	Building	Building	1
what is flr airport	Airport	Airport	1
where is the xfinity theater?	Building	Building	1
where do you mail form 941 qua rterly report	Non−geographic	Building	0
what county is paulden az in	GPE	GPE	1
what county is idaho falls, idaho	GPE	GPE	1
wow where is the timewalking v endor	Non−geographic	Non−geographic	1
where is the control center for r espiration located	Non−geographic	Non−geographic	1
where do they film broadchurch	Film Location	Film Location	1
which ocean zone is the biggest	Locators	Locators	1
where is strange oklahoma	GPE	GPE	1
what county is medford, mass	GPE	GPE	1
what maryland county is andrew s afb located	GPE	GPE	1
where is the acl?	Non−geographic	Non−geographic	1

Test Result			
Text	Predicted	Topic	Accuracy
what part of florida is newport beach	GPE	Geographic Feature	0
where is paulsboro new jersey	GPE	GPE	1
where is trotwood ohio	GPE	GPE	1
what county is san pablo ca	GPE	GPE	1
what county is doral, fl	GPE	GPE	1
what county is johns island sc	GPE	GPE	1
where is fume on the map	GPE	Non−geographic	0
what county is wolfville ns	GPE	GPE	1
where did american bandstand originate	Origination/Production	Origination/Production	1
where is the fuel tank pressures	Non−geographic	Non−geographic	1
what time zone is memphis tennessee	Locators	Locators	1
where is independence kentucky on the kentucky map?	GPE	GPE	1
what county is castlewood va located in	GPE	GPE	1
where is the deepest ocean	Geographic Feature	Geographic Feature	1
what province of brazil is brasilia in?	GPE	GPE	1
where is chianti found	Non−geographic	Origination/Production	0
where is coopers creek track thomson river at	Geographic Feature	Geographic Feature	1
where was the caton ave pit beef stand	Building	Building	1
where is alexandria alabama located	GPE	GPE	1
where is lacewood drive, halifax	GPE	GPE	1
where can i buy crown magnolia paint	Building	Building	1
what county is riverside ca in	GPE	GPE	1
where is the us alamo	GPE	GPE	1
what county is kennard tx in	GPE	GPE	1
where is bard college located	Building	Building	1
what county is johnson city tx	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where is nasa's main engineering location?	Building	Building	1
enumclaw is in what county wa	GPE	GPE	1
where is baumholder germany located	GPE	GPE	1
county where wild horse ranch is	GPE	GPE	1
where is internet explorer favorites bar stored	Non-geographic	Non-geographic	1
which county is daresbury warrington in	GPE	GPE	1
where to find tapestries	Origination/Production	Building	0
what town is near windham maine	GPE	GPE	1
what bases with adenine in dna	Non-geographic	Non-geographic	1
where was goonies filmed	Film Location	Film Location	1
where is dr. nowzaradan from	Place of Birth/Raised	Place of Birth/Raised	1
what county is ottumwa iowa in?	GPE	GPE	1
what county is anthon, iowa in	GPE	GPE	1
what county is spring valley il in	GPE	GPE	1
what continent is madagascar in	GPE	GPE	1
what county is tunica ms located in	GPE	GPE	1
where do minks live	Habitat	Habitat	1
des moines what co	GPE	GPE	1
where do the goths come from	Origination/Production	Origination/Production	1
where is tonga	GPE	GPE	1
in what county is pontiac, mi	GPE	GPE	1
what county is springfield gardens	GPE	GPE	1
where is leonard cohen from nationality	Place of Birth/Raised	Place of Birth/Raised	1
what county is winslow nj in	GPE	GPE	1
where is cricklewood	GPE	GPE	1
where is texas a&m university	Building	Building	1

Test Result			
Text	Predicted	Topic	Accuracy
seafood buffet in colonial williamsburg	Building	Building	1
where, – is nome texas, – located – – – ? –	GPE	GPE	1
where was robert m green born	Place of Birth/Raised	Place of Birth/Raised	1
where is solvang ca	GPE	GPE	1
where is los santos	GPE	GPE	1
where is the pku gene located	Non–geographic	Non–geographic	1
is samoa a country	GPE	GPE	1
which airport code is ewr	Airport	Locators	0
county is iowa is unionville	GPE	GPE	1
where is bullnose tile used	Origination/Production	Non–geographic	0
jackson is what county in tn	GPE	GPE	1
where was gossip girl filmed	Film Location	Film Location	1
what county is harrison township indiana in	GPE	GPE	1
where is claptrap in salt flats	Geographic Feature	Non–geographic	0
where is felandaris in dragon age inquisition	Non–geographic	Non–geographic	1
what county is loomis california	GPE	GPE	1
what county is cedar point in	GPE	GPE	1
biggest treasure ever found	Geographic Feature	Non–geographic	0
where is willits ca	GPE	GPE	1
where else can you use a victoria secret credit card at	Building	Building	1
where is columbia city indiana	GPE	GPE	1
where are chiggers found	Habitat	Habitat	1
where is your right flank located	Non–geographic	Non–geographic	1
what is the deepest place in the ocean called	Geographic Feature	Geographic Feature	1
what county is bristol indiana in	GPE	GPE	1
where is santa ana	GPE	GPE	1
where is the cochlear duct	Non–geographic	Non–geographic	1
what county is happy camp in	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where is thunder hole in maine	Geographic Feature	Geographic Feature	1
where does fort name originated from	Non-geographic	Origination/Production	0
where is bgf airport	Airport	Airport	1
what beaches are near san jose california airport	Airport	Geographic Feature	0
where is brain sand found	Non-geographic	Non-geographic	1
what county is bayfield co in	GPE	GPE	1
where is soil moisture found?	Non-geographic	Non-geographic	1
what county hartwell, ga in	GPE	GPE	1
where is mandarin spoken outside of china	Origination/Production	Non-geographic	0
where is the sun above the earth during solstice	Non-geographic	Locators	0
what area do jaguars live	Habitat	Habitat	1
where is danscomp located	GPE	Building	0
what county is herndon va	GPE	GPE	1
where is acuna mexico	GPE	GPE	1
where is beirut lebanon	GPE	GPE	1
what county is vega tx in	GPE	GPE	1
where is east asian	GPE	GPE	1
what county is snow hill md	GPE	GPE	1
where are particles with virtually no mass found in the atom	Non-geographic	Non-geographic	1
what town is meteora in greece	GPE	GPE	1
which countries gems come from	Origination/Production	Origination/Production	1
where is my colon located in my body	Non-geographic	Non-geographic	1
what county is state center, iowa in	GPE	GPE	1
in what county is leland, nc located?	GPE	GPE	1
where is the bacterial cell's dna found	Non-geographic	Non-geographic	1
which ocean is dendy street beach	Geographic Feature	Geographic Feature	1

Test Result			
Text	Predicted	Topic	Accuracy
which belongs in the domain euk arya	Non−geographic	Non−geographic	1
where is gironde in france	GPE	GPE	1
where is the snaefellsnes penins ula	GPE	Geographic Feature	0
where do the ukranian people co me from	Origination/Product ion	Origination/Product ion	1
where did b boying originate fro m	Origination/Product ion	Origination/Product ion	1
where is preproinsulin cleaved	Non−geographic	Non−geographic	1
where is girona airport in spain	Airport	Airport	1
where is casuarina	GPE	GPE	1
what county is raleigh, nc	GPE	GPE	1
where is cesare borgia family fro m	Origination/Product ion	Origination/Product ion	1
where does a thesis statement g o in mla format	Non−geographic	Non−geographic	1
where do most metamorphic roc ks form	Geographic Feature	Geographic Feature	1
what casino is in elgin illinois	Building	Building	1
where is dunstable, ma located	GPE	GPE	1
what county includes edgerton, k s	GPE	GPE	1
where is asheboro, nc	GPE	GPE	1
what county is carlisle ky in	GPE	GPE	1
where is the gist settlement	GPE	GPE	1
what county is garden city li in	GPE	GPE	1
what county is flushing new york located	GPE	GPE	1
where is wtae broadcast station	Building	Building	1
where is the pacific ocean locate d on the ap	Geographic Feature	Geographic Feature	1
where is the picketwire	Building	Geographic Feature	0
what is the flag 4 honduras	Non−geographic	Non−geographic	1
what town is near banff canada	GPE	GPE	1
where was basf founded	Origination/Product ion	Origination/Product ion	1

Test Result			
Text	Predicted	Topic	Accuracy
where is cathedral quarter belfas t	GPE	Building	0
what county is germany valley w v in	GPE	GPE	1
where is the robot used	Non−geographic	Non−geographic	1
which states in the u. a produce watermelons	Origination/Product ion	Origination/Product ion	1
where is prattsville, ny	GPE	GPE	1
what county is arden nc	GPE	GPE	1
where to find the dinka	Building	Non−geographic	0
where is asl originate from	Origination/Product ion	Origination/Product ion	1
where is coca cola located	Building	Building	1
where can australians go to find i nfos about alcohol and staying sa fe	Building	Building	1
where was fences filmed	Film Location	Film Location	1
where are vitamins found	Non−geographic	Non−geographic	1
where is cranbourne	GPE	GPE	1
is canaan columbia county	GPE	GPE	1
what county is iowa falls iowa in?	GPE	GPE	1
where is los alamos	GPE	GPE	1
what us state is farthest east	GPE	GPE	1
where is itasca city	GPE	GPE	1
where is the valle dei templi	Geographic Feature	GPE	0
where does witch hazel grow	Habitat	Habitat	1
what region is boracay island?	GPE	GPE	1
which city does the iditarod take place in	Locators	GPE	0
what state of germany is ulm in	GPE	GPE	1
where is josh gates from	Place of Birth/Raise d	GPE	0
what county is hartsfield ga in	GPE	GPE	1
where is new castle pa	GPE	GPE	1
where are c	GPE	Non−geographic	0

Test Result			
Text	Predicted	Topic	Accuracy
where is roscommon michigan	GPE	GPE	1
where does the african elephant live	Habitat	Habitat	1
where is gladys va	GPE	GPE	1
what county is clewiston in?	GPE	GPE	1
where and what is tabor academ y	Building	Building	1
what county in iowa is leclaire in	GPE	GPE	1
what county is apple springs tx i n	GPE	GPE	1
where is young jeezy from	Place of Birth/Raise d	Place of Birth/Raise d	1
what county is allentown, nj	GPE	GPE	1
what county is dunnsville va in	GPE	GPE	1
augusta nj is in what county	GPE	GPE	1
where is australia located geogra phically	GPE	GPE	1
where is residence heracles in b ordeaux france	Building	Building	1
what region is kenya in	GPE	GPE	1
where is chewy pet supplies loca ted	Building	Building	1
where is greenworks located	Building	Building	1
where is zip code what county is 19083	Locators	GPE	0
what is the abbreviation for anno ny	Non-geographic	Non-geographic	1
where is galilee located	GPE	GPE	1
where did the phenomenon of de vil's footprints occur?	Non-geographic	GPE	0
where is olathe	GPE	GPE	1
what town is university of north carolina main campus located at	Building	GPE	0
where is journey lead singer fro m	Place of Birth/Raise d	Place of Birth/Raise d	1
where is mexico city located at	GPE	GPE	1
what county is lynden wa is	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
The Hawaiian Islands were formed from: mountains deltas volcanoes	Geographic Feature	Non−geographic	0
canada club hunt club in walpole island	GPE	GPE	1
where is nat pagle	GPE	Non−geographic	0
where is preston idaho	GPE	GPE	1
where is new zealand new zealand	GPE	GPE	1
where is pontypandy	GPE	GPE	1
what airport is near lee university	Airport	Airport	1
where is lawrenceville georgia	GPE	GPE	1
what county is downingtown pa in	GPE	GPE	1
what county is wyncote, pa	GPE	GPE	1
what county is wando river	Geographic Feature	GPE	0
where is yokosuka	GPE	GPE	1
where was st patrick born in scotland	Place of Birth/Raised	Place of Birth/Raised	1
where do new oceans form, quizlet	Non−geographic	Non−geographic	1
what county is sevierville, sc in	GPE	GPE	1
what county is ridgeway wi in	GPE	GPE	1
where is arkport ny	GPE	GPE	1
where is westford	GPE	GPE	1
what county is framingham mass in?	GPE	GPE	1
where is sultan washington	GPE	GPE	1
what time zone is henderson, nv	Locators	Locators	1
where is douglas, az	GPE	GPE	1
what places are near residence inn marriott danvers ma	Building	Building	1
where is sardinia?	GPE	GPE	1
where is the north valley	Geographic Feature	GPE	0
what continent is poland in	GPE	GPE	1

Test Result			
Text	Predicted	Topic	Accuracy
where was robert fulton from	Place of Birth/Raised	Place of Birth/Raised	1
what is canada country code	Locators	Locators	1
where is bluffton ohio located	GPE	GPE	1
where is ningbo	GPE	GPE	1
what county is minooka illinois in	GPE	GPE	1
where is columns in word 2013	Non-geographic	Non-geographic	1
where is mount saint helens located	Geographic Feature	Geographic Feature	1
where is bevier mo	GPE	GPE	1
where is piedmont	GPE	GPE	1
what county is alma mi in	GPE	GPE	1
what territory is ontario in	GPE	GPE	1
where is angers france	GPE	GPE	1
where is grove hill alabama	GPE	GPE	1
where is ford taurus made	Origination/Production	Origination/Production	1
where does water come in to the sump pit	Non-geographic	Non-geographic	1
where does bismuth oxychloride come from	Non-geographic	Non-geographic	1
what airport serves niagara falls	Airport	Airport	1
where is orangeville ny	GPE	GPE	1
what county is spring valley mn	GPE	GPE	1
what county is california city, ca in?	GPE	GPE	1
where is the navigator of the seas	Non-geographic	Geographic Feature	0
what is the name of the town outside london where the swans inn is located	Building	GPE	0
where did emily carr study art	Building	Building	1
where is maxim's fiscal calendar	Non-geographic	Non-geographic	1
where did the saying cross your fingers come from	Non-geographic	Origination/Production	0
where is hm cut in pokemon brickbronze	Non-geographic	Non-geographic	1

Test Result			
Text	Predicted	Topic	Accuracy
where is langley field located	Origination/Production	GPE	0
where is rice grown	Origination/Production	Habitat	0
where was the choice filmed	Film Location	Film Location	1
where is cave of the winds	Geographic Feature	Geographic Feature	1
what state is tasmania in	GPE	GPE	1
where is mary rice incarcerated	Place of Birth/Raised	Building	0
where are polar bears from	Habitat	Origination/Production	0
where is portage michigan	GPE	GPE	1
what county is north port fl in	GPE	GPE	1
which zone would phytoplankton be found	Habitat	Non−geographic	0
where does the speaker keep the sabbath	Non−geographic	Non−geographic	1
where do luau leaves come from	Origination/Production	Origination/Production	1
where is sam houston university	Building	Building	1
where is the 326 area code	Locators	Locators	1
where was the movie silverado filmed	Film Location	Film Location	1
where was the movie, the safe haven filmed ?	Film Location	Film Location	1
where was horse whisperer filmed	Film Location	Film Location	1
where was superman filmed in michigan	Film Location	Film Location	1
where was lars van triers movie antichrist filmed	Film Location	Film Location	1

Train Validation Metrics											
Epoch 01		***Epoch 11***		***Epoch 21***		***Epoch 31***		***Epoch 41***		***Epoch 51***	
Training loss	*1.28465427*	*Training loss*	*0.0326224*	*Training loss*	*5.57E-05*	*Training loss*	*0.01893636*	*Training loss*	*0.00377009*	*Training loss*	*1.51E-06*
Validation loss	*0.95295945*	*Validation loss*	*1.1618198*	*Validation loss*	*1.3818998*	*Validation loss*	*1.3225457*	*Validation loss*	*1.64320305*	*Validation loss*	*1.65863517*
F1-Score	*0.71692939*	*F1-Score*	*0.84244684*	*F1-Score*	*0.85495523*	*F1-Score*	*0.87368531*	*F1-Score*	*0.85931994*	*F1-Score*	*0.85976598*
Epoch 02		***Epoch 12***		***Epoch 22***		***Epoch 32***		***Epoch 42***		***Epoch 52***	
Training loss	*0.62420018*	*Training loss*	*0.00039765*	*Training loss*	*5.02E-05*	*Training loss*	*0.00657085*	*Training loss*	*0.00995968*	*Training loss*	*1.44E-06*
Validation loss	*0.65514147*	*Validation loss*	*1.10368578*	*Validation loss*	*1.53490739*	*Validation loss*	*1.4317836*	*Validation loss*	*1.58054497*	*Validation loss*	*1.66775941*
F1-Score	*0.84645885*	*F1-Score*	*0.8651704*	*F1-Score*	*0.84412457*	*F1-Score*	*0.87111875*	*F1-Score*	*0.86485034*	*F1-Score*	*0.85976598*
Epoch 03		***Epoch 13***		***Epoch 23***		***Epoch 33***		***Epoch 43***		***Epoch 53***	
Training loss	*0.34023603*	*Training loss*	*0.01084797*	*Training loss*	*0.00127058*	*Training loss*	*0.00807903*	*Training loss*	*2.97E-06*	*Training loss*	*1.31E-06*

Train Validation Metrics											
Validation loss	*0.74400103*	*Validation loss*	*1.16546328*	*Validation loss*	*1.47223321*	*Validation loss*	*1.42821367*	*Validation loss*	*1.58886298*	*Validation loss*	*1.67457449*
F1−Score	*0.84838846*	*F1−Score*	*0.86308076*	*F1−Score*	*0.85279541*	*F1−Score*	*0.86002023*	*F1−Score*	*0.86485034*	*F1−Score*	*0.85976598*
Epoch 04		***Epoch 14***		***Epoch 24***		***Epoch 34***		***Epoch 44***		***Epoch 54***	
Training loss	*0.22002558*	*Training loss*	*0.02006465*	*Training loss*	*0.01709253*	*Training loss*	*1.31E−05*	*Training loss*	*0.00100492*	*Training loss*	*0.00112493*
Validation loss	*0.86016928*	*Validation loss*	*1.20652516*	*Validation loss*	*1.48865138*	*Validation loss*	*1.43650846*	*Validation loss*	*1.7312295*	*Validation loss*	*1.72436293*
F1−Score	*0.85219446*	*F1−Score*	*0.85557159*	*F1−Score*	*0.85661183*	*F1−Score*	*0.86582712*	*F1−Score*	*0.86037614*	*F1−Score*	*0.85512992*
Epoch 05		***Epoch 15***		***Epoch 25***		***Epoch 35***		***Epoch 45***		***Epoch 55***	
Training loss	*0.12761631*	*Training loss*	*0.00425748*	*Training loss*	*0.00628658*	*Training loss*	*6.79E−06*	*Training loss*	*0.00811626*	*Training loss*	*1.14E−06*
Validation loss	*0.93126456*	*Validation loss*	*1.44943219*	*Validation loss*	*1.3789298*	*Validation loss*	*1.40393497*	*Validation loss*	*1.68294666*	*Validation loss*	*1.72929652*
F1−Score	*0.84902319*	*F1−Score*	*0.83548911*	*F1−Score*	*0.8603707*	*F1−Score*	*0.87138875*	*F1−Score*	*0.86036675*	*F1−Score*	*0.85512992*

Train Validation Metrics											
Epoch 06		***Epoch 16***		***Epoch 26***		***Epoch 36***		***Epoch 46***		***Epoch 56***	
Training loss	*0.10324723*	*Training loss*	*0.00611134*	*Training loss*	*0.00790245*	*Training loss*	*5.67E−06*	*Training loss*	*4.11E−06*	*Training loss*	*1.59E−05*
Validation loss	*0.94586648*	*Validation loss*	*1.26429237*	*Validation loss*	*1.3604804*	*Validation loss*	*1.42312703*	*Validation loss*	*1.70126212*	*Validation loss*	*1.72192361*
F1−Score	*0.84506394*	*F1−Score*	*0.85358069*	*F1−Score*	*0.86719527*	*F1−Score*	*0.87138875*	*F1−Score*	*0.85089026*	*F1−Score*	*0.85987208*
Epoch 07		***Epoch 17***		***Epoch 27***		***Epoch 37***		***Epoch 47***		***Epoch 57***	
Training loss	*0.07966039*	*Training loss*	*0.03425391*	*Training loss*	*0.00712719*	*Training loss*	*5.57E−06*	*Training loss*	*2.13E−06*	*Training loss*	*1.04E−06*
Validation loss	*0.92640453*	*Validation loss*	*1.24786831*	*Validation loss*	*1.41065752*	*Validation loss*	*1.42138984*	*Validation loss*	*1.70868573*	*Validation loss*	*1.73885995*
F1−Score	*0.85692216*	*F1−Score*	*0.8561391*	*F1−Score*	*0.86841934*	*F1−Score*	*0.8757496*	*F1−Score*	*0.85089026*	*F1−Score*	*0.85735138*
Epoch 08		***Epoch 18***		***Epoch 28***		***Epoch 38***		***Epoch 48***		***Epoch 58***	
Training loss	*0.05981024*	*Training loss*	*0.00339888*	*Training loss*	*0.01487084*	*Training loss*	*1.61E−05*	*Training loss*	*1.93E−06*	*Training loss*	*9.70E−07*

Train Validation Metrics											
Validation loss	*1.03144806*	*Validation loss*	*1.53427871*	*Validation loss*	*1.47722173*	*Validation loss*	*1.45376526*	*Validation loss*	*1.71486276*	*Validation loss*	*1.74237238*
F1-Score	*0.84877681*	*F1-Score*	*0.84769226*	*F1-Score*	*0.86855615*	*F1-Score*	*0.86448037*	*F1-Score*	*0.85089026*	*F1-Score*	*0.85735138*
Epoch 09		***Epoch 19***		***Epoch 29***		***Epoch 39***		***Epoch 49***		***Epoch 59***	
Training loss	*0.04595679*	*Training loss*	*0.01387798*	*Training loss*	*0.00587361*	*Training loss*	*9.84E-06*	*Training loss*	*0.00127797*	*Training loss*	*9.38E-07*
Validation loss	*1.04595129*	*Validation loss*	*1.4166894*	*Validation loss*	*1.35279275*	*Validation loss*	*1.52238503*	*Validation loss*	*1.64346275*	*Validation loss*	*1.74492875*
F1-Score	*0.85156951*	*F1-Score*	*0.85403659*	*F1-Score*	*0.87797493*	*F1-Score*	*0.86487874*	*F1-Score*	*0.85980539*	*F1-Score*	*0.85735138*
Epoch 10		***Epoch 20***		***Epoch 30***		***Epoch 40***		***Epoch 50***		***Epoch 60***	
Training loss	*0.02374367*	*Training loss*	*0.03263084*	*Training loss*	*0.00160188*	*Training loss*	*0.0090539*	*Training loss*	*1.62E-06*	*Training loss*	*9.48E-07*
Validation loss	*1.12243125*	*Validation loss*	*1.46142914*	*Validation loss*	*1.60407966*	*Validation loss*	*1.52393331*	*Validation loss*	*1.65066686*	*Validation loss*	*1.74596875*
F1-Score	*0.8491981*	*F1-Score*	*0.83871921*	*F1-Score*	*0.85418501*	*F1-Score*	*0.85604454*	*F1-Score*	*0.85980539*	*F1-Score*	*0.85735138*

3. *Bibliography*

Allam, A. M. N., & Haggag, M. H. (2012). The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences*, *2*(3).

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., & Wang, T. (2018). *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset* (arXiv:1611.09268). arXiv. http://arxiv.org/abs/1611.09268

Cortes, E. G., Woloszyn, V., Binder, A., Himmelsbach, T., Barone, D., & Moller, S. (2020). An Empirical Comparison of Question Classification Methods for Question Answering Systems. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 5408–5416.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Edwardes, A. J., & Purves, R. S. (2007). Eliciting concepts of *place* for text-based image retrieval. *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, 15–18. https://doi.org/10.1145/1316948.1316953

González-Carvajal, S., & Garrido-Merchán, E. C. (2023). Comparing BERT against traditional machine learning text classification. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE3202838

Hamzei, E. (2021). *Place-related Question Answering:From Questions to Relevant Answers*. Department of Infrastructure

Engineering The University of Melbourne.

Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S., & Tomko, M. (2020). Place Questions and Human-Generated Answers: A Data Analysis Approach. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Mansourian (Eds.), *Geospatial Technologies for Local and Regional Development* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-14745-7_1

Hamzei, E., Tomko, M., & Winter, S. (2022). Translating Place-Related Questions to GeoSPARQL Queries. *Proceedings of the ACM Web Conference 2022*, 902–911. https://doi.org/10.1145/3485447.3511933

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Khilji, A. F. U. R., Manna, R., Laskar, S. R., Pakray, P., Das, D., Bandyopadhyay, S., & Gelbukh, A. (2020). Question Classification and Answer Extraction for Developing a Cooking QA System. *Computación y Sistemas*, *24*(2). https://doi.org/10.13053/cys-24-2-3445

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Li, X., & Roth, D. (2002). Learning question classifiers. *Proceedings of the 19th International Conference on Computational Linguistics -*, *1*, 1–7. https://doi.org/10.3115/1072228.1072378

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., & De Rijke, M. (2004). Creating the DIS

EQuA Corpus: A Test Set for Multilingual Question Answering. In C. Peters, J. Gonzalo, M. Braschler, & M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems* (Vol. 3237, pp. 487–500). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30222-3_47

Mai, G., Janowicz, K., Zhu, R., Cai, L., & Lao, N. (2021). Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions. *AGILE: GIScience Series*, *2*, 1–21. https://doi.org/10.5194/agile-giss-2-8-2021

Mallikarjuna, C., & Sivanesan, S. (2022). Question classification using limited labelled data. *Information Processing & Management*, *59*(6), 103094. https://doi.org/10.1016/j.ipm.2022.103094

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv. http://arxiv.org/abs/1310.4546

Mishra, A., Mishra, N., & Agrawal, A. (2010). Context-Aware Restricted Geographical Domain Question Answering System. *2010 International Conference on Computational Intelligence and Communication Networks*, 548–553. https://doi.org/10.1109/CICN.2010.108

Mohasseb, A., Bader-El-Den, M., & Cocea, M. (2018). Question categorization and classification using grammar based approach. *Information Processing & Management*, *54*(6), 1228–1243. https://doi.org/10.1016/j.ipm.2018.05.001

Oguz, B., Chen, X., Karpukhin, V., Peshterliev, S., Okhonko, D., Schlichtkrull, M., Gupta, S., Mehdad, Y., & Yih, S. (2022). UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering. *Findings of the Association for Computational Linguistics: NAACL 202*

2, 1535–1546. https://doi.org/10.18653/v1/2022.findings−naacl.115

Punjani, D., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., Beris, T., Bilidas, D., Ioannidis, T., Karalis, N., Lange, C., Pantazi, D., Papaloukas, C., & Stamoulis, G. (2018). Template−Based Question Answering over Linked Geospatial Data. *Proceedings of the 12th Workshop on Geographic Information Retrieval*, 1–10. https://doi.org/10.1145/3281354.3281362

Salton, G., & Buckley, C. (1988). Term−Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306−4573(88)90021−0

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2021). Geo−analytical question−answering with GIS. *International Journal of Digital Earth*, *14*(1), 1–14. https://doi.org/10.1080/17538947.2020.1738568

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). *How to Fine−Tune BERT for Text Classification?* (arXiv:1905.05583). arXiv. http://arxiv.org/abs/1905.05583

Sundblad, H. (2007). *Question Classification in Question Answering Systems*. Department of Computer and Information Science Linköpings universitet.

Wang, Z., & Qu, Z. (2017). Research on Web text classification algorithm based on improved CNN and SVM. *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 1958–1961. https://doi.org/10.1109/ICCT.2017.8359971

Xu, H., Hamzei, E., Nyamsuren, E., Kruiger, H., Winter, S., Tomko, M., & Scheider, S. (2020). Extracting interrogative intents and concepts from geo−analytic questions. *AGILE: GIScience Series*, *1*, 1–21. https://doi.org/10.5194/agile−giss−1−23−

2020

Xu, H., Nyamsuren, E., Scheider, S., & Top, E. (2022). A grammar for interpreting geo-analytical questions as concept transformations. *International Journal of Geographical Information Science*, *37*(2), 276-306. https://doi.org/10.1080/13658816.2022.2077947

Yang, J., Jang, H., & Yu, K. (2023). Analyzing Geographic Questions Using Embedding-based Topic Modeling. *ISPRS International Journal of Geo-Information*, *12*(2), 52. https://doi.org/10.3390/ijgi12020052

Zhang, D., & Lee, W. S. (2003). Question Classification using Support Vector Machines. *SIGIR*, 26-32.

Zhou, T. C., Lyu, M. R., & King, I. (2012). A classification-based approach to question routing in community question answering. *Proceedings of the 21st International Conference on World Wide Web*, 783-790. https://doi.org/10.1145/2187980.2188201

4. Abstract (Korean)

GeoQA 시스템 성능 향상을 위한 장소 관련 질문 분류

서울대학교 공과대학

건설환경공학부 대학원

이 서 형

질문 분류(Question Classification, QC)는 질문의 주제를 식별함으로써 사용자에게 정확하고 관련성 높은 답변을 제공하는 데 중요한 역할을 한다. 질문 응답(QA) 시스템에서 정확한 질문 분류는 사용자의 문의 사항에 정확히 대응하고 정보 검색 (Information Retrieval, IR) 과정에서 관련성 높은 답변을 가져오는 데 필수적이다. 이와 같이 질문을 세분화하여 분류함으로써 IR 과정에서 더 정확한 정보 검색이 이루어질 수 있도록 다양한 도메인에서 질문 분류를 개선하는 데 상당한 연구가 진행되고 있다. 그러나 GeoQA(지리 질의 응답) 시스템에서는 다양한 주제를 포괄하는 장소 관련 질문은 분류하는데 어려움이 있다.

GeoQA 관련 분야에서는 질문의 지리적 속성을 확인하고 질문 생성을

위해 지리적 질문의 구조적 패턴을 분석하며, 잠재적 주제에 기반한 지리적 질문을 분류하는 노력이 이루어졌지만, 장소 관련 질문 내에서 특정 주제를 식별하는 것에 대한 연구는 아직 공백이 남아있다. 본 연구는 이러한 공백을 메우기 위해 장소 관련 질문에서 특정 주제를 정확하게 분류하는 방법론을 개발함으로써 GeoQA 시스템에서의 질문 분류 성능 향상에 기여하고자 한다.

본 논문은 미리 세분화하여 정의한 주제를 활용하여 지리 분야에서 장소 관련 질문을 분류하기 위한 방법론을 제시한다. 목표는 장소 관련 질문 내에서 주제를 정확하게 식별하는 것으로, 이는 지리적 위치에 대한 자연어 질의에 관련성 높은 정보를 제공하는 데 중요하다. 예를 들어, “Henderson, zone”과 같은 질문이 GeoQA 시스템에 들어오게 되면 사용자가 Henderson, TN의 시간대에 대한 것인지 혹은 한계구역(hardiness zone)에 대한 것인지 알 수 없으므로 이 질문을 “Locator” 주제로 분류하는 것이 중요하다. 따라서, 본 연구에서는 세분화된 주제를 활용하여 GeoQA의 성능을 향상시킬 수 있는 장소 관련 질문 분류를 수행하고자 한다.

연구를 진행하기 위해, MS MARCO 데이터셋에서 임의로 선택된 3,025개의 장소 관련 질문을 42개의 세분화된 주제로 라벨링 작업을 진행했다. 그리고 BERT 모델을 활용하여 사용자 관심사 기반의 장소 관련 질문을 분류하기 위해 Fine-Tuning 작업을 수행했다. 본 연구에서 수행한 질문 분류 모델은 학습 정확도 87.9%와 테스트 정확도 86.6%를 달성하여 장소 관련

질문 분류에 효과를 입증하였다.

본 연구에서 제안한 다중 클래스 질문 분류 모델은 GeoQA 시스템에 중요한 기여를 하는데, MS MARCO 데이터셋에서 장소 관련 질문의 관련 주제에 대한 질문 분류 방법론을 제시한다. 본 연구는 장소 관련 질문에 관련한 공간 관련 질의 모델 (GeoQA) 시스템에서 정보 탐색이 이루어지기 전에 질문에 대한 분류가 이루어짐으로써 GeoQA의 성능 향상 대한 새로운 접근 방식을 제공하여 지리 공간 관련 자연어 처리 분야에 중요한 기여를 한다.

Keywords: GeoQA, GeoQA dataset, Close-domain QA System, Question Classification, Multi-Class Question Classification

Student Number: 2021-22601