공학석사 학위논문

다양한 문장형태로 표현된 지리공간질의에 대한 기계학습기반 지리공간분석절차 변환기법 개발

Development of Machine Learning Based Geographic Analysis Workflow Transduction Technique for Geographic Questions with Various Sentence Type

2023년 2월

서울대학교 대학원

건설환경공학부

채 희 진

다양한 문장형태로 표현된 지리공간질의에 대한 기계학습기반 지리공간분석절차 변환기법 개발

Development of Machine Learning Based Geographic Analysis Workflow Transduction Technique for Geographic Questions with Various Sentence Type

지도 교수 유 기 윤

이 논문을 공학석사 학위논문으로 제출함

2022년 10월

서울대학교 대학원

건설환경공학부

채 희 진

채희진의 공학석사 학위논문을 인준함

2023년 1월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

국문초록

문서로부터 질문에 대한 답을 간결하고 명확하게 도출하는 질의응답(question answering, 이하 QA) 분야의 발전에도 불구하고 연간 20% 내외 증가하는 지리공간정보(geographic information)와 관련된 질의를 답하는 시스템은 부족한 상태다. 이를 해결하기 위해 등장한 연구 분야가 지리공간 질의응답(geographic QA)이고 이 중 지리공간분석 질의응답(geographic analysis question answering, 이하 Geo-analytical QA)은 지리공간질의(geographic question)를 지리공간분석절차로 변환하고 이를 수행하기 적합한 데이터와 도구를 탐색하는 연구 분야다. 현실적인 Geo-analytical QA를 수행하기 위해서는 다양한 문장 형태를 가진 질의를 지리공간분석절차로 변환할 수 있어야 하지만 기존 연구에서 제안한 방법은 제한된 문장형태에 대해 규칙 기반 방식을 통해 문장을 분석절차로 변환하기 때문에 현실적인 Geo-analytical QA를 수행하기 어렵다는 한계를 지닌다. 따라서 본 연구에서는 현실적인 Geo-analytical QA를 수행하기 위해 다양한 문장 형태를 가지는 지리공간질의를 지리공간분석절차로 변환하는 방안을 제시하고자 한다. 또한, 지리공간분석을 실제로 수행하기 위해서는 지리공간연산함수를 파악하는 것이 중요하기 때문에 도출한 지리공간분석절차가 지리공간연산함수를 분석 의도에 맞게 순서대로 포함하도록 설정했다. 지리공간질의를 분석절차로 변환하기 위해서 본 연구에서는 문장 분류(text classification)기법을 적용했고, 문장 분류 기법을 이용하기 위해서는 문서를 목적에 맞게 모아 놓은 말뭉치(corpus) 선정, 말뭉치를 라벨링 해 데이터셋 생성, 데이터셋을

분류모델(classification model)의 입력값으로 만들기 위해 말뭉치에 등장하는 질의를 임베딩(embedding)하는 과정, 그리고 각 임베딩과 라벨로 이루어진 데이터셋을 이용해 분류모델을 학습하는 과정이 필요하다. 질의를 답하기 위해 다양한 지리 공간 분석절차를 이용해야 하는 것으로 알려진 GeoAnQu 말뭉치를 대상 말뭉치로 선정하고 분석해서 고유한 분석절차를 도출한 후 해당 분석절차에 고유 번호를 부여했다. 해당 고유번호를 기준으로 GeoAnQu 말뭉치에 등장하는 질의에 대해 라벨링을 수행해 데이터셋을 확보한 후 다양한 문장형태 생성 및 데이터셋 증강을 위해 어휘변용(paraphrase)을 실시했다. 그 후 해당 데이터셋을 Glove(global vectors), BERT(bidirectional encoder representations from transformers), RoBERTa(robustly optimized BERT pre-training approach), SBERT(sentence-BERT)를 이용해 문장 임베딩을 수행하고 각각 임베딩을 linear SVM(support vector machine), 랜덤포레스트(random forest)을 이용해 학습시켰다. 최종적으로 SBERT 문장 임베딩을 linear SVM에 학습시킨 모델이 가장 높은 성능을 보이는 것을 확인할 수 있었고, 해당 모델을 통해 다양한 문장형태를 가지는 지리공간 질의를 지리공간분석절차로 변환할 수 있었다. 또한 해당 결과의 한계점을 분석해 향후 연구 방향을 제시했다.

주요어 : 지리공간 질의응답, 문장분류, 지리공간분석 질의응답, 지리공간 말뭉치, 지리공간분석절차, 문장 임베딩

학 번 : 2021-26777

목 차

표 목 차

그 림 목 차

용 어 정 의

■ 지식베이스(knowledge base)

: 특정 분야 온톨로지를 기반으로 정보를 가공하여 지식으로 만들어 저장하는 데이터 셋의 일종이다. 지식베이스와 지식그래프(knowledge graph)를 같은 의미로 혼용해서 사용하는 경우가 많기 때문에 본 논문에서는 지식베이스로 용어를 통일해서 사용한다.

■ 문장 임베딩(sentence embedding)

: 문장 임베딩은 자연어처리(natural language processing) 분야에서 사용되는 용어로, 문장을 벡터공간(vector space)에 대응시켜 숫자의 조합인 벡터(vector)로 표현하는 것을 뜻한다.

■ 언어모델(language model)

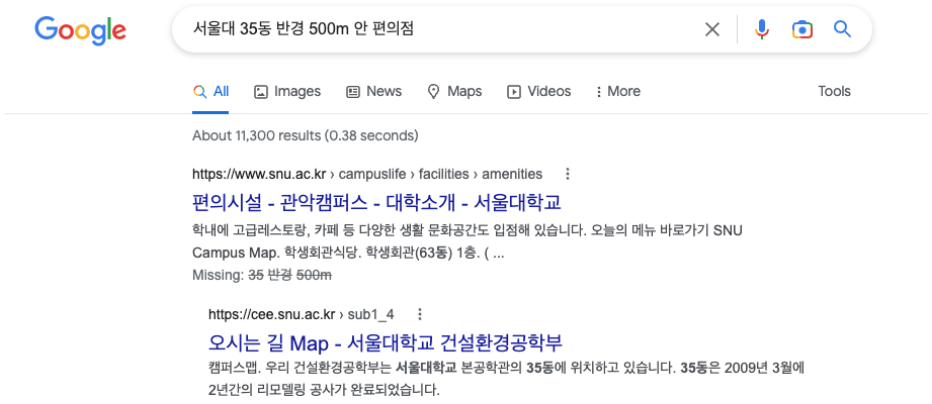
: 단어 시퀀스(sequences of words)에 확률을 부여하는 모델을 뜻한다. 단어 시퀀스에 확률을 부여함으로써 단어 시퀀스가 주어졌을 때 확률 기반으로 다음 단어 예측을 수행하거나, 문장에 빈 단어를 예측하는 등의 업무를 수행할 수 있다.

1. 서론

1.1 연구 배경 및 목적

기하급수적으로 늘어나는 문서로부터 사용자가 원하는 답을 간결하고 정확하게 도출할 수 있도록 만들기 위한 노력은 질의응답(question answering, 이하 QA)분야의 발전을 가져왔다(Hirschman & Gaizauskas, 2001; Hovy *et al.*, 2000). 한편, 지리공간정보(geographic information)의 양은 연간 20% 내외 증가하고 있고 웹상의 데이터 중 상당량을 차지함(Lee & Kang, 2015)에도 불구하고 QA 방법론을 사용하고 있는 구글(Google) [1]을 통해서도 지리공간질의(geographic question)을 답하지 못하는 경우가 존재한다. 그 이유는 구글이 작동하는 방식은 지리공간연산을 수행해 답을 도출하는 것이 아닌 다량의 웹 문서를 참조해 해당 질의에 대한 답과 가장 유사한 문서를 찾는 방식으로 작동하기 때문이다(Google, 2022). 가령, “서울대학교 35동 반경 500m 안에 있는 편의점 알려줘”라는 질의는 버퍼(buffer) 연산 등을 사용하면 답을 도출할 수 있지만 현재 구글을 통해 해당 질의에 대한 답을 도출할 수 없다([그림 1-1]).

[1] https://www.google.com/



[그림 1-1] Google을 통한 GeoQA 수행 한계

위와 같은 문제를 해결하기 위해 지리공간정보시스템(geographic information system, 이하 GIS) 분야에 특화한 QA를 연구할 필요성이 대두되었고 이를 지리공간 질의응답(geographic question answering, 이하 GeoQA)이라 한다. Hamzei, Winter, *et al.* (2022)은 GeoQA를 "지리공간 질문에 대한 답을 도출함으로써 질문자의 정보요구를 만족시킬 수 있도록 돕는 방법 또는 알고리즘"이라고 정의했다. 현재 보고되는 연구를 통해 GeoQA를 수행하기 위한 방법으로 지리공간 지식베이스 질의응답(Geographic Knowledge Base Question Answering, 이하 GeoKBQA)와 지리공간분석 질의응답(Geo-analytical Question Answering, 이하 Geo-analytical QA) 두 연구가 수행되는 것을 확인할 수 있다.

GeoKBQA는 Wikipedia [2] 등 일반적인 정보를 온톨로지(ontology)에 기반해 정제한 지식베이스(knowledge base, 이하 KB)에 지리공간정보를 혼합해 지리공간 지식베이스(geographic

[2] https://www.wikipedia.org/

knowledge base, 이하 GeoKB)를 구축하고, 자연어를 통해 QA를 수행하는 방법론이다. 이때 온톨로지란 정보를 특정 분야의 지식체계에 맞게 구조화 할 수 있도록 만드는 체계를 뜻한다(Guarino *et al.*, 2009). 일반적으로 GeoKB를 구축할 때 웹 표준을 제정하는 단체인 W3C(world wide web consortium)[3]에서 제안하는 RDF(Resource Description Framework)[4] 기반의 온톨로지를 채택하고 있다(Hamzei, Tomko, *et al.*, 2022; Punjani *et al.*, 2018; Stadler *et al.*, 2012). 구축된 GeoKB상에서 RDF 표준 쿼리 언어(query language)인 SPARQL[5] 또는 개방형 공간정보 컨소시엄(Open Geospatial Consortium, 이하 OGC)[6]이 제공하는 GeoSPARQL[7]을 이용해 일반적인 속성 조회 뿐 아니라 GeoSPARQL이 제공하는 15가지 벡터(vector) 기반 지리공간연산 및 지리공간 속성조회를 할 수 있다. 그러나 지리공간 질의를 답하기 위해서는 래스터(raster) 기반 연산을 사용하거나 GeoSPARQL이 지원하는 15가지 벡터 기반 연산 이외의 다른 연산을 필요로 하는 경우도 존재하기 때문에 이를 보완하는 연구의 필요성이 대두되었다.

GeoSPARQL에 새로운 기능을 추가해서 GeoSPARQL의 한계를 보완하려는 연구가 보고되고 있지만(Almobydeen *et al.*, 2022), 새로운 기능을 추가하기 위해서는 온톨로지를 새로 정의하고 기능을 개발해야

---

[3] https://www.w3.org/

[4] https://www.w3.org/RDF/

[5] https://www.w3.org/TR/rdf-sparql-query/

[6] https://www.ogc.org/

[7] https://www.ogc.org/standards/geosparql

하므로 기존 지리공간분석 도구를 사용하는 방법에 비해 비효율적인 방식이다.

Geo-analytical QA는 지리공간 개체(geographic entity), 지리공간 개념(geographic concept), 지리공간 관계(geographic entity)를 포함하는 질의인 지리공간질의(Mai *et al.*, 2021)를 지리공간분석절차로 변환하고 해당 분석 절차를 수행하기 적합한 지리공간분석 도구 및 데이터를 탐색하는 연구분야이다(Scheider *et al.*, 2021). GeoKBQA 연구에 비해 초기 단계의 연구가 보고되고 있고, 여기에 해당하는 연구는 개념적인 프레임워크(framework)를 제안하는 연구(Gao & Goodchild, 2013; Scheider *et al.*, 2021) 및 의문문 문장 형태에 대해 지리공간분석절차를 도출하는 연구(Xu *et al.*, 2022)가 있다.

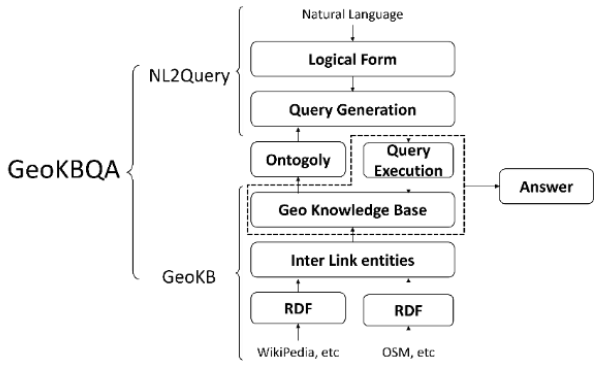
하지만 사람들이 실제로 질의하는 문장 형태 의문문뿐만 아니라 명령문, 문법에 맞지 않는 표현 등 다양하기 때문에 현실적인 Geo-analytical QA를 수행하기 위해서는 다양한 문장 형태의 질의를 지리공간분석절차로 변환할 수 있는 시스템이 필요하다. 또한 지리공간분석을 수행하기 위해서는 지리공간연산함수를 파악하는 것이 중요하기 때문에 도출한 지리공간분석절차가 지리공간연산함수를 분석의도에 맞게 순서대로 포함해야 한다. 따라서 본 연구에서는 다양한 문장 형태로 이루어진 지리공간질의를 분석절차로 변환할 수 있는 알고리즘을 개발하는 것을 목표로 하고, 해당 분석절차가 분석의도를 충족할 수 있는 지리공간연산함수를 순서에 맞게 포함하도록 설계하는 것을 목표로 한다.

1.2 관련 연구

1.2.1 GeoKBQA

GeoKBQA는 GeoQA의 한 분야로서 GeoKB를 구축하는 것과 자연어를 GeoKB 표준 쿼리 언어로 변환해 QA를 수행과정을 포함한다. 본 연구에서는 현재 진행되고 있는 GeoKBQA연구를 분석해 해당 분야의 한계점을 도출하고 이를 보완하는 연구의 필요성을 설명한다.

현재 보고되고 있는 다수의 연구(Hamzei, Tomko, *et al.*, 2022; Punjani *et al.*, 2018; Stadler *et al.*, 2012)를 통해 GeoKBQA를 수행하기 위해서는 RDF 형식으로 표현된 일반적인 지식을 다루는 KB에 지리공간정보를 RDF 형식으로 변환한 정보를 혼합하여 GeoKB를 구축하는 과정과 자연어로 구성된 질의를 RDF 표준언어로 변환하는 과정이 필요하다([그림 1-2]). 이때 사용되는 쿼리 언어로는 일반적인 정보를 조회하는 데 쓰이는 SPARQL과 15가지 벡터 기반 지리공간연산 및 지리공간 속성조회를 지원하는 GeoSPARQL이 존재한다.



[그림 1-2] 일반적인 GeoKBQA아키텍쳐

Stadler *et al.* (2012)은 오픈스트리트맵(OpenStreetMap, 이하 OSM)[8], GeoNames[9] 등 에서 취득한 지리공간정보를 RDF 형식으로 변환하고 Wikipedia 정보를 RDF 형식으로 변환한 DBPedia 등과 통합해서 GeoKB인 LinkedGeoData를 구축했다. 이때 사용된 OSM은 다양한 지리공간정보를 포함하고 있는 오픈 소스(open source) 방식의 무료지도 서비스의 일종이고, GeoNames는 100만 개 이상의 지명을 포함하는 오픈소스 데이터베이스다. 해당연구에서 정보의 크기가 과도하게 커지는 것을 방지하기 위해 OSM정보 중 일부를 필터링(filtering)해서 GeoKB를 구축했고, 자연어를 표준 쿼리 언어로 변환하는 방법에 대해서는 다루고 있지 않다. 비록 자연어를 표준 쿼리언어로 변환하는 방법을 제시하지 않았지만 해당연구에서 제시한 GeoKB 구축 방법론을 이후 연구(Hamzei, Tomko, *et al.*, 2022; Punjani *et al.*, 2018)에서 일반적으로 채택하고 있다.

Punjani *et al.* (2018)은 전 세계 행정구역 정보를 담은 GADM(global administrative areas)[10] 과 OSM 등에서 추출한 지리공간정보를 RDF 형식으로 변환하고 DBPedia와 통합하여 GeoKB인 LinkedGeospatialData를 구축하고 템플릿과 규칙 기반 방식을 이용해 자연어를 SPARQL 또는 GeoSPARQL로 변환해 GeoKBQA를 수행하는 방법을 제시했다. 해당 연구에서 제시한 방법을 통해 거리 관계(distance relations), 위상학적 관계(topological relations), 방위 관계(cardinal direction relations)을 묻는 질의를 답할

[8] https://www.openstreetmap.org/

[9] https://www.geonames.org/

[10] https://gadm.org/

수 있으나, aggregate, sorting과 같은 기능을 지원하지 않아 '몇 개'와 '가장 가까운'과 같은 질의를 답할 수 없는 한계를 지닌다.

Hamzei, Tomko, *et al.* (2022)은 YAGO2(yet another great ontology 2)[11]와 OSM, GADM 정보 및 영국 및 아일랜드 지역에 해당하는 100만 개의 지명 정보를 통합해 YAGO2Geo를 구축하고 템플릿을 이용해 자연어를 논리적인 구조로 변환해 GeoKBQA를 수행하는 방법을 제시했다. 여기서 쓰인 Yago2는 Wikipedia 정보를 RDF 형식으로 변환한 YAGO에 GeoNames 및 대규모 영어 어휘 데이터베이스인 WordNet 정보를 통합해서 Yago를 발전시킨 KB의 일종이다. 자연어를 논리적인 구조로 변환하는 방법을 통해 Punjani *et al.* (2018)의 연구에서 다루지 못한 aggregate, sorting 등에 해당하는 질의를 답할 수 있게 되었다. 위 연구를 표로 정리하면 [표 1-1]과 같다

[표 1-1] GeoKBQA 선행연구

연구자	데이터형식	GeoKB	NL2Query
Stadler *et al.* (2012)	RDF	LinkedGeoData	-
Punjani *et al.* (2018)	RDF	LinkedGeospatialData	템플릿
Hamzei, Tomko, *et al.* (2022)	RDF	Yago2Geo	Logical Form

위 에서 제시한 GeoKBQA 시스템을 통해 일반적인 정보에 조회뿐만 아니라 지리공간정보에 대한 벡터 연산 및 지리공간정보

[11] https://yago-knowledge.org/

속성조회를 할 수 있다. 이는 데이터를 RDF 형식으로 통합하고 RDF 표준 쿼리 언어인 SPARQL과 SPARQL을 지리공간정보 분야에 특화한 GeoSPARQL을 사용하기 때문에 가능한 결과로 볼 수 있다. 그러나 현재 GeoSPARQL을 통해 벡터 기반의 7가지 위상학적(topological) 연산과 8가지 비 위상학적(non-topological)연산을 수행할 수 있기 때문에([표 1-2]) 그 외의 연산을 수행해야 하는 질의를 답하지 못하는 한계를 지닌다. 따라서 이러한 GeoSPARQL에 기반한 GeoKBQA의 한계를 보완할 수 있는 연구의 필요성이 대두되었다.

[표 1-2] GeoSPARQL에서 지원하는 연산 목록

GeoSPARQL Functions	
Topological	Non-topological
EQUALS	DISTANCE
DISJOINT	BUFFER
INTERSECTS	CONVEXHULL
TOUCHES	INTERSECTION
WITHIN	DIFFERENCE
CONTAINS	SYMDIFFERENCE
OVERLAPS	ENVELOPE
	BOUNDARY

1.2.2 Geo-analytical QA

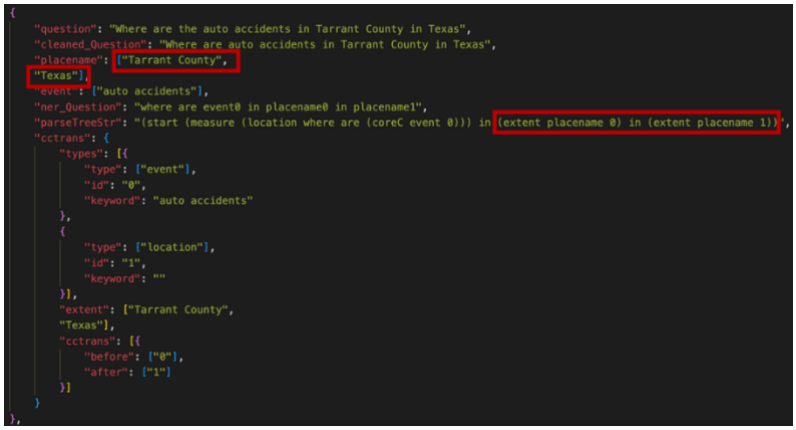
Geo-analytical QA는 지리공간질의를 지리공간분석절차로 변환하고 해당 분석절차를 수행하기 적합한 지리공간분석 도구 및 데이터를 탐색하는 연구이다(Scheider *et al.*, 2021). Geo-analytical QA 분야에서 핵심 연구 주제는 지리공간질의를 지리공간분석절차로 변환하는 것이다. 여기서 분석절차는 지리공간질의를 답하기 위해 필요한 지리공간 연산함수를 적합한 순서로 나열한 형태를 뜻한다.

Gao and Goodchild (2013)는 질의에 나타난 지리공간 개체의 데이터타입과 질의를 답하기 위해 필요한 지리공간연산함수를 도출하는 개념적인 프레임워크를 제안했다. 지리공간질의를 자연어처리(natural language processing) 기술을 이용해 질의에 등장하는 단어를 기반으로 데이터 타입과 연산함수를 도출하는 방식이다. 해당 프레임워크를 통해 연산함수를 도출할 수 있을 것으로 기대할 수 있으나 실제로 적용된 사례가 아직 보고되지 않았다.

Scheider *et al.* (2021)은 지리공간질의에 나타난 지리공간 개체를 폴리곤(polygon)과 같은 데이터 타입이 아닌 Kuhn (2012)이 제안한 공간정보핵심개념(core concept of spatial information)을 이용해 지리공간질의를 추상화하는 방안을 제시했다. 여기서 공간정보핵심개념은 학제 간 통용될 수 있는 공간정보개념으로서, 공간에 관한 개념인 location, neighbourhood, field, object, network, event와 정보에 관한 개념인 granularity, accuracy, meaning, value가 여기 해당한다(Kuhn, 2012). 공간정보핵심개념을 이용해 데이터 추상화(abstract data type)를 하는 것이 폴리곤과 같은 데이터 타입을 이용하는 것보다 GIS 분야의 현상을 이해하기 쉽기 때문에 이를 이용해

추상화할 것을 제안했다.

Xu *et al.* (2022)은 공간정보핵심개념을 이용해 GeoAnQu 말뭉치(corpus)에 대해 지리공간정보 개체를 추상화하고 질문어(WH-words)와 개체 사이에 나타나는 단어를 규칙 기반 방식을 이용해 지리공간질의를 분석절차로 변환하는 연구를 수행했다. 이때 말뭉치란 자연어 연구를 수행하기 위해 모아놓은 질의 또는 단어의 집합을 뜻한다. 지리공간정보 개체를 추상화하기 위해 이름을 가진 개체를 인식하는 개체명 인식(named entity recognition), 문장 구조를 이해하는 의존성 분석(dependency parsing) 등 자연어 처리 기법을 이용해 지리공간 개체를 탐지하고 공간정보핵심개념을 이용해 탐지한 개체를 추상화했다. 그 후 질문어와 추상화된 지리공간정보 개체 그리고 개체 사이에 등장하는 단어를 규칙 기반 방식을 통해 지리공간분석절차를 도출했다. 가령 "Where are the auto accidents in Tarrant County in Texas"라는 문장을 개체명 인식과 의존성 분석을 통해 'where are event0 in placename0 in placename1'로 변환하고 가장 처음 나타나는 단어인 'where'와 공간핵심개념 간 주술 관계를 규칙 기반 방식을 통해 질의를 분석절차로 변환한다. 해당 연구는 GeoAnQu 말뭉치를 대상으로 진행되었다. 그러나 해당 연구에서 제안한 방식은 규칙 기반으로 'which', 'what', 'where', 'how'와 같은 질문어와 추상화된 지리공간 개체 간 주술 관계에 의존하기 때문에 명령문 형태로 바꾸거나 미리 정의한 주술관계 이외의 질의를 입력하면 분석절차로 변환할 수 없는 한계를 확인했다. 가령 'Tarrant County in Texas'라는 지명을 'show'로 시작하는 질의에서는 탐지해내지 못하는 등의 문제가 여기에 해당한다([그림 1-3]).



(a)‘where’로 시작하는 질의에 대한 분석절차변환 성공

(b)‘show’로 시작하는 질의에 대한 분석절차변환 실패

[그림 1-3] 분석절차변환 성공 및 실패 예시

1.2.3 지리공간질의 말뭉치(Geographic question corpus)

지리공간질의란 지리공간 개체, 지리공간 개념, 지리공간 관계를 포함하는 질의를 뜻하고, 여기서 지리공간 개체란 서울, 대전과 같은 특정 지명을, 지리공간 개념이란 건물, 도시, 시군구 등과 같은 피쳐 유형(feature type)을, 지리공간 관계란 방위, '가까운', '사이에' 등을 뜻한다(Mai *et al.*, 2021). 따라서 지리공간질의 말뭉치란 지리공간 개체, 지리공간 개념, 지리공간 관계를 포함하는 질의를 모아 놓은 집합을 뜻한다.

MSMARCO 데이터 세트는 사용자들이 웹 검색엔진인 Bing[12] 에 실제로 검색한 약 100만 개의 질의(questions)와 해당 질의에 대한 답을 유추할 수 있는 구절(passages), 사람이 직접 질의와 구절을 이해하고 기재한 답(answers) 등으로 이루어진 데이터 세트이다. 해당 데이터 세트를 배포할 때 데이터 세트의 질의를 질문어 또는 질의 유형을 기준으로 분류한 결과를 함께 제시했다. 그 결과 질문어 기준 3.46%가 'where'를 포함하고 있고, 질의 유형 기준 6.17%가 위치(location)를 묻는 질의로 나타났다. 실제 사람들이 검색한 내용을 바탕으로 만들어진 데이터 세트이기 때문에 이를 사용해 현실적인 상황을 고려한 QA시스템을 구축할 수 있다(Bajaj *et al.*, 2016). MSMARCO 데이터 세트 중 'where'를 포함하는 질의와 위치를 묻는 질의를 지리공간질의 말뭉치로 분류할 수 있고 해당 질의 중 일부를 [표 1-3]을 통해 확인할 수 있다.

[12] https://www.bing.com/

[표 1-3] MSMARCO 질의 중 일부

weather in **washington dc** in march
homeplace restaurant in **catawba, va**
distance between warsaw in and **Indianapolis**
val thorens **where** is

Punjani *et al.* (2018)은 해당 연구에서 제안한 GeoKBQA 성능을 검증하기 위한 용도로 지리공간질의 말뭉치인 GeoQuestions201을 제작했다. 연구자가 판단하기에 간단한 질의를 인공지능 배경지식이 있는 학부생들에게 제시하고 이와 유사한 질의를 학생들이 제작하도록 요청한 후 제작한 질의를 연구자가 다시 정제하는 방식으로 말뭉치를 제작했다. 해당 말뭉치에 등장하는 질의는 5가지 유형 중 하나로 분류할 수 있고, 지리공간 개체 또는 개념의 위치를 묻는 유형, 지리공간 관계를 예/아니오로 답할 수 있는 유형, 지리공간 관계에 해당하는 지리공간 개체 또는 개념을 묻는 유형, 지리공간관계에 특정 조건을 부여한 유형, 수량 계산 또는 aggregate를 이용해 답할 수 있는 유형이 여기 해당하고 이들 유형이 명확히 분리되지는 않는다([표 1-4]).

[표 1-4] GeoQuestions201 질의유형별 예시

질의 유형	예시
개체 또는 개념의 위치를 묻는 유형	Where is Loch Goil located?
관계를 예/아니오로 답할 수 있는 유형	Is Liverpool east of Ireland?
관계에 해당하는 개체 또는 개념을 묻는 유형	Which counties border county Lincolnshire?
관계에 특정조건을 부여한 유형	Which mountains in Scotland have height more than 1000 meters?
수량 계산 또는 aggregate를 이용해야 하는 유형	Which is the largest county of England by population which borders Lincolnshire?

Xu *et al.* (2020)은 질의를 답하기 위해 지리공간분석절차를 필요로 하는 질의를 모아놓은 GeoAnQu 말뭉치를 제작하고 MSMARCO, GeoQuestions201 말뭉치와 비교해서 해당 말뭉치의 특징을 설명했다. GeoAnQu 말뭉치는 429개의 질의로 이루어져 있고, 해당 질의는 GIS 튜토리얼(tutorial) 교재 및 GIS 논문에서 발췌한 후 정제한 내용이다. 해당 연구에서 MSMARCO, GeoQuestions201, GeoAnQu 세 말뭉치를 단어, 구문, 문장 수준에서 분석했다 단어 수준에서 분석한 결과 GeoAnQu 말뭉치에 포함된 질의 중 45%의 질의를 답하기 위해서는 공간분석 또는 통계분석 방법론을 사용해야 하는 것으로 나타났다. 여기서 공간분석 또는 통계분석을 사용해 답을 도출해야 하는 개체에 해당하는 단어로는 distribution, pattern, change, accessibility 등이 있다. 따라서 해당 연구에서는 “What is the spatial distribution of probabilities of robberies in Salvador, Brazil?”과 같은 질의를 지리공간분석을 통해 답할 수 있는 질의로 분류했다. Xu *et al.* (2022)은 해당 말뭉치를 대상으로 Geo-analytical QA 연구를 수행했다.

GeoAnQu말뭉치 중 일부를 [표 1-5]에서 확인할 수 있다.

[표 1-5] GeoAnQu 말뭉치 중 일부

What is the **kernel density** of traffic accidents in Pasadena, California
Where are the **hot spots** of traffic accidents in Pasadena in California
What is the **driving time** from the windfarm company headquarters to the windfarm areas in Colorado
What is the **Euclidean distance** to the rivers in Crook, Deschutes, and Jefferson count

1.2.4 지리공간연산함수(geospatial operation) 분류체계

지리공간분석절차는 지리공간연산함수를 분석 의도에 맞게 나열된 형태로 이루어져야 하므로 현재 쓰이고 있는 지리공간연산함수를 체계적이고 광범위하게 포함하고 정확하게 설명하는 분류체계가 필요하다.

Albrecht (1998)은 사용자 친화적인 지리공간연산함수를 분류하는 연구를 수행하고 해당 연구의 결과로 7개 분류에 속하는 20개의 지리공간연산함수를 제시했다. 해당 20개의 지리공간연산함수를 도출하기 위해 국제 컨퍼런스(conference) 참여자를 대상으로 설문을 진행했고 그 방식은 144개 지리공간연산함수 중 피 설문자가 중요하다고 생각하는 함수에 대해 평점을 부여하고 이 평점을 종합하는 것이다. 위 연산함수를 이용하면 대부분의 연산을 수행할 수 있다고 주장했다.

Stefanakis and Sellis (1998)은 지리공간분석을 위해 쓰이는 지리공간연산함수를 5개의 대분류에 따라 분류했고, 데이터 전처리 및 변경, 데이터 분석, 데이터 시각화, 데이터 베이스관리가 위 5개 대분류에 해당한다. 데이터 전처리 및 변경은 데이터를 원하는 형태로 변환하는 기능, 데이터 분석은 공간 분석 및 속성 분석 기능, 데이터 시각화는 분석 결과를 시각화하는 기능, 데이터 베이스 관리는 원하는 데이터를 변경하고 저장하는 기능을 포함하고 있다.

Li and Stefanakis (2020)은 새로운 그리드 시스템에 적용하기 위한 지리공간연산함수 개발을 위해 Stefanakis and Sellis (1998)가 제안한 분류체계와 Meaden and Do Chi (1996)가 제안한 분류체계 그리고 OGC가 제안하는 필수 기능을 참고해 28개의 중분류와 5개의 대분류를

기준으로 지리공간연산함수 분류체계를 제안했다([표 1-6]). 해당분류체계는 데이터베이스, 데이터 전처리, 데이터 연산, 데이터 시각화, 지리공간분석 대분류 아래 광범위한 지리공간연산함수를 다루고 있어 Stefanakis and Sellis (1998)의 분류체계에서 다루지 못한 연산함수를 포함할 뿐만 아니라 각 연산함수에 대한 설명이 비교적 자세하다.

[표 1-6] Li and Stefanakis (2020)가 제안한 지리공간연산함수 분류

지리공간연산함수	
대분류	**중분류**
Database techniques	Data storage and retrieval
	Data editing
	Communication with other systems
Data pre-processing and manipulation	Data validation
	Data model conversion
	Geometric conversion
	Integration
	Generalization
	Classification
Data computation	Cloud computing
	Parallel processing
Data visualization	Theme maps
	Statistics and reports
	Application
Spatial analysis and data interpretation	Data queries
	Overlay analysis
	Buffer
	Geometry measurement
	Network analysis
	Image algebra
	Terrain data storage and representation
	Topography
	Hydrology
	Geostatistics
	Sampling
	Geocoding
	Predictive modeling
	Workflows and pipelines

1.2.5 시사점 및 소결론

선행연구 분석을 통해 GeoKBQA를 수행하기 위해서 지리공간정보 및 일반적인정보를 RDF 형식으로 변환해 GeoKB를 구축하는 과정과 자연어를 쿼리 언어로 변환하는 과정이 필요한 것을 확인할 수 있다. 이때 사용되는 쿼리 언어는 RDF 표준 언어인 SPARQL 또는 지리공간정보 속성조회 및 지리공간연산을 지원하는 GeoSPARQL이 존재한다. GeoSPARQL을 통해 15가지 벡터연산을 수행할 수 있지만 실제 분석을 위해 사용되는 연산은 15가지 벡터연산 이외의 벡터연산 및 래스터 연산을 수반하는 경우가 있기 때문에 기존 GeoKBQA를 보완하는 연구가 필요하다.

Geo-analytical QA는 지리공간질의를 지리공간분석절차로 변환하고 변환한 분석절차를 수행하기 적합한 도구 또는 데이터를 탐색하는 연구로 개념적인 프레임워크 제안(Gao & Goodchild, 2013; Scheider *et al.*, 2021) 및 질문형 문장 형태에 대해 규칙기반 방식을 이용해 질의를 분석절차로 변환하는 연구(Xu *et al.*, 2022)가 수행되었다. 질의를 분석절차로 변환하기 위해서는 말뭉치가 필요하며 Geo-analytical QA에서 사용하는 대표적인 말뭉치는 GeoAnQu가 존재한다.

사람들이 실제로 질의하는 문장 형태는 질문형뿐만 아니라 명령형 등 다양한 형태를 띨 뿐만 아니라 문법에서 벗어난 형태를 띠기도 하기 때문에 현실적인 Geo-analytical QA를 수행하기 위해서는 다양한 문장 형태에 대해 분석 절차를 도출할 수 있는 알고리즘이 필수적이다. 이를 규칙 기반 방식을 통해 달성하기 위해서는 모든 경우에 대해 규칙을 정의해야 때문에 비효율적 방식이며 기계학습(machine learning)은

규칙을 정의하지 않아도 데이터 패턴을 통해 답을 도출하는 알고리즘(El Naqa & Murphy, 2015)이기 때문에 다양한 문장형태를 가지는 질의에 대해 분석절차를 도출할 수 있어 효율적인 방식이다.

따라서 본 논문에서는 기계학습 방식을 이용해 다양한 문장 형태로 구성된 지리공간질의를 지리공간분석절차로 변환할 수 있는 알고리즘을 개발하는 것을 목표로 한다. 또한 지리공간분석을 실제로 수행하기 위해서는 지리공간연산함수를 분석 의도에 맞게 순서대로 실행하는 것이 중요하기 때문에 도출한 지리공간분석절차는 지리공간연산함수를 순서대로 포함하는 것을 목표로 한다. 지리공간연산함수를 분류하는 방식은 연구 관점에 따라 차이를 보이기 때문에 선행 연구 중 광범위한 연산함수를 체계적으로 다루는 분류체계에서 다루는 지리공간연산함수를 사용해 질의를 분석절차로 변환하는 것을 목표로 한다.

1.3 연구 범위 및 방법

본 연구는 다양한 문장 형태를 지니는 지리공간질의를 지리공간분석절차로 변환하는 것을 목표로 한다. 이를 수행하기 위해 두 가지 과정이 필요하다. 첫 번째는 지리공간질의 말뭉치를 선정하고 지리공간분석절차를 기준으로 라벨링 및 말뭉치에 대한 어휘변용을 수행해 데이터 세트를 확보하는 것이다. 두 번째는 기계학습 방식을 이용해 질의를 분석절차로 변환하는 것이다.

본 연구에서는 GeoAnQu 말뭉치를 연구 대상 말뭉치로 선정했고 그 이유는 해당 말뭉치에 등장하는 질의를 답하기 위해서 다양한 분석을 수행해야 하는 것으로 알려져 있기 때문이다. 해당 말뭉치를 라벨링하기 위해서는 라벨링을 위한 지리공간분석절차가 필요하며 분석 절차는 다수의 지리공간연산함수를 분석 의도에 맞게 포함하고 있어야 한다. Li and Stefanakis (2020)이 제안한 지리공간연산함수 분류체계는 비교적 광범위한 연산함수를 체계적으로 다루고 있기 때문에 해당 연산함수를 기준으로 선정했다. 즉, GeoAnQu의 각 질의를 답하기 위해 Li and Stefanakis (2020)의 지리공간연산함수를 각 질의 분석 의도에 맞게 나열하는 방식을 이용해 각 질의를 분석절차로 변환했다. 이때 중복되는 분석 절차도 존재하기 때문에 이들을 종합해 고유한 분석절차에 대해 고유 클래스 번호를 지정해 다시 각 질의를 라벨링 하는 방식으로 데이터 세트를 구축했다. 그 후 어휘변용(paraphrase)을 실시해 다양한 문장 형태의 질의를 생성했다. 이를 통해 데이터 세트를 증강할 수 있었고, 알고리즘이 실제로 다양한 문장 형태에 대해서 작동하는지 검증할 수 있었다. 이때 사람이 직접 어휘변용을 수행하는 작업은 시간이 많이 소요되기 때문에 Damodaran (2021)이 제안한 어휘변용

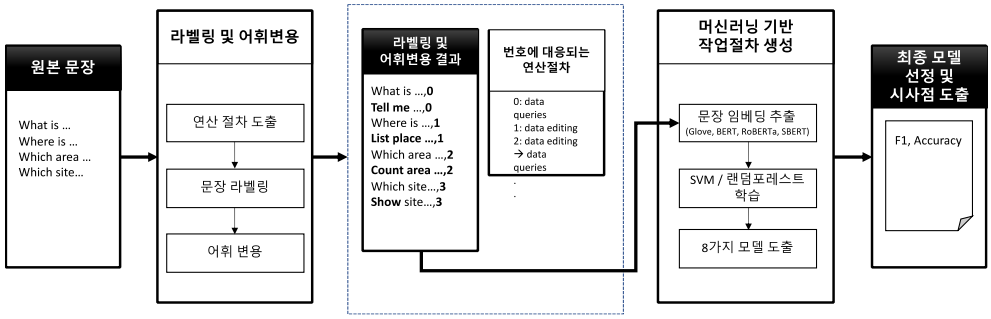
모델을 사용했다.

기계학습 방식 중 문장분류 방식을 이용해 질의를 분석절차로 변환하는 방식을 채택했다. 그 이유는 문장분류 방식을 사용하면 적은 데이터로도 높은 성능을 기대할 수 있기 때문이다. 말뭉치에 포함된 질의를 분류모델의 입력값으로 사용하기 위해서는 질의를 벡터공간(vector space)에 대응시켜 숫자의 조합인 벡터로 바꾸는 과정이 필요하고 이 과정을 문장 임베딩(sentence embedding)이라고 한다. 문장 임베딩을 수행하기 위해서는 문장의 의미를 반영할 수 있는 언어 모델을 사용할 필요가 있고, 분포가설을 따르는 모델이 이에 해당한다. 분포가설에 의하면 유사한 문맥(context)에서 나타나는 단어는 유사한 의미를 가지며 이 가설에 기반한 모델로는 Word2vec(Mikolov, Chen, *et al.*, 2013; Mikolov, Sutskever, *et al.*, 2013), Glove(global vectors for word representation)(Pennington *et al.*, 2014), BERT(bidirectional encoder representations from transformers)(Devlin *et al.*, 2018), RoBERTa(robustly optimized BERT pre-training approach)(Liu *et al.*, 2019), SBERT(sentence-BERT)(Reimers & Gurevych, 2019) 등이 존재한다. 본 논문에서는 Word2vec을 개선한 모델인 Glove, 문맥에 따라 단어 또는 문장에 대한 임베딩 값이 변하는 방식을 채택하는 BERT, BERT를 개선한 모델인 RoBERTa, BERT를 문장 임베딩에 특화한 SBERT를 이용해 총 4가지 문장 임베딩을 수행해 데이터 셋을 분류 모델의 입력값으로 전환할 수 있었다.

분류모델로는 초평면(hyperplane)을 이용해 이진 분류(binary classification)를 수행하는 SVM(support vector machine)(Noble, 2006) 중 가장 간단한 커널 트릭(kernel trick)인 linear 방식과 여러

개의 결정트리(decision tree)를 생성하고 각 결정트리의 결과를 종합하는 방식으로 결정트리를 개선한 모델인 랜덤포레스트(random forest)를 이용했다. SVM은 기본적으로 이진 분류를 수행하는 방식이기 때문에 K개의 클래스에 대해 K개의 SVM을 생성하고, 하나의 클래스 k의 라벨을 1, 나머지 클래스를 0으로 가정하는 one-against-all 방식(Franc & Hlavac, 2002)을 이용해 분류를 수행했다.

최종적으로 4가지 문장 임베딩과 2가지 분류모델을 통해 8가지 결과를 도출할 수 있었고, 이들 결과를 비교분석해서 가장 높은 성능을 보이는 조합을 도출할 수 있었고, 해당 결과를 분석해 시사점을 도출했다. 위 연구 흐름을 그림으로 살펴보면 [그림 1-4]과 같다.



[그림 1-4] 연구 흐름도

2. 연구 방법

2.1 데이터 세트 구축

2.1.1 지리공간질의 말뭉치 선정 및 지리공간분석절차도출

본 연구를 수행하기 위해 GeoAnQu 말뭉치를 연구대상 지리공간질의 말뭉치로 선정했다. 그 이유는 Xu *et al.* (2020)의 연구에서 GeoAnQu 말뭉치에 등장하는 질의 중 45%가 분석을 통해 답할 수 있는 개체를 포함하는 것으로 나타났기 때문에 질의를 답하기 위해 다양한 분석절차를 필요로 하는 말뭉치로 볼 수 있기 때문이다. 말뭉치는 단순히 질의로만 이루어져 있기 때문에 이들을 분석해 고유한 지리공간분석절차를 도출하고 도출한 지리공간분석절차를 이용해 라벨링을 하는 과정이 필요하다. 본 연구에서는 Li and Stefanakis (2020)가 제안한 지리공간연산함수를 이용해 해당 말뭉치에 대한 지리공간분석절차를 도출했다. 즉, 말뭉치에 나타난 각 질의를 확인하고 이를 답하기 위해 필요한 지리공간연산함수를 분석 의도에 맞게 나열해 분석절차를 도출하는 방식이다. 연산함수는 데이터 타입에 의존적이기 때문에 OSM 온톨로지를 참고해서 분석절차를 도출했다. 가령, ‘What is house for sale in Utrecht?’ 라는 질의에 대해 ‘house’에 대한 태그(tag)인 ‘building=house’는 존재하고 ‘for sale’이라는 태그는 존재하지 않기 때문에, ‘for sale’이라는 속성을 부여하고(data editing) 해당 태그에(building=house) 해당하는 폴리곤 데이터를 OSM에서 가져와 Utrecht의 경계로 clip하는 분석절차를 도출할 수

있다. 또한, 지리공간질의를 답하기 위해 산술연산(arithmetic operation)을 해야 하는 경우가 있지만 Li and Stefanakis (2020)의 지리공간연산함수 분류체계는 해당 연산을 포함하고 있지 않기 때문에 해당 분류체계를 이해한 후 산술연산을 분류체계 중 적합한 하위항목에 대응시키는 과정이 필요하고 이를 반영하였다.

2.1.2 말뭉치 라벨링

말뭉치를 라벨링 할 때 지리공간분석절차를 기준으로 라벨링 하는 과정이 필요하며, 분류모델을 사용하기 위해서는 라벨이 숫자로 표현되어야 하기 때문에 각각 고유한 지리공간분석절차를 숫자로 대응시키는 과정이 필요하다. 따라서 GeoAnQu 말뭉치에 등장하는 각 질의를 분석해 도출한 지리공간분석절차를 지리공간연산함수의 개수를 기준으로 오름차순으로 정렬해 고유 번호를 부여한 후 이 번호를 이용해 각 질의를 라벨링 하는 방식을 사용했다.

2.1.3 어휘 변용

어휘 변용은 자연어처리 분야에서 데이터 증강을 위해 쓰이는 방법 중 일종이며, 어휘 변용을 수행하는 방법으로는 문장에 등장하는 어휘를 동의어로 바꾸는 방법, 단순히 어순을 무작위로 바꾸는 방법, 언어모델을 이용해 원 문장의 의미를 유지하는 범위 내에서 어순을 바꾸는 방법이 존재한다(Bayer *et al.*, 2022; Shorten *et al.*, 2021; Wei & Zou, 2019). 본 논문에서는 Damodaran (2021)이 제안한 언어모델(language model)을 사용해 어휘 변용을 실시했다. 해당 언어모델은 웹상의 20TB 크기의 말뭉치로 학습된 T5(text-to-text transfer transformer)모델에 기반하고 있고, 원 문장과 의미 유사도, 생성된 문장의 다양성 및 어휘 유창성을 고려 어휘 변용을 수행하는 모델이다. 따라서 해당 모델을 사용하면 원 문장의 의미를 유지하는 선에서 특정 단어를 동의어로 바꾸거나 어순을 바꿔 어휘 변용을 수행할 수 있다.

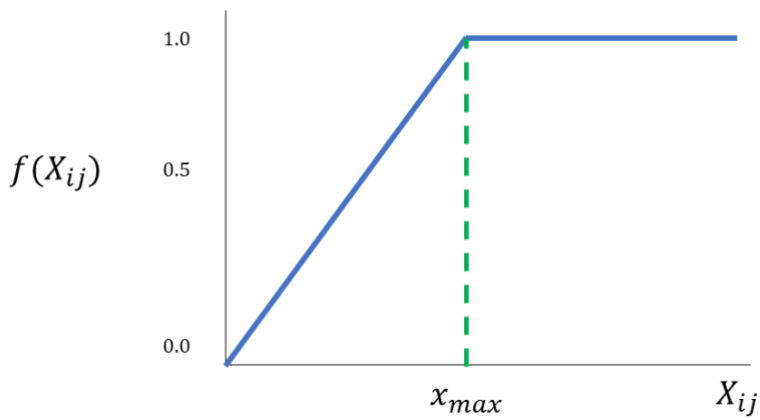
2.2 문장 임베딩(sentence embedding) 언어모델

문장 임베딩은 문장을 벡터공간에 대응시켜 숫자의 조합인 벡터로 표현하는 것을 뜻한다. 분포가설은 의미가 유사한 문장 또는 단어는 유사한 문맥(context)내에서 등장한다는 가정(Firth, 1957; Harris, 1954)이며 해당 가설을 채택해서 만든 언어모델을 사용하면 언어의 의미론(semantic) 또는 구문론(syntactic)적 의미를 표상(represent)하는 임베딩을 수행할 수 있다. 해당 가설을 채택하는 모델은 Word2vec 이후에 등장하는 모델로 볼 수 있고, 본 논문에서는 Word2vec을 발전시킨 Glove, Glove의 단점인 정적인 임베딩(static embedding)을 발전시켜 문맥에 따른 임베딩을 수행하는 방식인 contextualized representation을 사용하는 BERT, BERT를 더 오래 학습시키는 방식 등으로 개선한 RoBERTa, 문장 유사도 측정을 이용해서 BERT를 문장 임베딩에 특화한 SBERT를 문장 임베딩 모델로 선정했다.

2.2.1 Glove

Pennington *et al.* (2014)은 Word2vec의 비효율적인 학습방식을 해결하기 위해 Glove를 제안했다. Word2vec의 학습 방식이 비효율적인 이유는 'the'와 같은 관사가 자주 등장할 때, 중요도와 상관없이 빈도수에 비례해 학습을 진행하기 때문이다. 이와 같은 문제를 해결하기 위해 가중치 함수(weighting function)를 도입했다. 가령 'the'라는 단어가 말뭉치 안에 특정 횟수 이상 등장한다면 100회까지 빈도만 손실함수(loss function)에 반영시켜 효율적인 학습을 가능하게 만든 모델이다. X_{ij} 를 단어 간 동시 발생 빈도라고 할 때, 식 (2-1)에서 $f(X_{ij})$는 가중치 함수 역할을 한다. 이 가중치 함수를 그림으로 나타내면 [그림 2-1]와 같다. 해당 식에서 w는 단어 임베딩, b는 상수를 뜻한다.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2 \qquad (2\text{-}1)$$



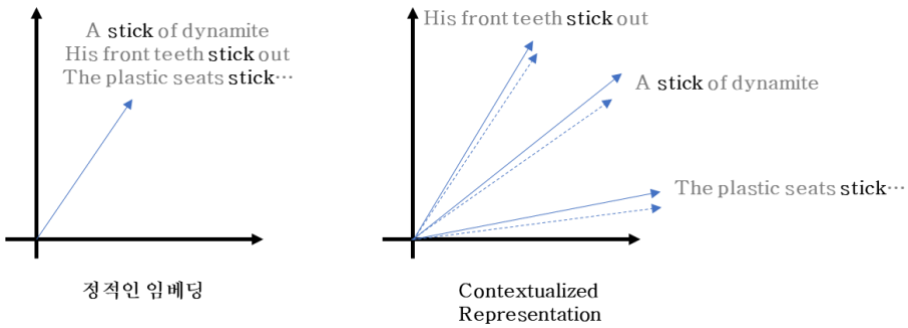
[그림 2-1] Glove 가중치 함수

또한 Word2vec에서는 n개의 연속적인 단어 나열인 n-gram을 문맥(context)으로 가정하고 한 단어와 다른 단어 간의 관계를 학습시켜 모델을 생성하지만, Glove에서는 전체 문서를 문맥으로 가정하고 학습시켰다. 위 방법을 통해 학습시킨 Glove를 이용해 단어 유사도(word similarity), 개체명 인식 태스크(task)에서 Word2vec보다 높은 성능을 보였다.

Glove는 각 단어를 임베딩하는 방식이기 때문에 해당 모델을 사용해 문장 임베딩을 수행하기 위해서는 각 단어를 임베딩 시킨 후 단어 임베딩 값을 평균 내는 방식을 사용한다.

2.2.2 BERT

Word2vec과 Glove는 분포가설을 따르는 모델이지만 정적인 임베딩을 수행하는 모델로, 정적인 임베딩이란 특정 단어가 문맥에 상관없이 같은 벡터값으로 대응되는 것을 뜻한다(Ethayarajh, 2019). 가령 단어 'stick'은 문맥에 따라 '막대기'를 의미하거나 '붙이다'를 의미하게 된다. 따라서 문맥에 따라 'stick'의 임베딩 값은 달라져야 하지만 정적인 임베딩을 수행하는 모델은 'stick'에 대해 항상 같은 벡터값으로 표현하게 된다. Contextualized representation 방식은 주어진 문장에 따라 임베딩을 달리해 정적인 임베딩을 개선한 방식(Ethayarajh, 2019)으로 BERT는 해당 방식을 따르는 언어모델이다. 정적인 임베딩 방식과 contextualized representation 방식을 이용해 'stick'을 벡터 공간상 대응시킨 모습을 개념적으로 표현하면 [그림 2-2]과 같다.



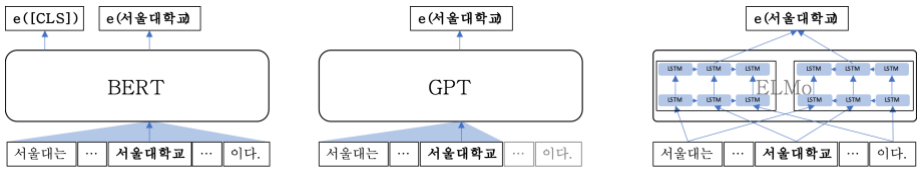
[그림 2-2] 정적인 임베딩 및 contextualized representation

Devlin *et al.* (2018)은 ELMo(embeddings from language model)(Peters *et al.*, 2018) 와 OpenAI GPT(generative pre-training

transformer)(Radford *et al.*, 2018) 모델의 사전학습(pre-training) 방식을 개선한 언어모델인 BERT를 제안했고 여기서 사전학습이란 다량의 말뭉치에 라벨링을 수행하지 않은 채 언어모델을 학습시켜 문장 또는 단어 패턴을 이용해 모델을 학습시키는 방식을 뜻한다. ELMo는 단어의 왼쪽과 오른쪽으로부터 시작하는 시퀀스를 이용해 단어 임베딩을 수행하는 bi-LSTM(long short-term memory)에 기반한 모델로 문맥에 민감한(context-sensitive) 임베딩을 수행할 수 있지만, 각 방향 LSTM의 결괏값인 잠재 상태(hidden state)를 단순히 concatenate해서 각 단어에 대한 임베딩을 수행한다. 또한 LSTM은 직전 시퀀스를 참조해 현재 시퀀스의 결괏값을 계산하는 방식이기 때문에 시퀀스가 길어지면 문맥상 단어의 중요도에도 불구하고 멀리 위치한 단어를 반영하지 못하는 한계가 있다.

예를 들어, '**서울대**는 국립대학법인 **서울대학교**다' 라는 문장과 '**서울대**는 1948년 명칭이 서울대학교로 바뀌었고, 국립대학법인으로 전환(2011)되어 현재 국립대학법인 **서울대학교**가 정식명칭이다' 라는 문장에서 '서울대' 와 '서울대학교' 는 서로 연관성이 있지만 LSTM을 이용하면 전자의 문장에 비해 후자의 문장에서 연관성을 도출하기 어렵다. 반면 Transformer 모델은 거리와 상관없이 단어간 상관성을 계산하는 방식인 attention 기법을 이용해 위와 같이 단어와 단어 사이의 거리가 멀어졌을 때 서로간 연관성을 반영하지 못하는 문제를 개선했다. OpenAI GPT는 Transformer모델 중 한 종류로 attention 기법을 사용할 때 한 단어가 등장할 때 이전에 등장한 시퀀스만 참조하는 방식이다. 앞의 예시 문장인 '**서울대**는 1948년 명칭이 서울대학교로 바뀌었고, 국립대학법인으로 전환(2011)되어 현재 국립대학법인 **서울대학교**가 정식명칭이다' 에서 '서울대학교' 에 대해

학습을 진행할 때 GPT는 그 이전까지 나온 단어를 attention 기법을 이용해 참조해 사전학습을 수행하는 방식이다. BERT는 attention 기법을 양방향에 적용한 방법으로 각 단어에 대한 임베딩을 *e(단어)*라고 표현한다면 [그림 2-3]를 통해 BERT, GPT, ELMo의 차이점을 볼 수 있다.



[그림 2-3] BERT, GPT, ELMo 아키텍쳐

Devlin *et al.* (2018)이 제안한 방식은 문장에 등장하는 단어 중 일부를 가리고 양쪽 시퀀스를 반영해 해당 단어를 예측하는 방식으로 이 방식을 MLM(masked language model)이라 한다. 해당 방식을 적용하기 위해 다음 문장을 예측하는 용도인 [CLS] 토큰, 단어를 예측하는 용도인 [MASK] 토큰, 문장과 문장 사이를 표시하는 [SEP] 토큰을 사용하게 된다. 가령 'The woman arrived at lab. She turned on the light'라는 문장 중 'The woman arrived at lab'에 다음 나올 문장을 예측할 때 [CLS] 토큰을 사용하고 중간 단어를 [MASK]로 가려 각 문장과 단어에 대한 임베딩을 학습시킬 수 있다([표 2-1]).

[표 2-1] BERT에서 다음문장 예측 테스크 예시

Input	[CLS] the woman [MASK] at lab [SEP] She turned on the [MASK] [SEP]
Label	IsNext
Input	[CLS] the woman [MASK] at lab [SEP] seoul is [MASK] of korea [SEP]
Label	NotNext

BERT를 이용해 문장을 임베딩 시킬 때 [CLS] 토큰의 임베딩을 이용해 문장 임베딩을 수행할 수 있고, 본 연구에서는 BERT의 [CLS] 토큰을 이용해 문장 임베딩을 수행한다.

2.2.3 RoBERTa

Liu *et al.* (2019)은 언어모델을 구성할 때 어떤 하이퍼 파라미터(hyper parameter) 또는 모델의 구성요소를 바꿀 때 모델 성능이 개선되는지 알아보기 위해 BERT를 이용해 실험을 진행하고 이를 개선한 모델인 RoBERTa를 제안했다. BERT를 개선하기 위해 더 많은 데이터와 시간을 사용한 학습하는 방법, 한번 학습 시 사용하는 데이터 크기인 배치 크기(batch size) 증가시키는 방법, 더 긴 문장에 대해 학습시키는 방법, 문장 중 일부를 가리는 방식인 마스크 패턴 다양화하는 방법을 각각 테스트하고 종합하여 BERT를 개선한 모델을 제안했다. 예를 들어, BERT를 학습시키기 위해서 16GB의 데이터가 사용되었다면 RoBERTa를 학습시키기 위해서, BERT에 쓰인 16GB의 데이터를 포함해 뉴스데이터 76GB, 인터넷 커뮤니티인 Reddit 글을 이용한 38GB 데이터, 기타 31GB를 통해 약 145GB 데이터를 추가 확보해 학습시켰다. 또한, BERT를 사전학습 시킬 때 한번 정한 [MASK] 토큰을 학습 시 바꾸지 않았다면 RoBERTa에서는 같은 학습데이터에 대해 10가지 다른 위치에 [MASK] 토큰을 정해 학습을 진행했다. 그리고 배치 크기를 BERT에서 사용한 256에서 8K(8000)로 확장했다. 위 방식을 적용한 RoBERTa를 이용해 질의응답, 문장의 긍정 부정을 분류하는 감성분석(sentiment analysis), 문장에 포함된 가정을 추론하는 텍스트 함의(textual entailment) 등을 종합 테스트하는 GLUE(general language understanding evaluation)(Wang *et al.*, 2018)에 포함된 9가지 테스크 중 4가지 테스크에서 가장 좋은 성능(state-of-art)을 보였다. 본 연구에서는 RoBERTa의 [CLS] 토큰을 이용해 문장 임베딩을 수행한다.

2.2.4 SBERT

Reimers and Gurevych (2019)은 BERT를 개선해 의미론적인 내용을 반영해 문장임베딩을 수행할 수 있는 언어모델인 SBERT를 제안했다. BERT를 이용해 1만 개 문장에 대해 의미론적으로 유사한 문장 쌍을 찾기 위해서는 (10000)(10000−1)/2 회의 연산을 수행해야 하고, 해당 연산을 딥 러닝(deep learning)용 그래픽 카드인 V100을 이용해 수행할 때 약 65시간이 걸린다. 그러나 SBERT를 이용하면 같은 테스크를 5초 이내에 수행할 수 있다. 이는 하나의 데이터를 기준으로 다른 데이터와의 유사도를 측정하는 기법인 siamese and triplet networks(Schroff *et al.*, 2015)를 적용했기 때문이다. 해당 기법은 기준 문장(anchor)의 임베딩 s_a 와 해당 문장과 같은 라벨(positive)을 가지는 문장의 임베딩 s_p , 다른 라벨(negative)을 가지는 문장의 임베딩 s_n 에 대해 s_a 와 s_p 는 유사한 의미를 지녔기 때문에 임베딩간 거리가 가깝고 s_a 와 s_n 은 다른 의미를 가지고 있어 임베딩간 거리가 멀다는 점을 이용해 손실함수를 계산하는 방식이다. 해당 손실함수를 식 (2−2)와 같이 표현할 수 있다.

$$\max\left(||s_a - s_p|| - ||s_a - s_n|| + \alpha, 0\right) \tag{2-2}$$

여기서 α 는 마진(margin)값으로 $||s_a - s_p||$ 와 $||s_a - s_n||$ 의 차이에 더해 손실함수를 크게 만드는 데 사용되고, 이를 통해 학습을 효율적으로 진행 시켰다.

미세조정(fine−tune)은 다량의 데이터에 대해 사전학습 된 모델을 용도에 맞게 학습시키는 것을 뜻하며, SBERT는 사전학습 된 모델인

BERT를 미세조정한 모델이다. 이때 사용한 데이터는 57만 개의 문장쌍에 대해 서로 함의 관계를 유사(entailment), 다름(contradiction), 중립(neutral)로 라벨링 한 SNLI(standford natural language inferencing)[13] 데이터 세트와 다양한 분야에서 수집한 문장쌍에 대해 서로 함의 관계를 라벨링한 MultiNLI(multi-genre natural language inference)[14] 데이터 세트다.

해당 모델은 문장 임베딩에 특화된 모델이기 때문에 BERT 또는 RoBERTa처럼 [CLS] 토큰을 이용하는 것이 아닌 모델을 통해 바로 문장에 대한 임베딩을 수행할 수 있다.

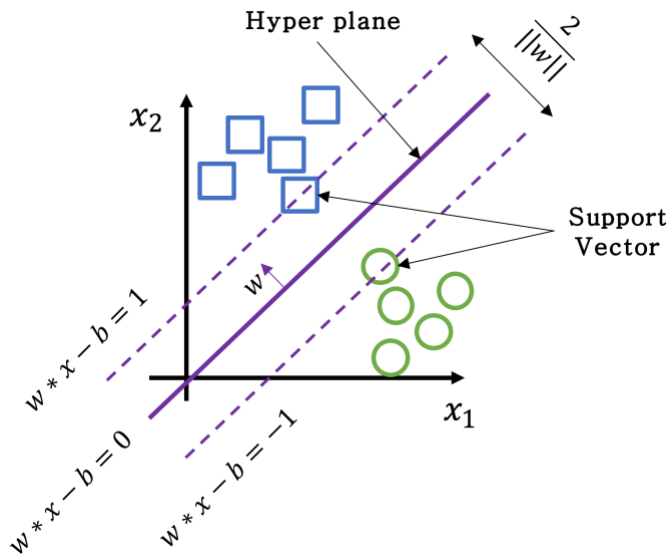
[13] https://nlp.stanford.edu/projects/snli/

[14] https://cims.nyu.edu/~sbowman/multinli/

2.3 분류모델학습

2.3.1 SVM

SVM은 주어진 객체의 라벨을 분류하는 컴퓨터 알고리즘으로, 금융 사기 탐지, 손 글씨 인식 등 다양한 분야에서 사용되고 있다(Noble, 2006). 주어진 객체의 라벨을 분류하기 위해 벡터 공간상에 초평면을 생성해 분류를 수행하며 초평면의 법선 벡터 방향으로 일정 거리 이상 위치한 데이터를 참값으로 반대방향으로 일정거리 이상 위치한 데이터를 거짓 값으로 분류하게 된다([그림 2-4]).



[그림 2-4] SVM 개념도

이때, 초평면과 가장 가까운 데이터와의 거리를 margin이라고 하고, SVM 학습 시 margin을 최대화하는 방향으로 학습이 진행된다.

전체 데이터를 M이라고 할 때, 초평면은 식 (2-3)으로 정의할 수 있고, 각 데이터는 식 (2-4)를 만족해야 한다(Widodo & Yang, 2007).

$$f(x) = w^T x + b = \sum_{j=1}^{M} w_j x_j + b = 0 \tag{2-3}$$

$$y_i f(x_i) = y_i(w^T x_i + b) \geq 1 \text{ for } i = 1,2,\ldots,.M \tag{2-4}$$

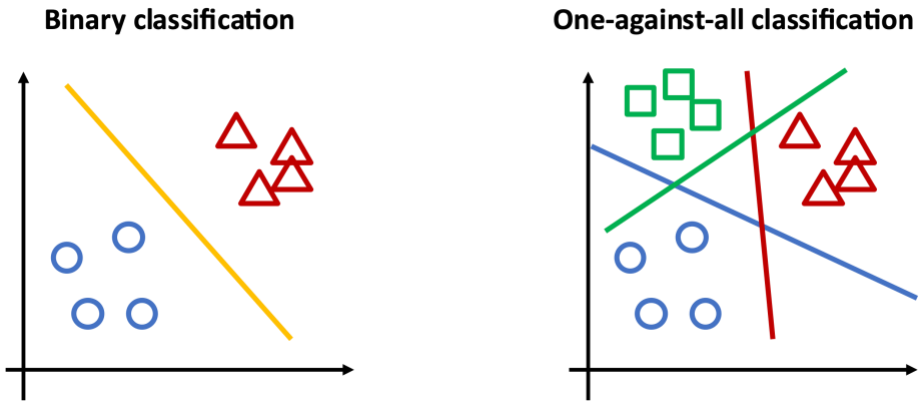
SVM에서 소프트 마진(soft margin)은 참값과 거짓 값을 완벽히 분류하는 초평면이 존재하지 않을 때, 오분류를 허용하며 초평면을 정의하는 방식으로(Cortes & Vapnik, 1995) margin으로부터 오분류된 데이터까지 거리를 뜻하는 느슨한 변수(slack variable) ξ_i 을 도입해 오분류를 허용하는 초평면을 정의 할 수 있다. ξ_i을 도입한 손실함수를 식 (2-5)로 정의할 수 있다.

$$minimize\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{M} \xi_i \tag{2-5}$$

끝으로 SVM을 통해 분류를 수행할 때 각 데이터를 다른 차원으로 대응하는 커널 트릭을 사용할 수 있다. 커널 트릭은 데이터를 다른 차원으로 대응하는 방식으로 이를 이용해 비선형 분류를 수행할 수 있다.

본 논문에서는 SVM 중 가장 간단한 커널 트릭 방식인 linear 방식을 채택해 랜덤 포레스트와의 결과 비교를 위해 사용했다. 또한 SVM은 기본적으로 이진 분류를 수행하기 때문에 K개 클래스에 대해 K개의 SVM을 생성하며, 하나의 클래스를 k라 하고 해당 클래스 라벨을

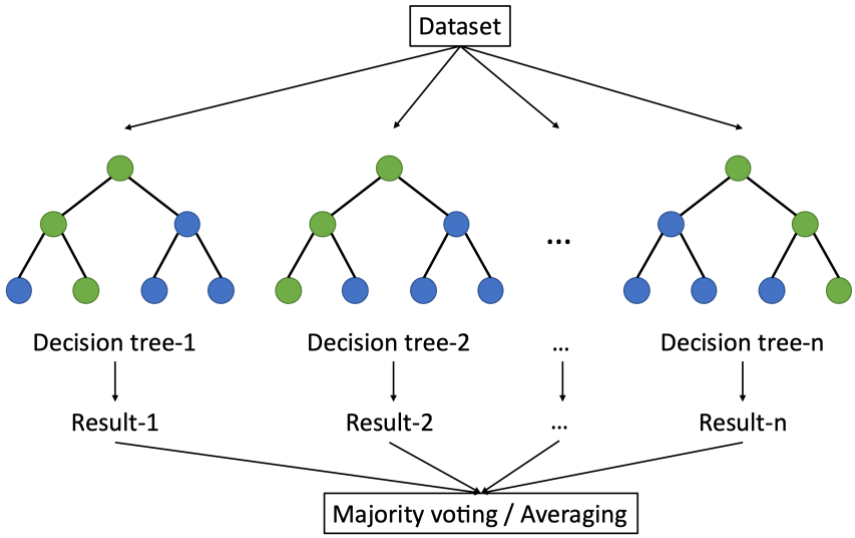
1이라 하면 나머지 클래스 라벨을 0으로 가정하고 SVM을 생성하는 one-against-all 방식(Franc & Hlavac, 2002)을 채택했다([그림 2-5]).



[그림 2-5] 이진분류 및 one-against-all 방식 SVM

2.3.2 랜덤포레스트

랜덤 포레스트 알고리즘은 무작위 의사결정 트리를 생성한후 각 의사결정 트리에서 예측한 값의 평균을 이용해 분류를 수행하는 앙상블 기계학습 방법의 일종이고 (Biau & Scornet, 2016) 개념도는 [그림 2-6]과 같다.



[그림 2-6] 랜덤포레스트 개념도

랜덤 포레스트 알고리즘은 의사결정나무 모델의 단점인 과적합(overfit) 문제를 해결할 수 있는 모델로 다음 과정을 통해 의사결정 나무를 만들고 학습시킨다. 학습데이터의 중복을 허용하며 샘플링하는 방식인 부트스트래핑(bootstrapping)방식을 이용해 학습데이터와 동일한 데이터 크기를 가지는 하위 샘플(subsample)을 생성하고 각 하위 샘플의 특징을 임의로 선정해 의사결정 나무를 생성하고 학습시킨다. 그 후 각 의사결정 나무의 결과를

종합(aggregate)해서 모델을 생성하게 된다(Livingston, 2005). 만약 한 개의 의사결정나무만 학습하게 된다면 노이즈에 민감하게 되지만, 여러 의사결정나무 결과를 종합한 방법을 사용해 강건한 모델을 구축할 수 있다.

2.4 평가방법

2.4.1 기존연구의 알고리즘과 비교

본 연구의 알고리즘이 기존 연구인 Xu *et al.* (2022)의 알고리즘에 비해 개선된 점을 확인하기 위해 같은 테스트셋에 대해 기존 알고리즘과 비교를 수행한다. 해당 알고리즘은 개체명 인식, 의존성 분석 등을 통해 질의를 분석절차로 변환하기 때문에 이 과정 중 오류가 발생하면 분석절차 변환 실패로 가정하고 성능을 비교했고, 후술할 정확도(accuracy)를 기준으로 평가했다.

2.4.2 평가지표

본 연구에서는 4가지 임베딩과 2가지 분류모델을 사용하기 때문에 총 8개의 결과가 생성되게 된다. 이들의 성능을 비교하기 위해서는 평가지표가 필요하다. 대표적인 평가지표는 정밀도(precision), 재현율(recall), 정확도(accuracy), F1-score가 있다. 다중 클래스 분류(multi class classification) 시 사용되는 F1-score는 micro, macro, weighted average가 존재하고 이 중 weighted average F1-score는 각 클래스에 대해 F1-score를 계산한 후, 전체 데이터 수 중 각 클래스에 속하는 데이터 수(support)의 비율을 곱해 더한 값으로 데이터 수가 불균형할 때 사용된다. [표 2-2]를 예로 들어 설명하면, 해당 표는 class b를 기준으로 TP(true positive), TN(true negative), FN(false negative)을 표시한 것이다.

[표 2-2] weighted average F1-score 계산 예시

		Predicted class				
Actual class	Classes	a	b	c	d	Support
	a	TN	FP	TN	TN	n_a
	b	FN	TP	FN	FN	n_b
	c	TN	FP	TN	TN	n_c
	d	TN	FP	TN	TN	n_d

이를 식 (2-7)을 통해 정밀도, 식 (2-8)을 통해 재현율을 구한 후 이 두 값을 식 (2-9)에 대입해 F1-score를 구할 수 있다. 해당 과정을 모든 클래스 k에 대해 반복하게 된다. 클래스 k에 대한 F1-score를 $F1_k$라 하고, 각 $F1_k$에 대한 비중을 $w_k = \frac{n_k}{total\ number\ of\ data}$라 할 때, 식 (2-10)에 의해 weighted average F1-score를 구할 수 있다.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + FN} \quad (2\text{-}6)$$

$$Precision = \frac{TP}{TP + FP} \quad (2\text{-}7)$$

$$Recall = \frac{TP}{TP + FN} \quad (2\text{-}8)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2\text{-}9)$$

$$weighted\ average\ F1 - score = \sum (F1_k \times w_k) \quad (2\text{-}10)$$

3. 실험 적용 및 결과분석

3.1 실험환경

실험환경 OS(Operation System)는 “Ubuntu 18.04.3 LTS”, 그래픽카드는 NVIDIA GeForce GTX 1080 Ti, Python 버전은 3.7로 실험을 진행했다. Glove를 통한 문장 임베딩을 수행하기 위해 Patel *et al.* (2018)이 제공한 패키지와 glove.6B.100 모델을 이용했고, BERT, RoBERTa, SBERT는 다양한 언어모델을 제공하는 프로젝트인 허깅페이스(huggingface)[15]에서 제공하는 모델을 사용했다.

[15] https://huggingface.co/

3.2 데이터 세트 구축 결과

3.2.1 지리공간분석절차 도출

GeoAnQu 말뭉치와 OSM 온톨로지를 이용한 방식을 통해 총 59개의 고유한 분석절차를 도출할 수 있었다. 이들 분석절차를 연산함수 개수를 기준으로 오름차순으로 정렬해 각 분석절차에 고유 클래스번호를 부여했다. 이중 각 클래스에 속하는 데이터가 1개인 경우 학습용데이터와 테스트용 데이터로 분리할 수 없기 때문에 해당 클래스는 실험에서 제외하였다. 최종적으로 도출된 23개 클래스를 [표 3-1]을 통해 확인할 수 있다.

[표 3-1] GeoAnQu말뭉치 분석을 통해 도출한 분석절차

분석절차	클래스 번호
속성조회	0
데이터베이스 조작 -> 속성조회	2
속성조회 -> 지리공간 통계 기법 수행	5
속성조회 -> Overlay 연산	6
속성조회 -> Buffer 연산 -> Overlay 연산	8
데이터베이스 조작 -> 속성조회-> 지리공간 통계 기법 수행	11
속성조회 -> 네트워크 연산 ->결과 쿼리	13
속성조회 -> Overlay 연산->결과 쿼리	14
속성조회-> 일반화 ->지리공간 통계 기법 수행	15
속성조회 -> Overlay 연산-> 지리공간 통계 기법 수행	16
속성조회 -> 지오메트리 측정 -> 속성조회	18
데이터베이스 조작 -> 속성조회 -> Buffer 연산 -> Overlay 연산	19
속성조회 -> Buffer 연산 -> Overlay 연산 -> 데이터 쿼리	21
데이터베이스 조작 -> 속성조회 -> Overlay 연산 -> 데이터 쿼리	27
데이터베이스 조작 -> Overlay 연산-> 속성연산 -> 데이터 쿼리	28
데이터베이스 조작-> 속성조회-> Overlay 연산->지리공간 통계 기법 수행	32
속성조회 -> 지오메트리 측정 -> 속성연산 -> 데이터 쿼리	33
네트워크 연산 -> 데이터 쿼리-> Buffer 연산 -> Overlay 연산 -> 데이터 쿼리	39
속성조회 -> 네트워크 연산 -> 클래시피케이션 -> 데이터 쿼리-> Overlay 연산	42
지형연산 -> 클래시피케이션 -> 데이터 쿼리 -> 데이터 형식 변환 -> Overlay 연산	44
지리공간 통계 기법 수행 ->클래시피케이션 -> 데이터 쿼리 -> 데이터 형식 변환 -> Overlay 연산	46
속성조회->네트워크 연산 -> 클래시피케이션 -> 데이터 쿼리-> Overlay 연산->데이터 쿼리	50
속성조회->네트워크 연산 -> 데이터 쿼리->네트워크 연산-> 클래시피케이션->데이터 쿼리-> Overlay 연산	53

3.2.2 말뭉치 라벨링 및 어휘 변용

위 23개 클래스의 클래스 번호를 기준으로 GeoAnQu말뭉치에 대해 라벨링을 수행했다. 이 과정을 통해 총 227개의 데이터셋을 확보할 수 있었다. 그 후 데이터세트 증강과 알고리즘 일반화 검증을 목적으로 어휘변용을 해 494개의 데이터를 확보할 수 있었고 [표 3-2]를 통해 각 클래스에 대한 데이터 수를 확인할 수 있다. 각 클래스에서 어휘변용으로 늘어난 질의수가 차이를 보인다. 이는 각 질의에 포함된 단어 또는 문장형태에 중 Damodaran (2021)의 언어모델이 어휘변용을 수행할 수 없는 단어 또는 문장 형태를 포함하고 있기 때문이다.

[표 3-2] 어휘변용 전/후 데이터 수

클래스 번호	어휘변용 전 데이터 수	어휘변용 후 데이터 수 (괄호 안은 어휘변용 전에 비해 늘어난 질의 수)
0	11	25(+14)
2	11	44(+33)
5	13	22(+9)
6	6	20(+14)
8	30	62(+32)
11	5	6(+1)
13	7	25(+18)
14	11	27(+16)
15	6	21(+15)
16	5	11(+6)
18	12	31(+19)
19	5	12(+7)
21	7	17(+10)
27	7	31(+24)
28	8	30(+22)
32	4	6(+2)
33	4	10(+6)
39	6	8(+2)
42	4	9(+5)
44	4	13(+9)
46	7	29(+22)
50	9	17(+8)
53	6	17(+11)
합계	**227**	**494(+267)**

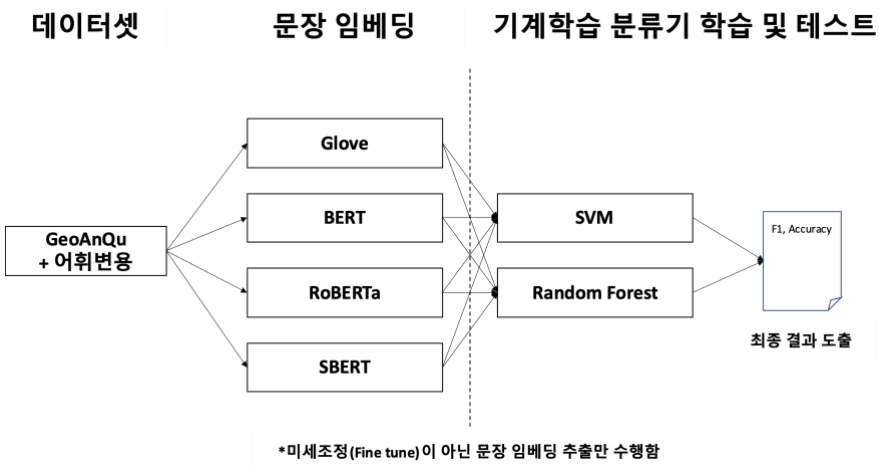
Damodaran (2021)이 제안한 언어모델을 사용하면 의문문을 명령문으로 바꾸는 방식, 단어를 동의어로 바꾸는 방식, 어순을 바꾸는 방식 등이 동시에 일어나게 된다. 또한 해당 모델은 단어를 소문자로 바꾸는 특징이 있다. 각 클래스에서 어휘변용된 질의를 [표 3-3]에서 확인할 수 있다.

[표 3-3] 어휘변용 된 질의 예시

클래스 번호	지리공간질의
0	What is the land use in the Happy valley resort?
0	tell me the use of land in the happy valley resort
2	What is the cervix cancer mortality rate of white women in each city in the western USA from 1970 to 1994?
2	show the cervix cancer mortality rates of white females for each city in the western us from 1970 to 1994
6	Which areas in Houston are not classified as flood plains?
6	which areas in houston are not flood plains?
14	What are the land use inside the flood zones in Oleander?
14	list the uses of land in the flood zone of oleander
15	What are weighted average coordinates of bank branches in Oleander?
15	give me weighted average coordinates of the branches of oleander bank
16	What is the weighted coordinate average of library patrons for each district in Oleander
16	calculate the weighted coordinate average of library patrons for each district in oleander

3.3 모델구성 및 학습

사전학습 된 모델을 사용하는 방식은 크게 두 가지로, 모델을 사용해 임베딩만 수행해 해당 임베딩을 기계학습 분류기의 입력값으로 사용하는 방식과, 사전학습모델의 파라미터를 연구목적에 맞게 재학습시키는 미세조정 방식이 있다. 미세조정 방식을 사용하게 되면 모델을 연구목적에 맞게 조정할 수 있는 장점이 있지만 학습을 위해 많은 데이터와 시간이 필요하며 최근 보고되는 연구에 따르면 미세조정시 모델 성능이 불안정해질 수 있다(Zhang *et al.*, 2020). 본 연구에서 쓰이는 데이터의 총수는 494개이고, 해당 데이터로는 유의미한 미세조정 방식을 수행할 수 없다. 따라서 문장 임베딩 수행 후 임베딩을 기계학습 분류기의 입력값으로 사용하는 방식을 채택했다([그림 3-1]).



[그림 3-1] 모델구성 및 학습 개념도

3.3.1 문장 임베딩

문장 임베딩에 앞서 어휘변용 및 라벨링을 통해 확보된 데이터 세트를 계층적 샘플링(stratified sampling)방식을 이용해 학습과 테스트 세트으로 나눴다. 계층적 샘플링은 지정한 비율에 따라 각 클래스를 학습과 데이터 세트로 나누는 방식으로 본 연구에서 사용한 데이터 세트는 데이터 불균형이 심해 랜덤샘플링(random sampling)을 시행할 시, 특정 클래스의 모든 데이터가 학습데이터로 분류되어 테스트를 진행할 수 없는 경우가 발생하기 때문에 계층적 샘플링방식을 사용했다. 계층적 샘플링 방식을 사용해 325개의 학습데이터와 169개의 테스트 데이터로 분리했다. 그 후 해당 데이터 세트를 각 언어모델을 이용해 임베딩을 수행해 학습 및 테스트에 쓰일 문장 임베딩을 확보했다.

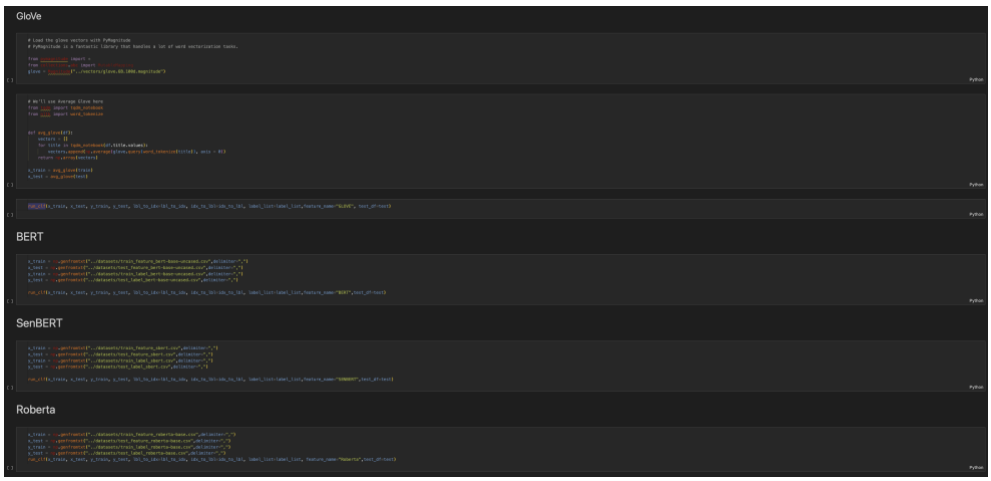
임베딩 된 질의는 모델에 따라 다른 벡터 차원(vector dimension)을 갖게 된다. 본 연구에서는 Glove, BERT, RoBERTa, SBERT를 이용해 임베딩을 수행했고, 해당 모델의 임베딩 차원은 각각 100, 768, 768, 768이다. 이 중 Glove를 이용해 ‘what areas are not wetlands in houston’ 라는 문장을 임베딩한 결과를 [표 3-4]를 통해 확인할 수 있다.

[표 3-4] Glove이용 'what areas are not wetlands in houston' 임베딩 결과(크기:100)

−0.0434770,	0.0325313,	0.0720024,	−0.0506995,	0.0121384,
0.0109590,	−0.0280939,	0.0576082,	−0.0096959,	−0.0118115,
−0.0232287,	−0.0395202,	0.0636854,	−0.0257063,	−0.0095234,
−0.0745400,	0.0480692,	0.0644448,	−0.0936422,	0.0570374,
0.0204296,	0.0584049,	0.0627301,	0.0204027,	−0.0112105,
−0.0204238,	0.0106391,	−0.0865625,	−0.0700105,	−0.0165072,
0.0007625,	0.0183587,	0.0282352,	0.0338914,	−0.0015282,
0.0228518,	0.0103286,	0.0803609,	−0.0284651,	−0.0121807,
−0.0588501,	−0.0617917,	−0.0156630,	−0.0355193,	0.0279047,
0.0161853,	0.0532929,	−0.0333124,	−0.0187346,	−0.1249969,
0.0016099,	−0.0158532,	0.0166877,	0.1239255,	−0.0261837,
−0.3320132,	0.0410837,	−0.1080591,	0.2612525,	0.1055176,
−0.1051268,	0.1524288,	0.0026136,	−0.0084181,	0.1710316,
0.0126081,	0.1062805,	0.0306764,	0.0308606,	−0.0380624,
−0.0144943,	−0.0571105,	−0.0211613,	−0.0540755,	0.0280905,
−0.0530129,	0.0059179,	0.0260564,	−0.1274586,	0.0298583,
0.0905459,	−0.0014158,	−0.1305073,	0.0673642,	−0.1942515,
0.0005768,	0.0269813,	−0.0280566,	−0.0302671,	−0.0186793,
−0.0031148,	−0.0624604,	−0.0455469,	0.0085596,	−0.1186768,
0.0689448,	−0.0556519,	−0.0543895,	0.0799255,	0.0446410

3.3.2 분류모델학습

각 언어모델을 사용해서 수행한 임베딩 값을 분류모델의 입력값으로 사용해 분류모델을 학습시켰다. 본 연구에서는 SVM 중 linear SVM과 랜덤포레스트를 분류모델로 선정해 학습시켰다. 이때 Pedregosa *et al.* (2011)이 제안한 프레임워크인 Scikit-learn을 사용했다. 해당 프레임워크 사용 시, 각 임베딩에 대해 분류모델을 재사용할 수 있도록 Python 함수(function)로 정의한 후 학습을 진행했다([그림 3-2]).

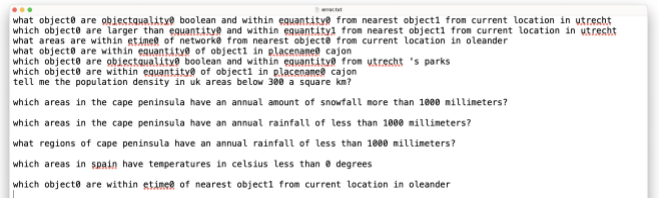


[그림 3-2] 각 임베딩 값에 대한 분류모델학습 코드

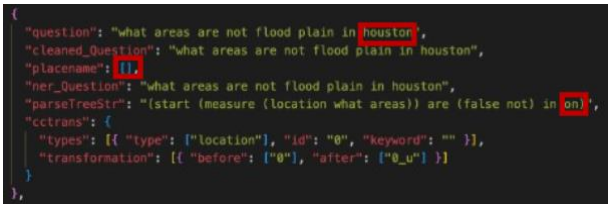
3.4 실험결과 분석

3.4.1 기존연구 알고리즘 적용 결과

Xu *et al.* (2022)이 제안한 알고리즘을 테스트해본 결과 모든 질의를 분석절차로 변환할 수 없는 것으로 나타났다. 그 이유는 해당 연구에서 'what', 'where' 등 질문어를 기준으로 규칙을 설정했기 때문이고, 또한 소문자로 변환된 지명을 탐지하지 못하는 한계도 확인했다. 테스트로 사용한 169개 질의 중 12개 질의는 오류로, 그 외 질의에 대해서는 지명을 제대로 탐지하지 못해 분석절차로 변환할 수 없었다([그림 3-3]). 따라서, 어휘 변용된 질의뿐만 아니라 단순히 소문자로 변환된 질의에 대해서도 해당 알고리즘이 작동하지 못하는 것을 확인할 수 있었다.



(a) 오류로 분류된 경우



(b) 지명 탐지 실패한 경우

[그림 3-3] Xu *et al.* (2022)의 알고리즘 적용 결과

3.4.2 모델성능 비교

각 임베딩 방식과 이를 기계학습방식을 이용해 분류한 결과는 [표 3-5]와 같다. Glove 성능은 [표 3-8], BERT 성능은 [표 3-9], RoBERTa 성능은 [표 3-10], SBERT 성능은 [표 3-11]을 통해 확인할 수 있다. SVM 및 랜덤포레스트에서 SBERT, Glove, BERT, RoBERTa를 이용한 임베딩 순서로 높은 성능을 보여주는 것을 확인할 수 있다. 이는 SBERT가 문장의 의미를 잘 나타낸 결과로 볼 수 있고, 다양한 문장 임베딩 방식을 이용해 성능비교를 연구 결과와 일치한다 (Reimers and Gurevych, 2019). 각 임베딩에 대한 분류모델 중 SVM 성능이 높은 것을 확인할 수 있다.

[표 3-5] 최종모델 성능

	SVM (Weighted avg F1 / Accuracy)	Random Forest (Weighted avg F1 / Accuracy)
RoBERTa	0.703 / 0.710	0.502 / 0.527
BERT	0.843 / 0.858	0.675 / 0.705
Glove	0.914 / 0.923	0.862 / 0.882
SBERT	**0.925 / 0.935**	**0.888 / 0.899**

각 클래스에 해당하는 어휘변용 된 문장과 원 문장에 대한 예측 결과를 표로 나타내면 [표 3-6]와 같고, SBERT와 Glove가 같은

의미를 가지는 다른형태 문장에 대해 동일한 클래스로 분류하는 모습을 확인할 수 있다. 해당 모델을 통해 문장의 의미를 반영한 임베딩을 수행한 결과로 해석할 수 있다. 반면 BERT와 RoBERTa도 높은 성능을 보였지만 어휘변용 된 문장에 대해 다른 클래스로 분류했기 때문에 문장 특징 추출 시 의미론적 유사도를 제대로 반영하지 못했다고 볼 수 있다.

[표 3-6] Linear SVM을 이용한 분석절차 변환 결과

질의	실제 값	예측 값			
		GLOVE	BERT	RoBERTa	SBERT
show some areas in houston that are not considered green belts?	6	6	6	6	6
which are not green belt areas in houston?	6	6	6	6	6
list the areas in amsterdam within 1000 meters of the major transport routes?	8	8	8	16	8
what areas are within 1000 meters of the major transport routes in amsterdam?	8	8	8	8	8
tell me the weighted average coordinates of bank branches in oleander	15	15	15	15	15
what is the average weighted coordinates of bank branches in oleander?	15	15	15	15	15
tell me the number of elderly people in each neighborhood of amsterdam?	27	27	27	27	27

질의	실제 값	예측 값			
		GLOVE	BERT	RoBERTa	SBERT
what is the number of elderly people for each neighborhood in amsterdam?	27	27	27	27	27
tell me the density of cycling destinations in the metro vancouver region of canada?	32	32	2	0	32
tell me the point density of cycling destinations in the metro vancouver region in canada?	32	32	2	2	32
tell me the population density in california?	33	33	5	5	33
what is population density in california?	33	33	33	33	33

*질의가 소문자로 표현되는 것은 사용한 어휘변용 모델의 특징임

각 클래스에서 임의로 문장을 선택해 비교분석을 실시했고, [표 3-7]을 통해 결과를 확인할 수 있다. Glove를 제외한 BERT, RoBERTa, SBERT는 유사한 계열 모델임에도 불구하고 주어진 문장을 분류하는 성능이 큰 편차를 보임을 알 수 있다. 예를 들어 “which houses have construction year between 1990 and 2000 in utrecht” 라는 문장은 0번 지리공간 분석절차에 해당하지만 Glove, BERT, RoBERTa, SBERT는 각각 2,14,14,2로 분석절차를 예측했다([표 3-7]). 0번 분석절차는 가장 간단한 분석절차로 단순히 공간객체의 속성을 조회하는 방식이고 2번은 데이터베이스에 없는 속성정보를 보충한후 공간객체 속성을 조회하는 지리공간 분석절차다. 따라서 0번 분석절차와 2번 절차는 0번 절차와 14번 절차에 비해 높은 유사도를 가지는 분석절차로 정의할 수 있다. 따라서 Glove와 SBERT를 통해 도출한 작업절차는 비록 정답인 0번은 아니지만 0번과 가장 유사한 2번 분석절차로

변환했기 때문에 14번에 비해 답에 근접한 결과로 해석할 수 있다.

[표 3-7] 각 클래스별 결과 랜덤 샘플링

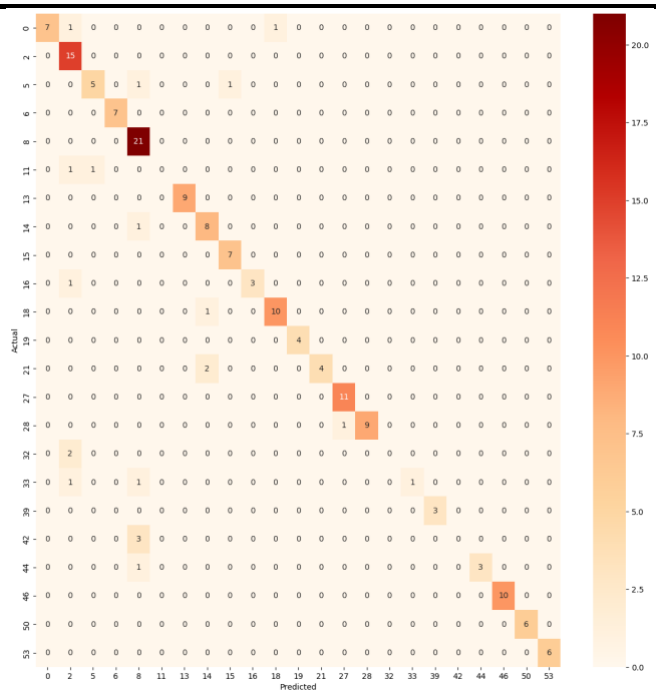
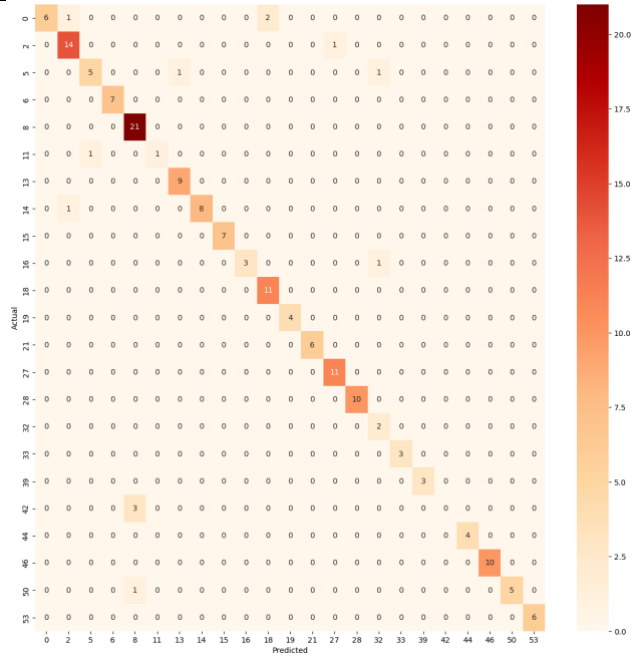
질의	실제 값	예측 값			
		GLOVE	BERT	RoBERTa	SBERT
which houses have construction year between 1990 and 2000 in utrecht	0	2	14	14	2
show the auto accidents in tarrant county in texas?	2	2	0	0	2
what is the euclidean distance to the rivers in crook, deschutes, and jefferson county	5	13	11	42	5
what areas are not conatined as green belt areas in houston	6	6	6	6	6
what areas are within 2000 meters of the playgrounds in oleander?	8	8	8	8	8
what is the interpolated surface of ozone concentration in california	11	5	11	5	5
which are the two fire stations nearest each school in utrecht?	13	13	13	2	13
which land use contains meteorological stations in netherlands	14	2	0	6	0
what is the central feature of bank branches in oleander	15	15	15	15	15

질의	실제 값	예측 값			
		GLOVE	BERT	RoBERTa	SBERT
what is the mean center of the fire calls for each alarm territory in fort worth in 2017	16	32	32	46	32
which park is the biggest in utrecht?	18	18	18	18	18
what areas are within two miles of urban land use in loudoun county in the us?	19	19	19	19	19
what houses are for sale and within 0.5km from the main roads in utrecht	21	21	21	14	14
what is the number of election votes for each precinct in dallas?	27	27	2	2	2
tell me the average rating of the street pavement for each borough in new york city?	28	28	28	14	28
tell me the population density in uk areas below 300 a square km?	33	33	5	5	33
what houses are less than 30 square meters and within 1km from the nearest school (from my current location) in utrecht	39	39	39	39	39
which areas are within 60 minutes of the airports in crook deschutes and jefferson counties?	42	8	8	42	8
which areas in spain have altitudes between 700 and 2000 meters?	44	44	44	18	44

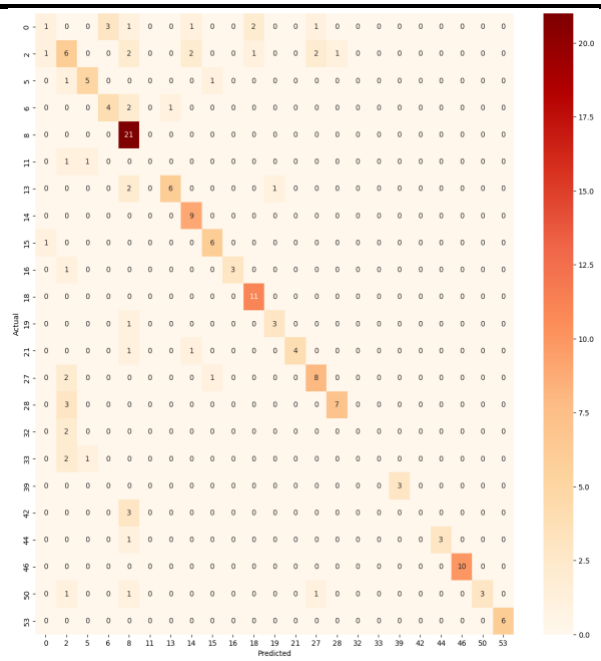
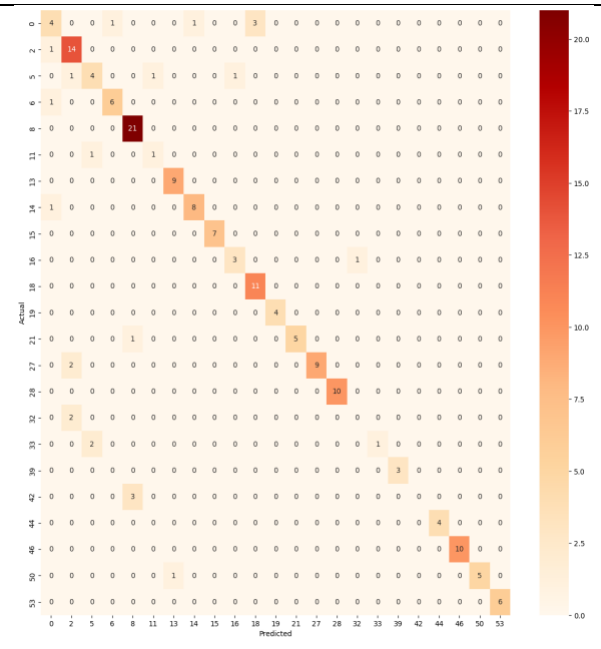
질의	실제 값	예측 값			
		GLOVE	BERT	RoBERTa	SBERT
which areas in spain have temperatures in celsius less than 0 degrees	46	46	46	46	46
list the four fire stations within 3 minutes of a fire in san francisco?	50	50	13	14	50
which areas are accessible within 3 minutes by car from the nearest fire station from my current location in oleander?	53	53	53	53	53

*질의가 소문자로 표현되는 것은 사용한 어휘변용 모델의 특징임

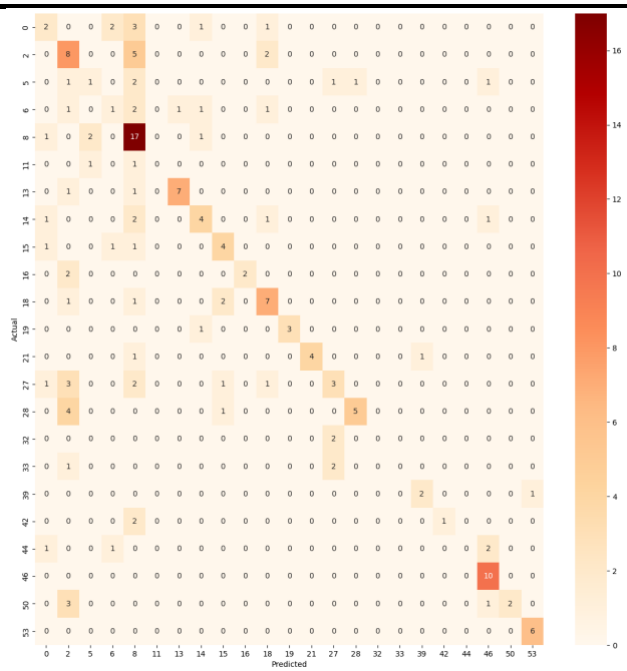
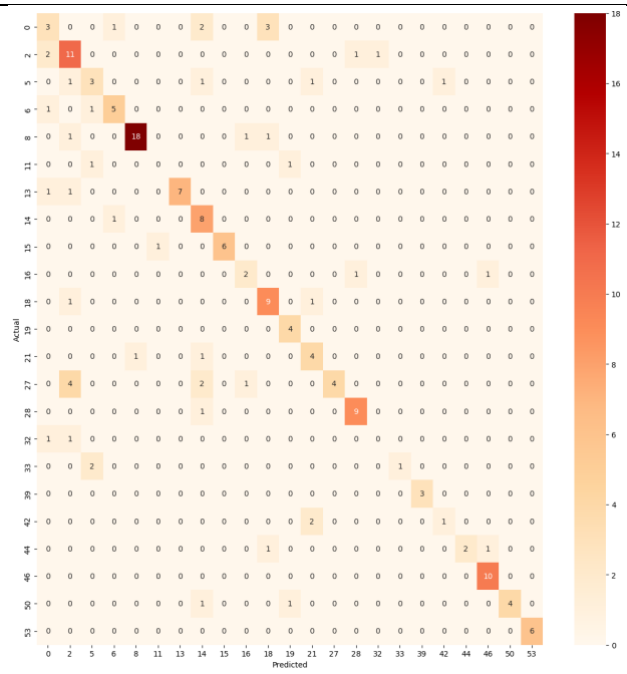
[표 3-8] Glove 임베딩 사용 분석절차 변환 confusion matrix

Glove	
분류모델	**결과**
Random Forest	
SVM	

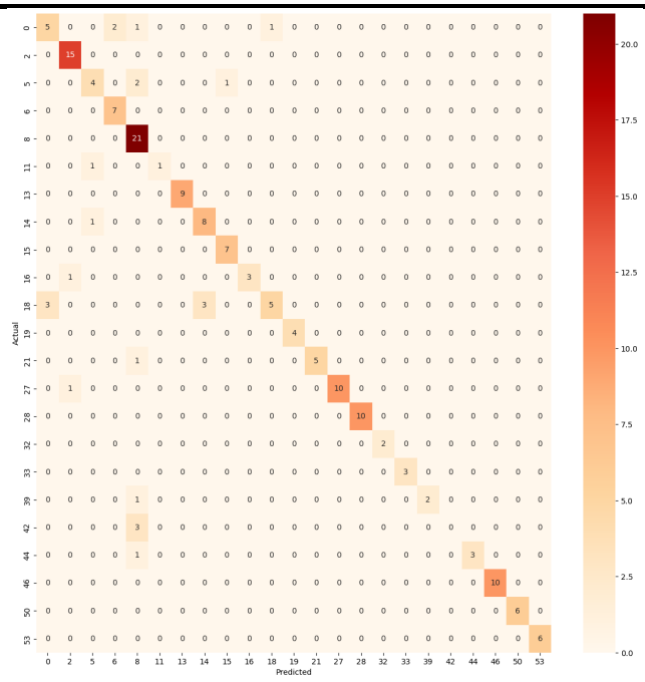
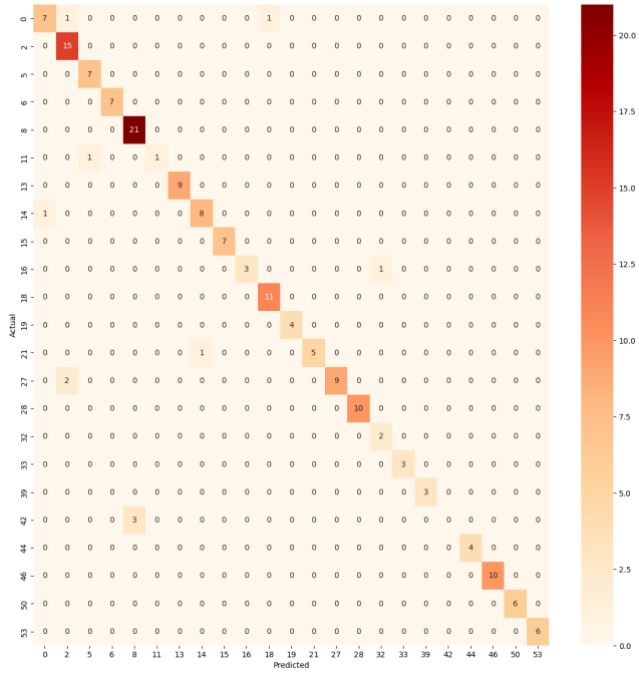
[표 3-9] BERT 임베딩 사용 분석절차 변환 confusion matrix

BERT	
분류모델	결과
Random Forest	
SVM	

[표 3-10] RoBERTa 임베딩 사용 분석절차 변환 confusion matrix

RoBERTa	
분류모델	결과
Random Forest	
SVM	

[표 3-11] SBERT 임베딩 사용 분석절차 변환 confusion matrix

SBERT	
분류모델	결과
Random Forest	
SVM	

4. 결론

본 논문에서는 지리공간질의를 기계학습 방식을 이용해 지리공간 분석절차로 변환하는 연구를 수행했다. 기존 지리공간 분석절차 변환 연구는 제한된 키워드에 대해 작동하는 방식이고 지리공간연산함수를 표현하지 못하는 한계를 지녔다. 또한 기존 연구에서 제안한 방식은 제한된 키워드에 의존하는 방식이기 때문에 실제로 사용자가 구사하는 다양한 질의형태를 답하지 못하는 한계점을 가지고 있다. 따라서 본 연구에서는 다양한 질의형태에 대응할 수 있는 분석절차 변환 연구를 수행하기 위해 기계학습 방식을 도입했고, 지리공간연산함수를 중심으로 라벨링을 실시하여 최종적으로 도출된 결과가 지리공간연산함수를 포함하도록 설계했다. 지리공간 연산자를 선정할 때 기존 GIS 도구에서 쓰이는 지리공간 연산자를 최대한 반영할 수 있는 분류체계를 참조하여 진행했다. 다양한 분류체계가 존재하는 것을 확인할 수 있었지만, 비교적 많은 연산자에 대해 정확한 설명을 포함하고 있는 연산함수 분류체계인 Li and Stefanakis (2020)의 연산함수 분류체계를 기준으로 라벨링을 실시했다. 지리공간질의 말뭉치로는 GeoAnQu를 선정했고, 그 이유는 해당 말뭉치에 등장하는 질의를 답하기위해 다양한 분석을 필요로 하는 것으로 알려져 있기 때문이다. GeoAnQu 말뭉치를 분석해서 고유한 분석절차를 도출하고 해당 분석절차에 고유번호를 부여해서 이를 라벨링시 사용했다.

또한 데이터가 부족한 한계를 극복하고 다양한 질의 형태를 발생시키기 위해 언어모델의 도움을 받아 어휘변용을 실시해 충분한 수의 데이터를 확보하고 제안하는 방법이 다양한 문장형태에 대해서 작동하는지 검증하였다. 말뭉치를 기계학습에 사용하기 위해서는 문장의

의미를 잘 전달할 수 있는 문장 임베딩 수행이 필요하고 본 연구에서는 문장의 의미를 잘 표현할 수 있는 것으로 알려진 Glove, BERT, RoBERTa, SBERT를 이용해서 문장 임베딩을 수행 후 기계학습 분류모델의 입력 값으로 사용했다. 위 과정을 거쳐 모델 성능을 비교해본 결과 SBERT을 통한 문장 임베딩을 이용한 성능이 가장 높음을 확인할 수 있었다. 일반적인 문장에 대해 사전학습한 모델임에도 불구하고 GIS와 같이 특정 분야에 대한 질의를 분류할 때도 우수한 성능을 보이는 것을 확인했다. 최종적으로 SBERT와 linear SVM을 조합한 모델이 가장 우수한 성능을 보이는 것을 확인했고, 해당 모델을 통해 다양한 질의 형태를 분석절차로 변환할 수 있었다.

본 연구의 한계점은 다음과 같다. 본 연구에서 질의를 분석절차로 변환하기 위해 문장분류 방식을 택했다. 따라서 본 연구에서 제안한 방식은 임의의 질의가 주어졌을 때 23가지 분석절차 중 하나로 대응시키는 것이다. 만약 특정 질의를 답하기 위한 분석절차가 23가지 분석절차 중 하나에 해당하지 않을 때, 해당 라벨이 존재하지 않기 때문에 비록 모델은 결괏값을 생성하지만 질의를 답하기 충분한 분석절차라고 볼 수 없다. 이를 보완하기 위해서는 문장분류 방식이 아닌 자연어 처리 분야에서 사용하는 방식 중 하나인 텍스트 생성(text generation)기법을 도입할 필요가 있다. 해당 기법을 사용하게 되면 질의를 거의 모든 분석절차로 변환할 수 있을 것으로 기대할 수 있다. 그러나 해당 기법을 사용하기 위해서는 충분한 수의 말뭉치를 확보해야 하며, 말뭉치 확보를 위해 사용할 수 있는 방법 중 하나로는 KB의 온톨로지를 이용해 자연어를 생성하는 방법(Agarwal *et al.*, 2020)이 존재하며, 기 구축된 GeoKB에 해당 방법론을 적용해 말뭉치를 확보할 수 있을 것으로 기대 할 수 있다.

다른 한계점은 본 연구는 영어를 대상으로 진행되어 한국어와 같은 다른 언어에 대한 성능을 검증하지 못했다는 점이다. 한국어에 대해 본 연구와 유사한 연구를 진행하기 위해서는 한국어 말뭉치를 확보하고 한국어 언어모델을 사용할 필요가 있다. 현재 한국어 말뭉치로는 Korpora(Korean corpora archives)[16], 국립국어원에서 제공하는 모두의 말뭉치 [17] 등이 존재하고, 해당 말뭉치에서 지리공간질의를 분류해 지리공간질의 말뭉치를 확보해야 할 것이다. 언어모델의 경우, 허깅페이스에서 한국어 BERT, RoBERTa, SBERT모델을 제공하는 것을 확인할 수 있고 각각 34, 37, 2개의 모델을 제공하기 때문에 해당 모델 성능을 비교 분석해 사용할 필요가 있다. 위 과정을 통해 말뭉치와 언어모델을 확보한다면 본 연구와 유사한 연구를 한국어에 대해 진행할 수 있을 것으로 기대 할 수 있다.

[16] https://ko-nlp.github.io/Korpora/

[17] https://corpus.korean.go.kr/

참고 문헌

Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2020). Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *ArXiv*. https://doi.org/10.48550/ARXIV.2010.12688

Albrecht, J. (1998). Universal analytical GIS operations: A task-oriented systematization of data structure-independent GIS functionality. *Geographic information research: Transatlantic perspectives*, 577-591.

Almobydeen, S. B., Viqueira, J. R., & Lama, M. (2022). GeoSPARQL query support for scientific raster array data. *Computers & Geosciences*, *159*, 105023. https://doi.org/10.1016/j.cageo.2021.105023

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., & Nguyen, T. (2016). Ms marco: A human generated machine reading comprehension dataset. *ArXiv*. https://doi.org/10.48550/ARXIV.1611.09268

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*. https://doi.org/10.1145/3544558

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/bf00994018

Damodaran, P. (2021). Parrot: Paraphrase generation for NLU. *v1.0*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*. https://doi.org/10.48550/ARXIV.1810.04805

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In *Machine Learning in Radiation Oncology* (pp. 3-11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *ArXiv*. https://doi.org/10.48550/ARXIV.1909.00512

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1-32.

Franc, V., & Hlavac, V. (2002). Multi-class support vector machine. In *2002 International Conference on Pattern Recognition* (Vol. 2, pp. 236-239). IEEE. https://doi.org/10.1109/ICPR.2002.1048282

Gao, S., & Goodchild, M. F. (2013). Asking Spatial Questions to Identify GIS Functionality. In *2013 Fourth International Conference on Computing for Geospatial Research and Application* (pp. 106-110). IEEE. https://doi.org/10.1109/comgeo.2013.18

Google. (2022). *In-depth guide to how Google Search works*. Retrieved 12.07 from https://developers.google.com/search/docs/fundamentals/how-search-works

Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In *Handbook on Ontologies* (pp. 1-17). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-92673-3_0

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

Hamzei, E., Winter, S., & Tomko, M. (2022). Templates of generic geographic information for answering where-questions. *International Journal of Geographical Information Science*, *36*(1), 188-214. https://doi.org/10.1080/13658816.2020.1869977

Harris, Z. S. (1954). Distributional Structure. *Word*, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, *7*(4), 275-300. https://doi.org/10.1017/s1351324901002807

Hovy, E. H., Gerber, L., Hermjakob, U., Junk, M., & Lin, C.-Y. (2000). Question Answering in Webclopedia. *TREC*, *52*, 53-56.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, *26*(12), 2267-2276. https://doi.org/10.1080/13658816.2012.722637

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, M., & Stefanakis, E. (2020). Geospatial Operations of Discrete Global Grid Systems—a Comparison with Traditional GIS. *Journal of Geovisualization and Spatial Analysis*, *4*(2), 1-21. https://doi.org/10.1007/s41651-020-00066-3

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*. https://doi.org/10.48550/ARXIV.1907.11692

Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 1-13.

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

Meaden, G. J., & Do Chi, T. (1996). *The functioning of a GIS*. Retrieved 10.17 from http://www.fao.org/3/W0615E/W0615E06.htm#ch6

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*. https://doi.org/10.48550/ARXIV.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*. https://doi.org/10.48550/ARXIV.1310.4546

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Patel, A., Sands, A., Callison-Burch, C., & Apidianaki, M. (2018). Magnitude: A fast, efficient universal vector embedding utility package. *ArXiv*. https://doi.org/10.48550/ARXIV.1810.11190

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://doi.org/10.48550/ARXIV.1201.0490

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv*. https://doi.org/10.48550/arXiv.1802.05365

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. In *Proceedings of the 12th Workshop on Geographic Information Retrieval (GIR'18)* (pp. 1–10). Association for Computing Machinery. https://doi.org/10.1145/3281354.3281362

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*. https://doi.org/10.48550/ARXIV.1908.10084

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

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823). https://doi.org/10.1109/CVPR.2015.7298682

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00492-0

Stadler, C., Lehmann, J., H□ffner, K., & Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, *3*(4), 333–354. https://doi.org/10.3233/SW-2011-0052

Stefanakis, E., & Sellis, T. (1998). Enhancing Operations with Spatial Access Methods in a Database Management System for GIS. *Cartography and Geographic Information Systems*, *25*(1), 16–32. https://doi.org/10.1559/152304098782441723

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *ArXiv*. https://doi.org/10.48550/ARXIV.1804.07461

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *ArXiv*. https://doi.org/10.48550/ARXIV.1901.11196

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical systems and signal processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

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*, 1–31. https://doi.org/10.1080/13658816.2022.2077947

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020). Revisiting few-sample BERT fine-tuning. *ArXiv*. https://doi.org/10.48550/ARXIV.2006.05987

Abstract

Development of Machine Learning Based Geographic Analysis Workflow Transduction Technique for Geographic Questions with Various Sentence Type

Heejin Chae
Department of Civil and Environment Engineering
The Graduate School
Seoul National University

Despite the advance of the question answering(QA), which derives succinct and clear answers to questions from documents, there is a lack of a system to answer questions related to geospatial information, which increases by around 20% annually. The research field emerged to solve this problem is named geographic QA. Geo-analytical QA, a subfield of geographic QA, is a study to convert geographic question into geospatial analysis workflow and find the suitable tool and data to perform the analysis workflow. In order to perform realistic Geo-analytic QA, questions with various sentence type must be converted into geospatial analysis workflow. But it is difficult to perform realistic Geo-analytical QA through the method

proposed in the previous study because it is rule based approach that fits into limited sentence type. Therefore, to perform realistic Geo-analytical QA, this study proposes a method to convert geospatial questions with various sentence type into geospatial analysis workflow. In addition, in order to perform geospatial analysis, it is important to understand the geospatial operators, so the derived geospatial analysis workflow was set to include the geospatial operators in order according to the analysis intention. In this study, sentence classification techniques were applied to convert geospatial questions into analysis workflow. To use sentence classification techniques, it is necessary to select corpus, label corpus to create datasets, embed questions in corpus to make datasets as input values for classification models, and to learn classification models. The GeoAnQu corpus, known to require various geospatial analysis workflow, was selected and analyzed as the target corpus to derive its own analysis workflow, and then a unique number was assigned to the analysis workflows. Based on the unique number, the questions appearing in the GeoAnQu corpus was labeled to secure a dataset, and then paraphrase was performed to generate various sentence types and increase the data size. After that, sentence embedding was performed using Glove (global vectors), BERT(bidirectional encoder presentations from transformers), RoBERTa(robustly optimized BERT pre-training approaches) and SBERT(sentence-BERT) and then those embeddings were used to learn random forest and linear support vector machine(SVM) respectively. Finally, it was confirmed

that the model that trained with SBERT sentence embedding in linear SVM showed the highest performance, and the model was able to convert geospatial questions with various sentence type into geospatial analysis workflow. In addition, the limitations of the results were analyzed and future research directions were presented.

Keywords : GeoQA, Geo-analytical QA, text classification, geographic questions corpus, geospatial analysis workflow, sentence embedding
Student Number : 2021-26777