Universiteit Utrecht

Master Thesis

Comparing deep learning methods for concept recognition in geo-analytic questions

Aristoteles Kandylas, 7723822

Supervisors: Simon Scheider, Haiqi Xu

Utrecht University

Applied Data Science

Utrecht, July 1, 2022

Contents

Abstract

Named Entity Recognition (NER) is an important process of NLP systems for relation extraction, information retrieval and machine translation. Although various NER systems researched and improved for many decades, more accurate and advanced NER systems, which exploit deep learning techniques have emerged in the NLP domain, only the last few years. These newly emerged NER systems, due to the word embeddings and the non-linear transformations of data, lead to significantly improved performance. They are capable of tagging and classifying semantic entities such as person, location, organization, time, quantities, etc. more easily and accurately. For interpretation of geo-analytical questions, these NER systems should detect GIS-related semantics such as geographic phenomena, place names and temporal information. The last two pieces of information can be recognized by the current NER models, but none of them can identify and categorize geographic phenomena. To this end, this study presents two deep learning-based NER systems to extract geographic phenomena from geo-analytical questions and classify them into core concepts of spatial information that conceptually model and distinguish spatial information. The NER systems are trained by BERT and Bi-LSTM models on 278 geo-analytical questions and tested on 31 validation questions, from a corpus that contains 309 questions in total. The evaluation and comparison results showed that the BERT model had higher accuracy, precision, recall and F1-score on recognizing core concepts in geo-analytical questions, compared to Bi-LSTM.

Keywords: Geo-analytical questions, Natural language processing (NLP), Named Entity Recognition (NER), deep learning, GIS, core concepts of spatial information, geo-computation

1. Introduction

During the last decade, the research about question answering systems has gained a lot of attention among enterprises and scientific institutions. Most question-answering systems which can be found today on Web Search Machines are focused on answering simple questions such as “Where is Utrecht?” and not questions such as “What houses are for sale within 1km from the nearest school in Utrecht?”. The reason for this is that questions such as the latter one, do not have an a-priori answer, but instead the answer needs to be estimated using geo-spatial analysis (Xu et al., 2022; Xu et al., 2020). These questions are known as geo-analytical questions, a kind of more sophisticated questions that can be answered with the use of Geographic Information Systems (GIS) tools after generating analytic workflows. To achieve that, firstly the geo-analytical questions should be translated into core concept transformations from which useful information can be retrieved. With this information, it is possible to extract pertinent data and choose appropriate GIS tools, to generate the corresponding answer to geo-analytical questions (Xu et al., 2022).

To this end, the core concepts are very important, for the geo-analytical question answering systems, as words and phrases of a question can be annotated as core concepts. In general, the core concepts are concepts needed to understand and interpret the spatial information included in GIS-related text. Every core concept has specific properties (e.g., ‘Events’ happen in time in specific location, ‘Networks’ are relations between objects) to which, particular GIS operations can be applied. Through these operations, the answer to a geo-analytical question can be generated and visualized in GIS software (Xu et al., 2022; Scheider et al., 2020). For example, in a geo-analytical question answering system, the core concepts of a geo-analytical question can be recognized and annotated in the following way:

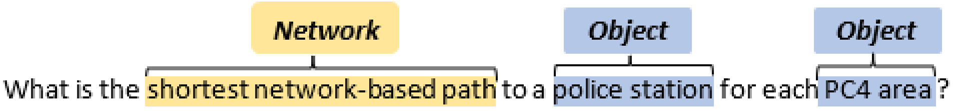

Network
Object
Object
What is the shortest network-based path to a police station for each PC4 area ?

Figure 1. Annotation of core concepts in a geo-analytical question

The answer to this geo-analytical question can be visualized in GIS software via a analytic GIS workflow as shown in Figure 2:

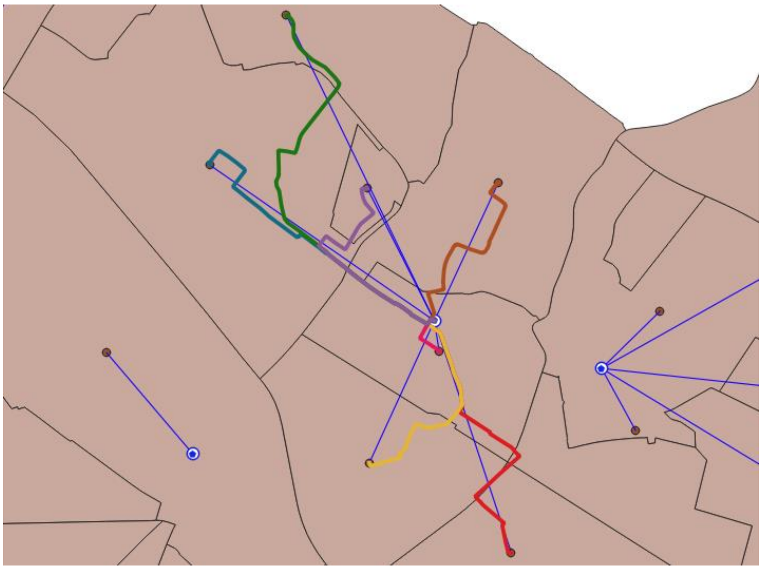


Figure 2. Shortest network-based paths to a police station for specific PC4 areas

In this direction, the development of a Named Entity Recognition (NER) system is the first step toward the automatic recognition of GIS-related text for information extraction and retrieval (Dai et al., 2018), text clustering (Chen et al., 2018) and automatic text summarization (Enríquez et al., 2017). GIS named entity recognition (NER) (can also be referred to as spatial core concept and element identification from GIS texts) can be an important process in GIS-related language processing that will involve the use of pertinent terms (single words and multiword phrases) to identify and annotate information into predetermined categories-core concepts (e.g., object, field, event, network, etc.).

Consequently, it is necessary to develop innovative and scalable models and methods that can automatically recognize core concepts related to various GIS phenomena. Once these newly developed models achieve satisfactory results in core concept recognition tasks, they could be implemented, for question answering (QA) with GIS and in extent for GIS tasks. However, up until now, the pre-existing techniques and models for automatic interpretation of core concepts from GIS-related text were limited, regarding their capabilities, in comparison to recent NER models for the following reasons. Firstly, in comparison to general narrative text, GIS-related texts have the issue that they need to be interpreted on a conceptual level, which is not immediately obvious from the words used in the text. Thus, there is an interpretation step needed to relate words/tokens in a text with the concepts relevant to geographic information. Secondly, each noun-core concept is interpreted differently in different geo-analytical questions. For instance, the noun "*areas*" have disparate meaning when it is presented in a question, isolated (a region or part of the world) and following the word e.g. urban (a town, a city, or the suburbs of a city) or catchment (an area which attracts people based on its services and human activities located in it). Thirdly, the pre-existing NER techniques (rule-based or supervised learning-based) necessitate in many cases a considerable amount of manual labor to create a comprehensive and cohesive dataset of representative core concepts for each word/token or to annotate the text which will be used for the training of a model and dealing with the high variability in new text patterns. The major restriction in the training process of a supervised learning model is that it necessitates undue human workload to label manually the training dataset. Howbeit, there is limited availability of such labeled datasets in the GIS domain. The annotation of a dataset is an arduous and time-consuming process, as annotators need not only specific domain expertise (e.g., geoscience), but also a natural language processing (NLP) background (Liu et al., 2022; Qiu et al.,2019).

Nowadays, the recent development of deep learning (DL), expanded the capabilities of the NLP field with the inclusion of more advanced neural network models. With these models, DL has become extremely popular among researchers, due to their ability to learn representations from data without requiring sophisticated feature engineering and their state-of-the-art performance in most tasks (Gao et al., 2019). More specifically, for the NER tasks, the deep learning techniques intend to empower the model to an autonomously feature learning process from numerous annotated data (Liu et al., 2022). One popular neural network model which integrates this deep learning process is the Bidirectional-Long Short-Term Memory (Bi-LSTM). This model uses attention mechanisms to detect the correct context in sentences, but it demands a substantial amount of manually annotated datasets for its training (Gao et al., 2019; Liu et al., 2022). Another model which has emerged recently and is widely used for NLP tasks is the Bidirectional Encoder Representations from Transformers (BERT). This model utilizes the transformers-based architecture, and it is a pretrained language model. This means that compared to the Bi-LSTM model, BERT has been trained previously on extensive unlabeled text corpora (which is computationally expensive), so the user can only fine-tune the model with fewer resources to optimize its performance on particular NER tasks (Ezen-Can, 2020; Devlin et al., 2019).

To tackle the aforementioned challenges of the older NLP methods on GIS NER tasks, in this paper, the performance of these two deep learning models (BERT and Bi-LSTM) is tested and evaluated in classifying and tagging spatial core concepts in GIS-related text. Through the evaluation of these models, useful conclusions are drawn regarding their strengths and limitations in capturing GIS phenomena in geo-analytical questions. The present work aims to answer the following questions:

1. Which DL methods/approaches are suitable for detecting core concepts in geo-analytical questions?
2. What is the performance quality of NER classifiers based on such methods?
3. What are their weaknesses and how could they be improved?

The rest of this paper is organized as follows: Section 2 details the related work from the NER domain and the two models (BERT and LSTM) background. Section 3 presents the used data and its pre-processing steps. In Section 4 the proposed deep learning models-algorithms are introduced. Section 5 reports and analyzes the two models' performance results. Then, in Section 6 the experimental results are discussed and directions for future research are provided. Finally, Section 7, is the conclusion of my study.

2. Background and Related Research

2.1 Core concepts of spatial information

The core concepts of spatial information were firstly described by Kuhn (2012), as specific concepts through which a GIS environment can be studied (Scheider et al., 2020). In the recent version of Kuhn's research (Kuhn and Ballatore, 2015), five core concepts that generalize the geographical information in terms of fundamental GIS phenomena, are included:

- **Locations** are used to answer questions such as, *where* spatial phenomena are located and to calculate the geometric properties (e.g., size, height).
- **Objects** answer questions regarding the properties and relations of objects. Objects correspond to spatially constrained regions that have their own identities and spatial, temporal, and thematic properties (qualities). For instance, the different municipalities in the Netherlands are considered objects, as each one of them has its bounded spatial region, a distinct identity (e.g., Utrecht is the 4th largest city in the Netherlands) and a population as quality.
- **Fields** answer questions regarding the value of a phenomenon in space. Fields are certain functions whose domain are locations for which the distance can be measured and do not have a predefined range. For example, the air temperature of a country, the slope and the elevation of an area are considered fields.
- **Events** are entities that happen in a specific time (they have a particular duration) and have locations, fields, objects, and networks as participants. For instance, a hurricane and a rainfall are considered events as both take place in a location and have a particular duration.
- **Networks** are considered quantified relationships between objects. In this way, networks provide information on whether two objects in space are linked or not. The walking distance from the Utrecht University campus to the city center or the driving time from a residence to the work are two examples of networks (Xu et al., 2022; Scheider et al., 2020; Kuhn and Ballatore, 2015; Kuhn, 2012).

However, the five aforementioned core concepts cannot cover all the core concept transformations in the GIS domain. Due to this limitation, the use of two further concepts, namely proportion and amount is necessary (Xu et al., 2022).

- **Proportion** is a quantity ratio derived from amounts. Different proportions are generated by different combinations of amount categories. For example, the crime rate is a ratio of two content amounts, the crime and the population count. In contrast to the amount, the proportion is measured as intense ratio scale (IRA).
- **Amount** quantifies core concepts or their properties and can be distinguished into two categories-types the content amounts (the aggregated outcomes of core concepts and their properties in space) and the coverage amounts (quantify the 'coverage' of core concepts in space). An example of content amount is the household income (object content amount) while the total area of a park is an example of coverage amount. Although the object and event content amounts are measured on count level, the field and concept content amounts are measured on extensive ratio level (ERA) (Xu et al., 2022).

2.2 Measurement levels in the GIS domain

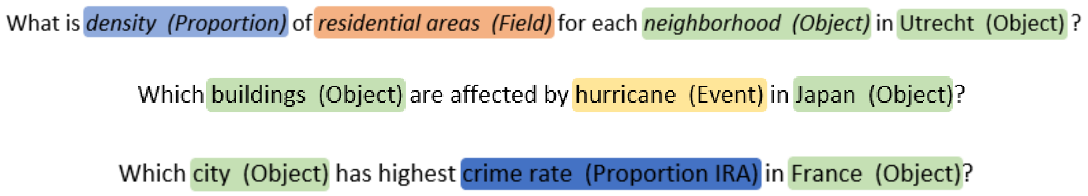
The different core concepts can be classified with particular measurement levels-scales. The measurement level of a core concept affects the type of analytical tools that may be used in GIS analysis (Chrisman, 1998). There are 4 measurement scales:

- **Nominal:** On a nominal measurement scale, numbers are employed to establish identity. Examples of Nominal scales are the zip codes or the telephone numbers. Any mathematical operations on the nominal scale will not produce a meaningful result. Adding two zip codes will lead to a meaningless number.
- **Ordinal:** The number in this scale establishes order. The ordinal scale can be used to classify the most popular restaurants in a city based on the number of people visiting them every day. Any mathematical operations on the ordinal scale will also produce a meaningless result.
- **Interval:** The difference between the values-numbers is meaningful but the scale lacks a real origin. The average summer temperature, measured in degrees of Celsius is an example. The interval scale can have also negative values.
- **Ratio:** On this scale, the difference between the values is important and the measurements can have an absolute zero value. The water level is a graspable example. Water level can take negative value and information about the data relations is known, for instance, an area that is located 20m above the water level has a higher altitude than an area located -10m beneath the water level.

In addition, **count** can be considered as a special case of the Ratio measurement scale, where integer numbers represent count of a core concept. For instance, the number of votes during elections will be counted. Since the individual objects are counted, this scale can take values equal to zero or positive (Chrisman, 1998).

2.3 Core concepts in geo-analytical question-answering systems

As it has been mentioned previously, core concepts are crucial for the formation and interpretation of questions. In the following example questions (Fig. 3), I recognize and annotate core concepts of geo-analytical questions, to make it easily observable how the core concepts can be identified in a question:



What is *density (Proportion)* of *residential areas (Field)* for each *neighborhood (Object)* in Utrecht (Object) ?

Which buildings (Object) are affected by hurricane (Event) in Japan (Object)?

Which city (Object) has highest crime rate (Proportion IRA) in France (Object)?

Figure 3. Recognized and annotated core concepts in 3 geo-analytical question

The above example questions show that nouns or noun phrases in a geo-analytical question can be interpreted as core concepts and their corresponding measurement levels. For each core concept category, there are relevant GIS processes, which are used to study it. These processes, receive as inputs and produce as outputs similar annotations (core concepts). For example, as it is illustrated in Figure 4, in GIS, the Euclidean distance tool receives an object input and outputs a distance field. In the same manner, the Zonal statistics tool receives an object input and converts it into an object content amount (Xu et al., 2022).

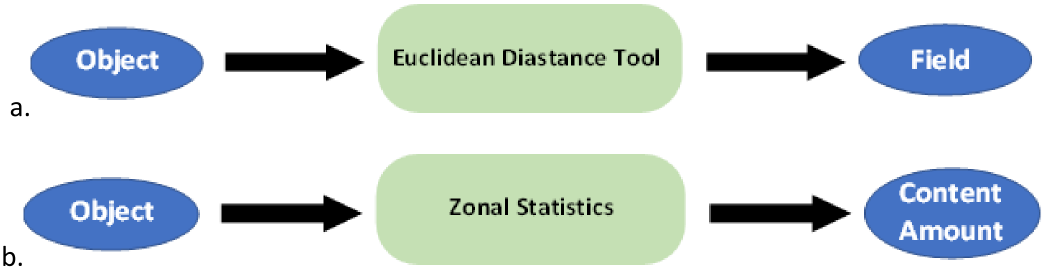


Figure 4. GIS tools with their corresponding inputs and outputs. a. Euclidean Distance tool, b. Zonal Statistics (modified image from Xu et al., 2022)

Based on this commonality between the question components and the GIS processes, it is possible to feed the components of a question as inputs in GIS workflows (multiple GIS processes). These inputs will be analyzed with the appropriate GIS workflows and the outputs from this analysis will be the answers to the geo-analytical questions (Xu et al., 2022).

2.4 NER and Deep Learning models

Automatic named entity recognition (NER) has been a widely researched topic nowadays, with a large number of researches, devoted to the development and improvement of NER methods and tools (Liu et al., 2022; Li et al., 2022; Van et al., 2021; Jin et al., 2019; Qiu et al., 2019). The term NER refers to an NLP task, used to locate and classify accurately and correctly named entities in a text (Li et al., 2022; Van et al., 2021; Martins et al., 2008). The first official NER system was introduced by Grishman and Sundheim (1996) at the Sixth Message Understanding Conference (Yadav & Bethard, 2019; Marrero et al., 2013; Grishman & Sundheim, 1996). Grishman and Sundheim (1996) proposed a NER system that recognized PER (person), ORG (organization) and LOC (location) (Grishman & Sundheim, 1996). Since then, many studies on different scientific domains such as medicine (Wen et al., 2021; Bose et al., 2021; Wang et al., 2018; Finkel et al., 2004), geology (Qiu et al., 2019;) and geography (Zhao et al., 2018; Ortega et al., 2009) have focused on developing methods and tools for NER tasks.

Rule-based, statistical and deep learning (DL) are the three well-known methods for NER tasks (Li et al., 2022; Van et al., 2021). However, the most popular one today, among these methods, are the deep learning methods, as they do not require extensive human labor for feature engineering and extensive additional resources (Li et al., 2022; Jin et al., 2019). Deep learning has gained a lot of research attention in recent years because it has presented new techniques for solving NLP challenges. Due to the drawbacks of feature engineering, DL has been recommended as a valuable methodology, for automatic learning, deep feature mining and allocated representation of words. Deep neural networks are employed in DL, to substitute classical machine learning's feature engineering (Li et al., 2022; Jin et al., 2019). Their major privilege is their capacity for end-to-end learning. It signifies that the network can learn sequence labeling rules from a pre-labeled dataset without the need for human interference (Van et al., 2021).

2.5 NLP and NER in the Geoscience domain

Many studies have presented the benefits of the NLP methods implementation and tools in the geoscience domain and more specifically in the GIS domain (Perea-Ortega et al., 2013; Lampoltshammer & Heistracher, 2012; Calì et al., 2011). According to the work of Calì et al., (2011), an NLP-based GIS interface has multiple advantages compared to a traditional GIS interface, as it can make GIS more approachable and usable to people with no previous knowledge about the domain. In the same direction, Lampoltshammer & Heistracher, (2012) analyzed the impact of the NLP exploitation on three selected research domains from GIS literature, human-computer interaction, geographic information retrieval (GIR) and location-based services (LBS). In addition, several NLP query reformulation methods linked to the alteration and enlargement of both thematic and geospatial aspects which are often identified in a geographical query have been presented by Perea-Ortega et al. (2013).

Moreover, previous research has shown that NER techniques can be implemented for the recognition of geographical, geological, and spatial entities. Perea-Ortega et al., (2009) presented a system called Geo-NER, which was used for the detection and recognition of geographic name entities. Geo-NER was built on a generic entity tagger, which has been supplemented with Wikipedia-generated geographic resources. However, the ability to consider geographic data from other sources to help in the recognition of places from text was lacking (Perea-Ortega et al., 2009). In other work, NER systems have been also utilized to detect spatial relationships between places and to identify location properties (Lima & Davis, 2017).

In addition to this, NER systems have been implemented to improve the pre-existing techniques in the GIR (geographic information retrieval) domain (Acheson & Purves, 2021; Buscaldi & Rosso, 2009). In the research of Buscaldi and Rosso (2009), a NER system was employed to find location names (toponyms). After the toponym was found, the corresponding coordinates were added by their system (Buscaldi & Rosso, 2009). Similarly, Acheson and Purves (2021) exploited a NER system to extract location names from scientific articles and represent spatially these locations by geocoding them (Acheson & Purves, 2021). Nevertheless, these studies have focused on the use of NER systems for the recognition of geographic places (toponyms), while the present research's focus is on geographic entity types (names for geographic categories, e.g., distance networks, temperature fields, etc.)

Nowadays, the implementation of deep learning algorithms in geoscience has gradually revealed the advantages of this technology for the domain. Newly proposed NER systems exploited the benefits of deep learning methods, for the recognition of geological name entities. Qui et al., (2019) addressed the GNER (Geological NER) issue in the geoscience domain and presented a detailed framework of how the GNER which incorporates DL methods can be extended through the fine-tuning process in other scientific domains (Qui et al., 2019). Moreover, some researchers went a step further and applied particular deep learning techniques (BERT and LSTM) for the extraction of geographic information. Specifically, Shin et al., (2020) introduced a BERT-based spatial information extraction model for spatial information extraction and an R-BERT model for the extraction of spatial relationships. Correspondingly, the LSTM model has been proved a useful tool for the recognition of geographical information and the identification of spatio-temporal relations between different Points-of-Interest (POI) (Zhao et al., 2018).

Recently, the importance of spatial core concept recognition for the interpretation of geo-analytical questions was evaluated in the research of Xu et al., (2022). In this study, the authors proposed a question parsing method to convert geo-analytical questions to core concept expressions, which conform to GIS workflows. Beyond that, the authors suggested in their work, that deep learning approaches such as the Bi-LSTM and BERT could be used to train NER models which recognize automatically, different and complex core concepts in questions (Xu et al., 2022).

However, none of the aforementioned studies have applied explicitly deep learning methods and tools for the automatic recognition of core concepts of spatial information in GIS-related texts. Following the suggestion of Xu et al., (2022), this study attempts to evaluate and compare the performance of the two deep learning algorithms (BERT and Bi-LSTM) for NLP tasks in the geoscience domain, in this direction.

3. Data

In this section, the source, format and the pre-processing of the data used in this research are described. The data of this research consisted of two datasets: a geo-analytical question corpus and a core concept dictionary.

3.1. Geo-analytical question corpus

The geo-analytical question corpus (GeoAnQu) is a question corpus created by Xu *et al.*, (2020). The corpus initially included 429 questions in the English language, which had been generated from various sources such as GIS literature, scholarly papers and coursebooks (Xu et al., 2020). More specifically, a large number of scholarly papers was gathered during a Master's Thesis at Utrecht University using Scopus. The papers were selected based on three criteria: 1) they were in the discipline of Human Geography, 2) they contained GIS analysis, and 3) they had been published between 2009 and 2018. Additionally, GIS literature and coursebooks were searched for questions that were included in GIS tutorials and exercises. All the questions found in the GIS literature and coursebooks were included in the corpus, although some questions were reformulated as they had ambiguous meaning (Xu et al., 2020). Nevertheless, for the purpose of this study, a reformulated version of this corpus (429 questions) was used and contains 309 distinctive and explicit GIS questions (Xu *et al.*, 2022). Of these, 196 questions are from the GIS literature, 76 are from scholarly papers, and the others are from GIS coursebooks (Xu *et al.*, 2022). A percentage (90%) of these questions will be used as the training dataset for the BERT and Bi-LSTM models, presented in this study.

ID	Source	Authors	Year	Title	Page	Question
1	Competency questions	Haiqi, Nyamsuren	2019	IAOA Summer Institute on Places and Things		Which houses are for sale in Utrecht
227	Journal of Urban Planning and Development 144(4),04018047	Romanillos, GarcÃƒÂa-Palomares	2018	Accessibility to schools: Spatial and social imbalances and the impact of population density in four European cities		What is the network distance to primary schools for children aged between 4 and 12 in Multifunctional Urban Area of Rotterdam
483	GI Minor	Simon	2019	GI Minor course		What is the proportion of people over 65 for each PC4 area in Amsterdam
499	GIS textbook	Allen	2013	GIS tutorial 2: spatial analysis workbook	P38	What is the number of Hispanics for each census block in Tarrant County in Texas
520	ArcGIS Pro/Tools	esri		ArcGIS Pro Multi-Distance Spatial Cluster Analysis (Ripley's K Function)		What is the degree of clustering of 911 calls for each distance band from 300 to 900 meters by 60 meter increments in Portland
542	Master thesis	Romay	2020			What is the percentage of population between 16 and 24 years to the total

					population per neighborhood in Amsterdam
563	Proceedings - 2011 19th International Conference on Geoinformatics, Geoinformatics 2011 59801000	Wang, Wu & Yu	2011	Analyzing spatio-temporal distribution of crime hot-spots and their related factors in Shanghai, China	Where are the hot spots and cold spots of thefts in Shanghai in December 2009

Table 1. Geo-analytical question corpus structure

3.2 Core concept dictionary

The core concept dictionary (CCD) is a dataset proposed by Xu *et al.* (2022) and consists of keywords- core concepts that are related to spatial phenomena. The CCD was generated by analyzing the corpus (GeoAnQu) nouns and noun phrases and manually annotating such nouns and noun phrases with core concept types (CCT) and measurement levels (MSRLV). The CCD contains 11 different core concept types (e.g., 'object', 'field', 'event', 'network', 'proportion', 'amount', 'location' etc.) and 9 measurement levels (e.g., 'interval', 'nominal', 'count', 'ratio', 'boolean', 'ordinal' etc.), which have been used to create the core concept IOB tags for the training of the BERT and Bi-LSTM model.

Tokens	CCT	MSRLV
house	object	
urban areas	field	nominal
construction year	object quality	interval
population	conamount	count
traffic flow	network quality	ratio
accidents	event	
percentage	proportion	ira
network based path	network	

Table 2. Example tokens with their corresponding core concept type and measurement levels from the CCD

3.3 Data Preprocessing

3.3.1 Tokenization and Part of Speech (POS)

A basic preprocessing step was the tokenization of the 309 questions from the GeoAnQu corpus. Tokenization is the process of dividing a raw text into individual words, called tokens. This process is important for the meaning interpretation of the text through analysis of the sequence of the words in the text (Webster & Kit, 1992). In this study, tokenization allowed the easier comparison between the tokens of each question and the keywords-nouns included in the CCD and in extend the generation of the appropriate IOB tag for each token. In addition, the tokens of each question along with their corresponding IOB tags were used as the train dataset for the training process of both DL models. Figure 5 shows an example question before and after the tokenization step. The final total number of tokens was 4096.

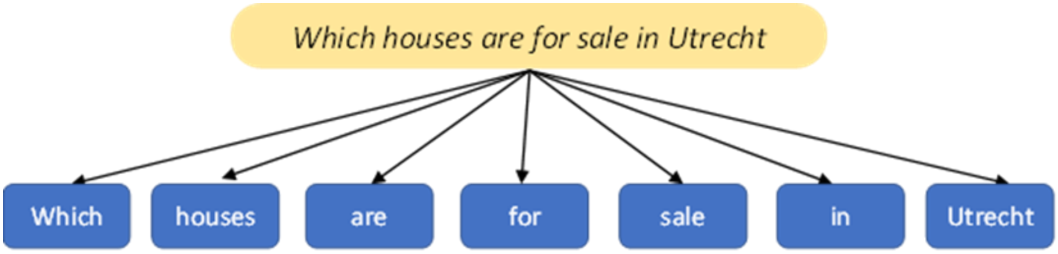


Figure 5. Example question before and after tokenization

Additionally, to the tokenization, the Part of Speech (POS) tags of the tokens was generated. In general, the POS tags, facilitate the comprehension of human language for the NLP tasks (Chiche & Yitagesu, 2022). In this study, POS tags are used to distinguish the different meanings of the same words in several questions. For example, the word “sale” can be a noun or a verb in a sentence, according to its location. By knowing the syntactic structure around each word (e.g., determiners and adjectives come before nouns and verbs after nouns) the NER models can detect named entities in texts (Schmid, 1994; Cutting et al., 1992). Figure 6 presents the tokens of a question along with their generated POS tags. The final total number of POS tags was also 4096.

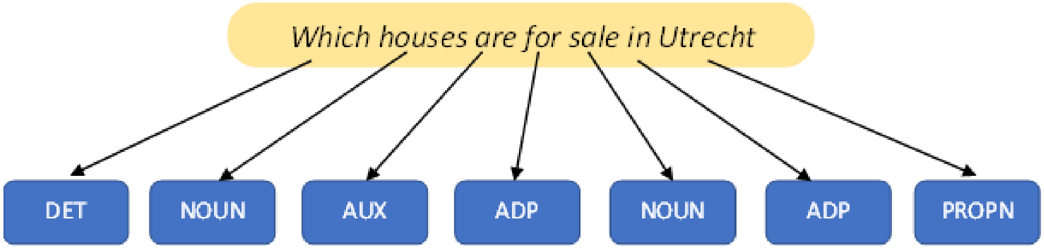


Figure 6. Example question before and after POS tags generation

3.3.2 IOB2 Tagging

For every Named entity recognition (NER) task the use of a particular annotation scheme at the word level is important (Alshammari & Alanazi, 2021). Among various annotation schemes (e.g., IOB, IOE, BIIE, BIES, BILOU), the one which is commonly used with natural language processing (NLP) models such as BERT or Bi-LSTM, is the IOB-tagging scheme (Alshammari & Alanazi, 2021; Hwang et al., 2021; Luoma & Pyysalo, 2020; Hakkani-Tür et al., 2016). The name of the scheme (IOB) denotes whether the corresponding word-token is located, Inside(I), Outside(O), or at the Beginning(B) of a particular name entity (Alshammari & Alanazi, 2021).

The IOB tags of the nouns or noun phrases (tokens) of the GeoAnQu corpus were constructed by combining the core concept type and measurement level of each noun or noun phrase from the CCD in a unified form (core concept type + measurement level) (Table 3). For example, if a noun has as core concept type ‘field’ and measurement level ‘nominal’ the unified tag has generated as ‘FLDN’.

Tokens	U_Tag	Core Concept Type	Measurement Level
house	OBJ	Object	
construction year	OBJQI	Object Quality	Interval
land use	FLDN	Field	Nominal
hurricane	EVE	Event	
driving time	NETQR	Network Quality	Ratio
population	CNAC	Content Amount	Count
ozone concentration	FLDR	Field	Ratio
temperature	FLDI	Field	Interval
shortest path	NET	Network	
mean direction	CVAL	Coverage Amount	Location
severity	EVEQO	Event Quality	Ordinal
total area	CVAER	Coverage Amount	Extensive Ration Level (ERA)
income	CNAER	Content Amount	Extensive Ration Level (ERA)
crime rate	PROPIR	Proportion	Intense Ratio Level(IRA)
noise level	FLDO	Field	Ordinal
walkability	OBJQR	Object Quality	Ratio
wind speed	EVEQR	Event Quality	Ratio
political leaning	OBJQN	Object Quality	Nominal
locations	LOC	Location	
for sale	OBJQB	Object Quality	Boolean
rating	OBJQO	Object Quality	Ordinal

Table 3. Examples of unified tags, core concept types, measurement levels and their respective tokens

Afterward, for the automatic generation of the appropriate IOB2 tagging format (i.e. B-tag or I-tag), a python script was written. This script compares each token from the GeoAnQu corpus with the tokens in the CCD. If the tokens are the same (e.g. houses-houses) then the algorithm puts the corresponding U_Tag in the noun or noun phrase. However, according to the IOB2 tagging scheme, for the single nouns-tokens the tag has the following format 'B-U_Tag' while for the noun phrases', the first token's tag is formatted as 'B-U_Tag' and the rest tokens' tags of the phrase, are formatted as 'I-U_Tag', regardless of its length, i.e., a noun phrase consisted of two, three, four or even more words-tokens (Table 4).

ID	question	token	pos	tag
1	What is crime density within buffer area of shortest path from home to workplace in PlaceName0	What	PRON	O
		is	AUX	O
		crime	NOUN	B-PROPIR
		density	NOUN	I-PROPIR
		within	ADP	O
		buffer	NOUN	O
		area	NOUN	O
		of	ADP	O
		shortest	ADJ	B-NET
		path	NOUN	I-NET
		from	ADP	O
		home	NOUN	B-OBJ
		to	ADP	O
		workplace	NOUN	B-OBJ
		in	ADP	O

		Amsterdam	PROPN	O

Table 4. Example question of the training dataset after the completion of preprocessing

The final number of name entities with generated IOB tags, based on the above format (B-tag, I-tag excepting ‘O’) was 1228. The code and methods implemented for the data preprocessing (tokenization, POS tags and IOB tags generation) can be found on GitHub: https://github.com/AristotleKandylas/GIS-NER-ADS-Thesis-Code

3.3.3 Training and Test dataset

For the comparison of BERT’s and Bi-LSTM’s performance, the models were trained on the same percentage of training set and evaluated on the same percentage of validation/test set. Since the training dataset (corpus) contains a limited number of questions, the dataset was randomly split into train and test sets by using 90% of the corpus (278 questions) as the train set and 10% (31 questions) as the test set for the two models.

4. Methods

4.1 BERT

4.1.1 Introduction

The BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model was chosen in this study, for the name entity recognition task of geographical core concepts, mainly due to its satisfactory results and state-of-the-art performance in NER tasks. It is a recently developed and advanced deep learning model, compared to RNNs and CNNs, which is based on transformers machine learning techniques. BERT has been designed for the pretraining of deep bidirectional representations from unlabeled data, by learning both left and right context simultaneously across all layers. This model’s architecture is ideal for token classification tasks and allows to make predictions at the token or sequence level. One important advantage of BERT is the use of pre-trained word embeddings as inputs, which can be further fine-tuned or kept fixed during the training process of a NER model (Li et al. 2020; Devlin et al., 2019).

Here, the BERT model is applied on token level, to predict the IOB tags of the core concepts included in geo-analytical questions. The implementation of BERT includes two steps: pre-training and fine-tuning (Fig. 7) (Devlin et al., 2019).

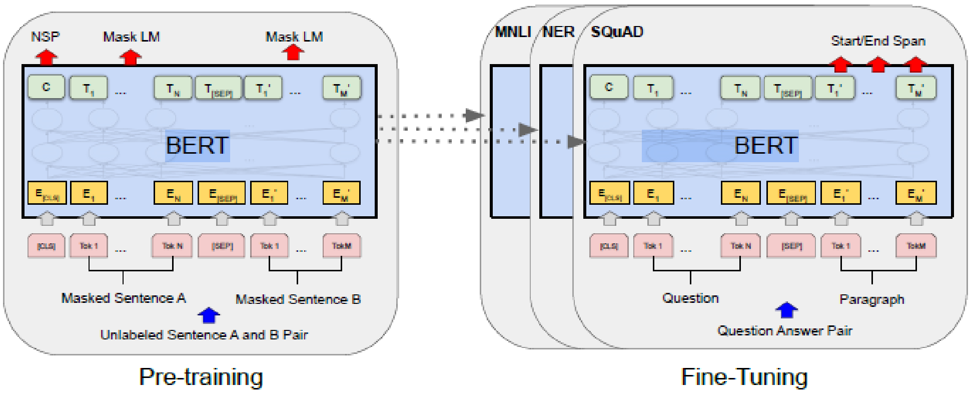


Figure 7. The two steps of BERT model implementation (Devlin et al., 2019).

However, for the implementation of BERT in this study, the model was not pretrained but was fine-tuned only. Instead, the 'bert-base-cased' pre-trained model was used, to fine-tune the model on the NER training dataset (corpus). The 'cased' version was chosen, in order the model to be case-sensitive to the text of every question, previously to the tokenization step e.g., 'Houses' ≠ 'houses'. In this manner, the existence of different cases in the text was taken into consideration.

The 'BERT-base' model was already pretrained in an unsupervised way on larger corpora (Wikipedia and BooksCorpus), based on the process presented by Devlin et al., (2019), on 2 tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). During the first task, a [MASK] token was used to replace ~15% of the words in each sequence, previously to word sequences feeding into the model. Then, based on the context supplied by the rest of, the non-masked words in the sequence, the model tried to predict the original value of the masked token. In the second task, the model was trained, by receiving pairs of sentences as input, to predict whether the second sentence in a pair is the next in the original document. In the training step, 50% of the inputs were in pairs for which, each second sentence was the next sentence in the original document, while the other 50% was a random sentence from the corpus. It was assumed that the random sentence would be detached from the first sentence. Before feeding the model, the input was treated in the following manner (Fig. 8) to assist the model in differentiating between every two sentences. Firstly, at the start of the first sentence, a [CLS] token was inserted, and at the end of each sentence, a [SEP] token was placed. Then, each token-word had a sentence embedding which indicates whether it was Sentence A or Sentence B. Sentence embeddings are similar to token embeddings in concept, but with a vocabulary of two (Sentence A and B). Finally, each token was given a positional embedding to denote its location in the sequence (Devlin et al., 2019).

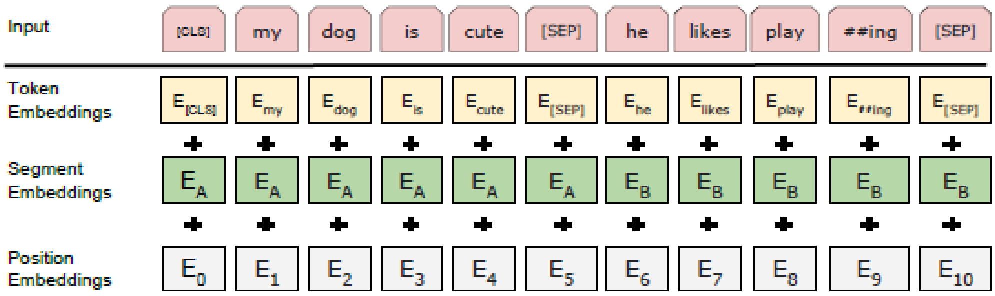

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 8. BERT input representation (Devlin et al., 2019)

4.1.2 BERT Fine-Tuning

In the scope of the presented experiments, the 'bert-base-cased' model was fine-tuned (supervised learning step) for the required core concept recognition task, using a NER training dataset that contains labeled data (i.e., IOB tags) (Devlin et al., 2019). The training dataset was split into sentences and tokenized at word level. These were the inputs for the fine-tuning process.

For computational reasons, a maximum sequence length of 128 was used to incorporate maximum cross-sentence context and the simple version of the Adam optimizer with a learning rate (5e-05) was employed during the model's tests. For the training and validation, a batch size of 2 was chosen, as the tests showed that smaller numbers of batch sizes, improved the performance of the model. The maximum gradient normalization value was set at 10, to prevent the "exploding gradients" problem and no warmup was used over the first training epoch. The model was trained for 4 epochs and evaluated after every epoch on the validation set using entity-level precision, recall and F1-scores. The best-performing checkpoint is used as the final prediction model.

4.1.3 BERT Hyperparameter Tuning

The best hyperparameters values for the fine-tuning process of our BERT model were selected, by using as a reference, the grid search presented in the study of Devlin et al., (2019). Specifically, the learning rate, batch size and epochs values in section 4.1.2, were selected after tuning manually the hyperparameters of the BERT model, using the following values presented in Table 5:

Hyperparameters	Fine- Tuning Values	Optimal Values
Learning rate	5e-5, 3e-5, 2e-5	5e-5
Batch size	2, 4, 8, 16, 32	2
Epochs	2, 3, 4, 6, 8, 10	4

Table 5. Optimized parameters of BERT model

A total number of 90 trials were run, one for each combination of hyperparameters. The best hyperparameter values were chosen, based on the validation accuracy, F1-score, recall and precision of the model after each trial, which was assessed with a python script from the *seqeval metrics* evaluation package. Afterwards, these optimal values were used for the fine-tuning of the presented model in this study.

4.2 Bi-LSTM

4.2.1 Introduction

In comparison to the BERT model, the Bi-LSTM is an earlier developed model, based on a recurrent neural network. According to previous research, it has proven its capability on NLP and more specifically on NER tasks. Bi-LSTM can achieve high accuracy on POS, chunking, and NER datasets without relying much on word embedding like other RNN models (Le et al., 2018; Qin & Zeng, 2018; Huang et al., 2015). Furthermore, recent research has presented that, for a small dataset, Bi-LSTM models outperform BERT models substantially and can be trained in considerably less time than fine-tuning a pre-trained model (Ezen-Can, 2020).

Although the Bidirectional Long-Short Term Memory (Bi-LSTM) is used in this paper for the same recognition and classification task of core concepts like BERT, it presents some differences in its architecture from the latter. The Bi-LSTM is a model architecture in which the previous (backward) and future (forward) sequence information is used in the output layer (Fig. 9) (Sun et al., 2018; Shahid et al., 2020).

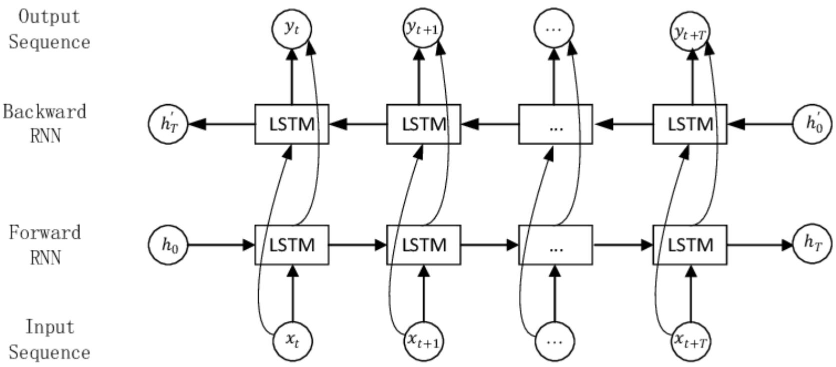


Figure 9. Bi-LSTM network architecture (Xiang et al., 2020)

In the Bi-LSTM the input information moves in two directions, differentiating it from the original LSTM. In the original LSTM, the input can move in one direction, either forwards or backwards (Fig. 10). This difference makes the Bi-LSTM model's implementation, ideal for sequence-to-sequence tasks, such as speech recognition, text classification and forecasts (Shahid et al., 2020; Sun et al., 2018).

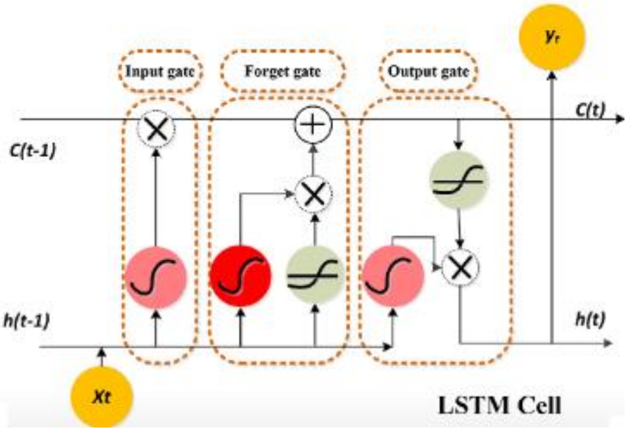


Figure 10. Single LSTM cell architecture (Shahid et al., 2020)

4.2.2 Bi-LSTM Architecture

In the present study, the Bi-LSTM is applied for text classification- NER task. The implementation of Bi-LSTM as with any other neural network requires the design of the network architecture and the definition of the input and output dimensions for every layer. For our recognition and classification task, the many-to-many architecture was chosen as there are multiple inputs and outputs in our model. In this architecture, three layers (the Embedding, Bi-LSTM, and LSTM layer) are considered and the final (4th) layer is a TimeDistributed Dense layer, which is used to output the final predicted tags. In Figure 11, the input and output dimensions of each layer can be seen.

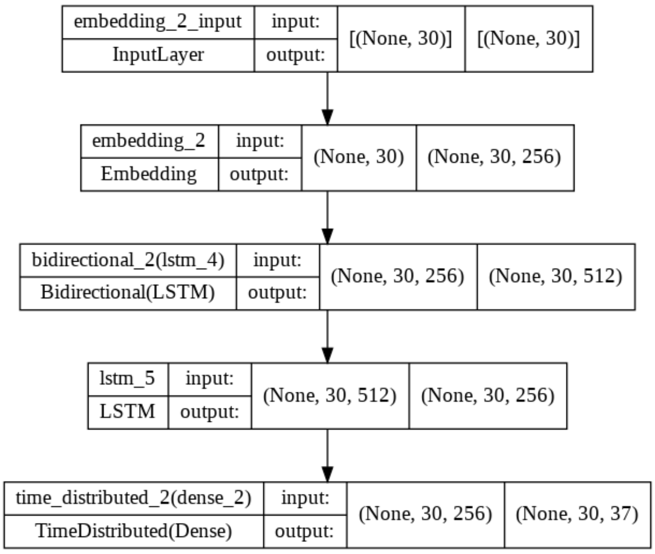


Figure 11. Input and output dimensions of each layer of Bi-LSTM model

The first layer in our neural network is the embedding layer (Fig. 11). In this layer, the maximum length (30) of the padded sequences is specified. It is important to mention that the first dimension (None) in the plots represents the batch size. In our case the batch size is not specified, that's why the plot shows None in the first dimension. After the network's training, the embedding layer converted each token into an n-dimensional (n=256 in our case) vector. Then in the second layer (Bidirectional LSTM), a recurrent layer (e.g., the 1st LSTM layer) was used as a parameter in the Bi-LSTM. The output from the preceding embedding layer was employed in this layer (30, 256). Provided that, this was a Bi-LSTM, it had both forward and backward outputs. There are several ways (multiplication, summation, concatenation and average) to amalgamate these outputs before feeding them to the next layer. In this network, the outputs were concatenated, which doubled (512) the outputs to the next layer. Afterwards, the 3rd layer, which is the LSTM layer, receives the output dimension (None, 30, 512) and outputs (None, 30, 256) from the preceding Bi-LSTM layer. To avoid a single final output in our network, the TimeDistributed layer was added. This layer guaranteed that the Dense (fully connected) operation, will be implemented on each output on each time step. This layer received the preceding LSTM layer's output dimension (None, 30, 256) and outputs the maximum sequence length (30) and the total number of tags (37).

Following the definition of the neural network's architecture, the model was trained on the train set and its performance was evaluated on the corresponding test set. The experiments were run for a different number of epochs and batch sizes. The model which combined high performance and fewer iterations were chosen, as the ideal one for our purpose. According to the experiments, training the model for 4 epochs with batch size 2 resulted in the best performance.

4.2.3 Bi-LSTM Hyperparameter Tuning

In order to select the best parameter values for our Bi-LSTM neural network, the hyperparameters of our model were tuned, using the Bayesian optimization technique. This technique selects optimal hyperparameters by learning from previous trials-evaluations. Being an informed learning method means that, it utilizes mostly, values from the parameter space which can lead to a better (accuracy, loss, etc.) model in the next trials (Aslam et al., 2021; Frazier, 2018). The following table (Table 6) presents the values of the hyperparameters which were optimized for the proposed Bi-LSTM model. Five parameters have been tuned using the Bayesian optimization method. The table also displays the optimal hyperparameter values, after the tuning process.

Hyperparameters	Tuning Values	Optimal Values
Activation	relu, softmax	softmax
Dropout	0.0, 0.1, 0.2, 0.4	0.4
Learning rate	1e-2, 1e-3, 1e-4	1e-2
Recurrent dropout	0.0, 0.1, 0.2, 0.4, 0.5	0.0
Units	Min:32, Max: 512	256

Table 6. Optimized parameters of the Bi-LSTM model

A total number of 10 trials were run for 20 epochs each one of them. The best hyperparameter values were chosen, based on the validation loss of the model after each trial. Afterwards, the optimal values were used for the construction of the presented model in this study.

4.3 Evaluation

After the training step, the two models' performance is evaluated on their predictions on 31 randomly selected questions, which are included in the test set. The evaluation of the prediction-recognition performance of both models was based on the precision, recall, F1-score and validation accuracy metrics. According to Grishman and Sundheim (1996), for each NER category, the *precision* was defined as the number of accurately predicted entities divided by the number of predicted entities by the system.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

The *recall* was defined as the number of entities accurately predicted by the system divided by the number correctly recognized by humans.

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

The *F1-score* was defined as the harmonic mean of the model's precision and recall (Grishman & Sundheim, 1996).

$$\text{F}_1\, score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Lastly, the *accuracy* of the models was defined as the ratio of accurately predicted entities divided by the total number of entities (Halteren et al., 2001).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

The above equations are calculated on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) which are described in detail as follows (Li et al., 2022):

- TP: the entities that are identified by NER models and match ground truth.
- TN: the entities which are not identified or completely missed by the NER models.
- FP: the entities that are identified by NER models but do not match ground truth.
- FN: entities annotated in the ground truth that are not identified correctly by NER models.

The code and methods implemented in the methods section is available under open licenses on GitHub: https://github.com/AristotleKandylas/GIS-NER-ADS-Thesis-Code

5. Results

With the completion of the experiments on the two deep learning models, useful conclusions can be drawn regarding their performance. The precision, recall and F1-score for the models trained solely on the training dataset, using the best hyperparameters, are presented in Table 7.

Model	Accuracy	Precision	Recall	F1-score
BERT	0.9522	0.78	0.76	0.77
Bi-LSTM	0.9391	0.41	0.46	0.43

Table 7. Models' performance according to the 4 evaluation metrics on validation/test dataset

According to the accuracy metric, we can conclude that the BERT model outperforms the Bi-LSTM model in this spatial core concept recognition task. However, the difference between the models is not significant, as BERT's accuracy is 95% and Bi-LSTM's is around 94% (Table 7).

Although accuracy is a very good first metric to evaluate the performance of the two models, this metric cannot guarantee exclusively, that the model with the higher accuracy is the best one especially if the dataset is not balanced-symmetric. In the training dataset, there are tags (Fig. 12) such as FLDO, FLDI, OBJQN, CVAER, OBJQI, etc., which are under-presented as each one of them corresponds to 1-6 keywords, while other tags such as OBJ, EVE, FLDN, PROPIR, etc. are over-presented in the dataset, as each tag corresponds to more than 29 keywords.

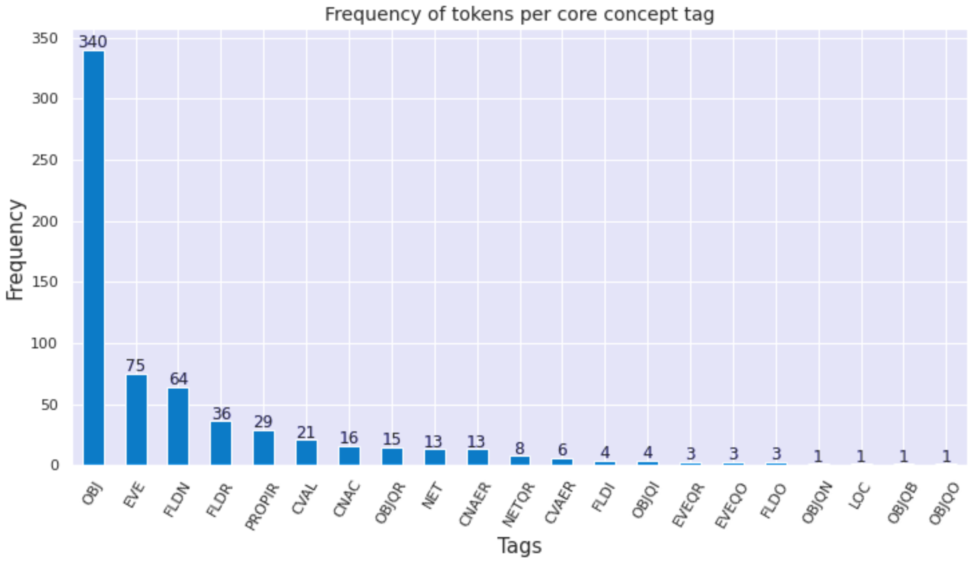


Figure 12. Frequency of tokens per core concept tag in the CCD

Given the uneven class distribution in the training dataset and taking into consideration that the recognition and tagging of the core concepts is a multi-label problem, the other three measures (Precision, Recall and F1-score) in Table 7, provide a more accurate and comprehensive picture, regarding the performance of the two models.

By comparing the precision, recall and F1-score of the two models, it is observable that BERT exceeds the performance of the Bi-LSTM model in all three metrics. More specifically, the overall precision of BERT is 0.78, while the precision of Bi-LSTM is 0.41, conveying that the BERT performs better, as its precision is closer to the optimal value of 1. The higher precision of BERT is also related to a low false positive rate, meaning that in the case of the Bi-LSTM model, the user might lose correctly predicted core concepts. Correspondingly, the recall (0.76) of the BERT model is higher than the one of Bi-LSTM, representing a good value as it is above 0.5. On the contrary, the recall of the Bi-LSTM model is lower (0.46) than 0.5, denoting that out of the core concepts which should have been labeled correctly, only a limited number of core concepts have been tagged correctly or not tagged at all.

Despite the importance of the aforementioned measures for the evaluation of the models' performance, F1-score is more useful, especially in our case, where there are imbalanced classes in the training dataset, as it considers as a metric, both the false positives and false negatives. According to the experimental results, the F1-score of the BERT model (0.77) is higher, compared to the F1-score of Bi-LSTM (0.43). Provided that, F1-score makes more balanced predictions than the precision and recall metrics, it proves that the BERT model can predict more correct positive results in the experiments than Bi-LSTM, supporting robustly the better accuracy result of the BERT model and its superior performance. The detailed efficiency, based on the three previous measures, of each model on the IOB tags recognition task, can be seen in the next two tables, Table 8 and Table 9 :

IOB TAGS	Precision	Recall	F1-score	Support	Total number of tag instances
CNAC	0.75	0.75	0.75	4	5
CVAL	1	1	1	3	12
EVE	0.89	0.89	0.89	9	15
FLDI	0	0	0	0	1
FLDN	0.67	0.86	0.75	7	16
FLDR	1	0.86	0.92	7	9
NETQR	1	1	1	1	2
OBJ	0.76	0.7	0.73	37	54
OBJQB	1	1	1	1	2
OBJQR	0	0	0	2	3
PROPIR	0.75	0.75	0.75	4	8
micro avg	0.78	0.76	0.77	75	127
macro avg	0.71	0.71	0.71	75	127
weighted avg	0.79	0.76	0.77	75	127

Table 8. Precision, Recall and F1-score of BERT model per recognized IOB tag on the validation dataset

IOB TAGS	Precision	Recall	F1-score	Support	Total number of tag instances
CNAC	0	0	0	0	5
CVAL	0.67	0.25	0.36	8	10
EVE	0.43	0.43	0.43	7	8
FLDI	0	0	0	0	1
FLDN	0.11	0.09	0.1	11	16
FLDR	0	0	0	7	7
NET	0	0	0	0	2
NETQR	0	0	0	0	10
OBJ	0.79	0.64	0.71	47	52
OBJQB	0	0	0	0	2
OBJQR	0	0	0	0	1
PROPIR	0.2	0.1	0.13	10	9
micro avg	0.46	0.41	0.44	90	123
macro avg	0.18	0.13	0.14	90	123
weighted avg	0.54	0.41	0.46	90	123

Table 9. Precision, Recall and F1-score of Bi-LSTM model per recognized IOB tag on the validation dataset

It should be clarified that only the included tags in the validation dataset are presented in the above tables. Therefore, the rest tags which are not present in the two tables were not included during the splitting in the validation set. One important conclusion from the two tables is that both models were able to recognize the OBJ and EVE tags efficiently. The main reason for this result is that these two tags appear with high density in the training dataset.

Furthermore, the training and validation loss graphs (Fig. 13). of the BERT and Bi-LSTM model, confirm the high effectiveness of the BERT to fit the training data and perform well on the validation set as its training and validation loss curve begins at a lower point (0.2686 and 0.0985) than the ones of Bi-LSTM model (0.7044 and 0.4513). However, the loss curves of both models decrease by the number of epochs until the point where they are closer to each other. This illustrates that both models behave correctly as their losses are reduced after each epoch without presenting overfitting or underfitting.

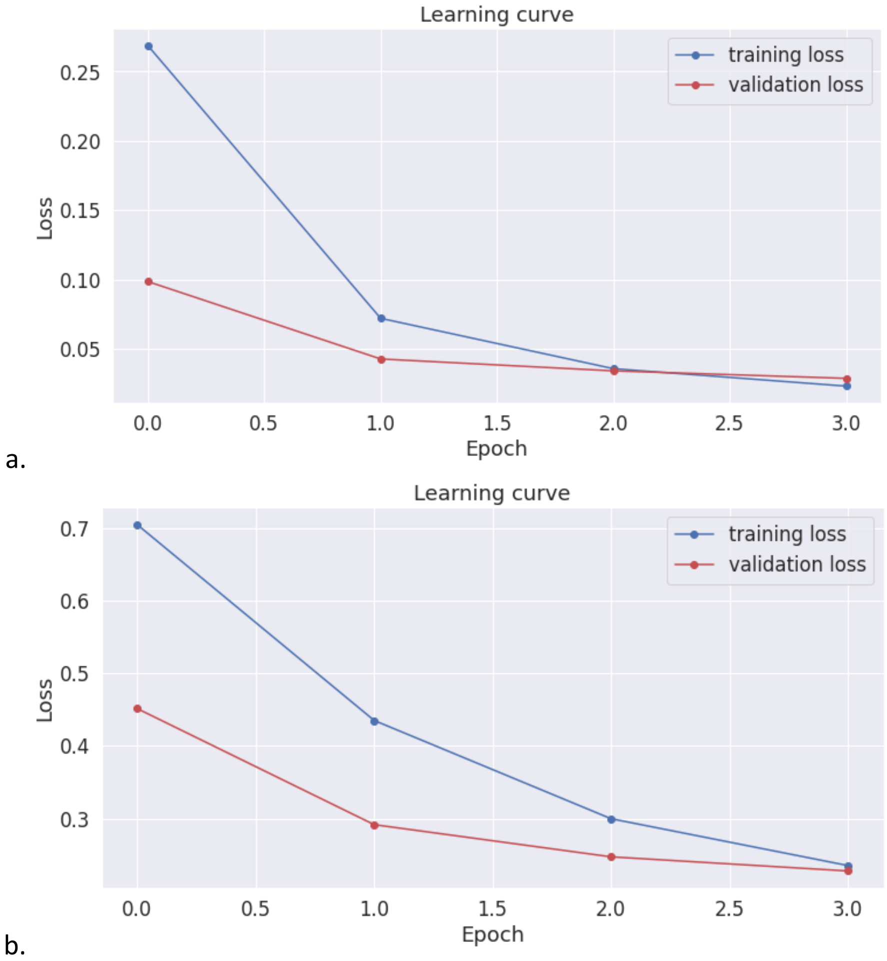
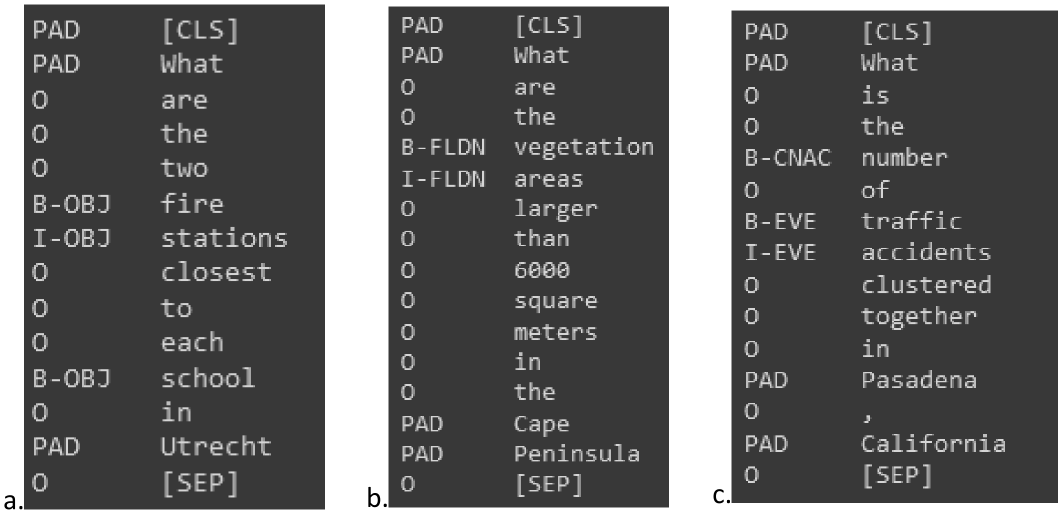


Figure 13. The training and validation loss of: a. BERT model and b. Bi-LSTM model

Finally, the BERT-based NER model was evaluated on three selected new questions, from another geo-analytical question corpus proposed by Xu, et al. (2022). As shown in Figure 14 a, fire stations and school are correctly recognized as objects. Also in Figure 14 b, the model correctly captures the vegetation areas as field nominal. Finally, as it is presented in Figure 14 c, the model recognizes accurately in the third question, the number as conamount count and the traffic accidents as events.



a.

```
PAD      [CLS]
PAD      What
O        are
O        the
O        two
B-OBJ    fire
I-OBJ    stations
O        closest
O        to
O        each
B-OBJ    school
O        in
PAD      Utrecht
O        [SEP]
```

b.

```
PAD      [CLS]
PAD      What
O        are
O        the
B-FLDN   vegetation
I-FLDN   areas
O        larger
O        than
O        6000
O        square
O        meters
O        in
O        the
PAD      Cape
PAD      Peninsula
O        [SEP]
```

c.

```
PAD      [CLS]
PAD      What
O        is
O        the
B-CNAC   number
O        of
B-EVE    traffic
I-EVE    accidents
O        clustered
O        together
O        in
PAD      Pasadena
O        ,
PAD      California
O        [SEP]
```

Figure 14. Predicted IOB tags of the BERT model on three new validation questions. a. Detection of OBJ core concepts, b. Detection of FLDN core concepts and c. Detection of CNAC and EVE core concepts

6. Discussion

This study is the first which attempts to recognize automatically core concepts from geo-analytical questions by training DL based NER model. The experimental results reveal that a more recent and advanced deep learning model such as BERT can achieve higher performance on spatial core concepts recognition task, compared to deep learning neural networks such as Bi-LSTM. The most important aspect which contributes to these results is that Transformers which are included in BERT, do not process necessarily data sequences in a particular order, as RNNs (e.g., Bi-LSTM) and CNNs do. On the contrary, they can understand the context and ambiguity of human language in any order. This enables BERT to identify easier the full context of a word-core concept in a sentence, by comparing the relation of every given word to all other words in the sentence (Devlin et al., 2019). In the present study, this becomes perceivable, as BERT can match the right IOB tags in the corresponding core concepts with the right order. For example, as Figure 14a illustrates, in the core concept *"fire stations"*, the word *"fire"* which is the begging of the phrase took the *B-OBJ* tag, while the word *"station"* is inside the phrase and took the *I-OBJ* tag.

To this end, it also contributes the fact that BERT is a pre-trained model in contrast to the Bi-LSTM. This means that there is a large amount of linguistic a-priori knowledge encoded in BERT that is lacking in the other model. This linguistic knowledge of the co-occurrence of words (which is what MLM encodes in terms of vector embeddings), is required to solve the task of ambiguous words for core concept recognition. That is why in this research BERT model achieves higher accuracy in core concept recognition compared to the Bi-LSTM model which is not pre-trained. Conversely, the RNNs networks are based on word embeddings which demand extensive training on labeled data and in the end, they still fail to recognize as accurately as BERT, the different context included in the geo-analytical questions. This is related to the fact that words in RNNs networks are defined only by pre-fixed (vectors) identities-semantics (Huang et al., 2015).

Despite the high performance of the BERT model, the approach to test and evaluate these two deep learning models encompassed some limitations. A significant limitation in this research was the restricted number of geo-analytical questions (309) from which only 4096 tokens were produced, with the tokenization process. Although according to previous research, both deep learning models can achieve satisfactory results with small datasets and in some cases Bi-LSTM model can outperform the BERT (Boudjellal et al., 2021; Souza et al., 2020; Ezen-Can, 2020), I am inclined to believe that a larger dataset with geo-analytical questions might improve the core concept identification performance of both models. Ideally, future research could investigate whether a more extensive corpus with geo-analytical questions can lead to better training and in extent to better efficiency of these proposed models.

In addition, as it has been mentioned in the result section, one serious drawback in the corpus is that some tags are over-presented while others are extremely under-presented, creating an imbalance among the tags' distribution in the training dataset. This can be noticed by looking at tables 8 and 9 where both models have the highest percentages in all the over-presented tags (e.g., OBJ, EVE, etc.) which correspond to more than 20 keywords in the dataset and fail to identify accurately all the tags which are under-presented and correspond to less than 20 keywords. The issue is more substantial in the Bi-LSTM model than in the BERT. The latter achieves to identify, with high accuracy, even tags such as NETQR and CNAC with less than 21 corresponding tokens (Table 8). At this point, it is important to clarify that during the experiments of this study, there have been attempts to overcome this limitation, by using a new version of the firstly created core concept dictionary, where particular subcategories of IOB tags (e.g., "OBJCONOBJCOVPRO", "EVECONOBJCOVPRO", "EVECONOBJCONPRO", "OBJCONOBJCONPR" etc.) which appeared limited times (i.e., 1) in the dictionary, were combined into the main category e.g., PROPIR (proportion IRA). These modifications increased the number of corresponding IOB tags in specific categories and consequently improved the training dataset. Hence, any feature studies should focus on including, a more balanced number of tags in each core concept category, by implementing the aforementioned techniques in this paragraph.

Along with the previous, one aspect deserving more study is how the prediction performance of the two deep learning models is affected if synonyms/hyponyms/hypernyms/ similar words will be used in the geo-analytical questions to replace the original words. For instance, in the question "What is the intensity of a hurricane in Texas", the keyword "hurricane" could be replaced by "storm" and the keyword "intensity" could be replaced by "wind speed". By replacing the initial keywords with synonyms, a similar question could be created, with the only difference that the new keywords (e.g., "storm" and "wind speed") might belong to new core concepts. In such a way, the keywords of under-presented core concepts will be extended in number and as a result the accuracy of the models might be increased too. Another aspect for future research would be, the use of combined and different deep learning models' architectures. For example, a combination of the two models BERT+Bi-LSTM or Bi-LSTM+CRF or a more advanced Bi-LSTM architecture which includes several layers for the model's training (Ashrafi et al., 2020; Anh et al., 2017). Towards this direction, a detailed hyperparameter tuning research for both models can also be done, to further improve their performance.

Finally, any future work in the field should focus on the appropriate evaluation of the two models' performance by using an independent validation corpus. In this study, the two models were evaluated, by separating a proportion (10%) of the questions training dataset and using it as the test-validation dataset. However, what would be interesting is to train the models in the whole training dataset and use another corpus, which includes completely new geo-analytical questions to assess the models' effectiveness. In this way, it would be possible to evaluate more accurately and explicitly the accuracy, precision, recall and F1-score of these two models in recognizing core concepts in different questions.

7. Conclusion

As the use of deep learning models becomes increasingly popular in the named entity recognition (NER) tasks it is important to evaluate their efficiency in different datasets. A NER system for geoscience texts is a fundamental step for GIS-related information extraction. This research tests and compares the performance of two popular deep learning models for the recognition of spatial core concepts from geo-analytical questions. To this end, two deep learning models were developed, a BERT and a Bi-LSTM model and trained on the same training dataset. Afterwards, their performance was compared and evaluated on a test-validation dataset which was separated by the initial training corpus. The experimental results designated that a more recent and advanced deep learning model such as BERT is significantly more effective in recognizing core concepts in geo-analytical questions, compared to an older deep learning architecture such as Bi-LSTM. Additionally, this study showed that BERT can achieve remarkable results in core concept identification, even when the size of the training dataset is relatively small and there is an imbalanced distribution of tags. However, the performance validation was conducted on a limited number of questions from the training set, and this cannot provide a comprehensive overview of the model's actual efficiency on a new question corpus. Several limitations including the previous one could be addressed and exceeded in future research. Then, these deep learning models could potentially be implemented, for geo-analytical question answering systems and in general for GIS mapping tasks.

8. Acknowledgements

With the completion of this thesis, I would like to express my deepest gratitude to my university supervisors, Professor Simon Scheider and Mrs. Haiqi Xu.

I am thankful to Mrs. Haiqi Xu, my daily supervisor, for her understanding, support and guidance through this thesis research. This research has been significantly challenging for me, as I have faced some difficult moments. Nonetheless, Mrs. Xu has provided me with her help along the way and I have managed to overcome any of the project's "obstacles".

I would like also to thank Professor Scheider, my project supervisor, for helping with any of my problems directly or indirectly through Mrs. Haiqi Xu, for providing useful feedback, answering all my questions regarding the thesis and describing any formal requirements.

9. References

- Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. PLOS ONE, 16(1), e0244918. https://doi.org/10.1371/journal.pone.0244918
- Alshammari, N., & Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. Egyptian Informatics Journal, 22(3), 295–302. https://doi.org/10.1016/j.eij.2020.10.004
- Anh, L. T., Arkhipov, M. Y., & Burtsev, M. S. (2017). Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition (arXiv:1709.09686). arXiv. http://arxiv.org/abs/1709.09686
- Ashrafi, I., Mohammad, M., Mauree, A. S., Nijhum, G. Md. A., Karim, R., Mohammed, N., & Momen, S. (2020). Banner: A Cost-Sensitive Contextualized Model for Bangla Named Entity Recognition. IEEE Access, 8, 58206–58226. https://doi.org/10.1109/ACCESS.2020.2982427
- Aslam, M., Lee, S.-J., Khang, S.-H., & Hong, S. (2021). Two-Stage Attention Over LSTM With Bayesian Optimization for Day-Ahead Solar Power Forecasting. IEEE Access, 9, 107387–107398. https://doi.org/10.1109/ACCESS.2021.3100105
- Bose, P., Srinivasan, S., Sleeman, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. Applied Sciences, 11(18), 8319. https://doi.org/10.3390/app11188319
- Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., & Dai, L. (2021). ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. Complexity, 2021, 1–6. https://doi.org/10.1155/2021/6633213
- Buscaldi, D., & Rosso, P. (2009). Using GeoWordNet for Geographical Information Retrieval. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, & V. Petras (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access (pp. 863–866). Springer. https://doi.org/10.1007/978-3-642-04447-2_113
- Calì, D., Condorelli, A., Papa, S., Rata, M., & Zagarella, L. (2011). Improving intelligence through use of Natural Language Processing. A comparison between NLP interfaces and traditional visual GIS interfaces. Procedia Computer Science, 5, 920–925. https://doi.org/10.1016/j.procs.2011.07.128
- Chen, H., Wei, B., Liu, Y., Li, Y., Yu, J., & Zhu, W. (2018). Bilinear joint learning of word and entity embeddings for Entity Linking. Neurocomputing, 294, 12–18. https://doi.org/10.1016/j.neucom.2017.11.064
- Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. Journal of Big Data, 9(1), 10. https://doi.org/10.1186/s40537-022-00561-y
- Chrisman, N. R. (1998). Rethinking Levels of Measurement for Cartography. Cartography and Geographic Information Systems, 25(4), 231–242. https://doi.org/10.1559/152304098782383043
- Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing -, 133. https://doi.org/10.3115/974499.974523
- Dai, H., Tang, S., Wu, F., & Zhuang, Y. (2018). Entity mention aware document representation. Information Sciences, 430–431, 216–227. https://doi.org/10.1016/j.ins.2017.11.032

- 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
- Enríquez, J. G., Domínguez-Mayo, F. J., Escalona, M. J., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. Expert Systems with Applications, 80, 14–27. https://doi.org/10.1016/j.eswa.2017.03.010
- Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus (arXiv:2009.05451). arXiv. https://doi.org/10.48550/arXiv.2009.05451
- Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., & Sinclair, G. (2004). Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), 91–94. https://aclanthology.org/W04-1217
- Frazier, P. I. (2018). A Tutorial on Bayesian Optimization (arXiv:1807.02811). arXiv. https://doi.org/10.48550/arXiv.1807.02811
- Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-Dependent Sentiment Classification With BERT. IEEE Access, 7, 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594
- Grishman, R., & Sundheim, B. (1996). Message Understanding Conference- 6: A Brief History. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. COLING 1996. https://aclanthology.org/C96-1079
- Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L., & Wang, Y.-Y. (2016). Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. Interspeech 2016, 715–719. https://doi.org/10.21437/Interspeech.2016-402
- Halteren, H. van, Zavrel, J., & Daelemans, W. (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. Computational Linguistics, 27(2), 199–229. https://doi.org/10.1162/089120101750300508
- Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging (arXiv:1508.01991). arXiv. https://doi.org/10.48550/arXiv.1508.01991
- Hwang, W., Yim, J., Park, S., Yang, S., & Seo, M. (2021). Spatial Dependency Parsing for Semi-Structured Document Information Extraction (arXiv:2005.00642). arXiv. http://arxiv.org/abs/2005.00642
- Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., & Wang, R. (2019). LSTM-CRF Neural Network With Gated Self Attention for Chinese NER. IEEE Access, 7, 136694–136703. https://doi.org/10.1109/ACCESS.2019.2942433
- 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
- Kuhn, W., & Ballatore, A. (2015). Designing a Language for Spatial Computing. In F. Bacao, M. Y. Santos, & M. Painho (Eds.), AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities, 309–326. Springer International Publishing. https://doi.org/10.1007/978-3-319-16787-9_18
- Lampoltshammer, T. J., & Heistracher, T. (2012). Natural Language Processing in Geographic Information Systems – Some Trends and Open Issues –. International Journal of Computer Science & Emerging Technologies, 3(3), 81-88.
- Le, T. A., Arkhipov, M. Y., & Burtsev, M. S. (2018). Application of a Hybrid Bi-LSTM-CRF Model to the Task of Russian Named Entity Recognition. In A. Filchenkov, L. Pivovarova, &

J. Žižka (Eds.), Artificial Intelligence and Natural Language (pp. 91–103). Springer International Publishing. https://doi.org/10.1007/978-3-319-71746-3_8

- Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314
- Lima E.B, & Davis C.A. (2017). Geographic information extraction using Natural Language Processing in Wikipedia texts. Proceedings XVIII GEOINFO, 122—127.
- Liu, H., Qiu, Q., Wu, L., Li, W., Wang, B., & Zhou, Y. (2022). Few-shot learning for name entity recognition in geological text based on GeoBERT. Earth Science Informatics, 15(2), 979–991. https://doi.org/10.1007/s12145-022-00775-x
- Luoma, J., & Pyysalo, S. (2020). Exploring Cross-sentence Contexts for Named Entity Recognition with BERT (arXiv:2006.01563). arXiv. http://arxiv.org/abs/2006.01563
- Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004
- Martins, B., Manguinhas, H., & Borbinha, J. (2008). Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. 2008 IEEE International Conference on Semantic Computing, 1–9. https://doi.org/10.1109/ICSC.2008.86
- Perea-Ortega, J. M., García-Cumbreras, M. A., & Ureña-López, L. A. (2013). Applying NLP Techniques for Query Reformulation to Information Retrieval with Geographical References. In T. Washio & J. Luo (Eds.), Emerging Trends in Knowledge Discovery and Data Mining (Vol. 7769, pp. 57–69). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36778-6_6
- Perea-Ortega, J. M., Santiago, F. M., R´aez, A. M., and L´opez, L. A. U. (2009). Geo-NER: un reconocedor de entidades geograficas para ingles basado en GeoNames y Wikipedia. Procesamiento del Lenguaje Natural, 43:33–40. https://www.redalyc.org/articulo.oa?id=515751743004
- Qin, Y., & Zeng, Y. (2018). Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF. Journal of Shanghai Jiaotong University (Science), 23(3), 392–397. https://doi.org/10.1007/s12204-018-1954-5
- Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2019). GNER: A Generative Model for Geological Named Entity Recognition Without Labeled Data Using Deep Learning. Earth and Space Science, 6(6), 931–946. https://doi.org/10.1029/2019EA000610
- Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020). Geo-analytical question-answering with GIS. International Journal of Digital Earth, 14(1), 1–14. https://doi.org/10.1080/17538947.2020.1738568
- Schmid, H. (1994). Part-of-Speech Tagging with Neural Networks (arXiv:cmp-lg/9410018). arXiv. https://doi.org/10.48550/arXiv.cmp-lg/9410018
- Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals, 140, 110212. https://doi.org/10.1016/j.chaos.2020.110212
- Shin, H. J., Park, J. Y., Yuk, D. B., & Lee, J. S. (2020). BERT-based Spatial Information Extraction. Proceedings of the Third International Workshop on Spatial Language Understanding, 10–17. https://doi.org/10.18653/v1/2020.splu-1.2
- Souza, F., Nogueira, R., & Lotufo, R. (2020). Portuguese Named Entity Recognition using BERT-CRF (arXiv:1909.10649). arXiv. https://doi.org/10.48550/arXiv.1909.10649

- Sun, Q., Jankovic, M. V., Bally, L., & Mougiakakou, S. G. (2018). Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network (arXiv:1809.03817). arXiv. https://doi.org/10.48550/arXiv.1809.03817
- Van, T. T., Sy, K. V., Anh, T. T., Duc, V. H., Quang, T. L., Xuan, P. H., Viet, H. L., Quang, H. B., & Bao, S. P. (2021). Design of an GIS-based Investment Heatmap System using Topic Classification and NER. 2021 13th International Conference on Knowledge and Systems Engineering (KSE), 1–5. https://doi.org/10.1109/KSE53942.2021.9648730
- Wang, X., Yang, C., & Guan, R. (2018). A comparative study for biomedical named entity recognition. International Journal of Machine Learning and Cybernetics, 9(3), 373–382. https://doi.org/10.1007/s13042-015-0426-6
- Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics. COLING 1992. https://aclanthology.org/C92-4173
- Wen, C., Chen, T., Jia, X., & Zhu, J. (2021). Medical Named Entity Recognition from Unlabelled Medical Records based on Pre-trained Language Models and Domain Dictionary. Data Intelligence, 3(3), 402–417. https://doi.org/10.1162/dint_a_00105
- Xiang, J., Qiu, Z., Hao, Q., & Cao, H. (2020). Multi-time scale wind speed prediction based on WT-bi-LSTM. MATEC Web of Conferences, 309, 05011. https://doi.org/10.1051/matecconf/202030905011
- 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
- Yadav, V., & Bethard, S. (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models (arXiv:1910.11470). arXiv. http://arxiv.org/abs/1910.11470
- Zhao, P., Zhu, H., Liu, Y., Li, Z., Xu, J., & Sheng, V. S. (2018). Where to Go Next: A Spatio-temporal LSTM model for Next POI Recommendation (arXiv:1806.06671). arXiv. https://doi.org/10.48550/arXiv.1806.06671