Prompt Engineering for Domain-Specific Geo-spatial Named Entity Disambiguation

Anonymous ACL submission

Abstract

Despite the scarcity of employing transformer approaches for toponym resolution, this study leverages oral and transcribed text data to address the disambiguation of diverse named entities, including place names such as camps, ghettos, and streets. We utilise generative AI techniques, incorporating prompt engineering, to effectively disambiguate these named entities within geographical contexts.

Our methodology aims to demonstrate how leveraging prompt engineering from general large language models (LLMs) can be effectively employed for less commonly addressed topics, such as toponym resolution in the field of Natural Language Processing (NLP). We have evaluated the few-shot chain of thought (COT) prompting approach combining the knowledge base (KB) as a retriever to provide the fewshots required for the reasoning process of LLM. This technique illustrates the efficacy of these advanced approaches in accurately identifying and resolving toponyms in complex textual datasets, thereby contributing valuable insights to the field of geographic information systems and digital humanities.

1 Introduction

In the geospatial domain, ambiguities in words are widespread and can present significant challenges, particularly in sensitive historical contexts such as the Holocaust. Spoken language, with its diverse dialects, accents, and linguistic nuances, further complicates the resolution of toponyms, placenames or geographic locations. Identifying these toponyms accurately is crucial for understanding historical events. Over time, geographic locations may have been referred to by different names in textual documents, adding to the complexity. These discrepancies pose formidable obstacles to the analysis of historical documents, underscoring the need for robust toponym resolution methods in Holocaust research. In the process of automatic information extraction, resolving toponyms presents a significant challenge that remains largely unaddressed. This task is particularly crucial in the context of named entity recognition (NER), where accurately identifying and categorising geographic locations mentioned in transcribed text, especially within sensitive historical domains like the Holocaust, is paramount.

In comparing spoken and transcribed data with written language, various ambiguities arise in speech data. Disambiguating location-based named entity tags in speech data is particularly challenging compared to written text due to the inherent complexities of speech, including variations in pronunciation, accents, and dialects, as well as the absence of punctuation and grammatical cues found in written language. These factors contribute to difficulties in accurately identifying and resolving named entities related to locations in speech data. In Holocaust research, oral testimonies play a pivotal role in preserving survivors' experiences. These testimonies often mention concentration camps, ghettos, and other geographical locations, using consistent naming conventions. This consistency in naming conventions accentuates the need for robust NER systems capable of resolving toponyms accurately, thereby enhancing our understanding of historical narratives. While there has been some related research, we found that most of the existing approaches are unable to deliver satisfactory results because of the following reasons. For a clearer explanation, please refer to Figure 1.

- Referring the same name for different contexts
- Different spelling referring to the same place
- Symbols refer the geographical location

With the recent advancement of Large Language Models (LLMs), which are trained using billions of parameters, promising results have been achieved for various Natural Language Processing (NLP) tasks compared to previously existing machine learning models in the general domain. These models, primarily developed with contextual understanding, have shown (including in recent studies conducted by the authors) that they outperform rule-based approaches. However, more research needs to be conducted within domain-specific approaches to evaluate the adaptability of context-specific methodologies. In this study, we experiment with the adaptability of the LLMs and transformer models for the toponym resolution.

More specifically, we propose a novel approach which employs LLMs for toponym resolution, comparing different traditional approaches and seeking to answer the following research questions.

- RQ1: Does structural similarity of sentences effect in toponym resolution?
- RQ2: Are general task LLMs able to identify the toponyms discussed in the oral and transcribed texts?
- RQ3: Can advanced prompt engineering techniques, combined with lexicon knowledge, recognise domain-specific toponyms?

The rest of this paper is organised as follows. We describe previous studies in Section 2. We present our methodology in Section 3. In Section 4, we describe our experiments and report the results. Section 5 offers an error analysis, and a brief conclusion is provided in Section 6.

2 Related Work

Even though different traditional approaches, such as hand-crafted rules and heuristics, heuristics of rule-based systems as features in supervised machine learning models to predict geospatial labels for place names were employed. In previous studies, deep learning methodologies have been employed for toponym resolution to model the textual elements by combining bidirectional Long Short-Term Memory (LSTM) units with pre-trained contextual word embeddings (i.e., static features extracted using either the Embeddings from Language Models (ELMo) or the Bidirectional Encoder Representations from Transformers (BERT) methods. A limitation of these studies is that they discuss only the general named entity tags such as LOC GPE but not the domain-specific entities such as concentration camps (CAMP), ghettos (GHETTO), streets (STREET), etc.

Additionally, several studies have leveraged deep neural network architectures for toponym resolution (Cardoso et al., 2019; Kulkarni et al., 2021). For example, Gritta et al. proposed a network architecture called the CamCoder system, which aims to disambiguate place references by detecting lexical clues within the context surrounding the mention. The authors also introduced a sparse vector representation named MapVec, which encodes prior geographic probabilities associated with locations based on coordinates and population counts (Cardoso et al., 2019). Similarly, Cardoso et al. (Kulkarni et al., 2021) utilised a combination of context-aware word embeddings (Peters et al., 1802) and a recurrent neural network based on Bidirectional LSTMs (Huang et al., 2015). The above studies have covered not only English but also other languages such as Spanish.

Transformer-based techniques have recently had a substantial impact on toponym resolution methodologies. The current approaches can be broadly classified into two categories: localisation-based and ranking-based. The localisation-based approach primarily focuses on the direct prediction of geographic coordinates or areas from the given textual input. For instance, Radford's method (Radford, 2021) utilises DistilRoBERTa for end-to-end probabilistic geocoding. Similarly, Cardoso et al. (Cardoso et al., 2022) employ Long Short-Term Memory (LSTM) networks with BERT embeddings to predict probability distributions over spatial regions. In a sequence-to-sequence framework, Solaz and Shalumov (Solaz and Shalumov, 2023) use the T5 Transformer model to translate text into hierarchical encodings of geographic cells. Another notable study by Gomes et al. (Gomes et al., 2024) proposes a method that leverages the adaptation of SentenceTransformer models, initially designed for sentence similarity tasks, for toponym resolution. The authors fine-tune the models on geographically annotated English news article datasets, including Local Global Lexicon, GeoWebNews, and TR-News.

One of the major challenges in transformer-based toponym resolution methods is the absence of domain-specific fine-tuning. Pre-trained transformer models such as BERT (Devlin et al., 2019) are optimised to generate embedding for tasks like masked language modelling and next-sentence prediction. Therefore, it is plausible that models trained on larger datasets have a greater capacity to identify the correct toponym.

Example 01: Referring the same name for different contexts

We	were	taken	to	**Theresienstadt**	transit	camp	to	Majdanek
O	O	O	O	B-CAMP	O	O	O	B-CAMP

All	of	us	stayed	in	**Theresienstadt**	for	three	nights
O	O	O	O	O	B-GHETTO	O	O	O

Example 02: Different spelling referring to the same place example (**Auschwitz- Birkenau is a one camp**)

who	had	to	come	to	**Auschwitz**	in	1942	from	Slovakia
O	O	O	O	O	B-CAMP	O	B-DATE	O	B-GPE

those	unfit	for	further	experiments	were	sent	back	to	**Birkenau**	or	gassed
O	O	O	O	O	O	O	O	O	B-CAMP	O	O

Example 03: Symbols refer the geographical location

They	were	transported	to	**KZ**	Flossenbuerg	in	Bavaria
O	O	O	O	B-CAMP	I-CAMP	O	B-GPE

Figure 1: Sample examples for each scenario.

Another significant issue with machine learning-based toponym resolution methods is the geographic bias, which arises due to the imbalance in the geographic distribution of training datasets. Liu et al. (Liu et al., 2022) make the point that models tend to favour locations that are overrepresented in the training corpora. The scarcity and lack of diversity in geotagged datasets further intensifies this bias (Gritta et al., 2018).

Our review revealed a notable gap in the current body of research: no studies have employed state-of-the-art generative language models for toponym resolution. Despite the advancements in generative language models like GPT, which have demonstrated significant potential in other natural language processing tasks, their application to toponym resolution and disambiguation of place names remains unexplored.

3 Methodology

According to the previous studies, it is evident that knowledge of Large Language models is effective and can be used for domain-specific tasks using proper computational techniques (Chang et al., 2024; Zhao et al., 2023). In the present study, we designed the prompts to leverage the general task language model for the name-entity recognition task in the geospatial arena. The absence of fine-tuning or training of base models in our approach is intentional, and we attempt to utilise prompt augmentation techniques to reframe the prompts to suit the downstream tasks, such as toponyms resolution.

3.1 Data and Dataset creation

Oral and transcribed versions of Holocaust testimonies were used for the following experiments described in this study. These data were manually annotated according to the BIO (Beginning-Inside-Outside) tagging scheme. For the annotation process we employed UBIAI tool. The training samples were manually annotated by human annotators, resulting in an inter-annotator agreement of 0.76. More details about the data used for this study are reported in (Anuradha Nanomi Arachchige et al., 2023). Refer Figure 1 for annotation style.

3.2 Baseline Approaches

The baseline approaches were designed from scratch to determine whether it is possible to identify toponyms correctly without considering contextual knowledge.

Rule-Based Approach: For this approach, we selected the SpaCy NER model and augmented it with vocabularies specific to concentration camps and ghettos. Additionally, we defined rules to extract street names and ghettos, which were combined with the SpaCy transformer (trf) NLP model to form domain-specific NER model. Some of these defined rules are shown in Table 1.

Table 1: Examples for the defined regular expression for entity mining.

<table>
<tr><th>Entity</th><th>Regex Expression</th><th>Match</th></tr>
<tr><td>Street</td><td>If name followed by street semantically identical word ([A-Z][a-z]*(strasse|straße|straat)|([A-Z][a-z]*(Street|St|Boulevard|Blvd|Avenue|Ave|Place|Pl)()*))</td><td>Hauptstraße</td></tr>
<tr><td>Ghetto</td><td>Search on the lexicon consist Ghetto names or either name followed by ghetto [A-Z]w+((-|)*[A-Z]w+)* (g|G)hetto</td><td>Anyksciai</td></tr>
</table>

Structural similarity:

N-Gram Approach: In this approach, we consider the n-grams surrounding the target word. We experiment with different window sizes and various n-gram combinations. Subsequently, we attempt to identify the most similar n-grams in conjunction with our target word to determine the most common and probable entity. However, this approach does not perform well due to the nature and unstructured of the dataset.

Part-of-Speech Tags: In this approach, we generate part-of-speech (POS) tags for every word in the corpus, along with their respective sentences. We then analyse the presence of similar POS tag patterns in sentences containing the target word associated with a toponym. By identifying sentences with the most similar POS tag combinations to the target word, we select the most frequently occurring sentences with similar POS tags and use them to calculate the probability of the word being the correct toponym. Unfortunately, the proposed method proved ineffective due to the highly unstructured nature of oral and transcribed text data. Although we could identify common POS tag combinations with the target word, it was challenging to find a sufficient number of instances meeting our threshold. Specifically, we required at least three similar sentences to predict the label as a particular toponym based on the POS tag structure, but this criterion was seldom met within the dataset.

3.3 Prompt creation

The labelling of geo-entities in relation to historical data remains largely unexplored. Incorporating a proper KB that includes both historical and geospatial data is crucial for accurate modelling. However, data scarcity and the unavailability of properly labelled data are significant issues in the Holocaust domain. To address these challenges, we have explored the integration of Large Language Models (LLMs) with different prompt engineering techniques to bridge the knowledge gap.

We evaluated our approaches in two primary phases: zero-shot Chain-of-Thought (COT) prompting to explore the model's accuracy and establish a baseline value, and few-shot COT prompting using a labelled KB as the retriever to refine the model based on the value of geospatial data. Throughout the evaluation, GPT-4o served as the base model. It was set to operate with a temperature of 0 and a maximum token limit of 1500 per output, primarily functioning as a 'helpful assistant' for identifying geospatial entities in Holocaust testimonies.

3.3.1 Zero-shot COT prompting

Identifying an optimal prompt is considered to be a crucial point in the prompt engineering process related to LLM inferencing. The leading researcher of the study incorporated multiple prompt augmentation techniques to identity the optimal prompt required for the initial study. Task-specific prompting approaches and fact-checking approaches are thoroughly explored. The used prompt is depicted in the Table 2.

Table 2: Zero-shot COT Prompt.

Zero-shot COT Prompt

Consider the year from 1936-1944. You are going to identify name entity tags for holocaust-specific tags. The list of name entity tags should be {list_of_tags}. Each tag is as follows: {tags_meaning}. Now do the below tasks.

1. Try to identify the most suitable Name entity tag for the word 'NAMEENTITY' in the GIVEN SENTENCE based on the below criteria:

- Analyse the word in front of the 'NAMEENTITY' tag before you tag.
- Understand the complete sentence and try to identify specific factors discussing the word you want to tag.

The GIVEN SENTENCE: {sentence}.

2. Return only the GIVEN SENTENCE after assigning the identified tags instead of the word 'NAMEENTITY'. Do not add additional data.

Use the following format for the output:
"<Updated sentence with correctly identified name entity tags>"

In the prompt, {tag meaning} contains information related to geospatial entities. We have used the following information to enhance geospatial knowledge within the prompt during the inference process:

- LOC: Locations except countries or cities.
- GPE: Geographical locations such as countries or cities.
- CAMP: Concentration camps (Extermination, Transit, Labour)
- GHETTO: Ghettos, the Jewish quarters in cities.
- STREET: Pathways or roads.

{sentence} is the sentence that contains values that need to be tagged with geospatial entities. During the evaluation process, it was noticed that GHETTO, LOC, and CAMP need notable improvement. Therefore, we proposed a few-shot COT prompt to address this issue.

3.3.2 Retrieval Augmented Generation (RAG)

The zero-shot approach or the baseline study fails to evaluate the entity labelling accurately for Geospatial labels like GHETTOS (see Table 2). To overcome this issue, we used a retrieval-augmented generation (RAG) pipeline to share the geospatial knowledge during the prompting process. The RAG approach is mainly designed using two phases: vector store generation and the retriever with response generation with the GPT 4o model.

- Vector store generation and Embedding

  The 'BGE small' model from Huggingface has been used as the embedding for the study, while Chroma DB is utilised to store the vectors related to the labelled geospatial data. To preserve contextual meaning during chunking, a recursive character text splitter from LangChain has been incorporated to create the necessary data chunks with 2500 tokens, overlapping 50 tokens. These chunks are stored in the vector store once embedded using the embedding model.

- Retriever and prompting

  Retrieval QA has been utilised to build the retriever, with search_kwargs(k) set to '2' and the search_type set to 'similarity'. The similarity search uses cosine similarity to extract the vectors closest to the input sentence we want to tag. This approach allows us to feed data with a similar labelled context to the model, enriching the response generation task with geospatial knowledge. The designed few-shot COT prompt is employed here, with minor adjustments to fit it into the process.

The major drawback of this approach is that the retriever relies on similarity score measurements to retrieve related data based on the context provided in the sentence. Consequently, it may retrieve sample chunks where the target word is absent. We have redesigned the word level retriever to tackle this concern using a KB.

3.3.3 Few-shot COT prompting

During this phase, our primary aim was to improve prediction accuracy by incorporating pre-labelled knowledge into the inference process. We organised the labelled data in a knowledge graph based on value and geospatial entities, from which the few-shots required for inference are retrieved. The tree structure of the knowledge graph is designed with the place as the root node and geospatial entities as the first-level parent nodes. Leaf nodes are implemented using a list structure containing sample instances of labelled datasets. This approach has effectively improved the retrieval time of example phrases required for few-shot learning, which can be performed in a constant time. This knowledge-sharing method has enhanced the geospatial knowledge during the response generation process.

The presence of the target word in the retrieved sentences is considered mandatory for efficient labelling in the few-shot approach. Utmost five instances for each entity are retrieved from KB. If the word is absent, the prompt will function in a zero-shot manner. The detailed workflow of the approach is presented in figure 4.

The prompt in the table 2 is amended by sharing the additional information retrieved for the second phase. The below line is introduced as the first chain of thought to the prompt.

- Examine the below examples and learn about the appropriate Name entity tags for the words based on the context. Examples are '{result}'.

{result} tag contains the extracted knowledge from the KB. This approach has shown promising results in handling the GHETTO, LOC and CAMP.

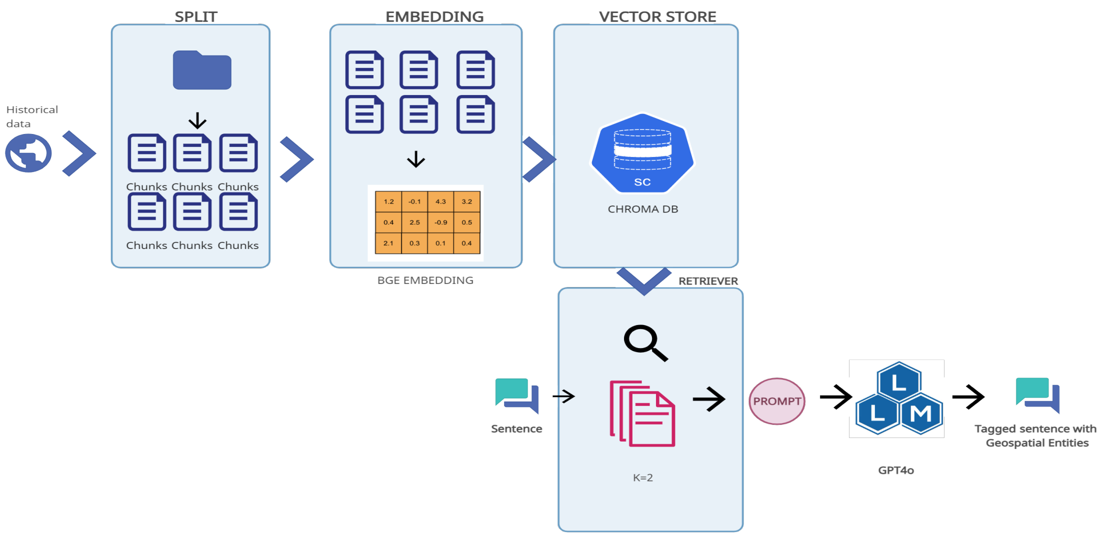

Figure 2: Data-flow of the RAG pipeline.

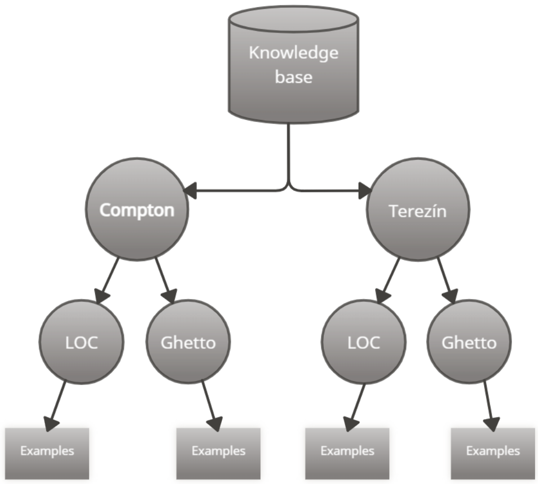

Figure 3: Knowledge base arrangement.

The detailed analysis of the results is discussed in the 4 section.

The code associated with this research will be made publicly available as part of the supplementary materials accompanying the final version of this paper upon its acceptance for presentation at the conference.

4 Results and Discussion

Due to the nature of oral interviews and testimonies, it is often necessary to disambiguate toponyms rather than perform straightforward geocoding. Many people were transported to or travelled through various locations, often unknown or unspecified, which complicates the identification of precise geographic references.

Table 3: Performance of baseline study using Rule-based approach (Spacy Transformer Model).

Entities	**Baseline: Rule based**		
	Precision	**Recall**	**F1score**
LOC	1.0	0.10	0.18
GPE	0.64	0.83	0.72
CAMP	0.77	0.51	0.62
GHETTO	0.00	0.00	0.00
STREET	0.67	0.51	0.58

As a baseline for the study, we evaluated a rule-based approach using SpaCy to perform geospatial entity labelling. During the initial baseline study, we concluded that contextual and pragmatic relations between words are crucial for disambiguating geospatial entities like GHETTO and LOC. This is evident from the results shown in the table 3. From the results, it is evident that GHETTO and LOC are misinterpreted as GPE in most cases, highlighting the importance of identifying the contextual meaning and the involvement of the word in the testimony.

Table 4 presents the evaluation of the baseline model with zero-shot prompting, RAG pipeline and the Hybrid approach with few-shot prompting and KB. For each approach, a carefully crafted prompt was selected through an iterative evaluation process employing prompt augmentation techniques. The GPT-4o model modestly classifies GPE, CAMP, and STREET entities in the zero-shot prompting

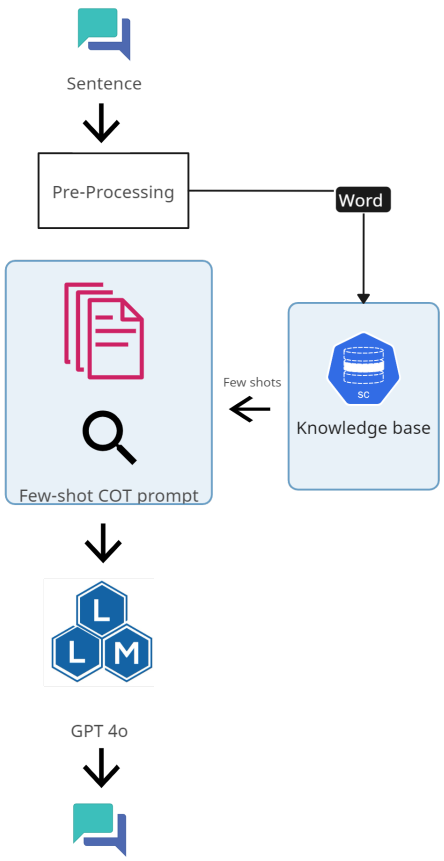

Figure 4: Data-flow of the few-shot COT pipeline.

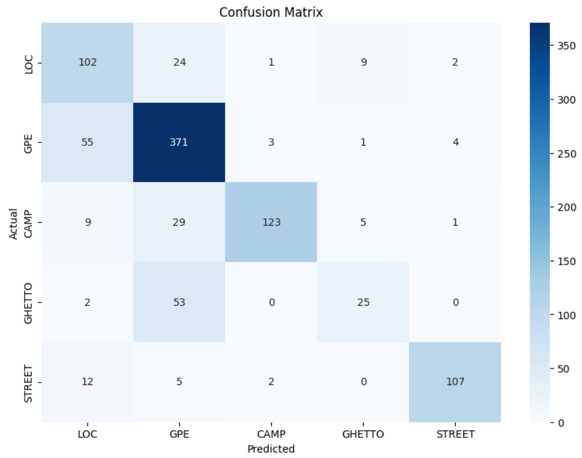

Figure 5: Confusion Matrix for zero-shot prompting approach.

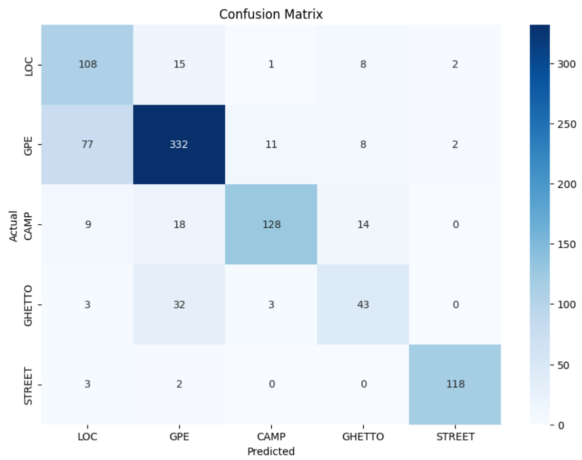

Figure 6: Confusion Matrix for RAG based Approach.

approach. However, it significantly underperforms at classifying GHETTO, showing a notable prediction loss. To address this issue, we propose an RAG pipeline which targets sentence-level retrievers using the cosine distance. Compared to the baseline approach, GHETTO tagging shows a 0.15 improvement in F1 score while other entities show a slight improvement in the tagging. This approach has shown that proper retrieval would improve the performance of the tagging process. A few-shot chain-of-thought (COT) based approach is proposed to handle the geospatial knowledge scarcity. The target word-orientated retriever, which uses a tree structure, is incorporated to extract the most appropriate few-shots required to infer GPT 4o. The few-shot COT approach has shown significant improvements, particularly for the GHETTO category, with an increase of 0.19 in the F1 score, and for the LOC category, with an increase of 0.08 in the F1 score. These results show that well-crafted prompts, along with a knowledge-sharing approach, can drive the general purpose language models to specific tasks like Name entity recognition in the Geo-spacial domain.

5 Discussion

In this section, we will discuss the findings from our experiments.

5.1 RQ1: Does structural similarity of sentences effect in toponym resolution?

Since this is a novel problem, we explored various methods to assess whether structural similarity can be leveraged to identify toponyms and disambiguate the given named entities accurately. We employed several tasks, as discussed in the 'Methodology' section; however, none yielded robust results. This indicates that structural similarity alone is insufficient for effectively detecting highly unstructured oral and transcribed data.

Table 4: Performance comparison between Models.

Entities	Baseline GPT 4o			RAG with GPT 4o			Few-shot COT Prompting		
	Precision	**Recall**	**F1score**	**Precision**	**Recall**	**F1score**	**Precision**	**Recall**	**F1score**
LOC	0.57	0.74	0.64	0.54	0.81	0.64	0.63	0.84	0.72
GPE	0.77	0.85	0.81	0.83	0.77	0.80	0.89	0.82	0.85
CAMP	0.95	0.74	0.83	0.90	0.76	0.82	0.88	0.79	0.83
GHETTO	0.62	0.31	0.41	0.59	0.53	0.56	0.61	0.59	0.60
STREET	0.94	0.84	0.88	0.97	0.94	0.95	0.91	0.91	0.91

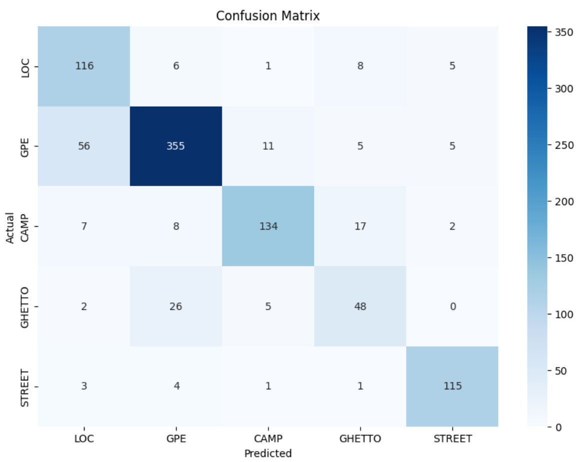

Figure 7: Confusion Matrix for Few-shot prompting approach.

5.2 RQ2: Will general task LLMs be able to identify the toponyms discussed in the oral and transcribed texts?

According to our second research question (RQ2), our experiments demonstrate that general-purpose LLMs were highly effective in identifying toponyms within domain-specific contexts. Despite the presence of code-mixing, where terms from different languages are interspersed within the transcribed texts, these LLMs successfully identified and accurately labelled the correct geospatial named entities. This capability highlights the robustness of general-purpose LLMs in handling multilingual and mixed-language scenarios, providing reliable results. We plan to extend this approach to open-source LLMs such as Mistral, Falcon and Llama.

5.3 RQ3: Can advanced prompt engineering techniques, combined with lexicon knowledge, recognise domain-specific toponyms?

Our experiments indicate that advanced prompt engineering techniques significantly enhance performance in domain-specific geo-spatial named entity disambiguation. Employing these advanced prompts not only improves accuracy but also reduces computational costs. This cost efficiency is particularly beneficial during both the initial pretraining of models and subsequent fine-tuning processes. By optimising prompt design, we can achieve more effective model training with lower computational requirements, thus streamlining the entire model development lifecycle.

In the future, this study can also be extended and generalised to prompts to address the disambiguation of other geographical named entities, including natural landmarks such as rivers, forests, and mountains. By expanding the scope to include a broader range of toponyms, we can enhance the model's ability to accurately identify and differentiate between various types of geographical entities. This extension will contribute to a more comprehensive and robust system for geographical named entity resolution, benefiting applications in fields such as geographic information systems (GIS), environmental monitoring, and digital humanities.

6 Conclusion

In this paper, we have explored the evolution from traditional methods to state-of-the-art LLMs for toponym resolution in oral and transcribed texts, particularly within the context of Holocaust studies. Our discussion highlights how these advanced approaches significantly improve accuracy and efficiency. We demonstrate how using labelled data as a knowledge base enriches the inference process, turning few-shot examples into a wealth of information to handle corner cases in geospatial disambiguation. Moreover, as detailed in the preceding sections, leveraging prompts within these models can yield high-quality results at a reduced cost, thereby enhancing the overall feasibility and effectiveness of toponym resolution efforts in this specialised domain.

Limitations

In our study, we have demonstrated promising results in an unexplored domain. However, several limitations exist. The study is exclusively centered on GPT-4o without any training or fine-tuning. The characteristics of the data used in GPT's initial training may have directly impacted the outcomes. Further refinement through fine-tuning the model with oral and transcribed data could enhance the process. Due to the constraints of the datasets, we utilized a limited number of records for the evaluation process, which warrants further exploration.

References

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2019. Using recurrent neural networks for toponym resolution in text. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II 19*, pages 769–780. Springer.

Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2022. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11(1).

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diego Gomes, Ross S Purves, and Michele Volpi. 2024. Fine-tuning transformers for toponym resolution: A contextual embedding approach to candidate ranking. In *GeoExT@ ECIR*, pages 43–51.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52:603–623.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2021. Multi-level gazetteer-free geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88.

Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, and M. Shi. 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3:9.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 1802. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*, 42.

Benjamin J. Radford. 2021. Regressing location on text for probabilistic geocoding. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 53–57, Online. Association for Computational Linguistics.

Yuval Solaz and Vitaly Shalumov. 2023. Transformer based geocoding. *Preprint*, arXiv:2301.01170.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.