Evaluating Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies

Isuri Anuradha Nanomi Arachchige
University of Wolverhampton/ UK
isurianuradha96@gmail.com

Le An Ha
University of Wolverhampton/ UK
ha.l.a@wlv.ac.uk

Ruslan Mitkov
Lancaster University / UK
r.mitkov@lancaster.ac.uk

Vinitar Nahar
University of Wolverhampton, UK
vinita.nahar@wlv.ac.uk

Abstract

Relationship extraction from unstructured data remains one of the most challenging tasks in the field of Natural Language Processing (NLP). The complexity of relationship extraction arises from the need to comprehend the underlying semantics, syntactic structures, and contextual dependencies within the text. Unstructured data poses challenges with diverse linguistic patterns, implicit relationships, contextual nuances, complicating accurate relationship identification and extraction. The emergence of Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer), has indeed marked a significant advancement in the field of NLP.In this work, we assess and evaluate the effectiveness of LLMs in relationship extraction in the Holocaust testimonies within the context of the Historical realm. By delving into this domain-specific context, we aim to gain deeper insights into the performance and capabilities of LLMs in accurately capturing and extracting relationships within the Holocaust domain by developing a novel knowledge graph to visualise the relationships of the Holocaust. To the best of our knowledge, there is no existing study which discusses relationship extraction in Holocaust testimonies. The majority of current approaches for Information Extraction (IE) in historic documents are either manual or Optical Character Recognition (OCR) based. Moreover, in this study, we found that the Subject-Object-Verb extraction using GPT3-based relations produced more meaningful results compared to the Semantic Role labeling-based triple extraction.

1 Introduction

Understanding unstructured texts using computational methods is considered a challenging task due to the complexity of the natural language. Holocaust testimonies are firsthand accounts provided by survivors, witnesses, and others who experienced or observed the atrocities of the Holocaust during World War II (Isuri A. Nanomi Arachchige, 2023). Holocaust testimonies also belong to the category of unstructured texts, which presents unique challenges for computational assessment. These testimonies are often emotionally charged and contain highly-sensitive and personal information which is scattered everywhere in the testimony. Moreover, Holocaust testimonies contain a range of linguistic complexities, such as archaic language, regional dialects, and highly specialised terminology related to the Holocaust, which can be challenging to parse using traditional NLP techniques.

Extracting relationships from Holocaust testimonies is essential for historians as these firsthand accounts provide valuable information about the Holocaust. By uncovering hidden connections and associations between entities from the testimonies, historians can gain deeper insights into the historical context, dynamics between individuals and groups, and the broader narrative of the Holocaust. This information helps to enhance the understanding of the events, identify patterns, and shed light on the social, political, and cultural aspects of this tragic period in history. However, existing approaches for IE in historic documents are mainly manual (reference), or a few based on advanced digitalised approaches such as OCR (Bryant et al., 2010). Recently, there have been some efforts made towards IE in historic documents using NLP (Blanke et al., 2012).

Relationship Extraction (RE) plays a crucial role in discovering meaningful connections and associations between entities from Holocaust testimonies to enhance our understanding of the historical context. However, the unstructured nature of these testimonies presents additional challenges when it comes to extracting relationships, making

Proceedings of Recent Advances in Natural Language Processing, pages 117–123
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_013

the task even more difficult for humans. Dependencies between words and phrases, captured by dependency parsing, provide valuable insights into the syntactic and semantic relationships within the text. There are many downstream applications which are based on extracted relations, such as Information Retrieval (Guo et al., 2020), Question Answering (QA) (Lan et al., 2021), and Knowledge Graph Construction (Zhang et al., 2022). A knowledge graph is a type of graph database that is designed to systematically organise and present knowledge in a structured format. In the context of unstructured data, knowledge graphs are used to extract and organise the information into a structured format using different cutting-edge NLP techniques. Further, graphical representations improve the accuracy and relevance of the search and retrieval results of information from unstructured texts. Visualising relationships between entities and events in the Holocaust using a graph enables people to identify all of the personal names, places, and locations mentioned in a collection of testimonies.

However, with recent advancements in large language models (LLMs) such as GPT3, there has been significant progress in uncovering hidden relationships within specific content (Xu et al., 2023; Haddad et al., 2023). These LLMs, trained on vast amounts of textual data, have demonstrated their capability to learn complex patterns and capture nuanced relationships between entities. LLMs have introduced a novel paradigm known as in-context learning (ICL) (Dong et al., 2022). This paradigm, as exemplified by studies such as (Brown et al., 2020), formulates NLP tasks as language generation problems, allowing the models to make predictions based on demonstrations provided within the context. Instead of relying solely on fine-tuning with labelled data, LLMs leverage the power of language generation to produce outputs such as Named Entity Recognition. The objective of this paper is to examine the performance of LLMs by employing ChatGPT on Holocaust testimonies for RE. Following are the contributions of the proposed paper.

- We evaluate the traditional dependency parser-based relation extraction method against the results of the GPT model.

- We conduct systematic analysis to provide valuable insights into the strengths and weaknesses of each traditional dependency extraction and relationships obtained from the GPT.

- We release the code of the experiments as an open-source GitHub project[1]

The rest of this paper is organised as follows. We critically analyse related work in Section 2. We present our methodology in section 3. In Section 4, we describe our experiments and report the results and Section 5 discusses the next steps of this research. Finally, a brief conclusion is provided in Section 6.

2 Related Work

In this section, we critically analyse existing research in the field of NLP for relationship extraction. We will discuss and establish the context of RE within historic documents in particular Historical testimonies. Previous studies have paid little attention to the computational approaches for information extraction in Holocaust testimonies. The valuable information embedded within these testimonies remains largely unexplored, representing a hidden knowledge source within historical data. Leeuw, D. et. al shed light on the existing digital infrastructure for Holocaust studies and underscored the significant limitations inherent in this domain (De Leeuw et al., 2018). They emphasised the pressing need for computational approaches to effectively address these challenges and overcome the limitations. Moreover, some rule-based computational approaches were performed on multi-source Holocaust victim reports to extract biographical information (Sagi et al., 2016).

To date, no study has been conducted specifically focused on identifying the relations and entities present in Holocaust testimonies. This research gap highlights the untapped potential for leveraging computational techniques. Relationship extraction is a common downstream task that is often performed in conjunction with named entity recognition in various domains, including biomedical (He et al., 2023), finance (Wu et al., 2023), and more. The goal of RE is to identify and extract meaningful connections or associations between entities mentioned in the text. According to previous studies, several approaches have been considered in identifying relationships.

[1] https://github.com/isuri97/infoextra

- **Existence of relationship between entities** classify whether a meaningful semantic relationship exists between two entities or if they are mentioned together without a specific named relationship.

- **Extracting predicate verb as relationship type** predicate does not consist of a closed set of possible classes. Any predicate verb that appears in a sentence and indicates a relationship between entities is considered a relationship type. The normalisation of relationship types is deferred for future processing or analysis.

These extracted relationships can then be used to build knowledge graphs (Milošević and Thielemann, 2023), which serve as representations of the extracted information.

The recent advancements in LLMs have led to their widespread adoption in various NLP tasks, including text classification (Sun et al., 2023). These studies have leveraged GPT models to improve the performance of text classification tasks. Furthermore, a recent study (Wan et al., 2023) explored the use of LLMs for relationship extraction. However, their findings indicate that LLMs reveal lower performance than fully-supervised baselines, such as fine-tuned BERT. Despite the application of transformer-based models in relationship extraction across various domains, there is currently a lack of annotated datasets specifically tailored for the Holocaust domain. This poses a challenge in leveraging the power of these models for extracting relationships from Holocaust testimonies and gaining domain-specific insights. Addressing this gap by creating annotated datasets tailored to the Holocaust domain would greatly contribute to the development of more accurate and contextually relevant relationship extraction models.

Despite the promising performance of LLMs in various NLP tasks, the application of In-Context Learning (ICL) for relation extraction (RE) still presents challenges. RE involves identifying the semantic relationship between two entities mentioned in a sentence, which requires a comprehensive understanding of natural language. Recent research by (Carrino et al., 2022) has explored the application of GPT3 ICL for biomedical RE and evaluated the complete dataset, suggesting that there is room for improvement in this area for domain-specific contexts.

3 Methodology

In this section, we describe the proposed pipeline employed for creating the knowledge graph. As shown in Figure 1, our proposed knowledge graph consists of four components: 1) Data processing, 2) Coreference resolution, 3) Triple extraction 4) Visualisation. After the collection of Holocaust testimonies, the coreference resolution component identifies chains of entities and pronouns that refer to the same entity. The triple extraction component extracts relation triples from the text using open information extraction techniques and lastly, extracted relationships are visualised our findings on a graph database. The details of each component are presented below.

3.1 Data Processing

The collection of documents plays a pivotal role in our project, with a specific focus on extracting information from Holocaust testimonial transcripts. Our primary objective is to gather a comprehensive set of English-language testimonies sourced from diverse Holocaust testimonial archives. To accomplish this, we have employed web scraping techniques to gather data specifically from the Wiener Library website. Subsequently, we have undertaken appropriate pre-processing steps to ensure the data is prepared for further analysis and information extraction. Table 1 refers to the list of relations that we have taken into consideration.

As discussed in the introduction, in order to experiment with how RE works with GPT3, we have processed the same set of testimonies with the GPT3 API. Due to the limitation of the GPT3 API for processing long documents, we are required to divide the documents into smaller parts. This allows us to work within the constraints of the API and effectively process the content.

Though individual testimony consists of different types of relationships bonded with the environment, for this experiment only we have chosen the following relationships which describe the survivor experiences.

Relationship Category	Relationship
Biographical	born, die, learn, live, locate
Career	work, employ, travel, return
Holocaust Events	forced, transport, evacuate, arrest, deport, kill

Table 1: List of Relations

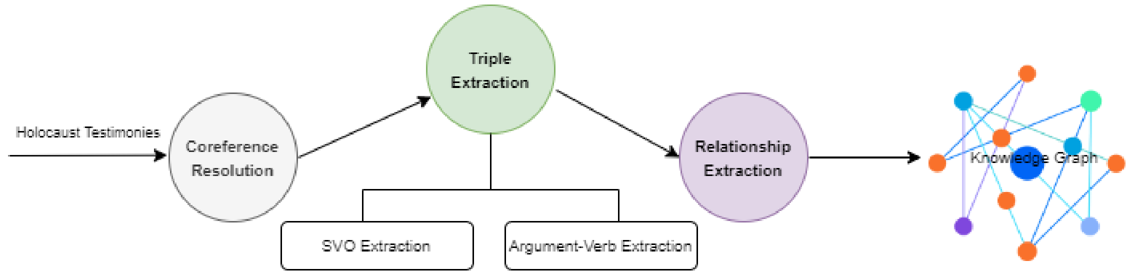

Figure 1: Proposed pipeline

3.2 Coreference Resolution

The coreference resolution component aims to identify and group together entities in natural language text that refer to the same entity. We have employed crosslingual coreference[2] Python library for this.

3.3 Triple Extraction

In this section, we conduct experiments using two methods for triple extraction and provide a detailed discussion of each method.

3.3.1 Method 01: Chunking based Extraction

From Chunking-based Extraction, we extract Subject, Object, and Verb extraction from the Holocaust testimony data. In this approach, we employed the chunking method to identify subject-verb-object (SVO) triplets to locate verb phrases, longer verb phrases and noun phrases. We define part-of-speech patterns, such as *"POS": "AUX"*, which help us identify the relevant components of the triplets. Table 2 refers to examples of method 01.

Subject	Verb	Object
Jews	taken	Auschwitz
Dr. Denes	transport	detention camp
Mrs Milman	employed	SS

Table 2: Examples for SVO Extraction (original text)

3.3.2 Method 02: Semantic Role Labelling based extraction

In this method, we employ the AllenNLP [3] model to determine the latent predicate-argument structure of a sentence and provide representations. After extraction of the Verb, we mapped with the arguments and relations defined in the sentence. To minimise the complexity of the arguments we consider only the First argument with verb either with another argument or else argument modifier. Table 3 refers to examples of method 02.

Argument	Verb	Argument1
Frau Meier	living	In 1936
Frau Morgenstern	escape	to Switzerland
Frau Gerard , a school principal	recommended	a young man

Table 3: Examples for SRL Extraction (original text)

3.4 Relationship Extraction with the GPT3

After applying the same set of testimonies to the GPT3 API, we retrieve the automatically generated relations from the model's output. To obtain these relations, we construct a prompt that describes the desired output, including the named entities specific to each testimony and the relationships observed by the GPT3 API.

3.4.1 Prompt Construction

In our approach, we create a specific prompt for each document, which is then inputted into the GPT model. The prompt is designed based on x, to provide the necessary context and information for the model to generate an appropriate response. It typically includes the following components:

Task Description x_{desc} We offer a concise summary of the task description for relationship extraction (RE) to think as a historian and present a predefined set of instructions to define Name entity tags. The task description is given as follows: *I*dentify the named entities with their named entity tags.

Demonstrations x_{demo} In the demonstration part, we reformulate each example by first showing the input prompt and then asking to generate the relation. The input prompt can be further enriched by asking to include the original

[2]https://pypi.org/project/crosslingual-coreference/
[3]https://demo.allennlp.org/semantic-role-labeling/semantic-role-labeling

Relationship Type	Original		GPT3	
	Method 01	Method 02	Method 01	Method 02
born	1,090	3	722	0
die	3,625	10	1,504	5
learn	2,715	75	272	17
live	9,513	234	2,998	136
work	7,148	322	3,317	229
travel	3,253	97	985	57
return	2,314	3	1,090	0
transport	1,626	7	1,122	5
find	203	74	66	43
locate	856	6	885	2
employ	985	13	241	6
forced	99	12	2,170	7
evacuate	841	4	353	2
arrested	723	28	2,134	23
deport	2,015	4	1,554	4

Table 4: Frequencies of the occurrences of relations

sentences.

$$(x^1_{demo}, y1^1_{demo}, y2^1_{demo}), .., (x^n_{demo}, y1^n_{demo}, y2^n_{demo})$$

where x^j_{demo}, 1 ≤ j ≤ k denotes the j^{th} input sequence and $y1^j_{demo}$, $y1^j_{demo}$ denotes the text which is remade from the label, e.g., list of named entity tag and the reformulated sentences

Test Input x_{input} Test input is the test text document needed to identify the relations. The prompt x_{prompt} for a Test input is constructed by concatenating the task description x_{desc}, a sequence of demonstrations

3.5 Visualisation in graph database

In this study, Neo4j [4] was utilised as a database management system to store and visualise the extracted relations in the form of a graph. In the triple, the subject/object pair or argument pair act as nodes in the graph and the verb act as the relation. Figure 2 illustrates the knowledge graph created for a set of triples.

4 Results and Comparative Analysis

In this section, we evaluate the results obtained from the methods described above.

We adopted the above-described methods to the original transcripts of testimonies and their GPT-derived relations. Table 4 describes the overall count of individual relationships identified according to the two methods. After obtaining the frequencies related to all relations, we manually access the relations as there is no computational method available to determine which relation is most relevant to the testimony, as a single relation may or may not be important for relationship extraction. We identified that the relations obtained using Subject-Verb-Object extraction (Method 01) have considerable random noises.

After conducting a comparative analysis between GPT3 results and the original testimonies relations, we identified that GPT3 results have less noise and they were properly arranged.

Furthermore, another finding from this research

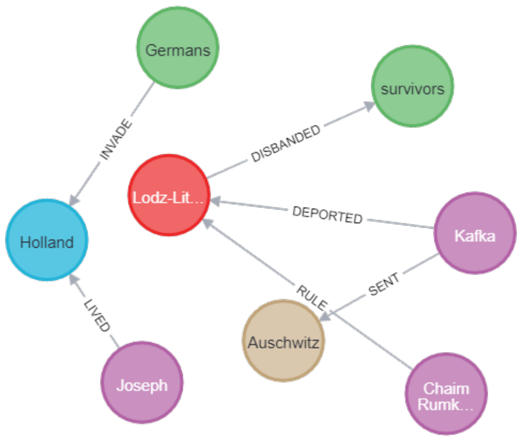

Figure 2: Knowledge graph visualisation of a sample set of triples

[4]https://neo4j.com/

is that the Argument-Verb Extraction method (Method 02) also failed to identify many relations in the context of the Holocaust, both in the original data and in the relations generated by the GPT model. This suggests that the Argument-Verb Extraction method may not be suitable for accurately capturing the full range of relations in this specific domain.

5 Discussion and Future Works

The primary contribution of this paper is the identification of relations through dependency-based SVO extraction, semantic role labelling, and GPT prompts in Holocaust testimonial data which describe about the survivor experience. These identified relations are then visualised in a graphical format, providing a clear representation of the relationships within the data. Based on our findings, we observed that the relations generated by the GPT3 API and the triplet extraction method based on subject-verb-object were able to provide the most accurate and effective results when identifying relations in Holocaust data.

Currently, our research primarily focuses on identifying relations from individual Holocaust testimonies. However, our future plans involve expanding this work to link individual testimonies together by establishing additional relations. This broader network of relations will enable a deeper understanding of the collective experiences, interactions, and events within the Holocaust, contributing to a more comprehensive and interconnected understanding of this historical period. Moreover, we plan to extend our experiment with name entity recognition combine with RE and integrate the results got with the SVO extractions as a part of triple integration to construct an *N-to-N* knowledge graph. By integrating the extracted triples from the Holocaust testimonies and mapping the predicates to a common schema, we aim to create a comprehensive and interconnected knowledge graph. This graph will capture the relationships, associations, and connections between entities, events, and concepts related to the Holocaust.

6 Conclusion

This research, evaluates and compares the performance of traditional rule-based dependency methods for relationship extraction with the recent advancements in LLMs. Through our proposed novel knowledge graph relationships can be visualised better than baseline approaches, hence proving the usefulness of the work specifically for the historians for better synthesis and presentation of the hidden information. This study represents the first-ever investigation into the domain-specific analysis of Holocaust text data. It focuses on examining the unique characteristics and challenges presented by this specific domain in the process of relationship extraction.

Acknowledgments

This paper is dedicated to the memory of Dr Vinita Nahar, who passed away while camera-ready version of this paper was being prepared.

References

Tobias Blanke, Michael Bryant, Reto Speck, and C Kristel. 2012. Information Extraction on Noisy Texts for Historical Research. *Digital Humanities*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Michael Bryant, Tobias Blanke, Mark Hedges, and Richard Palmer. 2010. Open source historical OCR: the OCRopodium Project. In *Research and Advanced Technology for Digital Libraries: 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings 14*, pages 522–525. Springer.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pre-trained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199.

Daan De Leeuw, Mike Bryant, Michal Frankl, Ivelina Nikolova, and Vladimir Alexiev. 2018. Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 58–66. IEEE.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A Survey for In-context Learning. *arXiv preprint arXiv:2301.00234*.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce

Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Amal Haddad Haddad, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2023. Deep Learning Methods for Extracting Metaphorical Names of Flowers and Plants.

Kai He, Yucheng Huang, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Virtual prompt pre-training for prototype-based few-shot relation extraction. *Expert Systems with Applications*, 213:118927.

Ruslan Mitkov Johannes-Dieter Steinert Isuri A. Nanomi Arachchige, Le An Ha. 2023. Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models. In *The 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 1–6, New York, NY, USA. Association for Computing Machinery.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A survey on complex knowledge base question answering: Methods, challenges and solutions. *arXiv preprint arXiv:2105.11644*.

Nikola Milošević and Wolfgang Thielemann. 2023. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, 75:100756.

Tomer Sagi, Avigdor Gal, Omer Barkol, Ruth Bergman, and Alexander Avram. 2016. Multi-source uncertain entity resolution at yad vashem: Transforming holocaust victim reports into people. In *Proceedings of the 2016 International Conference on Management of Data*, pages 807–819.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. *arXiv preprint arXiv:2305.08377*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. *arXiv preprint arXiv:2305.02105*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. Deep Learning Methods for Extracting Metaphorical Names of Flowers and PlantsHow to Unleash the Power of Large Language Models for Few-shot Relation Extraction? *arXiv preprint arXiv:2305.01555*.

Ningyu Zhang, Tao Gui, and Guoshun Nan. 2022. Efficient and robust knowledge graph construction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–7.