TECHNICAL UNIVERSITY OF MUNICH

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

Master's Thesis in Robotics, Cognition, Intelligence

Enabling Personal Communication for Voice-Based Health Assistants in Geriatric Care

Murilo Bellatini

TECHNICAL UNIVERSITY OF MUNICH

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

Master's Thesis in Robotics, Cognition, Intelligence

Enabling Personal Communication for Voice-Based Health Assistants in Geriatric Care

Ermöglichung persönlicher Kommunikation für sprachbasierte Gesundheitsassistenten in der Altenpflege

Author:	Murilo Bellatini
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	M.Sc. Phillip Schneider
Submission Date:	December 15, 2023

I confirm that this master's thesis in robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, December 15, 2023 Murilo Bellatini

Acknowledgments

I take this opportunity to express my gratitude to all those who have supported me during my academic journey, specifically during the composition of this master's thesis. I am especially thankful to Prof. Dr. Florian Matthes, whose guidance enabled me to work on such a significant and captivating topic. Additionally, I extend my appreciation to M.Sc. Phillip Schneider for his constant provision of valuable guidance and advice during the composition of this study. Lastly, I would like to express my gratitude to my family and friends for their unwavering support throughout my academic journey.

Abstract

This thesis explores the utilization of Language Models, such as BERT, ChatGPT-3.5-Turbo, and LLaMA, for automatically constructing Personal Knowledge Graphs from chat interactions in plain text, aiming to personalize communication in geriatric healthcare settings. While these Natural Language Processing technologies possess impressive capabilities, their ability to process complex relational data in personal interactions still hast to be explored using publicly available datasets. Our study strives to embed personalized responses into dialogue systems by focusing on structuring knowledge about individual users. We utilize Kitwood's psychological framework as a research lens for guiding personal relation extraction based on the geriatric communication literature. Additionally, we employed our novel SlideFilter technique, a custom data augmentation strategy for text, to enrich the DialogRE dataset. Although promising, this approach also revealed challenges in accurately and structurally organizing conversational data, indicating areas that require further development. Therefore, structuring personal user information within knowledge graphs is presently too intricate for existing Language Models utilizing public datasets. Our primary findings include insights into how Large Language Models handle relation extraction, why it is a complex process, and the series of rigorous experiments used to validate these insights. This study offers valuable insights into the capabilities and limitations of current Natural Language Processing technologies and guides future exploration of data structure simplification and hybrid model development. These discoveries significantly advance the personalization of communication tools in geriatric healthcare.

Contents

1. Introduction

1.1. Motivation

As the healthcare industry faces increasing hurdles, including an aging population, workforce shortages, and rising chronic diseases, the importance of technology in addressing these challenges has never been more critical [1]. The advent of conversational agents has emerged as a major advancement, gaining momentum both within the healthcare domain and beyond [2]. Intelligent computer systems can comprehend and generate human language with Natural Language Processing (NLP) technologies, operating as conversational agents or digital assistants.

These systems have significant potential not only to improve the quality of life for the elderly and people with chronic illnesses but also to offer various benefits in assisting patients, caregivers, and medical professionals [3]. Their abilities include conducting regular health evaluations, managing medication schedules, monitoring overall health, and even engaging in social interactions to foster the well-being of patients. Therefore, having automated systems able to handle these tasks, can also reduce workforce overload.

The elderly population faces unique challenges, such as reduced mobility and social isolation, which can be mitigated by technological advancements. Previous research has shown that social isolation and loneliness are associated with various negative mental and physical health outcomes [4, 5]. Improving the personalization features of voice assistants is essential for mitigating the detrimental effects of elderly loneliness. These advanced digital agents can offer companionship and foster a sense of connection, effectively addressing the core issue of social isolation. AI-powered interactions from these assistants can aim to decrease the risk of depression among elderly users by providing a responsive and empathetic communication companion. Ensuring this type of interaction is present is not only crucial for promoting the emotional well-being of this group, but also their mental fitness. It can encourage the constant recollection and narration of joyful experiences. Maintaining their overall health and emotional satisfaction entails ensuring they feel heard and comprehended.

On the other hand, the rise of conversational health assistants has already had a considerable impact on healthcare, offering benefits. However, there is still room for improvement. This study aims to enhance the capabilities of voice-based health assistants, making them more responsive and practical tools for geriatric care. Despite their functionality, current voice assistants have two major drawbacks: they provide

generic responses and lack personalization [1], especially in casual conversation settings. These challenges are particularly difficult for older adults with unique healthcare needs and a preference for interpersonal communication.

To effectively meet the healthcare needs of the elderly population, voice-activated digital assistants must do more than just provide basic functionality; they must also establish personal connections. By focusing on customized output through the collection of personal data, vocal assistants can be developed to cater to the healthcare needs of the elderly population without neglecting their emotional and psychological well-being.

The extraction of personal information and the creation of Personal Knowledge Graphs (PKGs) is essential in facilitating personalized dialogues. PKGs are not just auxiliary features; they play a vital role in providing personalized interactions, as suggested by Balog and Kenter [6]. Empathy, as the *Cambridge Dictionary* elucidates, is "the ability to share someone else's feelings or experiences by imagining what it would be like to be in that person's situation" [7]. Within this framework, empathy is expressed as the ability of the system not only to recognize and understand the emotional states and personal circumstances of users, but also to respond to them with appropriate sensibility. Accordingly, PKGs constitute a vital element of empathetic care as they provide the system with the necessary tools to formulate responses that are contextually and emotionally relevant to the user.

Designing and constructing PKGs, however, is not an easy task since it involves many challenges related to computational resources and algorithmic complexity. To successfully capture individuals' essence, the system must objectively understand their requirements, continuously monitor and extract mentions of people and places within the dialogue, understand their relationships, and store this information in a format that is both retrievable and updatable.

With a comprehensive understanding of the user's data, we can employ a variety of Retrieval-Augmented Generation (RAG) strategies to extract and use this information effectively, as suggested by Lewis et. al. [8]. In this context, Pre-trained Language Models (PLMs) and Large Language Models (LLMs) emerge as powerful tools due to their advanced text generation capabilities, as demonstrated by Brown et al. [9] and Wei et al. [10]. These models, particularly LLMs, are adept at processing large datasets, which enables them to generate accurate predictions for tokens that are both contextually relevant and coherent.

Moreover, we have successfully validated the entire pipeline in a proof-of-concept format, ensuring that the combination of structured knowledge and an LLM-based RAG engine is a promising framework for personalized conversations. The whole conversational experience was thoroughly tested for its effectiveness and practicality. Therefore, we believe that this framework holds great potential for achieving personalized conversations. However, the primary focus of this thesis will remain on constructing PKGs. This decision stems from the vital significance of PKGs in enabling genuinely personalized and empathetic dialogues [6]. It also results from the intricate

nature of this undertaking. In-depth research and development of text generation techniques will be postponed for future work due to their inherent complexities and challenges.

Given this, our study focuses on the implementation of sophisticated and controlled mechanisms to extract user-specific data from chat histories in plain text. In summary, the healthcare industry requires personalized and empathetic communication which most voice-based assistants lack. This is not only about increasing efficiency but also about placing emphasis on human-centric design and interaction. As we enter the age of AI and machine learning, there is potential for the development of these technologies to fulfill our fundamental human needs. This highlights how personal interaction between voice assistants and users is a topic worth researching.

1.2. Problem Statement

This thesis aims to address the significant difficulties concerning extraction of personal information from dialogues, therefore enabling the construction of PKGs customized to individual users. The goal is to seamlessly incorporate this graph into a data retrieval system, enabling more substantial and individualized follow-up queries.

Problem 1 - Domain Specificity The evaluation of the effectiveness and construction of PKGs closely relates to the availability and appropriateness of datasets. While diverse relation-type datasets are abundant, there are significantly fewer choices when it comes to dialogue-specific datasets. The problem worsens when focusing on personal relationships, and it becomes even more restricted in the field of geriatric care; refer to Figure 1.1. Thus, the main challenge in this thesis is to address the problem of insufficient data without undertaking the difficult and costly task of creating a customized dataset, which is unfeasible due to time constraints. Consequently, the emphasis is on identifying available datasets in published literature that are relevant to the field and can serve as dependable benchmarks for evaluation.

Problem 2 - Definition of Taxonomy Developing an effective taxonomy is vital to creating a PKG with relevant content. Therefore, a structured approach is necessary to overcome constraints and ensure a dependable taxonomy. However, this task is complicated by the domain-specific data, which limits the available datasets and relevant entities and relationships. This project aims to develop a taxonomy that accurately captures the quantity and types of relationships and entities that are specific to geriatric care. This refined taxonomy should be easily understood and enable effective management of data structures in future applications, including conversation management. An excessively complex classification system could lead to a graph that is hard to interpret, while a too simplistic one may not suffice for the aim of generating empathetic and personalized interactions.

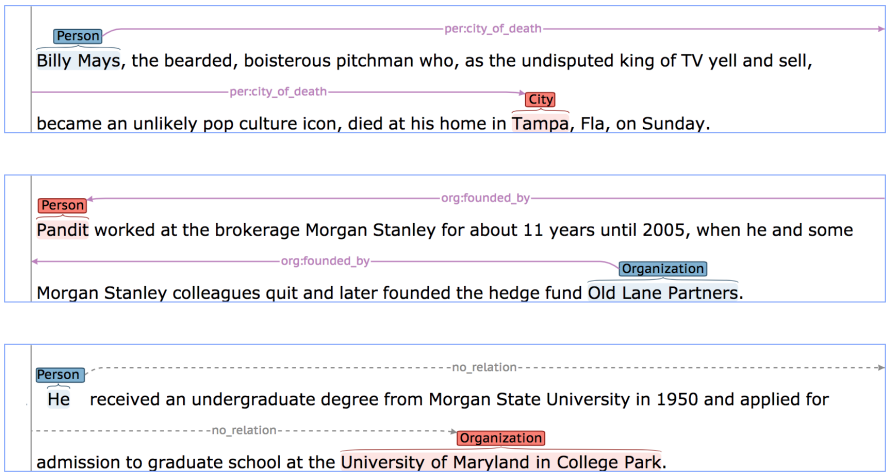


Figure 1.1.: Exemplary **TACRED** data, for instance, lacks data that captures personal relationships within our target domain of geriatric care [11]

Problem 3 - Entity Recognition Determining which entities to include or exclude in the PKG can be a difficult task, particularly when taking into account sensitivities specific to the field of elderly care. This selection process has a significant impact on the quality and usability of the PKGs. For instance, the incorporation of cardinal numbers like age might introduce unwanted complexity or noise into the graph. Thus, the challenge is to create criteria for filtering entities that meet the specific requirements of geriatric care while also maintaining a functional and easily interpretable PKG.

Nurse: Good **morning**, **Mr. Smith**! How are you feeling **today**? I see you'll be 78 in **two weeks**.

Patient: **Morning**! I'm feeling okay, thanks. I've been living in this **facility** for **4 years now**.

Figure 1.2.: Sample dialogue illustrating the challenge of entity recognition in geriatric care. Entities are highlighted in **purple**. This complexity is evident, even in brief exchanges, when extracting such entities for inclusion in the PKG.

Problem 4 - Relation Extraction (RE) The key to constructing a robust PKG lies in effectively extracting relationships between identified entities. This requires not just pairing entities but also precisely defining the nature of their relationship. The challenge comprises two main elements: first, discerning whether a relation exists

between two given entities, an undertaking that is subject to contextual nuances. Second, it is crucial to categorize the type of existing relationship, which demands a refined comprehension of both the domain and entities involved. These two subproblems are essential in developing an accurate and effective PKG that serves as the basis for future applications, including personalized conversational systems. This can be understood as the sub-problems below, also shown in Figure 1.3.

- Sub-problem 1 (Relation Identification (RI)): The task is to determine if there is a significant relationship between two individual entities within a particular context.
  - Example: In a dialogue, you hear, "My son spoke to Dr. Brown about my arthritis." The names "son" and "Dr. Brown" appear, but do they know each other professionally, or was it a one-time thing related to your care?
- Sub-problem 2 (Relation Classification (RC)): Once it's clear that a relationship exists, the challenge is to classify its nature.
  - Example: Continuing with the above dialogue, if it turns out that the son often consults with Dr. Brown about your health, then their relationship could be classified as "acquaintance". But could this be inferred from the dialogue alone?

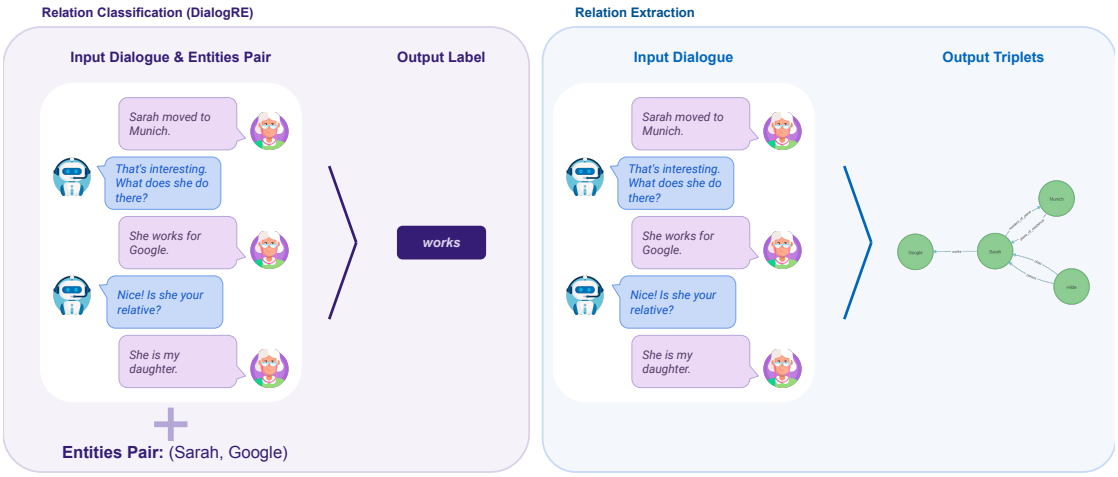


Figure 1.3.: Contrasting RC and RE: RC in the DialogRE paper [12] predicts a specific relation label (such as "works") for an entity pair given a dialogue. In contrast, RE generates a comprehensive list of all relationships existing in the text from the input dialogue. The list can be visually represented through a graph, providing a clear mapping of all entities and their corresponding relations.

Scope This thesis explores a proposed three-step iterative approach to construct and utilize a PKG in the field of geriatric care. The approach aims to facilitate dialogue, curate knowledge, and provide personalized interaction. While the comprehensive model intends to address numerous crucial aspects for the development of a dynamic and flexible PKGs, this study mainly concentrates on the second phase - knowledge curation. This complexity necessitates extracting relations from plain text, as demonstrated by the aforementioned issues, thus making it the cornerstone of the entire approach.

During the initial stage of the proposed approach, users would be prompted with a series of predetermined questions for dialog stimulation. These questions have been strategically designed to gather relevant data on elder care topics, including daily routines, social interactions, and personal preferences. This phase will establish the foundation for subsequent steps and aid in the initiation of the PKG.

The thesis focuses on the knowledge curation phase, where significant challenges exist. Advanced NLP techniques are employed utilizing PLMs for accessing prior knowledge derived from previous datasets. The aim is to facilitate the efficient extraction and mapping of entities' relationships within the PKG. The emphasis is on identifying and contextualizing personal relationships, organizations, and social groups with which individuals interact, rather than medical terminologies. These aspects are captured through casual, dialogue-based exchanges and are critical to developing a nuanced understanding of each user. Figure 1.4 illustrates the target task of our study, which involves extracting relationships from input dialogues. An example is provided to demonstrate this process.

The third phase, personalized interaction, is another aspect of the overall model under consideration, but it is not within the scope of this study. During this phase, the updated PKG is applied to offer customized follow-up questions and tailored care recommendations that meet the specific requirements and medical history of each user.

The iterative nature of this model enables continuous improvement of the PKG, presenting the possibility of enhanced user experiences over time. This approach is designed to evolve and enhance with each iteration, progressively boosting both user engagement and satisfaction.

1.3. Research Questions

RQ1: Essential Content for Personalized Communication in Elderly Care

Question: What are the essential concepts and entities for effective personalization and user engagement in geriatric care? We will conduct a comprehensive literature review to identify appropriate conversation topics that foster user engagement in geriatric care. Our objective is to thoroughly analyze communication protocols recommended by experts in medical and psychological fields to establish best practices. This approach

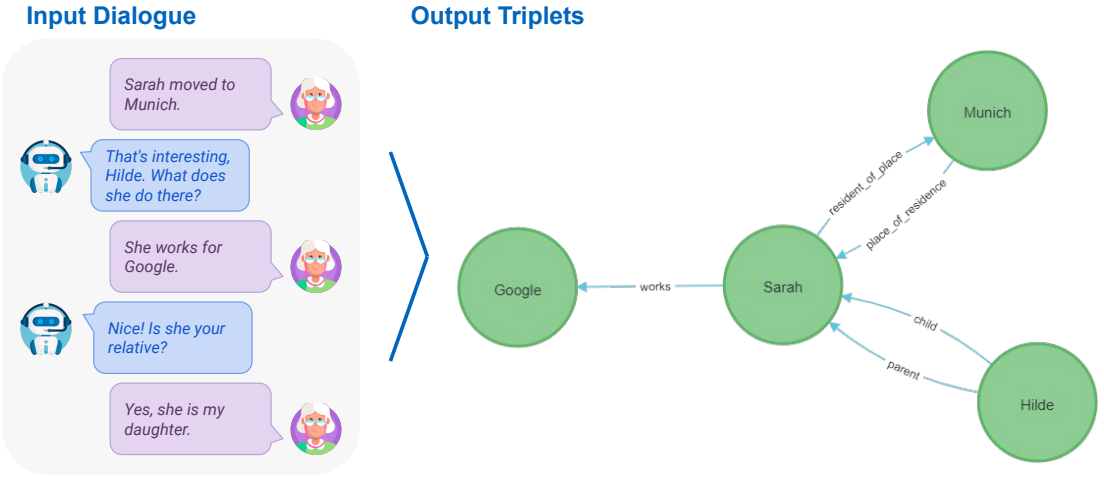


Figure 1.4.: The above figure depicts our study's objective to convert the provided input dialogue between a health assistant and an elderly user, who engages in a casual conversation with personal relevance, to the mentioned relation triplets. Concurrently, the assistant, while empathetic, extracts the relations in the shown graph format on the right-hand side.

enables the integration of technical requirements and medical necessities to develop a robust user model.

RQ2: Strategies for RE and PKG Construction

Question: What are the appropriate methods and data sets for populating the PKG with relevant information? This inquiry is essential to our thesis. Our objective is to utilize existing knowledge to meet the unique needs of elderly care. Therefore, we will investigate various RE approaches that apply to both personal and geriatric domains.

RQ3: Integration of Curated Knowledge into Conversations

Question: How can curated knowledge be integrated into the conversational model to facilitate personalized interactions? Although not the primary focus of this thesis, this inquiry remains pertinent in ensuring the constructed knowledge graph's utility for future personalized interactions. This is in line with the three-step approach depicted in Figure 1.5 and 1.6.

RQ4: System Performance Assessment

Question: What are effective evaluation methods for assessing the system's performance? This investigation aims to identify suitable metrics for measuring the effectiveness

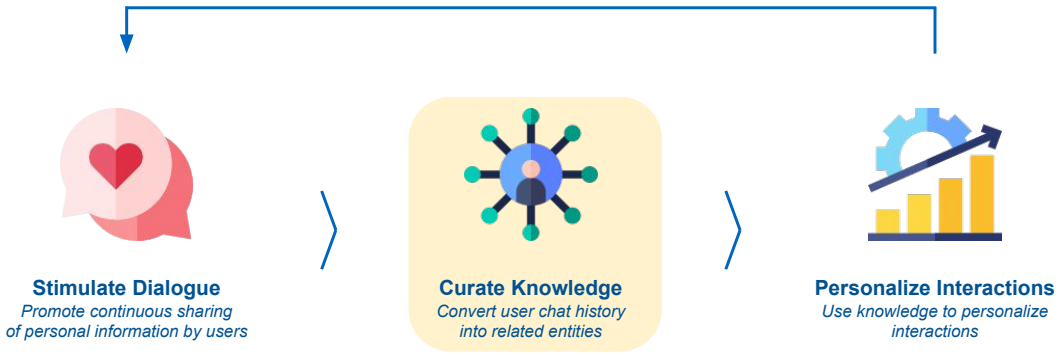


Figure 1.5.: High-level diagram presenting a three-part approach to constructing a PKG for geriatric care. The first phase aims to facilitate dialogue among users to encourage the sharing of relevant personal information. The second phase, the main focus of this thesis (highlighted in yellow), involves organizing this information into a meaningful knowledge graph through curation of knowledge. Lastly, the third phase involves customizing interactions by utilizing the curated knowledge for personalized user engagement.

of converting unstructured text to structured knowledge using PKG. To achieve this goal, we will review prior research on similar challenges and adapt their techniques to our specialized application of personalized communication for geriatric care. We have made the code for our experiments available in our GitHub repository [1].

1.4. Outline

To address the research questions, this thesis is structured as follows. Chapter 2 presents the foundational knowledge necessary applied throughout the thesis. Chapter 3 provides an overview of the related research. In Chapter 4, we discuss the methodology used to obtain our results. First, we explain the criteria for selecting Language Models (LMs). Then, we outline the methods and datasets used for RE in the personal domain. Chapter 5 presents the our findings. Chapter 6 offers a detailed analysis and insights derived from these results. Upon presenting the key findings, we conclude with a discussion of the limitations of our work. Finally, the last chapter provides a summary of the thesis and suggests avenues for future research.

[1] `https://github.com/murilobellatini/RelNetCare`

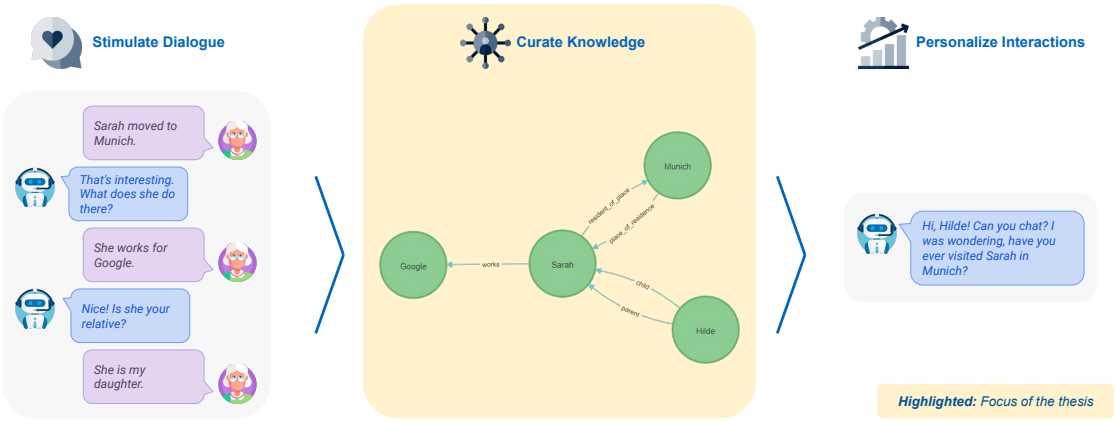


Figure 1.6.: A specific example of our envisioned three-part approach is presented, illustrating each step clearly. Our main focus is the thesis task, which generates a graph of relation triples using only the input dialogue.

2. Fundamentals

This chapter provides a necessary foundation for navigating the subsequent sections of this thesis. It introduces dialogue systems with emphasis on Conversational Personalisation in a clear and concise manner. Moreover, pivotal to our methodology, this chapter also introduces Knowledge Graphs (KGs) along with key concepts for their automated creation from plain text. Within this context, the chapter details the significance of Named Entity Recognition (NER), RI, RC and RE. The process of contributing to the structured representation of user knowledge is conceptualized. To conclude the chapter, a dedicated section explores PLMs, providing a introduction on transformers—the leading methodology for crafting LLMs—and analyzing the characteristics of different renowned language models.

2.1. Dialogue Systems

According to Deriu et al. [13], there are three main classes of Dialogue Systems (DSs). These are conversational, task-based, or question-answering systems. They are controlled by the user in natural language and can be used for different use cases such as virtual assistants or information retrieval systems. DSs provide the user with a text-based, speech-based, or multimodal interface. The dialogs are usually structured in sequences.

- *Conversational agents* strive to imitate human behavior and are tailored specifically for unstructured open-domain dialogues. An initial illustration of such a system is recognized as ELIZA [14].
- *Task-based systems* assist users in achieving a particular objective. They are tailored to a specific field and adhere to a predetermined format. For instance, an in-car virtual assistant that aids the driver in inputting their intended destination into the GPS.
- *Question Answering systems* attempt to answer natural language questions posed by the user. They typically use external knowledge bases to retrieve the required information and may be capable of single-turn or multi-turn interactions that support follow-up questions.

2.1.1. Conversational Personalisation

Maintaining a user-specific knowledge base is crucial for achieving sustained personalization. This is based on the fundamental notion that a system must retain records of past interactions to engage in meaningful future interactions. By leveraging historical data, the system can retrieve prior information to pose related follow-up questions and reference past discussions in future dialogues.

Among several methods to achieve this, KGs stand out, which we will delve into in the subsequent paragraphs.

KG The "knowledge graph" concept has various interpretations in academic circles [15, 16, 17]. For this discussion, it represents a data graph that encapsulates real-world knowledge. Nodes in this graph represent entities, while edges depict the relationships between them. These graphs often adopt data models such as directed edge-labelled graphs [18]. The scope of knowledge within the graph can vary from basic facts, like "Santiago is the capital of Chile", to more nuanced statements. While rudimentary facts are symbolized as edges, more sophisticated ones might necessitate advanced constructs like ontologies. Knowledge can either be imported from external sources or mined directly from the graph, with deductive and inductive techniques further augmenting this knowledge pool.

Balog and Kenter, in their research from Google [6], suggest that PKGs are powerful instruments for safeguarding user-specific data. Essentially, a PKG is a KG designed to house personal content. According to Balog and Kenter, utilizing PKGs for data storage has three significant benefits.

1. Catalyzing personalized search and content recommendations.
2. Enhancing personal information oversight.
3. Augmenting personal digital assistant functionalities.

Their studies emphasize that PKGs play a pivotal role in cataloging entities—like family connections and hobbies—and their associated ties to the user. This facilitates the continual growth and oversight of such graphs. An exemplar of a PKG as illustrated by Google Research is showcased in Figure 2.1.

2.2. Personal Knowledge Graph Construction

At the core of our project, as illustrated in Figure 1.5 and 1.6, is the aim to curate user knowledge by building PKGs. This involves establishing nodes and edges to generate a complex, web-like structure. Various fundamental concepts supporting this endeavor warrant further elucidation.

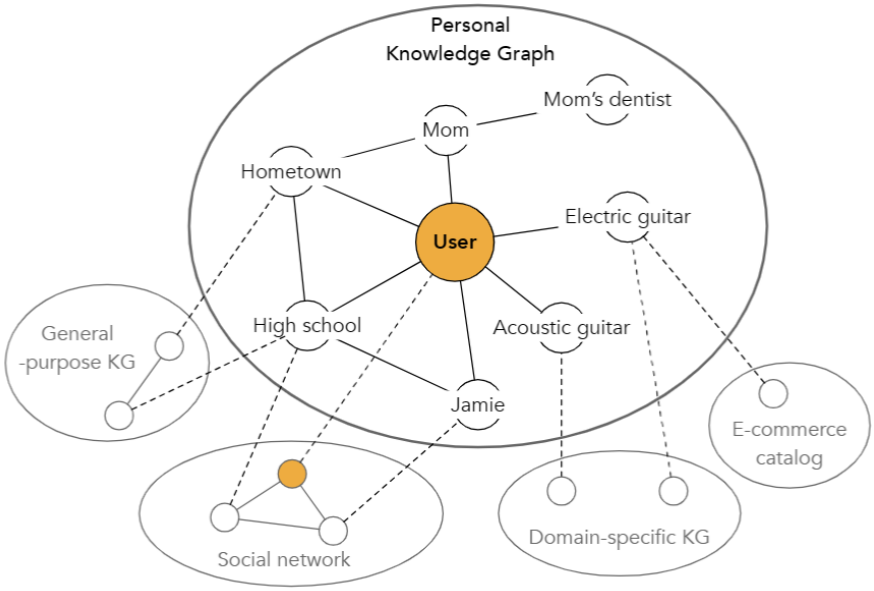


Figure 2.1.: PKG Example from Google Research [6]. The diagram depicts entities as nodes and their pairwise relations as edges. The structure showcases the interconnection of individuals and organizations in a personal context and demonstrates its applicability in other domains such as social networks and e-commerce catalogs. This structure provides an additional layer of control for applications aiming at precise user data manipulation.

NER NER discerns and categorizes named entities in text, assigning them to predefined classes such as people, organizations, or dates. [19]

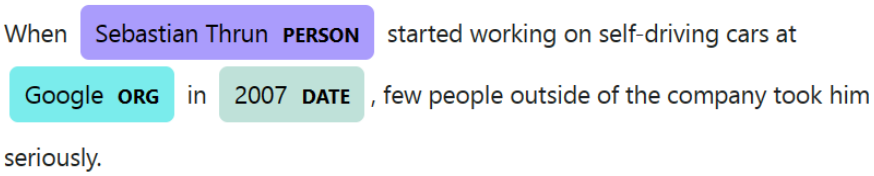


Figure 2.2.: Illustration of NER using the SpaCy visualizer [20].

Triplet A relation triplet involves two entities connected by a relation. Such triplets are often structured and hosted by knowledge bases (KBs) like Freebase [21], DBpedia [22], and Wikidata [23]. In a knowledge graph, nodes interconnect, and relations are often defined in pairs. This means that a relation may appear as a triplet, indicated by the format entity-relation-entity (e.g., Mary-isSpouseOf-Robert) [24] .

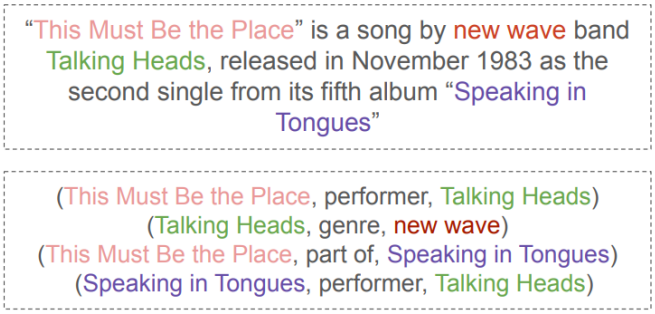


Figure 2.3.: RE example from the REBEL paper [25], where input text is transformed into an output list of triplets with the format node-relation-node.

RE Perceiving a knowledge graph as a collection of triplets (or relations) offers clarity. The main technique emphasized in this thesis involves extracting triplets from unstructured text. Consequently, RE converts plain text into a list of triplets.

RC Literature, such as DialogRE [12] and HiDialog [26], explores RC versus RE. RC predicts label attribution for a specific entity pair while presupposing their relation's existence. This prior assumption may lead to skewed outcomes since detecting the relation's presence initially is a more complex task. This inherent bias could reduce practicality in real-world scenarios.

Constructing a package is considered a sequence-to-sequence (seq2seq) task, wherein natural dialogue is converted into a sequence of relations. There are several ways to accomplish this task, one of which is by using an end-to-end pipeline. This process moves systematically, starting with extracting entities from the input text, followed by identifying and classifying relationships, and culminating in the creation of a set of triplets. Alternatively, addressing end-to-end tasks through seq2seq models, especially LLMs, presents a simplified solution. This method offers practicality and flexibility while reducing the number of required models. Therefore, it becomes easier and less burdensome to adapt, particularly when introducing new datasets. Chapters 4 and 5 provide further insights into these methodologies.

2.3. Pre-Trained Language Models

PLMs are machine learning models that have already undergone extensive training with large amounts of data. They offer the ability to handle general tasks or targeted fine-tuning for domain or task-specific instructions. To begin, we present an overview of the widely used transformer architecture for PLMs. Following that, we provide a survey of popular LLMs and their attributes.

2.3.1. Transformers

All information in this section and its subsections is drawn from [27]. Vaswani et al. introduced the transformer architecture in 2017 with their paper "Attention is all you need" [28]. This paper caused a paradigm shift, as most NLP tasks currently employ transformer models as state-of-the-art approaches. Using self-supervised learning, models are trained on extensive textual data sourced mostly from the internet. This learning technique eliminates the need for human labeling as tasks center around the texts as the reference point. A training method for self-supervised learning is exemplified by masking a word within a body of text, predicting its content, and comparing the predicted output to the original term. The models acquire statistical language proficiency through training on diverse texts. These models are referred to as foundational models or PLMs, possessing general language knowledge but not yet trained for specific duties. They can be tailored with task-related training data to optimize a PLM for a specific use case. For a detailed explanation of the transformer architecture, please refer to the original paper by Vaswani et al. [28]. In the following section, we outline three distinct types of transformer models.

Auto-Encoding Transformer If an architecture only includes the encoder component of the transformer, it is referred to as an auto-encoding transformer or encoder model. Within the architecture of transformers, there is a technique called attention that identifies the most crucial words to consider when predicting a token. Attention layers can access the whole input sentence for encoder models. Thus, when a word is obscured in a sentence, it can consider the terms to the left and right of the mask through bi-directional attention. Pre-training auto-encoding transformers can be accomplished using masked word prediction, which correlates with predicting a corrupt (masked) word in a sentence. This design is most suitable for sentence classification, named entity recognition, and extractive question answering. One of the most favored models in this category is Bert.

Auto-Regressive Transformer The auto-regressive transformer model employs only the decoder aspect of the transformer architecture. Unlike encoder models, its attention layers are limited to words preceding the current word. Consequently, during pre-training, it predicts the subsequent token of an incomplete sentence. Since the source text is known, the forecasts are matched against the true values to gauge the loss and modify the model weights. These models are best suited for generating textual content, with GPT-2 serving as one illustration from this category.

Sequence-to-Sequence Transformer Encoder-decoder models, also referred to as sequence-to-sequence transformers, utilize both components of the transformer architecture outlined in the original paper. with this approach, the attention layers of

the encoder have access to the full input sentence. Conversely, the attention layers of the decoder can only attend to the words that come before the given word. Multiple consecutive words can be replaced with a single mask when pretraining such models. The primary objective of this task is to predict the original words accurately. Encoder-decoder models are commonly utilized for summarization, translation, or generative question answering purposes. Two well-known models include BART and T5.

2.3.2. Popular Language Models

This section outlines the fundamental concepts underlying popular Language Models. We provide a succinct summary of the training strategies and datasets used, as well as an overview of key model properties. For a more detailed understanding, we recommend consulting the relevant papers.

BERT stands for Bidirectional Encoder Representations from Transformers, which was introduced by Google in 2018 [29]. Unlike traditional models that scan texts in a single direction, BERT reads texts bidirectionally, thus optimizing the understanding of context in both directions. It has achieved state-of-the-art results on various Natural Language Processing tasks. The model is pre-trained on two tasks.

- *Masked Language Modeling*: Predicting a word in a sentence by masking it out.
- *Next Sentence Prediction*: Identifying the correlation between two provided sentences.

BERT was trained on a total of 3,300M words, which were composed of two data sources - BooksCorpus (800M words) [30] and English Wikipedia (2,500M words).

LLaMA refers to a set of models created by Meta, which were made available with a non-commercial license for public use in February 2023. The models are available in four different sizes: 7B, 13B, 33B, and 65B. Specifically, the smallest LLM (7B) in this family uses an adapted transformer model architecture and has been pre-trained on publicly available and open-source data. The training includes multiple datasets:

- English CommonCrawl [31] and C4 [32]
- GitHub projects distributed under open-source licenses like MIT
- Wikipedia entries in 20 different languages
- Two Book datasets
- ArXiv for scientific data
- StackExchange for questions and answers

The 7B model was trained using approximately one trillion tokens, whereby most tokens were only utilized for one epoch during training [33].

LLaMA - Finetuned This model is based on LLaMA but has been fine-tuned for our specific tasks using the Low-Rank Adaptation of Large Language Models (LoRA) [34] approach instead of full fine-tuning. LoRA significantly reduces trainable parameters and GPU memory requirements while performing similarly to regular fine-tuning. We utilized the scripts provided by the git-cloner repository on GitHub [35], which are based on the Vicuna training code [36] but optimized for GPUs with low-resource capability. Our GitLab repository 1 contains the training datasets, as well as the exact command, including hyperparameters. The resulting models are capable of supporting a context length of 512 tokens and possess the same license as LLaMA.

GPT-3.5-Turbo is a closed-source model developed by OpenAI. Limited information is available regarding the number of model parameters or the datasets used for training as this has not been disclosed. However, it is expected that GPT-3.5-Trubo has a size of approximately 150B parameters; although, this is not officially confirmed. In addition, the standard model's supported context length is 4k tokens [37]. The model is an improvement of text-davinci-003 (another proprietary model from OpenAI) and optimized for chat use-cases [38].

3. Related Work

This chapter examines previous research on RE and classification, both key components of personalized conversational systems. It begins by outlining classic approaches before delving into the advancements of NLP technologies and their relevance in geriatric care. This chapter serves as the foundation for our thesis, which aims to address the unique communication needs of older adults.

RC The field of RC in NLP has seen extensive development recently, with researchers adopting progressively complex architectures. One example is Yan Xu and his team, who achieved an impressive F1-score of 86.1% on the SemEval-2010 Task 8 through the use of Deep Recurrent Neural Networks (DRNNs) combined with a data augmentation technique [39]. This method was successful in overcoming the limitations of shallower models. Conversely, Rui Cai and colleagues achieved noteworthy progress with the development of the Bidirectional Recurrent Convolutional Neural Network (BRCNN) [40]. This novel model combines both convolutional and recurrent layers, showcasing a great ability to identify the subtle relationships and their directions between entities. While much research within this sector has focused on textual corpora, recent endeavors have delved into the dialogue domain. Qi Jia and colleagues have made strides in this direction with "DDRel," which established a specialized dataset extracted from movie scripts to address unique challenges presented by dialogue media [41]. Perhaps most relevant to our discussion is the study by Dian Yu and colleagues, titled "Dialogue-Based RE" [12]. Although the title may be deceiving, the study centers on dialogic RC, rather than RE, and highlights the significance of speaker-based data in identifying connections. This pioneering work remains critical to the field. Importantly, this study utilizes BERT and demonstrates promising outcomes in detecting relationships within personal dialogues. These results provide a critical foundation for projects like our own, which seek to comprehend and promote significant interactions, particularly in geriatric communication settings.

RE The subfield of RE in Natural Language Processing is rapidly gaining popularity due to its broad applicability and flexibility. Chenguang Wang et al. made a significant contribution to this field with their paper on 'Zero-Shot Information Extraction' presented at EMNLP 2021 [42]. Their innovative methodology reframes extraction tasks as text-to-triple translations, utilizing pre-trained language models to accomplish notable zero-shot performance. This approach presents a substantial departure from customary

techniques, introducing the possibility of scalability and adaptability across multiple domains without necessitating task-specific training data. Pere-Lluís Huguet Cabot and Roberto Navigli's "REBEL" employs an end-to-end strategy, built on the BART architecture, to streamline pipelines typically used for high-quality RE [25]. This model adds another layer of sophistication to the process, making it particularly groundbreaking. The REBEL system has achieved state-of-the-art performance across multiple benchmarks, establishing a new standard in this field. Its remarkable adaptability sets it apart, as it can be fine-tuned to identify more than 200 different relation types, thus providing a much-needed level of flexibility that is lacking in traditional approaches. This attribute makes REBEL more than just an incremental step forward; it is truly a groundbreaking advancement in RE. Its capability to streamline intricate workflows without sacrificing top-notch performance makes it a promising tool for future research and practical applications alike.

Geriatric Communication Communication with elderly patients often neglects their unique psychological needs, resulting in suboptimal dialogue that fails to recognize them as individuals. In fact, general approaches toward geriatric patients often create a deficit-oriented view of aging, thus overlooking the need for patient-oriented communication tailored to their specific needs and abilities [43, 44]. Tom Kitwood's person-centered framework identifies five psychological needs - comfort, attachment, identity, occupation, and inclusion - that should guide our understanding of effective communication for geriatric patients [43]. These needs can be translated into entity-relation pairs of personal information, enabling the creation of dialogues that can fulfill these crucial psychological elements. This approach aligns with both the International Classification of Functioning, Disability, and Health (ICF) and the biopsychosocial model of health [45].

In this thesis, we conduct a comprehensive analysis of various techniques aimed at extracting relationships within the specialized field of geriatric care. Our primary objective is to customize conversational interactions for elderly patients by integrating available datasets and technologies with current research in the geriatric communication field. To achieve this goal, we begin by conducting an in-depth evaluation of methods that are ideal for our specific target population. Successful completion of this fundamental task will facilitate the application of these techniques in real-world scenarios, resulting in the production of consistent and credible responses that hold immense value for the end user. The DialogRE dataset [12] is expected to be particularly advantageous owing to its plentiful relations that conform to Kitwood's framework [46] for entity-relation pairs. It is crucial to recognize the constraints of a one-size-fits-all strategy. Such an approach endangers neglecting the different demands and abilities of each patient, going against the aims of patient-centered methods. Therefore, it is necessary that the datasets and methodologies used accommodate discussions that consider individual psychological needs, supported by existing research [44].

Figure 3.1.: Tom Kitwood's person-centered framework emphasizes the five essential psychological needs for effective communication with dementia patients, which can also be applicable to elderly people - comfort, attachment, identity, occupation, and inclusion. These core needs are crucial in ensuring effective communication and should be considered in geriatric care practices.

4. Methodology

4.1. Research Procedure

This chapter presents the research procedure utilized in this thesis. An initial literature review is conducted to identify existing studies on RC and RE utilizing LMs, essential components for constructing the PKG. Two PLMs, i.e. BERT and LLaMA, which were developed through varied training approaches, are subsequently selected. Our objective is to assess the influence of various types of PLMs with diverse attributes on downstream assignments such as identifying and extracting relationships.

We examine data preprocessing techniques to measure their impact on model performance for both individual tasks and the entire pipeline. Since our domain has limited data, this stage plays a crucial role in determining the importance of data for our particular tasks. In addition, we test prompting methods to enhance the output of the LLMs for the mentioned duties. As the field of LLMs and prompt engineering is rapidly developing, our primary objective is to implement established methods that increase the likelihood of its generalization with other models.

While establishing the models, employing effective preprocessing methods, and identifying suitable prompting techniques, we progressively fine-tune the LLMs. First, we start with RC having an explicit RI step, then we resolve both of them jointly, and finally RE. We conduct each experiment in an ablation study style, keeping all parameters constant except for one, to ensure understanding of the isolated impact of each parameter. This may entail testing different models on a single dataset or one model on various datasets. Standardized evaluation criteria are followed for each experiment to ensure comparable metrics. It should be noted that every experiment is guided by a hypothesis and provides insights into its strengths and weaknesses. This systematic approach is maintained throughout the thesis, providing a comprehensive strategy for addressing the encountered challenges.

4.2. Literature Review

This section details our methodology in answering our first and second research questions. We aim to locate relevant prior studies on 1) personalized geriatric communication in the context of elderly care, and 2) the utilization of PLMs for RC and RE. To achieve this, we conduct a comprehensive literature review employing scientific databases to identify suitable academic works. Firstly, we create a search query for

each of our two focus areas. Afterward, we utilize search queries across three academic databases: ACM Digital Library, IEEE Xplore, and Springer. Our search filters studies by title, keywords, and abstract to find matching results. Additionally, we conduct forward and backward snowballing based on preliminary findings. Forward snowballing involves reviewing work that cites relevant papers, while backward snowballing considers papers cited by relevant articles [47]. We establish criteria for inclusion and exclusion to guarantee that only pertinent academic work is incorporated. This assists in filtering the initial outcomes to obtain the ultimate set of considered works.

Inclusion and Exclusion Criteria We only consider peer-reviewed papers that have been published in academic journals or presented at conferences to ensure the quality of our sources. Additionally, the papers must be written in English and focus on English texts or datasets, as our research is centered around RI and RC in the English language. We require full access to the contents of the academic work. Otherwise, the text must be excluded as it is vital to consider all the information from the approaches under consideration. Lastly, the studies should closely relate to RC and RE using PLMs. This determination is made by screening the preliminary results title and abstract.

Search Queries We developed two search strings for querying scientific databases. Both employ Boolean operators to connect multiple keywords. The first search string centers on semantic parsing, specifically generating database queries. To achieve this, we include "relation classification" and "relation extraction" using the OR operator. Additionally, our focus lies solely on techniques involving PLMs. Therefore, the query should contain either the term "large language model", "pretrained language model", or "seq2seq".

("relation classification" **OR** "relation extraction") **AND**
("large language model" **OR** "pretrained language model" **OR** "seq2seq")

The second search string focuses on geriatric communication and elderly care. To specifically target medical guidelines on this topic, we include "guidelines", "psychological guidelines", and "medical guidelines" in our search query.

("geriatric communication" **OR** "elderly care communication") **AND**
("guidelines" **OR** "psychological direction" **OR** "medical specifications")

4.3. Selection of Language Models

Due to limited time and resources, we tested only a select subset of models. We executed a diverse selection in order to optimize our efforts and generate a more

comprehensive range of insights. In addition, the task to be solved also has a direct impact on the model selection. In our application for the elderly care domain, we are concentrating on RE. There are two options to tackle this problem, each with their own factors to consider when choosing the model:

Ensemble of Methods This approach combines a series of models to extract relationships from text fragments. The pipeline incorporates Named Entity Recognition (NER), RI, and RC. Although there are multiple models that can fulfill these roles, we chose SpaCy [48] for NER due to its extensive use in industry and its versatility. For identifying and classifying relationships, after feature-engineering, our top two choices were XGBoost [49] and BERT [29], respectivelly. XGBoost immediately showed satisfactory results and solidified its position in the RI task after several ablation studies, while BERT was retained for RC due to its previous success in the literature, e.g. on the DialogRE dataset [12].

End-to-end Task RE can be approached as a sequence-to-sequence (seq2seq) task. This method uses a transformer model, like BART [50] or T5 [51], to create an output sequence from an input sequence, much like machine translation or summarization. With this, one can input an input dialogue to these seq2seq models, and they can predict an output sequence of relations. This approach was successfully employed with the REBEL dataset [25], motivating our pursuit of this option. The dataset's adaptability allows for automatic model updates when modified, eliminating the need for multiple adjustments required in more complex pipelines. Within the context of the seq2seq framework, utilizing language model pretraining with prompt guidance is also an option, and it is our preferred approach in this study.

Beyond tasks, our model selection process also weighed several key factors, including:

Model Capacity The number of parameters in a model determines its ability to learn and recognize patterns. While larger models excel in generating text and answering questions, they also require greater computational resources. As parameter size increases, Wei et al. [10] have demonstrated that LLMs can acquire certain skills. However, the compromise is evident in higher hardware and electricity costs. Our infrastructure can reliably handle models with up to 7 billion parameters. While we primarily focus on models of this size, we also include larger models which can be commercially obtained through APIs due to their unique capabilities, such as OpenAI's ChatGPT [52]. This also enables appropriate benchmarks to be conducted.

Model Accessibility Open-source LLMs provide full control and the ability to be hosted on personal servers, ultimately promoting advanced reproducibility. In contrast, closed-source LLMs, typically hosted by companies and only accessible through APIs,

introduce uncertainties in research caused by potential undisclosed pre- and post-processing steps. Nevertheless, despite these obstacles, many of these closed-source models are among the most powerful available.

Training Data and Procedure The choice of training data and methods has a strong impact on the performance of a model. It is important to ensure a causal relationship between statements and to maintain a logical flow of information. Data sources can include factual databases such as Wikipedia or conversational platforms such as Reddit. Training methods can vary, ranging from pre-training techniques to task-specific optimizations such as fine-tuning or RLHF to adapt outputs to human preferences. We used Hugging Face's Open LLM Leaderboard [53] to evaluate the effectiveness of different models on different tasks.

4.4. Selection of Prompting Techniques

Prompt engineering is a field of study that examines how a model's abilities and resulting output are influenced by a variety of prompting methods and best practices. It is founded on the idea that the structure and wording of a prompt can have a significant impact on the quality of the output. Additionally, prompts may be augmented with additional information to direct the model towards producing responses that are factually accurate. This approach is referred to as "grounding" the responses, or "function grounding" [54]. Examples can be included in instructions to facilitate contextual learning. Prompt Engineering has gained recent popularity, prompting the use of grey literature such as blog posts, in addition to academic papers, to obtain a comprehensive understanding of established and emerging approaches. The following presents a concise overview of the most prominent prompting techniques and best practices we consider when selecting the prompts for this thesis.

Zero-shot prompting is a straightforward technique that involves providing a natural language description of a task in the prompt. According to Wei et al., this capability can be improved by instruction tuning [55]. Example:

```
Translate the following sentence into German.
EN: The quick brown fox jumps over the lazy dog.
DE:
```

Few-shot prompting involves incorporating input-output examples into the model to teach it how to solve a task in context. This strategy has yielded significant improvements in performance, sometimes matching fine-tuning approaches. When using a single example, it is also referred to as one-shot prompting [8]. Example:

> Translate the following sentence into German.
>
> EN: Hello, how are you?
> DE: Hallo, wie geht es Ihnen?
>
> EN: The quick brown fox jumps over the lazy dog.
> DE:

Chain-of-thought prompting aims to enhance the reasoning capabilities of the LLM. The model is encouraged to include the individual steps required to generate the result, rather than simply returning an answer. This can be accomplished by providing examples in the prompt that illustrate expected outputs [56]. Another variation of this method is known as zero-shot chain-of-thought prompting. Instead of providing examples, the sentence "let's think step by step" is added to the prompt [57]. Example:

> Q: Bob receives a gift of 10 apples from his grandmother. After sharing half of them with Alice, he consumes two. Unfortunately, a third of the remaining apples begin to rot. How many of Bob's apples are still in good condition?
> A: Let's think step by step.

For our research on RE, we mainly investigated Zero-shot and Few-shot prompting methods. We evaluated the performance of each approach through ablation studies, enabling us to make informed decisions. As part of this thesis, we created a detailed demonstration featuring a customized bot designed for elderly care. Before conducting thorough ablation studies, we conducted a qualitative assessment of the most reliable prompting techniques within the demonstration to guarantee their effectiveness in our intended application. This demonstration also tested prompts to assist the chatbot in providing empathetic replies and using the user's previous knowledge, specifically for data-to-text conversion. Nevertheless, these investigations go beyond the main objective of this thesis, which was limited to RE given its complexity.

4.5. Personal Knowledge Graph Construction

To construct PKGs from plain text, RE is the dominant method in current literature. Various approaches are available, ranging from using custom pipelines with model ensembles to treating it as a seq2seq task. Nonetheless, much of the present research concentrates on sources such as REBEL, SemEval, CONLL-2004 and Re-TACRED datasets [25, 58, 59, 60], which has a distinct data distribution from our intended target. It's vital to emphasize our specific aim of extracting personal relations from dialogues to personalize interactions with senior citizens. Given this context, it is essential to develop

methods specifically tailored to this data distribution. Additionally, the datasets may be even more critical. Further discussion of these considerations will follow in subsequent sections.

4.5.1. Datasets

After conducting a thorough review of available datasets for personal dialogue interactions, we have identified two viable options: DialogRE [12] and DDRel [41]. These datasets meet our criteria, as they focus on dialogues within the context of personal relationships. While they may not perfectly align with our specific needs, they offer significant benefits for our research. First, we could potentially modify the dataset through preprocessing to better align with our target distribution. Second, in order to make progress in the field, it is critical to build upon previous scholarly work, which serves as a standard for assessing the novelty of our research. Yet, it is vital to acknowledge and comprehend the characteristics and limitations of these datasets.

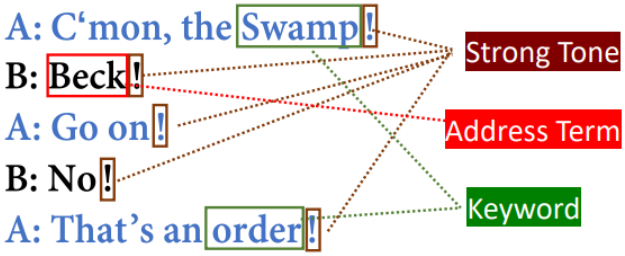


Figure 4.1.: An example from the DDRel dataset [41] illustrates a standard dialogue extracted from various movies. The dataset captures complex and sometimes implicit relationships, as showcased in the exchange between speakers A and B. Notably, our project primarily focuses on relations between elderly patients and other mentioned entities, rather than the relations between both dialogue participants emphasized by the dataset.

After closely examining the DDRel dataset, as showcased in Figure 4.1, its limitations indicate it may not be fitting for our objectives. The dataset solely identifies relations between speakers and overlooks entities mentioned within the dialogue. This misaligns with our goal as interactions between the user and healthcare assistant are secondary. Our principal concentration lies in the bonds between the user and referenced entities. Therefore, although DDRel may not be optimal for our primary requirements, it could potentially serve as a resource for enhancing our data.

On the other hand, DialogRE is the first dataset comprised of human-annotated relations extracted from dialogues. This dataset shows great promise for our targeted objective, since it consists of 1,788 dialogues extracted from 'Friends' scripts, which is a popular source in dialogue-based research. The dataset has 10,168 relational triples and

S1: Hey Pheebs.
S2: Hey!
S1: Any sign of your **brother**?
S2: No, but he's always late.
S1: I thought you only met him once?
S2: Yeah, I did. I think it sounds y'know big sistery, y'know, 'Frank's always late.'
S1: Well relax, he'll be here.

	Argument pair	Trigger	Relation type
R1	(Frank, S2)	brother	per:siblings
R2	(S2, Frank)	brother	per:siblings
R3	(S2, Pheebs)	none	per:alternate_names
R4	(S1, Pheebs)	none	unanswerable

Figure 4.2.: A sample from the DialogRE dataset [12], sourced from dialogues, entities, and relations in the TV show "Friends". The dataset is appropriate for our project due to its thorough categorization of personal relationships in dialogues. While the show's comedic style may make the dialogues intricate and lengthy, it shows promise for our objectives.

36 distinct types of relations annotated. One feature that stands out in this dataset is the annotation of the most precise text span that signifies each relation. Approximately 96.0% of these triples extend across multiple sentences, indicating that the dataset is suitable for studying cross-sentence RE. Moreover, 65.9% of the triples reference arguments from non-consecutive turns, highlighting the importance of multi-turn context in RE based on dialogue. The majority of relations in DialogRE, around 89.9%, describe a speaker's attribute or establish a relationship between two speakers. A sample example is shown in Figure 4.2, indicating promising features for our study.

In Figure A.1, all DialogRE relations are presented in detail, covering a wide range of personal connections. This includes the 'unanswerable' category, designated for instances where no connection exists due to inherent constraints. Specifically, "for which there is 'obviously' no relation between an argument pair in consideration of aspects such as argument type constraints (e.g., relation PER:SCHOOLS_ATTENDED can only exist between a PER name and an ORG name)" [12]. Moreover, if these relations are cross filtered with Kitwood's framework's five areas, the relation count can be reduced to 11 as shown in Figure 4.3. This approach will be utilized throughout our ablation studies in an effort to streamline the model's complexity for improved outcomes, as well as to establish more precise connections for our specific use case.

4.5.2. Data Augmentation

Due to datasets in the literature not fully reflecting our target distribution, the integration of data augmentation strategies is crucial in our pipeline. However, data augmentation in NLP presents distinctive challenges as it can potentially introduce excessive noise, leading to a decrease in model performance. Established strategies

Figure 4.3.: Schema illustrating selected relations from DialogRE, aligned with Kitwood's framework. The schema comprises 11 relations: RESIDENTS_OF_PLACE, PLACE_OF_RESIDENCE, VISITORS_OF_PLACE, VISITED_PLACE, OTHER_FAMILY, ACQUAINTANCE, PARENTS, CHILDREN, SIBLINGS, SPOUSE, and PET. The relation of health_status, in red, has been conceptualized for future research efforts and is not addressed in our current study. The nodes represent possible entity types, which allow these relation pairs, such as Person, Pet, Organization (ORG), **G**eo**P**olitical **E**ntities / Location (GPE) and Health.

encompass Paraphrasing, Noising, and Sampling [61]. It is important to apply Paraphrasing and Noising judiciously as excessive use may result in nonsensical data. The need for moderation in NLP distinguishes it from computer vision, wherein augmentation techniques are frequently used more freely.

Sampling Techniques Sampling entails crafting a customized dataset by carefully selecting from one or more sources. Although merging data from various sources, inclusive of DialogRE and DDRel, appears to be beneficial, it can result in deviations from the target distribution, as outlined in Subsection 4.5.1, and thus, necessitates careful consideration.

Our SlideFilter Technique A crucial contribution of this thesis is the creation of SlideFilter, a novel augmentation technique based on the sampling strategy. SlideFilter processes and segments extensive conversations into smaller and more manageable units, requiring simultaneous filtration of associated relations. An established window, such as three turns, enables the systematic deconstruction of a large dialogue into several smaller segments. This approach assumes that a dialogue's relation can only be identified when the involved entities are explicitly stated within the text. While

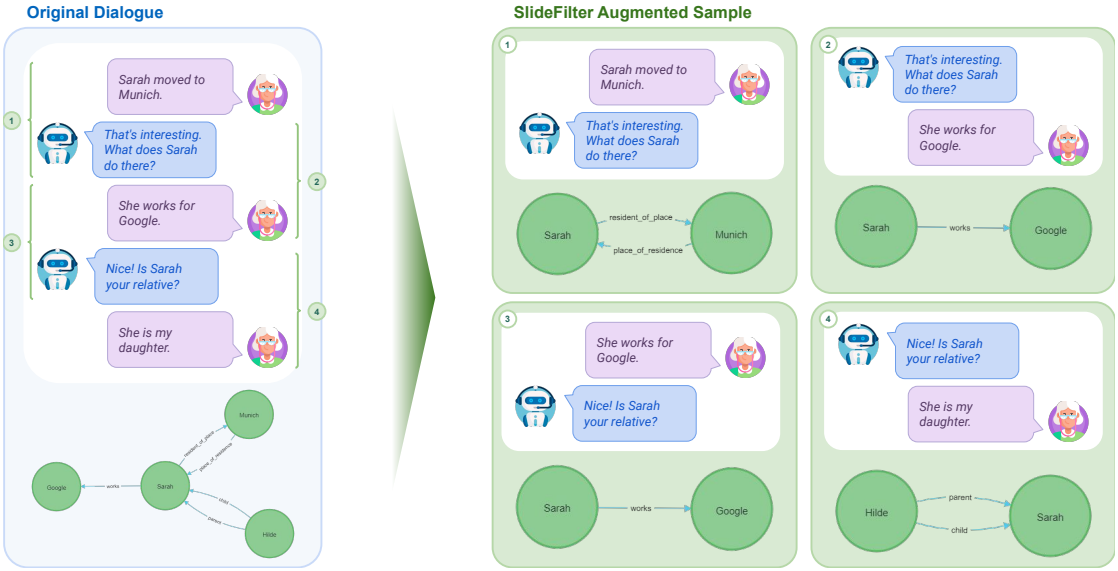


Figure 4.4.: The SlideFilter Method for Enhanced RE: The figure displays the SlideFilter technique implementation for dialogue RE. On the left, an original dialogue is paired with its corresponding relation graph. The four chunks visible in the diagram depict the application of the SlideFilter, which systematically processes the dialogue with a window covering 'n' turns at a time (n=2 here). The right side shows the resulting sub-samples, each representing a section of the dialogue with its relevant sub-graph of relations. The objective of segmenting into smaller, more focused dialogue samples is to reduce noise and improve relation detection accuracy. Nonetheless, there is a significant trade-off as a narrower window may increase accuracy but exclude broader context and relevant cues for relation prediction. Finding the right balance between focus and breadth of context is a crucial challenge when fine-tuning this method for dependable RE.

this assumption may not always hold true, Chapters 5 and 6 will demonstrate its effectiveness. Figure 4.5 demonstrates our SlideFilter method in action and the Figure 4.4 gives an concrete example to explain the methodology.

4.5.3. Ensemble Method

The ensemble method, depicted in Figure 4.6, incorporates multiple models from prior literature. This pipeline has a twofold objective: to appraise the effectiveness of conventional ML techniques in dealing with our tasks and to establish connections between our research and earlier studies. The task of extracting relations from personal dialogues is largely unexplored in existing literature. Nonetheless, our approach uses

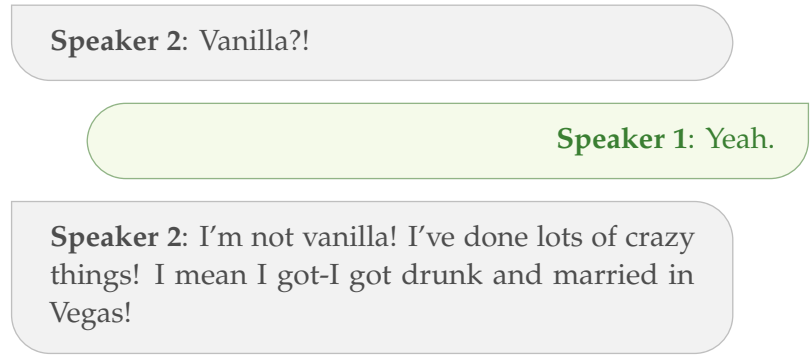


Figure 4.5.: Optimized DialogRE Subdialogue via SlideFilter: This example depicts the use of our SlideFilter approach on a DialogRE subdialog, implementing a three-turn count threshold to extract only those relationships explicitly referenced within the text. Previously, the dialogue comprised 27 turns and 12 relations, but our method condensed it into a brief 3-turn, 2-relation snapshot. This strategic refinement is designed to simplify and streamline the RI process.

DialogRE's RC as a fundamental benchmark. We build upon this foundation by adding more models and components, leading to a comprehensive end-to-end performance evaluation.

NER Given a text fragment that spans multiple turns, our initial step is to extract entities that are referred to in the text. While literature has extensively documented NER, we chose SpaCy's NER [48] because of its strong performance right out of the box. We streamlined our extraction process by omitting cardinals to avoid this explosion of complexity, as shown in the figure 1.2.

RI After identifying the entities in the conversation, we establish pairwise combinations among them. While some entity pairs may exhibit relationships, not all do, underscoring the importance of this phase in our analysis. Our exploration encompassed multiple models, including XGBoost and BERT, as well as feature engineering techniques, such as computing the shortest word distance between entities. We also considered an implicit identification approach where the absence of a relation is classified under a "no_relation" category during the RC phase. To determine the optimal model for our objectives, we conducted ablation studies in this stage.

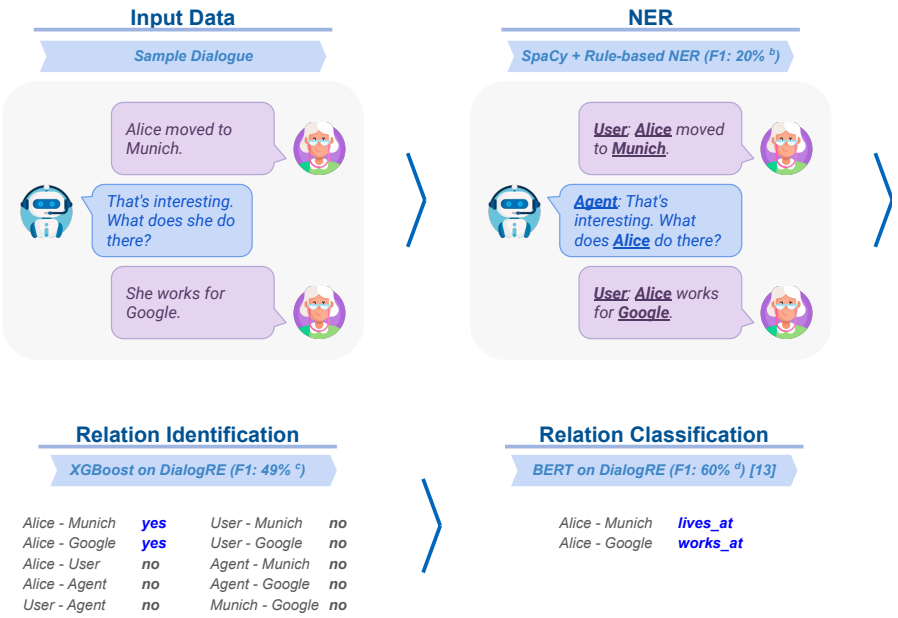


Figure 4.6.: The figure above illustrates our ensemble approach for extracting relations from dialogue data. The process begins with the input of a dialogue sample, followed by NER using SpaCy NER. XGBoost then assesses the connection between all potential pairs of entities. Lastly, BERT performs final classification for the confirmed relations.

RC This stage follows the procedures outlined in the DialogRE paper. For confirmed relations, we tested BERT and LLaMA to classify the connection between two entities based on the dialogue context. For operational efficiency, our ensemble primarily utilizes BERT. The integration of LLaMA may be unnecessary since its complex architecture is proficient enough to handle the entire task. This decision is supported by ablation studies, which will be further explained in subsequent chapters.

4.5.4. End-to-End Method

An alternative to conventional RE is the end-to-end approach, where a single model oversees the entire process. This strategy handles the input conversation and produces relation triplets directly. As a sequence-to-sequence (seq2seq) task, numerous models have been utilized for this objective, as evidenced by Huguet-Cabot and Navigli's use of BART [25]. The benefits of this technique are a simplified process and increased adaptability. Unlike approaches utilizing multiple models and processing stages, an end-to-end system streamlines adjustments, requiring modifications solely to the data or the model, instead of multiple components.

For this thesis, we utilize LLaMA [33], harnessing the increasing popularity and

effectiveness of LLMs across numerous tasks. Their remarkable performance in zero-shot and few-shot learning scenarios is notable as they often produce promising results without any training data [55, 8]. This capability is analyzed in the results section (refer to Chapter 5). Our decision to further investigate LLMs is driven by their capacity to utilize extensive prior knowledge for complex tasks that may pose a challenge to simpler models. As a result, prompt engineering becomes a crucial aspect of this methodology. Our ablation studies will concentrate on comparing Ensemble of Methods to LLMs, examining their respective advantages and limitations.

4.5.5. Prompt Design

This section outlines the components that each prompt should include and the methodology used to craft prompt templates for the end-to-end process using LLMs, as described in Chapter 5 for RE tasks.

An effective prompt must provide precise instructions and include all the essential information required for task completion. This comprises two critical components: 1) the input dialogue; and 2) a list of possible relations. Equipped with this information, the model is ready to tackle the task in a zero-shot manner. Previous research, including that of Lewis et al. [8], suggests that one-shot examples can improve model performance without fine-tuning. However, their effectiveness diminishes with fine-tuning. As a result, our benchmarks largely rely on zero-shot techniques.

The prompt design process is guided by a predetermined gold standard of expected outcomes for one sample set. The process entails crafting a succinct prompt that merges

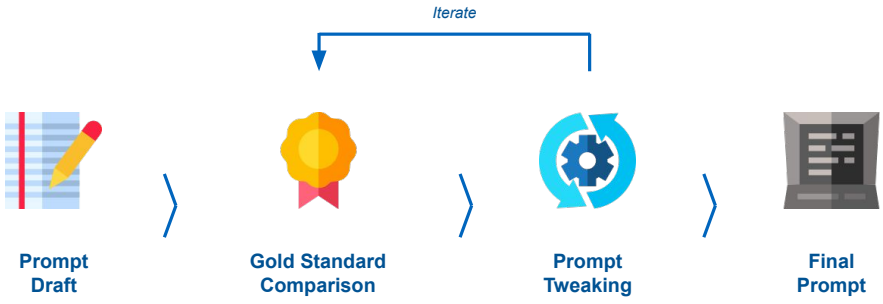


Figure 4.7.: Diagram of Prompt Design Workflow: The process begins with benchmark comparison using a selected example of input dialogue and desired prediction, followed by iterative refinements. In successive cycles, the prompt is adjusted with the LLM to match the required gold standard or format precisely. This ensures accurate label generation without any distortions or unwarranted explanations. This cycle persists until the optimal prompt is established as the standard, resulting in the highest fidelity output.

instructions, dialogue input, and relation ontology. This template is subsequently refined until the model's predictions are aligned with the gold standard or comply with the anticipated format, illustrated in the Figure 4.7.

4.6. Evaluation Metrics

Within this section, we explore the automatic evaluation metrics that form the basis of our RE model. Since our primary objective is RC with DialogRE [12], we utilize classification as our approach. We begin by explaining the metrics used for the basic scenario of RC. Later, we extend our evaluation framework to address the more general task of RE. Although classification principles are involved in both circumstances, the latter consists of a multi-label context, since a given dialogue may contain a variable amount of relation triplets. Moreover, a relation is considered accurately identified only when the entire triplet of 'subject', 'relation', and 'object' matches the ground truth. Therefore, our evaluation consolidates the metrics to reflect performance based on the 'relation' component alone. Please refer to the Figure 4.8 for clarification.

In the given instance, the prediction would be considered inaccurate because it fails to correctly identify the 'subject' aspect, even though it accurately identifies the 'object' and 'relation'. Consequently, despite partial correctness, the evaluation metric would categorize the entire prediction as erroneous. This mistake would be classified under the label 'spouse' for evaluation purposes. This follows the boundaries evaluation in RE proposed by Taillé et al. [62], since the relation types as ignored here.

4.6.1. Classification Metrics

The metrics presented in this paper are based on the categorizations outlined by A. Géron [63]. The desired outcome, namely the gold standard (or ground-truth), allows

Max: Hey, love! Did you pick up the kids already?

Leni: Not yet, honey. I have to call Sophia first...

Ground Truth: {'object': 'Max', 'relation': 'spouse', 'subject': 'Leni'}
Prediction: {'object': 'Max', 'relation': 'spouse', 'subject': 'Sophia'}

Figure 4.8.: An example dialogue shows a ground truth and a prediction. The evaluated label is 'spouse,' and the prediction is marked as a false positive due to the wrong identification of the subject. For metric calculation, we only consider the 'relation' label. However, to be considered as a true positive, the entire relation triplet must match with the ground truth.

for the classification of results into four distinct categories. This is valid for RC, which is the case on DialogRE paper [12].

- **True Positive (TP):** Entities that appear in both the gold standard and the obtained results.
- **True Negative (TN):** Entities that are absent in both the gold standard and the results.
- **False Positive (FP):** Entities present in the actual results but not in the gold standard.
- **False Negative (FN):** Entities that are missing in the actual results yet are part of the gold standard.

These categories enable the computation of the succeeding metrics.

- **Precision:** This formula calculates the percentage of accurate predictions out of all supplied results.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.1}$$

  Precision is deemed perfect if the outcome solely consists of pertinent elements, disregarding any absent entities from the benchmark.

- **Recall:** This measures the proportion of gold standard entities that are included in the actual results, as illustrated by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

  Recall is deemed perfect if all gold standard entities are incorporated in the result.

- **F1-Score:** The F1-score integrates precision and recall, addressing their distinctive constraints, into a unified metric. Its maximum value of one corresponds to optimal precision and recall performance.

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{\text{FN}+\text{FP}}{2}} \tag{4.3}$$

- **Accuracy:** Applied in binary classification, this metric evaluates the proportion of accurate predictions out of all predictions, expressed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.4}$$

4.6.2. Adaptation for Multi-Label Classification

For RE, the output varies. A dialogue may contain no relation, one relation, or multiple ones. To address this variability, we approach the problem through a multi-label classification based on the relation key of each relation triplet.

Evaluation metrics in multi-label classification scenarios need adaptation because instances may belong to multiple classes simultaneously. The standard definitions of true positive (TP), false positive (FP), and false negative (FN) remain unchanged. However, true negative (TN) loses its significance as the absence of one label may not indicate a true negative owing to the existence of other accurate labels.

Zhang et al. [64], identified two categories of classification performance measures: example-based and label-based. The former measures the accuracy of a learning system per instance, taking into account all of its labels, while the latter evaluates the accuracy of each label across the dataset. For our work in RE, which requires distinct labeling for each relation type, we utilize example-based metrics due to their effectiveness in capturing contextual nuances and the full range of relations in each text segment. This is essential for our analysis.

To account for this, we introduce the following metrics specific to multi-label classification (denoted as *exam* for example-based):

- **Multi-label Precision**: The precision for each instance is determined by the proportion of correctly predicted labels to the total number of predicted labels. Specifically, for the i^{th} case, this ratio is represented as $\frac{|Y_i \cap h(x_i)|}{|h(x_i)|}$, where Y_i refers to the set of actual labels, $h(x_i)$ indicates the set of labels predicted by the classifier, and p represents the total number of instances in the test set. To obtain the multi-label precision, this ratio is averaged across all cases, as depicted in the following equation:

$$\text{Precision}_{\text{exam}}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(x_i)|}{|h(x_i)|} \tag{4.5}$$

- **Multi-label Recall**: Recall measures the proportion of correctly predicted actual labels. For each instance, recall is computed as the intersection of predicted and true labels divided by the number of true labels. Y_i represents the true labels and $h(x_i)$ represents the predicted labels by the model. The following formula presents the average recall over all instances, yielding the multi-label recall.

$$\text{Recall}_{\text{exam}}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(x_i)|}{|Y_i|} \tag{4.6}$$

- **Multi-label F1-Score**: The F1-score is the harmonic mean of precision and recall, achieving balance between both metrics. It is especially valuable when precision

and recall hold equal importance. The F1-score, as defined for the context of multi-label classification, is as follows:

$$F^1_{\text{exam}}(h) = 2 \cdot \frac{\text{Precision}_{\text{exam}}(h) \cdot \text{Recall}_{\text{exam}}(h)}{\text{Precision}_{\text{exam}}(h) + \text{Recall}_{\text{exam}}(h)} \tag{4.7}$$

It's important to note that in multi-label classification, the objective isn't solely to maximize individual label predictions, but to enhance the prediction performance across all labels for each instance. As a result, these adapted metrics offer a more comprehensive view of a multi-label classifier's performance.

Utilizing Confusion Matrices in Multi-Label Classification A confusion matrix is essential for evaluating classifiers in multi-class tasks because it represents classification overlaps clearly. However, in multi-label classification where each instance may have multiple labels, a standard confusion matrix is not defined due to the variability and complexity of label count. To address this challenge, we employ the multi-label confusion matrix (MLCM) method detailed by M. Heydarian et al. [65]. This technique provides a customized approach for visualizing classifier performance in multi-label contexts. Our study simplifies matters by concentrating only on the relation key, neglecting the subject and object keys, which inevitably results in an overestimation in the confusion matrix. When used alongside metrics such as F1-score, precision, and recall, this method generates a speedy and informative overview of the model's performance, pinpointing potential failure modes and effectively guiding further refinements.

5. Results

5.1. RQ1: Concepts and Entities for Data Model

In this section, we will examine the basic concepts and entities that must be integrated into our data model to realize effective and personalized communication in geriatric care. The primary results of this study will be documented here.

Achieving Personalized Communication in Elderly Care Driven by the goal of effective customized communication in elderly care, the initial research question of this thesis explores the fundamental concepts and entities necessary for personalization and engagement. This inquiry is supported by two primary challenges: identifying an appropriate dataset that captures the intricacies of personal relationships and establishing communication guidelines that align with the distinct needs of the elderly.

Identification of Data and Communication Guidelines The first phase of the study entailed a thorough assessment of the available literature and datasets with the objective of identifying appropriate resources to construct a model of customized communication. The DialogRE dataset [12], which contains personal relationships and entities discussed in a conversation format, exhibited the potential to serve as an excellent point of departure. This dataset, comprising dialogues from the TV series "Friends", partially resembles the target distribution for our research by capturing the essence of personal interactions within a narrative context, as depicted in the Figure 4.2.

Application of Kitwood's Framework Tom Kitwood's person-centered framework provides a high-level lens for this research [43]. It emphasizes the fundamental need for individuals, especially the elderly, to express their identity, relationships, and competencies through dialogue. The framework outlines five psychological needs: comfort, attachment, identity, occupation, and inclusion, depicted in the Figure 3.1. These requirements served to streamline and prioritize the relationships and entities within the DialogRE data set, aligning them to meet the particular needs of geriatric care.

Mapping and Filtering of Data Model The mapping process comprehensively reviewed all relations in the DialogRE dataset to identify the five areas outlined by Kitwood. By doing so, relevant relations were filtered and selected to facilitate the

development of a comprehensive data model. The resulting data model establishes a fundamental framework for upcoming experiments and applications in tailored elderly care communication. It leverages the synergies between Kitwood's framework and the practical implementation of the DialogRE dataset. After filtering, our data model contained 11 relationships, a significant reduction from the 36 relationships in DialogRE. This allowed us to streamline our approach and implement a new strategy to improve results in our experiments for PKG construction. The filtered data model can be seen in Figure 4.3. Details on these experiments are provided in the following sections.

5.2. RQ2: Information Extraction Techniques

This section presents the key empirical findings of our thesis, specifically addressing the effectiveness of extracting relational data from textual dialogues. We conducted a comprehensive suite of 37 experiments, which covered three distinct tasks: RC, RI, and RE. To accomplish this, we utilized various preprocessing techniques, including standard filtering methods and our proprietary SlideFilter, which are explained in detail in Section 4.5.2.

5.2.1. Methodological Ablation Studies

Experiment Overview Tables The tests are organized by task and presented in Tables 5.1, 5.2, and 5.3. We have curated a subset of the most significant experiments for an in-depth analysis and provided their key performance metrics due to the extensive nature of the tests. Each of these experiments was chosen based on a hypothesis. The findings of one experiment led to new hypotheses, which were validated in the next batch of experiments, as outlined in Section 4.1.

Waterfall Chart Analysis of Ablation Studies To assess the critical phases of our research, we employed the macro F1-Score, which assesses precision and recall, taking into account class imbalance. This metric is exhibited in the cascade diagrams of Figures 5.1, 5.2, and 5.3. These diagrams are particularly informative as they present our experiments in the form of an ablation study, with each step altering only one parameter. Additionally, we have prepared three separate charts to align with the scope of our research. These charts correspond to the main tasks of RC, RI, and RE, enabling a focused comparison of model performances across these tasks.

Quantitative Experiment Breakdown For a detailed breakdown of the metrics, refer to Tables 5.4 and 5.5. These tables offer a comprehensive view that goes beyond the Macro F1-Score. Precision and Recall macro averages are provided for each label type: 'No Relation,' 'Null Relation,' and 'All Other Relations.' This selection of experiments has undergone a thorough evaluation. In cases where a label is missing, as in experiment

Table 5.1.: Comprehensive Experiments on RC: An overview of our sequential experimentation starting from DialogRE replication, advanced preprocessing, and model evaluation with BERT, LLaMA, and BART. Additionally, comparative insights from ChatGPT are included. The emphasis is on the promising performance of BERT and LLaMA.

Id	Detailed Study	Description	Model
e00		Reproduce DialogRE paper pipeline	bert-tiny
e01b		BERT Baseline Reproduction of DialogRE, w/o Per Label Metrics	bert-base
e01	✓	BERT Baseline Reproduction of DialogRE, w/ Per Label Metrics	bert-base
e03	✓	BERT 'No_Relation' Comparison	bert-base
e05	✓	BERT Focus-Relations Assessment	bert-base
e07	✓	BERT Focus-Relations 'No_Relation' Comparison	bert-base
e04b	✓	GPT3.5 'No_Relation' Comparison	gpt-3.5-turbo-0613
e02	✓	LLaMA Comparison	llama-7b-hf
e04	✓	LLaMA 'No_Relation' Comparison	llama-7b-hf
e06		LLaMA Focus-Relations Comparison	llama-7b-hf
e06b	✓	LLaMA Focus-Relations 'No_Relation' Comparison	llama-7b-hf
e07b		BART 'No_Relation' Comparison	bart-large

Table 5.2.: Comprehensive Experiments on RI: This table presents the sequence of experiments carried out to assess explicit RI in our pipeline. The process involves a BERT-based pipeline for binary RI of a pre-processed dataset, different preprocessing techniques for a three-label classification scheme, and traditional XGBoost methods. We also conducted a comparative assessment with the LLaMA model. Particularly encouraging results have been observed using BERT and XGBoost approaches.

Id	Detailed Study	Description	Model
e08a	✓	Fine-tune BERT	bert-base
e10a		Assess Three Label Signal with BERT (no, with, and inverse relation)	bert-base
e10b		Assess Three Label Signal with BERT Undersampled	bert-base
e10c		Assess Three Label Signal with BERT Oversampled	bert-base
e10d		Assess Two Label Signal with BERT Oversampled	bert-base
e09a	✓	Train XGBoost with Engineered Features	xgboost
e09b	✓	Train XGBoost Undersampled (50/50 Split)	xgboost
e10e		Fine-tune LLaMA	llama-7B-hf

Table 5.3.: Comprehensive Experiments on RE: This table outlines our experience in developing a reliable RE pipeline. We integrated various preprocessing techniques such as adjusting null relation ratios, applying speaker and turn count filters, using diverse data augmentation methods, and implementing our unique SlideFilter approach. Additionally, we utilized several architectures including BERT Ensemble, LLaMA, REBEL, and BART. Our evaluation also includes a comparison with ChatGPT. Special attention is focused on the most efficient settings that incorporate BERT Ensemble and LLaMA with the SlideFilter enhancement.

Id	Detailed Study	Description	Model
e11	✓	BERT Ensemble w/ Explicit RI	ensemble-11cls
e12	✓	BERT Ensemble w/ Implict RI	ensemble-12cls-implitRelIdent
e13	✓	LLaMA Comparison	llama-7b-hf
e14	✓	ChatGPT3.5 Comparison	gpt-3.5-turbo-0613
e27		REBEL Comparison	rebel-large
e17		BART Comparison	bart-base
e21		BART Comparison w/o Null Relations	bart-base
e22		BART Comparison w/o Null Relations	bart-large
e24		BART Comparison with Null Relation Tweak	bart-large
e19		BART Comparison with DDRel Augmentation w/o Data Shuffle	bart-large
e20		BART Comparison with DDRel Augmentation	bart-large
e25		LLaMA Comparison with Insufficient Null Relation Tweak	llama-7B-hf
e26		LLaMA Comparison with w/ 2 Speaker Filter	llama-7B-hf
e28		LLaMA Comparison with DDRel Augmentation	llama-7B-hf
e29		LLaMA Comparison w/o Null Relations	llama-7B-hf
e15	✓	LLaMA with SlideFilter	llama-7b-hf
e16	✓	LLaMA w/ SlideFilter & Null Relation Tweak	llama-7b-hf
e23		BERT Ensemble w/ SlideFilter & Null Relation Tweak	ensemble-11cls

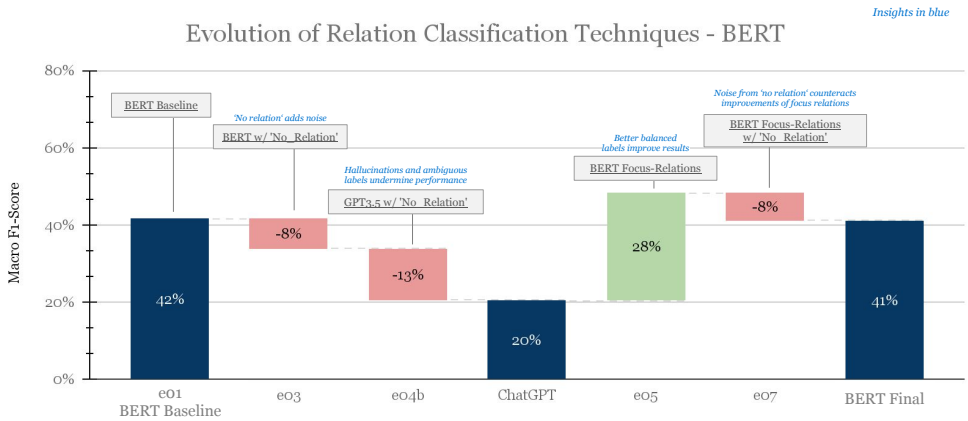


Figure 5.1.: BERT Model Performance in RC: This study presents the macro F1-scores for BERT in different settings on the DialogRE dataset, indicating improvement from the baseline to focused modifications and a comparative analysis with ChatGPT. Blue annotations highlight the impact of label adjustments. The refined BERT model performs similarly to the baseline, while also modeling a distribution that aligns more closely with our target. This result validates our iterative methodology. *Note: Distinct test sets were created for each model iteration to enable a thorough evaluation.*

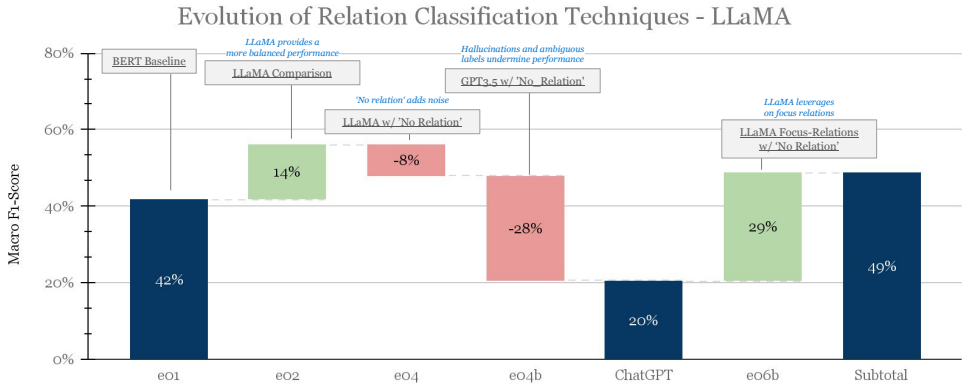


Figure 5.2.: LLaMA Model Performance in RC: The bar chart compares the macro F1-scores of LLaMA, BERT, and GPT-3.5 with various configurations on the DialogRE dataset. It illustrates LLaMA's improved performance by removing 'No Relation' and its balanced approach compared to GPT-3.5's reduced capability. Notably, the final iteration of LLaMA shows a significant advancement, underscoring its effectiveness for RE duties. *Note: Distinct test sets were created for each model iteration to enable a thorough evaluation.*

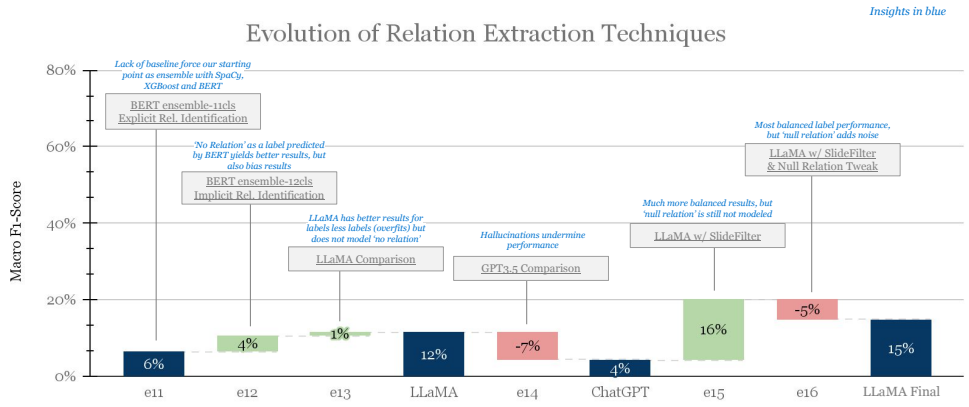


Figure 5.3.: Ensemble and LLaMA Performance in RE: This chart depicts the macro F1-scores of BERT ensemble and LLaMA models on the DialogRE dataset. LLaMA improves upon starting points with BERT ensemble benchmarks despite the issue of overfitting. It is worth noting that GPT-3.5 has problems with hallucination. LLaMA's and SlideFilter's integration demonstrates optimized performance, culminating in an ultimate model that emphasizes the synergy of their combined capabilities for RE tasks. *Note: Distinct test sets were created for each model iteration to enable a thorough evaluation.*

Table 5.4.: Experiment Results for RC

Id	Model	Dataset	Macro Average			No Relation			Others (Avg.)		
			P	R	F1	P	R	F1	P	R	F1
e01	bert-base	dialog-re-llama-37cls (baseline)	49%	43%	42%				49%	43%	42%
e03	bert-base	dialog-re-37cls-with-no-relation-undersampled	36%	35%	34%	47%	56%	51%	36%	34%	33%
e05	bert-base	dialog-re-11cls	47%	55%	49%				47%	55%	49%
e07	bert-base	dialog-re-12cls-with-no-relation-undersampled	43%	43%	41%	33%	85%	47%	44%	40%	41%
e02	llama-7B-hf	dialog-re-llama-37cls-clsTskOnl-instrB-shfflDt	64%	56%	56%				64%	56%	56%
e04	llama-7B-hf	dialog-re-37cls-with-no-relation-undersampled-llama-clsTskOnl	68%	49%	53%	48%	76%	59%	68%	48%	53%
e06b	llama-7B-hf	dialog-re-12cls-with-no-relation-undersampled-llama-clsTskOnl	55%	50%	49%	65%	25%	37%	64%	61%	60%
e04b	gpt-3.5-turbo	dialog-re-37cls-with-no-relation-undersampled-llama-clsTskOnl	25%	28%	22%	36%	18%	24%	25%	28%	22%

e01, the related metric fields are purposefully left blank. It's essential to differentiate between 'No Relation,' which implies the absence of a connection between two entities, and 'Null Relation,' which indicates that a dialogue doesn't have any relations at all. This subtle difference is crucial for accurately interpreting the information presented in the tables.

5.2.2. Strategic Enhancements and Visualized Outcomes

Comparative RI Results In our investigation of model ensembles for determining relations, we first considered that RE's complexity, especially with the inclusion of a 'no' relation, was excessively high, as demonstrated by the noise outlined in Table 5.1. However, subsequent discoveries revealed that implicit RI surpassed explicit techniques

Table 5.5.: Experiment Results for RE

Id	Model	Dataset	Macro Average			Null Relation			Others (Avg.)		
			P	R	F1	P	R	F1	P	R	F1
e11	ensemble-11cls	dialog-re-12cls-with-no-relation-undersampled-llama	9%	5%	6%	12%	23%	16%	13%	10%	7%
e12	ensemble-12cls-implicitRelIdent	dialog-re-12cls-with-no-relation-undersampled-llama	9%	26%	11%	63%	45%	52%	3%	32%	5%
e13	llama-7B-hf	dialog-re-12cls-with-no-relation-undersampled-llama	12%	13%	12%	0%	0%	0%	25%	20%	20%
e14	gpt-3.5-turbo	dialog-re-12cls-with-no-relation-undersampled-llama	3%	2%	3%	5%	60%	8%	6%	5%	4%
e15	llama-7B-hf	dialog-re-llama-11cls-rebalPairs-rwrtKeys-instrC-mxTrnCp3-skpTps	20%	21%	20%	0%	0%	0%	26%	37%	27%
e16	llama-7B-hf	dialog-re-11cls-llama-rebalPairs6x-rwrtKeys-instrC-mxTrnCp3-shfflDt-skpTps	14%	15%	14%	15%	80%	25%	23%	16%	16%

in performance. Despite this, it is advantageous to present our approach to RI here. Figure 5.4 shows that XGBoost outperforms BERT in this context, offering two main benefits: reduced complexity and improved performance. This improvement is mainly due to the inclusion of the minimum word distance feature, which, although simple, has a significant impact on the model's effectiveness. They reflect the experiments e08 and e09.

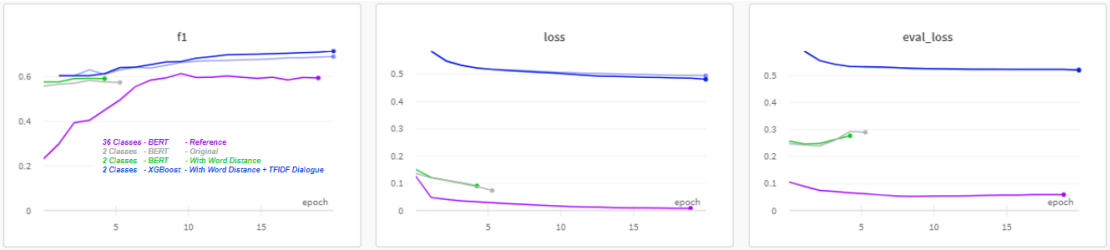


Figure 5.4.: Comparative Analysis of BERT and XGBoost on RI: The presented graphs depict epoch-by-epoch performance evaluations of models trained to identify relationships within dialogues (experiments e08 and e09). The F1 score graph on the left serves as a reference for the BERT model's performance with 36 classes, compared to its binary classification (2 classes) performance with and without word distance features. The XGBoost models show significant improvement in their F1 score when enhanced with word distance and TFIDF dialogue features, evenly classifying related and unrelated pairs. Additionally, the loss and evaluation loss graphs (center and right) showcase the better training stability and efficiency of XGBoost over BERT, featuring a notable reduction in loss metrics. The selection of XGBoost was confirmed not only for its reduced complexity and computational expenses but also for its better per-label metrics. Thus, it is the favored model for this binary classification job.

Efficiency Analysis of SlideFilter Augmentation Figure 5.5 presents the efficacy of the SlideFilter Augmentation technique. It shows the impact of optimizing the window size hyperparameter (mxTrnCnt). The Token Count histogram illustrates this optimization, displaying a significant shift in the distribution towards the lower range after applying SlideFilter. This reduces variance and trims extended dialogues into

more manageable segments. By comparison, the utilization of SlideFilter produces a more focused distribution of dialogue lengths. This suggests reduced variability among the conversations. To validate this assumption, we examined the model's performance using a bar chart. It became evident that setting a maximum turn count of 3 during hyperparameter tuning resulted in the most substantial boost to the F1 Score. This outcome highlights the crucial role of SlideFilter in improving the accuracy of the model.

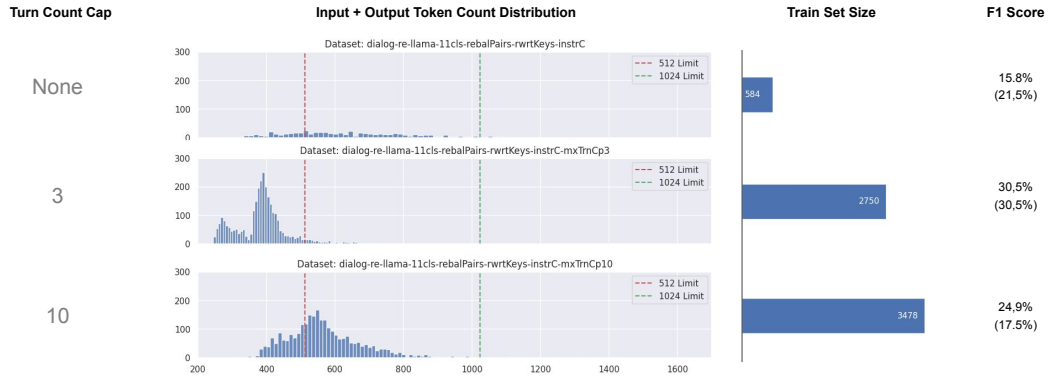


Figure 5.5.: SlideFilter Augmentation's Impact on Model Metrics and Distribution: The histograms display the distribution of input and output token counts across datasets without a turn count limit, with a limit of 3, and a limit of 10. By incorporating SlideFilter Augmentation, which segments lengthy dialogues into shorter sequences, we observed a broader distribution within the token count limits (512, 1024). The training set size was significantly increased through this method, as demonstrated in the central bar graph, which corresponds to the F1 score enhancements depicted on the right. By extending the maximum token length of the model, more detailed dialogue features were revealed, resulting in an improved model performance. Note: Metrics for the mxTrnCp3 dataset are presented, along with their initial distribution in brackets.

Visualized Confusion Matrices Our experimental performance can be better understood through the use of confusion matrices, as shown in Figures 5.6 and 5.7 for RC; and 5.8 for RE. We have utilized these matrices to gain further insight into the results of our experiments. These figures plot the predicted and actual labels, providing a clear overview for single label classification and multi-label extraction scenarios. It is worth noting that in multi-label instances, where traditional confusion matrices are not applicable, we utilize the methodology proposed by Heydarian et al. [65], as detailed in Section 4.6. The differences in the quantity and layout of labels are a result of the distinct scopes and training datasets of each model, which influence the labels they are

trained and assessed upon.

5.2.3. Evolution of Prompt Design

In this subsection, we present our findings on prompt design for RC and RE.

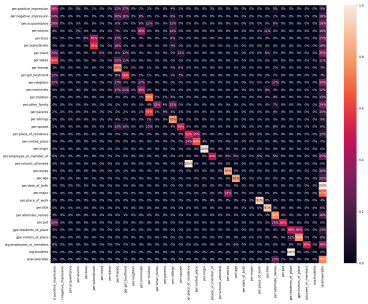
Foundational One-Shot Template Our exploration of prompt engineering commenced with the fundamental one-shot template shown in Figure 5.9. This template played a crucial role in establishing the framework for extracting relationships by utilizing dynamic placeholders to accurately represent entities and their associations within dialogues using a instruction-based approach with OpenAI's ChatGPT. The utilization of this framework denotes our initial endeavor in exploiting the potential of conversational AI when interacting with elderly patients, which establishes a foundation for more focused advancements.

Optimized Extraction and Comparative Analysis The initial experiments provided insights that led to the development of a more precise and efficient entity-RE prompt. The new prompt, depicted in Figure 5.11, was created to fulfill the specific needs of LLaMA, which is a less complex model than ChatGPT. Concurrently, Figure 5.10 emerged as a critical tool for benchmarking our model against the LLaMA framework, following the standards outlined in the DialogRE paper. This comparative analysis was essential in validating our model's performance and guiding subsequent refinements, ensuring that our approach remained congruent with the latest advancements in the field. Leveraging fine-tuning in our methodology eradicates the need for one-shot examples while preserving model performance, as supported by Wei et al. [55].

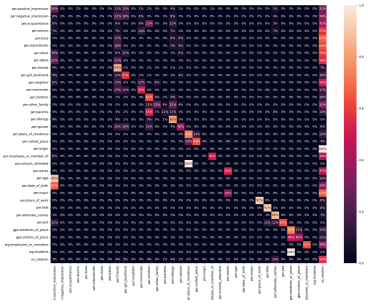
5.3. RQ3: Preliminary Exploration of Knowledge Integration

This section explores the concept of knowledge integration, which was only briefly mentioned in this thesis due to its primary focus on RE tasks. The insights presented derive from our prototype system that combines the capabilities of ChatGPT and a Neo4j database, as referenced in Section 1.1. We present a qualitative summary of our observations below. Our prototype employed ChatGPT to identify relationships which were structured into a Neo4j database graph. This graph, representing the bot's memory, enabled the generation of customized follow-up questions.

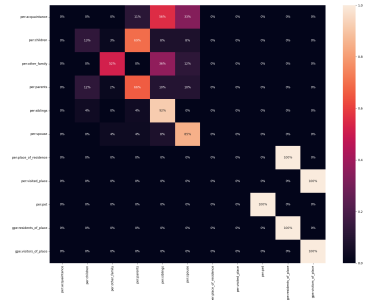
Memory Recall Implementation Follow-up message generation relied on a simple graph search algorithm. The relationships analyzed by ChatGPT aided entity selection for queries with precision. To search for subgraphs showing connections between an entity and the user, we randomly selected a relation and incorporated it. To avoid cyclic paths, we restricted the search to subgraphs with a limited number of entities. The



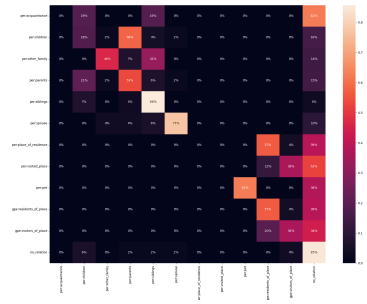
(a) e01 BERT Baseline



(b) e03 BERT w/ 'No_Relation'



(c) e05 BERT Focus-Rels



(d) e07 BERT Focus-Rels w/ 'No_Relation'

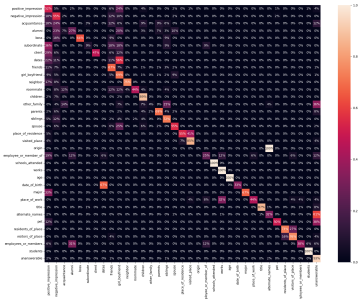
Figure 5.6.: **Comparative Analysis of BERT's RC across Various Pre-processing Techniques:** While the matrices demonstrate BERT's potential, they also indicate the need for a more nuanced architecture to improve classification accuracy.

(a) Experiment e01 presents baseline BERT model with a robust RC with a distinctive diagonal pattern, indicating good performance.

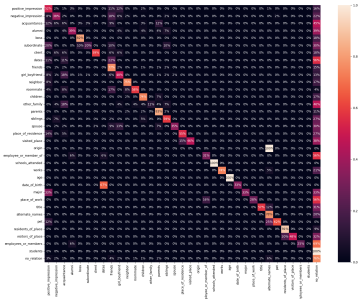
(b) Experiment e03 showcases the inclusion of the 'No Relation' label and causes noteworthy classification diffusion, particularly in the rightmost column of the matrix, indicating increased noise.

(c) Experiment e05 narrows down to focus relations, showing a strong diagonal, but also highlighting the model's limitations with certain classes, as seen by the near-black diagonal entries, indicating labels with almost zero performance.

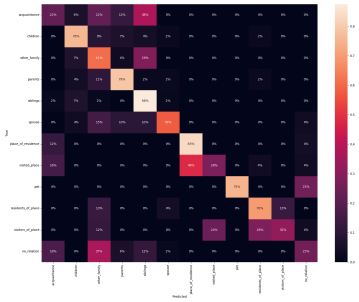
(d) Experiment e07, which adds 'No Relation' to the focused relations, further exacerbates these classification challenges.



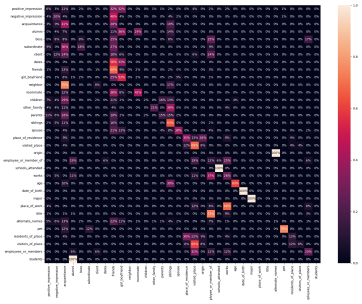
(a) e02 LLaMA



(b) e04 LLaMA w/ 'No_Relation'



(c) e06b LLaMA Focus-Rels w/ 'No_Rel.'



(d) e04b ChatGPT3.5 w/ 'No_Relation'

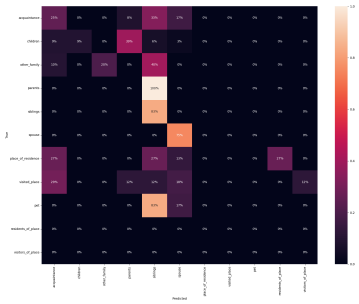
Figure 5.7.: **Comparative Analysis of LLMs' RC across Various Pre-processing Techniques:** This demonstrates how LLMs outperform BERT in general, and how fine-tuning greatly enhances performance.

(a) Experiment e02 uses LLaMA with the original DialogRE dataset, demonstrating a strong diagonal pattern indicative of high accuracy in RC and superior performance to BERT, as referenced in Figure 5.6(a).

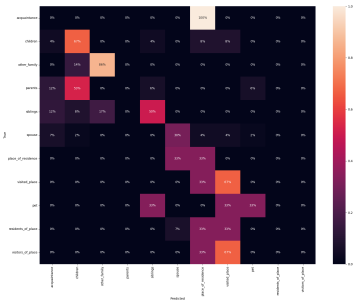
(b) Experiment e04 introduces the 'No Relation' label with LLaMA, maintaining a clear diagonal and effectively managing noise, showing improved outcomes over BERT, seen in Figure 5.6(b).

(c) Experiment e06b highlights LLaMA's focused relationships, achieving extensive classification without any zero-performance instances, thus outperforming the focused relationships of BERT in Figure 5.6(d).

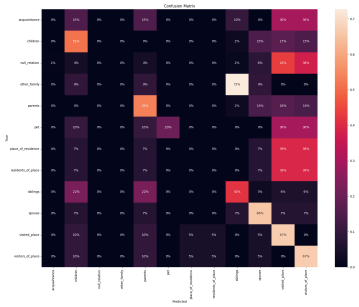
(d) In contrast, Experiment e04b features ChatGPT-3.5-Turbo with 'No Relation', which displays a reduced diagonal intensity and overall lower performance compared to LLaMA, underscoring the importance of fine-tuning.



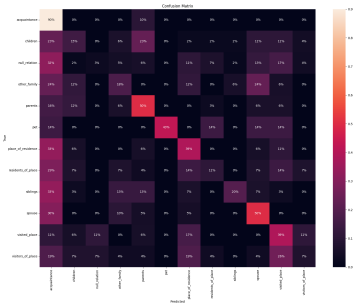
(a) e11 BERT ensemble-11cls



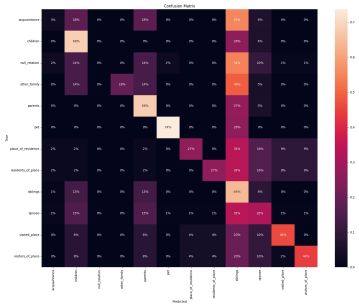
(b) e12 BERT ensemble-12cls-implIdent



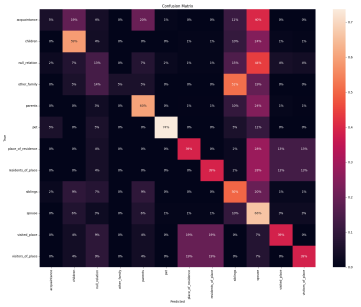
(c) e13 LLaMA



(d) e14 ChatGPT-3.5



(e) e15 LLaMA SlideFilter



(f) e16 LLaMA SlideFilter NullRel Tweak

Figure 5.8.: **Confusion Matrices for RE across Experimented Architectures and Preprocessing Techniques**: This matrix array displays the results of various models and techniques, where LLaMA, using the SlideFilter and null relation tweak (e16), is the most promising architecture. The robust diagonal and fewer misclassifications demonstrate its superiority. However, some labels, such as "acquaintance," show no performance, indicating a need for further investigation. Note that the presented values are normalized for comparison. However, it is crucial to note that these analyses utilize Heydarian's Multi-Label Confusion Matrix (MLCM) methodology [65]. Additionally, reducing each triple to its relation label may lead to an overestimation similar to a label co-occurrence matrix, potentially affecting the clarity of model performance distinctions.

Extract personal relevant entities, and their relations. Return only the jsonl format list.

Ontology:
- relations: {{"acquaintance", "children", "other_family", "parents", "siblings", "spouse", "place_of_residence", "visited_place", "pet", "residents_of_place", "visitors_of_place"}}
- types: {{"ORG", "GPE", "PERSON", "DATE", "EVENT", "ANIMAL"}}

Input:
(
"User: My daughter, Emma, recently moved to London.",
"Agent: That's exciting! Does she like it there?",
"User: Yes, she loves it! She even adopted a cat named Whiskers.",
)

Output:

```
[
    {{"x": "User", "x_type": "PERSON", "y": "Emma", "y_type": "PERSON", "r": "children"}},
    {{"x": "Emma", "x_type": "PERSON", "y": "London", "y_type": "GPE", "r": "place_of_residence"}},
    {{"x": "London", "x_type": "GPE", "y": "Emma", "y_type": "PERSON", "r": "residents_of_place"}},
    {{"x": "Emma", "x_type": "PERSON", "y": "Whiskers", "y_type": "ANIMAL", "r": "pet"}},
    {{"x": "Whiskers", "x_type": "ANIMAL", "y": "Emma", "y_type": "PERSON", "r": "pet"}},
]
```

Input:
{input_dialogue}

Output:

Figure 5.9.: One-Shot RE Prompt Template: This template represents our preliminary endeavor in RE tasks utilizing the ChatGPT model. Variables are denoted in blue as a reflection of their dynamic nature. In the earlier versions of this prompt, we used "x" and "y" as subject and object keys, respectively, following the DialogRE schema. We have since switched to using "subject" and "object" due to better empirical results.

Pick one ontology label describing the subject-object link. Only the label.

Ontology:
- Relations: ("acquaintance", "age", "alternate_names", "alumni", "births_in_place", "boss", "children", "client", "date_of_birth", "dates", "employee_or_member_of", "employees_or_members", "friends", "girl/boyfriend", "major", "negative_impression", "neighbor", "origin", "other_family", "parents", "pet", "place_of_birth", "place_of_residence", "place_of_work", "positive_impression", "residents_of_place", "roommate", "schools_attended", "siblings", "spouse", "students", "subordinate", "title", "unanswerable", "visited_place", "visitors_of_place", "works")

Input Dialogue: {input_dialogue}

Subject: {input_subject}
Object: {input_object}
Relation:

Figure 5.10.: Optimized Prompt Template for RC: This template was crucial in comparing the performance of the LLaMA model to other RC frameworks like BERT and XGBoost. Dynamic variables within the template are in blue. This prompt configuration was identified as the most accurate after extensive experimentation.

Extract entities and relations from the dialogue. Return a Python list of JSON objects, each fitting this schema:

```
{
    "subject": "<Entity>",
    "relation": "<RELATION_TYPES>",
    "object": "<Related Entity>"
}
```

No additional text or explanations. Return an empty list if no relevant entities or relations are found. Stick to the provided relations. You are like an API, you don't speak you only return JSON objects. Dialogue: {input_dialogue}

Figure 5.11.: Streamlined Entity-RE Prompt Template: This template is essential for enhancing RE tasks utilizing the LLaMA model. Variables are marked in blue, representing their dynamic nature. Extensive testing has demonstrated this format to produce the highest performance. Note: RELATION_TYPES is a placeholder for a string of all possible relationships separated by a slash, such as 'siblings/spouse'.

conversation segments that led to the deduction of these connections were extracted and subsequently employed to create follow-up questions informed by memory.

Generating Personalized Responses With our dual representation of memory, as subgraphs of relation triplets and associated dialogue turns, we have the ability to proactively construct follow-up questions. We have developed the optimal prompt template, as shown in Figure 5.12, based on our proof of concept. Although this approach is effective for ChatGPT, it may require alternative adaptations for deployment with smaller models. Beyond the custom follow-up inquiries, our system is programmed to engage with users in a cordial and lighthearted manner utilizing the prompt template showcased in Figure 5.13.

5.4. RQ4: Evaluation Methodologies

This section details the evaluation methodologies employed in this thesis, offering insights into their relevance and effectiveness in our research scope. Evaluation is critical to the development and refinement of dialogue systems, with a traditional emphasis on human-centered assessments such as surveys and user feedback. These methods can be resource-intensive in terms of time and cost, as highlighted by Deriu et al. [13]. Thus, our evaluation strategies for RE in this thesis were designed with efficiency in mind.

Our Choice for Classification Metrics We prioritized metrics for classification that assess the reconstruction accuracy of relation triples, as explained in Section 4.6. This decision aligns with our goal of enhancing the precision of RE in dialogue systems. The current results of our system, which is in the developmental stage rather than production-ready, justify our use of these classification metrics. They provide a practical and efficient method of measuring progress and identifying areas for improvement. As the system evolves and potentially reaches a suitable stage for deployment, we can reevaluate and expand upon these metrics. Further discussion of the results' implications and directions for system improvement will be explored in the subsequent discussion section.

You're an AI named {bot_name}, focused on engaging in friendly, lighthearted conversations. Your task is to create a follow-up question, based on the input knowledge of the user, named {user_name}. {user_name} is an elderly person.

Input (Topic: places):

```
[{ "subject": "Bob", "relation": "visited_place", "object": "Stuttgart" }]
```

{
'{bot_name}: Hello, Bob, it's {bot_name} here! Can we talk now? Tell me about a cherished memory of yours. I'd love to hear it!',
'Bob: I loved this time I went to Stuttgart.',
}

Output:
{bot_name}: Hi, Bob, it's {bot_name} again! Can we chat? I was thinking about when you told me about Stuttgart. Tell me more!

Input (Topic: {topic}):
{relation_list}
{chat_history}

Create a follow-up question for the example below. Keep it concise up to 20 words. You MUST ASK if the user has time to chat. Be very specific with the information in the input. Make a statement while mentioning the info in the input.

Output:
{bot_name}:

Figure 5.12.: Enhanced Prompt Template for Memory-Based Follow-Up Questions: This template was developed to generate context-aware follow-up questions and demonstrated effective use of prompt engineering during our proof of concept phase. To utilize the AI bot's memory for creating more personalized interactions, we integrated OpenAI's ChatGPT with a Neo4j Database. In blue are the variables to fill upon every new inference step.

You're an AI named {bot_name}, focused on engaging in friendly, lighthearted conversations.

For example:

Chat 1 (user wants to talk)
{bot_name}: Hi, {user_name}, it's {bot_name} again! Can we chat? I want to know if your back is better.
{user_name}: I still feel pain, even though Phillip applied some pain cream.
{bot_name}: I'm sorry you're still in pain. But I'm sure it will get better. Who's Phillip, if I may ask?
{user_name}: Thanks. He's my husband.
{bot_name}: That is great! How long have you been together?

Chat 2 (user does not want to talk)
{bot_name}: Hi, {user_name}, it's {bot_name} again! Can you talk now? I wanted to know how your back is doing.
{user_name}: No...
{bot_name}: No worries! I hope your back improves soon. I'm here when needed.

Chat 3 (user does not understand message)
{bot_name}: Hi, {user_name}, it's {bot_name} again! Can you talk now? I wanted to know how your back is doing.
{user_name}: What? Who are you? Why are you asking me that?
{bot_name}: I'm {bot_name}, designed to track your health. Sharing more about you helps us boost your well-being together!

Keep is as brief as you can, always try to reply with up to 20 words.

Remember, your priority is to know who mentioned people are first.

Try ask about the last mentioned entity or person by the user, {user_name}.

Say the user name, {user_name}, often.

Figure 5.13.: Preliminary One-Shot Response Generation Template: This template aims to guide structured conversations between our agent and an elderly patient and to integrate historical dialogue into the ChatGPT API call's system message. Such integration ensures that responses comply with the established conversation guidelines based on either customized follow-ups or a predetermined set of conversation starters. In blue are the variables to fill upon every new inference step.

6. Discussion

This chapter presents a thorough analysis of the study's findings. To ensure coherence and ease of understanding, the structure of this discussion mirrors that of the Results chapter (refer to Chapter 5).

6.1. RQ1: Concepts and Entities for Data Model

This section focuses on the insights gained regarding our data model, which were originally presented in Section 5.1. We will critically evaluate these findings, considering their implications and potential applications.

Interpretation and Impact of Streamlined Data Analysis The significant decrease in the number of relationships observed in the DialogRE dataset, which was reduced from 36 to 11, is a vital step towards streamlining data analysis for geriatric care communication. Guided by Kitwood's framework, this filtration process not only made analysis more manageable, but also enhanced the importance and specificity of the relationships within the context of elderly care. Consequently, this method improves the relationship metrics, which ensures a more focused and efficient analysis, as will be further discussed in following subsections.

Limitations of the Dataset and Its Implications One significant limitation of this study is the use of the DialogRE dataset, which originates from the TV show "Friends." While this dataset offers a formatted structure for dialogue relationships, it varies from the intended distribution of dialogues within the geriatric care setting. The humorous and occasionally senseless quality of these dialogues presents a difficulty in obtaining pertinent and contextually suitable relationships for elderly care. Furthermore, most conversations involve more than two participants, further deviating from our target distribution. This discrepancy emphasizes the necessity for a dataset that is more accurately aligned with the communication dynamics of geriatric care.

Practical Implications in Geriatric Care Despite its limitations, the filtered data model may provide significant implications for geriatric care. By organizing information in a structured format, it enables the development of customized interactions based on specific user relationships, including spouse, sibling, or pet. This approach permits a

controlled manner to personalize conversations with elderly individuals, enhancing the overall quality of care and interaction.

Comparison with Existing Literature The lack of prior literature that integrates Kitwood's psychological needs framework with computational data models presents a challenge when assessing the results of this study. Nonetheless, this underscores the novelty and possible impact of this research in geriatric care, paving the way for more comprehensive and data-driven investigations in this field.

Figure 6.1.: An example conversation illustrates how DialogRE deviates from our desired distribution. While it may be appropriate for a TV show, it lacks context without accompanying footage and is not representative of our intended geriatric communication target distribution.

Future Research Directions Further research could explore alternative strategies to augment the dataset following Kitwood's framework or other pertinent psychological theories. Time constraints prohibited an investigation of these aspects in this thesis, but they offer promising avenues for boosting the model's usefulness and efficacy in personalized communication for elderly care.

The integration of Kitwood's framework with the DialogRE dataset represents a important advancement in the merging of psychological requirements and computational dialogue analysis for the elderly care sector. Despite limitations posed by the dataset utilized, this study establishes a basis for prospective research endeavors in this domain, potentially culminating in more refined and compassionate communication approaches in geriatric care.

6.2. RQ2: Information Extraction Techniques

6.2.1. Methodological Ablation Studies

Experiment Overview Tables As depicted in tables 5.1, 5.2 and 5.3, the thesis began with an investigation into RC, starting with replicating the DialogRE paper utilizing BERT-Tiny and BERT-Base models. This crucial foundational step established a baseline and integrated necessary components into the training pipeline, including per-label metrics. The addition of the 'no relation' label proved to be a crucial expansion, which brought the model closer to real-world situations but also introduced expected inaccuracies into the prediction. This initial phase brought to light the model's inherent tendency to favor certain labels, an essential aspect that requires further examination. The research was conducted with an objective approach, employing various architectures, including BERT, BART, and LLaMA, along with diverse preprocessing techniques, such as oversampling, undersampling, filtering, and feature engineering. This iterative process highlighted the importance of methodology. The study found that BERT and LLaMA were particularly effective, with LLaMA showing promising results in RE.

The experiments revealed a nuanced understanding of RE, specifically in personal relationships within elderly care. Through exploring various models and preprocessing methods, the study provided insights into the advantages and disadvantages of each approach. For example, LLaMA demonstrated enhanced performance, while the use of BERT facilitated a connection to existing literature, although with relatively worse results. The study also revealed a significant challenge presented by the limited dataset, which seemed to lack a strong signal for the intended tasks.

The outcomes of this thesis establish a foundation for forthcoming research, primarily in the territory of data amplification and tailored dataset annotation. The application of the SlideFilter approach for data amplification materialized as a promising tactic, proposing the possibility for more targeted datasets in the future. We also contributed to

prompt engineering, if the strategy of LLMs is still to be explored in the future. Further studies should focus on developing datasets that reflect the target distribution within the elderly care domain, thereby improving the model's applicability and accuracy. Moreover, refining the balance of the 'no relation' count and exploring additional models could yield valuable insights. The conclusion of this thesis asserts that although noteworthy progress has been achieved, there is still considerable room for further exploration and enhancement in the realm of AI-powered RE for elderly care.

Waterfall Chart Analysis of Ablation Studies In our ablation studies, we utilized a waterfall chart format to present the macro F1 scores, providing a visual comparison of each model's performance across stages (see figures 5.1, 5.2, and 5.3). Our exploration began with RC, utilizing BERT and LLaMA as our primary models. BERT fine-tuned on DialogRE performed as a baseline, setting the standard for subsequent enhancements. The implementation of the 'no relation' label, while aiming to align the model with real-world situations, resulted in added noise that negatively impacted the accuracy of the model. It was observed that despite these challenges, BERT's fine-tuned performance exceeded that of ChatGPT 3.5 Turbo by a considerable margin, as shown in Figure 5.1. This could be attributed to the latter's tendency towards hallucinations and ambiguities within the given dataset.

Shifting attention to LLaMA, a noteworthy performance improvement was observed compared to BERT in Figure 5.2. This improvement emphasizes the efficiency of LLaMA in extracting and labeling relational data. However, similar to BERT, the addition of the 'no relation' label resulted in a dip in performance. Nonetheless, LLaMA remained ahead of ChatGPT by a significant margin. This comparative analysis highlights the robustness of LLaMA in RC, despite the challenges posed by more complex label structures.

Our study's critical phase was marked by an ensemble approach, which combined BERT and LLaMA, as presented in Figure 5.3. We went beyond the scope of DialogRE and developed a novel approach to explicit RI by using spaCy, XGBoost, and BERT to classify 11 potential labels, i.e., solving the broader RE task. Our study uncovered a complex interplay between explicit and implicit RI approaches. While avoiding XGBoost, the implicit method delivered marginally better results. However, it is noteworthy that the LLaMA model has a tendency to overfit in RE tasks, which is different from its behavior in RC. This overfitting poses a significant challenge when working with a small number of examples as it results in inaccuracies in classifying instances where no relation exists. Despite this limitation, our BERT-ensemble and LLaMA architectures consistently outperform ChatGPT. This superiority, however, was limited by the relatively modest overall results attained by both models.

The implementation of the SlideFilter technique was a significant improvement, particularly in balancing the dataset and enhancing overall performance. However, there is still a persistent issue in accurately predicting 'no relation' labels that needs

further refinement. Our experiments with LLaMA have shown promising avenues for future research, particularly in the context of a more evenly distributed dialogue turn and fewer relations. These findings highlight the potential of custom datasets and tailored models to improve AI-driven RE, particularly in the field of elderly care. In summary, our research not only enhances comprehension of RE methods but also establishes a foundation for future advancements in the field.

Quantitative Experiment Breakdown The results from tables 5.4 and 5.5 illustrate distinct behaviors in the performance of models on RC and Extraction. In RC, LLaMA's macro average F1 score of 49% (e06b) reflects an 8 percentage point improvement over BERT (e07), indicating LLaMA's effectiveness in relation modeling post-fine-tuning. However, the results also reveal a challenge common to both LLaMA and ChatGPT in the RE task, particularly in experiments e14 and e16, where both models exhibit high recall but low precision for the 'no relation' classification.

This set of findings emphasizes the nuanced balance required when fine-tuning LLMs to predict 'no relation' instances. Our efforts to adjust the threshold for 'no relation' classifications expose significant sensitivity, causing the models to often alternate between underpredicting and overpredicting 'no relation'. This sensitivity indicates that LLMs may tend towards one extreme or another: failing to recognize instances of 'no relation' which leads to a high rate of false negatives, or frequently predicting 'no relation' resulting in an abundance of false positives. Achieving an optimal threshold that attains a precise balance poses a nuanced challenge that demands careful consideration to prevent compromising model precision.

The observation has two-fold implications. First, the data implies that the 'no relation' category is inherently complex and noisy, which presents a significant challenge for LLMs that aim to make precise predictions. Second, it raises concerns about the LLMs' appropriateness for tasks that require high precision in RE. While the use of fine-tuning has demonstrated improvement in overall performance, the persistent problem of high recall but low precision highlights the need for the model's approach to 'no relation' occurrences to be recalibrated. Future research could consider implementing sophisticated methods, including cost-sensitive learning or negative sampling, to enhance the predictive performance of the models. It is essential to determine whether this pattern is an attribute of the model's design or a manifestation of the intrinsic noise in the 'no relation' class, which is vital for the progress of LLMs in real-world scenarios that depend on accurate RE.

6.2.2. Strategic Enhancements and Visualized Outcomes

Comparative RI Results Investigating different methods for identifying relations has shown that an implicit approach (e12), incorporating a 'no relation' class in BERT's RC, performs better than explicit approaches (e11) that require a BERT ensemble in

combination with XGBoost as illustrated in Figure 5.3. This finding was surprising because we had expected the ensemble method to enhance our feature engineering capabilities, which we utilized with XGBoost. Interestingly, XGBoost achieved similar results to BERT while using engineered features for RI, contradicting the assumption that BERT's contextual understanding would lead to superior outcomes as shown in Figure 5.4. This performance parity suggests that the classification of 'no relation' may contain intrinsic noise, emphasizing the need for careful feature selection and model training to increase the signal-to-noise ratio in relation recognition tasks. This issue could be tackled in future research.

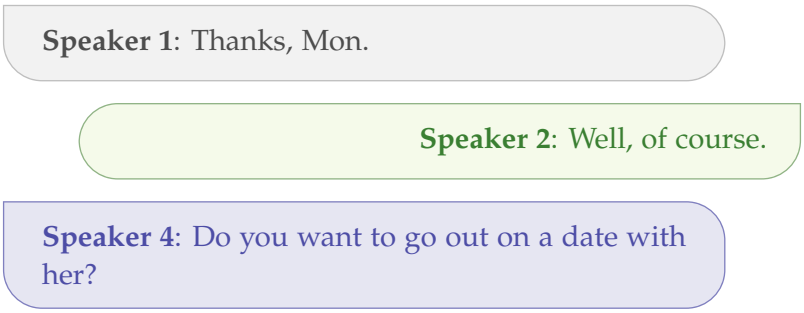
Efficiency Analysis of SlideFilter Augmentation The SlideFilter augmentation greatly improved the model's ability to extract relationships. This increase is verified by a boost of 11% in F1 scores, rising from 21.5% to 30.5%, when comparing the 3 Turn Cap SliderFilter against its absence. The impact is illustrated in figure 5.5. The token distribution across inputs and outputs was narrowed using this approach, resulting in a more consistent dataset that more closely adhered to the model's length limitations. Limiting dialogue samples to three turns streamlined the learning process and enhanced the interpretability of data, revealing a correlation between sample simplicity and ease of understanding. The SlideFilter's ability to reduce complexity presents a compelling avenue for future research, making it a notable advancement in our methodology. Further studies should encompass human evaluation to verify the efficacy of the SlideFilter, particularly in detecting relationships lacking explicit entity mentions and covering prolonged discussions, which our method can not handle. By scrutinizing these aspects, we can enhance our approach to ensuring extensive RE that accommodates a broader range of conversational contexts.

Visualized Confusion Matrices

RC The confusion matrices in Figures 5.6 and 5.7 provide informative per-label metrics for several experiments in our RC system. It is worth noting that the inclusion of the 'no relation' label introduces noise, as demonstrated by the considerable misclassifications appearing in the 'no relation' column (experiments e01 vs e03, e05 vs e07, and e02 vs e04) . The high prominence of this column within the matrices implies a significant number of misclassifications of other labels as 'no relation', which aligns with our expectation of increased noise.

The strength and prominence of the diagonal in these matrices are indicative of the system's performance - the stronger and more prominent the diagonal, the better the system performs. Ideally, the matrices should exhibit high values along the diagonal signifying correct classifications and low values elsewhere. Certain labels, such as 'acquaintance', 'place of residence', and 'visited place', are particularly challenging to model, as depicted in Figure 6.2. Experiment e05 highlights this difficulty, as these

labels are scarcely represented even when analysing BERT with Focus Relations only. The inclusion of 'no relation' only exacerbates the confusion (e07).



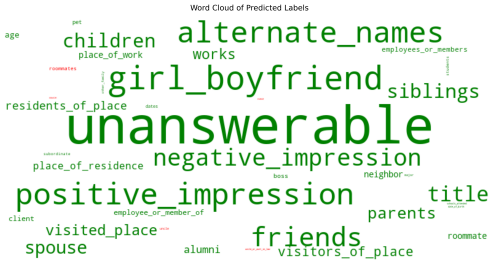
Ground Truth:

```
[ {"subject": "Speaker 1", "relation": "acquaintance", "object": "Speaker 2"} ]
```

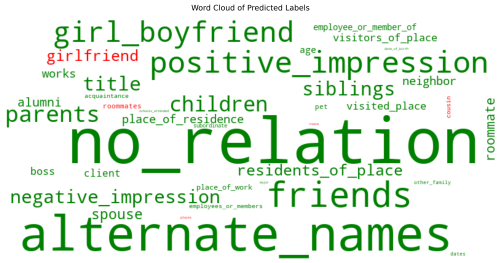
Figure 6.2.: Example dialogue illustrating the complexity of identifying the 'acquaintance' label. The interaction between Speaker 1 and Speaker 2 could easily be misinterpreted since it lacks context, highlighting the challenges of accurately classifying relational contexts.

By contrast, in Experiment e06b with LLaMA trained on focused relations, the diagonal appears more robust, suggesting a more balanced classification across labels, despite notable errors. This suggests that LLaMA may offer a more equitable distribution of attention across labels that BERT may overlook. Nonetheless, labeling issues persist, particularly with the 'acquaintance' label, frequently confused with 'family', 'parents', and 'siblings' even under LLaMA's classification. Interestingly, ChatGPT demonstrates reasonable performance even without fine-tuning. Finally, there are a few confabulated labels, as shown in the word clouds of Figure 6.3. It is important to acknowledge that this analysis may be biased because of the absence of labels beyond our predetermined ontology.

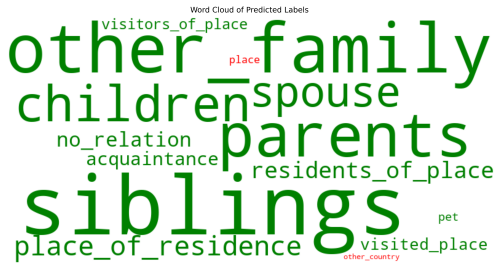
RE When examining Figure 5.8 regarding RE, it is important to exercise caution when interpreting the confusion matrices. The matrices concentrate mainly on relation labels, disregarding the subjects and objects of the triplets. As a result, they function more as label co-occurrence matrices than as precise performance indicators. For BERT models e11 and e12, it is noted that the implementation of an implicit RI technique results in an increase in diagonal strength, indicating better extraction performance. Conversely, with ChatGPT (e14), the diagonal becomes sparser and suggests a decrease in performance, despite a seemingly strong diagonal. The scarcity is especially noticeable in labels such as 'spouse', 'visited_place', and 'visitors_of_place'



(a) e02 LLaMA



(b) e04 LLaMA w/ 'No_Rel.'



(c) e06b LLaMA Focus-Rels w/ 'No_Rel.'

(d) e04b ChatGPT3.5 w/ 'No_Relation'

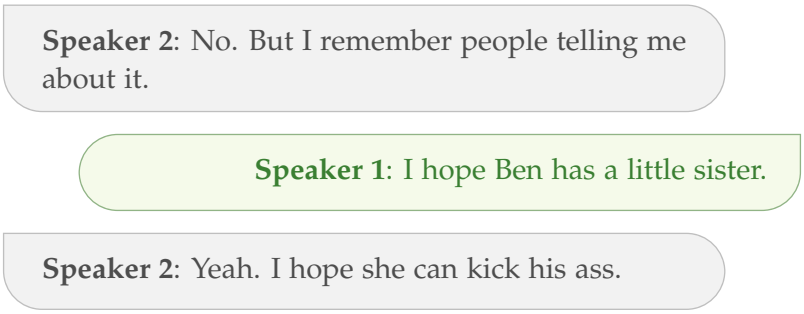
Figure 6.3.: Word Clouds for RC: Red labels indicate confabulations.

with LLaMA (e13), which have high scores but also significant misclassifications. This pattern, likely attributable to LLaMA's skilled handling of wider contexts, results in similar misclassifications in other labels.

ChatGPT demonstrates a strong diagonal trend in e14, but this is offset by a notable occurrence of hallucinations, as evidenced in the accompanying word clouds of Figure 6.5. Additionally, the use of the 'acquaintance' label presents inconsistencies, as it is distributed among various focus labels. The LLaMA with SlideFilter (e15) encounters challenges in accurately modeling 'no relation' and produces a convoluted portrayal of the 'siblings' label despite its data simplification efforts, revealing the limitations of the filter's context reduction ability, as depicted in Figure 6.4. After tweaking the null relation (e16), LLaMA exhibits an enhanced modeling of 'no relation', although certain labels such as 'visitor', 'place_of_residence', and 'resident_of_place' experience a decrease in accuracy. Despite its trade-offs, e16 appears as the most well-balanced option, pointing towards a potential route for enhancing methodological practices in RE. Therefore, future research could use SlideFilter along with post-human evaluation for faster dataset curation.

6.2.3. Evolution of Prompt Design

Foundational One-Shot Template In the foundational one-shot template for RE shown in Figure 5.9, the initial prompt design used generic placeholders 'x' and 'y'



Ground Truth:

```
[ {"subject": "Speaker 1", "relation": "siblings", "object": "Speaker 2"},
  {"subject": "Speaker 2", "relation": "other_family", "object": "Ben"},
  {"subject": "Speaker 2", "relation": "siblings", "object": "Speaker 1"},
  {"subject": "Ben", "relation": "other_family", "object": "Speaker 2"} ]
```

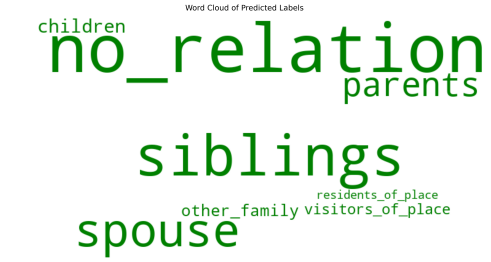
Predictions:

```
[ {"subject": "Ben", "relation": "siblings", "object": "Speaker 2"} ]
```

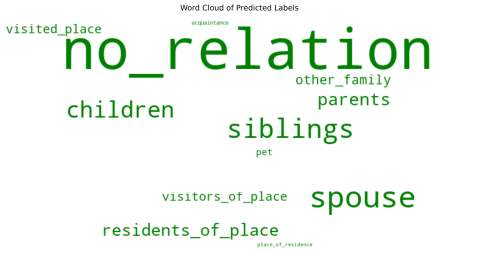
Figure 6.4.: This dialogue example demonstrates misclassification by SlideFilter, which uses a rule-based approach to determine relationships without contextual filtering. The ground truth indicates complex family relationships, which the model oversimplifies by combining multiple relationships into a single "sibling" relationship between Ben and Speaker 2. This highlights the need for more nuanced processing to accurately handle such complex relational dynamics.

for subjects and objects, which proved effective for larger models such as ChatGPT 3.5. Nevertheless, smaller models, like LLaMA, showed enhanced learning curves when employing semantically descriptive keys, indicating the benefit of semantic prompts in boosting performance. Initially using a one-shot template due to the lack of fine-tuning in ChatGPT, later optimization eliminated this need, in agreement with findings by Wei et al. [55] that fine-tuning reduces the reliance on one-shot examples. This optimization simplified the prompts, preserving processing resources while also maintaining informative ontologies and type lists within the prompts to create dataset-specific mappings from dialog input to JSON output.

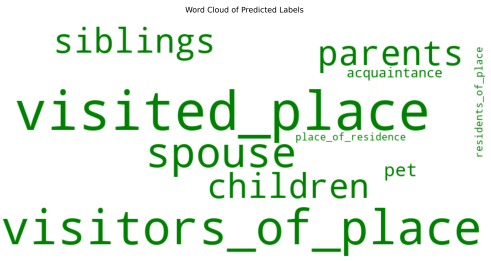
Optimized Extraction and Comparative Analysis The second phase involved refining prompts to benchmark LLaMA against BERT for the RC task, as demonstrated in 5.10. The fine-tuned LLaMA performed well with prompts that listed possible relations, followed by the subject and object, eliminating the one-shot approach. The resulting output was exclusively the relation label, indicating LLaMA's promising effectiveness in our dataset. Further advances in Figure 5.11 for RE resulted in LLaMA's prompts



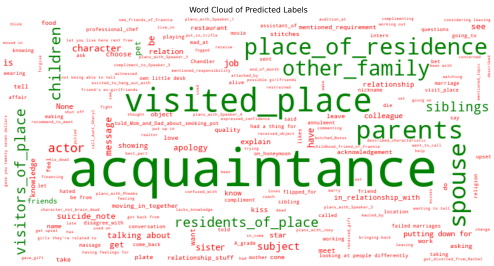
(a) e11 BERT ensemble-11cls



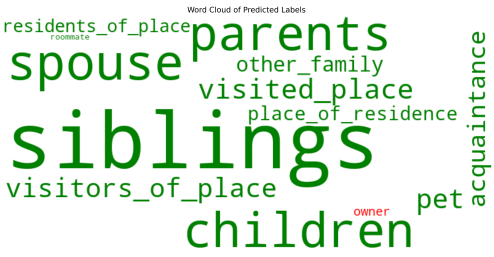
(b) e12 BERT ensemble-12cls-implIdent



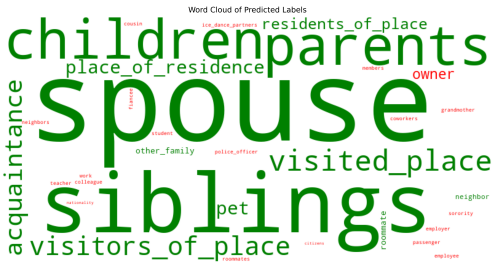
(c) e13 LLaMA



(d) e14 ChatGPT-3.5



(e) e15 LLaMA SlideFilter



(f) e16 LLaMA SlideFilter NullRel Tweak

Figure 6.5.: Word Clouds for RE: Red labels indicate confabulations.

incorporating a schema for desired relationships, leading the model to make accurate predictions without hallucinations, albeit still with mistakes. This iterative evolution highlights LLaMA's potential and establishes a benchmark for future investigations into other finely-tuned models like Zephyr or Orca [66, 67], which could provide improved relational comprehension through their training on diverse datasets.

6.3. RQ3: Knowledge Integration

Memory Recall Implementation Our developed graph search method effectively obtained relation triplets and their corresponding conversation turns that generated them, enabling us to avoid errors and achieve satisfactory results during our demo.

However, this search methodology has not undergone extensive testing, providing opportunities for future investigations to refine these approaches, assuming that the graph structure continues to be an integral part of the knowledge base.

Generating Personalized Responses We have devised a prompt utilizing subgraph data to produce tailored follow-up question, illustrated in Figure 5.12. Our approach entails incorporating relation-specific triplets and chat history into a one-shot example to formulate contextually applicable questions. While this approach proved effective with more complex models, it is advisable that future work use a smaller, fine-tuned, large language model with zero-shot prompts to solve this task. As depicted in Figure 5.13, ChatGPT's response policy was often adaptable and concise. However, maintaining policy conformity without fine-tuning may pose a challenge for models with smaller models. Loh et al. [68] propose that fine-tuning LLMs on the EmpatheticDialogues dataset [69] holds promise for generating empathetic responses. This approach may prove more valuable for future research than precise prompt engineering.

6.4. RQ4: Evaluation Methods

Our Choice for Classification Metrics Our system's classification performance was assessed using the F1 score, along with precision and recall, as explained in subsections 4.6.1 and 4.6.2. These metrics have been effective in measuring our knowledge graph's reconstruction abilities and provide a strict standard for evaluating our multi-class and multi-label classification tasks. However, these metrics may be considered too inflexible, leading to the exploration of more adaptable options in future research. Possible alternatives could involve utilizing embedding spaces for outputs or adopting BLEU (BiLingual Evaluation Understudy) scores, which assess deviations in contextuality and sequence generation more loosely. Employing such metrics as the foundation for a novel loss or regularization term could provide a more nuanced framework for model learning.

Additionally, the efficacy of a knowledge graph can be ascertained by its performance in the full conversational system. Possible areas of future research could incorporate indirect indicators of effectiveness, such as both qualitative and quantitative feedback from users, in order to evaluate how well the knowledge graph supports the retrieval of accurate and relevant contextual memory. Investigating these possibilities will produce a comprehensive evaluation framework that covers both the immediate output quality and the wider effects on user interaction.

7. Conclusion & Outlook

This chapter presents a summary of the research results, acknowledges the difficulties encountered throughout the study, and proposes potential avenues for future research.

7.1. Summary

Throughout the study, as described in Chapter 6, we aimed to comprehend and utilize state-of-the-art LLMs to create knowledge graphs for organizing personal data. This approach is also relevant in geriatric care. Our findings demonstrate that current LLMs, despite their high level of sophistication, encounter considerable difficulties meeting the intricate requirements of these knowledge graphs, when using the publicly available datasets for fine-tuning. The core is a careful exploration of the limits of current NLP technologies for the task of RE.

7.1.1. Contributions

Insights into Data Limitations A major contribution of this thesis is the comprehensive review of the constraints of state-of-the-art NLP architectures emphasized in Sections 6.1 and 6.2. Our trials utilizing models such as BERT, LLaMA, and ChatGPT indicated certain difficulties in managing intricate relational data structures. The DialogRE dataset, which is based on the TV show "Friends", illustrates the challenges that LLMs encounter when extracting pertinent relationships for elderly care due to its diverse and often humorous nature, as depicted in Figure 6.1.

We suspect that the poor performance is because knowledge is fragmented across multiple turns and speakers in DialogRE. This, combined with significant variation in dialog length and number of relationships, makes extracting relationships a challenge. In contrast, conventional assistant-to-human conversations exhibit more direct interactions, with queries and utterances being more self-contained and contextually independent. Each exchange typically hones in on specific information with a greater chance of including all necessary details within one turn, enabling clearer and more effective communication. This observation highlights the necessity of creating a customized dataset that aligns with our specific needs, one that reflects the direct and self-contained manner of interactions crucial for successful personalized communication.

Novel Approaches Our study aimed to expand current methods of personal RE by integrating Kitwood's psychological framework with computational models. This innovative approach is explained in detail in Section 6.1. The existing datasets have also been customized to serve as baseline comparisons. We utilized dialogue datasets from the literature, including DialogRE, to pursue our objectives. Although dialog-based, DialogRE was augmented by our SlideFilter method to better match actual speaker interactions. Our novel approach, outlined in Section 6.2, involved exploring a range of LM architectures, including BERT, LLaMA, and ChatGPT, resulting in significant findings in RE techniques.

This adaptation of DialogRE utilizing SlideFilter yielded valuable insights into our method's failure modes. One noteworthy limitation we discovered was that the filter is not entirely reliable, highlighting the difficulties in extracting and structuring conversational data with precision, as depicted 6.4. Nonetheless, this limitation presents opportunities for future research. The SlideFilter, although not without limitations, has the potential to preprocess datasets into more manageable segments effectively. However, there is a trade-off to consider in terms of window size for chunking samples during this segmentation process. Shorter window sizes lead to more focused and less noisy dialogues, but may not capture all of the complexities that define relations in a dialogue. Longer window sizes result in longer dialogues, creating less focused and noisier samples. Contextual intricacies for RE are present, but the noise may hinder the task. Thus, achieving a proper balance of window size is essential for effective application of this technique. Moreover, when combined with subsequent human evaluation, this methodology has the potential to be an effective approach for improving the precision and applicability of data in communication models for personal RE in geriatric care.

Practical Impacts In practical terms, our study demonstrates a pivotal finding: when fine-tuned to publicly available datasets, current LLMs lack the necessary capabilities to proficiently execute the RE task that is vital to the formation of knowledge graphs, particularly in the intricate personal domain required for geriatric care. Section 6.2 elaborates on this, as we discovered the obstacles that models like BERT, LLaMA, and ChatGPT encounter when accurately identifying and extracting relational information.

We suggest that choosing a simpler structure, rather than a complicated KG, could result in more efficient outcomes. Starting with the use of rule-based systems, regular expressions (regex), to produce a key-value structure of conversation snippets could result in more dependable and efficient of recalling information. Although less intricate than LLMs, such a methodology could offer lower latency and fewer errors in structuring memory data. Furthermore, this method enables a methodical and gradual advancement towards more complex systems while maintaining control and response time as indispensable elements. This approach could be promising as a primary step towards forthcoming advances in the field, balancing present-day technological

capabilities with the detailed requisites of geriatric care communication.

7.2. Future Work

The insights obtained from this thesis provide opportunities for future research in several directions.

- **Simplifying Data Structures**: Future research may concentrate on simplifying the memory data structures as our findings indicate that current models are unable to precisely automate the intricate knowledge graph constructions. To achieve this, we could loosen the constraints of existing ontologies, as discussed in Section 6.1 by dropping the strict list of possible relations from DialogRE as utilized in 6.2. Instead, we could focus on a basic data structure, such as a key-value structure of conversation snippets, or even a co-occurrence KG. Although simpler, this could still enable memory to be extracted in a controlled manner due to its structured nature.
- **Exploring Hybrid Systems**: Expanding on the findings presented in Sections 6.2 and 6.3, this research suggests a potential opportunity for the development of hybrid systems that fuse the control of rule-based approaches with the adaptability of LLMs. The aim is to leverage the advantageous attributes of both paradigms. Therefore, employing rule-based systems to structure the bot's memory in a controllable and low-latency specification is recommended. Furthermore, utilizing LLMs in generating instruction-based responses as RAG, i.e. where they excel, is advised.
- **Collecting a Custom Dataset**: Leveraging an appropriately curated dataset, we find that this task can be effectively addressed using LLMs, as evidenced by promising results in the existing literature, notably the REBEL framework [25]. Therefore, we suggest that compiling a dataset from real human-assistant interactions is a meaningful direction for future research. Developing a simple conversational assistant and involving human annotators to identify relational dynamics could be a solution. Utilizing tools such as SpaCy to extract entity pairs can optimize the process and mitigate the annotation labor intensity. However, it is important to acknowledge that this undertaking is both time-consuming and requires significant resources.

In summary, this thesis advances understanding of the capabilities and limitations of current NLP technologies to construct KGs with social relational content that could later be used to personalize geriatric care conversations. The findings suggest that the journey is ongoing, but offers a promising path towards innovation and improvement.

A. Figures

A.1. DialogRE Relation Types

ID	Subject	Relation Type	Object	Inverse Relation
1	PER	per:positive_impression	NAME	
2	PER	per:negative_impression	NAME	
3	PER	per:acquaintance	NAME	per:acquaintance
4	PER	per:alumni	NAME	per:alumni
5	PER	per:boss	NAME	per:subordinate
6	PER	per:subordinate	NAME	per:boss
7	PER	per:client	NAME	
8	PER	per:dates	NAME	per:dates
9	PER	per:friends	NAME	per:friends
10	PER	per:girl/boyfriend	NAME	per:girl/boyfriend
11	PER	per:neighbor	NAME	per:neighbor
12	PER	per:roommate	NAME	per:roommate
13	PER	per:children*	NAME	per:parents
14	PER	per:other_family*	NAME	per:other_family
15	PER	per:parents*	NAME	per:children
16	PER	per:siblings*	NAME	per:siblings
17	PER	per:spouse*	NAME	per:spouse
18	PER	per:place_of_residence**	NAME	gpe:residents_of_place
19	PER	per:place_of_birth**	NAME	gpe:births_in_place
20	PER	per:visited_place	NAME	gpe:visitors_of_place
21	PER	per:origin*	NAME	
22	PER	per:employee_or_member_of*	NAME	org:employees_or_members
23	PER	per:schools_attended*	NAME	org:students
24	PER	per:works	NAME	
25	PER	per:age*	VALUE	
26	PER	per:date_of_birth*	VALUE	
27	PER	per:major	STRING	
28	PER	per:place_of_work	STRING	
29	PER	per:title*	STRING	
30	PER	per:alternate_names*	NAME/STRING	
31	PER	per:pet	NAME/STRING	
32	GPE	gpe:residents_of_place**	NAME	per:place_of_residence
33	GPE	gpe:births_in_place**	NAME	per:place_of_birth
34	GPE	gpe:visitors_of_place	NAME	per:visited_place
35	ORG	org:employees_or_members	NAME	per:employee_or_member_of
36	ORG	org:students*	NAME	per:schools_attended
37	NAME	unanswerable	NAME/STRING/VALUE	

Figure A.1.: List of all relation types in the DialogRE dataset. The table presents all relationships derived from the dataset. The 'Inverse Relation' column shows the corresponding reverse relationship for each type. [12]

List of Figures

List of Tables

Acronyms

DS Dialogue System. 10

KG Knowledge Graph. 10, 11, 65, 66

LLM Large Language Model. 2, 15, 20, 22, 24, 31, 46, 56, 57, 63–66, 68

LM Language Model. 8, 20, 65

LoRA Low-Rank Adaptation of Large Language Models. 16

NER Named Entity Recognition. 10, 12, 29, 30, 68

NLP Natural Language Processing. 1, 6, 14, 17, 26, 27, 64, 66

PKG Personal Knowledge Graph. 2–8, 11, 12, 20, 24, 37, 68

PLM Pre-trained Language Model. 2, 6, 10, 13, 14, 20, 21

RAG Retrieval-Augmented Generation. 2, 66

RC Relation Classification. 5, 10, 13, 17, 20–22, 29, 30, 32, 33, 37, 38, 40, 41, 43–46, 49, 55–58, 60, 61, 68–70

RE Relation Extraction. 4, 5, 7, 8, 10, 13, 17, 18, 20–22, 24, 26, 28, 30–32, 34, 37, 39–44, 47–50, 55–62, 64, 65, 68–70

RI Relation Identification. 5, 10, 20–22, 29, 37–39, 41, 42, 56–59, 68, 70

Bibliography

[1] P. Schneider, N. Rehtanz, K. Jokinen, and F. Matthes. "Voice-Based Conversational Agents and Knowledge Graphs for Improving News Search in Assisted Living". In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '23. Corfu, Greece: Association for Computing Machinery, 2023, pp. 645–651. ISBN: 9798400700699. DOI: 10.1145/3594806.3596534. URL: https://doi.org/10.1145/3594806.3596534.

[2] M. McTear, Z. Callejas, and D. Griol. *The Conversational Interface: Talking to Smart Devices*. Cham: Springer, 2016. ISBN: 978-3-319-32965-9. DOI: 10.1007/978-3-319-32967-3.

[3] M. Valizadeh and N. Parde. "The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6638–6660. DOI: 10.18653/v1/2022.acl-long.458. URL: https://aclanthology.org/2022.acl-long.458.

[4] Crewdson and J. Crewdson. "The Effect of Loneliness in the Elderly Population: A Review". In: *Healthy Aging Clinical Care in the Elderly* 8 (Mar. 2016), p. 1. DOI: 10.4137/HACCE.S35890.

[5] C. Luanaigh and B. Lawlor. "Loneliness and the health of older people". In: *International journal of geriatric psychiatry* 23 (Dec. 2008), pp. 1213–21. DOI: 10.1002/gps.2054.

[6] K. Balog and T. Kenter. "Personal Knowledge Graphs: A Research Agenda". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '19. Santa Clara, CA, USA: Association for Computing Machinery, 2019, pp. 217–220. ISBN: 9781450368810. DOI: 10.1145/3341981.3344241. URL: https://doi.org/10.1145/3341981.3344241.

[7] Cambridge Dictionary. *Empathy*. https://dictionary.cambridge.org/dictionary/english/empathy. Accessed: 04-12-2023. 2023.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell,

M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[10] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL].

[11] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. 2017, pp. 35–45. URL: https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

[12] D. Yu, K. Sun, C. Cardie, and D. Yu. "Dialogue-Based Relation Extraction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. URL: https://arxiv.org/abs/2004.08056v1.

[13] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. "Survey on evaluation methods for dialogue systems". In: *Artificial Intelligence Review* 54.1 (June 2020), pp. 755–810. DOI: 10.1007/s10462-020-09866-x. URL: https://doi.org/10.1007%2Fs10462-020-09866-x.

[14] J. Weizenbaum. "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine (Reprint)". In: *Commun. ACM* 26.1 (1983), pp. 23–28. DOI: 10.1145/357980.357991. URL: https://doi.org/10.1145/357980.357991.

[15] M. K. Bergman. *A Common Sense View of Knowledge Graphs*. Adaptive Information, Adaptive Innovation, Adaptive Infrastructure Blog. 2019. URL: http://www.mkbergman.com/2244/a-common-sense-view-of-knowledge-graphs/.

[16] W. Fan, X. Wang, and Y. Wu. "Diversified Top-k Graph Pattern Matching". In: *Proc. VLDB Endow.* 6.13 (2013), pp. 1510–1521. DOI: 10.14778/2536258.2536263. URL: http://www.vldb.org/pvldb/vol6/p1510-fan.pdf.

[17] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti. "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web (Dagstuhl Seminar 18371)". In: *Dagstuhl Reports* 8.9 (2018), pp. 29–111. DOI: 10.4230/DAGREP.8.9.29. URL: https://doi.org/10.4230/DagRep.8.9.29.

[18] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A. N. Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, and A. Zimmermann. "Knowledge Graphs". In: *CoRR* abs/2003.02320 (2020). arXiv: 2003.02320. URL: https://arxiv.org/abs/2003.02320.

[19] D. Jurafsky and J. H. Martin. *Speech and language processing*. 2. ed., [Pearson International Edition]. Prentice Hall series in artificial intelligence. London [u.a.]: Prentice Hall, Pearson Education International, 2009, 1024 S. ISBN: 0-13-504196-1, 978-0-13-504196-3. URL: http://aleph.bib.uni-mannheim.de/F/?func=find-b&request=285413791&find_code=020&adjacent=N&local_base=MAN01PUBLIC&x=0&y=0.

[20] spaCy Developers. *Named Entity Recognition Visualizer*. Accessed: [Today's Date]. 2023. URL: https://spacy.io/usage/visualizers#ent.

[21] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*. Ed. by J. T. Wang. ACM, 2008, pp. 1247–1250. DOI: 10.1145/1376616.1376746. URL: https://doi.org/10.1145/1376616.1376746.

[22] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. "DBpedia: A Nucleus for a Web of Open Data". In: *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*. Ed. by K. Aberer, K. Choi, N. F. Noy, D. Allemang, K. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux. Vol. 4825. Lecture Notes in Computer Science. Springer, 2007, pp. 722–735. DOI: 10.1007/978-3-540-76298-0_52. URL: https://doi.org/10.1007/978-3-540-76298-0%5C_52.

[23] D. Vrandecic. "Wikidata: a new platform for collaborative data collection". In: *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*. Ed. by A. Mille, F. Gandon, J. Misselis, M. Rabinovich, and S. Staab. ACM, 2012, pp. 1063–1064. DOI: 10.1145/2187980.2188242. URL: https://doi.org/10.1145/2187980.2188242.

[24] T. Nayak, N. Majumder, P. Goyal, and S. Poria. "Deep Neural Approaches to Relation Triplets Extraction: A Comprehensive Survey". In: *CoRR* abs/2103.16929 (2021). arXiv: 2103.16929. URL: https://arxiv.org/abs/2103.16929.

[25] P.-L. Huguet Cabot and R. Navigli. "REBEL: Relation Extraction By End-to-end Language generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational

Linguistics, Nov. 2021, pp. 2370–2381. DOI: 10.18653/v1/2021.findings-emnlp.204. URL: https://aclanthology.org/2021.findings-emnlp.204.

[26] X. Liu, J. Zhang, H. Zhang, F. Xue, and Y. You. "Hierarchical Dialogue Understanding with Special Tokens and Turn-level Attention". In: *arXiv preprint arXiv:2305.00262* (2023).

[27] H. Face. *The Hugging Face Course*. https://huggingface.co/course. Online; accessed September, 6, 2023. 2022.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805 [cs.CL].

[30] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11. URL: https://doi.org/10.1109/ICCV.2015.11.

[31] G. Wenzek, M. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data". In: *CoRR* abs/1911.00359 (2019). arXiv: 1911.00359. URL: http://arxiv.org/abs/1911.00359.

[32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. "LLaMA: Open and Efficient Foundation Language Models". In: *CoRR* abs/2302.13971 (2023). DOI: 10.48550/ARXIV.2302.13971. arXiv: 2302.13971. URL: https://doi.org/10.48550/arXiv.2302.13971.

[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685 (2021). arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.

[35] git-cloner. *Fine-tuning vicuna-7b on a single 16G GPU*. https://github.com/git-cloner/llama-lora-fine-tuning/tree/main. Accessed: 2023-08-10. 2023.

[36] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena". In: *CoRR* abs/2306.05685 (2023). DOI: 10.48550/ARXIV.2306.05685. arXiv: 2306.05685. URL: https://doi.org/10.48550/arXiv.2306.05685.

[37] OpenAI. *GPT-3.5-Turbo-0613*. https://openai.com/blog/function-calling-and-other-api-updates. Accessed: 2023-08-19. 2023.

[38] OpenAI. *Model index for researchers*. https://platform.openai.com/docs/model-index-for-researchers. Accessed: 2023-09-10. 2023.

[39] Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin. *Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation*. 2016. arXiv: 1601.03651 [cs.CL].

[40] R. Cai, X. Zhang, and H. Wang. "Bidirectional Recurrent Convolutional Neural Network for Relation Classification". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 756–765. DOI: 10.18653/v1/P16-1072. URL: https://aclanthology.org/P16-1072.

[41] Q. Jia, H. Huang, and K. Q. Zhu. *DDRel: A New Dataset for Interpersonal Relation Classification in Dyadic Dialogues*. 2020. arXiv: 2012.02553 [cs.CL].

[42] C. Wang, X. Liu, Z. Chen, H. Hong, J. Tang, and D. Song. "Zero-Shot Information Extraction as a Unified Text-to-Triple Translation". In: *CoRR* abs/2109.11171 (2021). arXiv: 2109.11171. URL: https://arxiv.org/abs/2109.11171.

[43] T. Kitwood and C. Müller-Hergl. *Demenz: der person-zentrierte Ansatz im Umgang mit verwirrten Menschen*. Huber, 2013. ISBN: 9783456853055. URL: https://books.google.de/books?id=MMH_ngEACAAJ.

[44] S. Fazio, D. Pace, J. Flinner, and B. Kallmyer. "The Fundamentals of Person-Centered Care for Individuals With Dementia". In: *The Gerontologist* 58.suppl$_1$ (Jan. 2018), S10–S19. ISSN: 0016-9013. DOI: 10.1093/geront/gnx122. eprint: https://academic.oup.com/gerontologist/article-pdf/58/suppl_1/S10/23563262/gnx122.pdf. URL: https://doi.org/10.1093/geront/gnx122.

[45] M. Meis, M. Krueger, P. Gablenz, I. Holube, M. Gebhard, M. Latzel, and R. Paluch. "Development and Application of an Annotation Procedure to Assess the Impact of Hearing Aid Amplification on Interpersonal Communication Behavior". In: *Trends Hear* 22 (Dec. 17, 2018). DOI: 10.1177/2331216518816201.

[46] Unknown. *Demenz Audit-FB1-Online*. Accessed: 2023-10-26. 2023. URL: https://www.dnqp.de/fileadmin/HSOS/Homepages/DNQP/Dateien/Expertenstandards/Demenz/Demenz_Audit-FB1-Online.pdf.

[47] B. A. Kitchenham, D. Budgen, and P. Brereton. *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC, 2015. ISBN: 1482228653.

[48] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020). DOI: 10.5281/zenodo.1212303.

[49] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *CoRR* abs/1603.02754 (2016). arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754.

[50] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL].

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683 (2019). arXiv: 1910.10683. URL: http://arxiv.org/abs/1910.10683.

[52] OpenAI. *Introducing ChatGPT and Whisper APIs*. https://openai.com/blog/introducing-chatgpt-and-whisper-apis. Accessed: 02/11/2023. 2023.

[53] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf. *Open LLM Leaderboard*. Accessed: 2023-11-01. 2023. URL: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

[54] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer. *Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning*. 2023. arXiv: 2302.02662 [cs.LG].

[55] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. *Finetuned Language Models Are Zero-Shot Learners*. 2021. arXiv: 2109.01652 [cs.CL].

[56] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2022. arXiv: 2201.11903 [cs.CL].

[57] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. *Large Language Models are Zero-Shot Reasoners*. 2022. arXiv: 2205.11916 [cs.CL].

[58] D. Roth and W.-t. Yih. "A Linear Programming Formulation for Global Inference in Natural Language Tasks". In: *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2004, pp. 1–8. URL: https://aclanthology.org/W04-2401.

[59] G. Stoica, E. A. Platanios, and B. Póczos. *Re-TACRED: Addressing Shortcomings of the TACRED Dataset*. 2021. arXiv: 2104.08398 [cs.CL].

[60] M. Verhagen, R. J. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. "SemEval-2007 Task 15: TempEval Temporal Relation Identification". In: *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*. Ed. by E. Agirre, L. M. i Villodre, and R. Wicentowski. The Association for Computer Linguistics, 2007, pp. 75–80. URL: https://aclanthology.org/S07-1014/.

[61] B. Li, Y. Hou, and W. Che. "Data augmentation approaches in natural language processing: A survey". In: *AI Open* 3 (2022), pp. 71–90. ISSN: 2666-6510. DOI: 10.1016/j.aiopen.2022.03.001. URL: http://dx.doi.org/10.1016/j.aiopen.2022.03.001.

[62] B. Taillé, V. Guigue, G. Scoutheeten, and P. Gallinari. "Let's Stop Incorrect Comparisons in End-to-end Relation Extraction!" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 3689–3701. DOI: 10.18653/v1/2020.emnlp-main.301. URL: https://aclanthology.org/2020.emnlp-main.301.

[63] A. Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017. ISBN: 978-1491962299.

[64] M.-L. Zhang and Z.-H. Zhou. "A Review on Multi-Label Learning Algorithms". In: *IEEE Transactions on Knowledge and Data Engineering* 26.8 (2014), pp. 1819–1837. DOI: 10.1109/TKDE.2013.39.

[65] M. Heydarian, T. E. Doyle, and R. Samavi. "MLCM: Multi-Label Confusion Matrix". In: *IEEE Access* 10 (2022), pp. 19083–19095. DOI: 10.1109/ACCESS.2022.3151048.

[66] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. *Zephyr: Direct Distillation of LM Alignment*. 2023. arXiv: 2310.16944 [cs.LG].

[67] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah. *Orca: Progressive Learning from Complex Explanation Traces of GPT-4*. 2023. arXiv: 2306.02707 [cs.CL].

[68] S. B. Loh and A. Sesagiri Raamkumar. "Harnessing Large Language Models' Empathetic Response Generation Capabilities for Online Mental Health Counselling Support". In: (Oct. 2023).

[69] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: 1811.00207 [cs.CL].