Bachelor thesis

Artificial Intelligence

Radboud University

Relation Extraction using Few-Shot Entailment on Conversational Data

Author:
Arne Steffen Wittgen
s1034858

First supervisor:
Faegheh Hasibi
Data Science Department
f.hasibi@cs.ru.nl

Second reader:
Serge Thill
Artificial Intelligence Department
serge.thill@donders.ru.nl

March 17, 2023

Abstract

Relation Extraction is an important task for personal Natural Language Understanding, and especially so in conversational applications where Knowledge Graphs, built from such relations, are essential for knowledge storage. Recent advances in Natural Language Understanding have shown that pre-trained Language Models tend to be the best at solving various Language Understanding tasks. Most of the time, they are fine-tuned on the specific downstream task, which relies on a large enough, high-quality dataset for the task. To overcome this issue and improve model efficiency, the novel approach of Few-Shot Learning is explored. Combined with the task of Natural Language Entailment, recent research has shown that models using Few-Shot Learning and Inference can outperform fully fine-tuned State-of-the-Art models on Relation Extraction tasks (and others). Since relations are so important for conversational applications, the question is in how far the approach of performing Relation Extraction with Entailment-based Few-Shot Learning can be applied to conversational domains. Therefore, this thesis investigates in how far this is applicable. The obtained results of the Few-Shot Entailment models tested do not reach state-of-the-art approaches on relatively comparable tasks. Still, one main conclusion is that entity type information is potentially an important factor for accurate relation extraction, which is recommended for further research.

Contents

Chapter 1

Introduction

When building a Conversational Agent to interact with a human, many design considerations concern information: What should the agent know about its conversation partner? How should the information be stored? How can it learn more about them? All these questions receive attention in their own right in research. This thesis will mainly focus on the third question: How an AI model can extract information from a conversation in an efficient manner. The main issue for such language understanding applications is the need for large, annotated datasets to fine-tune the model. Recent research has established an alternative method leveraging the information contained in pre-trained language models to avoid extensively fine-tuning them. The goal of this research is to test whether that approach also works on conversational data.

This section gives an overview of the field of Natural Language Understanding (NLU) and more concretely defines some relevant subfields. These are broadly introduced, while specific research accomplishments are discussed in section 2. Also, it discusses the state-of-the-art models, highlighting their shortcomings and presenting Few-Shot Learning as a novel approach. Then, the topic of NLU specifically in conversations is discussed, introducing the aim of conversational AI, the concept of Personal Knowledge Graphs and why the topics of Relation Extraction and Entailment are of high relevance for NLU in conversations. Lastly, the research question for this thesis is established based on these topics.

1.1 Natural Language Understanding in AI

1.1.1 The history and Aim of NLU

Natural Language Understanding (NLU) has been a field of research for a long time [1, 2], with one of its main aims being the creation of autonomous models capable of reproducing human-level NLU [2, 3]. There are many uses for such capabilities: when properly understanding natural language, AI can be used to automate complex tasks such as text classification or summarization, suggesting writing improvements that go beyond simple spelling and grammar checks, generate texts that seem increasingly human-like, or have proper conversations[2, 4]. Evidently, AI is still a long way off perfecting these abilities, which motivates the multitude of ongoing research in this field.

In this aspect, it is important to understand what NLU is, and what it is not. Formally, NLU is considered a subfield of Natural Language Processing (NLP): The main

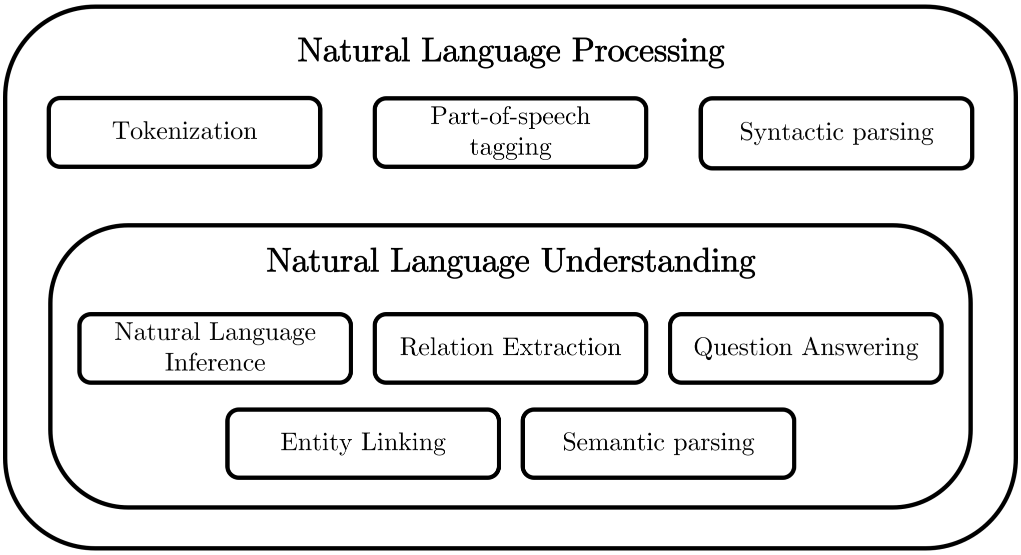

Figure 1.1: An overview of the topics included in the fields of Natural Language Processing and Natural Language Understanding. Examples are taken from [4, 3].

distinction made is that NLP is about processing text in any way, while NLU explicitly focuses on understanding language [1, 3]. To elaborate, NLP historically relied on rule- and statistic-based approaches [4, 5] and covers simpler, syntax-based tasks such as part-of-speech tagging or tokenization, which are only concerned with basic linguistic information [4, 5, 3]. In contrast, NLU aims at reproducing human understanding instead of simple analysis, being more concerned with semantic information and identifying an (potential) underlying meaning in language that is naturally ambiguous [3]. Overall, one could say that NLP generally contains any kind of text processing and analysis, while NLP focuses more on the complex processes emerging from such analysis.

Importantly, the distinction between these two fields is not often made in literature. The obvious reason is that NLU is a subfield of NLP meaning they often get grouped together, or the distinction is made more implicitly by distinguishing syntactic and semantic tasks (cf. [4]). Adding to this, the distinction between the fields has become less clear since both have progressed to employ Deep Learning (DL) approaches, which results in very similar (or even identical) models being used for both NLP and NLU tasks [4, 3].

1.1.2 Subfields of NLU

NLU itself contains a multitude of different topics; some examples are given in Figure 1.1. These topics range from broader goals (such as whole-text classification) to specific analysis techniques. Two of these are of particular interest for the research presented in this thesis: Natural Language Inference and Relation Extraction. In the following, the two topics are generally introduced while their relevance will be elaborated in the Conversation section of the introduction.

Natural Language Inference

Natural Language Inference (NLI) is a technique to confirm if information can be inferred from context: Given some text as the premise, it needs to be determined whether a hypothesis logically follows from it, i.e., if it is entailed [6]. Therefore, this is also known as

Table 1.1: Natural Language Inference examples for entailed, neutral and contradictory cases. Examples taken from [9].

He turned and saw Jon sleeping in his half-tent.	**Entailment**	He saw Jon was asleep.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	**Neutral**	All of the children love working in their gardens.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	**Contradition**	No one noticed and it wasn't funny at all.

Entailment Problem. These hypotheses are categorized as entailment, neutral, or contradiction depending on how their information content relates to the premise (cf. [7, 8]).

To illustrate this task, examples are given in Table 1.1: In the entailment example, it is obvious that the hypothesis "He saw Jon was asleep" is entailed since the fact is explicitly mentioned in the premise "He turned and saw Jon sleeping in his half-tent." The contradiction is similarly straight-forward: Since the premise states "someone else noticed it [...]" and "[...] it was really funny", the hypothesis "No one noticed and it wasn't funny at all" is obviously incorrect. In contrast, the neutral example highlights the complexity of this task: while the premise talks about the positive benefits the children have from gardening, it does not specify whether they actually like doing it or not – therefore, the hypothesis "All of the children love working in their gardens" is neither clearly entailed, nor contradictory to the premise.

NLI has received increased attention in the field of AI with the introduction of the PASCAL Recognizing Textual Entailment (RTE) Challenge in 2005 [6], and continued to establish its place among the core NLU tasks that contemporary research is concerned with [3, 4]. Until 2015, a great limitation of NLI research has been dataset size: while numerous datasets were available at high quality standards, their size of only a few hundred examples made it effectively impossible to use DL approaches due to the required amount of data [8, 4]. To overcome this roadblock, Bowman et al. [8] published the Stanford Natural Language Inference (SNLI) dataset with more than 500.000 data points, enabling DL for NLI and becoming the de-facto standard benchmark task for NLI models. Another dataset, MultiGenre Natural Language Inference (MNLI) by Williams et al. [10], builds upon SNLI, publishing NLI data for 10 different genres that can be used together with SNLI for an even larger corpus. Together, these datasets have been cited by over 6000 published papers[1], highlighting their importance for DL-based NLI approaches.

A relevant aspect of NLI is that many general NLP tasks can be reformulated as Entailment problem and thus solved as NLI task [11], which supports the significance of this area of research. In Figure 1.2, an example shows how the task of text classification can be rephrased as Entailment problem: for each output label (the possible document types), a natural language prompt is created. Together, these prompts represent the possible hypotheses that could be entailed by the premise (the text to be classified). Then, instead of having the model directly predict an abstract label (as would be done

[1] according to Google Scholar: 3404 citations for [8], 2756 for [10] (as of 14.02.2023).

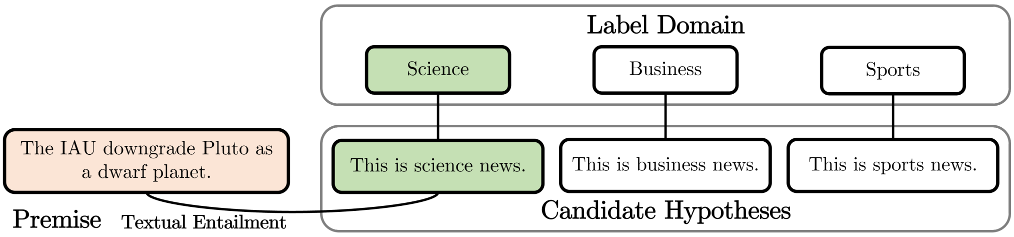

Figure 1.2: An example of how classification can be rephrased as entailment task. The labels are translated into natural language descriptions, and an NLI model can be used to select the most likely label (description) based on its entailment probability. Figure adapted from [11].

with typical classification), the entailment probability for each candidate hypothesis is compared and the one with the highest probability chosen. Lastly, the prompt can be easily mapped back to its formal label. This example demonstrates how an NLI model can be used to perform text classification, instead of training a dedicated model for the task.

Relation Extraction

Relation Extraction (RE) is a technique to determine how two entities (e.g., persons, objects, locations, etc) in a text are related. Generally, such relations can be anything: family relations, personal preferences, purely informational relations (such as locations or data points like age). The common denominator of these types is that they establish a link between two entities in a text. This technique can be helpful when the aim is to infer a relation from context that is not explicitly stated (cf. [12]).

The examples in Table 1.2 demonstrate the complexity that can be present in this task: determining the relation "(Irene Morgan Kirkaldy, city_of_birth, Baltimore)" in the first example is straight-forward as it is explicitly mentioned in text. In the second example, inferring that a person is adventurous and therefore extracting the relation "(I, misc_attribute, adventurous)" is reasonable if they mention a like for intense outdoor sports, but it requires some logical thinking to arrive at that conclusion. Lastly, the third example shows that there may be multiple, partially overlapping, relations in a single sentence, which is also something to consider.

Relation Extraction (RE) is considered a core task of NLU/NLP and is useful for building knowledge bases and graphs [4, 12, 13], the importance of which is discussed below in section 1.2.2. Like NLI, RE was historically performed using rule- and statistic-based methods, though kernel-based models became state-of-the-art for a longer period [14]. More recently, DL approaches have achieved top performance on commonly used benchmarks [14, 15].

Unlike NLI, there exists no dataset that dominates existing research as clearly as SNLI and MNLI. One reason for this is that not all RE research is concerned with fully identical tasks, as there are important differences such as sentence-level versus document-level RE [15], or research aims of integrating external information [16]. Naturally, these differences pose different requirements for the data used, which explains the existence of the numerous datasets available. Still, there are some datasets that are more commonly seen in contemporary research, such as SemEval 2010 Task 8, ACE 2004 and 2005, TACRED,

Table 1.2: Examples of Relation Extraction cases. The first example has a relation that stated explicitly in-text. The second example demonstrates a relation that needs to be inferred from context. The third example shows that one sentence can contain multiple examples. The examples are taken from [18, 19, 20], respectively.

Original Text	**Relations**
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	(Irene Morgan Kirkaldy, per:city_of_birth, Baltimore)
That's cool! I spent my weekends outdoors. You know, like surfing, hiking and rock-climbing.	(I, misc_attribute, adventurous)
My son. I bring him to church every Sunday with my Ford.	(I, has_children, son) (I, like_goto, church) (I, have_vehicle, ford)

and CoNLL [13, 15]. Of course, there exist several others that are also used in research, though not as frequently [17, 15].

1.1.3 State-of-the-art: large pre-trained LMs

Now, the question becomes how NLU is typically approached in the field. As shown by the examples of NLI and RE, rule- and statistics-based models were used for a longer time, though more recently DL approaches have taken over [4, 3, 17]. This development appears in the field of NLU in general, and the current state-of-the-art (SOTA) results are built upon large, pre-trained language models (LMs) [4, 3, 21]. Such models consume a huge amount of data [22], and the most performative ones are usually trained by industry-leading companies such as Google, Facebook, or OpenAI (cf. [23, 24, 11]) as these are the main institutions with access to the required amount of data and computational resources. The reason for such models defining the state-of-the-art is common among AI approaches: The more data a model can be trained on, the better the results it produces becomes.

Such SOTA models are all trained for text understanding but tend to be directed at specific tasks, such as question answering or information extraction [4, 3]. When using these models for even more specialized domains (e.g., for specific topics or applications in a certain field), they usually need to be fine-tuned using domain-specific data, which again requires availability of enough relevant data and enough computational resources [25].

1.1.4 A new direction: Few-Shot Learning

Overall, it becomes clear that being able to reach SOTA performance is constrained by access to enough resources, which begs the question if this dependency can be resolved or at least reduced. One novel approach for this is the concept of Few-Shot Learning (FSL). The underlying idea is to give a pre-trained LM a handful of examples of the task and domain it should perform, effectively priming it for the task at hand. This promises to greatly reduce the dependency on data and computational resources, as no extensive fine-tuning of the model is necessary – this becomes obvious in the fact that FSL usually keeps the model parameters fixed [24]. Research into this approach shows promising results, matching, or outperforming SOTA models on selected tasks (cf. [23, 24, 25]).

1.1.5 Combining Few-Shot Learning, Relation Extraction and Natural Language Inference

Recently, Sainz et al. [26] published a paper that applies the techniques of Few-Shot Learning and NLI to Relation extraction. In their work, they frame RE as an entailment task, and use Few-Shot Learning instead of fully re-training or fine-tuning the models. Testing their models on the widely used TACRED dataset [18], they produce results that match SOTA performance, with the SOTA models requiring 20 times more training data. Overall, this paper demonstrates that Few-Shot Learning is a realistic approach when it comes to Entailment-based RE on regular text.

1.2 NLU for Conversations

1.2.1 Definition and differences form classic NLU

Until now, NLU and its techniques have been discussed in general. Importantly, this thesis is specifically concerned with the more specific subtopic of conversations. This domain is distinctly different from regular text: conversations tend to contain more informal texts, are turn-based and therefore information flow is broken up compared to regular texts, and information contained tends to be more implicit [12, 27, 28, 29]. These differences present a challenge to regular NLU approaches, as techniques employed for regular texts are not guaranteed to translate into a conversational application [30, 31]. This means that approaches working well on regular texts cannot be assumed to work just as well in conversational settings.

1.2.2 The Aim of Conversational AI

Since the main topic of this work is conversations and conversational AI, it should be elaborated on what the aim of this field is, and what is necessary to reach it. The overarching goal of this domain is to develop conversational agents, which are AI models that can have a proper dialogue with humans [2, 29]. Possible applications for such agents are customer support, interactive FAQs or personalization. Of course, the development of such agents shares the goal of NLU in general, to improve language understanding. A markable difference is that agents may talk to the same person multiple times or are integrated into a customer database, so there needs to be a way to extract information from the conversations and store it adequately [12].

(Personal) Knowledge Bases

Such storage usually takes on the form of a knowledge base, which is commonly realized using Personal Knowledge Graphs (PKGs)[12]. These are based on triplet-relations, encoding a subject, an object, and the relation they have. Obviously, building a PKG based on these relations requires that a conversational agent can extract them from a conversation it has with a human. Therefore, the concept of RE is so relevant for conversational AI: it allows a model to extract the relevant knowledge triples from conversation, and use them to enrich the PKG [12].

Now, the question becomes what good approaches for RE for conversations are. It has been shown that simple rule-based approaches generally don't work well in conversations, due to their lack of generalization capabilities [32]. The current SOTA for RE in conversations mostly uses models that are either specialized for RE from the get-go, or re-trains large LMs (as shown in section 2.2). Their commonality is the dependence on enough training data and computational resources.

1.3 Research Approach

Firstly, it should be summarized what has been discussed so far. NLU is defined as a subfield of NLP, aiming to enable AI to better understand natural language by being aware of the information present. Current SOTA methods are resource-intense applying them to new tasks requires expensive re-training. To alleviate these problems, Few-Shot Learning has developed as novel research field to reduce the dependency on data. A pioneering approach combines the concepts of Few-Shot Learning and Entailment, exceeding SOTA results on RE tasks.

In conversational settings, the differences in language use pose a challenge when applying regular NLU techniques, and established techniques are not guaranteed to be applicable. Based on common application domains for conversational AI, RE is of high relevance due to its use for constructing PKGs. SOTA techniques for RE in conversations are defined by their dependency on enough training data.

Together, this begs the question if RE in conversations can be approached in a more efficient manner. Research on regular language establishes Few-Shot Entailment as promising solution, but due to the different language structure it is unclear whether it can be extended to conversations. Therefore, it is necessary to investigate its use in this different domain, which establishes the research question for this thesis:

How well does Relation Extraction using Few-Shot Entailment perform in conversational data?

This question will be investigated by compare pre-trained NLI LM models to SOTA results on a conversation RE task, and by examining how the low data scenarios impact model performance, as opposed to when the full dataset is available.

1.3.1 Contributions

This research aims to make contributions to the fields of NLU in Conversations and Data-efficient AI model training. It does so by examining whether the technique of Relation Extraction using FSL and NLI also works well in conversational settings, making the following concrete contributions:

- The RE approach by Sainz et al.[26] is examined on a conversation RE dataset, and the results are compared to SOTA approaches.
- The attribute extraction dataset by Wu et al.[20] is reformatted to be suitable for an RE task and made publicly available.

1.4 Outlook

In the following chapters, this thesis will cover the following content: In “Related Work”, the research approach is situated more clearly in the context of other research by highlighting specific developments and findings. In the “Methods” chapter, the experiment itself and the work leading up to it are explained, giving insight into the choices made. “Results” gives the outcomes of the experiment, analyzing them and comparing the findings with previous work. Lastly, the Conclusion and Discussion chapter elaborates on the theoretical implications of the findings, outlines the limitations of this experiment giving rise to possible further research, and situates them in the broader context of related topics.

Chapter 2

Related Work

2.1 Introduction

This section discusses the more specific research developments in the fields of Natural Language Inference, Relation Extraction, Few-Shot Learning, and Conversation Information Extraction with a focus on works that are closely related to this thesis. For each topic, a brief history of its development is given with the current State-of-the-Art (SOTA) approaches, while conversation-specific research is mentioned separately if relevant. Notably, the last section is a more elaborate discussion of available conversation datasets, as finding a suitable choice for this work has proven difficult.

2.2 Natural Language Inference/Entailment

As explained in the introduction, the two main NLI datasets seen in research are SNLI and MNLI, which have received much attention and seen many different research approaches. While going in-depth on the exact developments based on these datasets is out-of-scope for this thesis, for a long time Long Short-Term Memory (LSTM) model architectures have competed for the SOTA performance on NLI tasks (cf. [33, 34, 9]), though somewhat recently Transformer architectures have taken over and currently hold the top spots for SNLI and MNLI [35, 11]. Adding to this, a general literature survey by Otter et al.[4] establishes LSTM models as SOTA for NLP in general for a long time, though they have been recently overtaken by Transformers. Together, this this shows that such models seem most promising for NLI applications.

Looking at conversation data specifically, the first conversation-based NLI dataset was published by Welleck et al.[7], titled “DialogueNLI”; the aim for their research is to improve overall dialogue consistency. To build the dataset, they utilize the PersonaChat dataset[18] as basis for their work: PersonaChat was created by giving crowd-workers a persona description of a few sentences, which they subsequently used for short conversations with other workers. Welleck et al.[7] then annotate sentences from these conversations with relation triplets, and establish the inference class (Entailment, Neutral, or Contradiction) based on the (mis)match of the triplet data. The final dataset contains ca. 350.000 inference pairs, which is on the same scale as the SNLI [8] and MNLI[10] datasets, which have ca. 570.000 [8] and 433.000 [10] pairs, respectively. Unlike with SNLI and MNLI, not much research has been done with this dataset[1], and much of that

[1] According to Google Scholar: 169 citations for [7] (as of 14.02.2023).

research is focused or related to dialogue consistency as well (cf. [19]).

Overall, this work establishes Natural Language Inference as a core task for NLU which is actively researched, with two main State-of-the-Art datasets that see heavy use and help testing SOTA models. LSTM architectures were considered SOTA for a relatively long time, though recently Transformers have taken over the top spot in NLI benchmarks. Looking at conversations, there exists a reasonably large NLI dataset, though it sees much less use compared to the regular language datasets.

2.3 Relation Extraction

As mentioned in section 1.1.2, Relation Extraction (RE) is considered a core task of NLU/NLP and very useful for building knowledge bases and graphs. Initially, RE was performed using rule- and statistic-based methods, though kernel-based models became state-of-the-art for a longer period. More recently, Currently, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)/LSTMs are considered SOTA for RE tasks, though Transformers have started to receive more attention and tend to outperform CNNs and RNNs on selected benchmarks [36, 37, 38]. Overall, these models can be categorized into supervised, semi-supervised, distantly supervised, and unsupervised approaches [14, 38, 17]. Going in-depth on the precise differences of these approaches is out-of-scope, but there exist several surveys on RE covering this topic, such as [36, 39].

Revisiting the datasets mentioned in the introduction, the ones prevalent in contemporary research are SemEval 2010 Task 8, ACE 2004 and 2005, TACRED, and CoNLL [13, 15]. Of these, only the ACE datasets are suitable for document-level RE, while all other datasets contain sentence-level annotations only [15]. Therefore, much RE research naturally focuses on that type of RE.

Based on this, the two main limitations of current RE research become clear: reliance on data, and a focus on sentence-level analysis [39]. The second issue naturally derives from the nature of the datasets used, and therefore requires the collection of much new data, or at least the re-annotation of existing one. To elaborate on the first issue, it mainly comes down to the quality and size of the available datasets. Since human annotation of data is costly, most RE datasets can be split into two kinds: either, they are of high quality but comparably small and therefore unsuitable to train DL models (for example, ACE and SemEval), or they are large but (partially) annotated using automated techniques, which potentially introduces a lot of noise (for example, TACRED, or the New York Times dataset) [17, 15].

This naturally results in the current research avenues for RE: Using more data, training models more efficiently, and handling complex relations and domains (e.g., relations between three or more entities, or recognition of relations unseen at training)[40, 28]. This thesis focuses on the second direction, which currently sees a lot of progress in the form of Few-Shot learning.

2.4 Few-Shot Learning

As mentioned in the introduction, Few-Shot Learning (FSL) is a recent development in the field of Deep Learning, leveraging the information contained in pre-trained Language

Models (LMs) to fine-tune them more efficiently on specific tasks. According to Wang et al.[39] there are three typical reasons to use FSL: Imitating human learning, improving learning for rare cases by achieving better generalization, and reducing data-dependency for fine-tuning. Of these, the last is most important for this thesis. With respect to data efficiency, research on FSL shows that it can easily compete with, or even outperform, SOTA results from models that require drastically more data for fine-tuning on the task at hand (cf. [23, 24, 11]). Examples of this is for example the work of Chowdery et al.[23] who achieve new SOTA performance on several tasks using an FSL-tuned model. Similarly, Brown et al.[24] used the widely used GPT-3 model with FSL on several tasks, achieved results competing with regular fine-tuning approaches on several tasks.

FSL is usually performed by rephrasing the downstream task using a natural language prompt that the LM can predict using its language modelling capabilities [11]. Typically, this is done by using masked language modelling objectives, for example, cloze tasks [25, 24]. These approaches utilize large LMs such as GPT-3 (cf. [24]), and their results clearly match or outperform classic fine-tuning, thus establishing Transformer-based models as SOTA for FSL approaches.

Wang et al.[11] take an alternative approach to FSL: instead of using masked language modelling, they propose a task reformulation with entailment (as demonstrated in 1.2). They utilize a smaller LM, RoBERTa, and compare it to previous FSL and standard fine-tuning approaches on several benchmarks. The model is tested both pre-trained specifically on MNLI data and in its vanilla configuration. Their results clearly show that this approach outperforms prompt-based FSL SOTA results by 12%, and even if the model is not pre-trained with NLI data the entailment approach still competes with regular FSL. Together, these results show that entailment-based FSL is potentially more potent than prompt-based FSL, with the added benefit of using smaller LMs which improves efficiency.

The first to perform RE using entailment FSL were Obamuyide and Vlachos[32]. They used ESIM as entailment model and evaluated its performance on TACRED compared to SOTA fine-tuned models, with promising results for entailment FSL. Building on that work, Sainz et al.[26] use entailment-based FSL to perform RE on the TACRED dataset. Their main change compared to [32] is the use of pre-trained language models instead of ESIM, namely RoBERTa and DeBERTa. Using the model versions pre-trained on NLI data, they outperform fine-tuned SOTA models both in few-shot and full training scenarios, clearly showing the strength of Entailment-based RE using FSL.

Looking again more specifically at conversations, Madotto et al.[22] evaluate prompt-based FSL for dialogue systems. Aiming to improve both data efficiency and generalizability of models, they show that an FSL-fine-tuned model competes with regularly trained SOTA results on 9 different benchmark tasks. Notably, these benchmarks are less about specific technical tasks (like Relation Extraction) and more about conversation quality as a whole.

2.5 Conversation Information Extraction

The last important related topic is research into conversation data specifically. The two main points of focus in this section are the research approaches, and the available datasets.

Overall, Information Extraction (IE) in Conversations is a reasonably new area of research, with most papers having been published since 2019. As with other NLU tasks, the prevalent approach is to use some sort of DL model, where Transformers [31, 29, 22, 27], RNNs/LSTMs [29, 37], and CNNs [37, 41] achieve top performances. This shows that under the aspect of model selection regular text and conversation data are similar. Obviously, there is some overlap in the approaches when it comes to specific tasks such as RE, so this is no surprise.

In this aspect, the most similar work to the research in this paper is done by Wu et al.[20]. They perform user attribute extraction from conversations, which is closely related to the task of RE as the attributes are extracted as relation triple. Due to this similarity, their dataset is a good evaluation option for this work (as elaborated on in section 2.5.1). In their work, they compare a novel approach to established SOTA baselines for attribute extraction on a dataset they created based on DNLI [7]. This approach differs significantly from the entailment-based one used by Sainz et al. [26]: Instead of using pre-trained LMs, they build a custom two-stage attribute extractor. This model uses a context encoder, after which a memory network is used as classifier to predict which relations are present in a given sentence. Then, an entity generator determines which entities are related with the previously determined relation. In a way, this work takes an inverse approach to RE: instead of first determining the entities and then finding the most probably relation, the relation is inferred from context and then the associated entities are determined.

2.5.1 Conversation Datasets

When looking at the datasets used for conversation-based Information Extraction, a noticeable difference is that there are no as widely used conversation-based benchmark tasks for tasks such as NLI or RE. Therefore, many researchers resort to creating their own dataset for the task at hand, which results in a breath of datasets with different sources [37]. The dataset creation methods differ, though most are obtained either by using social media data or movie/series scripts, or by using crowdsourcing where humans have conversations following a specific task [37]. This results in over 100 different conversation datasets [37, 30], which can make it challenging to find the right one for a research project.

With regards to this experiment, there are three important factors for choosing a dataset: that it comes with proper relation annotations, that its domain is general enough to provide a suitably large number of available relations, and that the conversation origin is relevant enough. Of the datasets considered, several can be discarded since they do not fulfill either of those criterions: datasets like ConEL [30], ConEL-2 [31] or the original PersonaChat [18] do not have any annotations, or it is not possible to derive relations from them. Datasets such as FiRe [27] or MovieChat [12] only have a small number of relations and are thus not suitable. Lastly, sets like MovieChat ([12] and DDrel [42] have been excluded as their data is partially based on fictional movie scripts and therefore results obtained on them may not translate to real-world applications [27].

Ultimately, three datasets have been considered more closely for this work: DialogueNLI (DNLI)[7], DialogRE [28], and the Attribute Extraction dataset by Wu et al.[20], henceforth referred to as G2KY (”Getting to know you”, after the title of the paper). While DialogRE is desirable since it is already designed for RE specifically, it has not been chosen for three main reasons: The annotations are dialogue-level instead of sentence-level, which is less similar to existing RE research (cf. Relation Extraction), the entities are not necessarily existing in-text but can be, for example, “Speaker 1” which would make using natural language templates less natural, and it does not contain explicit instances of no relation (which is one of the strengths of the TACRED dataset [18]. Looking at the differences between DNLI and G2KY, the latter is the better choice since the task is very similar to RE (cf. section 2.4) and it contains no relation instances (in the form of sentences that are not annotated).

Chapter 3

Methods

This section describes the experimental approach to the research question stated in section 1. The general intuition behind the approach is discussed, followed by a theoretical formulation of the task to be solved. Then, the concrete experimental setup is described: first, a comparison of the TACRED dataset used by Saint et al. [26] and the G2KY dataset used here is made, followed by the data pre-processing procedure, and the evaluation scenarios. Second, the approach for template creation is described. Third, the models chosen for evaluation are discussed.

3.1 General Approach

The underlying approach for this research is to reproduce the work of Sainz et al.[26] for a conversation RE dataset. On a theoretical level, the advantage is that it establishes a more direct link between their results and ours, allowing for a level of qualitative comparability. On a practical level, this enables a straight-forward implementation of the experiments.

Since Sainz et al.[26] provide the necessary code base to run the experiments, the main addition made is finding a suitable conversation-based dataset and adapting it to work with the available framework. Importantly, this means that unless explicitly specified otherwise, the choices regarding the experimental setup have been based on the findings and choices from Sainz et al.[26].

3.1.1 Formulating Relation Extraction as Entailment task

The aim of this section is to give a mathematical formalization of the general approach that motivates this work: being able to reformulate RE as an NLI task. Welleck et al.[7] define NLI as follows: given a dataset D, each input of two sentences $(s1, s2)$ is associated with a class $y \in \{\text{entailment, neutral, contradiction}\}$. In NLI, each input s_j from the input space S usually is a natural language sentence. The input $(s1, s2)$ pair is described as the premise and hypothesis respectively, where the label is interpreted as the hypothesis being entailed, neutral to, or contradicting the premise. Thus, the computational problem comes down to learning a function $f_{NLI}(s1, s2) \rightarrow \{\text{entailment, neutral, contradiction}\}$ to predict the entailment type. Usually, NLI models can also predict an entailment probability, i.e., how likely it is that the hypothesis follows from the premise.

Sainz et al.[26] formalize RE using an NLI model as follows. To create natural language hypotheses for the relations, each relation r from the set of all relations R has one or multiple templates t_r written for it with placeholders for the entities, for example "{subj} have a degree in {obj}." Then, given two entities and a template, the hypothesis is created using a function $VERBALIZE(t, e1, e2)$ that replaces the placeholders with the concrete entities, for example "I have a degree in physics." This can be used to predict the most likely relation between two entities: assuming an input text x with two entities $e1$ and $e2$, their relation $\hat{r} \in R$ is predicted as $\hat{r} = argmax_{r \in R} P(x, e1, e2)$. Here, the function P_r corresponds to the entailment probability determined by the NLI model when applying a natural language template to the entities and relation: $P_r(x, e1, e2) = max_{t_r \in T} P_{NLI}(x, hyp)$, where $hyp = VERBALIZE(t, e1, e2)$.

Notably, Sainz et al.[26] employ another component to restrict the template choice based on entity type; since the G2KY dataset does not contain that information (cf. section 2.5.1), that component has been omitted.

To solve the problem of detecting no_relation cases, Sainz et al.[26] proposed two possible options: treating it as a relation with the verbalization \subj and obj are not related." or applying a threshold $\mathcal{T}$ to P_r where a positive relation is only predicted if it surpasses the threshold.

3.2 Experimental Setup

3.2.1 Comparing the G2KY and TACRED datasets

The evaluation dataset is G2KY [20]. G2KY and TACRED[18] are quite similar in terms of data structure, though some differences need to be taken into account. Firstly, TACRED[18] is explicitly designed to reflect the real-world distribution of relations present in text [18], while G2KY is based on a pre-existing conversation dataset and adds annotations if applicable, which results in a different data distribution. Connected to this, TACRED[18] contains explicit no_relation instances where the entities are present in-text. In comparison, G2KY only contains annotations for positive relations, and leaves sentences without one un-annotated, which is why these are treated as no_relation instances in the evaluation. Also, TACRED[18] contains exactly one relation per sentence, while G2KY annotates sentences with multiple relations if applicable (see for example the third sentence in Table 1.2). Content-wise, TACRED[18] contains only sentences in third person, while G2KY has both first- and third-person sentences.

The last main difference is that TACRED contains word type information for subject and object, which allows to constrain the template choice based on type. G2KY does not contain such information, so it is not possible to constrain the choice of templates based on that information. This is especially relevant since the relation set contains many similar relations, which are mainly distinct by the entity type that would be valid with them [19]. To illustrate why this is important, consider the following example: For the context "I enjoy watching golf on the weekend", the relations like_watching, like_sports, and have_vehicle - when referring to a VW golf car - are potential relations that make semantic sense. They would be verbalized as "I like watching golf.", "I like golf as a sport.", and "I have a golf.", respectively. If the type information is available, i.e., the model knows that golf refers to the sport, it can already discard like_watching (assuming that the relation requires a TV show or similar) and have_vehicle, before choosing the likeliest relation.

3.2.2 Data Pre-processing

The goal of the pre-processing is to ensure that all data points can be associated with a proper relation and thus properly templated. The G2KY dataset is available as a text file, which contains the full train, development, and test data from PersonaChat [43] including the persona descriptions, with annotations appended to the end of the corresponding lines. As mentioned, sentences without a relation are not explicitly annotated, which is done during pre-processing. The subject and object are replaced by dummy variables simply containing "dummy". The processing is performed as follows:

1. Sentences containing persona descriptions are removed
2. Each line is converted based on its annotations:
   - No annotation: annotate with ["dummy", "no_relation", "dummy"]
   - One annotation: convert into text and annotation
   - Multiple annotations: create one data point per annotation, consisting of the same text and one (different) annotation each

Then, the few-shot datasets are sampled. Sainz et al.[26] propose to perform stratified sampling to respect the original data distribution [26, p.4]. Instead, we argue that uniform sampling with the same amount of samples per positive relation is more realistic, because the goal is to evaluate model performance when only little (possibly hand-crafted) training data is available. Therefore, is likely that a few examples are collected for every relation, from which it may not be possible to deduce the original distribution. Therefore, in the few-shot scenarios each relation is sampled the same amount of times, with the exception on no relation which is sampled as often as there are positive relations to respect its disproportional occurrence in real-life texts (cf. [43]).

Table 3.1: Statistics regarding the different evaluation scenarios designed for G2Ky. For each scenario (original dataset without persona descriptions, full cleaned data, and Few-Shot scenarios) the total number of positive (relation) and neutral (no_relation) examples is given. Note that the full test set is used for evaluation on all scenarios, so the statistics do not change.

		Train		Development		Test	
Scenario		Positive	Neutral	Positive	Neutral	Positive	Neutral
Original		80344	66982	9789	7746	9084	7652
Cleaned		76111	66982	9250	7746	8559	7652
Few-shot	K=4	208	208	208	208	8559	7652
	K=16	944	944	826	826	8559	7652
	K=32	1888	1888	1595	1595	8559	7652

3.2.3 Data Scenarios

A total of 5 different scenarios are evaluated: Zero-Shot, 3 Few-Shot cases, and Full Training. Sainz et al.[26] also investigated whether it is better to detect no-relation cases using templates or a threshold and found strong evidence that threshold-based detection is superior. Therefore, this work omits this step and uses threshold-based detection. Table 3.1 summarizes the statistics for the different scenarios.

3.2.4 Zero-Shot

The Zero-Shot setting mainly serves as reference point to evaluate how well the Few-Shot scenarios improve performance. Here, the models do not receive any training or development data and the threshold for detecting no-relation cases is fixed to 0.5.

Few-Shot

The three Few-Shot cases evaluated contain 4, 16, and 32 examples per relation, except for no_relation which is sampled as often as the sum of all other relations. This results in samples of 0.27%, 1.24%, and 2.48% of the total available data. The development dataset is reduced in size accordingly, as is done by Sainz et al. [26].

Full Training

For Full Training, all available data is used. This scenario mainly serves as reference point to see how much the few-shot systems can improve when provided with more data.

3.2.5 Template creation

Overall, the template creation followed the same approach as used by Sainz et al.[26], i.e., the templates were not created using a fixed guide but rather based on the intuition of the researcher. Several relations in the G2KY dataset also exist in TACRED, for these the templates created by Sainz et al.[26] have been reused to allow for better comparability. The full list of templates if given in Appendix A.

One important difference in template creation when compared to TACRED is the existence of sentences in both first and third person. This implies that all relations need templates that can be expressed in first person for the subject “I”, as well as in third person, for example “my mother”. For templates taken over from TACRED, the existing templates are in third person, so a new template rephrasing the sentence in first person is added. For new templates the same sentence is used for first and third person, adjusted for proper grammar. To illustrate, for the relation have_pet, both the contexts `"I love my dog!"` as well as `"My mother got a cat three years ago"` need to be able to properly templated. Therefore, the relation has the templates `"subj have a obj."` and `"subj has a obj."`, which result in ”I have a dog.” and ”my mother has a cat.”, allowing for the right grammar.

Due to the high similarity of some of the relations in G2KY, the templates are very similar as well, for example with the relations `like_drink` and `like_food` (as well as other relations expressing a like for a certain concept): Their third person templates are `"{subj} likes to drink {obj}."` and `"{subj} likes to eat {obj}."` Such similarity means that for proper classification the LMs will be need to pick up on the fine-grained semantic differences for liking drinks versus liking foods, as compared to the differences between owning a car and liking drinks, for example.

3.2.6 Models

Following the setup by Sainz et al.[26], the model choice is constrained to those available on the Huggingface model hub [44]. Unfortunately, none of the best-performing models on either DialogRE [28], DNLI [7], or G2KY are available, so it was not possible to establish a direct comparison to SOTA results. Still, the results from Wu et al.[20] on G2KY for attribute extraction can be used for a qualitative comparison since the tasks are sufficiently similar in design.

Based on the findings by Sainz et al[26], the NLI RoBERTa and DeBERTa models are most suitable for this experiment. Due to computational constraints, it was only possible to use the RoBERTa model. It is both evaluated as-is (pre-trained on the MNLI dataset), and after fine-tuning using the full DNLI dataset[7]. The motivation behind fine-tuning on DNLI is to evaluate how model performance changes when the model gains knowledge directly related to the domain, as DNLI is a conversation dataset and derived from the same source as G2KY. Setup and technical details for the fine-tuning and experiments can be found in Appendix A B.

As the models are already trained for the NLI task at hand, they are not fine-tuned with the Few-Shot datasets and instead only receive the development data to adjust the threshold $\mathcal{T}$ for `no_relation` detection. While the methodology by Sainz et al.[26] does not clearly state how they evaluated the NLI models (i.e., by fine-tuning on the Few-Shot data or just providing the according development data), it is assumed that they did not perform additional fine-tuning, but only prompt-tuning via the templates. It is known that fine-tuning on small datasets results in model parameter instability and knowledge loss [45, 46, 47]; additionally, preliminary tests by fine-tuning the models using few-shot data result in results far worse than Zero-Shot performance.

Chapter 4

Results

To obtain the results, the models have been evaluated using the evaluation script provided by Sainz et al.[26]. The data and full code used for the experiments is made available (chapter 8). For each scenario, the models have been provided with the corresponding development set and were then evaluated on the test set. The metrics recorded are positive accuracy and F1-score, as these are evaluated by Sainz et al.[26] and Wu et al.[20] and thus allow for comparability.

Few-Shot vs. Zero-Shot. As shown in Table 4.1, both models improve their performance over the zero-shot scenario when provided with additional data. There is a clear jump from Zero-Shot to the first Few-Shot scenario (4 examples per relation): The DNLI-fine-tuned model achieves an F1-score of 10.69 as compared to 9.35, and the regular NLI model achieves 14.4 points F1 as compared to 12.95. When provided with additional data in the Few-Shot scenarios, performance does not noticeably increase: The DNLI model achieves an F1-score of 10.71 and 10.8 in the K=16 and K=32 scenarios respectively, while the regular model slightly decreases to 14.39 and 14.37 (K=16 and K=32, respectively).

Full development vs. Few-Shot. When provided with the full development data, both models improve marginally over their best Few-Shot performance. The DNLI model achieves an F1 of 10.83 in the full dev scenario compared to its 10.8 F1 in K=32, while the NLI model achieves an F1 of 14.42 compared to its 14.4 in K=4.

Effect of DNLI fine-tuning. As can be seen in Table 4.1, when the model is fine-tuned on the DNLI data, it performs worse than the regular RoBERTa model pre-trained on MNLI. In all scenarios, the F1-score is lower than the MNLI model.

Table 4.1: Experiment results for the MNLI pre-trained and DNLI fine-tuned model in the Zero-Shot, Few-Shot, and Full Development data scenarios. F1-score and accuracy are reported in percent.

	Scenario									
	Zero-Shot		K=4		K=16		K=32		Full Dev.	
Model	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
DNLI	10.61	9.35	9.19	10.69	9.24	10.71	10.12	10.8	9.48	10.83
NLI	10.61	12.95	10.49	14.40	10.06	14.39	10.01	14.37	10.08	14.42

Table 4.2: Qualitative comparison of positive accuracy and F1-score on the G2KY dataset. The top four models and results are given as reported by [20] on the attribute extraction task, the bottom two models are the full dev. scenario results reported in this paper.

Model	Acc.	F1
Seq2Seq	7.36	21.57
PG	11.80	22.99
KVMN	25.37	27.32
Wu et al.	26.52	28.68
DNLI	12.63	10.83
NLI	13.56	14.42

Accuracy Score. It should be noted that the evaluation script by Sainz et al.[26] only calculates the positive accuracy (i.e., accuracy on all cases where there is a relation present), instead of the overall accuracy. Therefore, this score was calculated separately. When looking at the accuracy scores for both models, they do not seem to correlate with F1 score performance: For example, the NLI model achieves a lower accuracy in the full development scenario compared to the zero-shot case, even though the F1 score improves.

Comparison to SOTA results. As explained in Section 2.4, their approach is fundamentally different as they first predict the relation from context (with a fine-tuned model) and then determine the entities [20]. Also, their work is not not related to researching (pre-trained) LMs. Together, this means that a comparison of the results can only be made to an extend, as the approaches are not directly comparable.

As shown in Table 4.2, when compared to the results obtained by Wu et al.[20], the two models tested in this work perform much worse than the ones by Wu et al.[20] in terms of F1 score, and still are 6 points behind the Seq2Seq model, which performs worst among the results by Wu et al.[20]. While the DNLI and NLI model are able to at least outperform Seq2Seq and PG in terms of reported accuracy, they still lack behind in overall performance.

Chapter 5

Analysis

This chapter discusses potential reasons as to why the results in this work fall short of SOTA results, as well as conducting a confusion analysis on the two LMs used in this work to better understand why the performance was sub-par.

5.1 Differences between our approach and SOTA models

When looking at the comparison of our results to SOTA performance, the main question becomes why the gap in performance is so significant, especially when considering that Sainz et al.[26] succeeded in outperforming SOTA models with the same approach on a different task. There are two probable reasons for this: The conversational domain is too different from the knowledge the LMs possess, and the approach by Wu et al.[20] is more specialized.

With regards to the domain, the G2KY dataset contains several relations that have a high semantic similarity [19], which may require specialized knowledge (in form of many training examples) to properly distinguish. Also, the evaluation dataset is not human-annotated, but derived from a different conversational dataset and annotations are created with distant supervision [20]. It should be noted that that aspect should not limit a relative comparison of models on the same dataset.

Looking at the approach by Wu et al.[20], it differs as a SOTA baseline from the ones used by Sainz et al.[26] in so far as that it doesn't utilize LMs as well, but rather a specialized, custom design for the task. Also, their model is fully fine-tuned from scratch, using the same dataset as used for the evaluation (obviously split in disjoint training and evaluation sets). This means that if the distant supervision approach consistently annotates certain relations incorrectly, or uses irrelevant semantic queues for certain relations, a fine-tuned model can theoretically pick up on these during training, as the same dataset is used. Since this work investigates in how far LMs can leverage their innate knowledge gained from pre-training on a large general language corpus, rather than the ability to fine-tune on a specific dataset, it is possible that this knowledge is simply not relevant enough to be applied to the evaluation dataset used, seeing how it has clear quality limitations and a very different domain.

5.2 Confusion Analysis

To better understand why the model performance is so low, it is useful to investigate how and which relations they tend to predict given the true label. For this, we conduct a confusion analysis utilizing the confusion matrix of the predicted versus true relation. Figures 5.1 and 5.2 give the confusion matrices[1] for the MNLI pre-trained and DNLI fine-tuned model when given the full development set, as these were the best-performing models (by a small margin).

For both models, the relatively weak but visible diagonal confirms that they predict relations with some consistency, but still make a lot of mistakes. Both models clearly struggle with correctly identifying relation cases versus no_relation cases and tend to over-predict no_relation, as can be seen by the strong first column, representing no_relation predictions. This observation also coincides with the results by Sainz et al.[26], who found that the Entailment-based model performance lacks when it comes to distinguishing a present versus no relation.

When looking at which classes the models predict (in)correctly, there are no classes the models seem to misclassify consistently, evident by the fact that no rows (true relations) are clearly visible in the matrices. Interestingly, both models seem to have a tendency to over-predict specific relations: the DNLI model tends to predict the relations `like_activity`, `job_status`, `favorite_activity`, `has_hobby`, `want`, `have_family`, and `favorite` (in order of appearance) significantly more often than any of the other relations. The MNLI model exhibits similar behavior, though with partially different relations: `have`, `like_activity`, `like_animal`, `misc_attribute`, and `favorite`. It is noteworthy that `misc_attribute` is predicted far more often and more consistently across all true relations than the other relations it tends to predict; the DNLI model does not have such a strong stand-out relation (albeit `job_status` seems to be predicted the most, though not with as much disparity as `misc_attribute` in the MNLI model).

Now, the obvious question becomes why the models tend to over-predict these relations. While we were not able to perform follow-up experiments to investigate this phenomenon, we can offer some plausible hypotheses for the reasons. Firstly, to look at the differences between the over-predicted relations between the two models, it seems likely that the DNLI training data causes the fine-tuned model to overfit on certain relations, or at least biases it towards relations that are over-represented in the training data. Fine-tuning pre-trained models is a known problem (cf. [48]), so it is not surprising that this approach did not work out without extensive revisions.

Looking at the reason for why the models over-predict relations in general, we hypothesize that this is due to the combination of their general language knowledge and the specific relations present in the dataset used: especially the MNLI model tends to predict general relations such as `like_activity, misc_attribute`, and `favorite`. Given that the dataset has many relations that are highly similar or potentially more precise versions of another one (e.g., `favorite_book` or `favorite_movie` could also be classified as `favorite` in a semantic sense), it may be that the knowledge in the model is not spe-

[1] Note that the matrix color scheme has been adjusted to make the model predictions more visible. For transparency, matrices with a color scheme scaling linearly with frequency of predictions are given in appendix C

cific enough to distinguish these relations. In this line, the circumstance that it tends to predict more general relations over more specific ones would be quite reasonable.

This issue also extends to one of the main differences in the G2KY and TACRED[18] datasets (cf. 2.5.1): Entity type annotations. While the TACRED[18] relations do not have as much overlap to begin with, the calculation of the most likely relation is supported by the entity type information that allows the model to automatically discard any relation that does not fit the entity types. Unlike, G2KY does not have such information: This means that relations which will have a logical meaning when verbalized but use different entity types will be competing. Obviously, this could be avoided if the entity type is known and thus certain relations can be excluded like with the TACRED dataset[18]. Since Wu et al.[20] are able to leverage this entity information due to their integrated entity extractor, it may additionally explain why the models tested in this experiment perform worse compared to others used on G2KY.

Altogether, it can be seen that the models struggle both from the already known problem of properly identifying relation vs. no_relation cases, as well as over-predicting certain relations which may be caused or influenced by the knowledge the models have, being biased by the additional training data, and/or not being able to utilize entity type information.

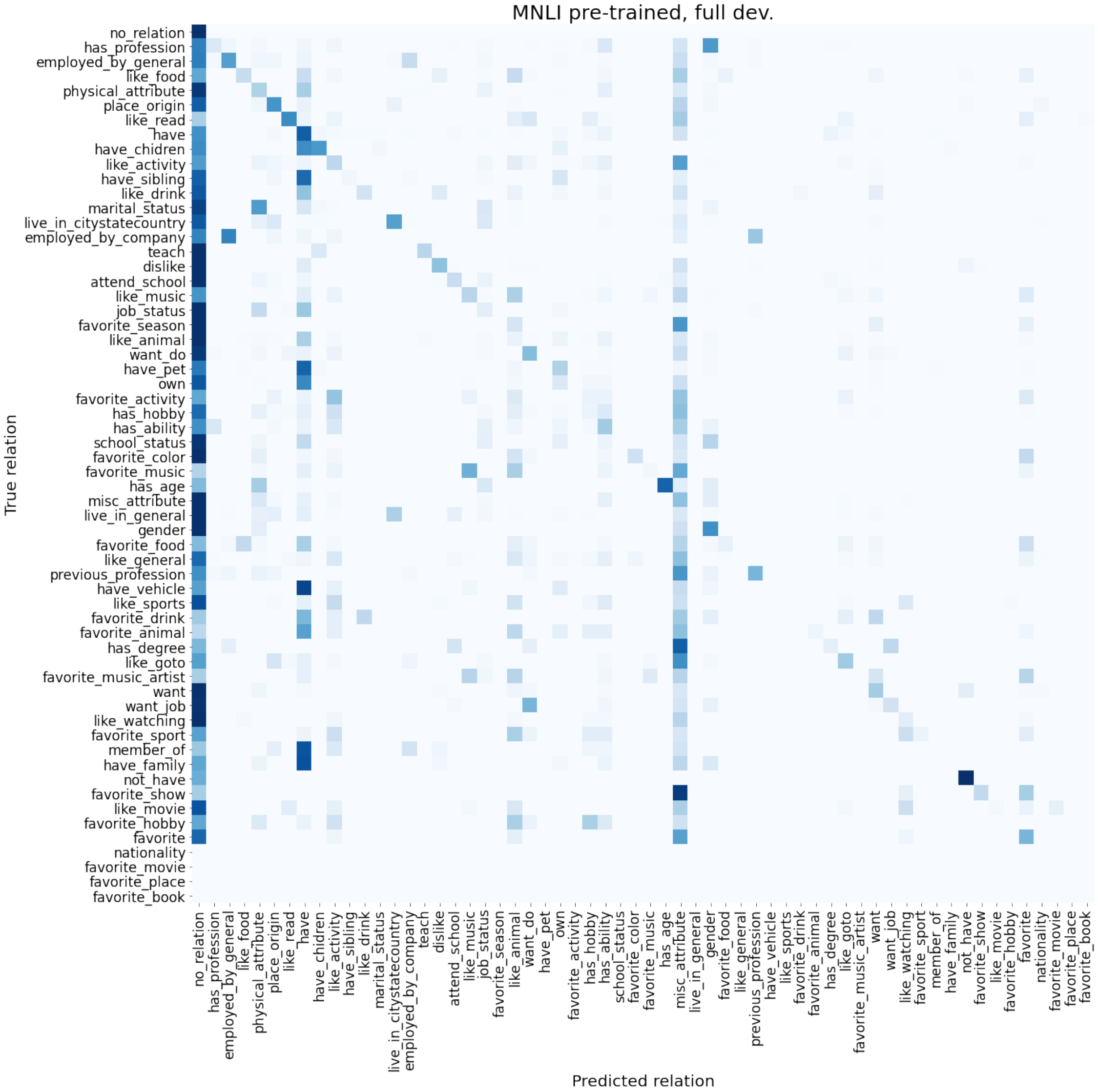

Figure 5.1: Confusion Matrix for the MNLI pre-trained model in the full dev. scenario. True relations are given in the rows, predictions in the columns. The matrix is row-wise normalized.

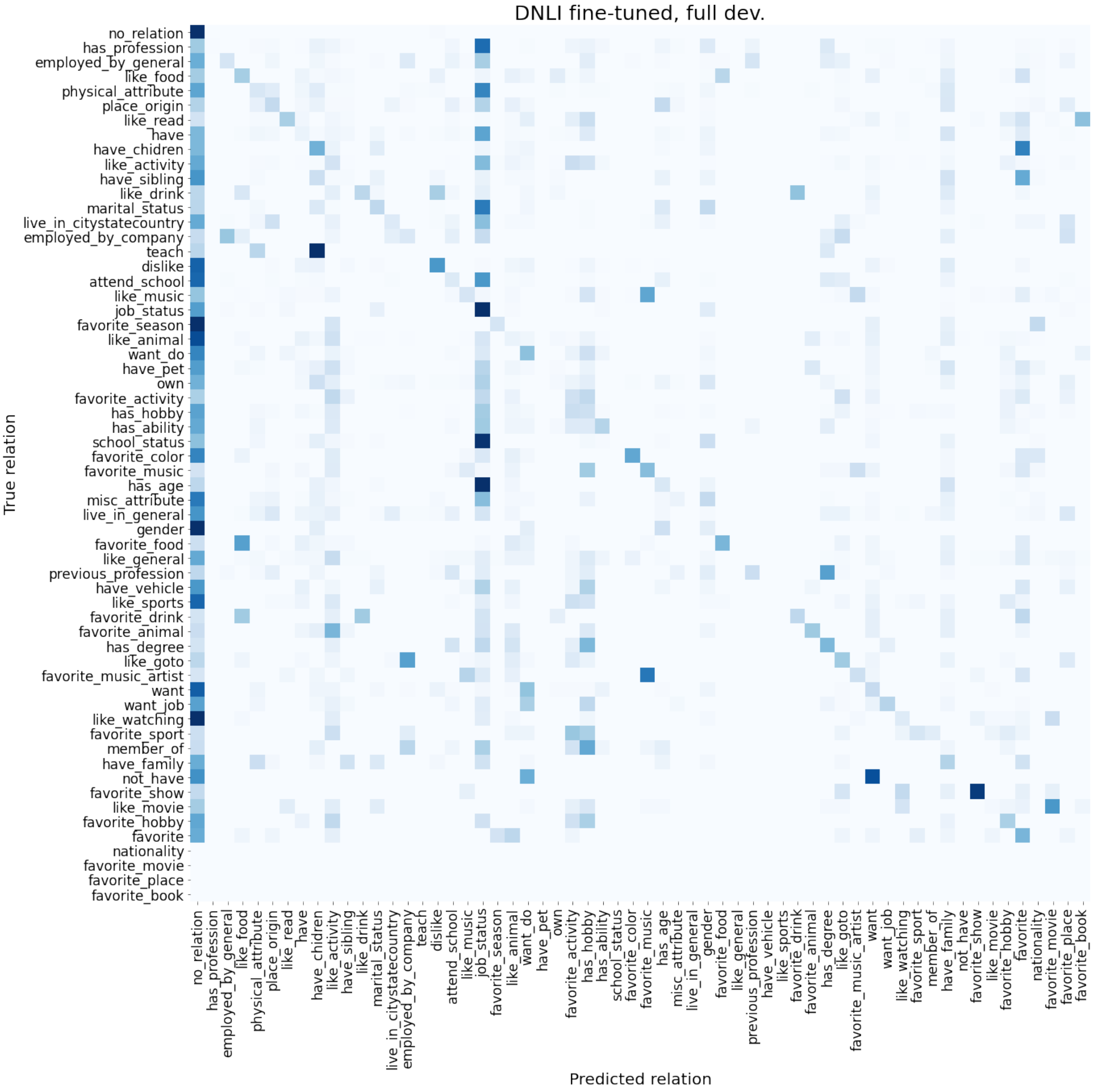

Figure 5.2: Confusion Matrix for the DNLI pre-trained model in the full dev. scenario. True relations are given in the rows, predictions in the columns. The matrix is row-wise normalized.

Chapter 6

Discussion

6.1 Conclusion

This work attempted to replicate the work of Sainz et al.[26] on Relation Extraction using Few-Shot Entailment on conversational data instead of regular text. To achieve this, the G2KY dataset was used to create a Relation Extraction test set, on which the MNLI pre-trained and DNLI fine-tuned RoBERTa model was evaluated. The results show mediocre performance clearly behind other models evaluated on the same dataset (albeit on a different but relatively comparable task), with the fine-tuned model achieving worse results on the test set even though it was hypothesized that it may perform better due to gaining knowledge on the conversation domain. The main cause for the performance seems to be the models' tendency to over-predict certain relations, which may possibly be caused by the missing entity type information in the dataset making predicting the right relation more difficult. Also, the overall setup was not optimal (cf. section 6.2): the conversation-based NLI dataset used for fine-tuning was not as good (in terms of quality and size) as other NLI datasets such as SNLI[8] and MNLI[10], and the computational resources limited model choice and fine-tuning performance.

To explicitly answer the research question, it is not clear how effective Relation Extraction using Few-Shot Entailment in conversational data can be based on the obtained results. The model performance was not convincing and clearly lacks behind comparable alternative approaches. Still, the analysis was able to uncover several potential causes for these observations. Further research, especially into entity type information, may be able to give more conclusive results and a clearer answer to the stated question.

6.2 Limitations and Future Work

There are two main limitations regarding this work: the dataset used, and the models chosen. Regarding the first issue, it was very difficult to find a suitable conversation-based relation extraction dataset, because of which the G2KY dataset was adapted to be usable as RE dataset. Presumably, this results in a lower data quality than a dataset that is dedicated for RE, like TACRED or DialogRE. Additionally, the dataset is created using distant supervision, which likely introduces a lot of noise. Lastly, there are no direct SOTA results to compare to as previous work on the same dataset was performed on the similar, but not identical, task of attribute extraction. Since the models used for work on G2KY or DialogRE are not available on the huggingface model hub, it was not possible to evaluate these models using the experiment setup in this work.

Regarding the model choice, ideally it would have been possible to use a better-performing DeBERTa architecture for this work. Unfortunately, due to the model size it was incompatible with the available computational resources, so the experiments used RoBERTa instead. In addition the resources limited training parameter choice (especially with regards to batch size) as well, which may have had an impact on the performance. Together, these reasons limit the conclusions drawn from the experiment in terms of their generalizability and applicability.

Looking at possible future work, there are three directions suggested: Firstly, to further investigate the findings in this work reproducing it using the larger DeBERTa model would be a simple approach to further research the usability of Few-Shot Entailment for RE in conversations. Adding to this, extending the work done here by annotating the entities with types (or using a Few-Shot model to extract the entities and types directly) may give insights into how far the missing entity type information influenced the results.

Secondly, it would be highly beneficial to NLU research on conversations to have a large-scale (comparable to TACRED), sentence-level annotated Relation Extraction dataset based on conversations. While this would be a challenging project, any attempt in this direction has the potential to benefit further research on this or related topics.

Thirdly, and a more tangential suggestion, there does not exist a general and extensive survey of conversation-based datasets. While Ni et al.[37] have created a comprehensive overview of the available datasets, they only touch on their properties superficially. Considering the diversity of available datasets, such a survey would be helpful for further research to choose an appropriate dataset more easily, or to determine if a new dataset is needed.

6.3 Summary

Altogether, this work aimed at researching the applicability of Relation Extraction using Few-Shot Entailment in conversational data. This was done by following the work of Sainz et al.[26] on a conversational dataset. The results were inconclusive, with performance lower than relatively comparable tasks. Overall, the conclusions that can be drawn from this research are limited due to the small scale of the experiment and the dataset used. Suggestions for future work include the creation of a more suitable conversation RE dataset, reproducing the experiment in this work with larger models or investigating entity type information, and creating a comprehensive survey of existing conversation datasets.

Chapter 7

Acknowledgments

First and foremost, I want to thank my supervisor Faegheh Hasibi for the continued support, theoretical and practical explanations and encouragement to finish this demanding thesis project. Sincere thanks also go out to Frank Leonè, the Bachelor Thesis coordinator, for interesting lectures and full understanding when it came to delays in the thesis process. I also want to thank Kim Bladder and Elèna Fromet for help during the thesis workgroups. Lastly, Helen Khorrami and my parents deserve a big thank you for supporting me fully during the whole semester (and then some extension time) it took to write this thesis, without their encouragement this project would have been much more difficult.

Chapter 8

Data and source code

The code used for this project is mainly based on the A2T Library by Sainz et al.[26]. A clone with the additional files created for this project, in addition with the converted G2KY dataset can be found in the Thesis Repository[1].

The code changes to the original library are explained in detail on the repository page. To summarize, changes were made to accomodate the different physical set-up and additional definitions were made to allow code use with a novel dataset. The G2KY dataset as used in this work is contained in the repository as well.

[1] bare-text link: https://github.com/ArneWittgen/Thesis_FS_Ent_DNLI

Bibliography

[1] A. Ram and K. Moorman, eds., *Understanding language understanding: computational models of reading.* Language, speech, and communication, Cambridge, Mass: MIT Press, 1999.

[2] J. Allen, *Natural language understanding.* Redwood City, Calif: Benjamin/Cummings Pub. Co, 2nd ed ed., 1995.

[3] R. Navigli, "Natural Language Understanding: Instructions for (Present and Future) Use," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, (Stockholm, Sweden), pp. 5697–5702, International Joint Conferences on Artificial Intelligence Organization, July 2018.

[4] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 604–624, Feb. 2021.

[5] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical natural language processing: a comprehensive guide to building real-world NLP systems.* Sebastopol, CA: O'Reilly Media, first edition ed., 2020. OCLC: 1164378473.

[6] B. MacCartney, *Natural language inference.* Stanford University, 2009.

[7] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue Natural Language Inference," Jan. 2019. arXiv:1811.00671 [cs].

[8] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," Aug. 2015. arXiv:1508.05326 [cs].

[9] N. Nangia, A. Williams, A. Lazaridou, and S. R. Bowman, "The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations," July 2017. arXiv:1707.08172 [cs].

[10] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," Feb. 2018. arXiv:1704.05426 [cs].

[11] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as Few-Shot Learner," Apr. 2021. arXiv:2104.14690 [cs].

[12] A. Tigunova, A. Yates, P. Mirza, and G. Weikum, "Listening between the Lines: Learning Personal Attributes from Conversations," in *The World Wide Web Conference*, (San Francisco CA USA), pp. 1818–1828, ACM, May 2019.

[13] H. Jiang, Q. Bao, Q. Cheng, D. Yang, L. Wang, and Y. Xiao, "Complex Relation Extraction: Challenges and Opportunities," Dec. 2020. arXiv:2012.04821 [cs].

[14] N. Bach and S. Badaskar, “A Review of Relation Extraction,” 2007.

[15] K. Liu, “A survey on neural relation extraction,” *Science China Technological Sciences*, vol. 63, pp. 1971–1989, Oct. 2020.

[16] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen, “KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction,” in *Proceedings of the ACM Web Conference 2022*, pp. 2778–2788, Apr. 2022. arXiv:2104.07650 [cs].

[17] M. Cui, L. Li, Z. Wang, and M. You, “A survey on relation extraction,” in *Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence: Second China Conference, CCKS 2017, Chengdu, China, August 26–29, 2017, Revised Selected Papers 2*, pp. 50–58, Springer, 2017.

[18] 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*, (Copenhagen, Denmark), pp. 35–45, Association for Computational Linguistics, 2017.

[19] Z. Wang, X. Zhou, R. Koncel-Kedziorski, A. Marin, and F. Xia, “Extracting and Inferring Personal Attributes from Dialogue,” Apr. 2022. arXiv:2109.12702 [cs].

[20] C.-S. Wu, A. Madotto, Z. Lin, P. Xu, and P. Fung, “Getting To Know You: User Attribute Extraction from Dialogues,” Aug. 2019. arXiv:1908.04621 [cs].

[21] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu, “Pre-trained models: Past, present and future,” *AI Open*, vol. 2, pp. 225–250, 2021.

[22] A. Madotto, Z. Lin, G. I. Winata, and P. Fung, “Few-Shot Bot: Prompt-Based Learning for Dialogue Systems,” Oct. 2021. arXiv:2110.08118 [cs].

[23] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “PaLM: Scaling Language Modeling with Pathways,” Oct. 2022. arXiv:2204.02311 [cs].

[24] 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,” July 2020. arXiv:2005.14165 [cs].

[25] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” July 2021. arXiv:2107.13586 [cs].

[26] O. Sainz, O. L. de Lacalle, G. Labaka, A. Barrena, and E. Agirre, “Label Verbalization and Entailment for Effective Zero- and Few-Shot Relation Extraction,” Sept. 2021. arXiv:2109.03659 [cs].

[27] A. Tigunova, P. Mirza, A. Yates, and G. Weikum, “PRIDE: Predicting Relationships in Conversations,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 4636–4650, Association for Computational Linguistics, 2021.

[28] D. Yu, K. Sun, C. Cardie, and D. Yu, “Dialogue-Based Relation Extraction,” Apr. 2020. arXiv:2004.08056 [cs].

[29] H. Liu, H. Peng, Z. Ou, J. Li, Y. Huang, and J. Feng, “Information Extraction and Human-Robot Dialogue towards Real-life Tasks: A Baseline Study with the MobileCS Dataset,” Oct. 2022. arXiv:2209.13464 [cs].

[30] H. Joko, F. Hasibi, K. Balog, and A. P. de Vries, “Conversational Entity Linking: Problem Definition and Datasets,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Virtual Event Canada), pp. 2390–2397, ACM, July 2021.

[31] H. Joko and F. Hasibi, “Personal Entity, Concept, and Named Entity Linking in Conversations,” June 2022. arXiv:2206.07836 [cs].

[32] A. Obamuyide and A. Vlachos, “Zero-shot Relation Classification as Textual Entailment,” in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, (Brussels, Belgium), pp. 72–78, Association for Computational Linguistics, 2018.

[33] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, “Annotation Artifacts in Natural Language Inference Data,” Apr. 2018. arXiv:1803.02324 [cs].

[34] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced LSTM for Natural Language Inference,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1657–1668, 2017. arXiv:1609.06038 [cs].

[35] 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,” July 2020. arXiv:1910.10683 [cs, stat].

[36] Z. Nasar, S. W. Jaffry, and M. K. Malik, “Named Entity Recognition and Relation Extraction: State-of-the-Art,” *ACM Computing Surveys*, vol. 54, pp. 1–39, Jan. 2022.

[37] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, “Recent advances in deep learning based dialogue systems: a systematic survey,” *Artificial Intelligence Review*, Aug. 2022.

[38] S. Pawar, G. K. Palshikar, and P. Bhattacharyya, “Relation Extraction : A Survey,” Dec. 2017. arXiv:1712.05191 [cs].

[39] H. Wang, G. Lu, J. Yin, and K. Qin, “Relation Extraction: A Brief Survey on Deep Neural Network Based Methods,” in *2021 The 4th International Conference on Software Engineering and Information Management*, (Yokohama Japan), pp. 220–228, ACM, Jan. 2021.

[40] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, M. Sun, and J. Zhou, “More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction,” Sept. 2020. arXiv:2004.03186 [cs].

[41] A. Tigunova, A. Yates, P. Mirza, and G. Weikum, “CHARM: Inferring Personal Attributes from Conversations,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 5391–5404, Association for Computational Linguistics, 2020.

[42] Q. Jia, H. Huang, and K. Q. Zhu, “DDRel: A New Dataset for Interpersonal Relation Classification in Dyadic Dialogues,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 13125–13133, May 2021.

[43] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing Dialogue Agents: I have a dog, do you have pets too?,” Sept. 2018. arXiv:1801.07243 [cs].

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, 2020.

[45] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” May 2018. arXiv:1801.06146 [cs, stat].

[46] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, “Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting,” Apr. 2020. arXiv:2004.12651 [cs].

[47] J. Wallat, J. Singh, and A. Anand, “BERTnesia: Investigating the capture and forgetting of knowledge in BERT,” Sept. 2021. arXiv:2106.02902 [cs].

[48] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, “Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning,” Apr. 2021. arXiv:2011.01403 [cs].

Appendix A

Verbalization Templates for G2KY

Table A.1 lists the templates used for the relation verbalization in the experiments in this work. For each relation the applicable templates are given.

Table A.1: Templates used for the relations present in the G2KY dataset. Note that the table extends over several pages.

Relation	Templates
no_relation	{subj} and {obj} are not related.
has_profession	{subj} work as {obj}.
	{subj} works as {obj}.
employed_by_general	{subj} am an employee of {obj}.
	{subj} is an employee of {obj}.
employed_by_company	{subj} am member of {obj}.
	{subj} is member of {obj}.
	{subj} am an employee of {obj}.
	{subj} is an employee of {obj}.
previous_profession	{subj} used to work as {obj}.
	{subj} used to have {obj} as a job.
member_of	{subj} am member of {obj}.
	{subj} is member of {obj}.
	{obj} joined {subj}.
teach	{subj} teach {obj}.
	{subj} teaches {obj}.
attend_school	{subj} studied in {obj}.
	{subj} graduated from {obj}.
has_degree	{subj} have a degree in {obj}.
	{subj} has a degree in {obj}.
	{subj} obtained a degree in {obj}.
school_status	{subj} am currently {obj}.
	{subj} is currently {obj}.
job_status	{subj} am currently {obj}.
	{subj} is currently {obj}.
place_origin	{subj} come from {obj}.
	{subj} comes from {obj}.

	{subj} was born in {obj}.
nationality	{subj} have a {obj} nationality.
	{subj} has a {obj} nationality.
	{obj} is the nationality of {subj}.
live_in_citystatecountry	{subj} live in {obj}.
	{subj} lives in {obj}.
	{subj} have a legal order to stay in {obj}.
	{subj} has a legal order to stay in {obj}.
live_in_general	{subj} live in {obj}.
	{subj} lives in {obj}.
	{subj} have a legal order to stay in {obj}.
	{subj} has a legal order to stay in {obj}.
physical_attribute	{subj} am {obj}.
	{subj} is {obj}.
gender	{subj} am a {obj}.
	{subj} is a {obj}.
	{subj}'s gender is {obj}.
have_chidren	{subj} am the parent of {obj}.
	{subj} is the parent of {obj}.
	{subj} am the mother of {obj}.
	{subj} is the mother of {obj}.
	{subj} am the father of {obj}.
	{subj} is the father of {obj}.
	{obj} is the son of {subj}.
	{obj} is the daughter of {subj}.
have_sibling	{subj} and {obj} are siblings.
	{subj} am brother of {obj}.
	{subj} is brother of {obj}.
	{subj} am sister of {obj}.
	{subj} is sister of {obj}.
have_family	{subj} and {obj} are family.
	{subj} am a brother in law of {obj}.
	{subj} is a brother in law of {obj}.
	{subj} am a sister in law of {obj}.
	{subj} is a sister in law of {obj}.
	{subj} am the cousin of {obj}.
	{subj} is the cousin of {obj}.
	{subj} am the uncle of {obj}.
	{subj} is the uncle of {obj}.
	{subj} am the aunt of {obj}.
	{subj} is the aunt of {obj}.
	{subj} am the grandparent of {obj}.
	{subj} is the grandparent of {obj}.
	{subj} am the grandmother of {obj}.
	{subj} is the grandmother of {obj}.

	{subj} am the grandfather of {obj}.
	{subj} is the grandfather of {obj}.
	{subj} am the grandson of {obj}.
	{subj} is the grandson of {obj}.
	{subj} am the granddaughter of {obj}.
	{subj} is the granddaughter of {obj}.
marital_status	{subj} am the spouse of {obj}.
	{subj} is the spouse of {obj}.
	{subj} am the wife of {obj}.
	{subj} is the wife of {obj}.
	{subj} am the husband of {obj}.
	{subj} is the husband of {obj}.
have	{subj} have {obj}.
	{subj} has {obj}.
	{subj} have a {obj}.
	{subj} has a {obj}.
have_pet	{subj} have a {obj}.
	{subj} has a {obj}.
has_hobby	{subj} have {obj} as a hobby.
	{subj} has {obj} as a hobby.
	{subj}’s hobby is {obj}.
has_ability	{subj} have the ability to {obj}.
	{subj} has the ability to{obj}.
	{subj} can {obj}.
	{subj} can do {obj}.
has_age	{subj} am {obj} years old.
	{subj} is {obj} years old.
have_vehicle	{subj} have a {obj}.
	{subj} has a {obj}.
not_have	{subj} don’t have {obj}.
	{subj} doesn’t have {obj}.
own	{subj} own {obj}.
	{subj} owns {obj}.
like_general	{subj} like {obj}.
	{subj} likes {obj}.
like_food	{subj} like to eat {obj}.
	{subj} likes to eat {obj}.
like_read	{subj} like to read {obj}.
	{subj} likes to read {obj}.
like_activity	{subj} like to do {obj}.
	{subj} like to {obj}.
	{subj} likes to do {obj}.
	{subj} likes to {obj}.
like_drink	{subj} like to drink {obj}.
	{subj} likes to drink {obj}.

like_music	{subj} like to listen to {obj}.
	{subj} likes to listen to {obj}.
like_animal	{subj} like {obj} as animals.
	{subj} likes {obj} as animals.
like_sports	{subj} like {obj} as a sport.
	{subj} likes {obj} as a sport.
like_goto	{subj} like to go to {obj}.
	{subj} likes to go to {obj}.
	{subj} like going to {obj}.
	{subj} likes going to {obj}.
like_movie	{subj} like the movie {obj}.
	{subj} likes the movie {obj}.
like_watching	{subj} like watching {obj}.
	{subj} likes watching {obj}.
dislike	{subj} don't like {obj}.
	{subj} doesn't like {obj}.
	{subj} dislike {obj}
	{subj} dislikes {obj}
favorite	{subj} have {obj} as a favorite.
	{subj} has {obj} as a favorite.
	{subj}'s favorite is {obj}.
favorite_season	{subj} have {obj} as favorite season.
	{subj} has {obj} as favorite season.
	{subj}'s favorite season is {obj}.
favorite_activity	{subj} have {obj} as favorite activity.
	{subj} has {obj} as favorite activity.
	{subj}'s favorite activity is {obj}.
favorite_color	{subj} have {obj} as favorite color.
	{subj} has {obj} as favorite color.
	{subj}'s favorite color is {obj}.
favorite_music	{subj} have {obj} as favorite music.
	{subj} has {obj} as favorite music.
	{subj}'s favorite music is {obj}.
favorite_food	{subj} have {obj} as favorite food.
	{subj} has {obj} as favorite food.
	{subj}'s favorite food is {obj}.
favorite_drink	{subj} have {obj} as favorite drink.
	{subj} has {obj} as favorite drink.
	{subj}'s favorite drink is {obj}.
favorite_animal	{subj} have {obj} as favorite animal.
	{subj} has {obj} as favorite animal.
	{subj}'s favorite animal is {obj}.
favorite_music_artist	{subj} have {obj} as favorite artist.
	{subj} has {obj} as favorite artist.
	{subj}'s favorite artist is {obj}.

favorite_sport	{subj} have {obj} as favorite sport.
	{subj} has {obj} as favorite sport.
	{subj}'s favorite sport is {obj}.
favorite_show	{subj} have {obj} as favorite show.
	{subj} has {obj} as favorite show.
	{subj}'s favorite show is {obj}.
favorite_hobby	{subj} have {obj} as favorite hobby.
	{subj} has {obj} as favorite hobby.
	{subj}'s favorite hobby is {obj}.
favorite_movie	{subj} have {obj} as favorite movie.
	{subj} has {obj} as favorite movie.
	{subj}'s favorite movie is {obj}.
favorite_place	{subj} have {obj} as favorite place.
	{subj} has {obj} as favorite place.
	{subj}'s favorite place is {obj}.
favorite_book	{subj} have {obj} as favorite book.
	{subj} has {obj} as favorite book.
	{subj}'s favorite book is {obj}.
want_do	{subj} want to do {obj}.
	{subj} wants to do {obj}.
	{subj} want to {obj}.
	{subj} wants to {obj}.
want	{subj} want {obj}.
	{subj} wants {obj}.
want_job	{subj} want {obj} as a job.
	{subj} wants {obj} as a job.
	{subj} want to work as {obj}.
	{subj} wants to work as {obj}.
misc_attribute	{subj} have {obj} as an attribute.
	{subj} has {obj} as an attribute.

Appendix B

Technical Experiment Details

The fine-tuning was performed on a cluster with 4 NVIDIA RTX 2080 GPUs. Fine-tuning took approximately 2 hours, inference testing between 30 minutes to 1.5h depending on development data size. The model was trained with a per-device batch size of 8 for 3 train epochs. The number of epochs is small as the model is already pre-trained on the NLI task. Following Sainz et al., the learning rate was chosen from {1e-6, 4e-6, 1e-5} based on its performance on the development set. The best learning rate proved to be 1e-6, with positive accuracy and F1-score for all learning rates as follows:

Learning Rate	Acc.	F1
1e-6	12.17	10.66
4e-6	9.85	9.01
1e-5	5.9	5.69

Table B.1: Experiment results for the different learning rates when evaluated on the G2KY development data.

Appendix C

Additional Confusion Matrices

Figures C.1 and C.2 (on the next two pages) give additional confusion matrices for the models evaluated in this work. The matrices analysed in section 5.2 are recolored to make the results more visible. Therefore, the matrices with a linear color scheme (i.e., linear to the frequency of predictions) are given here.

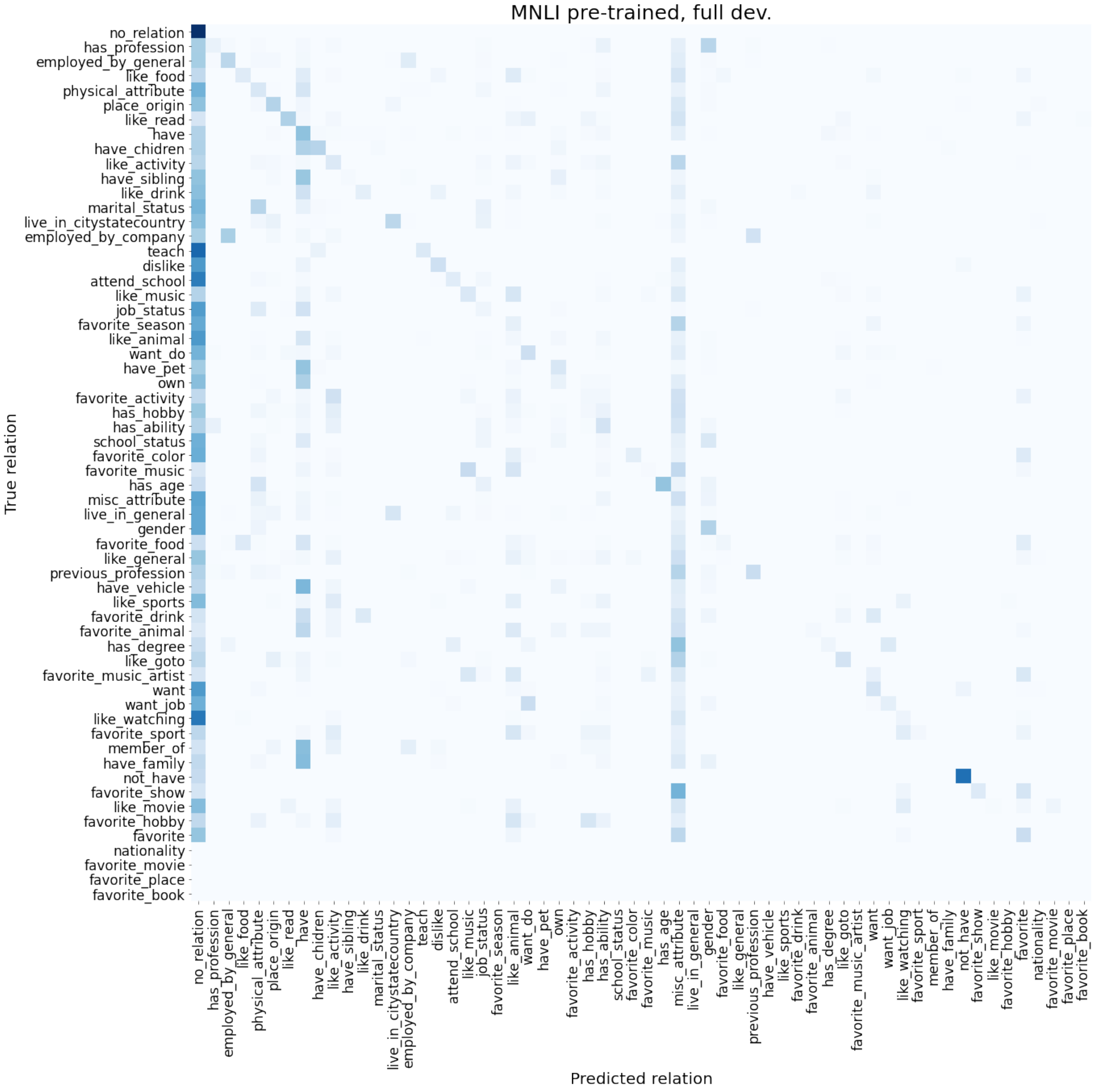

Figure C.1: Confusion Matrix for the MNLI pre-trained model in the full dev. scenario. True relations are given in the rows, predictions in the columns. The matrix is row-wise normalized. In this version, the color scheme is scaled linearly with the frequency of predictions.

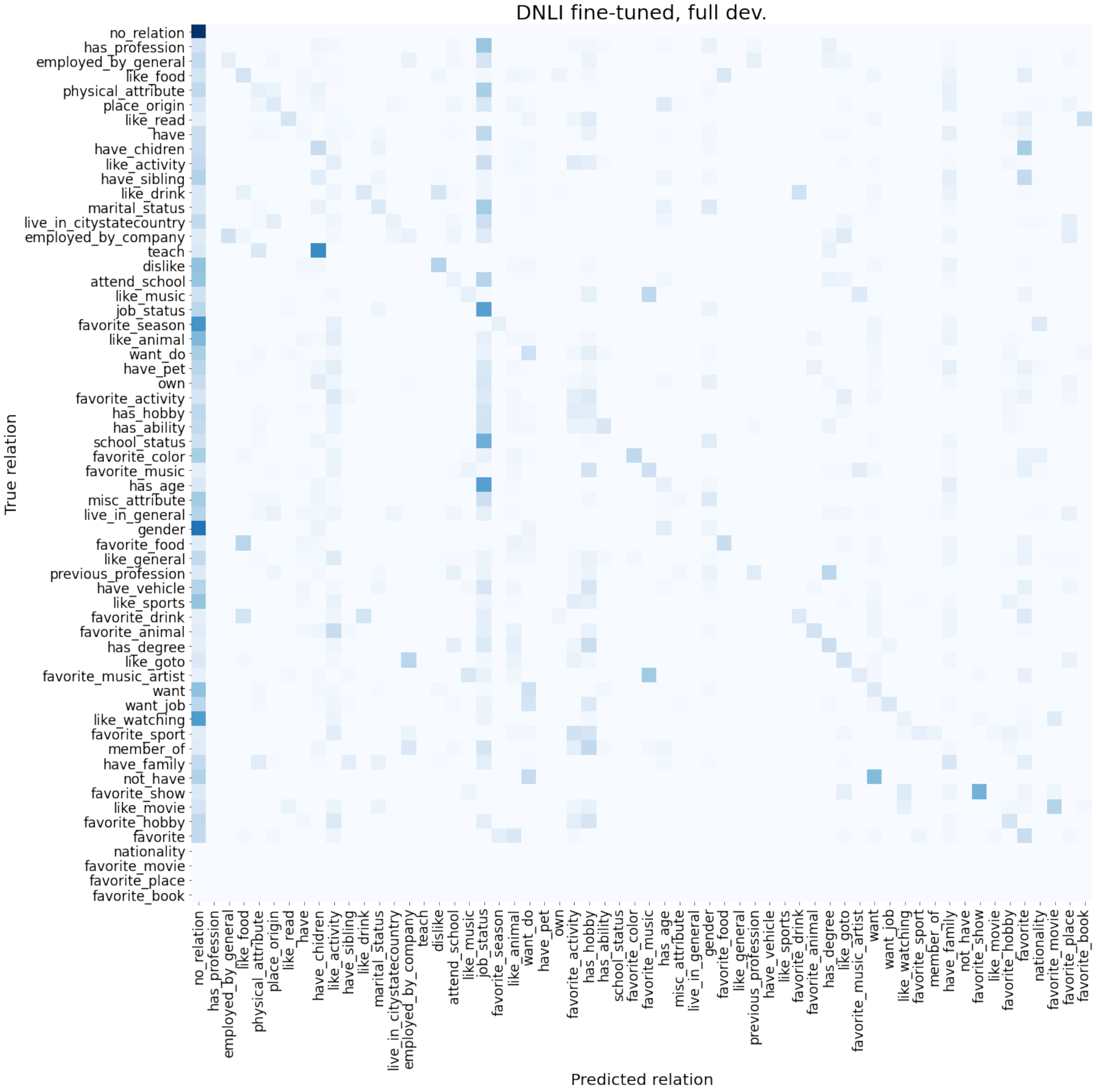

Figure C.2: Confusion Matrix for the DNLI pre-trained model in the full dev. scenario. True relations are given in the rows, predictions in the columns. The matrix is row-wise normalized. In this version, the color scheme is scaled linearly with the frequency of predictions.