Transformer Based Arabic Temporal Common Sense Understanding

Reem Alqifari

Doctor of Philosophy

University of York
Department of Computer Science

March 2024

Dedicated to

my parents

Abstract

Understanding temporal common sense is crucial for machines to make accurate predictions and judgments in various domains, including natural language processing, robotics, and more. Comprehending temporal text is complex, and grasping temporal common sense is essential for interpreting the inherent temporal aspects of the text. Currently, there is limited availability of datasets containing temporal information in English, and no Arabic dataset specifically designed for this purpose exists. Constructing an Arabic dataset is vital and would be a valuable resource for the Arabic-speaking community.

This thesis provides a detailed overview of existing datasets, comparing the temporal understanding capabilities of English and Arabic datasets, and aims to bridge the performance gap between them. Additionally, the use of pre-trained language models (PLMs) is explored as an alternative to the limited performance of traditional deep learning models, which are based on recurrent neural networks (RNNs) and attention layers, in Temporal Common Sense Understanding (TCU). The outcomes of this approach are discussed. The study also investigates various cross-lingual transfer learning approaches, yielding promising results.

A distinctive feature of this thesis is the implementation of error categorization and temporal classification as innovative methods to analyze models' understanding of temporal common sense. Through detailed error categorization and classification of temporal elements, the study establishes a comprehensive framework to elucidate the specific challenges PLMs face in TCU. This methodology underscores the urgent need for further research into PLMs' capabilities in the realm of TCU.

Given the recent achievements of large language models (LLMs) in various natural language processing tasks, it is necessary to assess their capabilities in the field of TCU. The findings indicate that LLMs currently fall short of human performance in this task, highlighting the challenges and the significant effort required for TCU in both Arabic and English.

Table of Contents

Abstract **v**

List of Figures **xi**

List of Tables **xiii**

Acknowledgements **1**

Declaration **3**

1 Introduction **7**

1.1 Temporal Common Sense Understanding 7

1.2 Motivation 10

1.3 Thesis and Research Questions 11

1.4 List of Contributions 11

1.5 Thesis Structure 12

2 Background: Deep Learning Models and Temporal Text Understanding **15**

2.1 Introduction 15

2.2 Deep Learning Models 15

2.2.1 Traditional Deep Learning Models 16

2.2.2 Word Embeddings 17

2.3 Transformer-Based Models 19

2.3.1 Pre-trained Language Models 20

2.3.2 Multlingual Pre-trained Language Models 21

2.3.3 Large Language Models 22

2.4 Textual Reading Comprehension 26

2.4.1 Evaluation 26

2.5 Arabic Language 28

2.6 Temporal Text 30
2.7 Temporal Reasoning 31
2.7.1 Logical Temporal Reasoning 33
2.8 Arabic Temporal Text 34
2.8.1 Challenges of Arabic Temporal Text Understanding 34
2.9 Temporal Question Answering System 35
2.10 Common Sense Understanding 37
2.11 Temporal Common Sense Understanding 38
2.12 Summary 41

3 Constructing Arabic Temporal Common Sense Dataset 43
3.1 Introduction 43
3.2 Temporal Related Datasets 44
3.3 Dataset Construction 44
3.4 Evaluation 46
3.5 Dataset Split 48
3.6 Summary 49

4 Deep Learning Models for Temporal Common Sense Understanding 51
4.1 Introduction 51
4.2 Experiments 51
4.3 Results 54
4.4 Summary 56

5 Temporal Common Sense Understandings Using Transformers 57
5.1 Introduction 57
5.2 Related Works 58
5.3 Experiments 58
5.4 Results 59
5.5 Summary 64

6 Zero-shot Cross-Lingual Transfer Learning 67
6.1 Introduction 67
6.2 Related Works 68
6.3 Standard Fine-tuning vs. Adaptive Fine-Tuning 69
6.4 Cross- Lingual with Standard Fine-tuning 70
6.4.1 Experiments 70

6.4.2 Results and Analysis 70
6.5 Cross-Lingual with Adaptive Fine-tuning 75
6.5.1 Results and Discussion 77
6.6 Summary 78

7 Assessing Temporal Common Sense Understanding: An Error Analysis and Evaluation of PLMs 81
7.1 Introduction 81
7.2 Error Analysis 82
7.2.1 Human Challenging Errors 82
7.2.2 Linguistic and Task Complexity Errors 86
7.3 PLMs Evaluation 88
7.3.1 Results and Discussion 88
7.4 Summary 89

8 Evaluating LLMs Temporal Common Sense Understanding in Zero-Shot Settings 93
8.1 Introduction 93
8.2 LLMs Promises and Challenges 93
8.2.1 Motivation 93
8.2.2 Challenges 94
8.2.3 Evaluation 94
8.3 Related Works 95
8.4 Method 95
8.4.1 Data set 95
8.4.2 Zero-Shot Prompts 95
8.4.3 LLMs 96
8.5 Results and Analysis 96
8.6 Summary 99

9 Conclusion 103
9.1 Introduction 103
9.2 Revisiting the Thesis and Research Questions 103
9.3 Contributions 105
9.4 Limitations 105
9.5 Future Work 105

References **107**

Appendix A Hyper-parameter **117**

Appendix B The Sample of the Datasets for Conducting LLMs Evaluations **119**

List of Figures

1.1 Example of a TCU challenge 8

2.1 Evolution of Deep Leaning Models 16
2.2 Reading Comprehension Task 27
2.3 Allens Temporal Logic 13 Base Relations 33

3.1 Percentage of Unique Question-Answer Pairs in each Temporal Category . 47

4.1 Symmetrical Architecture 52
4.2 Proposed Architecture 52
4.3 Results of the Different Deep Learning Models 55

5.1 Standard Error for Each Model 61
5.2 Models Results- F1 Score for Each Temporal Aspects 63
5.3 Predictions of AraBERT vs. CAMeLBERT 64

6.1 Models Results 73
6.2 Mono-Lingual vs. Cross-Lingual 74
6.3 X-MOd Adapter 76
6.4 MAD-X Adpater 77

7.1 The Results of TCU and Temporal Classification for Arabic and English . . 90
7.2 The Results of TCU and Temporal Classification for Cross-Lingual 90
7.3 Results of Arabic Temporal Classification Compared to TCU 91

8.1 Comparing the Average Results for Arabic and English 96
8.2 Comparing the Results of GPT4 on Arabic and English 100
8.3 Comparing the Results of Arabic and English 101

List of Tables

2.1 Multilingual Models 23
2.2 Comparisons between LLMs Models 25

3.1 List of the Available Datasets 44
3.2 Dataset Statistics 46
3.3 Temporal Category Statistics 46
3.4 Sample of the Dataset 50

4.1 Results of the Deep Learining Models 54

5.1 Comparisons between BERT models 60
5.2 Results of Applying PLMs on the Arabic Dataset 61
5.3 P-Value for F1 and EM 64

6.1 Cross-Lingual Results 70
6.2 Cross-Lingual Standard Error 71
6.3 Comparing the Cross-Lingual to the Mono-lingual Results 72
6.4 Adaptive Models 78

7.1 Subjective Event Understanding Errors Examples 84
7.2 Cultural Temporal Interpretation Errors Examples 85
7.3 Linguistic and Task Complexity Errors 87

8.1 Prompt in Arabic and English 97
8.2 Results for both English and Arabic for each Prompts 99

B.1 Sample of Dataset that Used for LLMs Evaluation 120

Acknowledgements

First and foremost, I praise and thank Allah, the Almighty, for helping me and giving me strength throughout all the challenging moments in completing this thesis. Without His blessings, this work would never have been completed.

I express my gratitude to Professor Simon O'Keefe, who provided valuable guidance and support during my research. I am honoured to be supervised by such an accomplished and dedicated mentor.

I extend my sincere appreciation to Professor Hend AlKhalifa for her unconditional support and her patience. Throughout my academic journey, Professor Hend has been a source of guidance and inspiration, always available to offer her expertise and insights, even before I started my Ph.D. studies.

I would like to thank Dr. Dimitar Kazakov, as a TAP member of my research degree, for his feedback and support during my research.

I am extremely grateful to my parents, Fahad and Faridah, for their endless love, prayers, sacrifices, and support throughout my life.

I would like to express my sincere appreciation to my husband Ameen for his understanding, constant encouragement, and support throughout this journey. My heartfelt gratitude goes out to my children, Mohammed and Leen. Despite being busy and away from home, they have been a constant support for me.

I would like to express my gratitude to my sisters Haifa, Sara, and Njoud for their support, motivation and believing in me. I also would like to extend my appreciation to my brothers Abdulaziz and Suliman. Their presence was always there for me whenever I needed it. I am grateful to my aunts for their prayers.

I would like to express my gratitude to my friends Noof and Alaa, who provided support during the good and the challenging times. I am truly grateful to have friends who were also studying and with whom I could share all my worries and joys and celebrate each milestone together.

I am thankful for the scholarship from the External Join Supervision Programme at King Saud University, which has made this endeavour possible. I express my gratitude to the Postgraduate Research Administration at the Computer Science Department for their assistance and continuous responsiveness to all my inquiries.

Lastly, I am grateful to all those who assisted me in any capacity to complete this thesis.

Reem Alqifari,
January 2024.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Related Publications

Reem Alqifari. *Question Answering Systems Approaches and Challenges*. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 69-75, 2019.

Reem Alqifari, Hend Al-Khalifa and Simon O'Keefe *Arabic Temporal Common Sense Understanding*. Submitted for publication, *Recent Advances on Computational Linguistics and Natural Language Processing*, 2024.

Abbreviations

NLP Natural Language Processing

NLU Natural Language Understanding

TCU Temporal Common Sense Understanding

MSA Modern Standard Arabic

QA Question Answering

RC-QA Reading Comprehension Question Answering

MRC Multiple-choice Reading Comprehension Question Answering

AI Artificial Intelligence

IA Interval Algebra

QSTR Qualitative Spatio-Temporal Reasoning

RNN Recurrent Neural Network

LSTM Long-Short Term Memory

Bi-LSTM Bi-directional Long-Short Term Memory

GRU Gated Recurrent Unit

Bi-GRU Bi-directional Gated Recurrent Unit

BERT Bidirectional Encoder Representations from Transformers

RoBERTa Robustly optimized BERT approach

DeBERTa Decoding-enhanced BERT with disentangled attention

XLM-RoBERTa multilingual version of RoBERTa

PLM Pre-trained Language Model

LLM Large Language Model

GPT Generative Pre-trained Transformer

Chapter 1

Introduction

1.1 Temporal Common Sense Understanding

The machine's comprehension of a text requires temporal understanding. This understanding can be hindered by the presence of explicit or implicit temporal features in the text. Although understanding the temporal text is complex, comprehending temporal common sense is crucial to understanding the inherent temporal aspects of the text. Although humans possess this ability, it is challenging for machines to replicate it. The scope of this research is textual reading comprehension, which employs the multiple-choice format. The MRC (Multiple-Choice Reading Comprehension) system is imperative for determining a suitable answer from a collection of potential responses, relying on the contextual information presented. To fulfil the requirements of the MRC task, which involves selecting the answer from multiple choice options, the proposed model must identify the correct response from a range of candidate alternatives. Multiple models have been applied to measure the ability of the recent deep learning model to understand temporal common sense.

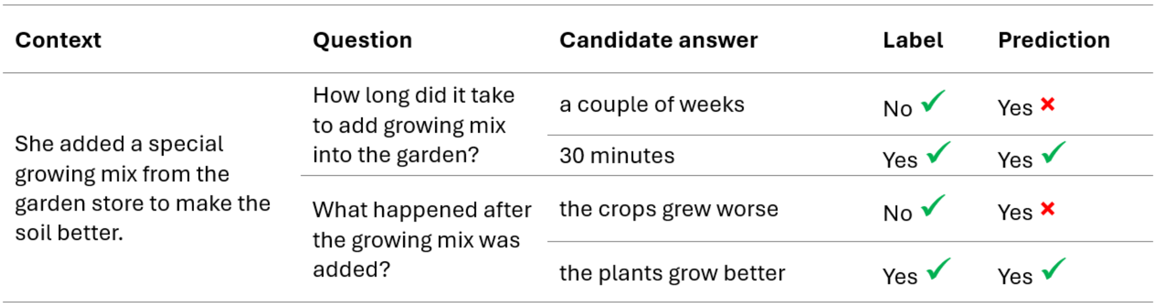
The model will take three inputs (context, question, and answer), and the predictor should learn to predict whether this answer is plausible. This model uses examples of textual training, where c is a context that is a passage of text, q is a question relevant to the context c, and a is an answer to question q. The model aims to learn a predictor l, which takes a context c, a corresponding question q and a candidate answer a as inputs and predicts the score that will be high if the answer is likely plausible and low otherwise. This predictor model could be formulated as the following formula:

$$Score_i = l(\{(c_i, q_i, a_i)\})_{i=1}^{n} \in [0,1] \tag{1.1}$$

This formula is suggested so the model can be applied to various MRC tasks with multiple correct answers.

The term ‘common sense’ can be defined as "the basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people." [65].

Understanding temporal common sense is a crucial aspect of human intelligence, enabling people to grasp the order in which events occur. Human beings innately understand that certain events precede others, such as falling ill before dying. However, machines face difficulties in acquiring this understanding, requiring complex algorithms, and programming to infer the chronological sequence of events. In addition to event ordering, temporal aspects include event duration. Humans find it relatively easy to predict the duration of activities such as eating, opening a door or walking. However, the limited datasets about the temporal understanding or extraction currently available primarily involve unusual events or unusual durations, posing a significant challenge for machines in anticipating the duration of routine events. Figure 1.1 illustrates an example of a TCU challenge and how the model can fail to validate the correct answer.



Context	Question	Candidate answer	Label	Prediction
She added a special growing mix from the garden store to make the soil better.	How long did it take to add growing mix into the garden?	a couple of weeks	No ✓	Yes ✗
		30 minutes	Yes ✓	Yes ✓
	What happened after the growing mix was added?	the crops grew worse	No ✓	Yes ✗
		the plants grow better	Yes ✓	Yes ✓

Figure 1.1 Example of a TCU challenge showing a scenario where the model fails to validate the correct answer.

The scope of this thesis is not focused on algorithmic approaches for temporal reasoning, such as dependency tree parsing or logical propositions. Instead, the primary goal is to apply deep learning models to assess their effectiveness, particularly for the Arabic language using a constructed dataset. The thesis will also involve a comparative analysis of the performance of these models between Arabic and English.

Challenges

The challenges will be classified into three main categories:

1. **Implicit Temporal Features:** Temporal reasoning is complicated because some events are vague. So, extracting and annotating temporal features will be a complex task [67] For instance, the sentence "She visited her friend after finishing her work." is ambiguous because it does not specify whether the visit happened immediately after finishing work, a few hours later, or even the next day. This ambiguity, where the exact timing of events is unclear, makes it challenging for models to accurately capture and process temporal information.

   Moreover, implicit temporal features add another layer of complexity. For example, in the sentence "Sara finished her breakfast and left for school," the temporal sequence must be inferred, as it implies that Sara left for work shortly after finishing her breakfast without explicitly stating the time interval. Similarly, the phrase "he often travels for work" implies a frequency of events without providing specific details on how often the travels occur. These implicit temporal cues require models to deduce the order and frequency of events from the context, further complicating the task of temporal reasoning.

   From the MC-TACO dataset [83], if the question concerns an event's duration, all candidate responses belong to a duration type. The challenge is how to validate whether there is a logical duration for this event. Because each candidate's answer has a different duration, categorizing the answers based on the temporal type of the question—for instance, duration—will not eliminate answers that do not fit into the category. Thus, the model should acquire temporal common sense knowledge. For example, the model should know that 30 seconds would be an illogical illness duration; that is, in this case, 30 seconds would be a valid duration but not a logical answer. Acquiring this knowledge is expensive and difficult.

2. **Limited Data:** While there are few datasets available in English, there is currently no dataset specifically designed for TCU in Arabic. One of the most widely used English datasets is MC-TACO, which is designed to evaluate models on TCU. MC-TACO is small, lacks a specific training split, and consists of only evaluation and test sets. In addition, the evaluation set is quite small, contains only 3,783 question-answer pairs. Moreover, to the best of our knowledge, there is no dataset in English that is designed to cover all temporal featuers except MC-TACO. This scarcity of datasets significantly affects the development of models for TCU.

3. **Lack of Knowledge:** According to existing research, current language models lag behind human performance in the task of common sense understanding. For example, this is evident from the MC-TACO leaderboard. Numerous studies have shown that

ths performance gap can be overcome by relying on external sources that encapsulate the common sense knowledge [65, 80, 76, 84]. For instance, as previously discussed, temporal reasoning involves understanding sequences of events, durations, and implicit time-related features, which are often not fully captured by existing datasets [76, 84]. As a result, models struggle to make accurate predictions. Therefore, insufficient data restricts the improvement of these models.

Existing models still struggle to understand the varying lengths of different the events. As, the duration of a verb describing an event can change depending on context. For example, the duration of the verb taking, the act of "taking a vacation" generally takes longer than "taking a shower". The latter can usually last for only a few minutes, whereas the former can last for several days or even weeks [83]. To address this issue, there should be a source of knowledge to help the model accurately capture this temporal context. The existing corpus that can be used for this purpose is skewed towards uncommon or unexpected event durations and rare events [83]. For example, the duration of "opening a door" is not mentioned unless it is longer than usual. Determining the duration of various events manually is expensive and time consuming.

Addressing this gap requires constructing comprehensive knowledge bases (KBs) specifically designed for temporal information, or alternatively, developing more advanced models and algorithms that can learn and infer temporal common sense from the limited data available.

1.2 Motivation

Despite the recognized importance of temporal comprehension in Natural Language Understanding (NLU), research in Temporal Common Sense Understanding (TCU) remains sparse, particularly beyond the English linguistic domain. This scarcity is pronounced in languages such as Arabic, largely due to the lack of necessary resources. This highlights an urgent need for dedicated Arabic datasets to foster advancements in NLU for the Arabic-speaking community.

While several temporal datasets exist for English, they do not have an equivalent in Arabic, limiting the development and testing of NLU systems in the Arabic context. Creating an Arabic-focused TCU dataset in Modern Standard Arabic (MSA) would address this gap and enhance the processing of temporal information in Arabic NLU applications. MSA is chosen over dialects because it is the standardized and widely understood form of Arabic used in formal communication, media, and education across the Arabic-speaking world.

Unlike regional dialects, which vary significantly and may not be mutually intelligible, MSA provides a consistent linguistic foundation for developing and evaluating NLU systems.

Furthermore, the application and evaluation of multilingual PLMs in TCU is an under-explored area. While some work addresses common sense reasoning, research specifically focusing on Arabic or using multilingual PLMs for TCU tasks is scarce. This presents a significant opportunity for pioneering studies that could improve NLU systems, particularly in Arabic.

The critical need for targeted research efforts in this area is clear. Applying multilingual PLMs to TCU with a focus on MSA offers the potential to uncover culturally and linguistically specific temporal reasoning. Bridging these research gaps is essential for enhancing synergy between English and Arabic NLU systems and narrowing the performance disparity. Extending TCU research to include MSA will contribute to a more nuanced understanding of temporal common sense across diverse linguistic landscapes.

1.3 Thesis and Research Questions

This research addresses the following questions based on the challenges outlined earlier:

1. Can an existing English dataset be utilized to develop an Arabic dataset for Temporal Common Sense Understanding?
2. How effective are the available multilingual PLMs in demonstrating temporal common sense? Can these models compete with their monolingual counterparts? Does the performance of these models vary between Arabic and English?
3. How effective is the use of cross-lingual transfer learning for this task? Can this approach mitigate the performance gap between English and Arabic?
4. What specific challenges and error patterns emerge from applying PLMs to TCU in Arabic and English contexts?
5. To what extent is Generative AI successful in TCU? How does the language of the dataset influence the model's performance?

1.4 List of Contributions

This thesis makes several significant contributions to the field of Temporal Common sense Understanding (TCU) using deep learning models. The contributions are delineated below in a descending order based on their influence on the field.

1. **Constructing of a TCU Arabic Dataset:** An Arabic dataset is constructed to serve the TCU task. This construction is expected to be highly impactful for the Arabic community and addresses the absence of such a resource. The dataset is based on an existing English dataset (Chapter 3).

2. **Benchmarking for Temporal Understanding:** To evaluate the ability of PLMs in understanding temporal features, a benchmark for temporal understanding was established (Chapter 7).

3. **Applying Large Language Models (LLMs):** Utilized state-of-the-art large language models with different prompts and zero-shot learning techniques to advance temporal common sense comprehension capabilities (Chapter 8).

4. **Applying Multilingual Pre-Trained Language Models (PLMs):** Examining the effectiveness of different multilingual PLMs to the MC-TACO (the original English dataset) and the Arabic dataset. Each model was assessed in terms of each temporal aspect (Chapter 5 and Chapter 6).

5. **Analysing Errors:** By analyzing the errors, a new classification is suggested to identify specific issues, which will help improve the understanding of PLMs (Chapter 7).

6. **Applying Different Deep Learning Models:** Various deep learning models were implemented, including architectures based on word embeddings, recurrent neural networks (RNNs), and attention mechanisms, to enhance temporal common sense comprehension (Chapter 4).

1.5 Thesis Structure

The rest of the thesis is organised as follows:

- Chapter 2 provides background information on temporal text and temporal common sense understanding, deep learning models, and reading comprehension tasks. Beyond this, Chapter 2 examines the existing literature on temporal comprehension.

- Chapter 3 constructs an Arabic dataset for temporal common sense understanding. This chapter provides a detailed overview of existing datasets related to temporal information in English and Arabic text.

- Chapter 4 presents an initial attempt to address the problem by applying a set of deep learning models specifically designed for temporal common sense comprehension.

The effectiveness of these models in this domain is detailed. These models, developed through this research, employ various architectures, including word embeddings, recurrent neural networks (RNNs), and attention mechanisms.

- In Chapter 5, the use of pre-trained large language models (PLMs) is explored as an alternative to the inadequate performance of basic deep learning models in this task. The outcomes of this approach are thoroughly examined and discussed.

- The performance of multilingual PLMs on Arabic lags behind that of these PLMs on English. To bridge this gap, Chapter 6 uses various cross-lingual approachs.

- Chapter 7 introduced a methodology for evaluating the effectiveness of multilingual PLMs by analysing the errors and trying to categorise them. Also, try to establish a benchmark for assessing PLMs in temporal classification to deeply understand the source of the challenge that affect PLMs performance in the TCU.

- Chapter 8 examines the effectiveness of various models on Arabic and English datasets, considering the recent success of LLMs in numerous NLP tasks.

- Chapter 9 concludes this thesis and indicates potential directions for future research.

Chapter 2

Background: Deep Learning Models and Temporal Text Understanding

2.1 Introduction

This chapter provides a comprehensive overview of the key concepts related to temporal text and deep learning models. Additionally, the chapter discusses the underlying principles of temporal text, such as temporal features, its distinctions from other text types and why it poses a challenge in natural language processing (NLP). Moreover, the chapter presents the difficulties associated with Arabic temporal text and the relevant research conducted in this area. Reading comprehension tasks are adopted to evaluate temporal common sense understanding. Therefore, a concise overview of these tasks is provided. Additionally, this chapter provides a detailed explanation of deep learning models, exploring their mechanisms and the types of models commonly used in NLP tasks. Ultimately, this chapter should provide a solid foundation of knowledge in this field.

2.2 Deep Learning Models

Deep learning (DL) is considered a subset of machine learning (ML). It allows computers to learn from experience using a layered structure of algorithms called an artificial neural network (ANN). This layered structure, often referred to as a neural network, mimics the way the human brain processes information. Deep learning reduces the need for manual feature engineering, which traditionally required significant effort to identify the most relevant features for a given task.

Deep learning includes both traditional models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as transformer-based models, which have recently gained prominence due to their superior performance in various tasks. Transformer models, introduced in the milestone paper "Attention Is All You Need" in 2017 [75], have revolutionized the field by enabling more efficient training and better handling of sequential data.

Since its inception in the early 2006 by Hinton et al.[35], deep learning has been widely used in different Natural Language Processing (NLP) tasks [66, 28], such as text classification, sentiment analysis, machine translation, and named entity recognition, consistently outperforming other methods. Recent advancements in Question Answering (QA) Systems have particularly followed a deep learning approach, leveraging the power of neural networks to understand and generate human language more effectively. The evolution of deep learning models has seen significant advancements over the years. Figure 2.1 illustrates the progression from traditional deep learning models to modern transformer-based models, highlighting key milestones and innovations in the field. The categorization of models into traditional DL, PLMs, and LLMs is based on [42, 52, 49].

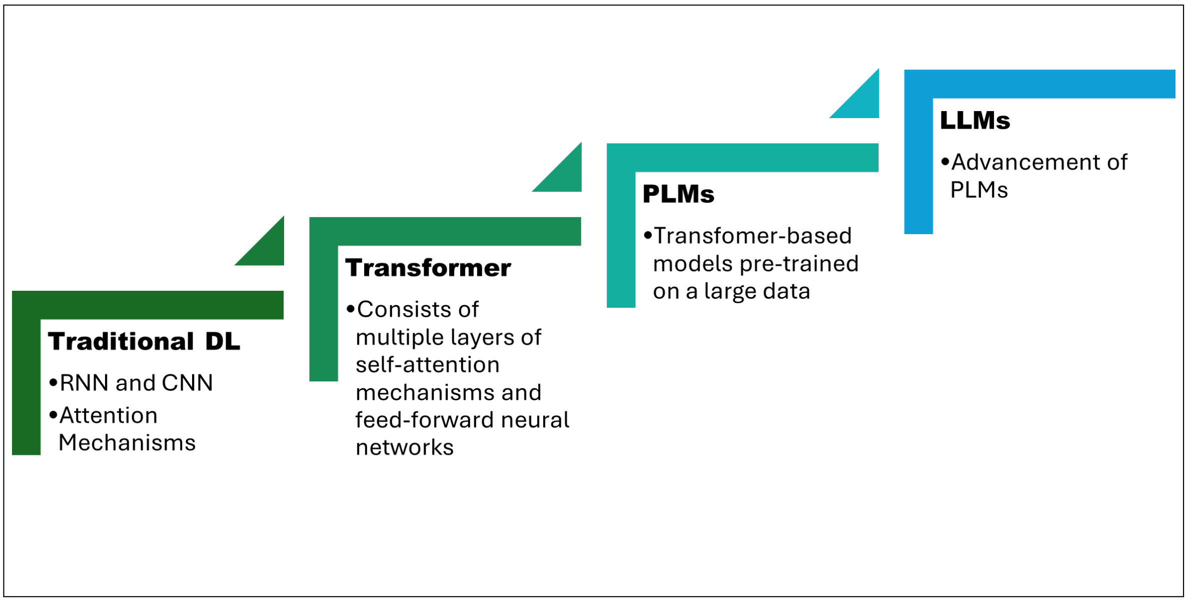

Traditional DL
•RNN and CNN
•Attention Mechanisms
Transformer
•Consists of multiple layers of self-attention mechanisms and feed-forward neural networks
PLMs
•Transfomer-based models pre-trained on a large data
LLMs
•Advancement of PLMs

Figure 2.1 Evolution of Deep Leaning Models

2.2.1 Traditional Deep Learning Models

There are different techniques for deep learning. For instance, convolutional neural networks (CNNs) and recurrent neural networks (RNNs). RNNs are popular for NLP tasks. CNNs are used for processing data that is known to have a grid-like topology. They contain one or more convolutional layers, which are pooling or fully connected, and use a variation of a multilayer perceptron. Convolutional layers apply a convolution operation to the input and pass the result to the next layer Recurrent Neural Networks (RNNs): An RNN models sequential interactions through a hidden state or memory, which is updated at each time step as it processes the input sequence. This hidden state captures information from previous time steps and propagates it forward, enabling the RNN to maintain context and dependencies across the sequence. The RNN can take up to N inputs and produce up to N outputs. It has some variations, including long short-term memory (LSTM) and gated recurrent units (GRUs).

- LSTM networks were invented to prevent the vanishing gradient problem in RNN by using a memory gating mechanism. Using LSTM units to calculate the hidden state in an RNN helps the network to efficiently propagate gradients and learn long-range dependencies.

- GRU is a simplified version of an LSTM unit with fewer parameters. Just like an LSTM cell, it uses a gating mechanism to allow RNNs to efficiently learn long-range dependency by preventing the vanishing gradient problem.The vanishing gradient problem occurs when the gradients used to update the network's weights during training become extremely small, effectively preventing the network from learning long-range dependencies because the early layers of the network learn very slowly. A GRU consists of a reset and update gate that determine which part of the old memory to keep or update with new values at the current time step.

- An RNN can also be bi-directional, which allows the hidden states to receive information from both the preceding and succeeding elements in the sequence, not only from the previous state. This can be applied in LSTM or GRU.

- Attention mechanisms are inspired by human visual attention, the ability to focus on specific parts of an image or a text. Attention mechanisms can help the network learn what to focus on when making predictions.

- Embedding maps an input representation, such as a word or sentence, into a vector. A popular type of embedding is word embedding, such as word2vec or GloVe.

2.2.2 Word Embeddings

The principal concept underlying word embeddings involves the representation of words in a manner that encompasses their meanings, semantic connections, and the diverse context in which they can be used. Representing the meaning of the word is crucial in any NLP application. Using WordNet is not adequate as it is not complete; some new words are missing. Moreover, it is difficult to compute accurate similarity, which is vital in most NLP tasks. The best way to represent the word is using a vector that can use distance measures such as Euclidian or cosine similarity. The first attempt was to use a discrete representation, which deals with a word as atomic and uses vector space to represent it, called "one-hot" [0,1,0,0,1,1]. The problem with this kind of representation is the sparsity, which requires massive storage that can be as extensive as the size of the vocabulary. Many techniques have been suggested to overcome the weakness of the "one-hot" representation. Word embeddings are a form of word representation that allows words with similar meanings to possess a similar representation. Words that are used in similar contexts tend to be closer to each other in the embedding space, making it possible to compute similarities between words based on their vector representations or even to produce analogies such as "king - man + woman = queen".

Word embeddings are a dense representation of words in a limited-dimensional expanse (usually several hundred dimensions), which overcomes the main issues of the one-hot encoding. The main advantage of this technique is that it can be used for a vast corpus that contains millions of words in the vocabulary.

There are different methods of word embeddings as listed below:

- **Word2Vec:**In 2013, Mikolov et.al. from Google introduced it. [51]. Neural networks are used to create word embeddings. However, there are two different approaches that are used: Skip-gram and Continues Bag of words (CBOW). CBOW is a model for predicting a target word from the surrounding context. In contrast, Skip-gram is a model for predicting surrounding context words given the target word. Soft-max and negative sampling are used for training.

- **GLOVE (Global Vectors for Word Representation):** The NLP group at Stanford University introduced this method in 2014, following the publication of word2vec by Pennington et al. [55]. This method primarily concentrates on the co-occurrence of words throughout the entire corpus. The embeddings associated with it represent the likelihood of two words appearing together.

- **FastText:** In 2017, a novel word embeddings technique was introduced by Facebook [29]. FastText has been proposed to mitigate the lack of generalisation for unknown words by Word2Vec. The idea of FastText is similar to Word2Vec. The distinction lies in the fact that the result produced by FastText is a combination of lower-level embeddings, including both word and character components.

The acquisition of embeddings in Word2Vec is based on establishing connections between target words and their corresponding context. However, it does not consider the difference in occurrence frequency among context words. In Word2Vec, a higher frequency of word co-occurrence leads to more training examples, but it does not provide additional information. On the other hand, GloVe recognizes the importance of co-occurrence frequency as valuable information that should not be ignored. Instead of following the Word2Vec approach, GloVe constructs word embeddings by directly linking a combination of word vectors to the likelihood of co-occurrence for those words in a given corpus.

2.3 Transformer-Based Models

The introduction of transformer-based models in 2017 marked a significant advancement in deep learning. Unlike traditional RNNs, transformers use self-attention mechanisms to process input data in parallel, making them highly efficient and effective for various tasks [75].The transformer architecture consists of multiple layers of self-attention mechanisms and feed-forward neural networks. There are three main types of transformer models:

- textbfEncoder-Decoder Models: These models utilize both the encoder and decoder parts of the transformer. They are particularly effective for sequence-to-sequence tasks such as machine translation, where the encoder processes the input text, and the decoder generates the corresponding output text. The original transformer model demonstrated state-of-the-art performance in translating between languages[75].

- textbfEncoder-Only Models: These models use only the encoder part of the transformer, like BERT[19]. They are designed for tasks that require understanding the context of the input text, such as text classification and named entity recognition.

- **Decoder-Only Models or Generative Pre-trained Transformer (GPT):** These models use only the decoder part of the transformer. They are optimized for text generation tasks, where the model needs to produce coherent and contextually relevant text from a given prompt. The pioneering paper presenting this architecture was published by OpenAI [61].

The introduction of transformers led to the development of Pre-trained Language Models(PLMs), such as BERT, which leverage large-scale pre-training on vast corpora followed by fine-tuning on specific tasks. PLMs have set new benchmarks in various NLP tasks by capturing contextual information more effectively than previous models. Building on the success of PLMs, Large Language Models (LLMs) like GPT-3 represent an improvement and extension of these models. LLMs can generate coherent and contextually relevant text, perform zero-shot learning, and handle a wide range of NLP tasks with minimal fine-tuning.

2.3.1 Pre-trained Language Models

PLMs are transformer-based models. Unlike traditional deep learning models and standard transformer models, they can be easily tailored to specific downstream tasks through the process of fine-tuning. This section explores the architectures of the language models and their advantages for enhancing NLU tasks. Below is a list of the PLMs applied in this research.

BERT (Bidirectional Encoder Representations from Transformers)

This model was introduced by Devlin et al. (2019) and is known as BERT [19]. This discovery represented a significant advancement in the domain of natural language processing. Pre-trained on a massive dataset of text and code. It is trained specifically on Wikipedia (2.5B words) and Google's BooksCorpus (800M words). BERT adopted a Transformer architecture with bidirectional attention. Employs Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP).

RoBERTa (Robustly optimised BERT approach)

Liu et al. introduced RoBERta, an improved version of BERT that incorporates enhanced training techniques [47]. It utilizes larger batch sizes, dynamic masking, and longer training sequences. One notable difference is that it eliminates the Next-Sentence Prediction (NSP) objective. In general, RoBERta demonstrates superior performance compared to BERT across a range of natural language processing (NLP) tasks. This is attributed to its optimized training approach, which makes it both more robust and efficient.

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

ELECTRA is a pre-training method for natural language processing that uses a generator and a discriminator [16]. Unlike traditional masked language models like BERT, ELECTRA's discriminator is trained to detect whether each token in a sequence is original or replaced, rather than just predicting masked tokens. This replaced token detection allows ELECTRA to learn from all tokens in the sequence, making it more efficient and requiring less computational resources. The pre-trained discriminator can then be fine-tuned for various downstream tasks, achieving comparable or superior performance to models like BERT with greater training efficiency.

DeBERTa (Decoding-Enhanced BERT with disentangled Attention)

He et. al. from Microsoft proposed DeBERTa [34]. This model presents two novel techniques that enhance the performance of RoBERTa. The first technique, disentangled attention, improves attention modeling by separating word content and position. The second technique, the enhanced mask decoder, utilises an improved output layer for predicting masked tokens. These enhancements result in state-of-the-art performance on various NLP benchmarks.

DeBERTa-v3

Microsoft has made further advancements in the DeBERTa model following its successful implementation, and in 2023, they introduced an upgraded version called DeBERTa-v3 [33]. DeBERTa-v3 is an improved version of DeBERTa that introduces some notable differences. One key distinction is that DeBERTa utilizes masked language modeling (MLM), where the model predicts randomly masked tokens in a text sequence. In contrast, DeBERTa-v3 incorporates replaced token detection (RTD), a more efficient pre-training task inspired by ELECTRA. This allows the model to distinguish between the original tokens and the randomly replaced ones. Additionally, DeBERTa-v3 introduces Gradient-Disentangled Embedding Sharing (GDES) to enhance the vanilla embedding sharing from ELECTRA, resulting in improved training efficiency and performance.

2.3.2 Multlingual Pre-trained Language Models

Each model of the above models, except ELECTRA, was trained on a multilingual corpus to improve its ability to understand and generalise across languages. A comprehensive description is provided in this thesis for the three models. Beyond this, a comparison between models is illustrated in Table 2.1, presenting a comprehensive analysis of the features and specifications of each model, as part of this research.

Multilingual Bert: Multilingual BERT is an enhanced version of BERT which can handle multiple languages. The training data for multilingual BERT is obtained from the complete Wikipedia dump for each language. However, the size of Wikipedia content varied significantly across different languages. Consequently, low-resource languages may not have enough representation in the neural network model, potentially impacting the model's effectiveness for low-resource languages.

XLM-Roberta: The Cross-Lingual Language Model Robustly Pretrained is an extension of RoBERTa focusing on cross-lingual applications. Trained on the Cross-Llingual Language Model (XLM) training data, including ClueWeb09, Common Crawl, and Wikipedia dumps in 100 languages.

mDeBERTa: (Multilingual DeBERTa)Multilingual DeBERTa is an extension of DeBERTa, a transformer model that aims to capture dependencies with greater accuracy, which supports multiple languages. This model has undergone training using the exact same set of training data used by XLM-RoBERTa.

The size of the Arabic dataset that was used to train the XLM-Roberta and mDeBERTa trains was 28 GB, while the English was 300.8 GB [59]. The Arabic dataset size was almost double that of the dataset used to train mBERT.

In this vein, mBERT is trained simultaneously with multiple languages, feeding text from various languages without providing additional information on the relationships between the languages [59]. Based on a survey by Pikuliak et al.[59], even in the absence of information on the interconnections between languages, mBERT can create representations which are partially independent of any specific language. Additionally, this model can understand the connections between languages without explicit instructions. Several hypotheses have been proposed to explain the underlying nature of this inherent ability of mBERT. Also, according to Pikuliak et al.[59], one hypothesis suggests that these models rely on a partial overlap of vocabulary between languages, which has been negated by other studies. The second hypothesis suggests that multilingual models can discover certain universal structural similarities between languages, which is reminiscent of the concept of universal grammar (Chomsky, 2007). Based on the aforementioned hypothesis, the disparity between English and Arabic will be substantial for cross lingual transfer learning, due to the presence of less common structural characteristics in these languages. However, there are models directly use some form of cross-lingual supervision, which means that they leverage data to establish connections between different languages. Examples of such models include XLM-RoBERTa [17] and mDeBERtav3 [33].

2.3.3 Large Language Models

LLMs are an advanced subset of PLMs characterized by their massive scale in terms of parameters and data, and their ability to perform a wide range of tasks with minimal task-specific fine-tuning. They leverage techniques like few-shot and zero-shot learning to adapt to new tasks. Zero-Shot Learning is a learning paradigm where a model is able to perform a

Feature	**mBERT**	**XLM-RoBERTa**	**mDeBERTa-v3**
Vocabulary size	120K	250K	250K
Vocabulary Type	–	Sub-word level	Full-word
Parameter size	178M	355M (base), 550M (large)	137M (base)
Pre-training objective	Masked Language Modeling (MLM)	Masked Language Modeling (MLM)	Replaced Token Detection (RTD)
Attention type	Masked self-attention	Cross-lingual attention	Disentangled attention
Languages	104	100	100
Training data	Wikipedia + books	Common Crawl (CC100)	Common Crawl (CC100)
Model size	Case - Uncased	Base, Large	Base
Performance	Varied	Better than mBERT on many tasks [17]	State-of-the-art in many cases [33]

Table 2.1 Multilingual Models

task without having been explicitly trained on any examples of that task. Instead, the model leverages its general knowledge acquired during pre-training to infer the correct output [78].

LLMs enable a deeper understanding and generation of human-like text. A brief description of the models applied in Chapter 8 is provided below. Table 2.2 summarises the main features of the models and compares them.

GPT 3.5 Turbo It is developed by OpenAI [14]. GPT-3.5 Turbo is an enhanced version of the GPT-3 model. it offers improvements over GPT-3 in terms of both speed and efficiency. It is designed to be faster while maintaining or even enhancing the quality of the generated text. GPT-3 is a very large model with 175 billion parameters, which are the parts of the model that are learned from the training data [14]. Its architecture is based on a transformer.

GPT 4 [1] It is the most powerful large -multimodal model released by OpenAI [54]. Despite the notable advances demonstrated by GPT-4 from previous GPT models, this model still shares certain limitations with its predecessors. These limitations include the generation of

[1] https://openai.com/gpt-4

inaccurate information, the potential to provide harmful guidance, a limited context window, and the absence of the ability to learn from past experiences [49].

Google Gemini [2] an AI model developed by Google, was introduced on December 6, 2023. This model comes in three varying sizes: Ultra, Pro, and Nano, listed in decreasing order. According to the report, Gemini Ultr has shown superior performance to GPT4. This study evaluated the models across various tasks, such as Massive Multitask Language Understanding (MMLU) and reasoning tasks, including reading comprehension tasks. Additionally, Google stated that Gemini is the first model to surpass the expertise of human professionals in MMLU. Gemini models are constructed based on transformer decoders. The models are trained to handle a context length of 32k, utilising efficient attention mechanisms such as multi-query attention. The sentence piece tokenizer is used [24].

AceGPT was launched in 2023 by the School of Data Science, the Chinese University of Hong Kong, Shenzhen (CUHKSZ), the Shenzhen Research Institute of Big Data (SRIBD) and the King Abdullah University of Science and Technology (KAUST) [36]. The model is based on LLaMA2, and its focus is localising the large-language models to the Arabic language. In the study, the model trained LLaMA2-7B with 30B data (19.2B tokens in Arabic and 10.8B in English) and LLaMA2-13B with 10B data (6B tokens in Arabic and 4B in English), prioritising a larger quantity of Arabic than English data.

Jais [3] was introduced in 2023 through a collaboration between Inception, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), and Cerebras Systems[68].

Generally, deep learning, as part of the broader field of machine learning, has significantly advanced the capabilities of NLP applications by automating the feature extraction process and providing robust models that can learn complex patterns from large datasets. This evolution has enabled the development of sophisticated systems that can perform tasks previously thought to be beyond the reach of machines.

[2] https://ai.google.dev/

[3] https://www.arabic-gpt.ai/

Model	**Dataset**	**Parameter Size**	**Model Architecture**	**Supported Languages**
GPT-3.5 Turbo	Books, articles, code, web text (estimated 300B words)	1.5T parameters	Transformer with Decoder-only architecture	Multilingual
GPT-4	Proprietary dataset (details unreleased)	Expected to be significantly larger than GPT-3.5	Likely a variation of Transformer architecture	Multilingual
Gemini	Similar to Bard, with additional focus on creative text formats	540B	Transformer with Encoder-Decoder architecture	Multilingual
AceGPT	Open source Arabic text 2022 and from Arabic Wikipedia, CC100, and OSCAR3. The English dataset is obtained from Slim Pajama. (estimated 1.5T words)	7B - 13B	Transformer with Decoder-only architecture	Arabic but can support other languages
Jais	Wikipedia corpora	13B - 30B	Likely a variation of Transformer architecture	Arabic, English

Table 2.2 Comparisons between LLMs Models

2.4 Textual Reading Comprehension

Understanding this type of QA is essential for this research. Most of the TCU dataset are designed as a reading comprehension question answering task. Textual comprehension of the language encompasses various classifications [82], namely:

1. Cloze-style (Fill in the Blanks): In this task, specific placeholders exist within the question. The MRC system must identify the most appropriate words or phrases that can be inserted into these placeholders based on the contextual content provided.
2. Multiple choice: In a task involving multiple choice, the Multiple-choice Reading Comprehension (MRC) system must choose the correct answer from a set of candidate responses provided per the given context.
3. Span prediction: In a task that involves span prediction, the answer corresponds to a specific section of text within the context. In other words, the MRC system must correctly determine the start and end of the answer text within the context. An example of this task is the SQuAd dataset.
4. Free-form answer: Tasks of the free-form answer variety allow the answer to take on any form of free-text expression. Consequently, the answer is not limited to a word or a specific section within the passage.

However, the above classification is incomplete as it does not account for hybrid tasks, such as multiple-choice span prediction. In such tasks, the MRC system must select the correct span from multiple provided options, combining elements of both span prediction and multiple-choice tasks. This highlights the need for a more nuanced classification system that can accurately capture the complexity and variety of modern MRC tasks.So, [82] proposed a new classification model that is summarised in Figuer 2.2. The proposed categories are built on the basis of available datasets, so the model might be updated as a new task is established.

2.4.1 Evaluation

Evaluating the state-of-the-art in question answering can be tackled from different aspects, including the accuracy of the system, speed, and scalability. The following list describes each one separately.

- What is the accuracy of the state of the art? How can I create a more accurate quality control system? Question-answer systems can be evaluated according to different

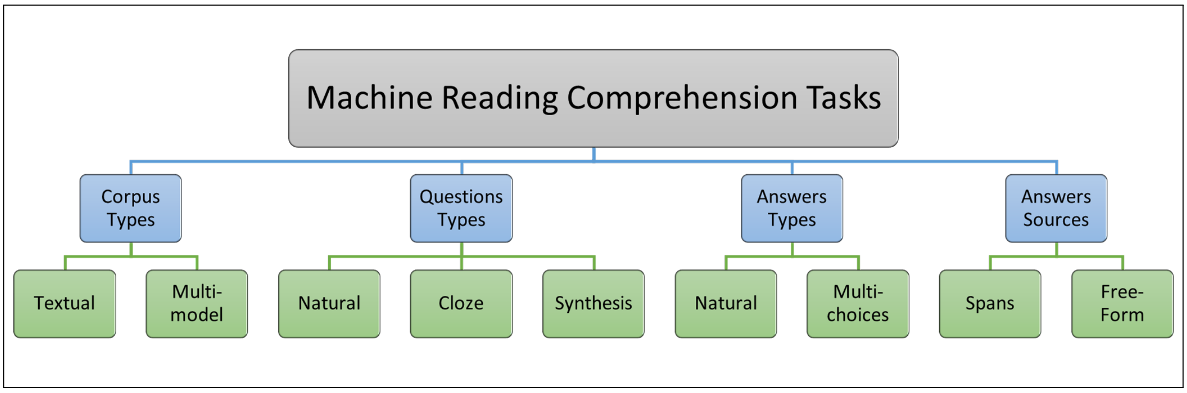


Figure 2.2 Reading Comprehension Tasks. Based on [82]

metrics. The widely used metrics are listed below. Other metrics will be discussed in detail in the next chapter.

1. Accuracy is the number of relevant items retrieved and the number of irrelevant items that are not retrieved divided by the number of all items.
2. BLEU stands for **bil**ingual **e**valuation **u**nderstudy. This method is commonly used to evaluate machine translation systems. It is a precision-based metric that uses a weighted average of variable length phrase matches (n-grams) against the reference sentence. It is commonly used for RC.

Accuracy is the vital aspect that has been used in each research to measure the performance of a system. However, the rest of the criteria are listed below with an explanation for each.

- Interpretability: This refers to the transparency and understandability of the model's predictions or output, especially in the context of questions with multiple constraints or complex questions. So, interpretability is the ability to track the model's thought process in case of wrong predictions, which improves DL models for NLU [85].
- Speed and Scalability: The efficiency and speed of the system can be measured by the time of training and the size of the parameters as well as the hyperparameters. According to [74], most recent systems focus on improving accuracy regardless of computational costs or efficiency of a system. The scalability of the system indicates the consideration of the broader context. Furthermore, it is necessary to consider the largest dataset on which the system can work without losing accuracy and efficiency.

- Robustness: How can systems generalise to other datasets and settings beyond the training distribution? This basically depends on the computational costs of the methodology.

- Data Set Creation: What are effective methods for building new datasets? How can a dataset be built in a different language (Arabic)?

- Dataset Analysis: What challenges do current datasets pose? For example, multi-sentence reasoning is required in NewsQA. Moreover, there is a vocabulary problem that is raised with KB-QA when the KB has limited vocabulary coverage. In fact, most recent research studies have tried to minimise it.

- Error analysis: What types of questions or documents are particularly challenging for existing systems? What types of question are unsolved? In the TCU task, which category poses the greatest challenge? Moreover, can the model accurately determine the appropriate time unit for a given duration, regardless of the specific number? For instance, if the model predicts that a vacation lasts for days, but the answer is stated as 1000 days, which is illogical, does the model fail to validate the number solely because the unit is days?

2.5 Arabic Language

Arabic, a Semitic language with over 400 million speakers, holds immense cultural, religious, and geopolitical importance [23, 30]. It is the religious language of Islam, spoken by Muslims worldwide, and is one of the six official languages of the United Nations [23, 30]. Arabic is also one of the oldest languages, with a history spanning over fifteen centuries.

Arabic has an extensive vocabulary with many words that can convey subtle nuances. For example, Arabic has numerous words for concepts like love, each with slightly different meanings and connotations. Addtionally, Arabic language boasts a wealth of synonyms and antonyms, providing speakers with a wide range of expressions to choose from.

The Arabic language can be divided into three primary categories [23, 30, 20], listed below:

- **Classical Arabic:** The language of the Quran, unchanged for over fifteen centuries, representing the historical and literary heritage of the Arab world.

- **Modern Standard Arabic (MSA):** An evolving variety of Classical Arabic used in contemporary media, education, and formal communication. MSA adapts to modern needs by incorporating new borrowings and innovations.

- **Arabic Dialects:** Arabic encompasses numerous dialects that can be categorized by region, country, or even city [2]. These dialects vary significantly, often to the point of mutual unintelligibility, reflecting the rich linguistic tapestry influenced by historical, geographical, and cultural factors.

Main Features and Complexity:

- **Morphological Structure:** Arabic features a rich system of prefixes, suffixes, and infixes, making it a heavily inflected language in which words change form based on their syntactic role.
- **Diglossia:** A complex diglossic situation where MSA is used in formal settings, while numerous regional dialects are used in daily communication. Diglossia involves the use of two distinct varieties of a language within a single language community, with MSA as the high variety and regional dialects as the low variety [20].
- **Code-Switching:** Frequent switching between MSA, dialects, and other languages like French, English, Spanish, and Italian, depending on the country and context. Code-switching refers to alternating between two or more languages or dialects within a conversation or even a single utterance.[20].
- **Ambiguity and Resource Challenges:** The lack of standard orthographies for dialects and the underspecified short vowels in written Arabic add to its linguistic ambiguity [20, 2].

Comparison to English: In comparison to English, Arabic is significantly different in several key aspects.

- Arabic is a Semitic language, characterized by its root-based morphology, where most words are formed from a set of three-letter roots that convey a core meaning. Languages that share similarities with Arabic include other Semitic languages, such as Hebrew and Amharic.
- Arabic script is written from right to left and includes 28 letters, which can change shape depending on their position within a word. Additionally, Arabic has a complex system of verb conjugation and noun declension [30], which is more extensive than in English.

While English is also a rich language, particularly due to its large vocabulary and capacity for creating new words, the richness of Arabic is often seen in its structural and morphological complexity. The richness of English comes from its flexibility, simplicity in structure, and extensive borrowing from other languages, which makes it adaptable and globally influential.

To summarize this section, Arabic's linguistic richness, cultural significance, and complexity present unique challenges and opportunities in natural language processing. The language's long history and continued relevance highlight its importance on the global stage.

2.6 Temporal Text

In order to gain a comprehensive understanding of temporal language, it is essential to have a clear grasp of its three fundamental components: events, time, and temporal relations [18]. A detailed definition will be provided for each of these components. Firstly, the concept of an event can be defined based on the Oxford English Dictionary as "a thing that happens or takes place, especially one of importance". Therefore, an event refers to any kind of occurrence or happening, whether it is a physical action, a mental process, or a change in state. Time, on the other hand, denotes the temporal framework in which these events occur and includes various temporal markers such as dates, hours, minutes, and seconds. There are different types of time expression [18].

- Absolute: The time is explicitly stated in the text, such as Sunday, February 11, 2018.
- Deictic: The period depends mainly on the time of speech or document creation. The phrase "three weeks ago" in a document can be determined by the time or date it was written.
- Anaphoric: The period is based on the time point mentioned in the text. Anaphoric temporal expression splits into three fragments: distance (3 months), direction (future or past), and anchor. For example, Ahmad arrived in Riyadh on 12 May, Fahad the day before. Therefore, Fahad's arrival date is contingent upon Ahamd's arrival, occurring one day prior.
- Duration: It describes an interval of time and how long the events take, for example, nine months.
- Set: Declares the regularly periodic times, for example, every Tuesday.

Finally, temporal relations describe the way in which events are related to one another in time, such as their sequence, duration, simultaneity, or causality. Temporal relations are categorized into three main categories [60]:

- Temporal Relation (TLINK): Represent the temporal relationship between two events, an event and a time or two times. For example: She submitted the report last week.
- Subordinate (SLINK): Used for modality, evidential, and factual realtions. For example: She refused to submit the report.
- Aspectual (ALINK): Only between two events, describing an aspectual connection. For example: She finished writing the report.

Temporal text understanding tasks are critical for many NLP applications, including information extraction, question answering, and summarization. The development of these tasks has progressed from rule-based systems to sophisticated deep learning models, significantly improving accuracy and scalability. Key temporal text understanding tasks include: temporal annotation, temporal normalization, event extraction and temporal relation extraction.

For automatic annotation of temporal expression in text, the most widely used method is TimeML[4] which was introduced in 2002. It is an XML-style markup for temporal information in natural language texts and has become an ISO standard. The four types of temporal expressions captured in the TIMEX3 (TimeML) are time, date, duration, and set. TimeML annotates temporal relations based on the three types described in the previous paragraph.

In order to evaluate the temporal expression annotation, precision and recall are used, as well as accuracy [18]. These metrics, along with their detailed explanations, are provided in Section 4 of this chapter.

2.7 Temporal Reasoning

Temporal reasoning involves deducing new relationships between events based on existing knowledge and their temporal connections. It is generally categorized into two main types based on how time is represented [27]:

- **Quantitative Temporal Reasoning:** This approach relies on precise, numerical time labels (or timestamps) to understand events. For example, it uses specific times like "event A starts at 01:31 and ends at 02:03" to make conclusions about temporal relationships.
- **Qualitative Temporal Reasoning:** This approach focuses on the relative timing of events rather than exact timestamps. It uses relative terms like "event A occurs before

[4] http://www.timeml.org

event B" or "event C happens during event B" to draw inferences about the sequence and overlap of events.

Allen’s Interval Algebra and Qualitative Spatio-Temporal Reasoning fall under the category of Qualitative Temporal Reasoning, as described below:

Allen’s Interval Algebra

Allen’s Interval Algebra is of the foundational formalisms in logical temporal. Allen’s interval algebra (IA), proposed by James F. Allen in 1983, provides a framework for representing and reasoning about temporal intervals and their relationships, for English [3]. It defines 13 basic relations that can exist between two time intervals, such as "Before," "After," "Meets," "Overlaps," and "During." These temporal relations help in understanding how different events are positioned relative to each other in time. Figure 2.3 demonstrates the basic relations and depicts how two events, x and y, are interconnected based on Allen’s framework. Real-line intervals are used to represent entities and encode potential relationships, with an emphasis on the qualitative relationships between time intervals.

Qualitative Spatio-Temporal Reasoning

Qualitative Spatio-Temporal Reasoning (QSTR) involves reasoning about spatial and temporal relationships using qualitative terms rather than precise measurements [70]. It aims to understand and infer spatial and temporal relationships in environments where numerical data may be sparse or unavailable. QSTR provides a framework with abstract terms such as ’before’, ’after’, ’near’ and ’far’. As a constraint language for qualitative time representation and reasoning, Allen [3] proposed Interval Algebra (IA) [70]. on the other hand, for the spatial, The Region Connection Calculus (RCC), as proposed by Randell et al. in [63], is a tool for mereotopological space reasoning and qualitative space representation. In particular, this theory is based on the primitive relation of connection and takes into account regions in any arbitrary topological space along with potential relationships among them [70].

In practical scenarios, pinpointing the exact start and end times of events can be challenging [27]. Instead, one might assess that event A occurs before event B with a probability or degree of certainty, such as 0.7. This estimation reflects the inherent uncertainty in defining temporal relationships. In the context of Temporal Common Sense Understanding (TCU), handling such uncertainties is crucial. TCU aims to model and infer temporal relationships between events, often dealing with imprecise or vague temporal information. Deep learning models play a significant role in this process by learning from large datasets to predict and understand temporal relationships with varying degrees of certainty. These models can be

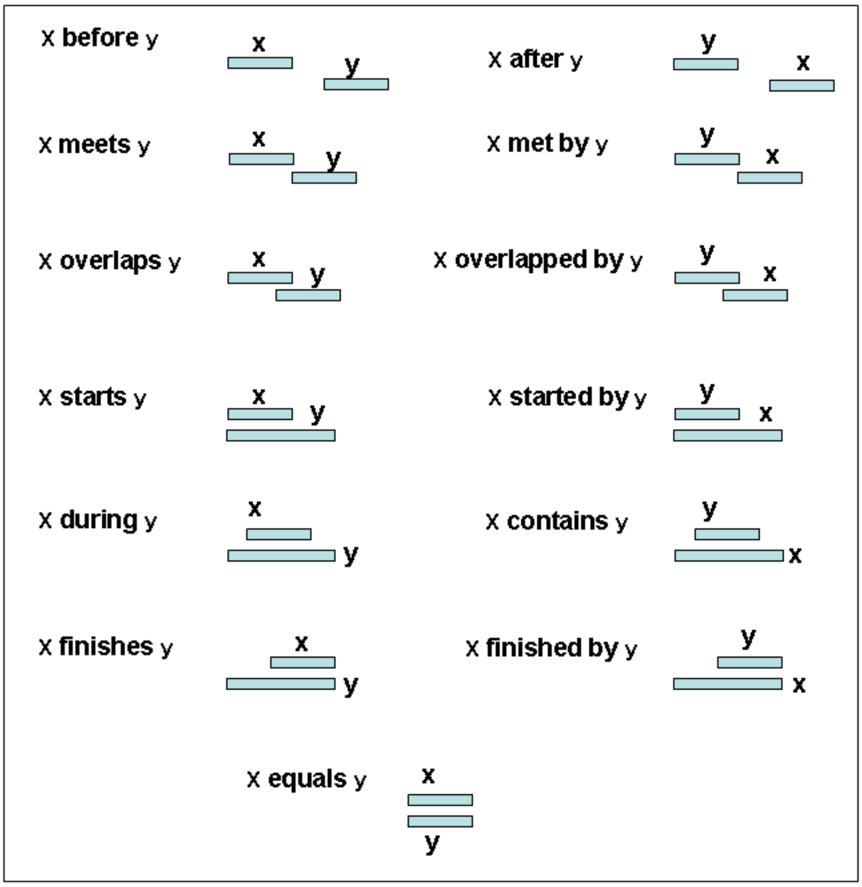


Figure 2.3 Allens Temporal Logic 13 Base Relations. Source [22]

trained to handle such probabilistic estimates and integrate them into their temporal reasoning capabilities, improving their performance in real-world scenarios where exact temporal data may not be available.

2.7.1 Logical Temporal Reasoning

Logical inference involving events in time and space might theoretically follow the same principles as other logical inference [27]. However, practical experience reveals that applying standard logical methods to temporal reasoning often leads to significant inefficiencies [27]. As, classical logic operates under the assumption of a static world where formulas have fixed truth values [27]. In contrast, real-world systems are dynamic, and temporal statements can change their truth values over time.

There are two types of tasks related to event temporal reasoning: information extraction (IE) and machine reading comprehension (MRC). Information extraction encompasses event extraction as well as relation extraction. In contrast, MRC focuses on selecting the correct answer for a temporally related quuestion.The scope of this thesis is MRC. The MRC task has been thoroughly described earlier in section 2.4.

2.8 Arabic Temporal Text

Constructing a robust framework to annotate temporal information in Arabic texts is essential for the enhancement of different NLP applications for Arabic content. TimeML standards for the Arabic language have been constructed by Haffar et al.[31], addressing the linguistic and cultural considerations. There are multiple efforts to improve the annotating standard for Arabic temporal text [31].

2.8.1 Challenges of Arabic Temporal Text Understanding

The difference between how Arabic and English express temporal information can make it challenging to compare temporal common sense understanding in the two languages. The Arabic language is known for its richness and complexity. Below are some of the challenges that machines may face in understanding Arabic temporal text.

- Arabic is a complex language in which diacritics represent short vowels, but in MSA they are often omitted. This lack of diacritics causes numerous ambiguities [30, 13]. For instance, the same word "ذهب" without diacritics, can have these two different meanings: “gold” if it is diacritised “ذَهْب” or go-went if it is diacritised ”ذَهَب” [13]. This issue, which leads to ambiguity, is not present in English as English does not use diacritics.

- Additionally, an Arabic date can be represented using either the Gregorian calendar, the Hijri calendar or both simultaneously. The Hijri calendar, also known as the Islamic calendar, is a lunar calendar that includes 12 months in a year with either 354 or 355 days. There are various methods to represent the Gregorian month names in Arabic, including using phonetically correct English or Arabic names [12].For example, "January" can be written as "يناير" (Janāyer) or phonetically as "جانيواري" (Jānyuwārī).

- Another challenge arises from the dual usage of Hijri month names as personal names [12]. The names Rajab, Shaaban, and Ramadan an refer to either a month or a person. Additionally the Eid, which is an Isalmic holiday can also refer to a person name. For example, in the provided sentence, the term "رمضان" is open to interpretation as it can denote either an individual’s name or the Islamic month of Ramadan: "The family is happy with the arrival of Ramadan"الأسرة سعيدة بدخول رمضان.

- Another difference between Arabic and English is the use of temporal adverbs. In this aspect, Arabic has a wider variety of temporal adverbs than English. For example, the Arabic adverb "قبل" (‘before’) can refer to events which happened before the present moment, events which occurred before a specific time, or events which happened before another event. For instance:

  - ذهبت إلى المدرسة قبل الساعة الثامنة, ("I went to school before 8 o’clock")
  - أنهى واجبه قبل أن يصل والده, ("He finished his homework before his father arrived")
  - كنت أعمل في الشركة قبل ثلاث سنوات, ("I was working at the company three years ago")
  - قرأت هذا الكتاب قبل عدة أشهر, ("I read this book several months ago")

  While the English adverb "before" can also order two events in the past or present, such as "I went to school before 8 o’clock" and "He finished his homework before his father arrived," it does not encapsulate all the nuances and contexts that "قبل" can in Arabic. For example, "before" does not as naturally express time periods without additional context, such as "I was working at the company three years ago" or "I read this book several months ago," where Arabic can use "قبل" directly to convey these time frames.

  As previously mentioned, Allen’s temporal relations provide a formal structure to understand these nuances. For instance, the "Before" relation in Allen’s framework can map directly to sentences like "I went to school before 8 o’clock." However, the richness of the Arabic adverb "قبل" extends to various temporal contexts, showing intervals spanning past periods. This highlights the complexity and depth of temporal understanding required for accurate natural language processing in different languages.

Although English also shows ambiguity, it appears in a different way because of its sequential morphology and spelling patterns. An example of ambiguity in English is that identical words can possess multiple meanings.

2.9 Temporal Question Answering System

According to the literature, most of the studies assessing machines’ ability to understand temporal text have been applied within the context of Question Answering (QA) systems.

Researchers have tackled different types of QA including reading comprehension and knowledge base QA systems. Temporal question answering means the ability to answer any temporal-based question. This encapsulates extracting the temporal information and requires some reasoning. Various questions can be made regarding temporal aspects, such as the time at which an event occurs or its frequency. Additionally, one may inquire about the duration of the event, the temporal relationship between events, or the order in which they occur. Questions that commence with terms such as "when," "how long," "how often," or "what happened after or before a specific event" are categorized astemporal questions.

Challenges Of Temporal Question Answering Temporal reasoning is challenging, and some questions require reasoning. Temporal reasoning is also complicated because some events are vague [67]. Meng et al. [50] proposed using Long Short-Term Memory (LSTM) networks to enhance temporal relation extraction by leveraging the shortest dependency path between entities in a sentence. Their approach involves applying a double-checking mechanism to ensure the accuracy of the extracted dependencies and using pruning techniques to remove irrelevant information. By focusing on the shortest dependency path, their method effectively captures the most relevant syntactic relations, leading to more accurate temporal relation extraction at that time. The task is extracting temporal relations from text, including (1) event extraction and (2) TLINK classification (Intra-sentence, cross-sentence, DCT (Document Creation Time), relation between TimeEXes). In order to evaluate the proposed methodology, Meng et al. [50] used two types of evaluation: intrinsic and extrinsic. Intrinsic evaluation measures the performance of the model on a specific, isolated task, such as the accuracy of temporal relation extraction. Extrinsic evaluation, on the other hand, assesses the impact of the model's performance on a broader application or end task, such as its effect on an overall NLP system. For the extrinsic evaluation, yes/no questions were used. This resulted in achieving the state of the art by a large margin for QA. This evaluation is about the accuracy for answering the targeted questions.

Jia et al.[40] has introduced TemQuestions dataset. The dataset has been extracted from three knowledge base question-answering (KB-QA) datasets: Free917 [15], WebQuestions [9] , and ComplexQuestions [7]. The answer sets for these datasets are based on Freebase, a large collaborative knowledge base that was created to enable users to build structured, searchable databases. Although Freebase is no longer maintained and its data has been migrated to Wikidata[5], it has historically been widely used as a source of ground truth for training and evaluating QA systems. The release of these datasets has been followed by an implementation of the Temporal QA system called TEQUILA by Jia et al.[40]. The main

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

limitation of TEQUILA is that it is a rule-based approach. TEQUILA has four stages: First, it detects temporal questions. Then, it decomposes and rewrites the question into subquestions and temporal constraints. After that, it retrieves the answers for the subquestions from the underlying KB-QA engine. Finally, it uses constraint reasoning on temporal intervals to compute the final answers to the main question.

2.10 Common Sense Understanding

Commonsense reasoning, which relies on world knowledge encompassing spatial and temporal relations, physical laws, causation, and social norms, is part of human intelligence [65]. Yet, embedding this type of common sense reasoning into artificial intelligence (AI) systems, poses a significant challenge. Although neural networks excel at learning patterns from extensive datasets, the essence of common sense reasoning for humans often transcends the need for such examples. Instead, humans acquire the knowledge necessary for common sense reasoning through everyday experiences and interactions with the world around them.

This domain has captivated numerous researchers within the AI community, leading to a significant surge in studies focused on integrating common sense reasoning into AI systems. The growing interest underscores the field's potential to bridge the gap between human and machine intelligence, heralding a new era of AI capabilities that mirror the depth and complexity of human thought [65].

The studies can be categorized into different types:

- Common sense representations.
- Analysing the PLMs in terms of the common sense understanding.
- Incorporate common sense knowledge into PLMs.

Knowledge Graphs (KG) or Knowledge Bases (KB) are used in Natural Language Understanding (NLU) works to enhance model performance. In particular, large knowledge bases (KBs) like ConceptNet [61, 62] is incorporated into the architectures of common sense reasoning works [64, 65, 66]. The following list emphasizes the knowledge bases particularly utilized for common sense reasoning, as well as those employed in temporal reasoning.

- **ConceptNet [71]:** Created by the MIT Media Lab in 2017 by Speer et al. [71]. ConceptNet is a semantic network created with the purpose of aiding in the comprehension of word meanings by computers, as utilized by humans. ConceptNet has been extensively used in common sense reasoning studies such as Bian et al. [10], where it was integrated with BERT for the common sense question answering task using the CommonsenseQA dataset [72], resulting in improved context understanding and reasoning capabilities. ConceptNet has also been utilized for temporal reasoning by providing relational knowledge about events and time.

- **ATOMIC [64]:** Developed by the Allen Institute for AI, ATOMIC is a large-scale common sense knowledge graph that provides inferential knowledge about everyday situations. Wang et al. [77] employed ATOMIC in various common sense question-answering datasets and suggested a method for data augmentation. The findings surpass the studies that were conducted up to that point.
- **COMET (Commonsense Transformers) [11]:** Based on ATOMIC and ConceptNet, COMET was developed by the Allen Institute for AI to generate common sense knowledge for NLP models. COMET was applied to generate common sense knowledge, significantly improving the performance of models on common sense reasoning tasks such as story completion and next event prediction [11].

2.11 Temporal Common Sense Understanding

Temporal common sense understanding refers to the task where the model must comprehend texts that encapsulate temporal information based on real-life knowledge not provided in the context. The temporal information can be categorized into the following types: event duration, event ordering, the typical time of the event, frequency of the event, and the stationarity of the event [83]. The list below explains each category:

- **Event Duration:** This refers to the length of time an event lasts. For example, "The conference lasted for three days" or "The movie runs for 120 minutes." Understanding the duration of the event helps to schedule and manage tasks.
- **Event Ordering:** This involves the sequence in which events occur. For instance, "She finished her homework before going to bed" or "The meeting was scheduled after lunch." Event ordering is crucial to understanding causality and planning.
- **Typical Time of the Event:** This category specifies the common or expected time at which an event usually occurs. For example, "Breakfast is typically served at 7 AM" or "The annual festival is held in October." Knowing the typical time of events helps to predict and organise activities.

- **Frequency of the Event:** This describes how often an event happens. Examples include "He goes jogging every morning" or "The company releases quarterly reports." Frequency information is important for routine planning and resource allocation.
- **Stationarity of the Event:** This pertains to whether an event remains consistent or lasts over a long term. For instance, "The store opens at 9 AM every day" indicates a stationary event, whereas "The store hours change every season" indicates a non-stationary event. Another example could be "The climate in this region has been stable for decades" versus "The project's timeline is frequently adjusted." Stationarity helps in understanding the stability and predictability of events over time.

There are a few studies in this field. Most of the studies are in English. The following chapter covers the list of datasets, specifically focusing on temporal reasoning and temporal common sense understanding. In this section, the studies that have been conducted in this area will be covered. Building on the research in common sense understanding, several studies have incorporated external knowledge bases to boost the temporal common sense comprehension of models. For instance, Gosal et al. [25] utilized ATOMIC and CONCEPT-NET to enhance the temporal common sense understanding (TCU), resulting in improved sentence ordering across multiple datasets.

Limited research has endeavored to construct language models that are especially tailored for TCU. This idea originated from the observation that the complexity of TCU tasks was not sufficiently addressed by the PLMs. It was discovered by researchers that PLMs frequently had difficulty identifying and acquiring temporal dimensions, which made explicit temporal data training models necessary. The following are instances of such initiatives, which seek to close this gap by enhancing language models' capacity for temporal reasoning through the use of specialized training methodologies and datasets.

- **TACOLM [84].:** The shortcomings of PLMs in handling the TCU task are highlighted by this study, especially in terms of their inability to detect and incorporate temporal dimensions. In order to bridge this gap, the researchers suggest adding a second pre-training phase that is intended to provide temporal-related data to the models. Two approaches were used to create the dataset for this enhanced pre-training phase: first, a large corpus of explicit mentions of TCS was created by designing syntactic rules to gather a significant amount of data from an unannotated corpus. Furthermore, a joint learning scheme was proposed to tackle the problem of reporting biases in common sense information extraction. These methods concentrate on the temporal reasoning abilities of the model. The duration, frequency, and typical time of TCS inference are three significant dimensions that are examined in this paper [84].

- **ECONET [32]:** This work aimed to develop temporal language models for the purpose of improving event sequencing tasks. The continuous training methodology used in this work is inspired by ELECTRA [16], with a particular focus on applying a focused masking technique to direct the model's attention to temporal elements. In the context of ECONET, this targeted masking strategy involves masking some temporal terms and related contextual terms so that the model can focus on temporal relationships during training. This method enhances the model's ability to identify and understand the sequence of events within a narrative [32].

- **A third language model:** [45] proposed for the TCU task also utilizes a continual training approach but introduces a different target masking strategy and employs various temporal-related datasets. It's important to note that this study did not construct its dataset but instead utilized available temporal datasets, which distinguishes its methodology from others [45]. This study's advantage over TACOLM and ECONET lies in its comprehensive coverage of all temporal dimensions. Additionally, it utilizes pre-existing temporal-related datasets instead of creating a new one.

Furthermore, the study by Virgo et al. [76] demonstrates that the recent PLMs have yet to reach the level of human performance in this particular task which is Event Duration task. Due to the limited training data, which only covers a finite number of events and their associated attributes. The study emphasize the crucial demand to incorporate external event duration information to enhance this task's effectiveness. Thus, a new QA dataset for event duration from an existing dataset have been constructed. Then, this dataset is used for the intermediate task in the adaptive fine tuning approach. While Kimura et al. [45] used the existing dataset as it is. Also, In study of virgo et al. [76], the focus is only on the event duration, while [45] studied all Temporal Understanding aspects.

Through the exploration of alternative training methodologies and the construction of specialized datasets, these studies seek to boost the models' proficiency in grasping and processing temporal information. Despite the challenges associated with building language models, nor constrcting dataset, the outcomes from all these proposed studies still fall short of the performance levels achieved by more advanced PLMs.

Furthermore, several adaptive fine-tuning techniques have been explored for the English MC-TACO dataset ([44] , [43]) and adversarial fine-tuning [56]. However, according to the leaderboard [6], these techniques perform worse than the DeBERTa-Large model [33], which utilizes the vanilla fine-tuning paradigm.

[6] https://leaderboard.allenai.org/mctaco/submissions/public

Some attempted to improve the performance by applying a preprocessing for the dataset. By analysing the outcome of the PLMs and finding that some errors could be mitigated by normalising the temporal feature. Two primary normalization techniques have been applied:

- Unit normalization is a process that extracts temporal expressions from candidate answers and converts them into standardized units [83]. For instance, the expression "30 months" would be transformed into "2.5 years." This normalization ensures that different temporal expressions are uniformly represented, facilitating more accurate comparisons and understanding by the model. Another example includes converting "7 days" to "1 week," or "60 minutes" to "1 hour." By standardizing these temporal expressions, the model can more effectively comprehend and process temporal information, leading to improved performance in tasks involving temporal common sense understanding.

- Duration normalization, particularly within the "Event Duration" category, often deals with the challenge that the exact duration of an event is not always precise but rather falls within a range [41]. For example, instead of stating that a meeting lasted exactly "60 minutes," it might be more accurate to say it lasted "about an hour" or "between 50 and 70 minutes." Similarly, the duration of a vacation could be described as "approximately two weeks" or "10 to 15 days." This approach acknowledges the inherent uncertainty and variability in event durations, allowing models to better handle and interpret temporal information by considering intervals rather than fixed values.

Unit normalization have used with BERT [83]. While in another study, time normalization, which includes unit and duration normalization, has been applied with T5 [41]. T5 is encoder - decoder architecture not like BERT [62]. Despite the fact that the margin of enhancement was not promising. The normalization is language dependant and also it is based on hand-coded rules.

All of the suggested techniques have been surpassed by DeBERTa-v3. Yet, DeBERTa's performance still falls significantly short of human performance on the same task. This gap emphasises the complexity of temporal reasoning in NLU and the ongoing need for research to refine and enhance the capabilities of language models in this critical area.

2.12 Summary

This chapter provides an overview of the essential concepts addressed in the thesis. The discussion begins by delving into the temporal features and the difficulties associated with temporal understanding in both Arabic and English, offering a comprehensive analysis. Furthermore, the evaluation of temporal understanding involves the utilization of a reading comprehension dataset, which is explained in this section. Additionally, the chapter explores the reading comprehension task, shedding light on its intricacies. In terms of the models employed in the thesis, a thorough explanation is provided for all the deep learning models that have been adopted. Lastly, the chapter offers a detailed presentation of the PLMs. To conclude, the selection of the model depends on the desired performance, the task's complexity, and the computational resources that are accessible.

Chapter 3

Constructing Arabic Temporal Common Sense Dataset

3.1 Introduction

Temporal Common Sense Understanding (TCU) is an integral component of the broader endeavor to comprehend common sense in natural language. Despite its importance, the availability of resources dedicated to this aspect in English is limited, and remarkably, there is an absence of datasets in Arabic tailored to this specific domain. Currently, the only dataset in Arabic that touches upon common sense understanding is essentially a translation from English, focusing on common sense validation [73]. This gap significantly limits the ability to effectively evaluate the performance of transformer-based models in grasping Arabic temporal common sense. For the objectives of this study, namely assessing and improving the efficacy of transformer-based models in Arabic TCU, it becomes imperative to construct a dataset.

Furthermore, the construction of an Arabic TCU dataset represents a crucial step forward. Such a dataset would not only facilitate the specific evaluation of transformer-based models in the Arabic linguistic context but also contribute significantly to the field by providing a foundational resource for future research endeavors. The construction of this dataset, while challenging, promises to be a valuable addition to Arabic NLU resources, enabling more nuanced and culturally relevant understanding of temporal common sense in Arabic.

Dataset	**Temporal Feature**	**Size**	**Task Description**
MATRES	Event Ordering	20 K	Classification
UDST-DurationQA	Event Duration	6 K	Classification
TimeDial	Typical Time	1.1 K	Multiple Classification
TNLI	Stationarity	10.7 k	Natural Language Inference
TimeQA	Typical Time	15 K	Question answering

Table 3.1 List of the Available Datasets

3.2 Temporal Related Datasets

Temporal Common Sense Datasets

This section presents a comprehensive list of datasets currently available for temporal common sense understanding in English. The list comprises different datasets which vary based on their main temporal features or the specific tasks for which they are designed.

- The MC-TACO dataset constructed in 2019 by Zhou et al. [83] covers all the temporal features. Further details about this dataset will be provided later in this chapter.

Table 3.1 lists and compares all the available temporal datasets.

3.3 Dataset Construction

The construction of a dataset from scratch is a notably resource-intensive task. This challenge is compounded in the context of Arabic, where there is a conspicuous absence of temporal-related datasets and a general scarcity of resources. Given these constraints, the decision to adapt an existing dataset from English to Arabic was motivated by both practicality and the unique requirements of our research focus.

The MC-TACO dataset [83] was selected for translation into Arabic due to several key factors that align with the research objectives. Firstly, MC-TACO is recognized,to the best of current knowledge, as the only dataset encompassing a comprehensive range of temporal characteristics, making it exceptionally relevant for our study on TCU. Secondly, the dataset's straightforward structure and use of simple sentences render it particularly amenable to translation, ensuring the preservation of semantic integrity in the process. MC-TACO was designed as a multiple-choice reading comprehension (MRC) task. The input of the model from the dataset consists of three components: an abstract or context, a question, and a corresponding answer. The model requires the output of a prediction score based on judgment of the plausibility value of the answer. The score should be close to one if the candidate answer is valid. The relatively concise nature of the information provided in the dataset, often encapsulated in three sentences, makes it feasible to employ a translation tool for the initial construction process. Google Translate was utilized for this purpose, with subsequent translations subjected to a thorough review by two native Arabic speakers specialized in proofreading to ensure accuracy and natural language use. The reviewers, who examined all the inputs individually, were from different Arabic countries, specifically Saudi Arabia and Morocco, as cultural differences might affect the understanding of the translated results. Finally, the overall results were reviewed to ensure consistency and accuracy.

The dataset encompasses approximately 13K question-answer pairs, spanning five temporal dimensions, thereby offering a rich resource for exploring various aspects of temporal reasoning. The temporal dimensions that are included in MC-TACO are explained in the list below:

1. **Event duration:** How long does an event last?
2. **Temporal ordering:** Typical order of events.
3. **Typical time:** When did an event occur?
4. **Frequency:** How often do events occur?
5. **Stationarity:** whether a state is maintained long-term or indefinitely.

Table 3.2 presents the statistical information for both the English and Arabic versions of the dataset. Table 3.3 presents statistics for temporal features. Furthermore, Figure 3.1 illustrates the distribution of question-answer pairs across different temporal aspects. The dataset predominantly consists of question-answer pairs related to event duration, with frequency being the second most common aspect. On the other hand, the coverage of the stationarity feature is notably low, comprising only 870 pairs (7%).

A sample of the data set is presented in Figure 3.4. This figure provides a comprehensive overview of different temporal categories, illustrating an example of each one. Additionally, this figure includes a question along with its corresponding set of answers. The correct answers are highlighted in bold.

Measures	**Arabic**	**English**
# of unique questions	1893	1893
# of unique question-answer pairs	13,225	13,225
avg. context length	15.2	17.8
avg. question length	6.5	8.2
avg. answer length	3	3.3

Table 3.2 Dataset Statistics

Category	# Unique Context	# Unique Questions	Avg. # of Candidates
Event Duration	135	440	9.4
Event Ordering	26	370	5.4
Frequency	229	433	8.5
Typical Time	43	371	6.8
Stationarity	73	279	3.1

Table 3.3 Temporal Category Statistics

3.4 Evaluation

Evaluating the state of the art in reading comprehension question answering can be approached from various angles. Accuracy has been the vital aspect used in each research project to measure system performance. Reading comprehension question-answering systems can be evaluated according to different metrics, especially those listed below:

1. Accuracy is a metric that measures how often the model correctly predicts the class label.

$$\text{Accuracy } = \frac{\text{Correct predictions}}{\text{All predictions}} \tag{3.1}$$

$$\text{Accuracy } = \frac{\text{True Positive } + \text{ True Negative}}{\text{Total Number of Predictions}} \tag{3.2}$$

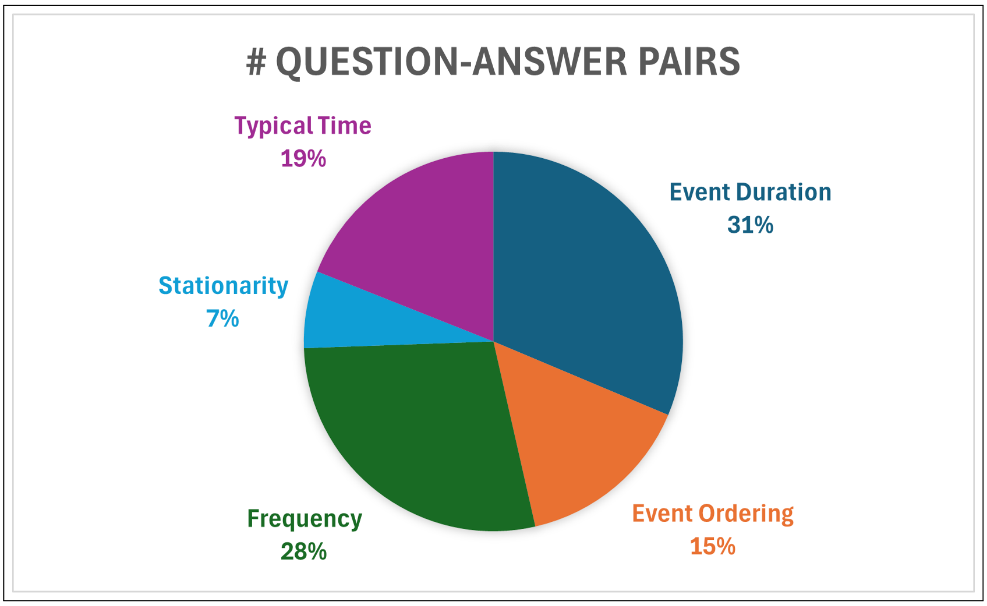


Figure 3.1 Percentage of Unique Question-Answer Pairs in each Temporal Category

2. Precision is known as a positive predictive value. This measure is the percentage of correctly predicted positives out of total positive predictions.

$$\text{Precision} \ = \frac{\text{True Positive}}{\text{True Positive} \ + \ \text{False Positive}} \tag{3.3}$$

3. Recall is considered the sensitivity or true positive rate, the percentage of correctly predicted positives out of the total number of actual positive examples.

$$\text{Recall} \ = \frac{\text{True Positive}}{\text{True Positive} \ + \ \text{False} \ \ \text{Negative}} \tag{3.4}$$

4. The F measure is a single metric that trades precision for recall. This factor is the weighted harmonic mean of precision and recall, where weight is denoted by a variable β. The default balanced F measure where β= 1 is commonly written as $F1$, which is short for $F_{\beta=1}$.

$$F = \frac{(1+\beta^2)*P*R}{(\beta^2*P)+R} \tag{3.5}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \ * \ \text{Recall}}{\text{Precision} \ + \ \text{Recall}} \tag{3.6}$$

5. Exact Match is a strict version of accuracy where all labels must match exactly for the sample to be correctly classified. For MC-TACO, the model must correctly predict all

answers to each question to be considered a correct prediction.

$$\text{Exact Match} = \frac{\text{Total Number of Questions that are Predicted Correctly}}{\text{Total Number of Questions}} \tag{3.7}$$

6. The standard error (SE)[1] is a statistical measure of the reliability and stability of the model's performance. When the model is trained multiple times with different seeds, the SE quantifies the degree to which the model's performance differs from one run to another due to the inherent randomness caused by the initialisation and data shuffling processes. A lower SE indicates that the model's performance is consistent and stable across different runs, indicating that the model is more robust and reliable. The SE provides a statistical basis for the precision of the average performance metric reported from the runs, enabling more model evaluation.

$$\text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{N}} \tag{3.8}$$

   Where N represents the number of values, which, in this case, corresponds to the number of runs.

Notably, although the dataset can be viewed as a binary classification task, where accuracy is a commonly used metric, accuracy may not be the most appropriate metric in this case. The distribution of labels for the candidate's responses is approximately one 'no' to two 'yesses', implying that a high level of accuracy (or a low error rate) can be achieved even by a model without real skill that simply predicts the majority class. Consequently, accuracy can be a misleading metric in this type of dataset. Therefore, based on [83] the two metrics that are used are: F1 score and the exact match.

3.5 Dataset Split

The dataset split, which is derived from the English dataset[83], consists of a set of development with 561 unique questions and 3,783 candidate responses. Additionally, the test set comprises 1,332 distinct questions and 9,442 candidate answers. Since there is no separate training split, the development set is utilised for training purposes. Despite the small size of the development set for training, it is hypothesised that PLMs do not necessitate a large dataset for training [83].

[1] https://en.wikipedia.org/wiki/Standard_error

A recent study by Ghosal et al. (2022) revealed that the most efficient strategy for a multiple-choice reading comprehension task is to evaluate each answer individually and determine its correctness [26]. This study utilised the mentioned approach. The decision was influenced by various factors, including the differing number of responses for each question in the dataset. Additionally, all prior research conducted on this dataset employed the same strategy, which [26] confirmed to be effective.

3.6 Summary

This chapter has discussed the creation of the Arabic dataset. Additionally, the chapter has provided a detailed explanation of the dataset statistics, the sample, and the evaluations.

Creating a dataset specifically designed for Arabic can significantly enhance the accuracy and performance of natural language processing models. However, the process of developing such a dataset can be demanding and resource-intensive, making it a costly endeavour for individual researchers.

Table 3.4 Sample of the Dataset.
Each row targets one temporal aspect from the five aspects covered by the original dataset. An example context for each aspect is provided from both the English and Arabic datasets. The English column is from the MC-TACO dataset and includes five different contexts, each representing one aspect. For each context, the question along with all candidate answers is provided, with the correct answers in bold. Note that there may be more than one correct answer for a question, and the number of answers for each question varies. The Arabic column is from the translated dataset.

	Arabic	English
Event Ordering	الفقرة: تدور أحداث فيلم لانق حول رجل يتجول في الشارع ليلة بعد ليلة . السؤال: ماذا يحدث بعد العثور على الرجل وهو يتجول في الشوارع ليلاً؟ • **أوقفت الشرطة الرجل عندما وجدته يتجول في الشوارع ليلاً** • **تم استجوابه من قبل الشرطة** • **تم استجوابه من قبل السلطات** • يأكل وجبة غداء	Context: Lang centers on a man who roams the street night after night . Question: What happens after the man is found to be roaming the streets at night? • **The man is stopped by police when he is found to be roaming the streets at night** • **He is questioned by police** • **He is questioned by authorities** • He eats lunch
Event Duration	الفقرة: توفي والد دورر عام 1502، وتوفيت والدته عام 1513. السؤال: كم من الوقت كانت والدته مريضة؟ • 30 ثانية • 6 قرون • 90 سنة • **6 أشهر** • **سنتين**	Context: Durer's father died in 1502, and his mother died in 1513. Question: How long was his mother ill? • 30 seconds • Six centuries • 90 years • **6 months** • **2 years**
Frequency	الفقرة: ذهب تومي وسوزي (أخ وأخت) إلى الملعب بعد ظهر أحد الأيام مع أمهما وأبيهما، جان ودين. السؤال: كم مرة ذهبوا إلى الملعب؟ • يذهبون للملعب مرتين في الليلة • مرتين في الدقيقة • **مرتان شهرياً** • **مرة في الأسبوع** • **يذهبون إلى الملعب مرة كل بضعة أيام**	Context: Tommy and Suzy (brother and sister) went to the playground one afternoon with their mom and dad, Jan and Dean. Question: How often did they go to the playground? • They go to the playground twice a night • Twice a minute • **Twice a month** • **Once a week** • **They go to the playground once every few days**
Stationarity	الفقرة: لقد كانت القضايا التي تعاملت معها على مر السنين تتعلق بمساعدة الناس في الحفاظ على أساسيات الحياة - المنزل والرعاية الصحية والوظائف والأسرة. السؤال: هل ما زالوا يساعدون الناس؟ • لا توقفوا قبل أسبوع • لا توقفوا بعد دقيقة • **أجل إنهم كذلك**	Context: The issues I've dealt with through the years have been on the side of helping people maintain the basics of life - home, health care, jobs and family Question: Are they still helping people? • No they stopped before a week • No they stopped after a minute • **Yes they are**
Typical Time	الفقرة: على سبيل المثال، ماذا لو وضعت كعكة في الفرن وتركتها لفترة طويلة؟ السؤال: في أي وقت ستضع الكعكة في الفرن؟ • 3 صباحاً • 12 صباحاً • **2 ظهراً** • **3 مساءاً** • **12 ظهراَ**	Context: For example, what if you place a cake in the oven and you leave it in too long? Question: What time would you put the cake in the oven? • 3:00 A.M. • 12 A.M. • **2 P.M.** • **3 P.M.** • **12 P.M.**

Chapter 4

Deep Learning Models for Temporal Common Sense Understanding

4.1 Introduction

This chapter is an exploratory study of the effectiveness of the deep learning models for the TCU task. Prior to the introduction of transformer-based models in 2019, natural language processing (NLP) tasks primarily relied on a wide range of deep learning models involving different architectures and word embeddings. The choice of the most suitable model for a particular task was crucial and depended on the task. Thus, NLP researchers and practitioners had to experiment with various models and embeddings to achieve the desired results. Nevertheless, at the time of writing, transformer-based models have gained popularity in various NLP tasks due to their remarkable performance, making them the preferred choice for most NLP tasks.

4.2 Experiments

The architecture proposed in this study is based on the symmetric architecture proposed by Nicula et al. [53], as shown in Figure 4.1. The proposed network architecture has been modified to effectively capture all relevant information from the three inputs. Specifically, cross-attention has been incorporated to capture the relevant information between the inputs. Additionally, the cosine similarity matrix, which is not applicable to the dataset because there is no similarity between the three inputs, has been removed. Unlike reading comprehension tasks, where the goal is to find similarity between the answer and the context or question, common sense tasks are focused on reasoning rather than similarity. To thoroughly explain the model's behaviour in temporal commonsense understanding, Figure 4.2 illustrates the proposed architecture.

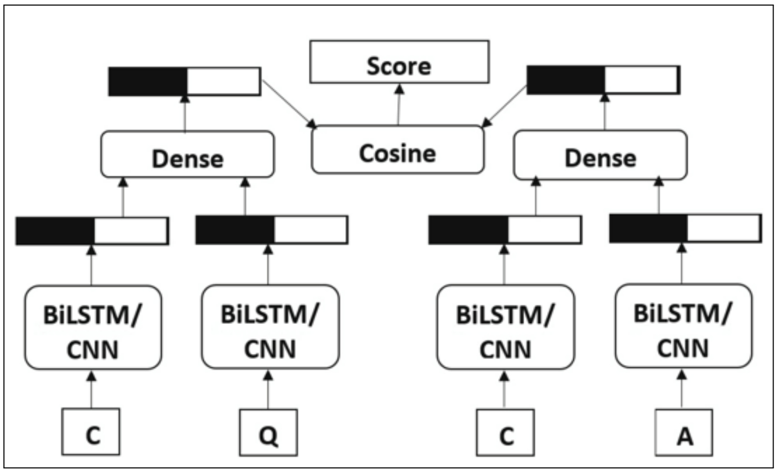


Figure 4.1 Symmetrical Architecture [53]. C: Context, Q: Question, and A: Candidate Answer.

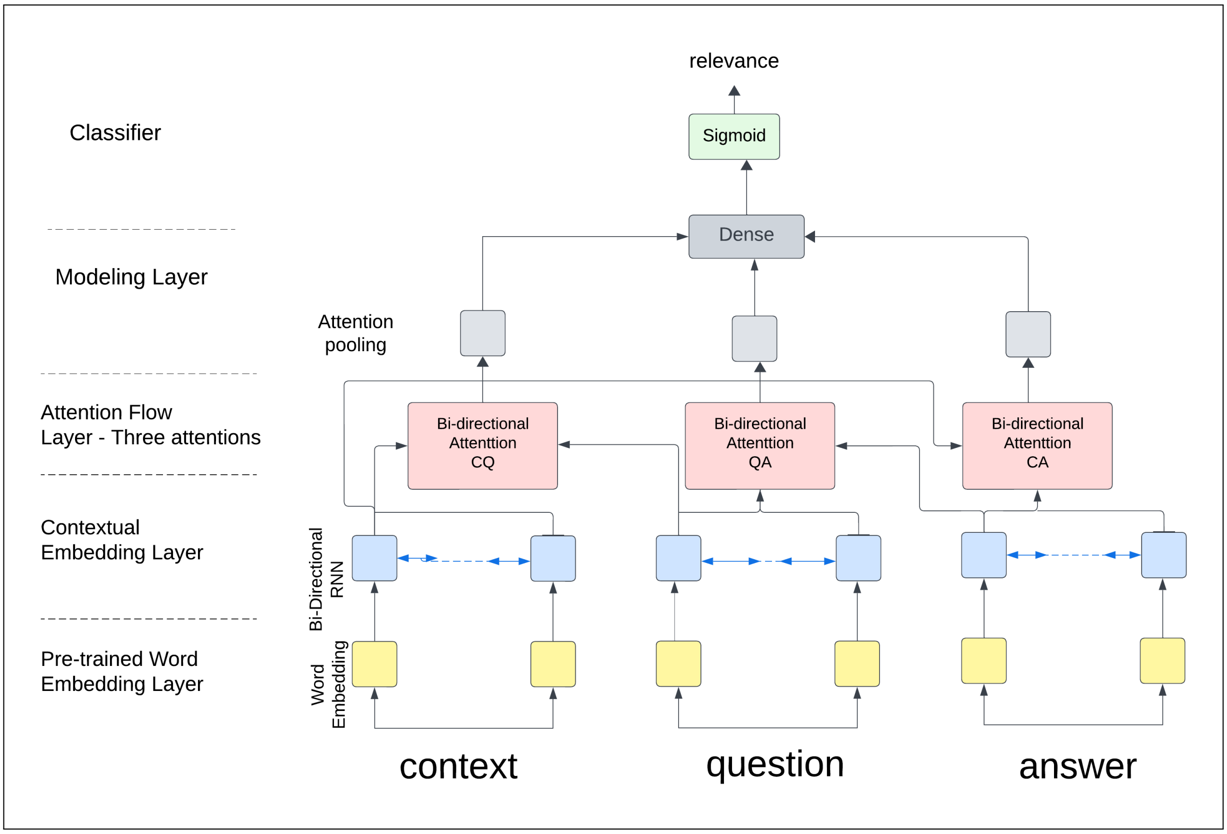

relevance
Classifier
Sigmoid
Dense
Modeling Layer
Attention pooling
Attention Flow Layer - Three attentions
Bi-directional Attenttion CQ
Bi-directional Attenttion QA
Bi-directional Attenttion CA
Contextual Embedding Layer
Bi-Directional RNN
Pre-trained Word Embedding Layer
Word Embedding
context
question
answer

Figure 4.2 Proposed Architecture

The components of the system are explained here in detail, thoroughly illustrating their underlying structure and functions.

1. **Embedding Layer:** responsible for mapping individual words to a vector space of high dimensionality. The shared parameters of this layer are utilised for context, question and answer. For embedding, pre-trained GloVe word vectors of 256 dimensions are employed, and these vectors remain unchanged throughout the training process. There are three levels of embeddings: character embeddings, word embeddings, or sentence embeddings. According to Baradaran et al., most reading comprehension studies apply word embedding [8]. Additionally, some investigations have demonstrated that using pre-trained word embeddings to initialise word vectors yields more precise outcomes than the arbitrary initialisation technique [8]. Beyond this, character embedding is appropriate for out-of-vocabulary words. Some studies have suggested concatenating word embeddings with character embeddings to improve performance. Recently, after the release of contextualised word embedding and large language models, concatenation has rarely been used. For the Arabic dataset, AraGloVe is used.

2. **The Encoding Layer:** contains three encoders for context, question and answer, respectively. Each encoder consists of an essential building block: a bidirectional LSTM or GRU followed by a dropout layer to avoid overfitting. A bidirectional encoder is applied, allowing the hidden states to receive information from the future (not only from the previous state). Experiments are conducted with GRU and LSTM to measure the effectiveness of both.

3. **Attention Layer:** captures the interaction between question, answer and context and represents it as a new context sequence. To achieve this sequence, three cross-attention layers were used:

   (a) Between Question and Answer

   (b) Between Question and Context

   (c) Between Context and Answer

   Two distinct attention mechanisms have been utilised, and the outcomes are compared. Luong's multiplicative style [48] and Bahdanau's additive style [6] draw attention. Both attention has been designed for the machine translation task.

4. **Relevance score:** The model outputs a score, which is relevant to the accuracy of the answer provided. That is, if the score is high, the label which classifies the candidate's answer is true.

Model	**English**		**Arabic**	
	F1	**Exact-Match**	**F1**	**Exact-Match**
Bi-LSTM + Additive attention	18.9	17.34	25.51	17.72
Bi-GRU + Additive attention	19.86	17.42	32.26	18.39
Bi-LSTM + Temporal Aspect + Additive attention	20.52	17.64	17.48	17.42
Bi-GRU + Temporal Aspect + Additive attention	29.98	17.04	17.42	17.52
Bi-LSTM + Multiplicative attention	19.22	17.64	18.74	17.72
Bi-GRU + Multiplicative attention	23.84	17.94	25.31	18.17
Bi-LSTM + Temporal Aspect + Multiplicative attention	21.86	17.79	18.81	17.79
Bi-GRU + Temporal Aspect + Multiplicative attention	20.08	17.87	17.42	17.42

Table 4.1 Results of the Deep Learining Models

4.3 Results

Table 4.1 displays the results obtained by the experiments. The results compared to the 'Always Negative', 'Always Positive' benchmarcks, and the BERT baseline. According to these findings, none of the models demonstrate a sufficient understanding of the task because all metrics are close to the 'Always Negative' benchmark. Although the input has been enhanced by incorporating the temporal aspect, the improvements are not significant enough to discourage further attempts at enhancement. Thus, it might be more beneficial to concentrate on transformer-based models because they have greater potential for improving task understanding than deep learning models.

The results of the models are presented based on the temporal categories in Figure4.3.

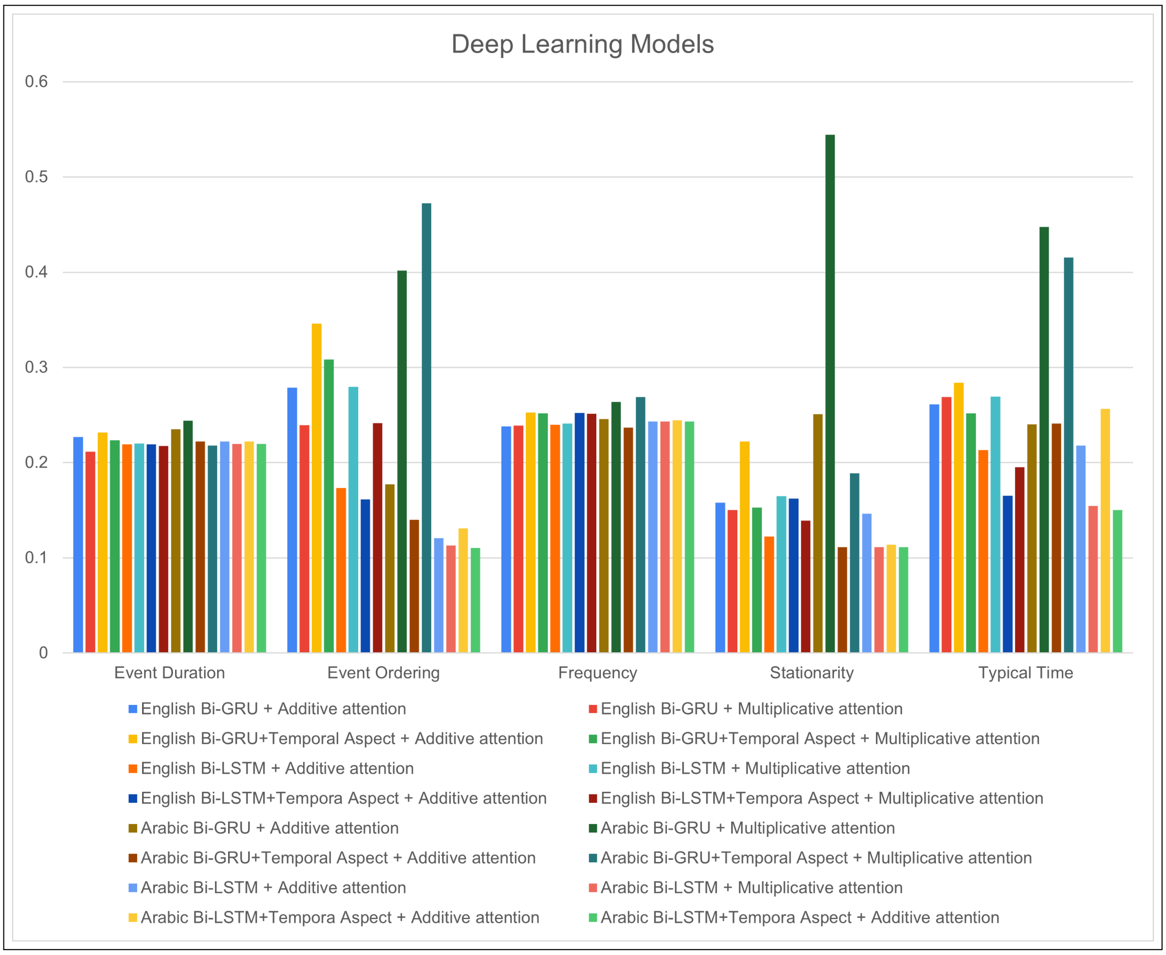


Figure 4.3 Results of the Different Deep Learning Models. Y represents F1 score.

From the results that depicted on Table 4.1 and Figuer4.3, it is evident that the model performs better in Arabic compared to English when temporal features are excluded from the input. This disparity can be attributed to the quality of the pre-trained embedding AraGloVe in contrast to its English counterpart. Arabic GloVe embeddings (AraGloVe) may offer richer and more informative representations for Arabic text due to their training on a larger corpus or better capturing the linguistic nuances of the Arabic language. Additionally, the larger embedding dimension of 300 in AraGloVe provides more expressive power compared to the English counterpart, which may contribute to the improved performance in Arabic classification tasks. Moreover, additive attention significantly contributed to the effectiveness of the models with the Arabic dataset. Arabic text often contains complex syntactic and semantic structures, and the additive attention mechanism helps the model focus on relevant parts of the input sequence, enhancing its ability to discriminate between important and less important information. This attention mechanism may be particularly beneficial for

Arabic due to its morphological complexity and varying word order. On the other hand, multiplication attention proved to be more effective for English. English text tends to have more straightforward syntactic structures and less morphological complexity compared to Arabic. Multiplication attention might be better suited for English due to its ability to capture interactions between different parts of the input sequence, which can be more prevalent and influential in English text. This attention mechanism enables the model to weigh the importance of different words or phrases based on their interactions, leading to improved performance in English classification tasks. The main findings from the extensive experiments are as follows:

- GRU demonstrates superior performance for both Arabic and English datasets. This may be because GRUs are better at capturing long-range dependencies and preserving information over longer sequences compared to LSTMs.
- Additive attention proves to be more effective for Arabic, whereas multiplication attention is more suitable for English. The preference for additive attention in Arabic may be due to its ability to better capture the contextual nuances and complex structures of Arabic text, while multiplication attention works better in English due to its capability to capture interactions between different parts of the input sequence.
- Incorporating temporal features improves performance in English, whereas it has the opposite effect in Arabic. This discrepancy may arise from differences in the temporal characteristics of English and Arabic text, highlighting the importance of considering language-specific features in model design and evaluation.

4.4 Summary

This section employed various deep learning models and diverse word embeddings. Despite extensive efforts, none of the outcomes satisfied the required standards. This shortcoming could be attributed to the difficulty of the task, the scarcity of data and the absence of temporal common sense knowledge. The dataset size was relatively small for the deep learning models. Furthermore, the proposed training split was only 3,783 question-answer pairs.

Chapter 5

Temporal Common Sense Understandings Using Transformers

5.1 Introduction

Understanding temporal common sense is difficult, as discussed in previous chapters. Certain studies have been conducted on English language datasets, such as the MC-TACO dataset [83]. However, there have been limited studies on this dataset, and a leaderboard has been established by Allen Institute to compare the performance of different models with human performance. There is a notable gap between submissions and human performance. Based on the rankings, DeBERTa-large surpasses all previous efforts. Consequently, it could be useful to investigate the performance of multilingual pre-trained language models (PLMs) on both Arabic and English datasets. In this vein, the multilingual model is a model capable of handling multiple languages and is trained on data from various languages. The performance of multilingual pre-trained language models (PLMs) varies depending on the specific language. This is because the dataset used for pre-training is not evenly distributed across all the languages covered by the models. Additionally, the quality of the dataset may also vary depending on the language. By comparing the results between the two languages and with the state of the art (SOTA), useful insights may be gained. Different models have been investigated, and the results of the two versions of the dataset were compared and analysed.

5.2 Related Works

The initial PLM utilised for Mc-TACO is BERT, as mentioned in [83]. The results obtained from BERT fell significantly below the performance achieved by humans. Consequently, a pre-processing step involving unit normalisation is implemented on the dataset, leading to an improvement in performance. Subsequently, RoBERTa is introduced and applied to the dataset without pre-processing [83]. The results obtained from RoBERTa surpass those obtained from BERT with unit normalisation. Since then, various PLMs have been tested. In a 2020 study conducted by Kaddari et al., T5 was used and produced better results than other models [41]. Furthermore, a rule-based duration normalisation was suggested and applied during data preprocessing. The results indicated that this strategy outperformed other models without duration normalisation. However, the enhancement of the proposed normalization was only by a slight margin.

Finally, applying DeBERTa-Large without any kind of preprocessing outperforms all models by a significant margin. Thus, PLMs might be successful even without rule-based processing. Additionally, applying rule-based pre-processing is highly dependent on language. Also, this process is prone to errors, and creating rules which encompass all possibilities is an intensive task due to the manual coding involved.

5.3 Experiments

Various experiments were conducted using different PLMs. Multilingual PLMs and Arabic versions of BERT were applied to understand the Arabic dataset. After this, a detailed comparison and analysis of the results of the models were presented. Two Arabic versions of BERT were adopted: AraBERTv2 and CAMeLBERT. AraBERTv2 is the latest version of AraBERT, which was initially introduced by Antoun et al. [4]. This model was selected over other Arabic versions of BERT due to its superior performance, as evidenced by the model card on Hugging Face [1] and research conducted by Alammary et al. [1]. This study involved text classification specifically for the Arabic language. In this investigation, AraBERTv2 exhibited better outcomes than XLM-RoBERTa. The CAMeLBERT model [38], developed by CAMeL Lab at New York University Abu Dhabi [2], was not included in the analysis conducted by [1]. It would be valuable to compare this model with the current leading Arabic BERT model. Furthermore, the performance of AraBERTv02 is better than that of CAMelBERT, according to [38]. However, CAMeLBERT was the second-best model among

[1] https://huggingface.co/aubmindlab/bert-base-arabert

[2] https://huggingface.co/CAMeL-Lab

all Arabic versions of BERT, based on research conducted by CAMeL Lab [38]. Therefore, it is worthwhile to compare these two models for this task. In this vein, CAMeLBERT has different versions. CAMeLBERT-msa was selected among all others because the target dataset is written in MSA Arabic. Notably, AraBERT and CAMeLBERT are different from any multilingual models because they are tailored for Arabic. AraBERT, CAMelBERT and multilingual BERT all have the same architecture because they are derived from the original BERT with some modifications. Table 5.1 shows simple comparisons between these three BERT models. According to inoue et al. [38], pretraining data size may not be an important factor in fine-tuning performance.

In this study, multilingual BERT was selected because it was the baseline model for the original dataset, and numerous versions of BERT have been designed specifically for the Arabic language. Additionally, mDeBERtav3 and XLM-RoBERTa were specifically chosen due to the effectiveness of their original models on the English dataset.

5.4 Results

AraBERT and CAMeLBERT were pretrained exclusively on Arabic datasets. Therefore, it was expected that these Arabic-specific models would outperform their multilingual counterparts and multilingual PLMs.However, the results were unexpected; while AraBERT and CAMeLBERT did outperform mBERT, they did not surpass the performance of XLM-RoBERTa or mDeBERTav3. The results of these experiments are presented in Table 5.2.

To assess the stability and variability of the system, additional runs were conducted, each with a different random seed. Three runs were conducted. Each run was independent, with its random processes, including data shuffling and GPU initialisation, influenced by its specific seed. From these three runs, performance metrics were observed, and the standard error was calculated to deduce how much performance varied with the change in seeds. Finally, the performance metrics reported were based on running the system with a default random seed, which was equal to 42.

The models were fine-tuned using identical hyperparameters to ensure a fair comparison. Nevertheless, distinct hyperparameters were employed for each model to guarantee optimization. Most of the models achieved their best performance with the chosen hyperparameter settings.

Based on these findings, all models except XLM-RoBERTa Large exhibited a standard error below 1. The maximum standard error of 11 was observed for XLM-RoBERTa. Although 11 is considered high, the F1 range is between 0 and 100, making this value

Model	Vocabulary size	# parameters	Dataset
mBERT	120K		Wikipedia
AraBERT-v2	60K	77GB	Arabic OSCAR corpus Arabic Wikipedia 1.5B words Arabic Corpus The OSIAN Corpus Assafir news articles.
CAMeLBERT-msa	30K	107GB	Arabic OSCAR corpus Arabic Wikipedia The OSIAN Corpus Arabic Gigaword Fifth Edition Abu El-Khair Corpus

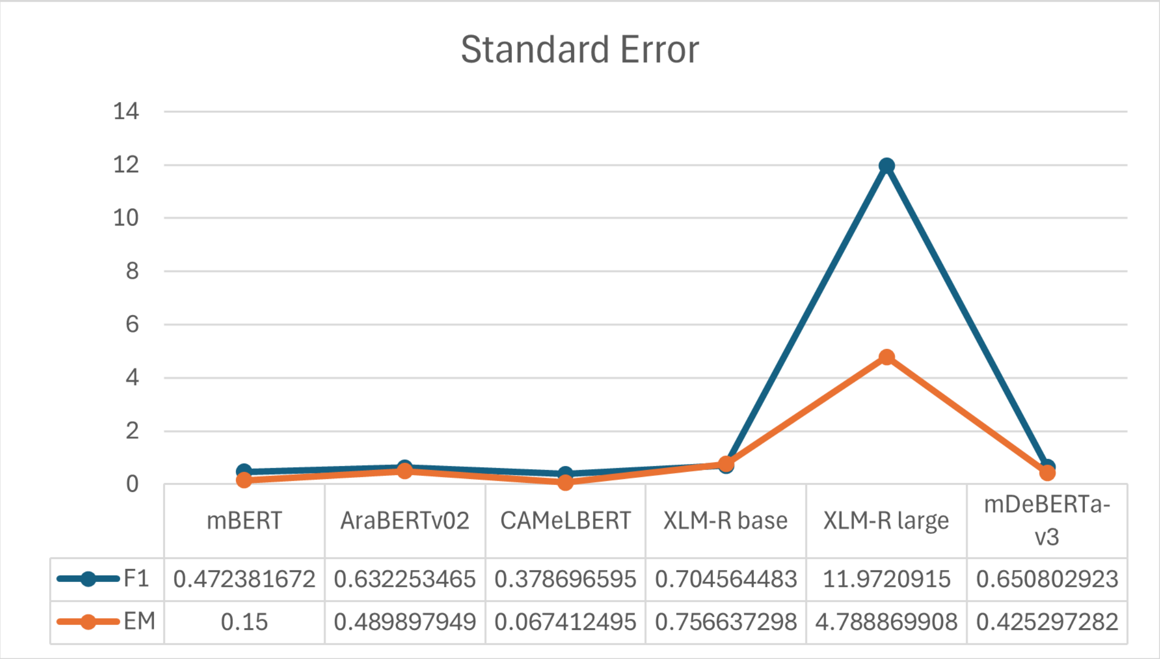
Table 5.1 Comparisons between BERT models

relatively low. Consequently, the models can be considered highly consistent and overall stable. Figure 5.1 depicts the standard errors for each model.

Based on the results presented in Table 5.2, the multilingual DeBERTa-v3 achieved the best performance, followed by XLM-Roberta Large. Although AraBERTv02 and CAMeLBERT were trained on Arabic datasets, but mDeBERTa-v3 outperformed them significantly. This may have occurred because the target task required common sense reasoning, suggesting that more advanced models like mDeBERTa-v3 could be necessary, explaining the performance discrepancy. Factors which can be attributed to the superior performance of mDeBERTa-v3 compared to other models are as follows:

Model	**F1**	**EM**
mBERT	58.12	28
AraBERT-v02	64.46	34.01
CAMeLBERT-msa	61.76	32.51
XLM-RoBERTa-Large	64.99	36.19
XLM-RoBERTa-base	61.77	31.53
mDeBERTa-v3	**67.98**	**38.66**

Table 5.2 Results of Applying PLMs on the Arabic Dataset



	mBERT	AraBERTv02	CAMeLBERT	XLM-R base	XLM-R large	mDeBERTa-v3
F1	0.472381672	0.632253465	0.378696595	0.704564483	11.9720915	0.650802923
EM	0.15	0.489897949	0.067412495	0.756637298	4.788869908	0.425297282

Figure 5.1 Standard Error for Each Model

1. The depth of the architecture in XLM-RoBERTa-large has 24 layers, twice the number of layers as mDeBERTa-v3, XLM-RoBERTa base, and mBERT, which all have 12 layers. This difference could indicate that the number of layers might not produce the best performance.

2. Although BERT and XLM-RoBERTa employ self-attention mechanisms, mDeBERTa-v3 may incorporate more advanced attention mechanisms specifically designed to capture precise linguistic dependencies. Consequently, this model can surpass the others in tasks requiring extensive linguistic analysis, such as the TCU.

3. This is also evidenced by applying English DeBERTa-v3 Large to MC-TACO, which achieved state-of-the-art results.

This study compared the performances of AraBERT-v02 and CAMeLBERT, both designed for the Arabic language. AraBERT-v02 showed better results than CAMeLBERT, possibly due to the former's vocabulary size, which is twice that of CAMeLBERT.

Figure 5.2 presents the outcomes of the various models applied to the Arabic dataset. The F1 score is displayed for each temporal aspect, facilitating an assessment of the effectiveness of each model. The figure reveals several significant findings, which are summarised below:

- The strength of all the models is the stationarity aspect. All the models scored above 74 in this aspect. mDEBERTa-v3 and XLM-RoBERTa large are scored the same. Upon analyzing the data, it appears that this particular aspect may be less difficult in comparison to other aspects. This is mainly because the majority of the responses for this feature are either yes or no. Furthermore, certain responses are evidently unrelated and were easily dismissed by the models.

- The duration of the event was the most challenging feature for all models, with a mean discrepancy of 10 units less than the overall F1-score of each model. Due to the challenging nature of this aspect, some studies, including Virgo et al. [76], have suggested that an external source is required.

- Overall mDeBERTa-v3 is the most effective model, but it did not outperform all models in all aspects, and mDeBERTa-v3 demonstrated superiority in event duration and frequency.

- AraBERt-v02 and CAMeL-msa demonstrated superior performance to all other models in the 'Typical Time' feature. Notably, the overall effectiveness of CAMeL was lower than that of the other models. Thus, it was necessary to identify the distinguishing factor of the typical time feature making the models trained on Arabic datasets perform better than other models.It might be because the typical time is closely related to the nature of the culture, making models fully trained on Arabic datasets more effective in this aspect.

- mBERT had the lowest performance in all categories

Although AraBERT and CAMeL models have different performance levels, it is important to analyse the differences in their predictions because they have numerous similarities in model architecture and pretraining datasets. The distribution of prediction percentages across

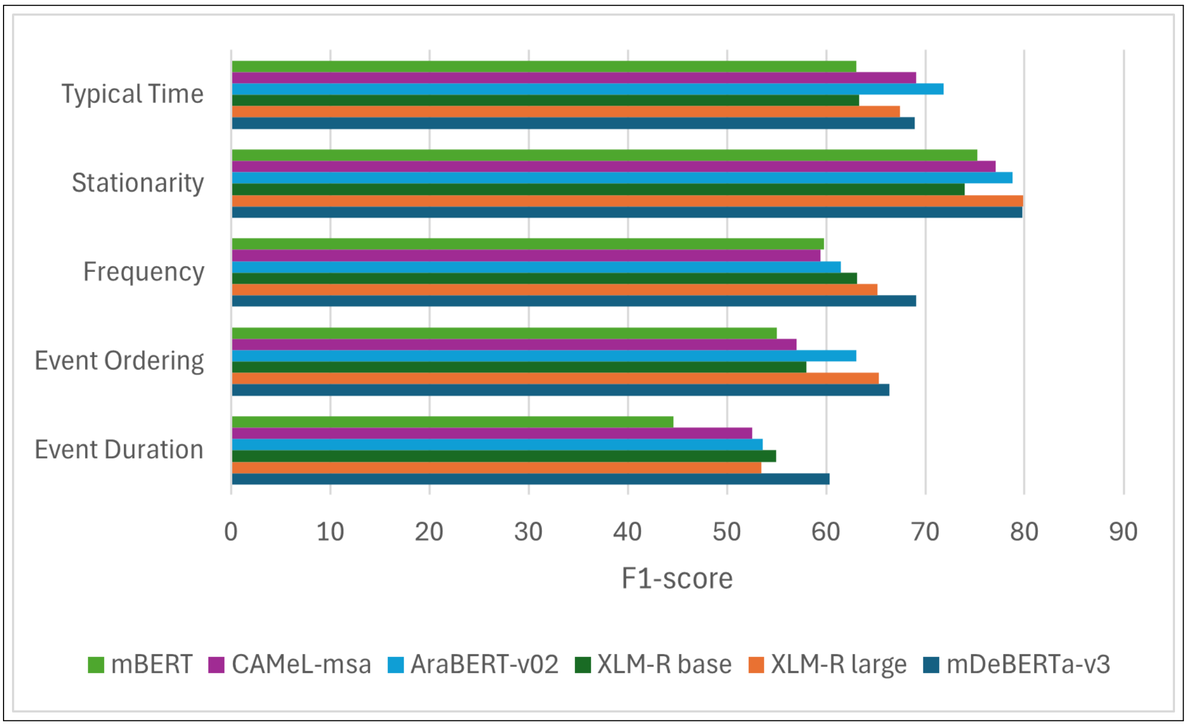


Figure 5.2 Models Results- F1 Score for Each Temporal Aspects

different temporal features is illustrated in Figure 5.3. The prediction of the probable order of events varied significantly between the two models, with most predictions differing by approximately 30%. Moreover, the percentage of identically incorrect predictions was the highest. The most similar predictions were those of frequency and event duration.

The results of multilingual PLMS for Arabic are significantly lower than the state-of-the-art (SOTA). To determine whether the performance differences are statistically significant, the t-test was conducted [21], comparing the F1-score and Exact Match (EM) metrics of the models on English and Arabic datasets. This comparison ensures that the performance differences between the PLMs on English and Arabic are statistically significant.

Hypotheses:

- ***Null Hypothesis (H0):*** There is no significant difference in performance between the English models and the Arabic models.
- ***Alternative Hypothesis (H1):*** There is a significant difference in performance between the English models and the Arabic models.

Interpretation: The independent t-test yielded the results that are displayed in Table5.3

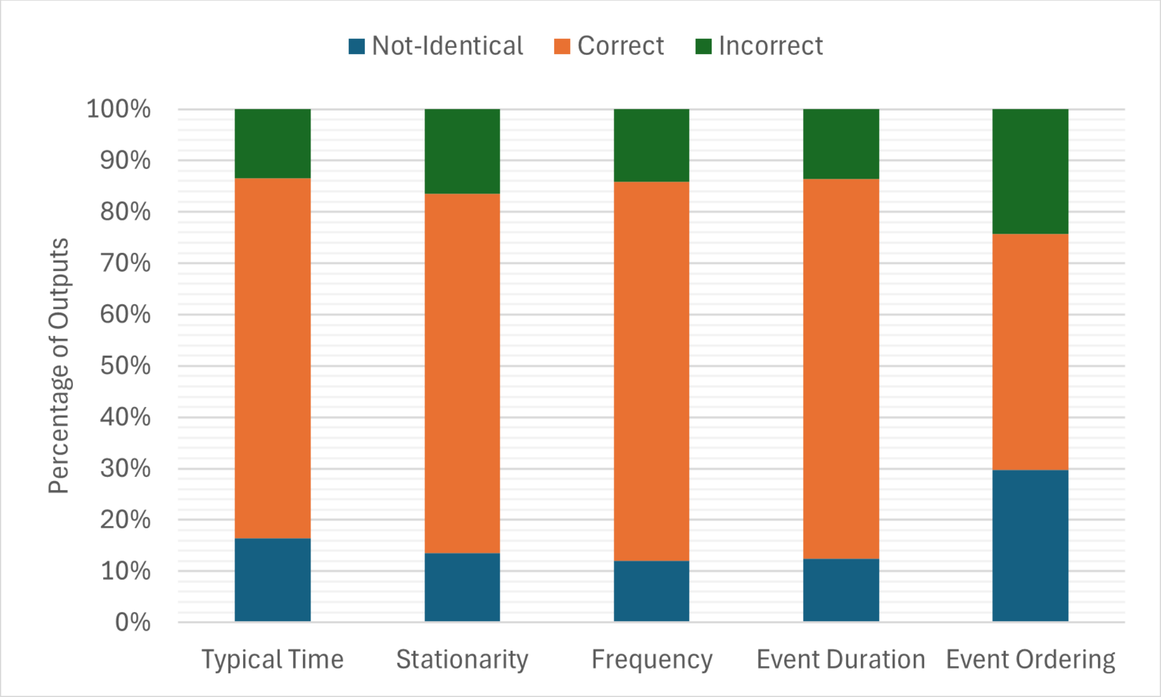


Figure 5.3 Predictions of AraBERT vs. CAMeLBERT. The y-axis represents the percentage of outputs based on the total examples that belong to a specific temporal feature.

	P-value	T-statistic
F1	0.008	6.34
EM	0.0034	8.52

Table 5.3 P-Value for F1 and EM

Given that the p-values for F1 and EM are significantly less than the commonly used threshold of 0.05 [21], the null hypothesis is rejected. This indicates that there is a statistically significant difference in performance between the English models and the multilingual PLMs.

5.5 Summary

This chapter covers applying various experiments on the newly translated Arabic dataset. The aim of the experiments to compare the effectiveness of different PLMs, including multilingual PLMs and models specifically trained on Arabic datasets, such as the Arabic versions of BERT.

The experiments demonstrated a noteworthy finding: multilingual PLMs outperformed their Arabic BERT counterparts when applied to the Arabic temporal common sense understanding tasks. This outcome suggests the potential of leveraging recent PLMs for Arabic language tasks without necessitating dedicated training on Arabic datasets, indicating that these multilingual models can effectively compete with language-specific models in the Arabic context.

The results of multilingual PLMS for Arabic are significantly lower than the state-of-the-art (SOTA). The independent t-test confirms that the performance of the English models is significantly better than the multilingual PLMs. This result underscores the challenges and the need for further research in improving model performance for temporal common sense understanding in the Arabic language. The findings highlight the significant performance gap between the two languages. Also, this suggests that the performance of multilingual PLMs on Arabic cannot be directly compared to models designed specifically for English. This disparity underscores a significant challenge: the performance of multilingual PLMs in Arabic does not directly align with that of models designed explicitly for English, reflecting a performance gap in cross-linguistic applicability.

In response to these findings, subsequent chapters of this study will explore a variety of approaches aimed at minimizing the performance disparity between English and Arabic. These efforts are directed towards enhancing the efficacy of multilingual PLMs within the Arabic linguistic domain, striving to bridge the gap and elevate the performance of these models to more closely match their English counterparts. Through these explorations, this study contributes to the broader goal of achieving linguistic parity in NLU systems, ensuring that advancements in language technology are as inclusive and universally applicable as possible.

Chapter 6

Zero-shot Cross-Lingual Transfer Learning

6.1 Introduction

A multilingual model can handle multiple languages and is trained on data from various languages. On the other hand, a cross-lingual model explicitly learns to understand and produce text in different languages by mapping representations between these languages [59]. The scope of this concept is extensive and can encompass various tasks. The research applies a cross-lingual approach, where the language used for training differs from the language of the testing data.

The motivation to examine the potential transferability of multilingual PLMs between different languages is to leverage the availability of English datasets not present in Arabic. Enhancing the model's comprehension of complex tasks necessitates a substantial dataset, which is currently unavailable. However, this issue can be addressed by utilising other temporally related datasets through multi-stage fine-tuning or continual fine-tuning. Given the unavailability of these resources in Arabic, one solution is to employ cross-lingual transfer learning to enhance the performance of Arabic tasks.

Due to the small size of the MC-TACO, many recent studies have adopted a multi-stage fine-tuning approach, including [45, 76, 43, 44]. This process involves using temporal-related datasets to enhance the model's understanding of temporal features. Additionally, external temporal knowledge has been utilised to improve model performance. However, there is currently no dataset available for the Arabic language. To address this gap, cross-lingual transfer learning can be employed.

There is a significant disparity in the availability of datasets and studies between English and other languages. Therefore, cross-lingual transfer learning is a viable approach to leverage the resources available for English and apply them to other languages for various tasks. Specifically, for the task of common sense understanding, cross-lingual transfer learning is crucial, as acquiring a common sense knowledge base is a challenging and time-consuming endeavour.

Various approaches and models are used and compared in this study. The first approach involves using the same task but with training and testing datasets from different languages. The other approach involves using adapters that are different from the standard fine tuning that have been applied in the first approach. Additionally, the results are compared to applying multilingual PLMs in monolingual settings.

No previous study has investigated cross-lingual transfer learning for the temporal domain rather than TCU. Thus, it is important to explore whether the differences in the specific representation of temporal features between Arabic and English have an impact on the performance of the models.

This chapter presents all the experiments utilising the standard fine-tuning approach for multilingual PLMs. In this approach, the model is first trained on the English dataset's training split and then tested using the Arabic test set to make predictions. The main obstacle faced by cross-lingual models is the difference in word order across Arabic and English. Also, the challenge of cross lingual between English and Arabic is that Arabic text is read from right to left, which is the opposite direction of English.

Recently developed multilingual PLMs, including mBERT [19], XLM-RoBERTa [17], mDeBERTa-v3 [33], have demonstrated significant cross-lingual capabilities, facilitating effective knowledge transfer between various cross-lingual natural language processing tasks.

6.2 Related Works

Cross-lingual learning is not a recent methodology, based on the survey by Pikuliak et al. [59], it was started in 2001 by Yarowsky et al.[81]. Following this, the methodology and techniques have been enhanced through the utilization of deep learning models and language models.

However, even the most advanced PLMs struggle to perform effectively when dealing with languages that are less represented. Additionally, the task of annotating sufficient training data in these languages is not feasible, thereby preventing studies of under-represented languages from benefiting from modern NLP capabilities. Therefore, several studies have try to overcome this gap by applying cross-lingual. However, applying this approach highlighted that a significant challenge for cross-lingual transfer is divergence in word order between different languages, which often leads to a marked decrease in performance [5].

The difference between the source and target languages, known as the domain shift or transfer gap, can be significant. Based on the survey by pikuliak et al.[59], this difference can manifest in various ways, such as distinct vocabularies, syntax, or even alphabets. To bridge this transfer gap, cross-lingual resources or technologies are commonly employed. These resources are crucial because they enhance the model with knowledge about the relationships between languages. Additionally, [79] studies the effectiveabilty of mBERT in the cross-lingual tasks. According to the extensive study, it has been indicated the potential for using linguistic features to optimize cross-lingual transfer. The study, emphasizing the importance of considering linguistic diversity for more efficient cross-lingual transfer. The authors highlight the potential use of linguistic properties to improve cross-lingual transfer and accommodate linguistic diversity. However, this research aims to explore transfer learning methods without relying on such resources or technologies. Instead, the focus is solely on utilising transfer learning methods to comprehend the domain shift.

6.3 Standard Fine-tuning vs. Adaptive Fine-Tuning

Standard Fine-tuning: This process involves taking a pre-trained model and fine-tuning it on a specific task by adjusting all its parameters using a new, often smaller, dataset relevant to that task. **Adaptive Fine-tuning:** It is considered as an advanced technique compared to the standard fine-tuning . Adaptive fine-tuning can be implemented by applying multi-steps of fine-tuning also limiting the parameters that are trained. In the multi-step, the model goes through an intermediate adaption phase and then is fine-tuned for the target task. The model is optimized or fine-tuned on an intermediate dataset that is related to the target task, and then to the target dataset [45]. This may enhance the model's performance by enabling it to more effectively adapt to the domain or features of the required task. Freezing some parameters is often used in this context to retain general features learned during pre-training while adapting more specific layers to the intermediate and target datasets [57, 58]. This can avoid catastrophic forgetting, where the PLMs lose previously learned knowledge when fine-tuning on a new task.

6.4 Cross- Lingual with Standard Fine-tuning

6.4.1 Experiments

Various multilingual models have been assessed for cross-lingual transferability. The models chosen for this task were Multilingual BERT [19], XLM-RoBERTa [17], and mDeBERtav3 [33]. Multilingual BERT was selected because it is the baseline model for the original dataset. Additionally, mDeBERtav3 and XLM-RoBERTa were specifically chosen based on a recent study by He et al. (2023), demonstrating that these models outperformed all other multilingual models for cross-lingual tasks involving the Arabic language [33]. Therefore, the researchers determined that these models would likely perform well for the task at hand.

6.4.2 Results and Analysis

The process of hyperparameter optimisation was employed to identify the most effective configuration for attaining good performance. The outcomes of the experiments are illustrated in Table 6.1.

Model	**F1**	**EM**
mBERT	37.34	18.62
XLM-RoBERTa-large	**65.09**	34.08
XLM-RoBERTa-base	59.61	28.15
mDeBERTa-v3	61.04	**36.26**

Table 6.1 Cross-Lingual Results

The stability and robustness of all the experiments were confirmed. The standard error for each model is presented in Table 6.2. Additionally, the Standard error was computed by executing the model three times, each with a different seed, revealing that a markedly low standard error.

Table 6.1 reveals that mDeBERTa outperformed all other models in terms of exact match and that XLM-RoBERTa Large achieved the highest F1 score. On the other hand, mBERT performed the worst and was not effective.

The inferior performance of mBERT is a direct result of a vocabulary size which is less half that of its counterparts. This vocabulary size is crucial for the model's comprehension, serving three purposes. First, large vocabulary size increases the coverage of rare words. Second, it improves the representation of nuanced meanings because many words have

Model	**F1**	**EM**
mBERT	0.17	0.14
XLM-R base	0.30	0.22
XLM-R large	0.40	0.72
mDeBERTa-v3	1.39	0.66

Table 6.2 Cross-Lingual Standard Error

multiple shades of meaning or synonyms. A larger vocabulary enables the model to capture these nuances and context-dependent variations more accurately, resulting in a better overall understanding. Lastly, a larger vocabulary enhances the handling of out-of-vocabulary words, even though the model may encounter completely new words despite its extensive vocabulary. Although XLM-RoBERTa and mDeBERTa-v3 are trained on the same multilingual dataset and have the same vocabulary size, there are noticeable differences in the performance of these two models. This discrepancy might be caused by the different architectures of the models. The following list enumerates these differences:

1. The mDEBERTa-v3 model incorporates a disentangled attention mechanism which separates content and positional information, improving its ability to capture complex linguistic features, including those specific to Arabic.

2. Regarding vocabulary, mDEBERTa-v3 focuses on full words rather than subwords, which may be more suitable for Arabic due to its rich morphology and word formation system.

3. For pre-training, mDEBERTa-v3 utilises the replaced token detection (RTD) objective, which has proved to be more effective than the MLM objective used in XLM-RoBERTa, particularly for languages like Arabic with rich morphologies.

The listed factors might explain why mDeBERTa can outperform XLM-RoBERTa in terms of exact match.

The outcomes of the cross-lingual fine-tuning were compared with those of the monolingual fine-tuning with the multilingual models. As shown in Table 6.3, there was a notable distinction between the monolingual approach and zero-shot transfer learning. All the models performed worse in cross-lingual settings, which occurred because English and Arabic are two separate languages. This divergence explains the discrepancy between the two languages, so achieving cross-lingual understanding between the two presents a significant challenge.

Similarly, generating comparative results for models fine-tuned in Arabic is also a difficult endeavor.

The analysis of the results presented in Table 6.3 can be outlined as follows:

- The effectiveness of mBERT showed the most flawed performance. Specifically, there has been a significant decline in the Exact Match and F1 scores. Thus, mBERT fails to meet the expected benchmarks in cross-lingual experiments, suggesting the need for further investigations.
- Fluctuations of the base and large XLM-RoBERTa were not significant because the margin of difference was almost five units. This is reflected in the size of the model. While it is more than 20 between mBERT and XLM-RoBERTa base. mBERT was not trained for cross-lingual tasks like XLM-RoBERTa, which might explain the results.
- Notably, XLM-RoBERTa large achieved almost the same F1 score in cross-lingual tasks and monolingual tasks for Arabic. Thus, this model might understand the task equally well in both languages, even though the model encountered difficulties with the testing set of Arabic. This decrease in the model's performance from that with English could be attributed to the linguistic differences between Arabic and English.
- Finally, mDeBERTa-based v3 surpassed all other models regarding exact match in all settings.

	English		**Arabic**		**Cross-lingual**	
Model	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
mBERT	63.62	33.11	58.12	28	37.34	18.62
XLM-RoBERTa-Large	**70.59**	42.04	64.99	36.19	**65.09**	34.08
XLM-RoBERTa-base	66.15	38.74	61.77	31.53	59.61	28.15
mDeBERTa-v3	70.47	**43.02**	**67.98**	**38.66**	61.04	**36.26**

Table 6.3 Comparing the Cross-Lingual to the Mono-lingual Results

Figure 6.1 illustrates the results categorised by temporal characteristics. The following list highlights some interesting points from the findings.

- Stationarity achieved the highest performance among all the features for XLM-RoBERTa and mDeBERTa-v3, but mBERT performed poorly in this regard, producing the worst results for at least 20% of all other performance characteristics.

- The most notable aspect of mBERT was its typical operating time. Additionally, its performance was close to that of XLM-RoBERTa.

- Like monolingualism, validating the duration of the event was the most challenging aspect for the TCU.

- Overall, mDeBERTa performed worse than the XLM-RoBERTa base in terms of frequency.

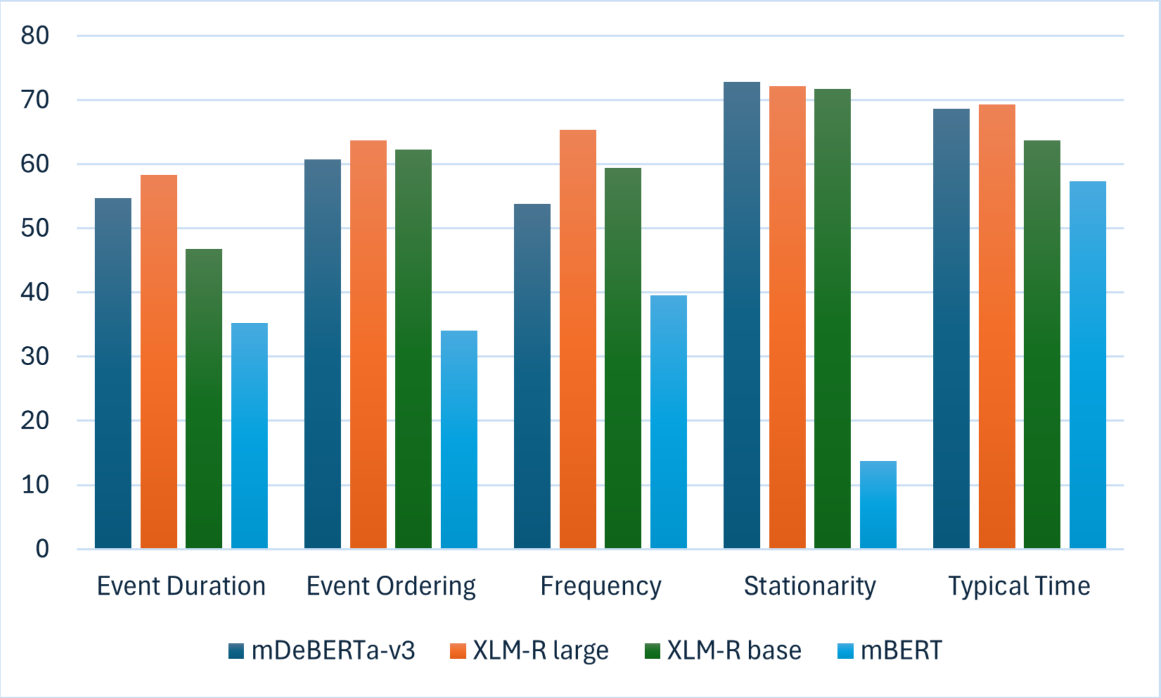


Figure 6.1 Models Results

Figure 6.2 compares monolingual and cross-lingual results with respect to temporal features. The following list will provide key points based on the figure.

- The performance of all models in a monolingual setting exceeded cross-lingual performance across all temporal categories.

- There was a noticeable difference in the performance of the stationarity characteristics , when applying mBERT, between monolingual and cross-lingual settings. This was the strongest feature in the monolingual settings but the lowest in the cross-lingual ones. The performance of mBERT in cross-lingual tasks across all aspects is 37, while it is only 15 in the stationarity category. Additionally, for monolingual tasks in Arabic, the performance of mBERT in the stationarity category is about 75. This indicates a large margin between monolingual and cross-lingual performance. The alignment of the stationarity feature between Arabic and English was not perfect, meaning that how this aspect is expressed can vary significantly between the English and Arabic languages. Thus, the model might not have been transferring knowledge about this feature effectively. As mentioned earlier, the training of mBERT involved multiple languages, but the relative representations between these languages were not considered. Therefore, the model's performance significantly declined in almost all aspects.

- XLM-RoBERTa performed consistently, suggesting that the transfer gap between languages is minimal.
- mDeBERTa-v3 showed similar performance for most features, except for frequency, where there is a decrease of approximately 18%.
- According to the analysis, XLM-RoBERTa demonstrated a 5% improvement over Arabic monolingual settings when validating the duration of the event.

Based on the comprehensive analysis, the language differences had varying impacts across the models. Generally, the event typical time was the most consistent feature.

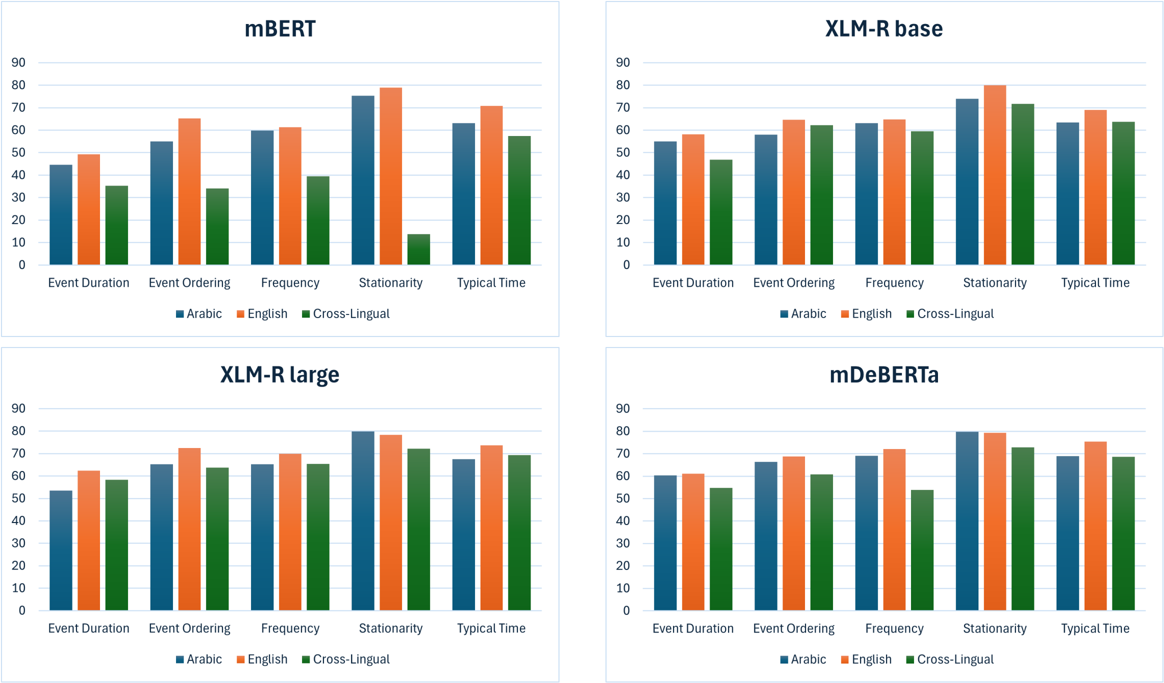


Figure 6.2 Mono-Lingual vs. Cross-Lingual

6.5 Cross-Lingual with Adaptive Fine-tuning

The previous section highlighted a crucial finding: multilingual models display a performance gap between monolingual and cross-lingual tasks. This discrepancy poses a significant challenge that needs to be addressed to enhance the efficiency of these models. Hence, searching for a method to bridge this gap and improve the overall performance of multilingual models is crucial. This chapter examines various fine-tuning techniques. Two primary factors determine technique choice: the effectiveness of parameter-efficient training and the efficiency of cross-lingual transferability. In the standard fine-tuning approach, all the model parameters are trained, but only a fraction of the parameters are trained using the proposed adapter techniques.

This study employed two approaches for adapter methodologies: X-Mod [57] and MAD-X [58].

X-MOD

In 2022, Pfeiffer et al. introduced the X-MOD model as an extension of the multilingual masked language model XLM-RoBERTa [57]. This model incorporates language adapters, modular components specific to the language, during the pre-training phase, and these language adapters remain frozen during fine-tuning. According to Pfeiffer et al. [57], multilingual pre-trained models often suffer from the curse of multilingualism, decreasing performance per language as more languages are included [57]. To address this issue, the X-MOD model introduces language-specific modules to increase the overall capacity of the model while maintaining a constant number of trainable parameters per language. Unlike previous approaches that learn language-specific components after pre-training, the X-MOD model pre-trains these modules from the beginning. Extensive studies in various NLP tasks have demonstrated that the X-MOD approach mitigates the harmful interference between languages and facilitates positive transfer, improving performance for both monolingual and cross-lingual tasks. Figure 6.3 depicts the model architecture.

Although X-MOD has demonstrated strong performance in various natural language processing (NLP) tasks, its effectiveness in common sense tasks has not yet been evaluated. To assess the efficacy of implementing the X-MOD adaptive model, experiments were conducted in three distinct scenarios, employing hyperparameter fine-tuning to optimize the model's performance. The model was tested on monolingual datasets in both English and Arabic. Additionally, the study evaluated the model's performance in a cross-lingual setting, where the model was fine-tuned in English and then assessed in Arabic.

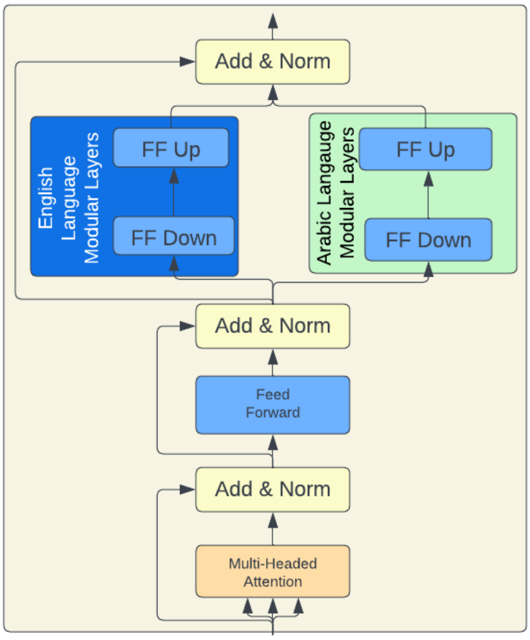


Figure 6.3 X-MOd Adapter. Inspired by [57]

MAD-X

In 2020, Pfeiffer et al. introduced MAD-X [58], an adaptive cross-lingual training method, prior to X-MOD. The study developed a collection of language adapters. The language adapters were trained on Wikipedia, and the transformer parameters and language adapters were kept frozen during the fine-tuning phase. To fine-tune the model, the source language and task adapters were used during the training stage, but only the task adapter was trained. Meanwhile, the source language adapters were substituted with the target language adapters during the prediction, and the task adapter was kept for the prediction. Figure 6.4 illustrates the architecture of MAD-X.

The adapter model was evaluated for various tasks, excluding common sense reasoning classification. Specifically, the model's performance was assessed in different languages but not in Arabic. The results showed a considerable improvement in the model output. This study evaluated the performance of this adaptive model in the TCU for both Arabic and English. Additionally, this research assessed the performance of MAD-X in monolingual settings, comparing the outcomes to those obtained using the multilingual base model without the adapter.

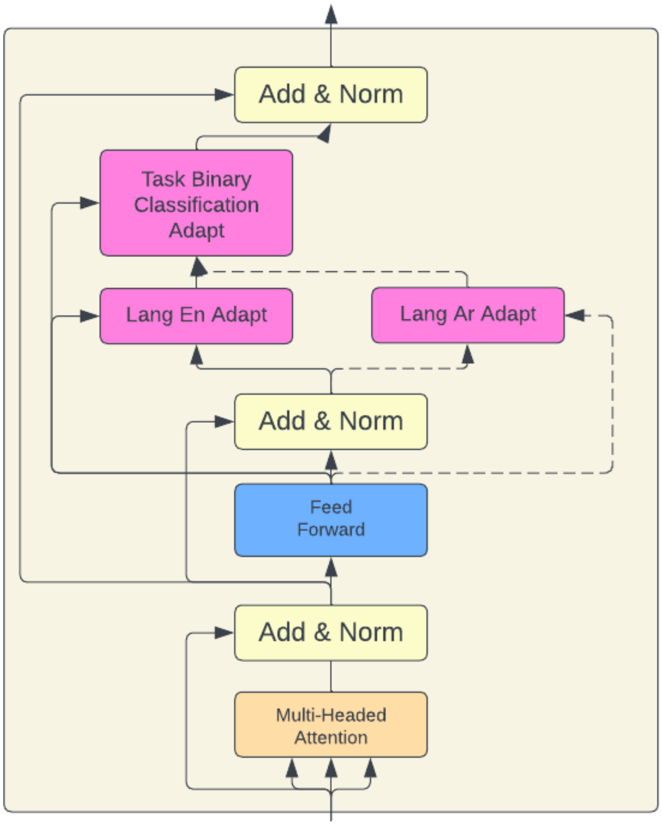


Figure 6.4 MAD-X Adpater. Inspired by [58]

Comparison

To compare MAD-X and X-MOD, one must understand that MAD-X incorporates pre-trained language adapters which are layered on top of PLMs, which are then fine-tuned with task heads to enhance performance. On the other hand, X-MOD is pre-trained using XLM-R with language modulariser layers, allowing it to support multiple languages. By specifying the language desired for the task, the language modulator is activated.

6.5.1 Results and Discussion

Table6.4 provides the results of applying adapters cross-lingually compared with those from monolingual settings. To ensure the reliability and robustness of the results, the experiments were conducted multiple times with different seeds, and the reported value represents the performance of the default random seed, which is 42. The standard error (SE) was approximately 0.4, quite low compared to the F1 metric, which ranged from 0 to 100, indicating the robustness of the results. Beyond this, the training data has been shuffled.

	Cross-Lingual		**English**		**Arabic**	
Model	F1	EM	F1	EM	F1	EM
XMOD	0.61	0.29	0.63	0.35	0.64	0.31
MAD-X	**0.69**	**0.4**	**0.73**	**0.46**	**0.69**	**0.44**

Table 6.4 Adaptive Models

Using MAD-X as the cross-lingual model resulted in superior performance to that of all other cross-lingual models. Furthermore, both MAD-X and X-MOD outperformed the XLM-RoBERTa model trained in Arabic. Consequently, this methodology for cross-lingual transfer learning is as effective as using multilingual language models. Findings from this part of the study are listed below:

- The F1-score of X-MOD in Arabic and English was very close, with only a 1% difference. This result suggested that the model outperformed any multilingual PLMs since there was a significant performance gap when using multilingual PLMs, indicating that the model is biased towards English. However, when X-MOD was used, the model gave equal attention to each language and achieved the same level of performance for both languages.
- The evidence from the cross-lingual settings demonstrated that X-MOD performs equally well in both mono-lingual and cross-lingual contexts, indicating the absence of a performance disparity typically seen when models are transferred from one language to another (known as a transfer gap).

6.6 Summary

The aim of cross-lingual transfer learning is to leverage models and resources developed for one language and apply them to another. In the context of zero-shot cross-lingual Temporal Common Sense Understanding (TCU), the objective is to train a model with an English dataset and then use it to predict labels for inputs in Arabic. This chapter has focused on adopting the zero-shot approach through the use of multilingual Pre-trained Language Models (PLMs), presenting and analyzing their performance. To further enhance cross-lingual capabilities, the chapter also introduces the utilization of adapters, which are specialized modules designed to tailor PLMs for specific languages or tasks without the need for comprehensive retraining. Despite these innovative approaches, the performance of the models still falls short of expectationsbehind the human performance and the SOTA for English dataset, highlighting a significant gap in effective language transfer.

Chapter 7

Assessing Temporal Common Sense Understanding: An Error Analysis and Evaluation of PLMs

7.1 Introduction

The aim of this chapter is evaluating the multilingual PLMs in the TCU task. Through a detailed analysis of errors generated by these models, the types of mistakes are classified to identify patterns and potential weaknesses inherent in the models' handling of temporal information. This error analysis aims to uncover the underlying challenges that impact the PLMs' effectiveness in TCU tasks.

Furthermore, benchmark is established for assessing PLMs in temporal classification tasks. This benchmark was suggested to identifying specific areas of difficulty in temporal reasoning. By deeply understanding the sources of these challenges, this chapter contributes to the broader goal of enhancing PLMs' ability to process and understand temporal information effectively.

In doing so, the aim of this study is to provide insight into the nuances of temporal common sense reasoning that are currently under-represented or misunderstood in multilingual PLMs. This investigation is pivotal for advancing the field's understanding of how temporal common sense is encoded in language models and for guiding future improvements in NLU systems, particularly those operating in multilingual and multicultural contexts.

7.2 Error Analysis

The task undertaken by TCU presents significant challenges, as evidenced by the results of the models in the previous section. This lead to manually investigate the inputs in order to understand the challenges. So, by studying MC-TACO dataset, certain questions pose difficulties even for humans, a fact substantiated by the dataset's human performance metrics (87%). Consequently, errors encountered in this context can be classified into two main categories: human challenging errors and linguistic and task complexity errors. Human challenging errors are those instances where both the model and humans struggle to validate the given answers. On the other hand, linguistic and task complexity errors denote situations where the model fails to grasp the intricacies of language use or the validating the answer plausibility. The following sections will present each type of error in detailes, providing a more detailed analysis.

7.2.1 Human Challenging Errors

As previously mentioned, the first type of errors, characterized by their complexity and the challenges they pose to both humans and models, can be approached with a degree of leniency compared to the second type. These errors inhabit an "uncertain zone," where they are neither fully correct nor entirely incorrect. For instance, consider the question related to the duration of an illness based on a provided context: "Dürer's father passed away in 1502, and his mother died in 1513. How long was his mother ill?" The provided answer is that she was ill for 30 years. Even though the gold standard label for this input is "no," suggesting the answer could be considered incorrect, it's also possible to argue "yes," indicating that a 30-year duration could indeed represent a period of illness. This example highlights the subjective nature of certain questions and answers, emphasizing the complexity and potential ambiguity inherent in evaluating temporal common sense understanding.

The human performance metrics displayed on the leaderboard of the MC-TACO, specifically the F1 score of 87.1% and the Exact Match (EM) rate of 75.8%. It's important to note, however, that these measures, derived from a subset of the dataset, may not fully encapsulate accuracy. Despite this limitation, it can be assumed that human challenging errors account for approximately 22% of the total error rate, providing insight into the extent of the challenges posed by these types of questions.

Capturing inputs that are considered challenging for humans within a test set comprising 9,442 items is an arduous task. Implementing a voting mechanism across the entire test set could offer insights into the difficulty level of each question. However, benchmark is derived from a sample of the total data [83], suggesting that a rough sampling of instances falling into this challenging category would be both practical and acceptable. This approach allows for the identification and analysis of particularly complex cases without the need to exhaustively review every item in the dataset, thereby providing a feasible method for gauging the extent of human challenging errors within the dataset.

Additionally, it's important to recognize that some answers may be culturally dependent, meaning that for certain cultures, an event or concept might be considered plausible, whereas for others, it might not. This variability introduces another layer of complexity, classifying such instances as human challenging errors. This cultural dimension underscores the necessity of incorporating a diverse perspective when evaluating answers, as it highlights the subjective nature of understanding and interpreting information. Recognizing the influence of cultural context on what is deemed correct or incorrect is crucial for accurately assessing the scope of human challenging errors within the dataset. For instance, within the dataset, a question regarding the appropriate time for an interview illustrates the impact of cultural differences. In Saudi culture, interviews can be scheduled on Sundays, a practice that might differ from norms in other cultures where the workweek typically begins on Monday. Furthermore, the start time for schools in Saudi Arabia is earlier than in many other countries, reflecting another aspect of cultural variance. Also, during Ramadan, eating late at night is very common to accommodate the fasting schedule. Although this contrasts with the dining habit for cultures that do not observe Ramadan. These examples highlight how cultural contexts significantly influence the interpretation of what constitutes a correct or plausible answer, thereby contributing to the categorisation of such instances as human challenging errors within the dataset.

Some research efforts have focused on the grounding of time expressions as a culturally dependent aspect. One notable study, by [69], analyzed time expressions across 27 languages, although Arabic was not included. This research aimed to define how the conceptual range of time periods, such as morning and noon, can vary significantly across different cultures. Additionally, it explored the impact of these cultural variations on PLMs. The findings of such studies are crucial because they highlight the challenges PLMs face in accurately understanding and generating contextually appropriate responses for time-related queries [69]. These variations in the perception of time can affect a model's ability to provide correct and culturally sensitive answers, underscoring the importance of incorporating diverse cultural understandings into the development and training of language models.

Accordingly, the human challenging errors can have two subcategories: "Cultural Temporal Interpretation" or "Subjective Event Understanding".

In the exploration of PLMs navigate the complexities of culturally dependent temporal expressions and subjective understandings, it becomes crucial to examine specific instances where errors occur. The following examples provides a comprehensive overview of various error types identified in this category, showcasing sample inputs alongside the expected responses and comparing these with the outputs generated by the PLMs. This comparative analysis not only highlights the discrepancies between expected and actual responses but also offers insights into the models' underlying challenges with cultural nuances and subjective interpretations. This examination can shed light on areas for improvement in PLMs in TCU task. Moreover, creating a dataset that can address this issue shows great potential.

Examples

Context	Question	Answer	Label
It was huge and inefficient, and she should never have spent so many pesos on a toy, but Papa would not let her return it.	How long did she spend at the store buying the toy?	She spent 7.5 minutes at the store buying the toy	yes
California was first to require smog checks for clean air, pass anti- tobacco initiatives and bike helmets laws.	How often are such initiatives passed?	One a month	no
Most of us have seen steam rising off a wet road after a summer rainstorm.	How long does stream rise after a summer rainstorm?	Steam rises for 30 minutes off a wet road before a summer rainstorm.	no

Table 7.1 Subjective Event Understanding Errors Examples

Table 7.1 presents various scenarios that may be classified as Subjective Event Understanding. The analysis of each example is discussed in the following list, ordered according to the same sequence as the examples in the table.

1. All the PLMs, in English and Arabic, predicted no. This outcome is indicative of a challenge for PLMs, touching upon human cognitive processes as well. The specificity of the duration-7.5 minutes-represents an atypical time frame that is not commonly associated with the activity described.

2. The response of all the PLMs, in English and Arabic is yes. This input could be considered human challenging, as it involves understanding the legislative process and the realistic pace at which laws and initiatives are typically passed, which varies significantly across different jurisdictions and over time.

3. Not all the PLMs able to predict the correct label. The scenario involving steam rising from a wet road after a summer rainstorm, with the duration specified as "steam rises for 30 minutes off a wet road before a summer rainstorm," presents a nuanced challenge that tests both temporal reasoning and contextual understanding. The question is inherently human challenging, not just because of the temporal aspect—quantifying the duration steam rises—but also due to the contextual misunderstanding in the provided response. The mention of steam rising "before" a rainstorm contradicts the common observation and understanding that steam typically rises "after" rain has fallen and heated by the warm road, creating a visual phenomenon observed by many.

Context	**Question**	**Answer**	**Label**
Most of us have seen steam rising off a wet road after a summer rainstorm.	How often does it rain in the summer?	a couple times every month	yes
It was huge and inefficient, and she should never have spent so many pesos on a toy, but Papa would not let her return it.	What time did she purchase the toy at the store?	Midnight	no
For example, what if you place a cake in the oven and you leave it in too long?	What time would you put the cake in the oven?	12:00 a.m.	no

Table 7.2 Cultural Temporal Interpretation Errors Examples

Furthermore, Table 7.2 presents various scenarios that may be classified as Cultural Temporal Interpretation. The analysis of each example is discussed in the following list, ordered according to the same sequence as the examples in the table.

1. All PLMs, in English and Arabic, predicted no. In fact, this case touches on cultural or geographical dependency. The perception of how often it rains in the summer varies significantly across different regions and climates, which makes this question inherently dependent on the cultural and geographic context. Where this information is not provided in the context.

2. Labelling this as no can be culturally dependent touches on broader themes of consumer behaviour, store operating hours, and possibly societal norms regarding appropriate times for shopping. The assumption that purchasing a toy at midnight is unusual or incorrect may indeed vary by culture and locale. In some regions or during certain times (like holidays or special sale events), late-night shopping can be common, while

in others, it might be seen as atypical due to differing social norms and operational hours of businesses.

3. Same as the previous case in some cultures and during holidays 12 a.m. can be possible for baking.

7.2.2 Linguistic and Task Complexity Errors

This category encompasses a broad range of challenges encountered by language models, making it the most diverse category of errors. These errors are predominantly language-dependent, influenced significantly by the specific tokenization methods employed during model training. Additionally, linguistic features such as morphology and overall complexity contribute to the difficulties faced by models in processing and understanding language inputs accurately. Finally, validating the likelihood of provided answers can be challenging, requiring common sense knowledge.

One key aspect of linguistic complexity errors is their relationship with the structure and nuances of language, including how words are formed and combined. The way a model tokenizes input—breaking down sentences into words, subwords, or characters—affects its ability to understand and generate coherent responses. Morphological complexity, which involves the structure of words and their relationship to one another within a language, further complicates comprehension, especially in languages with rich inflectional systems.

Moreover, these types of errors can often be mitigated through strategic pre-processing steps, such as time normalization (converting time expressions to a standard format) [41] or unit normalization (standardizing measurements) [83]. Additionally, models vary in their ability to interpret numbers, whether presented in word form or as numerals, which can lead to inconsistencies in understanding and answering questions accurately.

Some models demonstrate proficiency in identifying the type of answer required (e.g., a date, a quantity) but falter when it comes to providing the correct specific response. This discrepancy may stem from the models' training datasets, which might not adequately prepare them for the breadth of task complexity they encounter in real-world applications. This limitation suggests a need for more comprehensive training approaches that better encapsulate the linguistic diversity and complexity inherent in natural language. Moreover, enhancing the current dataset could assist the model in comprehending this complex task.

Table 7.3 illustrates a sample of errors in the complexity of language and tasks. The analysis of each example is discussed in the following list, ordered according to the same sequence as the examples in the table.

Context	Question	Answer	Label
However, more recently, it has been suggested that it may date from earlier than Abdalonymus' death.	How often did Abdalonymus die?	every two years	no
Setbacks in the 1930s caused by the European postwar slump were only a spur to redouble efforts by diversifying heavy industry into the machine-making, metallurgical, and chemical sectors.	How long did the postwar slump last?	decades	yes
Tommy and Suzy (brother and sister) went to the playground one afternoon with their mom and dad, Jan and Dean.	How often did they go to the playground?	twice a month	yes

Table 7.3 Linguistic and Task Complexity Errors

1. mBERT failed to predict this in both. For Arabic dataset, XLM-R base, AraBERT and CAMeL failed. While the human can easily validate the answer as no. For PLMs, however, this represents a more complex challenge. The models must not only process the natural language of the question but also apply logical reasoning and background knowledge to identify the irrationality of the premise—that a historical figure could die multiple times. This difficulty arises from the models' reliance on patterns and data within their training corpus, which are not explicitly cover every aspect of common sense or logical reasoning needed to immediately flag the question's premise as impossible.

2. All the PLMs, in English and Arabic, predicted no. The difficulty lies not just in interpreting historical events and their timelines but also in the nuanced understanding of the term "postwar slump" and its impact over time. The term "postwar slump" refers to the economic downturn following a significant conflict, in this case, likely World War I, considering the reference to the 1930s. The incorrect labelling of "decades" as possible answer by all PLMs could be referred to different reasons. The model's failure to recognize "decades" as a plausible duration may indicate a gap in understanding the prolonged effects of postwar economic conditions or the specific historical context.

3. All the PLMs, in English and Arabic, predicted no. Correctly interpreting and validating answers about the frequency of activities also involves common sense understanding and world knowledge, such as the typical behaviors of families with young children.

PLMs must leverage this broader knowledge to make informed inferences about habitual actions.

7.3 PLMs Evaluation

A novel hypothesis proposes assessing PLMs' proficiency in TCU tasks by categorizing inputs into five temporal dimensions. Under this framework, the model's ability to categorise questions correctly suggests that it grasps the temporal features and understands the type of temporal question being asked. However, failing to provide the correct answer implies that the model struggles with applying logical reasoning or lacks the necessary world knowledge to determine whether an answer is plausible within the given temporal context.

Drawing inspiration from traditional Question Answering (QA) systems, where question classification is crucial for high performance, based on [37], and [46], our approach adapts this methodology for a modern context. Unlike in QA systems, where classification directly aids in generating answers, here it serves as a diagnostic tool. This distinction is key, emphasizing that our goal is not to classify for classification's sake but to deepen our understanding of PLMs' handling of temporal information.

Implementing this strategy requires utilizing all PLMs applied to the dataset for temporal classification. This foundational step ensures models understand the temporal dimension of queries. Following classification, comparisons of the model performances on this task with their TCU task performance, focusing on each temporal aspect. This comparison is designed to unveil correlations or discrepancies, providing insights into the models' capabilities and areas needing refinement.

For example, a model proficient in classifying questions but faltering in providing accurate answers might indicate an understanding of temporal concepts but a lack in applying these in complex scenarios. Such findings highlight the importance of enhancing models' reasoning capabilities and understanding of temporal contexts.

Our methodology not only offers a nuanced assessment of PLMs' handling of temporal information but also points to potential improvements. By identifying specific weaknesses, this approach can guide the development of targeted training or fine-tuning strategies, enhancing PLMs' effectiveness in TCU tasks and beyond.

7.3.1 Results and Discussion

The classification task has been implemented both in monolingual and cross-lingual settings, examining models' performance across different languages. In monolingual settings, particularly for Arabic and English, the classification F1-score achieved an impressive rate of around 97%. This high level of F1 underscores the effectiveness of PLMs in understanding temporal information within a single language context.

Figure 7.2 depicted cross-lingual experiments, the F1-score remained robust, exceeding 90% in various models, with the notable exception of mBERT. This variance supports the hypothesis that multilingual PLMs are adept at processing temporal information across languages, as evidenced by their high F1 rates. The slightly lower performance of mBERT in these cross-lingual tasks can be attributed to its design and training objectives, which may not have been optimized for handling the cross lingual as effectively as models like XLM and mDeBERTa. The distinction in performance between mBERT and other multilingual models like XLM and mDeBERTa in cross-lingual settings highlights the importance of model architecture and training data in achieving high F1-score in cross-lingual settings. While mBERT has shown proficiency in a range of multilingual applications, its slightly lower F1-score in this particular context suggests a potential area for improvement or adaptation to better support cross-lingual understanding of temporal information. There are some studies trying to overcome this weakness in different NLP tasks.

Figure 7.1 provided distinctly showcases the significant gap between classification F1-score and performance on TCU task in monolingual settings. This discrepancy is notably pronounced, emphasizing the challenges PLMs face when transitioning from understanding temporal categories to applying this understanding in more complex TCU scenarios.

Given the similarity in outcomes between Arabic and English datasets, the analysis will focus in-depth on the Arabic dataset to explore this phenomenon further. This targeted approach allows for a nuanced examination of where and how PLMs struggle, particularly within the context of language-specific nuances and temporal reasoning.

Figure 7.3 breaks down the gap across various temporal aspects, shedding light on the specific areas where discrepancies in performance are most stark. This visual representation serves as a critical tool for identifying the dimensions of temporal understanding that require further refinement in PLM, offering insights into potential focus areas to improve model training and development.

7.4 Summary

Adopting error categorization can enhance our understanding of Pre-trained Language Models' (PLMs) behavior and pinpoint the sources of errors. This insight paves the way for further research aimed at augmenting model comprehension in TCU and across all aspects of Natural Language Understanding (NLU). By adopting the classification approach that is

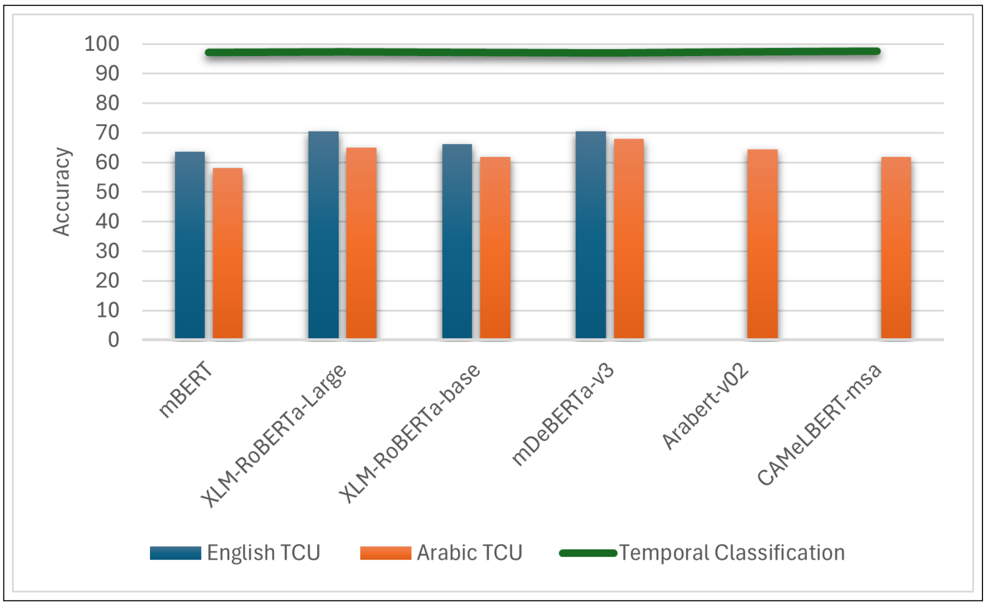


Figure 7.1 The Results of TCU and Temporal Classification for Arabic and English

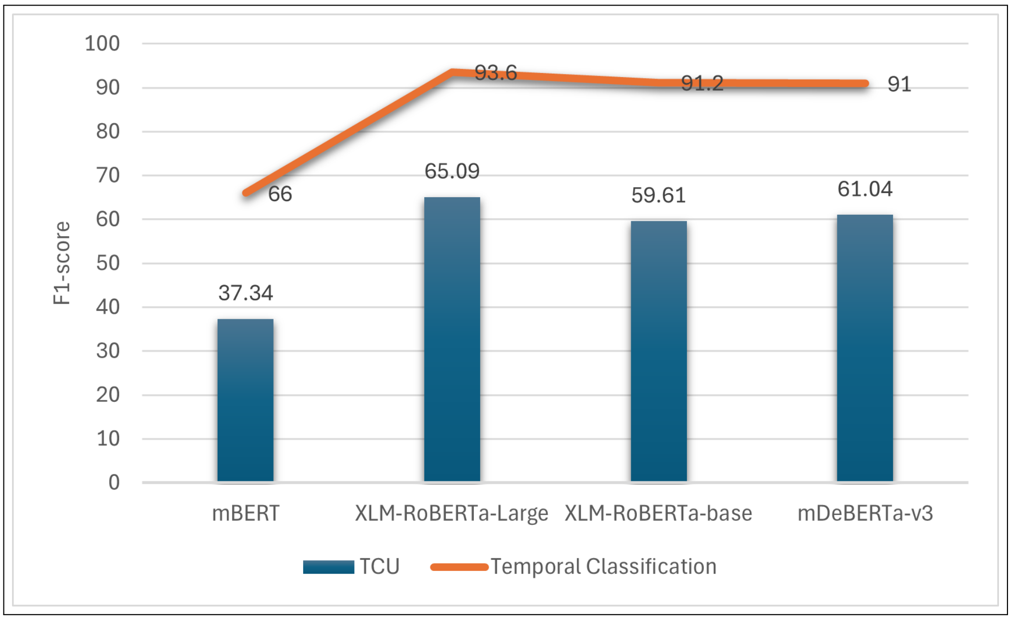


Figure 7.2 The Results of TCU and Temporal Classification for Cross-Lingual

used for PLMs evaluation, researchers can more precisely pinpoint where PLMs fall short in the TCU task, distinguishing between failures of basic temporal comprehension and failures of logical reasoning or knowledge application. This distinction is crucial for guiding future improvements in model training and development, directing efforts towards enhancing the models' reasoning capabilities and their understanding of the world.

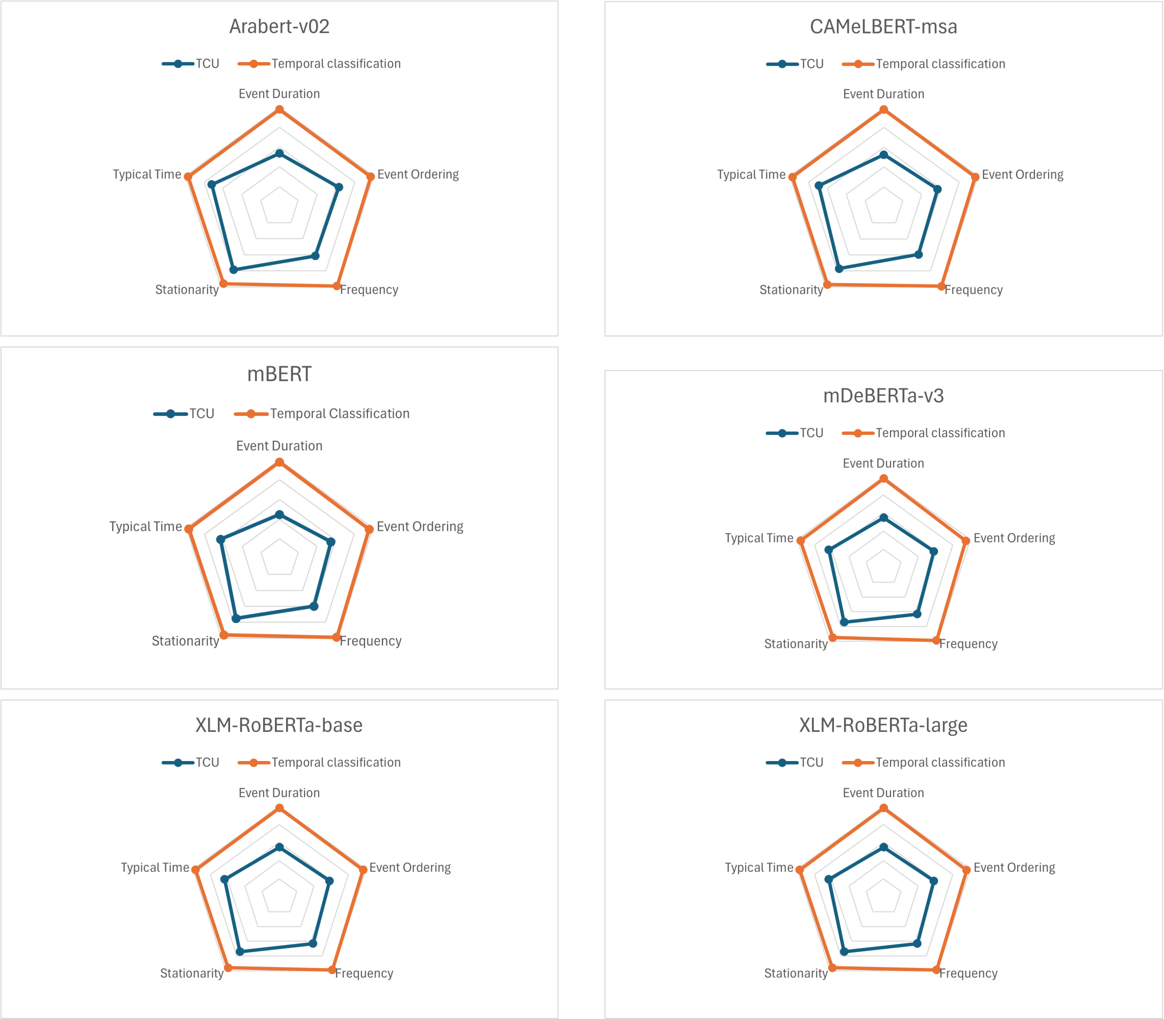


Figure 7.3 Results of Arabic Temporal Classification Compared to TCU

Chapter 8

Evaluating LLMs Temporal Common Sense Understanding in Zero-Shot Settings

8.1 Introduction

This chapter evaluates several recent LLMs using various prompts to assess their ability to reason about temporal common sense inputs in zero-shot settings. Both Arabic and English datasets were sampled. The study demonstrated various language models, including GPT3.5 Turbo, GPT4, Gemini, AceGPT and Jais. This is considered the first exploration of Google Gemini, AceGPT and Jais in the TCU field. Finally, this chapter compares and discusses the results. The research focuses on utilizing various prompts and analyzing the models' outputs accordingly.

8.2 LLMs Promises and Challenges

8.2.1 Motivation

Motivated by significant advancements in generative AI, this research examines the efficacy of large language models (LLMs) for temporal common sense understanding (TCU) in a zero-shot setting. The objective of this study is to investigate the complete capabilities of LLMs in TCU. As the researchers explore the potential of LLMs, the main objective is to comprehend LLMs' capacity for reasoning about temporal characteristics. This process involves examining the underlying mechanisms which enable these models to excel in this intricate undertaking. By examining these mechanisms, the researchers seek to uncover the factors contributing to the mechanisms' success and improve their performance

8.2.2 Challenges

This study's main challenge concerned the absence of an API in Jais to execute a Python script for the experiments. Therefore, this study limited to only 100 samples randomly selected from each language for testing purposes. To conduct experiments on AceGPT, more than 80 hours were dedicated using Google Colab, providing over 100 GPU units. On the other hand, for Jais, this author had to manually navigate the website and enter each instance of the sampled dataset then record the output one by one. After that the API for accessing Jais has been provided, so it can be accessed through a Python script. However, the researcher encountered a problem with the model's complexity, limiting input to only 17 samples per run. Analysing the results of 500 samples, 100 for each prompts, for Jais and AceGPT posed a challenge due to the occasional poor quality of the generated text. The output of all the models are manually annotated. However, annotating the output of Jais and AceGPT was an intensive task. Although the researcher attempted to apply the Chain-of-Thought (CoT) methodology, the complexity involved with Jais and AceGPT prevented them from completing the process within the given time frame. However, the author manually tested the methodology with a few samples, although the results were not promising. To ensure a fair comparison, the same methodology was applied to all language models (LLMs) which the researcher intended to evaluate.

8.2.3 Evaluation

The capabilities of understanding the temporal common sense of recent LLMs in Arabic and English were assessed through manual evaluation. The LLMs evaluated included ChatGPT 4 and 3.5 models, Jais, Accept, and Google's Gemini. Various temporal aspects were considered in the evaluation. To enhance the zero-shot learning performance of LLMs, the researchers developed new design prompt suggestions, and the effectiveness of this approach was demonstrated empirically. The accuracy metric is widely regarded as the most reliable measure for assessing the effectiveness of models. Finally, by quantifying model's accuracy, the researchers could assess model performance and make well-informed decisions using the available data.

8.3 Related Works

Several studies have examined the performance of language models (LLMs) in various tasks, including common sense reasoning. However, this study is the first to evaluate the understanding of temporal common sense using Jais, AceGPT and Google Gemini. In one recent study, Jain et al. investigated LLMs in different temporal datasets [39]. However, this study experimented with various prompts instead of only one. Additionally, the study of [39] did not mention sample size in their study. Furthermore, they evaluated the results using accuracy as the metric, without considering ChatGPT4. In contrast, the current study compared two versions of datasets using two recent bilingual LLMs, namely Jais and AceGPT. Additionally, the researcher incorporated prompts with a temporal aspect as hints to the LLM and examined their behaviour in the given task.

8.4 Method

8.4.1 Data set

A stratified sampling method was used to randomly select 100 samples from the MC-TACO dataset. The samples were selected in a way that ensured a representation of each temporal aspect of the dataset. Specifically, 20 unique questions were randomly chosen from each aspect. Additionally, half of the selected questions had correct answers, while the other half did not. The choice of sample size was influenced by the evaluation of human performance described by Zhou et al. [83]. Furthermore, according to Lopez-Espejel et al. [49], the sample size that is used to assess the reasoning skills for LLM is only 30 instances out of a larger dataset. This can indicate that 100 is a reasonable size. The second appendix contains the sample dataset as the one adopted in the experiments.

8.4.2 Zero-Shot Prompts

This study used five different prompts, some of which included the task, whereas others included the temporal aspects of the question. For example, if the question is asking for a logical duration of a specific event, the input for the model will be the context, question, answer, and the temporal feature which is event duration in this case.

The primary investigation of this experiment involved whether all the prompts provided the three inputs, exploring the logicality of the answer, rather than its correctness. Table 8.1 illustrates the list of proposed prompts. Since LLMs have shown improved zero-shot reasoning capabilities when given effective prompts [49], the researchers proposed five different and high-performance prompts, inspired by the prompts developed by [49]. Throughout the prompt engineering, the researcher tried to condense a task type into a single prompt. Additionally, contextual information, specifically the temporal aspect of the question, was included to assist the model in determining the plausibility of the answer based on the provided temporal feature. The purpose of this variation was to evaluate factors which can influence the model's reasoning ability.

8.4.3 LLMs

[id=RA, remark=moved to LR]A brief description of the models from this study is provided. Table 2.2 summarises the main features of the models and compares them.

8.5 Results and Analysis

This section examines the results of the different prompts for both the English and Arabic samples. Table 8.2 presents the outcomes of each prompt for each model, and the average across all prompts provides an overall result. Beyond this, Figure 8.1 illustrates the performance of the models in each language, and the chart is based on the average score. The analysis from the table is discussed in the following list:

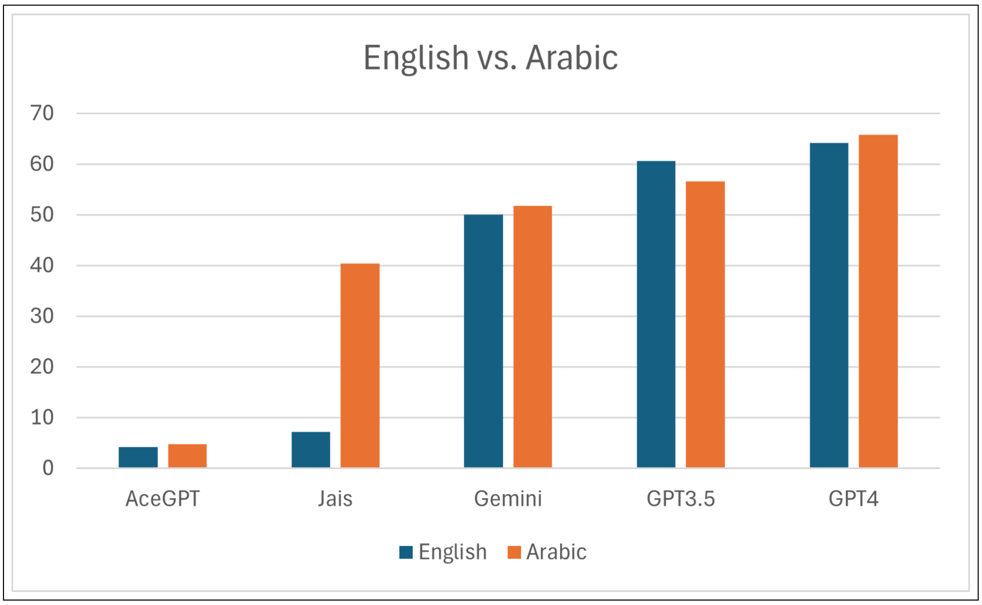


Figure 8.1 Comparing the Average Results for Arabic and English

- According to the results obtained from AceGPT, the model demonstrated a complete lack of understanding of the given task in both languages. The outputs produced by

English	Arabic
Given the input (context, question, answer): + {context} + {question}, and + {answer} determine if the provided answer is plausible or implausible.	بالنظر إلى المدخلات (السياق، السؤال، الإجابة): حدد ما إذا كانت الإجابة المقدمة معقولة أم غير معقولة.
The task is temporal common sense reasoning, Given the input (context, question, answer): + {context} + {question}, and + {answer} determine if the provided answer is plausible or implausible.	المهمة هي المنطق الزمني السليم ، بالنظر إلى المدخلات (السياق، السؤال، الإجابة): حدد ما إذا كانت الإجابة المقدمة معقولة أم غير معقولة.
Context: {context} Question: {question} Answer: {answer} Is this a logical answer?	(السياق، السؤال، الإجابة) هل هذه إجابة منطقية ؟
Context: {context} Question: {question} Answer: {answer} Is this a logical answer as {Temporal Aspect}?	(السياق، السؤال، الإجابة) هل هذه إجابة منطقية كجانب الزمني؟
Context: {context} Question: {question} Answer: {answer} determine if the provided answer is plausible {Temporal Aspect}?	(السياق، السؤال، الإجابة) حدد ما إذا كانت الإجابة المقدمة معقولة من ناحية الجانب الزمني؟

Table 8.1 Prompt in Arabic and English

running the model were largely unrelated to the provided inputs and prompts; the output consisted of only randomly generated text.

- It is evident that Jais struggles with English language understanding. While the model is meant to handle both English and Arabic. Also, according to the model's description, the numbers of tokens in Arabic and English are approximately the same. The results were unexpected. The model struggles with accuracy in Arabic, but performs even worse in English, indicating significant challenges in completing the task. The superior performance of Jais in Arabic compared to English can be attributed to several key factors. Firstly, as a transformer-based large language model, Jais integrates cutting-edge features such as ALiBi position embeddings, which allow the model to handle much longer inputs, thereby enhancing context handling and accuracy [68]. Additionally, the implementation of state-of-the-art techniques like SwiGLU and maximal update parameterization significantly improves the model's training efficiency and accuracy [68].

  A primary reason for the model's outstanding performance in Arabic is the unique, purpose-built dataset comprising 116 billion Arabic tokens [68]. This dataset is specifically designed to capture the complexity, nuance, and richness of the Arabic language, providing a robust linguistic foundation. In contrast, while the model also includes 279 billion English word tokens to boost cross-language performance, it is the focused and high-quality Arabic dataset that drives the model's performance in Arabic.

- The performance of GPT3.5, GPT4 and Gemini was almost the same for both languages.

- The accuracy of all the models in Prompt 5 was better in Arabic. This could be due to the fact that presenting the response type in Arabic assists the model in comprehending the input.

- GPT4 achieved the highest accuracy of 80% for Prompt 2 in English, while for Arabic, the best accuracy achieved was 74% for the same prompt. Providing a clear task outline in Prompt 2 likely contributed to the model's improved understanding and performance.

- The performance of Gemini was consistent across the five prompts in the Arabic sample.

According to this study, the optimal model for both Arabic and English is GPT 4, which. produced the most accurate predictions for the given Arabic and English sample data. Beyond

Language	Prompt	AceGPT	Jais	Gemini	GPT3.5	GPT4
English	Prompt 1	5	6	54	68	67
	Prompt 2	7	0	58	65	80
	Prompt 3	6	23	50	51	56
	Prompt 4	1	4	50	69	56
	Prompt 5	2	3	38	50	62
	Average	4.2	7.2	50	**60.6**	64.2
Arabic	Prompt 1	1	43	52	54	63
	Prompt 2	9	27	55	48	74
	Prompt 3	6	45	49	65	62
	Prompt 4	5	48	52	57	64
	Prompt 5	3	39	51	59	66
	Average	**4.8**	**40.4**	**51.8**	56.6	**65.8**

Table 8.2 Results for both English and Arabic for each Prompts

this, Prompt 2 yielded the highest accuracy among all prompts. A detailed analysis of the results is presented in Figure 8.2, which compares the predictions for both Arabic and English. Notably, there were 26 discrepancies out of 100 predictions. Event ordering is the most variable aspect between Arabic and English.

The results of models based on the temporal aspects of Arabic and English are shown in Figure 8.3. Although the models struggled in addressing stationarity, GPT4 demonstrated its superiority in the English and Arabic samples. deleteted[id=RA, remark=moved Bard]Bard achieved event ordering as the second notable feature in English.

8.6 Summary

This chapter assessed the capability of the LLMs to explain temporal common sense in Arabic and English. Notably, the models demonstrated satisfactory performance in both languages, a promising outcome. Thus, the models might share a comparable understanding of Arabic and English, but the performance of each is still inferior to human performance.

The accuracy of the NLP models is strongly affected by the data used to train them. Thus, the Arabic NLP models cannot be compared to the English models due to the limited data set

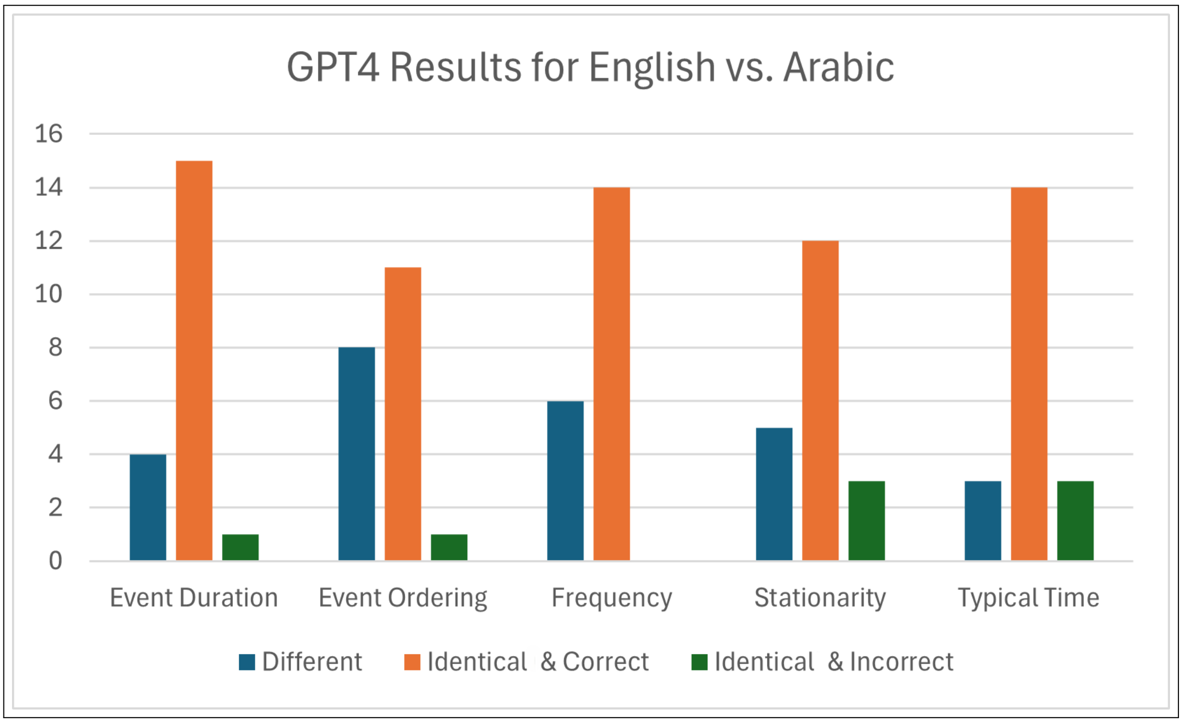


Figure 8.2 Comparing the Results of GPT4 on Arabic and English

and the low quality of the available Arabic corpus. An additional factor in the effectiveness of AI frameworks is that the meaning of Arabic words can vary significantly. To address these challenges, improving the annotated data and incorporating transfer learning techniques can improve the Arabic NLP models. However, further research is needed to better explain the complexities of Arabic content and improve NLP models, especially regarding common sense reasoning.

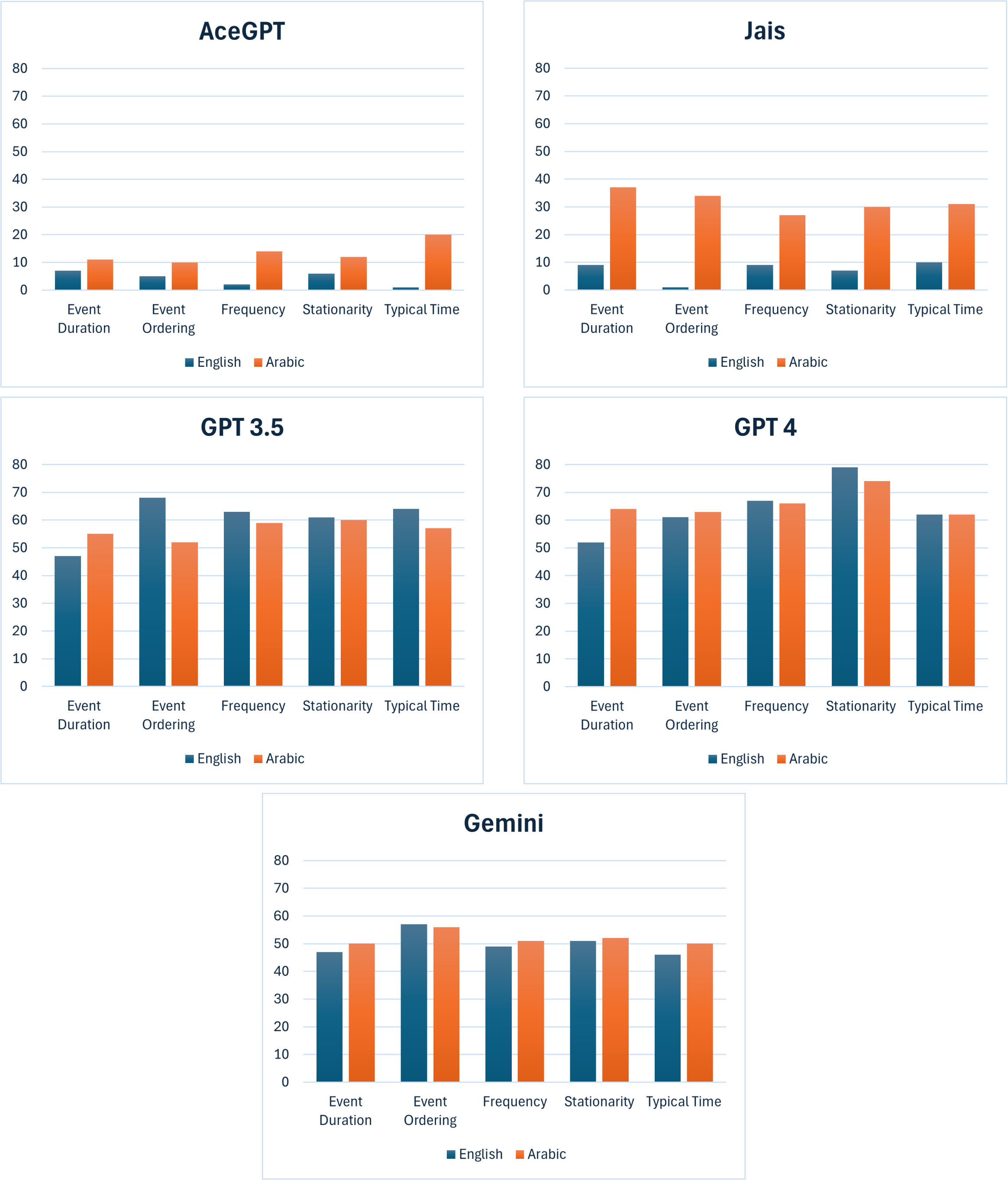


Figure 8.3 Comparing the Results of Arabic and English

Chapter 9

Conclusion

9.1 Introduction

This research contributed to Arabic temporal common sense understanding, adopting different LLMs for Arabic and English datasets and evaluating and analysing model performance.

This chapter concludes the research effort by analysing how each chapter has contributed towards addressing the research questions. Also, the research contributions and their significance in the field are disscussed. Furthermore, it outlines the study's constraints. Lastly, it gives an outline of potential avenues for future research.

9.2 Revisiting the Thesis and Research Questions

This study evaluated various questions to investigate the problem of the Arabic temporal common sense understanding. In this subsection, the research questions are revisited considering the results of the case studies.

- Can the existing English dataset be utilized to develop an Arabic dataset for Temporal Common Sense Understanding?

  To address this question, a translated version of MC-TACO was constructed. An Arabic dataset was generated using the most widely used English dataset for this task. Additionally, this dataset is comprehensive, encompassing all temporal aspects. The unavailability of an Arabic dataset for this task highlights the importance of constructing one. In this vein, it would be preferable to create a dataset specifically designed for Arabic, but the cost could be prohibitively expensive for individual researchers.

- How effective are the available multilingual PLMs in understanding temporal common sense? Can these models compete with their monolingual counterparts? Do the performances of these models vary across different languages?

  By examining various multilingual PLMs in both datasets, this study identified a notable performance gap between models operating in Arabic and those in English. Specifically, models originally designed for English exhibited a significant performance advantage over their multilingual counterparts. For models trained solely on Arabic datasets, a direct performance comparison with multilingual PLMs in English was not feasible. The analysis highlighted the complexity of the Arabic language and underscored a significant scarcity of suitable datasets for such endeavors. Addressing this gap will necessitate a collaborative effort.

- What is the effectiveness of applying cross-lingual transfer learning to this task?

  Applying cross-lingual transfer learning was proposed to overcome the gap in the performance of the models between the languages. This approach could be a potential solution for utilising existing English resources. However, the effectiveness of the results was limited due to the significant differences between Arabic and English. A comprehensive analysis and discussion of the results are provided.

- Can one mitigate the transfer gap between English and Arabic by applying different training techniques?

  Different approaches were adopted to improve cross-lingual performance and minimise the gap. Using pre-trained adapters has produced promising results to reduce the transfer gap between English and Arabic.

- To what extent is Generative AI successful in TCU? How does the language of the dataset influence the model's performance?

  Based on the evaluation of five distinct LLMs with various prompts, the findings suggest that the performance of both languages lags behind that of human performance. The models performed similarly to each other, with the exception of Jais, which proved to be ineffective for English.

- What specific challenges and error patterns emerge from applying PLMs for TCU in Arabic and English contexts? A new hypothesis has been applied to evaluate PLMs. This implies that the challenge lies not in the basic understanding of temporal concepts but in the model's capacity for deeper reasoning and validation of temporal information. Highlights the distinction between understanding the category of temporal information required (a relatively simpler task) and applying complex reasoning to validate the logical or factual consistency of answers (a more advanced and challenging task).

9.3 Contributions

Using deep learning models, this thesis advances the field of temporal common sense comprehension. In order to address the lack of resources, it creates an Arabic version of an extensive dataset based on an existing English dataset. To enhance temporal information processing across languages, the study employs various deep learning models, including word embeddings, RNNs, and attention mechanisms, and utilizes multilingual PLMs. Additionally, a set of LLMs has been evaluated for the task of TCU in both languages. In order to pinpoint areas that need work, the thesis also performs an error analysis and sets a standard for temporal understanding.

9.4 Limitations

This study's dataset was quite small and contained inputs that did not necessarily conform to Arabic cultural norms, which might have limited the efficacy of the models. To address this issue, a future study could construct an extensive Arabic gold dataset. However, this endeavour could require substantial investments in both resources and specialised knowledge.

9.5 Future Work

After examining various models and reviewing recent studies on natural language understanding, this researcher found it challenging to predict the performance of a model in a specific task. There is uncertainty regarding the effectiveness of the model for a specific task and the part of the model which contributes to these results. When performing an interpretability analysis, it might help to better understand the behaviour of the model.

Additionally, a thorough review of the available datasets should be undertaken, aiming to categorize errors based on the proposed categories. This process will not only refine our understanding of where current models falter but also pave the way for targeted improvements. Analyzing all the possible errors and categorize them, will be beneficial for identifying the weakness of the PLMs. Thus, achieving an effective strategies for enhancing the behavior of LLMs. This will ultimately leading to more robust and accurate natural language understanding systems.

Augmenting MC-TACO with a temporal common sense dataset specifically designed for Arabic can enhance model performance since input will surpass cultural disparities. Addressing the need for cultural specificity in model training could significantly reduce, the potential for errors or misunderstandings in output.

References

[1] Alammary, A. S. (2022). BERT Models for Arabic Text Classification: A Systematic Review. *Applied Sciences*, 12(11):5720. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Aliwy, A., Taher, H., and AboAltaheen, Z. (2020). Arabic Dialects Identification for All Arabic countries. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, WANLP@COLING 2020, Barcelona, Spain (Online), December 12, 2020*, pages 302–307.

[3] Allen, J. F. (1983). Maintaining Knowledge about Temporal Intervals. *Commun. ACM*, 26(11):832–843.

[4] Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. In Al-Khalifa, H., Magdy, W., Darwish, K., Elsayed, T., and Mubarak, H., editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

[5] Arviv, O., Nikolaev, D., Karidi, T., and Abend, O. (2023). Improving Cross-lingual Transfer through Subtree-aware Word Reordering. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 718–736, Singapore. Association for Computational Linguistics.

[6] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*.

[7] Bao, J., Duan, N., Yan, Z., Zhou, M., and Zhao, T. (2016). Constraint-Based Question Answering with Knowledge Graph. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514.

[8] Baradaran, R., Ghiasi, R., and Amirkhani, H. (2022). A Survey on Machine Reading Comprehension Systems. *Natural Language Engineering*, 28(6):683–732. Publisher: Cambridge University Press.

[9] Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

[10] Bian, N., Han, X., Chen, B., and Sun, L. (2021). Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12574–12582.

[11] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

[12] Boudaa, T., El Marouani, M., and Enneya, N. (2018). Arabic Temporal Expression Tagging and Normalization. In Tabii, Y., Lazaar, M., Al Achhab, M., and Enneya, N., editors, *Big Data, Cloud and Applications*, Communications in Computer and Information Science, pages 546–557, Cham. Springer International Publishing.

[13] Bousmaha, K. Z., Rahmouni, M. K., Kouninef, B., and Hadrich, L. B. (2016). A Hybrid Approach for the Morpho-Lexical Disambiguation of Arabic. *Journal of Information Processing Systems*, 12(3):358–380.

[14] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[15] Cai, Q. and Yates, A. (2013). Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 423–433, Sofia, Bulgaria. Association for Computational Linguistics.

[16] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30*.

[17] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[18] Derczynski, L. R. (2017). *Automatically Ordering Events and Times in Text*, volume 677 of *Studies in Computational Intelligence*. Springer International Publishing, Cham.

[19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[20] Diab, M., Habash, N., and Zitouni, I. (2017). NLP for Arabic and Related Languages. *Traitement Automatique des Langues*, 58(3):9–13. Place: France Publisher: ATALA (Association pour le Traitement Automatique des Langues).

[21] Dror, R., Peled-Cohen, L., Shlomov, S., and Reichart, R. (2020). Statistical Significance in NLP. In Dror, R., Peled-Cohen, L., Shlomov, S., and Reichart, R., editors, *Statistical Significance Testing for Natural Language Processing*, pages 23–33. Springer International Publishing, Cham.

[22] Elangovan, V. and Shirkhodaie, A. (2011). A survey of imagery techniques for semantic labeling of human-vehicle interactions in persistent surveillance systems. 8050:80501P. Conference Name: Signal Processing, Sensor Fusion, and Target Recognition XX ADS Bibcode: 2011SPIE.8050E..1PE.

[23] Farghaly, A. and Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Trans. Asian Lang. Inf. Process.*, 8(4):14:1–14:22.

[24] Gemini Team, G. D. G. (2023). Report: Gemini: A Family of Highly Capable Multimodal Models.

[25] Ghosal, D., Majumder, N., Mihalcea, R., and Poria, S. (2021). STaCK: Sentence Ordering with Temporal Commonsense Knowledge. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8676–8686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[26] Ghosal, D., Majumder, N., Mihalcea, R., and Poria, S. (2022). Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10158–10166, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[27] Goertzel, B., Geisweiller, N., Coelho, L., Janicic, P., and Pennachin, C. (2011). Temporal Reasoning. In Goertzel, B., Geisweiller, N., Coelho, L., Janičić, P., and Pennachin, C., editors, *Real-World Reasoning: Toward Scalable, Uncertain Spatiotemporal, Contextual and Causal Inference*, pages 79–97. Atlantis Press, Paris.

[28] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Google-Books-ID: omivDQAAQBAJ.

[29] Grave, , Joulin, A., Cissé, M., Facebook AI Research, D. G., and Jégou, H. (2017). Efficient softmax approximation for GPUs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1302–1310, Sydney, NSW, Australia. JMLR.org.

[30] Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[31] Haffar, N., Hkiri, E., and Zrigui, M. (2019). TimeML Annotation of Events and Temporal Expressions in Arabic Texts. In Nguyen, N. T., Chbeir, R., Exposito, E., Aniorté, P., and Trawiński, B., editors, *Computational Collective Intelligence*, Lecture Notes in Computer Science, pages 207–218, Cham. Springer International Publishing.

[32] Han, R., Ren, X., and Peng, N. (2021). ECONET: Effective Continual Pretraining of Language Models for Event Temporal Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[33] He, P., Gao, J., and Chen, W. (2022). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

[34] He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: decoding-Enhanced Bert with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

[35] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554.

[36] Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., Chen, Z., Alharthi, A., An, B., He, J., Liu, Z., Zhang, Z., Chen, J., Li, J., Wang, B., Zhang, L., Sun, R., Wan, X., Li, H., and Xu, J. (2023). AceGPT, Localizing Large Language Models in Arabic. arXiv:2309.12053 [cs].

[37] Huang, P., Bu, J., Chen, C., and Kang, Z. (2007). Question Classification via Multiclass Kernel-based Vector Machines. In *2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 336–341.

[38] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

[39] Jain, R., Sojitra, D., Acharya, A., Saha, S., Jatowt, A., and Dandapat, S. (2023). Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore. Association for Computational Linguistics.

[40] Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. (2018). TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM ’18, pages 1807–1810, New York, NY, USA. ACM.

[41] Kaddari, Z., Mellah, Y., Berrich, J., Bouchentouf, T., and Belkasmi, M. G. (2020). Applying the T5 language model and duration units normalization to address temporal common sense understanding on the MCTACO dataset. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–4.

[42] Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6:100048.

[43] Kimura, M., Kanashiro Pereira, L., and Kobayashi, I. (2021). Towards a Language Model for Temporal Commonsense Reasoning. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 78–84, Online. INCOMA Ltd.

[44] Kimura, M., Kanashiro Pereira, L., and Kobayashi, I. (2022a). Toward Building a Language Model for Understanding Temporal Commonsense. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 17–24, Online. Association for Computational Linguistics.

[45] Kimura, M., Pereira, L. K., and Kobayashi, I. (2022b). Effective Masked Language Modeling for Temporal Commonsense Reasoning. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–4.

[46] Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

[47] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*.

[48] Luong, T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

[49] López Espejel, J., Ettifouri, E. H., Yahaya Alassan, M. S., Chouham, E. M., and Dahhane, W. (2023). GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032.

[50] Meng, Y., Rumshisky, A., and Romanov, A. (2017). Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Copenhagen, Denmark. Association for Computational Linguistics.

[51] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

[52] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large Language Models: A Survey. Publisher: arXiv Version Number: 2.

[53] Nicula, B., Ruseti, S., and Rebedea, T. (2018). Improving Deep Learning for Multiple Choice Question Answering with Candidate Contexts. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 678–683, Cham. Springer International Publishing.

[54] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, , Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, , Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). GPT-4 Technical Report. arXiv:2303.08774 [cs].

[55] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[56] Pereira, L., Cheng, F., Asahara, M., and Kobayashi, I. (2021). ALICE++: Adversarial Training for Robust and Effective Temporal Reasoning. In Hu, K., Kim, J.-B., Zong, C., and Chersoni, E., editors, *Proceedings of the 35th Pacific Asia Conference on Lan-*

guage, Information and Computation, pages 373–382, Shanghai, China. Association for Computational Lingustics.

[57] Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., and Artetxe, M. (2022). Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

[58] Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

[59] Pikuliak, M., Šimko, M., and Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

[60] Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing Annotation Schemes: From Theory to Model. In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 21–72. Springer Netherlands, Dordrecht.

[61] Radford, A. and Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

[62] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

[63] Randell, D. A., Cui, Z., and Cohn, A. G. (1992). A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, KR'92, pages 165–176, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[64] Sap, M., Bras, R. L., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 3027–3035.

[65] Sap, M., Shwartz, V., Bosselut, A., Choi, Y., and Roth, D. (2020). Commonsense Reasoning for Natural Language Processing. In Savary, A. and Zhang, Y., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

[66] Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):420.

[67] Schockaert, S., Ahn, D., Cock, M. D., and Kerre, E. E. (2006). Question Answering with Imperfect Temporal Information. In *Flexible Query Answering Systems*, Lecture Notes in Computer Science, pages 647–658. Springer, Berlin, Heidelberg.

[68] Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., Afzal, O. M., Kamboj, S., Pandit, O., Pal, R., Pradhan, L., Mujahid, Z. M., Baali, M., Han, X., Bsharat, S. M., Aji, A. F., Shen, Z., Liu, Z., Vassilieva, N., Hestness, J., Hock, A., Feldman, A., Lee, J., Jackson, A., Ren, H. X., Nakov, P., Baldwin, T., and Xing, E. (2023). Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. arXiv:2308.16149 [cs].

[69] Shwartz, V. (2022). Good Night at 4 pm Time Expressions in Different Cultures. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

[70] Sioutis, M. and Meng, H. (2021). Towards Robust Qualitative Spatio-Temporal Reasoning for Hybrid AI Systems. In *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 426–430.

[71] Speer, R. and Havasi, C. (2013). ConceptNet 5: A Large Semantic Network for Relational Knowledge. pages 161–176, Berlin, Heidelberg. Springer Berlin Heidelberg. Book Title: The People's Web Meets NLP Series Title: Theory and Applications of Natural Language Processing.

[72] Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

[73] Tawalbeh, S. and AL-Smadi, M. (2020). Is this sentence valid? An Arabic Dataset for Commonsense Validation. arXiv:2008.10873 [cs].

[74] Tay, Y., Tuan, L. A., and Hui, S. C. (2018). Hyperbolic Representation Learning for Fast and Efficient Neural Question Answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 583–591, New York, NY, USA. ACM.

[75] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[76] Virgo, F., Cheng, F., and Kurohashi, S. (2022). Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457, Marseille, France. European Language Resources Association.

[77] Wang, W., Fang, T., Ding, W., Xu, B., Liu, X., Song, Y., and Bosselut, A. (2023). CAR: Conceptualization-Augmented Reasoner for Zero-Shot Commonsense Question Answering. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13520–13545, Singapore. Association for Computational Linguistics.

[78] Wang, W., Zheng, V. W., Yu, H., and Miao, C. (2019). A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):13:1–13:37.

[79] Xu, N., Gui, T., Ma, R., Zhang, Q., Ye, J., Zhang, M., and Huang, X. (2022). Cross-Linguistic Syntactic Difference in Multilingual BERT: How Good is It and How Does It Affect Transfer? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[80] Yang, Z., Du, X., Rush, A., and Cardie, C. (2020). Improving Event Duration Prediction via Time-aware Pre-training. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3370–3378, Online. Association for Computational Linguistics.

[81] Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

[82] Zeng, C., Li, S., Li, Q., Hu, J., and Hu, J. (2020). A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10(21):7640. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

[83] Zhou, B., Khashabi, D., Ning, Q., and Roth, D. (2019). “Going on a vacation” takes longer than “Going for a walk”: A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

[84] Zhou, B., Ning, Q., Khashabi, D., and Roth, D. (2020). Temporal Common Sense Acquisition with Minimal Supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

[85] Zhou, M., Huang, M., and Zhu, X. (2018). An Interpretable Reasoning Network for Multi-Relation Question Answering. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2010–2022, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendix A

Hyper-parameter

For all PLMs, the hyperparameters are as follows:

Epocs: 10

Learning-rate: 2e-5

Batch-size: 32

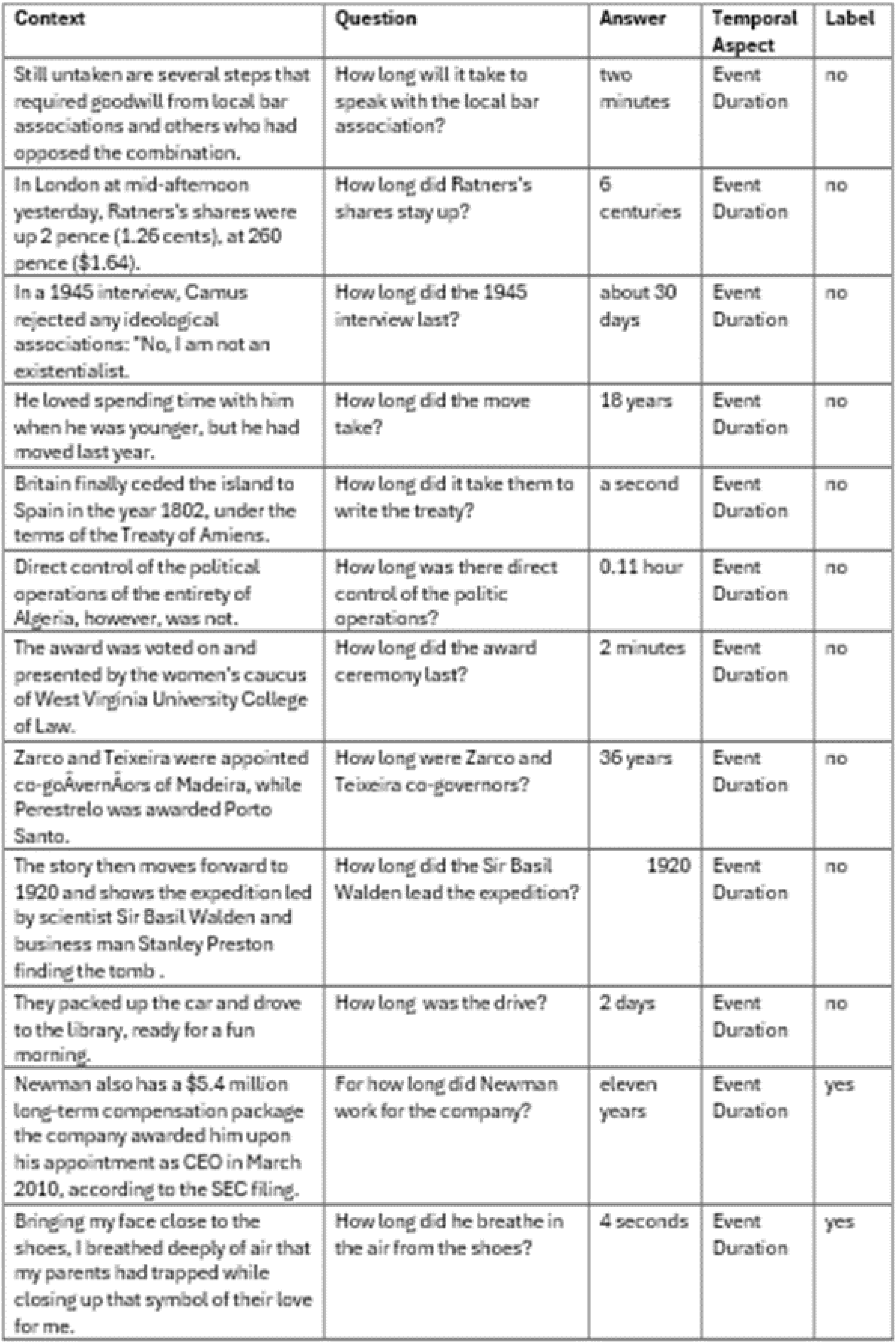
Batch-size for evaluation: 8

random seed: 42

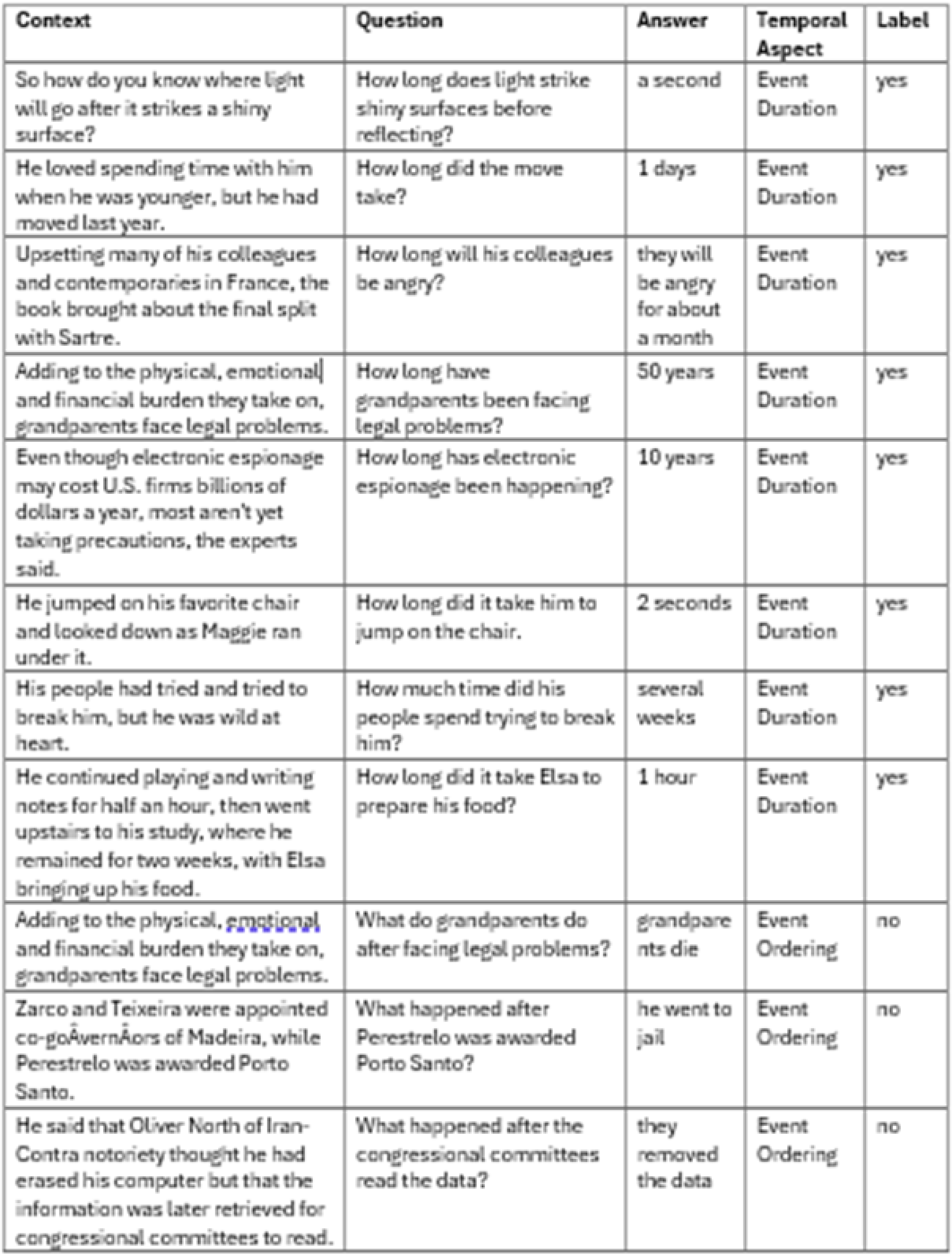
Appendix B

The Sample of the Datasets for Conducting LLMs Evaluations

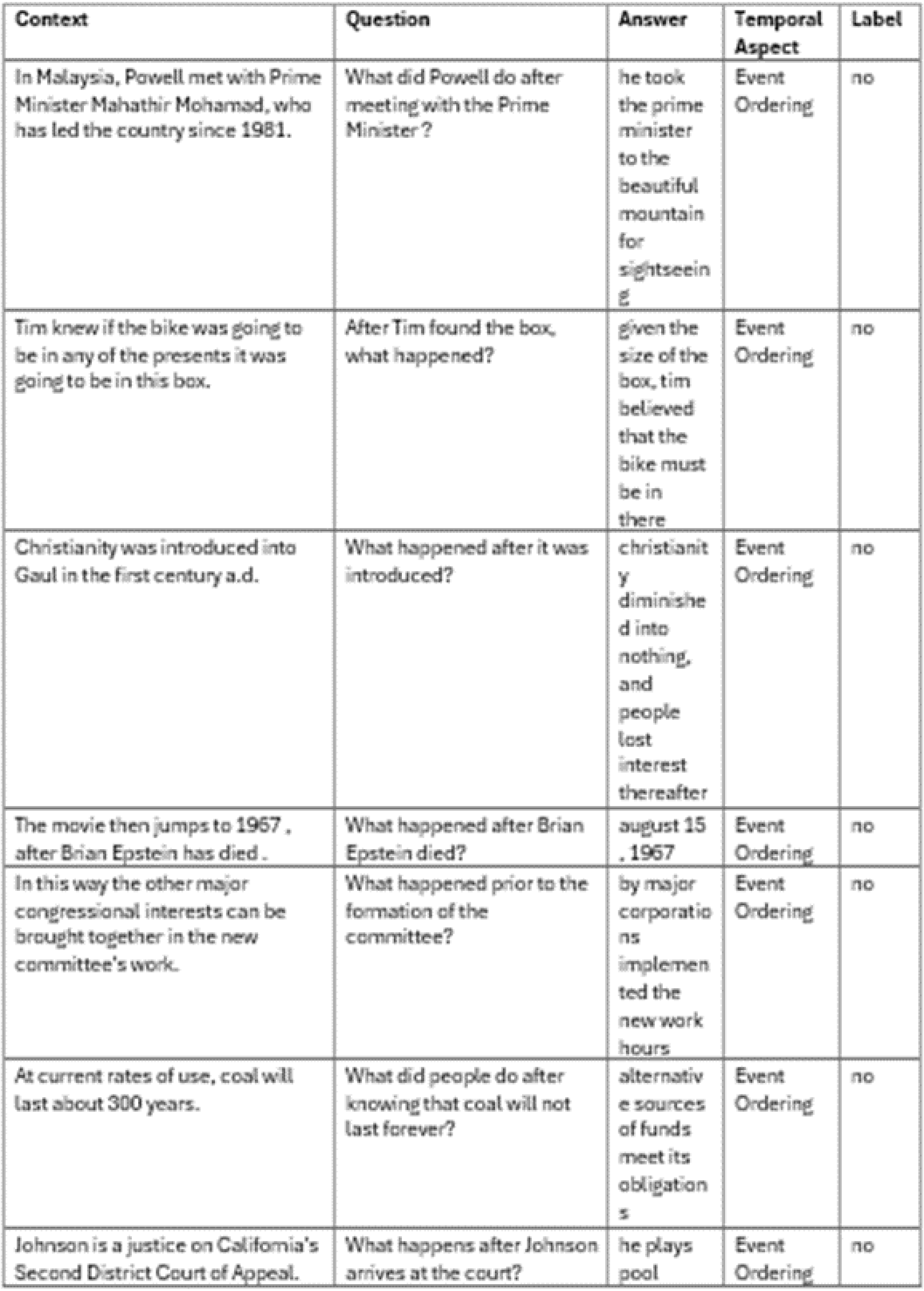
Table B.1 Sample of Dataset that Used for LLMs Evaluation



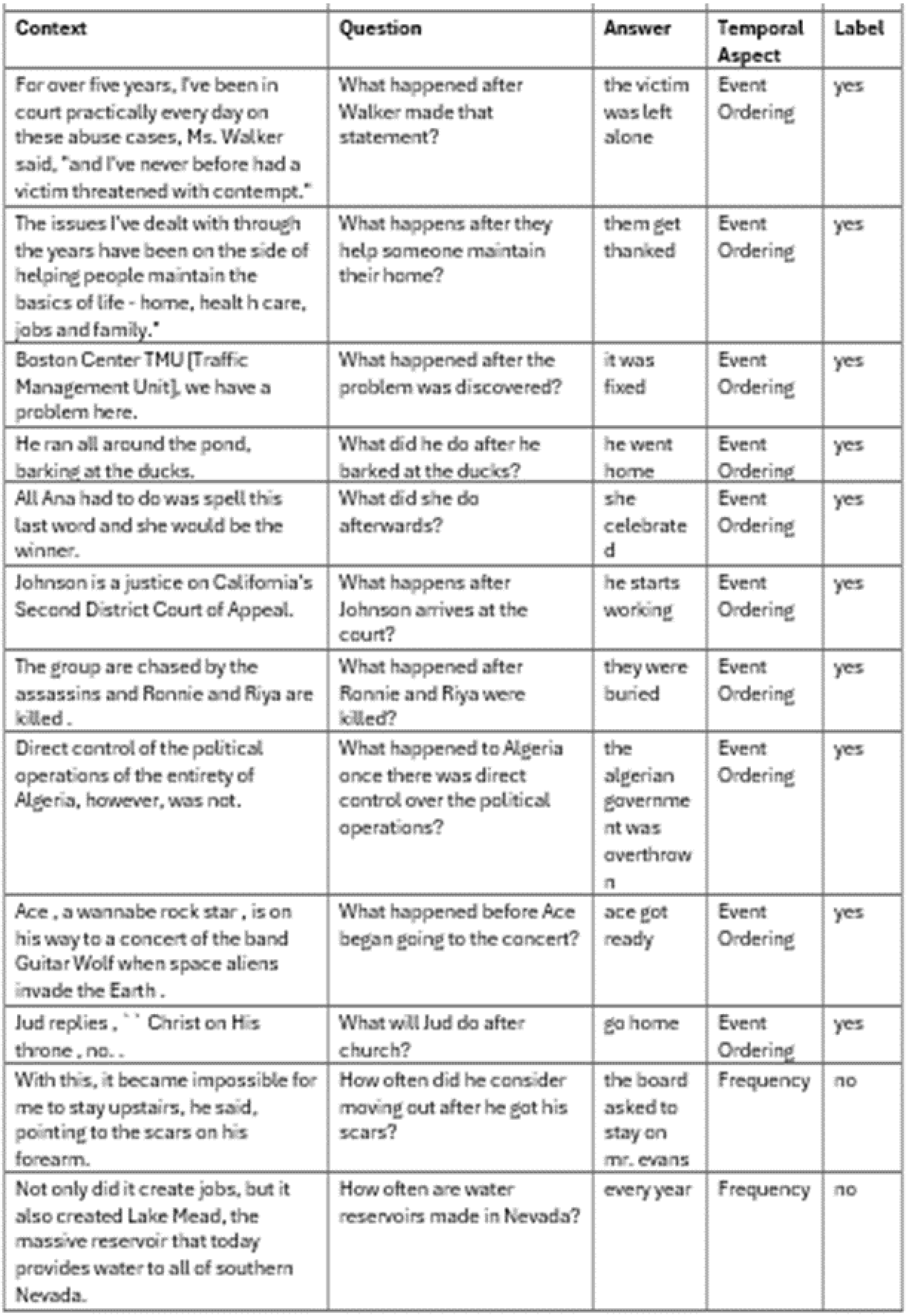
Context	Question	Answer	Temporal Aspect	Label
Still untaken are several steps that required goodwill from local bar associations and others who had opposed the combination.	How long will it take to speak with the local bar association?	two minutes	Event Duration	no
In London at mid-afternoon yesterday, Ratners's shares were up 2 pence (1.26 cents), at 260 pence ($1.64).	How long did Ratners's shares stay up?	6 centuries	Event Duration	no
In a 1945 interview, Camus rejected any ideological associations: "No, I am not an existentialist.	How long did the 1945 interview last?	about 30 days	Event Duration	no
He loved spending time with him when he was younger, but he had moved last year.	How long did the move take?	18 years	Event Duration	no
Britain finally ceded the island to Spain in the year 1802, under the terms of the Treaty of Amiens.	How long did it take them to write the treaty?	a second	Event Duration	no
Direct control of the political operations of the entirety of Algeria, however, was not.	How long was there direct control of the politic operations?	0.11 hour	Event Duration	no
The award was voted on and presented by the women's caucus of West Virginia University College of Law.	How long did the award ceremony last?	2 minutes	Event Duration	no
Zarco and Teixeira were appointed co-goÂvernÂors of Madeira, while Perestrelo was awarded Porto Santo.	How long were Zarco and Teixeira co-governors?	36 years	Event Duration	no
The story then moves forward to 1920 and shows the expedition led by scientist Sir Basil Walden and business man Stanley Preston finding the tomb .	How long did the Sir Basil Walden lead the expedition?	1920	Event Duration	no
They packed up the car and drove to the library, ready for a fun morning.	How long was the drive?	2 days	Event Duration	no
Newman also has a $5.4 million long-term compensation package the company awarded him upon his appointment as CEO in March 2010, according to the SEC filing.	For how long did Newman work for the company?	eleven years	Event Duration	yes
Bringing my face close to the shoes, I breathed deeply of air that my parents had trapped while closing up that symbol of their love for me.	How long did he breathe in the air from the shoes?	4 seconds	Event Duration	yes



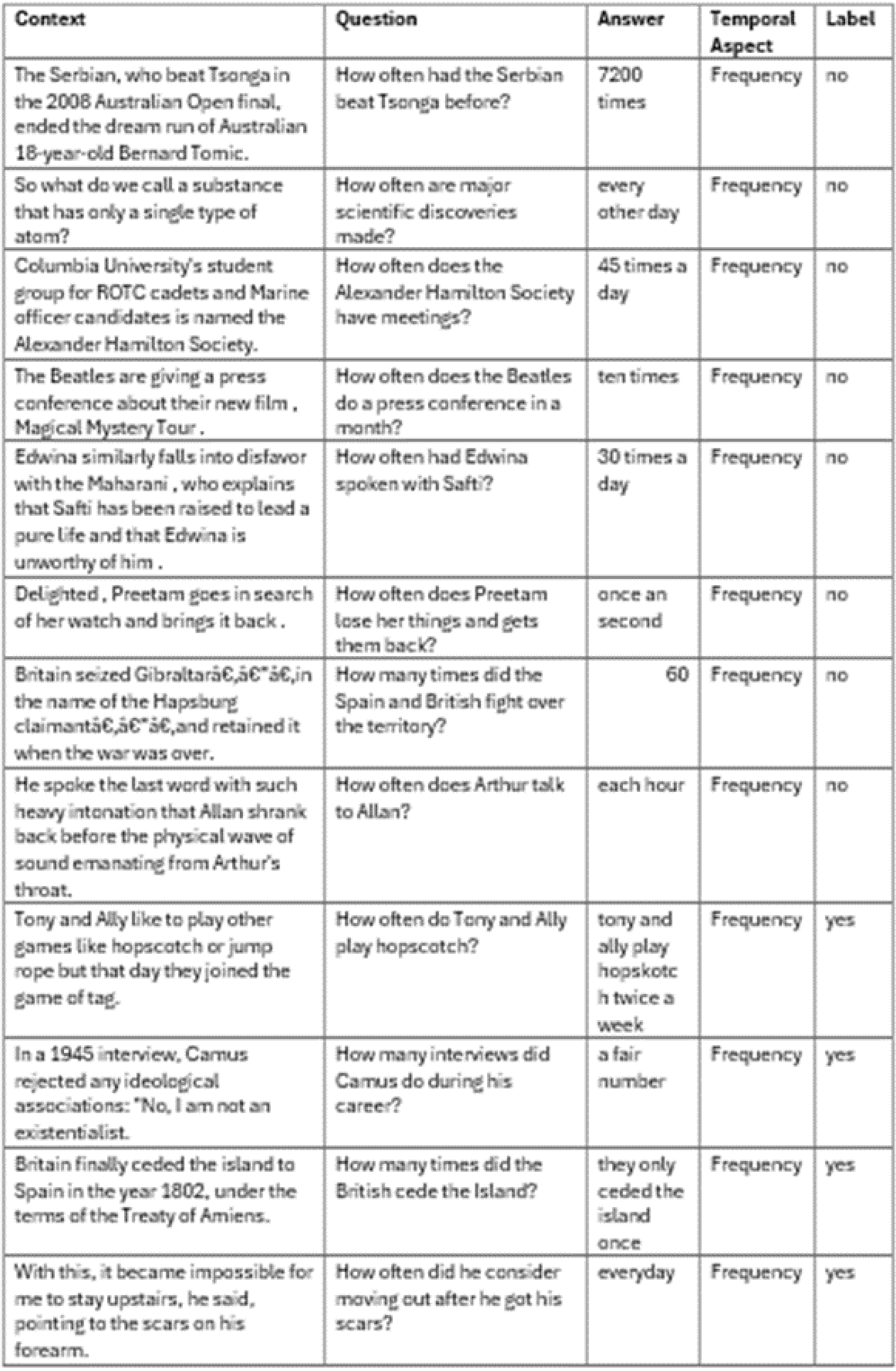
Context	Question	Answer	Temporal Aspect	Label
So how do you know where light will go after it strikes a shiny surface?	How long does light strike shiny surfaces before reflecting?	a second	Event Duration	yes
He loved spending time with him when he was younger, but he had moved last year.	How long did the move take?	1 days	Event Duration	yes
Upsetting many of his colleagues and contemporaries in France, the book brought about the final split with Sartre.	How long will his colleagues be angry?	they will be angry for about a month	Event Duration	yes
Adding to the physical, emotional and financial burden they take on, grandparents face legal problems.	How long have grandparents been facing legal problems?	50 years	Event Duration	yes
Even though electronic espionage may cost U.S. firms billions of dollars a year, most aren't yet taking precautions, the experts said.	How long has electronic espionage been happening?	10 years	Event Duration	yes
He jumped on his favorite chair and looked down as Maggie ran under it.	How long did it take him to jump on the chair.	2 seconds	Event Duration	yes
His people had tried and tried to break him, but he was wild at heart.	How much time did his people spend trying to break him?	several weeks	Event Duration	yes
He continued playing and writing notes for half an hour, then went upstairs to his study, where he remained for two weeks, with Elsa bringing up his food.	How long did it take Elsa to prepare his food?	1 hour	Event Duration	yes
Adding to the physical, emotional and financial burden they take on, grandparents face legal problems.	What do grandparents do after facing legal problems?	grandparents die	Event Ordering	no
Zarco and Teixeira were appointed co-goÂvernÂors of Madeira, while Perestrelo was awarded Porto Santo.	What happened after Perestrelo was awarded Porto Santo?	he went to jail	Event Ordering	no
He said that Oliver North of Iran-Contra notoriety thought he had erased his computer but that the information was later retrieved for congressional committees to read.	What happened after the congressional committees read the data?	they removed the data	Event Ordering	no



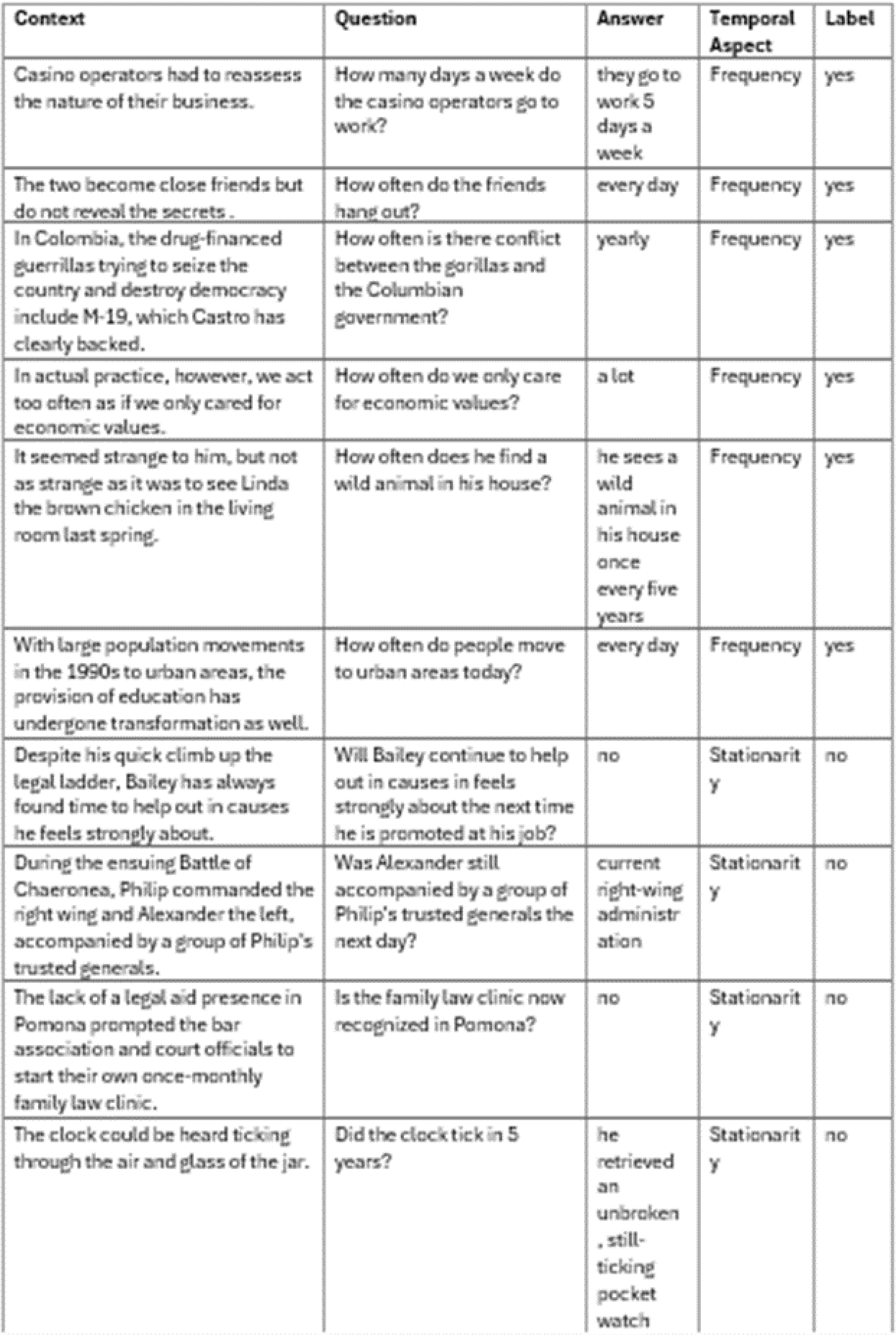
Context	Question	Answer	Temporal Aspect	Label
In Malaysia, Powell met with Prime Minister Mahathir Mohamad, who has led the country since 1981.	What did Powell do after meeting with the Prime Minister ?	he took the prime minister to the beautiful mountain for sightseeing	Event Ordering	no
Tim knew if the bike was going to be in any of the presents it was going to be in this box.	After Tim found the box, what happened?	given the size of the box, tim believed that the bike must be in there	Event Ordering	no
Christianity was introduced into Gaul in the first century a.d.	What happened after it was introduced?	christianity diminished into nothing, and people lost interest thereafter	Event Ordering	no
The movie then jumps to 1967 , after Brian Epstein has died .	What happened after Brian Epstein died?	august 15 , 1967	Event Ordering	no
In this way the other major congressional interests can be brought together in the new committee's work.	What happened prior to the formation of the committee?	by major corporations implemented the new work hours	Event Ordering	no
At current rates of use, coal will last about 300 years.	What did people do after knowing that coal will not last forever?	alternative sources of funds meet its obligations	Event Ordering	no
Johnson is a justice on California's Second District Court of Appeal.	What happens after Johnson arrives at the court?	he plays pool	Event Ordering	no



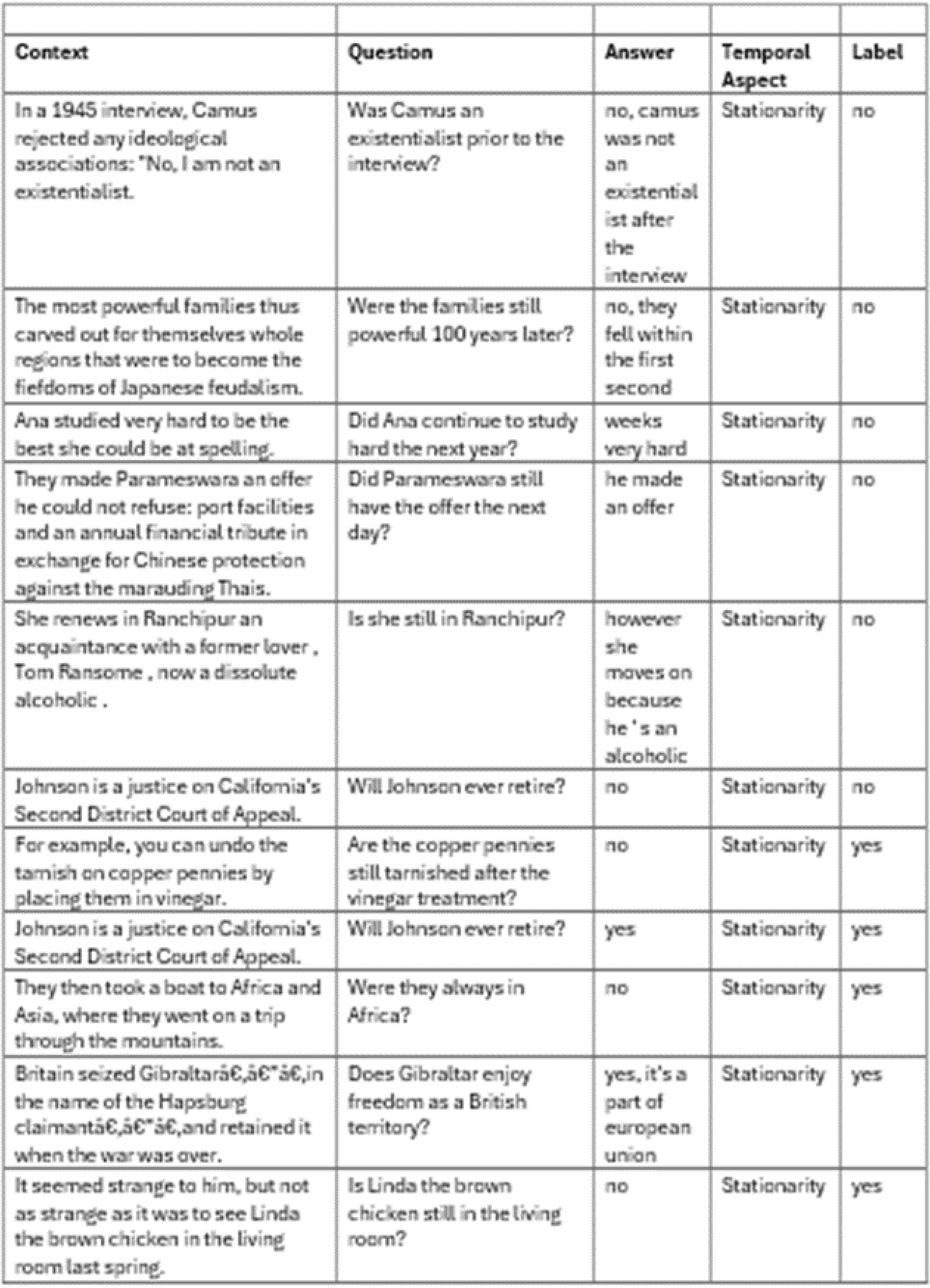
Context	Question	Answer	Temporal Aspect	Label
For over five years, I've been in court practically every day on these abuse cases, Ms. Walker said, "and I've never before had a victim threatened with contempt."	What happened after Walker made that statement?	the victim was left alone	Event Ordering	yes
The issues I've dealt with through the years have been on the side of helping people maintain the basics of life - home, healt h care, jobs and family."	What happens after they help someone maintain their home?	them get thanked	Event Ordering	yes
Boston Center TMU [Traffic Management Unit], we have a problem here.	What happened after the problem was discovered?	it was fixed	Event Ordering	yes
He ran all around the pond, barking at the ducks.	What did he do after he barked at the ducks?	he went home	Event Ordering	yes
All Ana had to do was spell this last word and she would be the winner.	What did she do afterwards?	she celebrated	Event Ordering	yes
Johnson is a justice on California's Second District Court of Appeal.	What happens after Johnson arrives at the court?	he starts working	Event Ordering	yes
The group are chased by the assassins and Ronnie and Riya are killed .	What happened after Ronnie and Riya were killed?	they were buried	Event Ordering	yes
Direct control of the political operations of the entirety of Algeria, however, was not.	What happened to Algeria once there was direct control over the political operations?	the algerian government was overthrown	Event Ordering	yes
Ace , a wannabe rock star , is on his way to a concert of the band Guitar Wolf when space aliens invade the Earth .	What happened before Ace began going to the concert?	ace got ready	Event Ordering	yes
Jud replies , ` ` Christ on His throne , no. .	What will Jud do after church?	go home	Event Ordering	yes
With this, it became impossible for me to stay upstairs, he said, pointing to the scars on his forearm.	How often did he consider moving out after he got his scars?	the board asked to stay on mr. evans	Frequency	no
Not only did it create jobs, but it also created Lake Mead, the massive reservoir that today provides water to all of southern Nevada.	How often are water reservoirs made in Nevada?	every year	Frequency	no



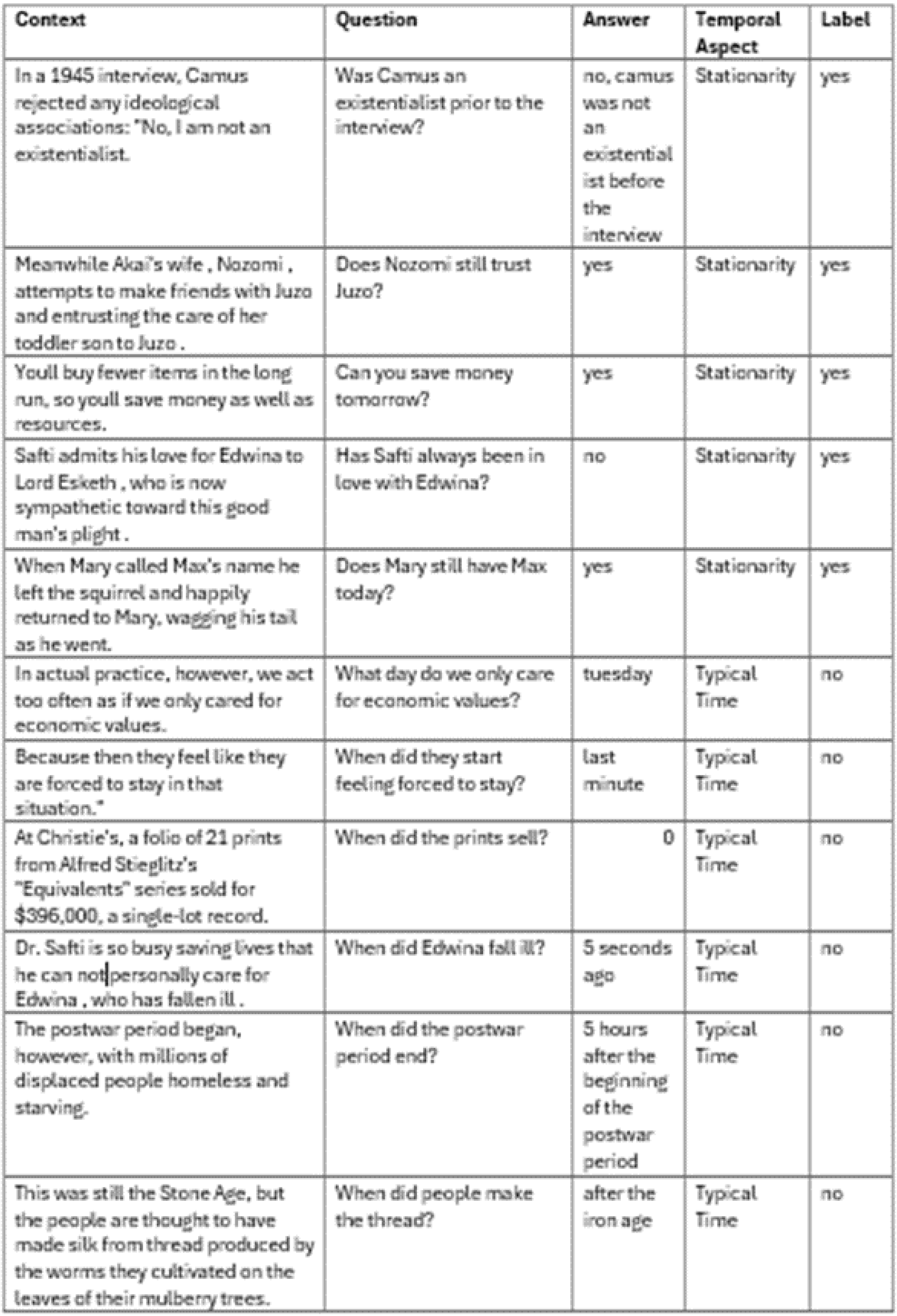
Context	Question	Answer	Temporal Aspect	Label
The Serbian, who beat Tsonga in the 2008 Australian Open final, ended the dream run of Australian 18-year-old Bernard Tomic.	How often had the Serbian beat Tsonga before?	7200 times	Frequency	no
So what do we call a substance that has only a single type of atom?	How often are major scientific discoveries made?	every other day	Frequency	no
Columbia University's student group for ROTC cadets and Marine officer candidates is named the Alexander Hamilton Society.	How often does the Alexander Hamilton Society have meetings?	45 times a day	Frequency	no
The Beatles are giving a press conference about their new film , Magical Mystery Tour .	How often does the Beatles do a press conference in a month?	ten times	Frequency	no
Edwina similarly falls into disfavor with the Maharani , who explains that Safti has been raised to lead a pure life and that Edwina is unworthy of him .	How often had Edwina spoken with Safti?	30 times a day	Frequency	no
Delighted , Preetam goes in search of her watch and brings it back .	How often does Preetam lose her things and gets them back?	once an second	Frequency	no
Britain seized Gibraltarâ€‚â€"â€‚in the name of the Hapsburg claimantâ€‚â€"â€‚and retained it when the war was over.	How many times did the Spain and British fight over the territory?	60	Frequency	no
He spoke the last word with such heavy intonation that Allan shrank back before the physical wave of sound emanating from Arthur's throat.	How often does Arthur talk to Allan?	each hour	Frequency	no
Tony and Ally like to play other games like hopscotch or jump rope but that day they joined the game of tag.	How often do Tony and Ally play hopscotch?	tony and ally play hopskotch twice a week	Frequency	yes
In a 1945 interview, Camus rejected any ideological associations: "No, I am not an existentialist.	How many interviews did Camus do during his career?	a fair number	Frequency	yes
Britain finally ceded the island to Spain in the year 1802, under the terms of the Treaty of Amiens.	How many times did the British cede the Island?	they only ceded the island once	Frequency	yes
With this, it became impossible for me to stay upstairs, he said, pointing to the scars on his forearm.	How often did he consider moving out after he got his scars?	everyday	Frequency	yes



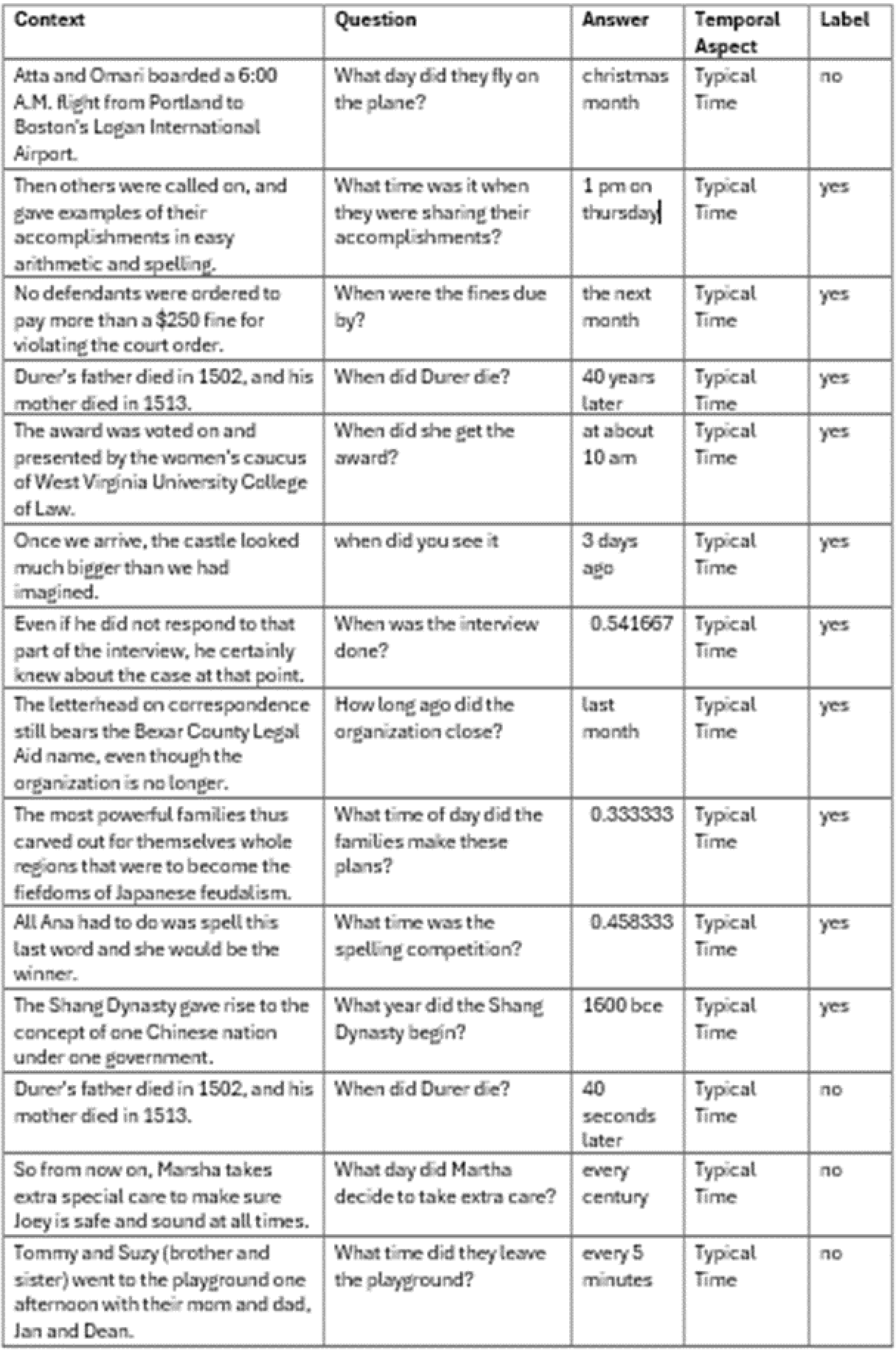
Context	Question	Answer	Temporal Aspect	Label
Casino operators had to reassess the nature of their business.	How many days a week do the casino operators go to work?	they go to work 5 days a week	Frequency	yes
The two become close friends but do not reveal the secrets .	How often do the friends hang out?	every day	Frequency	yes
In Colombia, the drug-financed guerrillas trying to seize the country and destroy democracy include M-19, which Castro has clearly backed.	How often is there conflict between the gorillas and the Columbian government?	yearly	Frequency	yes
In actual practice, however, we act too often as if we only cared for economic values.	How often do we only care for economic values?	a lot	Frequency	yes
It seemed strange to him, but not as strange as it was to see Linda the brown chicken in the living room last spring.	How often does he find a wild animal in his house?	he sees a wild animal in his house once every five years	Frequency	yes
With large population movements in the 1990s to urban areas, the provision of education has undergone transformation as well.	How often do people move to urban areas today?	every day	Frequency	yes
Despite his quick climb up the legal ladder, Bailey has always found time to help out in causes he feels strongly about.	Will Bailey continue to help out in causes in feels strongly about the next time he is promoted at his job?	no	Stationarit y	no
During the ensuing Battle of Chaeronea, Philip commanded the right wing and Alexander the left, accompanied by a group of Philip's trusted generals.	Was Alexander still accompanied by a group of Philip's trusted generals the next day?	current right-wing administr ation	Stationarit y	no
The lack of a legal aid presence in Pomona prompted the bar association and court officials to start their own once-monthly family law clinic.	Is the family law clinic now recognized in Pomona?	no	Stationarit y	no
The clock could be heard ticking through the air and glass of the jar.	Did the clock tick in 5 years?	he retrieved an unbroken , still-ticking pocket watch	Stationarit y	no



Context	**Question**	**Answer**	**Temporal Aspect**	**Label**
In a 1945 interview, Camus rejected any ideological associations: "No, I am not an existentialist.	Was Camus an existentialist prior to the interview?	no, camus was not an existentialist after the interview	Stationarity	no
The most powerful families thus carved out for themselves whole regions that were to become the fiefdoms of Japanese feudalism.	Were the families still powerful 100 years later?	no, they fell within the first second	Stationarity	no
Ana studied very hard to be the best she could be at spelling.	Did Ana continue to study hard the next year?	weeks very hard	Stationarity	no
They made Parameswara an offer he could not refuse: port facilities and an annual financial tribute in exchange for Chinese protection against the marauding Thais.	Did Parameswara still have the offer the next day?	he made an offer	Stationarity	no
She renews in Ranchipur an acquaintance with a former lover , Tom Ransome , now a dissolute alcoholic .	Is she still in Ranchipur?	however she moves on because he ' s an alcoholic	Stationarity	no
Johnson is a justice on California's Second District Court of Appeal.	Will Johnson ever retire?	no	Stationarity	no
For example, you can undo the tarnish on copper pennies by placing them in vinegar.	Are the copper pennies still tarnished after the vinegar treatment?	no	Stationarity	yes
Johnson is a justice on California's Second District Court of Appeal.	Will Johnson ever retire?	yes	Stationarity	yes
They then took a boat to Africa and Asia, where they went on a trip through the mountains.	Were they always in Africa?	no	Stationarity	yes
Britain seized Gibraltarâ€‚â€"â€‚in the name of the Hapsburg claimantâ€‚â€"â€‚and retained it when the war was over.	Does Gibraltar enjoy freedom as a British territory?	yes, it's a part of european union	Stationarity	yes
It seemed strange to him, but not as strange as it was to see Linda the brown chicken in the living room last spring.	Is Linda the brown chicken still in the living room?	no	Stationarity	yes



Context	Question	Answer	Temporal Aspect	Label
In a 1945 interview, Camus rejected any ideological associations: "No, I am not an existentialist.	Was Camus an existentialist prior to the interview?	no, camus was not an existentialist before the interview	Stationarity	yes
Meanwhile Akai's wife , Nozomi , attempts to make friends with Juzo and entrusting the care of her toddler son to Juzo .	Does Nozomi still trust Juzo?	yes	Stationarity	yes
Youll buy fewer items in the long run, so youll save money as well as resources.	Can you save money tomorrow?	yes	Stationarity	yes
Safti admits his love for Edwina to Lord Esketh , who is now sympathetic toward this good man's plight .	Has Safti always been in love with Edwina?	no	Stationarity	yes
When Mary called Max's name he left the squirrel and happily returned to Mary, wagging his tail as he went.	Does Mary still have Max today?	yes	Stationarity	yes
In actual practice, however, we act too often as if we only cared for economic values.	What day do we only care for economic values?	tuesday	Typical Time	no
Because then they feel like they are forced to stay in that situation."	When did they start feeling forced to stay?	last minute	Typical Time	no
At Christie's, a folio of 21 prints from Alfred Stieglitz's "Equivalents" series sold for $396,000, a single-lot record.	When did the prints sell?	0	Typical Time	no
Dr. Safti is so busy saving lives that he can not personally care for Edwina , who has fallen ill .	When did Edwina fall ill?	5 seconds ago	Typical Time	no
The postwar period began, however, with millions of displaced people homeless and starving.	When did the postwar period end?	5 hours after the beginning of the postwar period	Typical Time	no
This was still the Stone Age, but the people are thought to have made silk from thread produced by the worms they cultivated on the leaves of their mulberry trees.	When did people make the thread?	after the iron age	Typical Time	no



Context	Question	Answer	Temporal Aspect	Label
Atta and Omari boarded a 6:00 A.M. flight from Portland to Boston's Logan International Airport.	What day did they fly on the plane?	christmas month	Typical Time	no
Then others were called on, and gave examples of their accomplishments in easy arithmetic and spelling.	What time was it when they were sharing their accomplishments?	1 pm on thursday	Typical Time	yes
No defendants were ordered to pay more than a $250 fine for violating the court order.	When were the fines due by?	the next month	Typical Time	yes
Durer's father died in 1502, and his mother died in 1513.	When did Durer die?	40 years later	Typical Time	yes
The award was voted on and presented by the women's caucus of West Virginia University College of Law.	When did she get the award?	at about 10 am	Typical Time	yes
Once we arrive, the castle looked much bigger than we had imagined.	when did you see it	3 days ago	Typical Time	yes
Even if he did not respond to that part of the interview, he certainly knew about the case at that point.	When was the interview done?	0.541667	Typical Time	yes
The letterhead on correspondence still bears the Bexar County Legal Aid name, even though the organization is no longer.	How long ago did the organization close?	last month	Typical Time	yes
The most powerful families thus carved out for themselves whole regions that were to become the fiefdoms of Japanese feudalism.	What time of day did the families make these plans?	0.333333	Typical Time	yes
All Ana had to do was spell this last word and she would be the winner.	What time was the spelling competition?	0.458333	Typical Time	yes
The Shang Dynasty gave rise to the concept of one Chinese nation under one government.	What year did the Shang Dynasty begin?	1600 bce	Typical Time	yes
Durer's father died in 1502, and his mother died in 1513.	When did Durer die?	40 seconds later	Typical Time	no
So from now on, Marsha takes extra special care to make sure Joey is safe and sound at all times.	What day did Martha decide to take extra care?	every century	Typical Time	no
Tommy and Suzy (brother and sister) went to the playground one afternoon with their mom and dad, Jan and Dean.	What time did they leave the playground?	every 5 minutes	Typical Time	no