Toward Building a Language Model for Understanding Temporal Commonsense

Mayuko Kimura[1] **Lis Kanashiro Pereira**[2] **Ichiro Kobayashi**[3]
Ochanomizu University, Japan
[1,3]{g1720512,koba}@is.ocha.ac.jp
[2]kanashiro.pereira@ocha.ac.jp

Abstract

The ability to capture temporal commonsense relationships for time-related events expressed in text is a very important task in natural language understanding. However, pre-trained language models such as BERT, which have recently achieved great success in a wide range of natural language processing tasks, are still considered to have poor performance in temporal reasoning. In this paper, we focus on the development of language models for temporal commonsense inference over several pre-trained language models. Our model relies on multi-step fine-tuning using multiple corpora and masked language modeling to predict masked temporal indicators that are crucial for temporal commonsense reasoning. We also experimented with multi-task learning and build a language model that can improve performance on multiple time-related tasks. In our experiments, multi-step fine-tuning using the general commonsense knowledge task as an auxiliary task produced the best results. We obtained a significant improvement in accuracy over standard fine-tuning in the temporal commonsense inference task and on other time-related tasks.

1 Introduction

Commonsense reasoning is crucial for natural language processing (NLP). Commonsense is the basic level of practical knowledge that is commonly shared among most people[1]. A specific type of commonsense is temporal commonsense. Temporal commonsense refers to the common knowledge about various temporal aspects of events, such as duration, frequency, and temporal order.

Capturing temporal commonsense relations for time-related events expressed in sentences is a very important task in natural language understanding. However, pre-trained language models such as BERT (Devlin et al., 2019), which have recently achieved significant results in a wide range of NLP tasks, are still said to perform poorly in temporal reasoning (Ribeiro et al., 2020). For example, given two events, "going on a vacation" and "going for a walk," most humans know that "vacation is longer and occurs less frequently than walks," or that "going on a walk is more frequent than going on a vacation. However, it is difficult for computers to make inferences based on such commonsense knowledge.

In this paper, we focus on the development of a language model for understanding temporal commonsense. In a prior study (Kimura et al., 2021), BERT was used, and in this study, we also use RoBERTa (Liu et al., 2019b) and ALBERT (Lan et al., 2019), which are improved models of BERT. We use them for multi-step fine-tuning using multiple corpora and continual pre-training by performing the masked language modeling (MLM) task (Devlin et al., 2019) on the target dataset. MLM task is a fill-in-the-blank task that has been employed as a pre-training task for various language models.

For multi-step fine-tuning, we thought an additional stage of fine-tuning on an intermediate related supervised task might help improve performance because temporal datasets usually have only a small amount of training data available. For continual pre-training on the target dataset, we aimed to resolve the domain mismatch between the pre-trained models and the target task, and make the model better weight temporal indicators and event triggers for our downstream tasks. In addition, we apply multi-task learning to further improve our model's generalization performance.

Our contributions are summarized as follows:

- We propose a language model for understanding temporal commonsense that effectively leverages continual pre-training, multi-step fine-tuning, and multi-task learning.
- We conducted multi-step fine-tuning and continual pre-training by performing the MLM

[1]https://csrr-workshop.github.io/

Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 17–24

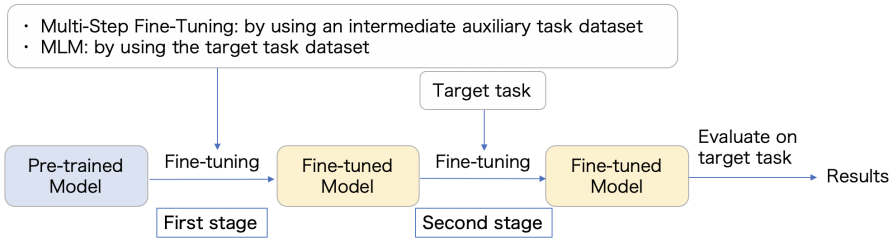

Figure 1: Overview of the multi-step fine-tuning and continual pre-training methods.

task on the target dataset on three pre-trained language models (BERT, RoBERTa, and ALBERT).

- We achieved the best performance with multi-step fine-tuning using the general commonsense knowledge task as auxiliary task on ALBERT.
- Although we focus on temporal commonsense reasoning, we also examined and confirmed the effectiveness of our multi-task learning model on several other temporal-related tasks.

2 Related Work

Although research on temporal inference has been conducted for a long time, in recent years, many studies have been proposed on temporal expression extraction (Lee et al., 2014; Vashishtha et al., 2019), temporal relation extraction (Ning et al., 2017, 2018b), and the construction of timelines (Leeuwenberg and Moens, 2018). As for temporal commonsense, there are studies focusing on the duration of events (Vempala et al., 2018; Vashishtha et al., 2019), the temporal order of events (Ning et al., 2018a), and so forth. Zhou et al. (2020) proposed methods for constructing language models that produce representations of events for relevant tasks such as duration comparison, parent-child relations, event coreference and temporal question-answering tasks.

In particular, some recent works have focused on the construction of challenging benchmarks for temporal commonsense inference. The Story Cloze Test (Mostafazadeh et al., 2016) dataset focuses on the typical temporal and causal relationships between events. TORQUE (Ning et al., 2020) is a machine reading comprehension dataset that focuses on the temporal ordering of events. MC-TACO (Zhou et al., 2019) is a challenging multiple choice temporal commonsense reasoning task that focuses on temporal properties such as duration and ordering of events. TIMEDIAL (Qin et al., 2021) is a dataset consisting of dialogues containing temporal information and is a complex temporal commonsense inference task using multi-turn dialogues.

In addition, pre-trained language models such as BERT have succeeded on broad-coverage probing benchmarks. However, in the case of domain mismatch between the pre-trained model and the target task, these models may still suffer catastrophic accuracy degradation.

In this study, we focus on temporal commonsense reasoning and attempt to improve the performance of the pre-trained language model for understanding temporal commonsense. Our model effectively leverages continual pre-training, multi-step fine-tuning, and multi-task learning. It substantially outperforms the standard fine-tuning approach.

3 Temporal Commonsense Reasoning Task: MC-TACO

MC-TACO is a dataset that entirely focuses on a specific reasoning capability: temporal commonsense. MC-TACO considers five temporal properties: (1) duration (how long an event takes), (2) temporal ordering (typical order of events), (3) typical time (when an event occurs), (4) frequency (how often an event occurs), and (5) stationarity (whether a state is maintained for a very long time or indefinitely). It contains 13k tuples, each consisting of a sentence, a question, and a candidate answer, that should be judged as plausible or not. The sentences are taken from different sources such as news, Wikipedia and textbooks. An example from this dataset is below. The correct answers are in **bold**.

Paragraph: He layed down on the chair and pawed at her as she ran in a circle under it.
Question: How long did he paw at her?
a) **2 minutes** b) 2 days
c) 90 minutes e) **7 seconds**
Reasoning Type: Duration

We mainly use the MC-TACO dataset for evaluating the performance of our model. In the later

sections, we also show evaluation on additional temporal-related tasks.

4 Methods

We focus on exploring different training techniques, i.e., multi-step fine-tuning, continual pre-training, and multi-task learning, for building our language model for understanding temporal commonsense. Each technique is detailed below.

4.1 Multi-Step Fine-Tuning

Multi-step fine-tuning (4.1) aims to supplement the language model pre-training with an intermediate fine-tuning stage on supervised tasks that are related to the target dataset. It has been shown to improve model robustness and performance, especially for data-constrained scenarios (Phang et al., 2018; Camburu et al., 2019). We first fine-tune models on carefully selected auxiliary tasks and datasets. This model's parameters are further refined by fine-tuning on the MC-TACO dataset.

4.2 Continual pre-training on the target dataset

As mentioned in Section 2, performing continual pre-training using the target dataset can be useful to adapt the pre-trained model to the target task. Based on this, we have applied the MLM task (Devlin et al., 2019) using MC-TACO on pre-trained language models before performing standard fine-tuning. The MLM task, which is used in the pre-training of language models, is performed by randomly replacing a subset of tokens by a special token (e.g., [MASK]), and asks the model to predict them.

An overview of the multi-step fine-tuning and continual pre-training methods is shown in Figure 1.

4.3 Multi-Task Learning

Multi-task learning (MTL) aims to improve the generalization performance of the model by learning multiple related tasks simultaneously. It has become increasingly popular in NLP because it can improve the performance of related tasks by exploiting their commonalities and differences (Zhang et al., 2022). In this study, we use MT-DNN (Liu et al., 2019a) to perform MTL and evaluate the model's performance on multiple time-related tasks. MT-DNN is a multi-task learning framework that can incorporate models such as BERT and RoBERTa as the shared text encoding layers (shared across all tasks), while the top layers are task-specific. We used the pre-trained BERT, RoBERTa, and ALBERT models to initialize its shared layers and refined them via MTL on multiple time-related tasks.

	BERT (large)	RoBERTa (large)	ALBERT (xxlarge)
Parameters	334M	355M	235M
Layers	24	24	12
Hidden	1024	1024	4096
Embedding	1024	1024	128
Pre-training data size	16GB	160GB	16GB

Table 1: Summary of each pre-trained language model used in our experiments.

5 Experiments

5.1 Text Encoders

In our previous study (Kimura et al., 2021), we used BERT-base as the text encoder. In this study, we explore the use of BERT-large, RoBERTa-large and ALBERT-xxlarge models. RoBERTa is an improved version of BERT, and has succeeded in significantly improving on BERT's accuracy by adjusting the hyperparameters, changing the pre-training method, and increasing the amount of data for training, while keeping BERT's mechanism intact. ALBERT is also an improved model of BERT, and is a lightweight, high-performance language model that has surpassed the accuracy of BERT by changing the type of pre-training task and how to handle parameters. The summary of each pre-trained language model is shown in Table 1. In pre-training, BERT and ALBERT use the English Wikipedia and BookCorpus, and RoBERTa uses CC-News, OpenWebText and Stories datasets in addition to them (Liu et al., 2019b).

5.2 Datasets

We use MC-TACO as the main training and evaluation dataset. In addition, we use the TimeML, CosmosQA, and SWAG datasets as auxiliary datasets in the multi-step fine-tuning setting. A summary of each dataset is provided below and in Table 2.

TimeML (Pan et al., 2006): This dataset is specifically about duration of an event in a span of text. The task is to decide whether a given event has a duration longer or shorter than a day. An example from this dataset showing a sentence with an event (in bold) that has a duration shorter than a day is below:

	train	val	test	huggingface model implementation	MT-DNN implementation
MC-TACO	-	3,783	9,442	*ForSequenceClassification	Pairwise Text Classification
TimeML	1,248	-	1,003	*ForSequenceClassification	Pairwise Text Classification
MATRES	12,716	-	838	*ForSequenceClassification	Single-Sentence Classification
CosmosQA	25,588	3,000	7,000	*ForMultipleChoice	Relevance Ranking
SWAG	73,546	20,006	20,005	*ForMultipleChoice	Relevance Ranking

Table 2: Summary of the datasets and their model implementations used in our experiments. We use huggingface for the multi-step fine-tuning and continual pre-training experiments, and MT-DNN for the multi-task learning experiments. The * symbol in the huggingface model implementation column stands for Bert, Roberta or Albert, depending on the text encoder we use. When using MT-DNN, we use the Single-Sentence Classification, Pairwise Text Classification, or Relevance Ranking implementations.

In Singapore, stocks **hit** a five year low.

CosmosQA (Huang et al., 2019): We propose to enrich the temporal commonsense reasoning task training by leveraging data from the general commonsense knowledge task. Since the commonsense reasoning task commonly also involves reasoning about temporal events, e.g., what event(s) might happen before or after the current event, we hypothesize that temporal reasoning might benefit from it. CosmosQA is a general commonsense knowledge task. This task focuses on reading between the lines of a story where the causes and effects of events are not explicitly mentioned and is a four-choice multiple-choice question. An example from the CosmosQA dataset is below. The correct answer is in **bold**.

Paragraph: Did some errands today. My prime objectives were to get textbooks, find computer lab, find career services, get some groceries, turn in payment plan application, and find out when KEES money kicks in. I think it acts as a refund at the end of the semester at Murray, but I would be quite happy if it would work now.
Question: What happens after I get the refund?
Option 1: **I can pay my bills.**
Option 2: I can relax.
Option 3: I can sleep.
Option 4: None of the above choices.

SWAG (Zellers et al., 2018): SWAG is also a general commonsense knowledge task. The task is to choose the correct ending among four options that leverages commonsense knowledge. An example from this dataset is below. The correct answer is in **bold**.

Question: On stage, a woman takes a seat at the piano. She
Option 1: sits on a bench as her sister plays with the doll.
Option 2: smiles with someone as the music plays.
Option 3: is in the crowd, watching the dancers.
Option 4: **nervously sets her fingers on the keys.**

	max seq_len	train batch_size	num train_epoch	learning rate
BERT				
standard fine-tuning	128	16	5	1e-5
TimeML	128	16	4	2e-5
CosmosQA	256	32	1	2e-5
SWAG	256	32	2	2e-5
MLM	128	32	3	3e-5
RoBERTa				
standard fine-tuning	128	16	20	1e-5
TimeML	128	16	6	2e-5
CosmosQA	512	16	1	1e-5
SWAG	256	32	2	1e-5
MLM	128	8	3	5e-5
ALBERT				
standard fine-tuning	128	16	6	1e-5
TimeML	128	16	6	2e-5
CosmosQA	256	16	2	1e-5
SWAG	256	16	1	1e-5
MLM	128	8	3	5e-5

Table 3: Hyperparameter settings.

5.3 Implementation Details

The hyperparameter settings used in our experiments are shown in Table 3. For each dataset, we select the best parameters based on validation experiments. The parameters for MLM using the target dataset are based on the values originally used in the pre-training of the language model.

The bert-large-uncased, roberta-large and albert-xxlarge-v2 models were used, and the Exact Match (EM) and F1 scores were employed as the evaluation metrics. The EM is the probability of correctly labeling all answers to each question, and the F1-

fine-tuned on	EM [%]	F1 [%]
BERT		
standard fine-tuning	42.6 (42.9)	70.9 (71.0)
TimeML→MC-TACO	44.8 (43.7)	72.8 (70.8)
CosmosQA→MC-TACO	**46.3** (43.6)	73.4 (70.7)
SWAG→MC-TACO	46.2 **(44.7)**	**73.6 (72.6)**
RoBERTa		
standard fine-tuning	53.8 (54.4)	75.3 **(77.6)**
TimeML→MC-TACO	51.3 (51.1)	75.7 (76.1)
CosmosQA→MC-TACO	**55.6 (55.2)**	**78.1** (77.3)
SWAG→MC-TACO	53.1 (53.9)	76.1 (77.3)
ALBERT		
standard fine-tuning	55.0 (54.6)	77.1 (77.9)
TimeML→MC-TACO	51.8 (51.3)	77.9 (75.5)
CosmosQA→MC-TACO	**59.5 (58.9)**	**80.3 (78.7)**
SWAG→MC-TACO	52.8 (51.3)	77.3 (74.6)

Table 4: Test results on multi-step fine-tuning. The 5-fold cross-validation results using the validation dataset are shown in parenthesis ().

	EM [%]	F1 [%]
BERT		
standard fine-tuning	42.6 (42.9)	70.9 (71.0)
MLM (MC-TACO)	45.2 (45.0)	72.5 (71.9)
RoBERTa		
standard fine-tuning	53.8 (54.4)	75.3 (77.6)
MLM (MC-TACO)	51.2 (54.4)	76.2 (77.5)
ALBERT		
standard fine-tuning	55.0 (54.6)	77.1 (77.9)
MLM (MC-TACO)	**59.2 (58.3)**	**79.9 (78.2)**

Table 5: Test results on MLM with target dataset. The 5-fold cross-validation results using the validation dataset are shown in parenthesis ().

score measures the average overlap between one's predictions and the ground truth (Zhou et al., 2020).

The model implementations we used in our experiments are speficied in Table 2. We use huggingface for the multi-step fine-tuning and continual pre-training experiments, and MT-DNN for the multi-task learning experiments.

5.4 Results

Multi-Step Fine-Tuning

The results of the multi-step fine-tuning experiments are shown in Table 4. The results show that changing the language model from BERT to RoBERTa and ALBERT improves accuracy. Overall, the best results were obtained when we used ALBERT.

Continual pre-training on the target dataset

Table 5 shows the results when we perform MLM on the target dataset. The results show that the accuracy also improved by changing the model used from BERT to RoBERTa and ALBERT. The best results were also obtained when ALBERT was used (with an EM score of 59.2% and an F1-score of 79.9% on the test set).

Multi-Task Learning

We used MC-TACO, TimeML, CosmosQA, and MATRES (Ning et al., 2018c) as auxiliary training data and evaluated on the time-related datasets (MC-TACO, TimeML, and MATRES). MATRES is a time-related task that focuses on the ordering of events in a sentence and events annotated with a temporal relation (BEFORE, AFTER, EQUAL, VAGUE). An example of a sentence from this dataset with two events (in bold) that hold the BEFORE relation is below:

> At one point , when it **(e1:became)** clear controllers could not contact the plane, someone **(e2:said)** a prayer.

We performed MTL using ALBERT, which obtained the best results in our previous experiments, shown in Table 4 and Table 5. These results are shown in Table 6. While there was an improvement in accuracy with MTL on MATRES, there were differences on MC-TACO depending on the auxiliary dataset used for training, and no improvement on TimeML.

5.5 Discussion

The experimental results show that changing the text encoder used from BERT to RoBERTa and ALBERT improves the accuracy of both multi-step fine-tuning using an auxiliary dataset (Table 4, with an EM score on the test set increasing from 46.3% to 55.6% and 59.5%, respectively) and of continual pre-training on the target dataset (Table 5, with an EM score on the test set increasing from 45.2% to 51.2% to 59.2%, respectively). These results indicate a significant improvement over the BERT baseline. This is a natural result considering that RoBERTa and ALBERT are improved models of BERT and have better performance than BERT on benchmarks such as GLUE.

RoBERTa is an improved model of BERT, with about 10 times the data size used for pre-training. We think that pre-training on a large amount of data improves performance in solving tasks that require commonsense.

The best results were obtained when ALBERT was used (with an EM score on the test set of 59.5%, in Table 4, and an EM score on the test set of 59.2%, in Table 5). The reason for this might also be the difference in its pre-training method. ALBERT's pre-training method employs Sentence Order Prediction (SOP) in addition to MLM. SOP is a binary classification task that determines whether two text segments are in the correct order, and focuses on

Train dataset \ Evalation dataset	MC-TACO EM [%]	MC-TACO F1 [%]	TimeML acc [%]	MATRES acc [%]
MC-TACO	57.6	80.6	-	-
MC-TACO, TimeML	**58.1**	79.7	81.0	-
MC-TACO, MATRES	57.3	80.1	-	**75.4**
MC-TACO, CosmosQA	**59.2**	80.4	-	-
MC-TACO, TimeML, MATRES	56.3	78.8	79.2	**76.3**
MC-TACO, TimeML, CosmosQA	53.0	76.5	79.9	-
MC-TACO, MATRES, CosmosQA	53.6	78.6	-	**76.8**
MC-TACO, TimeML, MATRES, CosmosQA	53.4	78.2	77.7	**76.8**
TimeML	-	-	81.1	-
TimeML, MATRES	-	-	79.4	**77.2**
TimeML, CosmosQA	-	-	80.4	-
TimeML, MATRES, CosmosQA	-	-	78.8	**76.2**
MATRES	-	-	-	74.6
MATRES, CosmosQA	-	-	-	**74.7**

Table 6: Test results on MTL using MT-DNN. Single-task learning results using MT-DNN are in blue, and those exceeding the accuracy of single-task learning are in **bold**.

modeling inter-sentence coherence. We hypothesize that this pre-training task enables the model to acquire additional temporal knowledge needed to solve the MC-TACO task.

Focusing on the results of multi-step fine-tuning using RoBERTa (Table 4), we can see that the proposed method improves the standard fine-tuning accuracy in many cases (with an EM score on the test set increasing from 53.8% to 55.6%, and a F1-score on the test set increasing from 75.3% to 75.7%, 78.1%, and 76.1%), but the increase in accuracy is smaller than that of BERT and ALBERT. The reason is that RoBERTa uses a much larger number of data for pre-training than BERT or ALBERT, and a large corpus is learned at the time of pre-training, thus multi-step fine-tuning may not be effective.

Note here that EM measures how many questions a system is able to correctly label all candidate answers (Zhou et al., 2019). EM is a stricter metric and we consistently obtain lower EM scores than F1 scores in our experiments.

The results of the MTL experiments (Table 6) were somewhat unstable, with the accuracy improving in some cases (e.g., an EM score on the test set of 59.2% with the model that trains with MC-TACO and CosmosQA and evaluates on MC-TACO) and worsening in others (e.g., an EM score on the test set of 53.0% with the model that trains with MC-TACO and TimeML and CosmosQA and evaluates on MC-TACO), depending on the dataset used. Task affinity is important for MTL, and performance may deteriorate if unrelated tasks are learned at the same time. In addition, we found it surprising that all multi-task settings lead to improved accuracy on MATRES. MATRES is a task that treats verbs in sentences as events and predicts their order. However, there are many temporal expressions other than verbs in natural language sentences (e.g., *before*, *after*, *when*, *first*, etc.), and in order to predict the order of events, not only verbs but also various parts of speech and other factors such as duration might be effective. We hypothesize this is why MTL improves the accuracy on MATRES. We think it is necessary to further analyze why these results are obtained in cases where accuracy improves and in cases where it does not.

6 Conclusion

In this paper, we focused on the development of a language model for temporal commonsense reasoning, and tried to develop a language model for understanding temporal commonsense. We conducted multi-step fine-tuning, continual pre-training, and multi-task learning on BERT, RoBERTa, and ALBERT, using several datasets. We confirmed that the multi-step fine-tuning model that uses the general commonsense knowledge task as an auxiliary task was often better than that obtained by ordinary fine-tuning and we were able to construct a language model that understands temporal commonsense. Comparing BERT, RoBERTa, and ALBERT, ALBERT produced the best results overall.

For future work, we plan to further investigate multi-task learning. In multi-task learning, we would like to visualize attention scores, for example, and pursue what setting can improve generalization performance. Also, we plan to construct a new general-purpose language model that performs well in a variety of time-related tasks.

References

Oana-Maria Camburu, Vid Kocijan, Thomas Lukasiewicz, and Yordan Yordanov. 2019. A surprisingly robust trick for the winograd schema challenge.

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

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. 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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Baltimore, Maryland. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246, Brussels, Belgium. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 841–851, New Orleans, Louisiana. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018c. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. 2006. Extending timeml with typical durations of events. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 38–45.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog.

In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Alakananda Vempala, Eduardo Blanco, and Alexis Palmer. 2018. Determining event durations: Models and error analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 164–168, New Orleans, Louisiana. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 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.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 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.