TA-MAMC at SemEval-2021 Task 4: Task-adaptive Pretraining and Multi-head Attention for Abstract Meaning Reading Comprehension

Jing Zhang, Yimeng Zhuang, Yinpei Su*
Samsung Research China-Beijing (SRC-B)
{jing97.zhang, ym.zhuang, yinpei.su}@samsung.com

Abstract

This paper describes our system used in the SemEval-2021 Task 4 Reading Comprehension of Abstract Meaning, achieving 1st for subtask 1 and 2nd for subtask 2 on the leaderboard. We propose an ensemble of ELECTRA-based models with task-adaptive pretraining and a multi-head attention multiple-choice classifier on top of the pre-trained model. The main contributions of our system are 1) revealing the performance discrepancy of different transformer-based pretraining models on the downstream task, 2) presentation of an efficient method to generate large task-adaptive corpora for pretraining. We also investigated several pretraining strategies and contrastive learning objectives. Our system achieves a test accuracy of 95.11 and 94.89 on subtask 1 and subtask 2 respectively.

1 Introduction

Machine reading comprehension (MRC) is one of the key tasks for measuring machines' ability of understanding human languages and reasoning, it can be used broadly in real world applications such as Q&A systems and dialogue systems. MRC often comes in a triplet style $\{passage, question, answer\}$, given a context passage, questions related with this passage is asked, and the machine is expected to give the answers. The question-answer form can be question-answer pair, where the answer text is to be provided by machines, or statement form where the answer is to be filled in as cloze or multiple choices selection. By the type of answer formation, MRC can be divided into extractive and generative MRC, the former takes segments from the passage as the answer and the latter requires answer text generation based on the understanding of the passage. Generative MRC is harder than extractive MRC, since it requires more on information integration and reasoning besides focusing on relevant information.

One of the classic MRC approach focuses on matching networks, various network structures have been proposed to capture the semantic interaction within passages/questions/answers. Recent years, pre-trained language models (LMs) have brought non-trivial progress to the performance on MRC, and there's a decline of complex matching networks (Zhang et al., 2020). Plugging matching networks on top of pre-trained LMs can see either improvements or degradation in performance (Zhang et al., 2020; Zhu et al., 2020). Multiple-choice MRC (MMRC) often lacks abundant training data for deep neural networks (this might be caused by the expensive human labelling cost) and it results in a limitation to take full advantage of the pre-trained LMs.

The SemEval-2021 task 4 Reading Comprehension of Abstract Meaning (Zheng et al., 2021), is a multiple-choice English MRC task, aiming at investigating the machine's ability to understand abstract concepts in two aspects: subtask 1, non-concrete concepts, e.g. service/economy compared with trees/red; subtask 2, generalized/summarized concepts, like vertebrate compared with monkey.

We propose an approach based on the pre-trained LM ELECTRA (Clark et al., 2020), with an ensemble of multi-head attention (Vaswani et al., 2017) multiple-choice classifier, and WAE (Kim and Fung, 2020) to get the final prediction. **First**, we conduct task-adaptive pretraining, which is transfer learning using in-domain data on the ELECTRA model. **Then** we fine-tune the ReCAM task using a multi-head attention multiple choice classifier (MAMC) on top of the ELECTRA model. **Finally** we enhance the system with WAE and ensemble them all to get the best generalization capability.

*Contribution during Intership in Samsung Research China-Beijing.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 51–58
Bangkok, Thailand (online), August 5–6, 2021.

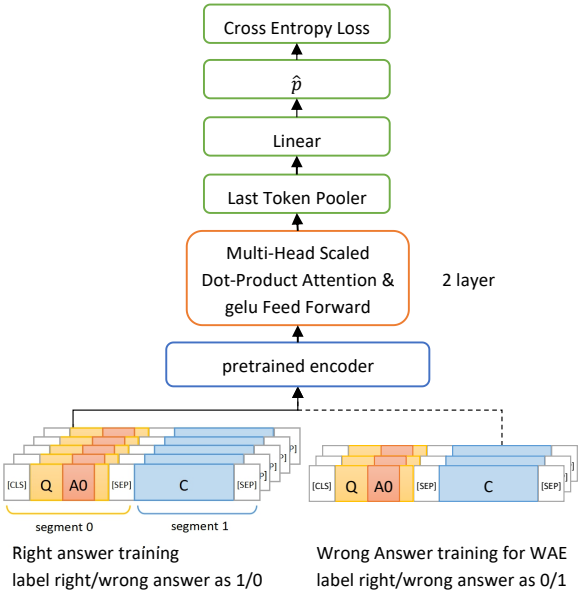

Figure 1: The overall architecture of our proposed system

In addition, we also investigated into transfer learning with natural language inference (NLI) tasks and contrastive learning objectives.

2 System Overview

Figure 1 illustrates the overall architecture of our system. The options are substituted into the query to form a complete context, rather than separate query/option segments, in order to get a less semantically ambiguous representation of the query and option. The option-filled query and context tokens are concatenated as in Figure 1, wrapped by [CLS] token and [SEP] tokens. Token embeddings are added up with segment embeddings and positional encodings to form the input for the pre-trained encoder. Then the representations from the encoder are put through a multi-head attention multiple choice classifier, which consists of 1) a 2 layer multi-head attention feed forward network to further capture the task specific query-context interactions, 2) a pooler and a linear transformation to get the final cross entropy loss. We first conduct task-adaptive pretraining on the system, and then fine-tune on the ReCAM dataset, the final model is an ensemble model by several generalization techniques including wrong answer ensemble.

2.1 Task-adaptive Pretraining

Pre-trained LMs and their downstream applications have definitely proved the power of transfer learning. The precondition of transfer learning is that the pretraining tasks have shared underlying statistical features with downstream tasks. Usually in-domain data brings more improvement on downstream tasks than out-of-domain data (Sun et al., 2019; Gururangan et al., 2020).

The genre of the ReCAM task dataset is news (confirmed by manual random checking), we argue that the task of news abstractive summarization provides high quality further pretraining dataset for ReCAM. The dataset comes in $\{article, summary\}$ pairs, the articles are crawled from formal online news publishers and the summaries are generated by humans and contain abstractive key information of the articles. News abstractive summarization aims at teaching machines to grasp the key information of the whole context by letting machines to generate the summary text.

We regenerate the ReCAM style multiple-choice dataset from the original news abstractive summarization dataset. Letting the article/summary be the passage/question, the regeneration strategy mainly includes 2 steps: 1) identify the abstract concepts in the news dataset, 2) generate gold and pseudo options. In step 1, we count the part-of-speech (POS) tags of all gold labels on the ReCAM training data as shown in Figure 2 (nouns, adjectives and adverbs are the most frequent option tags), and use a similar POS tag distribution to randomly sample word in the summary text that does not appear in the corresponding news article as gold option. In step 2, the gold option in the summary is replaced by the mask token and fed into the pre-trained LM. The LM predicts the mask token and we select some of the top ranking ones as pseudo options. Specifically, setting a high ranking threshold (e.g. top 5) would get words too similar with the gold option, which would bring extra ambiguity to the model, some relaxation on the ranking threshold would ease the problem. This method is automatic, cheap to apply on large dataset, while the abstract concept approximation in step 1 would bring some noise, such as person's names and geolocations are sometimes selected, but by our experiment result the overall pretraining performance is not hurt, the noisy samples should account for a small fraction.

In addition, it is reported that NLI task transfer

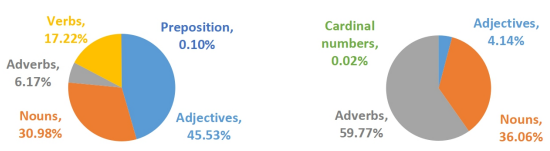

Figure 2: Subtask 1 (left) and subtask 2 (right) gold options POS Tag distribution

Dataset	# Passages	avg. doc len		avg. qry/smry len	
	training/dev/test	# words	# sent.	# words	# sent.
ReCAM subtask 1	3227/837/-	302.15	13.1	24.69	1
ReCAM subtask 2	3318/851/-	481.51	21.08	26.9	1
XSUM	20.3k/11.3k/1.1k	431.07	19.77	23.26	1
NEWSROOM	99.5k/-/-	658.6	-	26.7	-

Table 1: ReCAM/XSUM/NEWSROOM datasets statistics

learning performs well in several MMRC tasks (Jin et al., 2020). Therefore we also explored the MNLI (Williams et al., 2017) and RTE (Wang et al., 2018) tasks transfer learning for the ReCAM task, but it results in degradation. This indicates that NLI tasks are not generally fit for further pre-training in MMRC on pre-trained LMs.

2.2 Multi-head Attention Multiple Choice Classifier

The classifier takes the last layer hidden representations from the pre-trained encoder, applies the multi-head attention and feed forward non-linearity, each with a layer normalization (Vaswani et al., 2017). After that the last token is pooled, which is selecting the hidden vector from the hidden embeddings by the index of the last [SEP] token in the input, and then linearly transformed to get the probability of each $\{query_{option\_filled}, context\}$ candidate pair.

In addition, we also explored the contrastive learning objective. When humans do MMRC, they usually compare the options according to the passage, exclude the wrong ones and then analyze further on the indeterminate ones. Inspired by this, we experimented with triplet loss (Weinberger et al., 2006) (among $\{input_{nonfilled}, input_{gold}, input_{pseudo}\}$) and n-tuplet loss (Sohn, 2016) on all option-filled query and context within one sample. However the contrastive learning objective degrades the performance, suggesting these learning objectives are not as suitable for the ReCAM task as the MLE loss.

2.3 Wrong Answer Ensemble

Wrong Answer Ensemble (Kim and Fung, 2020) is a relatively simple yet effective method (Zhu et al., 2020). Kim proposed to train the model to learn the correct and the wrong answers separately and ensemble them to get the final prediction. In 2.2, the correct answer is labelled as 1 and wrong as 0 for correct answer training. Wrong answer training does the opposite labelling (correct/wrong answers as 0/1) and fine tune the model with binary cross entropy loss as below:

$$loss_w = -\sum ylog\hat{y} + (1-y)log(1-\hat{y}) \quad (1)$$

The two models's output, p_c and p_w are linearly combined to give the final prediction. A simple linear regression is leveraged to find the best value of weight w.

$$\hat{p} = p_c - w \cdot p_w \quad (2)$$

3 Experimental Setup

3.1 Dataset

We leverage external news abstractive summarization datasets for transfer learning, and then fine tune our model on the ReCAM dataset.

ReCAM. Dataset for the SemEval-2021 Task 4, consisting of news articles (verified by manually random checking) and multiple-choice questions.

XSUM. XSUM (Narayan et al., 2018) consists of 227k BBC articles from 2010 to 2017 covering a wide variety of subjects along with professionally written single-sentence summaries.

NEWSROOM. NEWSROOM (Grusky et al., 2018) is a dataset of 1.3 million news articles and summaries written by authors and editors in newsrooms of 38 major news publications between 1998 and 2017. After a coarse selection (filtering out lengthy articles/summaries, summaries duplicate with news articles, articles with unqualified pseudo options), about 229k article/summary pairs are used.

The data statistics are listed in Table 1, the 3 news datasets share similar article and query lengths.

3.2 Training Details

We compare the baseline performance of 3 kinds of Transformer-based models, BERT/ALBERT/ELECTRA, and select ELECTRA as our encoder. We adopt most hyper parameter settings from the ELECTRA large model, specifically our learning rate is 1e-5, batch size is 32 and gradient clip norm

Pre-trained model	subtask 1 dev acc.	subtask 2 dev acc.
BERT base	61.25	58.28
BERT large	66.31	67.33
ALBERT base	50.78	50.29
ALBERT large	80.88	79.08
ELECTRA base	76.82	76.97
ELECTRA large	90.20	90.13

Table 2: Baseline performance of different pre-trained Models

threshold is set to 1. In the task-adaptive data generation process, We set the threshold as top 10 for pseudo options selection, filtering out the word piece predictions(word pieces all start with a "#" in the vocabulary) and randomly select 4 words as pseudo options. See the appendix for hyperparameter details. Training was done on NVidia V100 GPUs. All the performance data is on the dev set.

4 Results

4.1 Pre-trained LM Selection and Task-adaptive Pretraining

The baseline performance of BERT, ALBERT and ELECTRA is tested by directly fine-tuning the ReCAM data on the pre-trained LMs. The results are shown in Table 2. ELECTRA outperforms the other two models with large margins. This may be caused by the learning objective difference among the models. The BERT/ALBERT models learn to predict the masked word from the vocabulary, while the ELECTRA model learns to predict whether each of the token in the input is replaced or not, which learns more about unreasonable co-occurrence knowledge besides reasonable co-occurrences and may help in digging deeper implicit semantic relations for ReCAM. Therefore the ELECTRA large model is selected as the encoder for further experiments.

The XSUM/NEWSROOM regenerated data (denoted as XN) is used for in-domain pretraining on the encoder, and the subtask 1 is fine tuned after pretraining. The prediction accuracy grows with more data fed, as shown in Figure 3. In the end of the task-adaptive pretraining, subtask 1 achieves dev accuracy 92.73, 2.80% higher than directly fine-tuning on the encoder, subtask 2 gets 92.95, increased by 3.13%.

Besides the task-adaptive pretraining and fine-tuning, we also tried multitask learning with

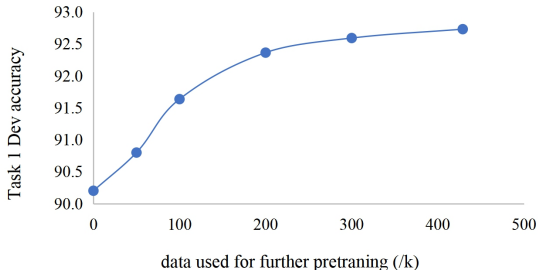

Figure 3: Subtask 1 fine-tuning performance increases with more data for further pretraining

Transfer learning setting	subtask 1	subtask 2
XN	92.73	92.95
ReCAM/$XN_{multitask}$	92.35	92.36
MNLI	78.14	81.67
RTE	88.53	89.36

Table 3: Dev accuracy for different transfer learning settings

XSUM/NEWSROOM and the ReCAM data together (up sampling the ReCAM data as 3:7 with the news dataset). The results in Table 3 shows that this approach outperforms the encoder baseline, while slightly worse than the full news data pre-trained model, this model is used for ensemble. Using MNLI/RTE for further pretraining hurt the ReCAM fine-tuning performance, especially MNLI pretraining brings about 10% accuracy decease than the baseline.

4.2 On-top Classifier and WAE

Adding MAMC on the top of the encoder helps increase accuracy on the ReCAM subtask 1 and subtask 2, the results are shown in Table 4. Further we applied the WAE to squeeze marginal increases on prediction accuracy. While option contrastive learning (OCL) does not bring performance improvement, worse than directly fine-tuning the encoder with multiple choice classifier.

Settings	subtask 1	subtask 2
Baseline	90.20	90.13
transfer learning	92.73	92.95
+ MAMC	93.64	93.79
+ WAE	93.94	94.07
OCL (triplet loss)	86.38	-
OCL (n-tuple loss)	85.32	-

Table 4: Dev accuracy on different transfer learning settings

Generalization Procedures	subtask1	subtask2
data repar. (3 sets)	93.72	93.65
	94.01	94.48
	93.82	94.36
task data aug.	93.29	93.36

Table 5: Dev accuracy of subtask 1/2 over generalization procedures.

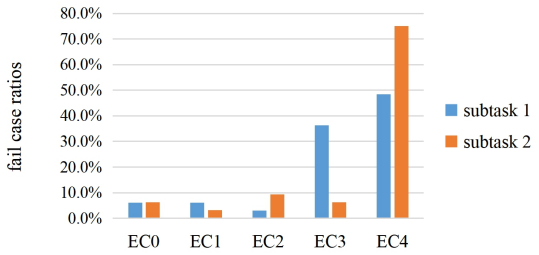

Figure 4: Subtask 1/2 fail case distribution

4.3 Improving Generalization

We mainly applied 3 procedures below for better generalization, and the ensemble of all the models have achieved test accuracy 95.11 on subtask 1 and 94.89 on subtask 2 on the ReCAM leaderboard.

1) Data repartitioning (mix the train/dev sets, and randomly split into new train/dev sets by 8:2 or 9:1) aims to smooth the distribution difference among different train/dev data partition. As is shown in the Table 5, the accuracy of different sets differs, with some higher than then original partition.

2) Augmenting the task data itself for fine-tuning, to mask different word than the original gold option (if there exists) using the method in 2.1. The accuracy remains almost the same after adding the task augmented data. This suggests that our automatic augmentation method makes lower quality samples than the labelling data, while not too noisy that it can contribute to the robustness of the model.

3) We also did Stochastic Weight Averaging (Izmailov et al., 2018) across multiple checkpoints in the same run to get better generalization (SWA dose not improve dev error but test error, so it's not listed in Table 5).

4.4 Fail Cases Analysis

We manually checked and categorized the fail cases on subtask 1 and subtask 2 into 5 classes (given roughly 850 dev cases, the total fail cases is around 50 for both subtask 1 and subtask 2). The detailed examples for each class can be found in the appendix.

- EC0, easy case. In these cases, the answer can be inferred from the query/context, while the model fails to give the correct prediction
- EC1, complicated coreference. Such cases has complicated coreference relations, though the answer can be inferred, the coreferences hinder the model from understanding correctly
- EC2, complex reasoning. In these cases, either the information related with the answer is sparse in the query/context, or the facets related with the answer is separated with intense unrelated noisy information
- EC3, external knowledge dependency. Only with the external knowledge can one give a correct answer
- EC4, ambiguity in sample cases. This category includes cases for which we think humans are not able to select the correct answer. Either the information is not enough to make a decision or there are more than one reasonable answers.

Figure 4 shows the ratios of each fail case class, the EC4 is the major class, 48.5% for subtask 1 and 75.0% for sutask 2. The following is EC3, 36.4% for subtask 1 and 6.3% for subtask 2. EC0 and EC1 are minor classes among all. With the system backbone being pre-trained LM with a matching network, it's not a surprise to see EC1 and EC3 failures, while the few EC0 and EC2 failures shows that our system learns well to capture abstract concepts within the query/article pair.

5 Conclusion

Our system takes the large pre-trained LM ELECTRA, and enhance it with in-domain transfer learning and a multi-head multiple-choice classifier on top. We compared the benchmark performance of different pre-trained LMs (BERT, ALBERT and ELECTRA) on the SemEval-2021 task 4, the result shows that different pretraining objective/dataset can lead to different inclination of model knowledge and large performance discrepancy on the downstream task. Task-adaptive pretraining has contributed the main improvement, and multi-head multiple-choice classifier and WAE bring marginal improvement. We also investigated into option contrastive learning and multitask learning, the degradation of performance suggests that triplet and n-tuplet contrastive loss is not suitable for this task and NLI is not generally beneficial for MMRC tasks.

References

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. 2020. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8010–8017.

Hyeondey Kim and Pascale Fung. 2020. Learning to classify the wrong answers for multiple choice question answering (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13843–13844.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! *Topic-aware Convolutional Neural Networks for Extreme Summarization. In*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Zhuosheng Zhang, Hai Zhao, and Rui Wang. 2020. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*.

Boyuan Zheng, Xiaoyu Yang, Yuping Ruan, Quan Liu, Zhen-Hua Ling, Si Wei, and Xiaodan Zhu. 2021. SemEval-2021 task 4: Reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Duma: Reading comprehension with transposition thinking.

Appendix

A Examples for each error case category

EC0	**easy case**
question	Two men have been arrested on suspicion of murdering a man who died after being pulled out of a fish **@placeholder** .
passage	The dead man went into cardiac arrest after rescuers responding to reports of a drowning found him in the water off St Michaels Road, Stoke-on-Trent. . .
options	0. term 1. boat 2. shop 3. pool 4. life
EC1	**complicated coreference**
question	Ceredigion council has failed to co-operate with an investigation into the **@placeholder** of a Llandysul residential home , a union has claimed
passage	Unison said the council had failed to provide answers for social care expert Tony Garthwaite, heading the investigation, and that he was not able to complete his report. Awel Deg care home was shut in February 2014. . . Awel Deg was closed following the suspension of 11 members. . . would re-open as a dementia home in spring 2015
options	0. creation 1. collapse 2. closure 3. safety 4. fate
EC2	**complex reasoning**
question	Six British teams **@placeholder** the draw for the Champions League group stage , which takes place on Thursday at 17:00 BST in Monaco .
passage	Premier League champions Chelsea, runners-up Tottenham and third-placed Manchester City are all in the draw. They will be joined by Europa League winners Manchester United, as well as Liverpool and Scottish champions Celtic who both came through qualifying. The group stages of the competition begin on 12-13 September. The last time six British teams qualified for the group stages was in 2007-08, when English sides Manchester United, Chelsea, Liverpool and Arsenal were joined by Scottish clubs Celtic and Rangers. The final saw Sir Alex Ferguson's United defeat Avram Grant's Chelsea on penalties. Scroll to the bottom to see the full list of teams and the pots they are in. . . Match day four: 31 October-1 November Match day five: 21-22 November Match day six: 5-6 December
options	0. announced 1. dominate 2. started 3. await 4. remains
EC3	**external knowledge dependency**
question	The M4 has been closed westbound near Newport after an overhead **@placeholder** became loose in high winds .
passage	The carriageway was shut from junction 24 Coldra to 28 at Tredegar Park. Officials said it led to very slow traffic as motorists were forced to come off the motorway on Friday night. A diversion using the A48 through Newport was put in place and the fire service tweeted that the M4 would stay closed until further notice while emergency repairs were carried out. Check if this is affecting your journey
options	0. wire 1. vehicle 2. link 3. valve 4. sign
EC4	**sample cases' ambiguity**
question	A book about Adolf Hitler by a University of Aberdeen historian is to be turned into a **@placeholder** television series.
passage	Prof Thomas Weber's book Hitler's First War, which was released in 2010, claimed his image as a brave soldier was a myth. The producers of the Oscar-nominated film Downfall - also about the Nazi leader - will make the show after a French TV network purchased the series. The show will be called Hitler. Production of the 10-hour series begins next year. . .
options	0. major 1. thrilling 2. special 3. planned 4. forthcoming

Table 6: Examples from each fail case category. Options in green denotes gold answers, red denotes our system predictions. Passages are truncated to reserve the most relevant parts to the questions

B Hyperparameter settings

Hyperparameter	Value
learning rate	1e-5
learning rate decay	linear
warmup fraction	0.1
Adam ϵ	1e-6
Adam $beta_1$	0.9
Adam $beta_2$	0.999
gradient clip norm	1.0
Weight Decay	0.01
Dropout	0.1
Batch Size	32
Train Epochs	10 for task-adaptive pretraining, 5 for fine-tuning

Table 7: System Hyperparameter settings