Noobs at Semeval-2021 Task 4: Masked Language Modeling for abstract answer prediction

Shikhar Shukla , Sarthak
Samsung Research Institute
Bangalore-560037, India
shikhar.00778@gmail.com,sarthak.j2709@gmail.com

Karm Veer Arya
ABV-Indian Institute of Information Technology & Management
Gwalior-474015, India
kvarya@iiitm.ac.in

Abstract

This paper presents the system developed by our team for Semeval 2021 Task 4: Reading Comprehension of Abstract Meaning. The aim of the task was to benchmark the NLP techniques in understanding the abstract concepts present in a passage, and then predict the missing word in a human written summary of the passage. We trained a Roberta-Large model trained with a masked language modeling objective. In cases where this model failed to predict one of the available options, another Roberta-Large model trained as a binary classifier was used to predict correct and incorrect options. We used passage summary generated by Pegasus model and question as inputs. Our best solution was an ensemble of these 2 systems. We achieved an accuracy of 86.22% on subtask 1 and 87.10% on subtask 2.

1 Introduction

There has been a lot of research in evaluating the performance of machine learning models to identify concrete concepts present in text and answer questions based on it (Hermann et al., 2015a). The organizers of ReCAM task at Semeval 2021 have provided a dataset to benchmark the models' performance on understanding the abstract concepts in the text in English language. The models are required to predict the missing words in a human written summary of the passage. This can help assess if the models can accurately capture the important concepts and meaning in the text.

The paper is organized as follows: In Section 2, we give a background on the problem and method that has been used, in Section 3, we present the proposed system architecture, in Section 4, we present the hyperparameters analysis done on the system, in Section 5, we present the results of the proposed architecture along with other approaches taken, and in Section 6, we conclude the paper.

2 Background

2.1 Dataset

The organizers provide two different datasets (Zheng et al., 2021) for two subtaks exploring two different definitions of abstractness (Spreen and Schulz, 1966; Changizi, 2008), *imperceptibility* and *nonspecificity*. Anything which can't be perceived is described as an *Imperceptible* concept (Example: culture,economy etc.) (Spreen and Schulz, 1966; Coltheart, 1981; Turney et al., 2011). *Nonspecificity*, as decribed by (Changizi, 2008), rather than looking at concrete things, focuses on generalizing the text (Example: hypernyms of words; vertebrate for whale). Subtask 3 explores the relationship between the two definitions.

2.2 Masked Language Modelling

Masked Language Models have played an important role in BERT(Devlin et al., 2019) and subsequent transformer models' success on different datasets. Masked Language Modelling is based on *Cloze* task (Taylor, 1953), which is described as filling the blanks in sentence using the surrounding context. Consider a Sequence **S** containing **n** tokens $(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, ..., \mathbf{w}_n)$. In Masked Language Models, a token $\mathbf{w}_t$ is replaced with a special token `[MASK]` and all the other tokens $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{t-1}, \mathbf{w}_{t+1}, ..., \mathbf{w}_n)$ are used to predict this token.

In our model, we use Roberta-Large's (Liu et al., 2019) Language Model(LMs), RobertaForMaskedLM, which randomly replaces 15% of tokens in a sequence and then tries to predict the masked word. Masked LMs perform better than left-to-right, right-to-left LMs, and concatenation of both. In a multi-layer Bidirectional LM, each word can indirectly see itself (from 2nd layer onwards) after the first layer, making the process redundant.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 805–809
Bangkok, Thailand (online), August 5–6, 2021.

Recently, Pegasus model, pretrained on C4(Raffel et al., 2020) and HugeNews dataset, (Zhang et al., 2020) has shown state-of-the-art results for abstractive summarization on many summarization tasks (Narayan et al., 2018; Hermann et al., 2015b; Koupaee and Wang, 2018). It has a transformer-based encoder-decoder architecture but has been trained with a novel self-supervised objective. It masks entire sentences in a corpus which are then generated later as one sequence capturing the abstractive summary.

3 System Overview

The key components in our proposed system are abstractive summarization of context using Pegasus, Roberta for Masked Language Modelling and Roberta for sequence classification. Algorithm 1 presents our system's algorithm for predicting the option using the given context and masked question.

Algorithm 1 MLM + Sequence Classification

```
function PREDICTOPTION(context, question, options)
    summary_model ← Pegasus-XSum
    MaskedLM ← RobertaForMaskedLM('Large')
    classifier ← RobertaForSequenceClassification('Large')
    question ← question.replace('@placeholder','[MASK]')
    context_summary ← summary_model(context)
    mlm_input ← concat(context_summary , question)
    top5_predictions ← MaskedLM(mlm_input)
    for i in top5_predictions do
        for j in options do
            if i = j then return i
            end if
        end for
    end for
    max_softmax ← 0
    answer ← 0
    for i in options do
        question ← question.replace('[MASK]' , i)
        input ← concat(context_summary , question)
        softmax_score ← classifier(input)
        if softmax_score[0] < softmax_score[1] then
            if softmax_score[1] > max_softmax then
                max_softmax ← softmax_score[1]
                answer ← i
            end if
        end if
    end for
    return answer
end function
```

Abstractive Summarization We use pretrained *Pegasus-XSum* (Zhang et al., 2020; Narayan et al., 2018) to capture the abstractive summary from context which relates closely to the task at hand. We also experimented by finetuning it for the given task, by giving the corpus as input and passing the text, obtained from question after replacing *@placeholder* with the correct option, as output. We further experimented with extracting three line summaries by splitting context into three parts and extracting summary for each.

RobertaForMaskedLM The task at hand can be converted into predicting the masked token by replacing the *@placeholder* in question with the `[MASK]` token. We use RobertaForMaskedLM to predict the masked token. To help the model get the context in question, we prepend the summary from Pegasus-XSum to the masked question text. To finetune the model, we train it by passing concatenation of summary and masked question as input and providing concatenation of summary and filling the question with correct blank as output. Once the model is trained, we use Huggingface's (Wolf et al., 2020) pipeline to predict the top 5 words for filling the masked token. Out of five predictions, word which has the highest probability of filling the blank and which is present in the given options is selected as the answer.

Roberta For Sequence Classification In some of the cases, we observed that the predicted words weren't present in the options. For handling such cases, we use Roberta for Sequence Classification. We convert the task into a binary classification task by filling the masked question with correct option and treating it as one class and by filling the mask with wrong options as another class. This leads to a class-imbalanced dataset (1:4 ratio). To handle this, we randomly selected equal number of wrong-option class. On the test set, option which achieves the highest softmax score out of all the given options, is selected as the answer.

4 Experimental Setup

All the experiments were conducted in a Google Colab system with 12GB RAM and T4-Nvidia GPU. We experimented with different settings in three components in our system: summary extraction, Masked Language Model (MLM) and handling cases when word predicted by MLM is not present in the options.

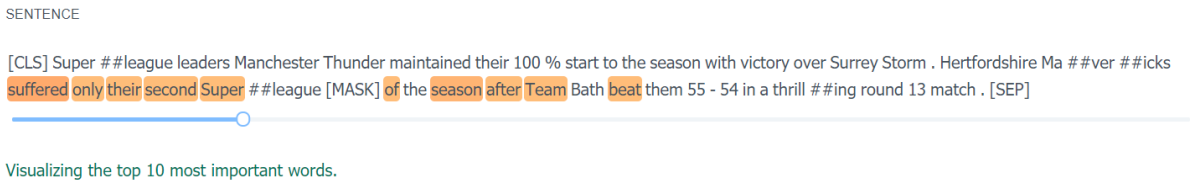

Figure 1: Salience map

Summary Extraction To identify the role of summary in the task, we experimented with three settings: no summary for prediction, extracting a one line summary from Pegasus-XSum and extracting a three line summary from Pegasus-XSum. To get a three line summary, we broke the context into three equal parts and fed into the summary model. Extracting one-line summary performed best for both the subtasks.

Finetuning RobertaForMaskedLM For Masked Language Model (MLM) to work, its very important for the model to identify and understand the context. Using the entire context meant that for some cases, the input size would exceed the limit (512 tokens) for model. So, we experimented with two settings, one line summary prepended to question, and using context when number of tokens wouldn't exceed the limit and one line summary when it did. For an untrained MLM, the latter setting performed better. Also, by training the MLM, a further improvement of 4-5% was observed. The MLM is trained with a batch size of 2 using the Adam (Kingma and Ba, 2017) optimizer with a learning rate of 1e-5.

Handling missing cases for Masked Language Modelling In few of the cases it was observed that the top 5 predictions made by MLM were not found in options. For such cases, we experimented with two settings: to find a pair of prediction and option which has the highest cosine similarity using Spacy embeddings (Honnibal et al., 2020), or to predict options using Sequence Classification (correct option as class 0, incorrect options as class 1). Input for Sequence classification was concatentating one line summary with question in which *@placeholder* is replaced with the options. If the option used to replace *@placeholder* is correct, output class is 0 else 1. Using Sequence Classification worked better than cosine similarity. Also, Roberta-Large outperformed other transformer models for sequence classification on this dataset. The sequence classification model is trained with a batch size of 8, sequence length of 128 and using the Adam (Kingma and Ba, 2017) optimizer with a learning rate of 1e-5.

5 Results

We have mentioned the accuracy of all the systems developed by us for subtasks 1 and 2 in Table 1 and 2. We have also plotted a salience map in Fig. 1 (with AllenNLP(Gardner et al., 2017) demo tool) to visualize the importance of each token in the prediction of the masked word. The example input from training set is: "Superleague leaders Manchester Thunder maintained their 100% start to the season with victory over Surrey Storm. Hertfordshire Mavericks [MASK] only their second Superleague loss of the season after Team Bath beat them 55 - 54 in a thrilling round 13 match." The top 5 predictions by the transformer model for the masked token are: "suffered", "recorded", "experienced", "sustained", "had". The correct option is "suffered". This demonstrates the effectiveness of using summary as context and formulating the problem as masked language modeling task.

5.1 Error Analysis

Across the 2 subtasks, masked language model made predictions which were present as part of options for 87% of the data. We present error analysis of the MLM over here:

a. Capturing more than one meaning

Context: (truncated) ... said he was too young to swim and should have still been in his mother's care ..... mother failed to show ... It tends to be when there's quite stormy weather the pups will get into trouble and they do get very tired, very hungry and very dehydrated and they just wouldn't survive without assistance. .....

Question: A @placeholder baby grey seal who was rescued from the rocks at Corbiere in Jersey will be flown to the UK on Friday .

System Description	Subtask 1	Subtask 2
Roberta-Large as binary classifier with summary and question as input	72.69	73.20
T5 large with passage, question, options concatenated	57.20	57.85
Untrained Roberta Large with MLM objective	72.64	73.36
Untrained Roberta Large with MLM objective and cosine similarity of word embeddings in failed cases	77.76	77.95
Trained Roberta Large with MLM objective	81.30	82.23
Trained Roberta Large with MLM objective and cosine similarity of word embeddings for failed cases	85.10	86.03
Trained Roberta Large with MLM objective and Roberta binary classifier for failed cases	**87.25**	**88.40**

Table 1: +-Accuracy of experimental setups on validation set

System Description	Trained On	Evaluated On Subtask-1	Evaluated On Subtask-2
Trained Roberta Large with MLM objective and cosine similarity of word embeddings for failed cases	Subtask-1	83.20	77.24
	Subtask-2	73.92	85.37
Trained Roberta Large with MLM objective and Roberta binary classifier for failed cases	Subtask-1	**86.22**	82.44
	Subtask-2	78.56	**87.10**

Table 2: Accuracy of experimental setups on Test set

Correct Option: lone
Predicted Options: rare, stranded,distressed,tiny

For this particular example, the number of tokens didn't exceed 512 and entire context was used. As suggested by the context, its a rare event that a pup would be found without mother, stranded since mother didn't show or couldn't find the pup and is distressed as well.

b. Failure to capture information in one line summary

Context: (truncated) ... "I am very saddened by this, but what matters most now is the well-being of our kids," he told People magazine. "I kindly ask the press to give them the space they deserve during this challenging time." Jolie, 41, filed for divorce from Pitt, 52, citing irreconcilable differences on Monday. Her lawyer, Robert Offer, said the decision had been made "for the health of the family". .....
Summary: Actor Brad Pitt has said he is "very saddened" after his wife Angelina Jolie filed for divorce.
Question: Actor Brad Pitt has said he is " very saddened " that his wife Angelina Jolie has filed for divorce and has asked for @placeholder on their children 's behalf.
Correct Option: privacy
Predicted Options: support, custody, forgiveness, protection, counseling

For this particular example, 1-line summary was used and as evident no information pertaining to children could be located in it. Our best guess for these predictions are based on the data on which model is trained and is pretty much commonsense seeing the presence of "custody" and "support".

c. More than one correct answer

Context: . A lunar eclipse is when the Moon is fully covered by the Earth's shadow. It is the second one this year. The Moon's surface showed up coppery orange or red because the light from all the Earth's sunsets and sunrises were reflected on to it during the eclipse. In this timelapse, the Moon can been seen re-appearing as the shadow moves away.
Question: A total lunar eclipse has been visible across much of the Americas and Asia , resulting in a @placeholder " Blood Moon " .
All Options: bizarre, special, dramatic, lunar, visible
Correct Option: dramatic
Predicted Options: rare, special, spectacular, unique, partial

With a context being concise and straightforward, it can also be termed as special or rare other than dramatic. As per our algorithm, our system predicted special.

6 Conclusion

We have described the systems developed by as to solve the Reading Comprehension challenge at Semeval 2021. In our best performing submission, we framed the problem as a masked language modeling task. We used the predictions from a separately trained binary classifier when the above system failed to generate words which were not part of the options. Our models were able to achieve high accuracy with a relatively simple setup. We were ranked 11th out of 23 participants in subtask 1 and 12th out of 28 participants in subtask 2. As part of future work, we aim to use information from knowledge bases such as ConceptNet. This can help extract broader concepts related to the words predicted by the masked language model.

References

Mark A. Changizi. 2008. Economically organized hierarchies in WordNet and the Oxford English Dictionary. *Cognitive Systems Research*, 9(3):214–228.

Max Coltheart. 1981. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015a. Teaching Machines to Read and Comprehend. *CoRR*, abs/1506.03340.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015b. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization.

Mahnaz Koupaee and William Yang Wang. 2018. WikiHow: A Large Scale Text Summarization Dataset.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Otfried Spreen and Rudolph W. Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5):459–468.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.

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)*.