Universiteit Leiden

Master Computer Science

A Multilingual Solution for Mental Health Evaluation with Full Context Components

Name:	Runda Wang
Student ID:	S2792265
Date:	10/05/2023
Specialisation:	Artificial Intelligence
1st supervisor:	Prof.dr. M.R. Spruit
2nd supervisor:	Dr. S. Verberne

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Text-based models for multi-purpose inspection and evaluation are absent in the mental health detection field. We propose a multilingual solution in the form of Multi-choice Machine Reading Comprehension (MMRC) tasks combined with modification of the network architecture, input format and training scheme for such a scenario. In our solution, we 1) introduce extra tokens for the identification of different components of an MMRC sample and reconstruct the whole sample as one input sequence, 2) insert a small transformer encoder network to fuse the information from every option and 3) transfer the modelling of relationships among all components obtained from high-resource language to low-resource languages. Through the above modifications, we input the text records, evaluation, and corresponding criteria in text form, and train the neural network to output a valid assessment. Experiment results suggest our methods can be efficient in the demand for adaptability and cross-language transfer.

Contents

1 Introduction

Mental health care has been getting increasing attention in recent years due to the prevalence of mental illness. According to the latest survey conducted by World Health Organization (WHO), 970 million people worldwide had been living with mental disorders, accounting for around 13% of the total population by 2019 [15].

Under this circumstance, the WHO report calls for wider support for mental health services, including affordable and accessible mental health care for all. However, the growth of professionals in this field relies on long-term training thus this situation can hardly be improved shortly. Therefore, the development of automatic processing mechanisms, which can assist medical professionals in diagnosing and treating, leads to a surge of machine learning application research in such field [25].

Detection of mental health conditions is a crucial premise for mental health care, which commonly relies on the manual analysis of free-text records and conversations between either patients and professionals or users and conversational agents [29]. **The popularity of smart devices and online mental health support applications contribute to the formation and easy acquisition of consultation records.** Furthermore, in the context of public mental health services, extra free-text resources including user-generated content on social media [1] are also taken into account for public and private mental health tracking.

By reviewing previous studies, we found that machine learning methods have been introduced to the detection and diagnosis phase such as intervening in suicide attempts based on notes [16], evaluating depression and post-traumatic stress disorder based on self-report symptoms [21], predicting mental conditions based on psychiatric notes [28] and diagnosing dementia based on cognitive function tests records [2].

Most of these applications, however, rely on the analysis of processed structural data instead of free-text format. Due to the difference in data processing and feature extraction methods for various intents, both models and methods need to be designed specifically. In practice, casting original text data into standardized evaluation forms (e.g. HAM-A [7], WHO-5 [27], Beck Depression Inventory, etc) for direct or further evidence-based diagnosis requires additional work for professionals, especially when they are collected from conversational agents and social media.

To alleviate this burden, we envision an automatic system that can directly process free-text records and is adaptable for different evaluations (e.g. intensity, frequency, likelihood, etc) of multiple inspections (e.g. anxiety, depression, suicidal thoughts, etc). Moreover, for different inspections and evaluations, the corresponding criteria could also vary (e.g polarity, scoring, etc) thus it is vital to enable the model to understand the measurement as well.

We propose a method for mental health detection problems in the form of MMRC (**M**ulti-choice **M**achine **R**eading **C**omprehension), as shown in Table 1 (right). By taking the free-text records

Consultation Record	**MMRC Task**
Record: Agent: How are you today? User: Not great... Agent: Looks like you're having a rough day. Tell me more about how you feel? User: I went out and felt very uncomfor--table with so many people around. Can't help feeling being gazed at. Agent: As always, let's start by turning our attention to our breath. ... Agent: It happens to all of us. Here are some techniques that can help.	**Passage:** Until we went to a playdate two weeks ago. Thea's mom is Serbian and crepes are apparently as common in Serbia as they are in France. We discussed the batter, the texture, the cooking process, the topping options and Dee generally brought me up to speed. Being not brave enough to just start throwing ingredients in a bowl as she did, I got a recipe of the internet for general proportions, ended up not using nearly as much water as was called for and successfully made crepes.
Inspection: Indication of User's anxious mood. (HAM-A item 1)	**Question:** Why did they discuss crepes?
Measurement of Evaluation: A.Not present. B.Mild. C.Moderate. D.Severe. E.Very severe.	**Options:** A. Because Thea's mom is Serbian. B. Because the writer got a recipe from the internet. C. Because the writer is interested in learning how to cook crepes. D. None of the above choices.

Table 1: An example of simulated consultation record (left) between a user and a conversational agent for mental health support, *Wysa*[1], combined with a HAM-A evaluation item and its corresponding measurement, and one MMRC task selected from the COSMOS-QA dataset (right).

as passages, and explicitly declaring inspection and evaluation measurement in the text as questions and options to formulate a complete sample, shown in Table 1 (left), we cast a sample of mental health evaluation task into the MMRC task format and use pre-trained language models to characterize the relationships within.

Previous methods for MMRC tasks are generally based on ptr-trained language models. Language models are for modelling the semantic rationality of text segments, where the semantic information of a text segment is contained in its words and word order in the text. In the context of natural language processing, especially in the language model based on neural networks, semantics are implied in the co-occurrence of tokens and their specific positions in a sequence.

[1]https://www.wysa.com/

Pre-trained language model creates dense representations for tokens and aggregates their information to evaluate the rationality of specific tokens appearing at specific positions in a text segment by pre-training on massive open-domain text corpora with the help of self-supervision characteristics in text data.[19, 6, 4]

We speculate that with the co-occurrences of records, inspection and evaluation measurement, pre-trained language models could adaptively identify inspections and measurements and directly model the relationships between them based on their semantics.

Additionally, considering the imbalanced development of medical resources in various countries and to respond to the call for affordable and accessible mental health care for all, we expect a multilingual-friendly model for vulnerable people who cannot use dominant languages and maximize the use of data in various languages through unified modelling.

Unfortunately, the datasets for mental health detection are mostly not publicly available. Therefore, we will use the open-source MMRC datasets with similar characteristics to our target data as an alternative and optimize the methods specifically to adapt to these characteristics.

Our contribution is two-fold: 1) We modify the modelling method of the MMRC task by introducing additional information through tag tokens and adding a fuser network to enhance the comparison between options; 2) We decouple task format from languages through the step-by-step training and enhance the performance of the target task in low-resource languages by intermediate training with samples in the identical format in high-resource language on pre-trained models.

In this thesis, we address the following research questions:

RQ-1 To what extent does the pre-trained language model establish relationships of all components based on their semantics with explicit indication of boundary and category information?

RQ-2 How much can a pre-trained multilingual language model learn relevant patterns from English datasets and transfer such knowledge to low-resource languages and domains?

This section introduces the existing circumstances of the applications of machine learning methods in the field of mental health detection, as well as our research question and corresponding solutions; In section 2, we will give a brief illustration of the related work on solving MMRC tasks and multilingual language models; In section 3, we provide a detailed explanation of the implementation of our methods; In section 4, we present our experiment data, settings and

results; In section 5, we discuss the reasons of some specific phenomena and results in the experiment; The conclusion will be shown in section 6.

2 Related Work

2.1 Multilingual Pre-trained Language Models

Transformers [30] have brought a significant impact on the Natural Language Processing (NLP) field since 2017, especially its parallel processing capabilities have allowed the surge of large-scale pre-trained language models such as BERT [6], RoBERTa [14], ALBERT [13], ELECTRA [4], GPT [19] etc. which are based on transformer architecture.

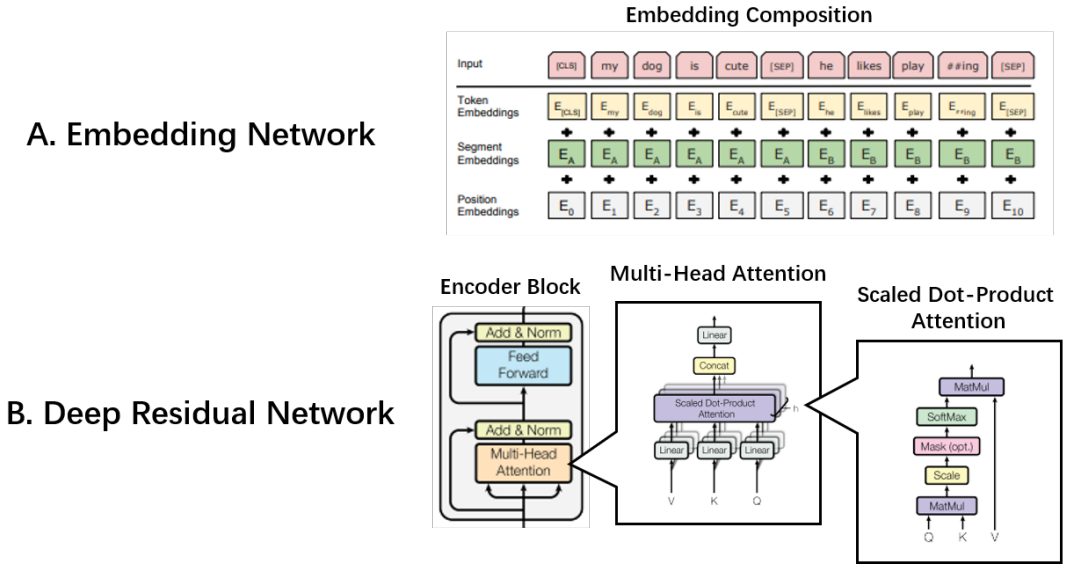

Figure 1: Universal structure of transformer auto-encoder models. (A) Embeddings are learnable and initialized in the pre-training process. (B) Each layer of the deep residual network is an encoder block which consists of a multi-head attention network and a feed-forward network. Images are collected from [6, 30]

These models mainly consist of two parts: 1) an embedding network for converting token indexes into dense vectors conveying superficial features, position information, and, according to the pre-training strategy, segment information; 2) a deep residual network for modelling contextual features by attending all tokens in a text sequence with the multi-head self-attention mechanism, as shown in Figure 1.

Among them, the multi-head self-attention endows the model with a strong ability to aggregate contextual information. The self-attention mechanism reconstructs contextual representation for them with the linear combination of all their embeddings by calculating scaled dot-product similarities between them as weights. In order to enhance the expression capability of token features, the multi-head self-attention method introduces different linear projections on top of self-attention, projecting embeddings into different subspaces before calculating similarity. Afterwards, the

embeddings would be fed into one single-layer neural network, namely a feed-forward network, for a non-linear transformation to refine their features.

Combined with residual connection and layer normalization, these structures form an encoder block. Original tokens would be converted into dense vectors by embedding network and processed by multiple stacked encoder blocks, in which they would be reconstructed with context information. As a result, the final output token embeddings could contain semantic information corresponding to the context they are located.

Although the improvement on downstream tasks made by large language models is very gratifying, the need for a massive amount of high-quality training samples obstructs the development of such models in low-resource languages. In contrast, available low-resource language corpora could hardly enable a monolingual model to achieve performance equivalent to high-resource models such as the English model [12]. Furthermore, in the fine-tuning stage, due to either the lack or the limited amount and size of MMRC datasets in low-resource languages, it would be difficult to train a monolingual language model of low-resource language for our target.

The pleasant surprise is that in recent years, many cross-lingual pre-training strategies have been proposed and validated that a language model can obtain multilingual capacity with training on a combination of the monolingual corpora, which might have orders of magnitude gap in size, from multiple different languages. These strategies require no modification of previous language models architecture and have brought many pre-trained multilingual language models to the stage such as mBERT [17] and XLMs [12, 5, 3], all trained on non-parallel data for more than 100 languages.

More importantly, when these multilingual models are fine-tuned with monolingual datasets of new tasks, they can transfer the task-relevant knowledge to other languages [17, 5]. This might indicate that the knowledge related to the task format could be independent of the languages. Therefore, training the multilingual models conforming to the target task format in the high-resource language could help the subsequent low-resource language target task.

We adopt their idea of using rich-resource language to enhance the training efficiency of low-resource language with a unified language-independent modelling method and apply multilingual pre-trained language models to extract the semantic representation of tokens from text segments in a multi-task fashion.

2.2 Multi-Choice Machine Reading Comprehension

MMRC is a subtask in Machine Reading Comprehension consisting of three main types of components: passage **p**, question **q** and answer options $\mathbf{o} = o_1, o_2, ..., o_n$. The purpose of such a task is to select the best matching answer o_i corresponding to question **q** from information provided by **p**. Normally, the questions and corresponding candidate options would vary accord-

ing to the passages, which coincides with our expectation of modelling different measurements for different evaluations and inspections.

Differing from Natural Language Inference(NLI) and other classification tasks at the sentence or document level, the illustration of the task objectives and the meaning of the corresponding class in MMRC tasks are expressed in explicit semantics by questions and options.

In 2019, Jin et al. [9] proposed a multi-stage multi-task training strategy and achieved state-of-the-art performance in many MMRC datasets with coarse-tuning on an NLI dataset and fine-tuning on supplement and target datasets in the form of multi-tasking with newly proposed multi-step attention network classifier. Later in 2020, Zhang et al. [31] and Zhu et al. [33] proposed DCMN+ and DUMA respectively, both focusing on modelling the relationship between passages and question-option pairs with extra cross-attention mechanisms. This methodology is brought to the ultimate attainment by Zhang [32] with HRCA+, which applied attention mechanism to model all 9 types of the correspondence relationship between three components.

These methods concatenate the passage $P = [p_1, ...p_{L_p}]$, question $Q = [q_1, ..., q_{L_q}]$ and one of the candidate options $O^i = [o^i_1, ..., o^i_{L_o}]$ as a sequence $S^i = concat(P, Q, O^i)$, where L_p, L_q and L_o are the length of P, Q and O^i, and $i = 1, 2, ..., N$ is the index of N candidate options, and using each sequence to retrieve features representing the rationality of itself for the downstream classifier to learn a probability distribution of $P(S^i|S^1, ..., S^N)$, which is selecting the best combination among N combinations.

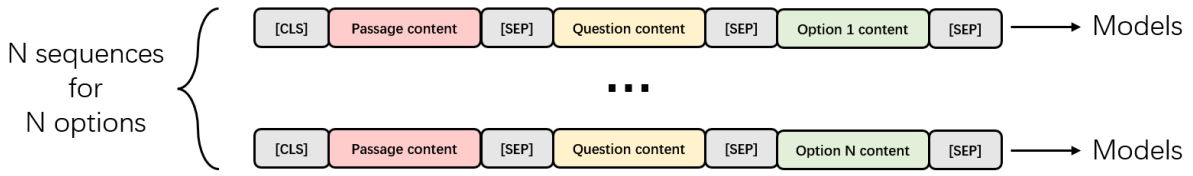

Figure 2: Input sequences for conventional modality. Each sequence consists of only one candidate option and it requires N sequences for tasks with N options.

This conventional modality, illustrated in Figure 2, which treats every permutation as a complete sequence, has achieved good results, yet has some remaining problems. An important issue of modelling every permutation respectively is the deficiency of semantically relevant information between options.

This problem could be critical when there exist references between options (e.g. 'none of the other options are correct', 'all of the options are correct', etc.), as shown in Table 1 (right). More importantly, when options are in the form of scoring, the models need the implicit semantics about intervals and gaps (e.g. In grades 1 to 5 and 1 to 10, the same grade 5 has different meanings) which could be recognized only with the co-occurrence of all options.

Although Ran et al. noticed this problem and proposed an Option Comparison Network [20], which takes the correlation between options based on token-level embeddings into account after retrieving each permutation respectively, this feature level aggregation still overlooked the potential direct semantic relation between options.

For the above issue, we believe there is a solution which is feeding the model all components of one MMRC task entry including passage, question, and all candidate options at once and using its encoding ability to directly encode the full context of such task. Meanwhile, this might require extra effort for the model to recognize different components and thus model the relationships between them.

3 Methods and Model

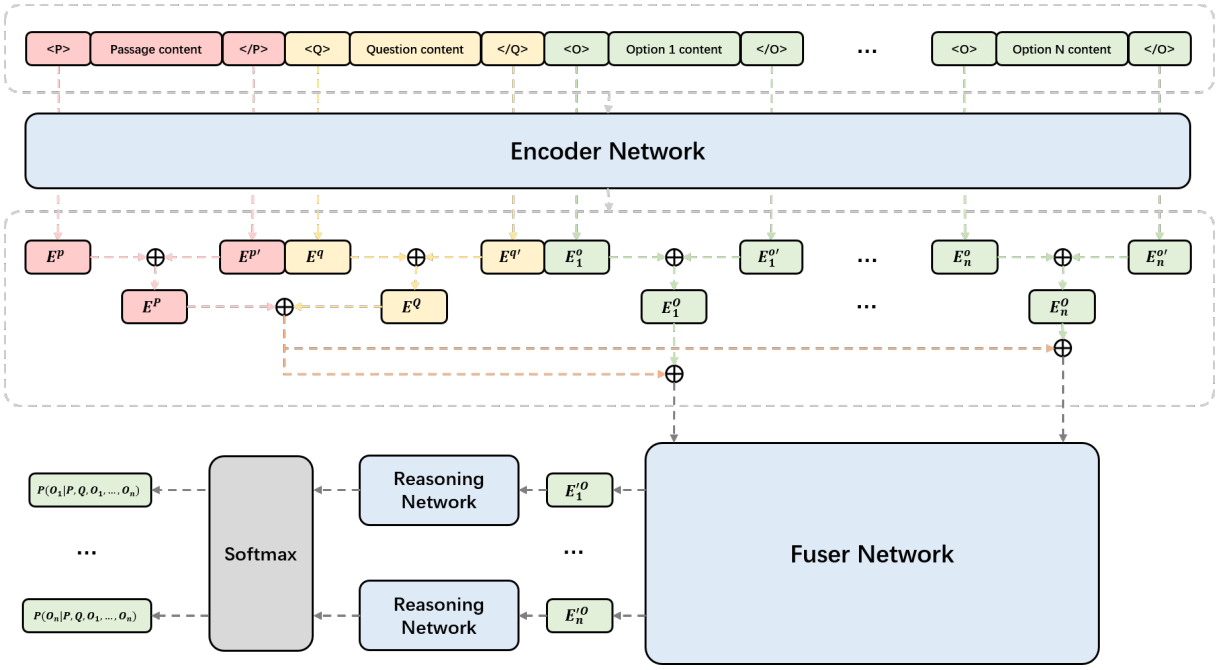

Figure 3: Model architecture. The encoder network is a pre-trained language model, the fuser network is a transformer encoder and the reasoning network is a multi-layer perceptron.

We propose a method, with extra tag tokens for introducing prior knowledge including components' boundaries and categories, to allow the models to read all components (passage, question, and all options) at once and model the semantic relevance between options while processing one sequence. The general architecture of our proposed method and model is illustrated in Figure 3.

3.1 Tags for Full Context Input

There are three categories of components in MMRC tasks: passages, questions, and options, while former methods allow only one component of each category in one sequence, as shown in Figure 2, with a fixed order. Moreover, they use $[CLS]$ token and $[SEP]$ token to mark the boundary of different components and retrieve either the embedding of $[CLS]$ token or mean pooling of the sequence as features for further processing.

Formally, the representations of these special tokens are contextualized and they jointly create an explicit distribution pattern for the input sequence. This inspires us that additional prior information could be added to the sequence by inserting particular special tokens in corresponding positions.

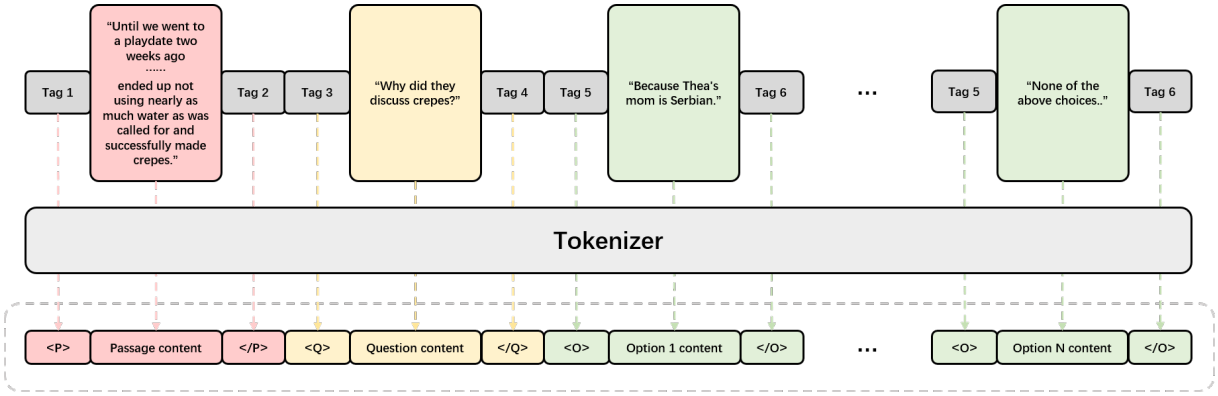

Figure 4: Tags have no actual semantics in human language and are added on both sides of a component. Correspondingly, we add an equivalent number of embeddings for tags in the embedding network of the pre-trained language model.

We expect that the model can identify different components and establish the semantic-based relationship between them according to their categories, especially when all components are fed together as one sequence. Hence we need identifiers to mark not only the boundary but the category of components in the input sequence for there exists an indefinite number of components from the options category, determined by the datasets. On account of the above-mentioned issues, we design learnable extra tag tokens for indicating the model to recognize the components, as shown in Figure 4.

We modify the modelling from $P(S^i|S^1, ..., S^N)$ to $P(O^i|P, Q, O^1, ..., O^N)$, where $i = 1, 2, ..., N$ is the index of N candidate options for corresponding passage P and question Q, by allowing the model to read all components at once to construct a full context with different start and end tags identifying every component in it. Finally, the tagged full context components S_{FCC} would be:

$$S_{FCC} = [[P_{start}]; P; [P_{end}]; [Q_{start}]; Q; [Q_{end}]; [O_{start}]; O^i; [O_{end}]; ...; [O_{start}]; O^N; [O_{end}]] \quad (1)$$

, where $[P_{start}]$,$[P_{end}]$,$[Q_{start}]$,$[Q_{end}][O_{start}]$,$[O_{end}]$ are tags for corresponding components and $[;]$ indicates row-wise concatenation. **It is worth mentioning that the tags for all options are identical to ensure the impact of tags on each option is equivalent and it does not require adjustment on the number of introduced extra tag tokens according to the number of options.**

In section 2, we have discussed how the pre-trained language models construct token embeddings that can convey contextual information. Associating with the application of special tokens, including $[CLS]$ for classification tasks, and $[MASK]$ for cloze tasks, we can infer that the intrinsic information of the token itself could be neglected by the network when it is trifling in downstream tasks while the task-relevant context feature would be reinforced. Therefore, we can consider the output embedding of each token to be a characterization of its context, rather than just its

own representation.

Similar to the usage of $[CLS]$ and $[MASK]$ tokens, we assume that the embedding of tags can represent the task-relevant information of each component in the input sequence and thus retrieve them as component features in full context for the downstream classifier. **The models would be trained in a supervised manner which makes use of the labels as supervision information.**

Our goal is to create appropriate representations for components according to the context. With extra knowledge about the boundaries and categories introduced by our tagging method, we expect the model to have different concerns about the content that appears in different components and aggregate the task-related features of them into the embeddings of tags.

3.2 Model Architecture

In addition to the above methods, we use neural networks to achieve our modelling goal. As shown in Figure 3, our model consists of three parts: encoder network, fuser network, and reasoning network.

3.2.1 Encoder Network

The Encoder network is originally a pre-trained language model, defined as f_{ENC}. It would become a tag-oriented feature extractor along with the training when we only retrieve the embeddings of tags. These embeddings carry the task-relevant features at the semantic level in the co-occurrences of full context components, which is enabled by the contextualized representation capability of the pre-trained language model.

By feeding the tagged full context sequence $S = concate(p, q, o_1, ..., o_n)$ into the encoder network, we can obtain the corresponding contextual embedding of every token. Considering the purpose of tag-oriented modelling, we retrieve only the embedding of tags as component features for further processing, defined as follows:

$$E^p, E^{p'} = f_{ENC}(t^p, t^{p'}|S), E^P = avg(E^p, E^{p'}) \tag{2}$$

$$E^1, E^{1'} = f_{ENC}(t^q, t^{q'}|S), E^Q = avg(E^q, E^{q'}) \tag{3}$$

$$E^o_i, E^{o'}_i = f_{ENC}(t^o_i, t^{o'}_i|S), E^O_i = avg(E^o_i, E^{o'}_i) \tag{4}$$

, where the $t^p, t^{p'}$ are start and end tag of passage component, $t^q, t^{q'}$ are start and end tag of question component and $t^o_i, t^{o'}_i$ are start and end tag of the i-th option component. **Based on our previous discussion, we adopt the representation of tags as features for different components and take the mean value of the tag embeddings on both sides to align the features of each dimension.**

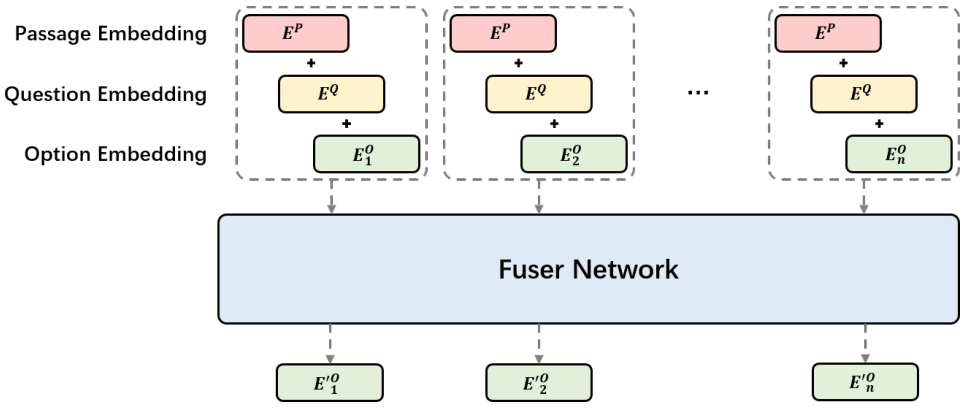

Figure 5: Illustration of embedding composition in fuser network. The network itself is a multi-layer transformer encoder.

3.2.2 Fuser Network

On top of the pre-trained language model, we insert another transformer encoder network to model the relationship between all options at the feature level, as shown in Figure 5, where we take the embeddings of E^P, E^Q and E^O_i to reconstruct the representation of every option by modifying their embedding compositions as follow:

$$E^{comp}_i = E^P + E^Q + E^O_i \tag{5}$$

E^{comp}_i is the new embedding of option i and we can have a new representation for all options $O = [E^{comp}_1, E^{comp}_2, ..., E^{comp}_n]$. These component embeddings are the result of the same transformation thus they are in the same subspace having a common basis and it is plausible to assume they are additive.

In particular, we removed the position embedding in this module since the option sets are in fact permutation-invariant and the transformer encoder is position-independent if position encoding is avoided. **We use it to strengthen the discrimination between options and directly formulate a new representation in consideration of the whole option set for each of them:**

$$E'^O_i = f_{FUSE}(E^{comp}_i | O) \tag{6}$$

Based on the structure of the transformer encoder, we can assume that this network can create representations of each option by attending to the representation of other options in the option set. Henceforth, we refer to this module as a fuser network while it fuses the information of different options.

3.2.3 Reasoning Network

Afterwards, we generate logit values for options as a confidence level with a multi-layer perceptron and apply softmax smoothing on them to obtain the probability distribution:

$$L_i^O = f_R(E_i'^O) \tag{7}$$

$$P(o_i|p, q, o_1, ..., o_n) = \frac{e^{L_i^O}}{\sum_{i=1}^{n} e^{L_i^O}} \tag{8}$$

Finally, we can regard the probability distribution of options are conditioned to all components in the corresponding MMRC task including passage, question, and all options on the semantic level, noted as $P(o_i|p, q, o_1, ..., o_n)$.

3.3 Unified Tuning

Previous research suggests further tuning pre-trained language models on similar tasks, in domains and forms, to target tasks is more conducive to their adaptation in the context of transfer learning. This method is known as intermediate task fine-tuning [18].

Due to the absence of publicly available mental health detection datasets and the anticipation of a shortage of collectable data, we propose to decouple the task format from our target and post-pre-train a language model that would be capable of solving tasks of the same form and scalable for new coming low-resource target data. For this, we construct a two-stage training which includes an intermediate tuning stage for format adaptation and a further fine-tuning stage for language and domain adaptation.

The encoder network would function as a tag-oriented feature extractor with our tagging method. In order to facilitate this transformation and adapt it to our target task, we construct our intermediate fine-tuning task with MMRC datasets which cover the different formal characteristics we emphasize for our target task in addition to ordinary MMRC tasks, including option semantic reference and dialogue format text. While there is not any single MMRC task that could contain all these characteristics, we would use a multi-task fashion training scheme in our intermediate fine-tuning phase to assemble the task-relevant knowledge.

This transformation and adaptation would be achieved via monolingual intermediate tuning on a multilingual model with high-resource language. As mentioned in section 2, the task-relevant formal characteristic learned from high-resource language datasets could be directly shared and could also be improved by further fine-tuning with target languages.

Our setting is based on the small number of our target task samples and the use of low-resource languages. In extreme cases, we expect our methods and model to have sufficient generalization ability only based on formal characteristics to adapt to zero-shot scenarios.

We assemble all datasets with different characteristics as one and construct every batch by sampling training data from it, illustrated as Figure 6, with a smoothing

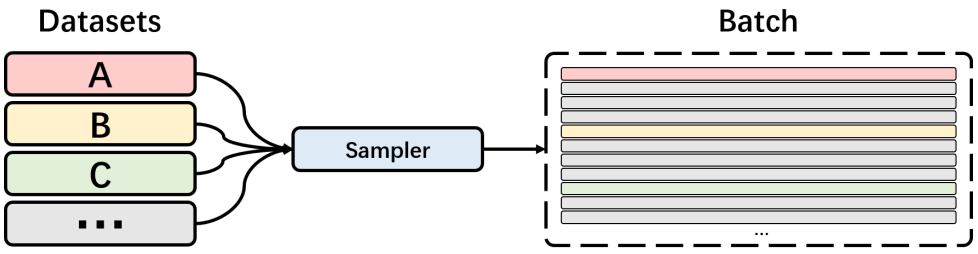

Figure 6: Batch construction for unified tuning. Each batch is constructed with samples collected from all available datasets.

method from applications in the previous study [12] of language sampling for multilingual models:

$$c_i = \frac{r_i^\alpha}{\sum_{k=1}^{N} r_k^\alpha} \text{ with } r_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \tag{9}$$

, where n_i, n_k is the number of training samples in the original corpus i and k, r_i, r_k denotes the ratio of sample amount in corpus i and k to the overall sample amount in all corpora, and c_i denotes the chance of sample in corpus i being sampled. In our experiments, we set $\alpha = 0.5$.

This sampling scheme eases the exposure bias of different formal characteristics. It significantly reduces the magnitude gap of their training samples. At the same time, its construction depends on our tagging method for full context input and can process tasks with different numbers of options without distinction.

We regard the unification of formal characteristics between intermediate tuning and the follow-up fine-tuning, and the indiscriminate sampling method for assembling these characteristics, for applying task-relevant knowledge to target tasks independent of languages as unified tuning.

4 Experiments

4.1 Datasets

Datasets	Intermediate Tasks			Target Tasks		
	RACE	COSMOS-QA	DREAM	C^3	SweQUAD-MC	MuSeRC
# of train	87,866	25,262	6,116	11,869	962	5,380
# of validation	4,887	2,985	2,040	3,816	126	993
# of test	4,934	6,963	2,041	3,892	102	-
Avg. P length	321.9	70.3	85.9	116.9	379.4	203.9
Avg. Q length	10.0	10.6	8.6	12.2	7.8	7.6
Avg. O length	5.3	8.1	5.3	5.5	4.3	5.3
Language	English	English	English	Chinese	Swedish	Russian
Characteristics	Ordinary	Reference	Dialogue	Dialogue	Ordinary	Ordinary

Table 2: Statistics of datasets involved.

As mentioned in the previous section, we construct our intermediate task with high-resource language datasets which consist of the formal characteristics we need. Firstly, considering the consultation record we will process in our target task, we need MMRC datasets with dialogues. Secondly, for adaptively recognizing the evaluation measurements, we would like to introduce datasets with semantic references between options to reinforce the modelling of their relationship.

Datasets with our required characteristics are rare even in English, a high-resource language. We filtered two datasets: DREAM [23], a multiple-choice dialogue-based reading comprehension examination dataset and COSMOS-QA [8], a commonsense-based reading comprehension dataset with option semantic references, to include the required formal characteristics. In addition, we entail a more general dataset collected from English examinations in China designed for middle school and high school students, RACE [11], to expand the sample capacity and thus improve the generalization ability in MMRC tasks.

To validate the effect of cross-lingual transfer, we selected three datasets with relatively low-resource language under the MMRC scenario including C^3 [24] which contains dialogues and more formally written mixed-genre texts in Chinese, SweQUAD-MC [10] which has very limited sample capacity in Swedish, and MuSeRC [22] which requires multi-sentences reasoning in Russian. For the record, MuSeRC is designed initially as a binary classification task defining whether the options are true or false according to passage and question. Since each passage and question pair corresponds to multiple options, we cast this dataset into MMRC format.

Statistics of these datasets are shown in Table 2. Examples are provided in Appendix A, Table 7, 8, 9, 10, 11 and 12. Our methods for full context components mainly focus on modelling the relevance of options, hence the performance of models on COSMOS-QA would be followed with the most interest. For unified tuning, we pay balanced attention to each target dataset.

4.2 Experiment Settings

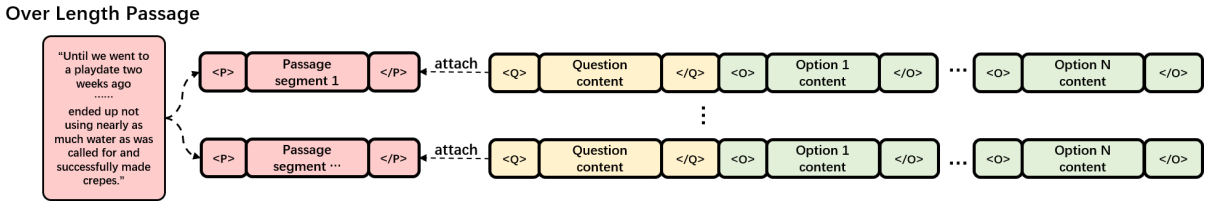

Figure 7: Illustration of truncation for over-length passages. The attached question and options for every segment are complete.

In all experiments, we set the maximum sequence length to 512 tokens including passage, question, all options, and extra tags to match the settings of selected pre-trained models. We truncate the passage which exceeds the length limit into segments and each of them is attached with corresponding question and options as input, as shown in Figure 7, and the statistics of truncation is shown in Table 3. The extracted features of passage, question, and options from different segments would be averaged.

Split	Statistics	RACE	COSMOS-QA	DREAM	C^3	SewQUAD-MC	MuSeRC
Train	Full	68648	25262	5855	11219	461	4897
	Truncated	19218	0	261	650	506	483
	Avg. truncation count	1.14	0	1.6	1.12	2.11	1.39
	Max. truncation count	6	0	4	3	14	5
Eval	Full	3839	2985	1962	3574	66	883
	Truncated	1048	0	78	242	61	110
	Avg. truncation count	1.12	0	1.42	1.12	1.79	1.22
	Max. truncation count	4	0	4	2	13	3
Test	Full	3983	6963	1943	3709	42	-
	Truncated	951	0	98	183	60	-
	Avg. truncation count	1.16	0	1.52	1.07	1.6	-
	Max. truncation count	4	0	3	2	9	-

Table 3: Statistics of truncation in all involved datasets.

We apply the *AdamW* optimizer and adopt the warming up and linear learning rate decay strategy in our training. Besides, our models are trained with automatic mixed precision provided by *PyTorch*, and we modify some hyperparameters of the optimizer accordingly to prevent value overflow.

We load only the pre-trained weights for the encoder network from *Transformers*[2] and initialize the weights of the fuser network and reasoning network for every training. For evaluation, we use accuracy as the indicator to measure the performances of our methods and models for all

[2]https://huggingface.co/

tasks, as common in MMRC tasks.

4.3 Results

Method	Intermediate Tasks				Target Tasks	
	RACE	COSMOS-QA*	DREAM	C^3	SweQUAD-MC	MuSeRC*
Single-Tasking						
baseline	79.44	73.79	73.33	79.34	87.60	91.72
+FCC and Tagging	78.36	74.15	71.42	76.85	89.58	91.38
+Fuser Network (Full Methods)	78.56	74.69	72.75	77.58	91.89	91.78
Unified Tuning						
Full Methods	78.68	75.33	81.80	78.17/82.73	92.16/98.29	86.61/94.39

Table 4: Result of the XLM-RoBERTa experiments. The number in this table represents the accuracy in percentage achieved on datasets with specific settings. The result of target tasks in unified tuning contains zero-shot/fine-tuning accuracy. Mark * indicates the result is evaluated on the development set since their test sets are hosted[3].

We apply our methods step by step on a pre-trained multilingual language model, XLM-RoBERTa, to assess their impact on all selected datasets respectively as the experiment results shown in Table 4. Note that in unified tuning, the model is jointly post-trained on all intermediate datasets and then fine-tuned on each target dataset respectively.

Taking full context components as input directly improves the model performances on COSMOS-QA and SweQUAD-MC compared to the baseline method by 0.36% point and 1.98% point. Semantic- and feature-level characteristics are both taken into account when a fuser network is added and it improves the performance compared to simply converting the input format as full context components on all selected datasets.

For full methods single-tasking, the improvement of accuracy on COSMOS-QA by 0.90% point and SweQUAD-MC by 4.29% point are also substantial, while the impact on DREAM and MuSeRC is relatively inconspicuous. However, it impairs the performances compared to the baseline method on RACE by 0.88% point and C^3 by 1.76% point.

We notice when DREAM is jointly trained with RACE and COSMOS-QA as an intermediate task, it achieves the most salient accuracy improvement compared to the single-tasking. For RACE and COSMOS-QA, their performances are also improved in joint training.

[3]https://leaderboard.allenai.org/cosmosqa/submissions/public,https://russiansuperglue.com/tasks/task_info/MuSeRC

Prominent zero-shot transfer capability emerges while unified tuning is applied. We notice on two of the three target datasets, C^3 and SweQUAD-MC, the zero-shot accuracy has exceeded their single-task performance. With further fine-tuning, a striking improvement arises on all target datasets of unified tuning, significantly surpassing their single-task results of baseline and full methods.

To assess the impact of the encoder network and compare our methods to previous research, we also experiment with our methods on the different encoders pre-trained merely in English, which have different pre-training strategies and architecture. From the results shown in Table 5, we can see another particular situation, which is our methods impairs the performances of all selected English datasets on BERT-Large. While on RoBERTa-Large and ALBERT-xxLarge, they can improve the performance of RACE and COSMOS-QA.

Model	RACE	COSMOS-QA	DREAM
Baselines			
BERT-large	72.0	67.1	66.8
RoBERTa-large	83.2	80.6	85.0
ALBERT-xxlarge	86.5	82.3	88.5
Our Methods			
BERT-large	68.8(-**3.2**)	63.2(-**3.9**)	55.1(-**11.7**)
RoBERTa-large	84.1(+**0.9**)	83.9(+**3.3**)	84.0(-**1.0**)
ALBERT-xxlarge	87.2(+**0.7**)	86.0(+**3.7**)	87.9(-**0.6**)
Human Performance	94.5	94.0	95.5

Table 5: Accuracy on RACE, COSMOS-QA, and DREAM. Baseline performances on RACE are from the previous research papers of language models [13, 14]. COSMOS-QA baselines are provided by its original paper [8] and Tian et al. [26]. DREAM baselines are from Jin et al. [9]

The performance variation on BERT, RoBERTa, and XLM-RoBERTa reflects the impact of pre-training tasks and data on our methods since all these three encoders have identical architecture (24 layers, 16 attention heads, and 1024 hidden dimensions) but different sizes of the vocabulary (30k on BERT, 50k on RoBERTa, and 250k on XLM-RoBERTa). From the overall trend represented by experiments of English encoders, the higher the training quality, the more suitable the model is for our method.

The results of current experiments suggest that in most cases our method can improve the performance of COSMOS-QA, which has obvious option relevance, and it is affected by the pre-training tasks and data of encoders. Besides, by comparing the result of English encoders and the multilingual encoder, taking full context components has an inconsistent impact on RACE. For DREAM, our methods always have a negative impact on single-tasking.

As for our target datasets, unified tuning can improve the performances of them all on XLM-RoBERTa. Compared to the baseline in single-tasking, it improves the accuracy of C^3 by 3.39% point, SweQUAD-MC by 10.69% point, and MuSeRC by 2.67% point. Additionally, the

performance of RACE, COSMOS-QA and DREAM are improved compared to applying our methods in single-tasking when they are jointly trained as an intermediate task, which is most outstanding on DREAM.

5 Discussion

Although we see the influence of full context components in previous experiment results, we cannot conclude that the improvement and deterioration of performances are caused by introducing option relations and cannot explain the decline of accuracy on DREAM and C^3 in single-tasking. In this section, we discuss the phenomenon that occurred in previous experiments and try to give explanations for them with extra experiments.

5.1 Visualization of Inferences

To analyze the modelling of options relations, we apply Integrated Gradients(IG) to our fine-tuned model and input samples for analyzing the contribution of tokens to the likelihood of options. This is accomplished by *Captum*[4] library based on *PyTorch* and we use its default visualization scheme. In our implementation, we select the embedding of $[PAD]$ token as baseline.

Figure 8: Attribution for all options of a sample from COSMOS-QA. Tokens marked in green indicate positive and in red indicate opposite for the option. We use coloured boxes to indicate the attribution target option and their corresponding text.

An example is shown in Figure 8. In this example, the total attribution score reveals the overall propensity of all components towards the corresponding option and the correct option, option D in this case, has the highest score. When we look into the word importance, which is the contribution of each token, we notice not only the tokens in the passage and question but also the tokens in other options are supplementing the result.

[4] https://captum.ai/

From the attribution of options A, B, and C, we can see that except for themselves, tokens in other options generally provide a negative contribution. As for correct option D, tokens in the passage, question, and other options would mostly contribute positively. This could be a piece of evidence for proving taking full context components could be beneficial to modelling option relevance where all options are contributing to the answer formation.

When we apply IG on more samples to all models from previous experiments, the tokens in each option still show a supplement to the result. This phenomenon is independent of the option characteristics of involved training data and the training schemes (single-tasking and unified tuning). From this, we can conclude that our method indeed introduces a comparison of information between options and can model the option relevance.

5.2 Effect of Unified Tuning

The result of unified tuning from the previous section suggests that it can decouple task format from specific task contents. Although the languages and genres are different in the intermediate task and target task, it still presents gratifying transfer capability.

This would be meaningful when we apply our methods to mental health detection scenarios. Predictably, the data under this scenario is typically low-resource and may span multiple languages in practice. After we have obtained the post-pre-trained model on the intermediate task, we could directly deploy it and optimize it in the mental health detection scenario with a modest amount of data.

The abrupt accuracy improvement of DREAM in the intermediate training stage is unexpected. We figure it might be related to its conversation format or test set partitioning. Considering C^3 also contains dialogue text as passages, we investigate the impact of such kind of textual form.

5.3 Impact of Dialogue Text

We conduct an extra experiment on C^3 to see the impact of our method on dialogue text. **We use its two original subsets: dialogue set and mixed-genre text set to conduct experiments to see how their performance change while they are trained separately and jointly.**

The results shown in Table 6 suggest that modelling dialogue text would be harder for both methods. We look into the difference between baseline and our methods, we notice joint training might expand the performance difference between the two methods on dialogue text from 1.00% point to 2.14% point while reducing it on mixed-genre text from 1.65% point to 1.40% point. Based on this result, we believe it is appropriate to conclude that the performance

Split	Separate		Joint	
	Dialogue	Mixed-genre	Dialogue	Mixed-genre
Ours	72.78	76.52	76.25	78.83
Baseline	73.78	78.17	78.39	80.23

Table 6: Accuracy of C^3 subsets with separate and joint training on XLM-RoBERTa.

difference on the whole C^3 dataset is mainly caused by the dialogue text.

When the dialogue subset is trained with the mixed-genre subset, the performance gap between them is reduced with both methods from 3.74% point to 2.58% point and from 4.39% point to 1.84% point respectively. This result agrees with the situation on DREAM, which could obtain a great improvement with joint training with RACE and COSMOS-QA.

Meanwhile, we notice that when COSMOS-QA is jointly trained with DREAM and RACE, its performance is also improved compared to all methods in single-tasking. This means that our method may still have positive benefits when applied to datasets containing dialogue text and option relevance.

5.4 Investigation of The Fuser Network

We can expect the depth of the fuser network to have a certain influence on the result, while the relationship between depth and effect is uncertain. To further investigate its influence, we explore the relationship between its depth and impact on performances. The performance change on the COSMOS-QA dataset caused by its depth is shown in Figure 9.

As we can see from the result, there is no clear pattern in either accuracy or standard deviation for deepening the fuser network. This result is frustrating since it could not give much information about the setting of it and the depth of it would remain a hyperparameter which needs to be adjusted according to the dataset.

The only thing we can be sure of is that adding a fuser network can indeed improve the model performance under full context components input. From a practical perspective, adding a single-layer fuser network ensure improved performance.

5.5 Reduction of Memory Occupancy

In forward propagation, the memory occupancy of all intermediate computation results could be regarded as linear to the length of the input token sequence. For a sample with n options

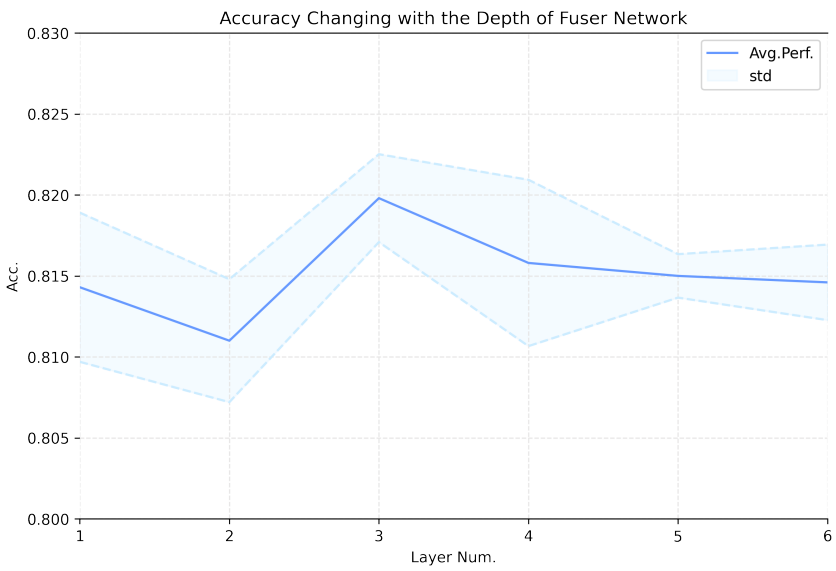

Figure 9: Accuracy on COSMOS-QA validation set with different depth of Fuser Network.

and the length of the passage, question, and each option is p_l, q_l, and o_l, we can have the total memory occupancy caused by this sample is linear to $n * (p_l + q_l + o_l)$ when the baseline method is applied.

As for full context components input, the memory consumption would be reduced to linear to $p_l + q_l + n * o_l$. We roughly apply the average length of all components to this equation and we can have a cursory estimation of diminution: 74% for RACE, 68% for COSMOS-QA, 63%for DREAM, 72% for C^3, 74% for SweQUAD-MC, and 73% for MuSeRC.

This means that taking the full context components as input could significantly reduce memory occupancy, especially when the samples have long passage text. It would be meaningful for expanding the batch size with limited memory capacity while having long text input. As a result, with full context components input format, the RoBERTa-Large model could be trained on the COSMOS-QA dataset with batch size 16 on a single RTX3090 GPU without any other memory trick.

6 Conclusions

In this work, we analyze the data processing requirements in the mental health detection scenario, including adaptability in multiple inspections and evaluations, capability in low-resource settings, and usability in cross-lingual. For the above analytical demands, we propose a multilingual solution in the form of an MMRC task and introduce new methods, **including inserting extra tags and applying a transformer encoder network to fuse information of options set**, for identifying components of inputs to enable the model to attentively recognize the inspections and corresponding evaluation criteria with explicit semantics. Moreover, by adopting a stepwise training scheme, we decouple the knowledge related to task format from languages and genres to make use of high-resource language data and transfer it to the low-resource scenarios.

As a result, our methods can be applied to MMRC tasks and improve over the performances of prior methods on datasets with similar characteristics to our target mental health detection data on pre-trained language models in most cases. It indicates that our methods help the models understand the relations among all the components of input, which can be regarded as beneficial to the modelling of evaluation criteria. Furthermore, the results of unified tuning show our methods can transfer the knowledge of the task format to other datasets, which complies with our demand.

Due to the absence of open-source mental health detection datasets in the expectation of our needs, all our experiments are conducted on available MMRC datasets. Although our results cannot be directly evaluated as valid for mental health detection, observations based on experimental phenomena indicate that they can meet some of the requirements for such tasks.

References

[1] Hayda Almeida, Antoine Briand, and Marie-Jean Meurs. Detecting early risk of depression from social media user-generated content. In *CLEF (working notes)*, 2017.

[2] Sheshadri Iyengar Raghavan Bhagyashree, Kiran Nagaraj, Martin Prince, Caroline HD Fall, and Murali Krishna. Diagnosis of dementia by machine learning methods in epidemiological studies: a pilot exploratory study from south india. *Social psychiatry and psychiatric epidemiology*, 53(1):77–86, 2018.

[3] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. Xlm-e: Cross-lingual language model pre-training via electra. In *ACL 2022*, June 2021.

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] M Hamilton. Hamilton anxiety scale. *Group*, 1(4):10–1037, 1959.

[8] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. *CoRR*, abs/1909.00277, 2019.

[9] Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-Tür. MMM: multi-stage multi-task learning for multi-choice reading comprehension. *CoRR*, abs/1910.00458, 2019.

[10] Dmytro Kalpakchi and Johan Boye. Bert-based distractor generation for swedish reading comprehension questions using a small-scale dataset. *CoRR*, abs/2108.03973, 2021.

[11] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[12] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[15] Tom L Osborn, Christine M Wasanga, and David M Ndetei. Transforming mental health for all, 2022.

[16] John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706, 2010.

[17] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *CoRR*, abs/1906.01502, 2019.

[18] Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. What to pre-train on? efficient intermediate task selection. *CoRR*, abs/2104.08247, 2021.

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[20] Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. Option comparison network for multiple-choice reading comprehension. *CoRR*, abs/1903.03033, 2019.

[21] Jessica Ross, Thomas Neylan, Michael Weiner, Linda Chao, Kristin Samuelson, and Ida Sim. Towards constructing a new taxonomy for psychiatry using self-reported symptoms. In *MEDINFO 2015: eHealth-enabled Health*, pages 736–740. IOS Press, 2015.

[22] Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. Russiansuperglue: A russian language understanding evaluation benchmark. *arXiv preprint arXiv:2010.15925*, 2020.

[23] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *CoRR*, abs/1902.00164, 2019.

[24] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Probing prior knowledge needed in challenging chinese machine reading comprehension. *CoRR*, abs/1904.09679, 2019.

[25] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

[26] Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yantao Jia, and Zhicheng Sheng. Scene restoring for narrative machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3063–3073, Online, November 2020. Association for Computational Linguistics.

[27] Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and psychosomatics*, 84(3):167–176, 2015.

[28] Tung Tran and Ramakanth Kavuluru. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148, 2017.

[29] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. Dcmn+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9563–9570, 2020.

[32] Yuxiang Zhang and Hayato Yamana. Hrca+: Advanced multiple-choice machine reading comprehension method. *LREC*, 2022.

[33] Pengfei Zhu, Hai Zhao, and Xiaoguang Li. Dual multi-head co-attention for multi-choice reading comprehension. *CoRR*, abs/2001.09415, 2020.

A Data Examples

Passage

When you buy a T-shirt, or a fur coat in a store, it often carries a label telling who made it or from what store it was bought. Indeed, some labels show the dress is famous and it is very expensive, so buyers who deal with the cheapest products would be pleased to do away with labels entirely. However, there is another label more important than the one showing from which store the dress was bought. When a person buys a fur coat, or a jacket, from a store, a label telling what the product is made of should be carried to it. This label is required by law. Besides telling what the product on show is made of, the label should be in clear English and be where one can find it easily. The information on the label must be the truth. The reason for this label is that most buyers today aren't expert enough to know exactly what kind of fur or material they are buying. The buyer must believe in the store that sells the products or in what the labels say.

Question

Which of the following is true?

Options

A. Not all buyers know the materials they are buying. ✓
B. A fur coat with a high price often carries a false label.
C. A label only says what material the product is made of.
D. A T-shirt seldom carries a label.

Table 7: An example of RACE dataset.

Passage

I polished some silver . I showered and attempted to beautify myself for my day despite my obvious lack of beauty sleep. I accepted my fate of scary dark under-eye circles. I am right now, to put it as succinctly as possible, a complete zombie .

Question

What may happen if you miss your beauty sleep?

Options

A. I would still look good.
B. I wouldn't be able to get anything done that day.
C. I would have to polish silver.
D. None of the above choices. ✓

Table 8: An example of COSMOS-QA dataset, with option relevance.

Passages

M: Good evening, Madam. Could you do me a favor?
W: Of course. What can I do for you?
M: I am looking for a hotel. Are there any hotels near here?
W: Yes, there are some in this street. The nearest one is next to the bank. It's quite modern.
M: You see. I'm leaving tomorrow morning. Do you think there're any hotels not too expensive?
W: Yes. Drive on for five minutes and you'll find a yellow building on your left. It's a family-style hotel, very comfortable, and the price is quite reasonable.
M: It sounds nice. Thank you very much for your help.
W: You are welcome.

Question

What can you learn from the conversation?

Options

A. The man has lost his way home.
B. The woman knows the place very well. ✓
C. The woman works in a modern hotel.

Table 9: An example of DREAM dataset, with dialogue passage.

Passages

女：快出来帮我拿一下儿东西。
(Woman: Come out and help me get something.)
男：怎么买这么多?
(Man: Why do you buy so much?)
女：今天的菜又新鲜又便宜，就多买了点儿。
(Woman: Today's dishes are fresh and cheap, so I bought more.)
男：这么多菜，什么时候能吃完呢?
(Man: There are so many dishes, when can we finish eating?)

Question

女的让男的帮什么忙?
(What do women ask men to do?)

Options

A. 买菜
(A. grocery shopping)
B. 拿东西✓
(B. take things)
C. 吃菜
(C. eat dishes)

Table 10: An example of C^3 dataset, with dialogue passage. English translation attached.

Passages

Miljö - och hälsoskyddsinspektör (Environmental and health protection inspector)
De allra flesta inspektörerna är anställda i kommunerna. En del arbetar även på länsstyrelser eller på myndigheter som Livsmedelsverket, Jordbruksverket och Naturvårdsverket.
(The vast majority of inspectors are employed by the municipalities. Some also work at county administrative boards or at authorities such as the Swedish Food Agency, the Swedish Agency for Agriculture and the Swedish Environmental Protection Agency.)
Arbetsuppgifter (Job duties)
Miljö- och hälsoskyddsinspektörer har en naturvetenskaplig utbildning och den kunskapsgrunden är viktig i yrket. Men arbetet har ett också ett mycket stort fokus på juridik och förvaltning.
(Environmental and health protection inspectors have a natural science education and that knowledge base is important in the profession. But the work also has a very large focus on law and administration.)
Arbetet handlar om att kontrollera att lagar och förordningar följs inom till exempel miljöbalken, livsmedelslagen eller i EU-förordningar. Miljö- och hälsoskyddsinspektörer besöker och utövar tillsyn bland annat på restauranger, industrier, bostäder, vattenverk och enskilda avloppsanläggningar.
(The work involves checking that laws and regulations are followed within, for example, the Environmental Code, the Food Act or EU regulations. Environmental and health protection inspectors visit and supervise, among other things, restaurants, industries, homes, waterworks and individual sewage plants.)
I små kommuner arbetar inspektörerna ofta med större delar av ansvarsområdet. I större kommuner är de ofta mer specialiserade.
(In small municipalities, the inspectors often work with larger parts of the area of responsibility. In larger municipalities, they are often more specialized.)
Arbetet är självständigt och innebär mycket kontakter med olika människor. Miljö- och hälsoskyddsinspektörer samarbetar till exempel med andra tjänstemän inom kommuner, företag och myndigheter. Ibland arbetar man ensam och ibland i arbetslag.
(The work is independent and involves a lot of contact with different people. Environmental and health protection inspectors collaborate, for example, with other officials within municipalities, companies and authorities. Sometimes you work alone and sometimes in teams.)
Rådgivning och information är en viktig del av arbetet. Inspektören informerar om de bestämmelser som gäller och ger råd, både till företag, myndigheter och allmänheten.
(Advice and information is an important part of the work. The inspector informs about the regulations that apply and gives advice, both to companies, authorities and the public.)
Miljö- och hälsoskyddsinspektören gör även utredningar och bereder ärenden som politiska nämnder ska fatta beslut om. Det ingår ofta att föredra ärendet för politikerna.
(The environmental and health protection inspector also conducts investigations and prepares cases for political committees to decide on. It often includes preferring the matter to the politicians.)
Arbetsmiljö (Working environment)
Arbetet är omväxlade och innebär både kontorsarbete och besök ute hos de verksamheter som man ska göra tillsyn på. Det finns oftast möjlighet att själv planera och strukturera sitt arbete.
(The work is varied and involves both office work and visits to the businesses that are to be supervised. There is usually the opportunity to plan and structure your work yourself.)

Question

Var är de flesta miljö- och hälsoskyddsinspektörer anställda?
(Where are most environmental and health protection inspectors employed?)

Options

A. enskilda avloppsanläggningar
(A. individual sewage plants)
B. i kommunerna ✓
(B. in the municipalities)
C. på restauranger
(C. in restaurants)

Table 11: An example of SweQUAD-MC dataset. English translation attached.

Passages

С согласия госпожи де Франваль Вальмон увозит Эжени, но Франваль догоняет их и убивает Вальмона.
(With the consent of Madame de Franval, Valmont takes Eugenie away, but Franval overtakes them and kills Valmont.)
Затем, дабы избежать кары правосудия, Франваль бежит в один из своих удалённых замков и берет с собой жену и дочь.
(Then, in order to avoid the punishment of justice, Franval runs to one of his remote castles and takes his wife and daughter with him.)
Узнав, что Эжени была похищена с ведома его жены, он решает отомстить госпоже де Франваль и поручает дочери отравить мать.
(Learning that Eugenie was kidnapped with the knowledge of his wife, he decides to take revenge on Madame de Franval and instructs his daughter to poison her mother.)
Сам же он вынужден бежать за границу, ибо ему вынесен смертный приговор.
(He himself is forced to flee abroad, for he has been sentenced to death.)
По дороге на Франваля нападают разбойники и отбирают у него все, что он имел.
(On the way, robbers attack Franval and take from him everything that he had.)
Израненный и измученный Франваль встречает Клервиля: достойному священнику удалось выбраться из застенков негодяя.
(Wounded and exhausted, Franval meets Clairville: the worthy priest managed to get out of the dungeons of the scoundrel.)
Однако, исполненный христианского смирения, Клервиль готов помочь своему мучителю.
(However, filled with Christian humility, Clairville is ready to help his tormentor.)
По дороге Франваль и Клервиль встречают мрачную процессию — хоронят госпожу де Франваль и Эжени.
(On the way, Franval and Clairville meet a gloomy procession - they bury Madame de Franval and Eugenie.)
Отравив мать, Эжени внезапно почувствовала столь жгучее раскаяние, что в одночасье умерла возле хладного тела матери.
(Having poisoned her mother, Eugenie suddenly felt such burning remorse that she died overnight near the cold body of her mother.)
Бросившись на гроб жены, Франваль закалывает себя кинжалом.
(Throwing himself on his wife's coffin, Franval stabs himself with a dagger.)
Таково преступление и «ужасные плоды его»...
(Such is the crime and "the terrible fruits of it" ...)

Question

Кто отравил жену Франваля?
(Who poisoned Franval's wife?)

Options

A. Мать
(A. Mother)
B. Людовик
(B. Louis)
C. Сам Франваль
(C. Franval himself)
D. Его дочь ✓
(S. His daughter)

Table 12: An example of MuSeRC dataset. English translation attached.

B Draft Condensed Version

A Solution for Mental Health Evaluation with Full Context Components

Anonymous EMNLP submission

Abstract

Text-based models for multi-purpose inspection and evaluation are absent in the mental health detection field. We propose a solution in the form of Multi-choice Machine Reading Comprehension (MMRC) tasks combined with modification of the input format and network architecture for such a scenario. In our solution, we 1) introduce extra tokens for the identification of different components of an MMRC sample and reconstruct the whole sample as one input sequence, and 2) insert a small transformer encoder network to fuse the information from every option. Through the above modifications, we input the text records, evaluation, and corresponding measurement in text form, and train the neural network to output a valid assessment. Experiment results suggest our methods can be efficient in MMRC tasks especially when option relevance exists.

1 Introduction

In 2019, the World Health Organization conducted an investigation on the global mental health care situation and revealed the universally scarce resources in mental support (Osborn et al., 2022). From the released data, only 2% of the total health budget goes to mental health care, two-thirds of countries are short of funds and human resources for implementing their mental health policy or plan, and around 970 million people, accounting for 13% of the population in the world could have suffered from a mental disorder during a certain period of their life. Moreover, there is a severe regional imbalance in the resources of mental health support, especially between developed and less developed regions. Henceforth, the WHO calls for wider support for mental health services, including affordable and accessible mental health care for all.

In addition to the efforts in the medical field, the development of applications on smart devices, especially online consultations and chatbots for mental care (Vaidyam et al., 2019) contribute to the accessibility of mental health support. Nevertheless, the evaluation of mental condition requires standardisation of these textual records in the form of standard scales such as the Hamilton Anxiety Rating Scale (HAM-A) (Hamilton, 1959), the World Health Organization 5 indicators for physical and mental health (WHO-5) (Topp et al., 2015), and the Beck Depression Inventory. Some of these standard scales need to be completed manually by professionals based on the conversation during the consultation, which requires additional human resources costs.

The surge of machine learning techniques has brought many automated applications to mental health evaluation, diagnosis and therapy. Beyond AI-based chatbots, regarding public mental health services, there are models for identifying specific symptoms in social media posts (Almeida et al., 2017). In medical scenarios, models work as an auxiliary to process the featured data and give diagnostic suggestions (Pestian et al., 2010; Ross et al., 2015; Tran and Kavuluru, 2017; Bhagyashree et al., 2018). Although these applications cover the evaluation of various mental conditions, the existing models always require a specific design for every specific inspection, while every standard scale consists of multiple different items corresponding to different inspections.

We propose a method for mental health detection problems in the form of MMRC (**M**ulti-choice **M**achine **R**eading **C**omprehension), as shown in Table 1, to enable a model to process multiple inspections in standard scales. By taking the free-text records as passages, and explicitly declaring inspection and evaluation measurement in the text as questions and options to formulate a complete sample, shown in Table 2, we cast a sample of mental health evaluation task into the MMRC task format and use pre-trained language models to characterize the relationships within.

Passage:
Until we went to a playdate two weeks ago. Thea's mom is Serbian and crepes are apparently as common in Serbia as they are in France. We discussed the batter, the texture, the cooking process, the topping options and Dee generally brought me up to speed. Being not brave enough to just start throwing ingredients in a bowl as she did, I got a recipe of the internet for general proportions, ended up not using nearly as much water as was called for and successfully made crepes.
Question:
Why did they discuss crepes?
Options:
A. Because Thea's mom is Serbian.
B. Because the writer got a recipe from the internet.
C. Because the writer is interested in learning how to cook crepes.
D. None of the above choices.

Table 1: An example MMRC task from the COSMOS-QA dataset.

Record:
Agent: How are you today?
User: Not great...
Agent: Looks like you're having a rough day. Tell me more about how you feel?
User: I went out and felt very uncomfor--table with so many people around. Can't help feeling being gazed at.
Agent: As always, let's start by turning our attention to our breath.
...
Agent: It happens to all of us. Here are some techniques that can help.
Inspection:
Indication of User's anxious mood. (HAM-A item 1)
Measurement of Evaluation:
A.Not present.
B.Mild.
C.Moderate.
D.Severe.
E.Very severe.

Table 2: An example of simulated consultation record between a user and a conversational agent for mental health support, *Wysa*[1], combined with a HAM-A evaluation item and its corresponding measurement.

Unfortunately, the datasets for mental health detection are mostly not publicly available. Therefore, we will use the open-source MMRC datasets with similar characteristics to our target data as an alternative and optimize the methods specifically to adapt to these characteristics.

Our contribution can be summarized as introducing additional information through tag tokens and adding a fuser network to enhance the comparison between options to allow the model to recognize the measurement of mental health evaluation, or the option relevance in the context of the MMRC task.

2 Related work

MMRC is a subtask in Machine Reading Comprehension consisting of three main types of components: passage **p**, question **q** and answer options $\mathbf{o} = o_1, o_2, ..., o_n$. The purpose of such a task is to select the best matching answer o_i corresponding to question **q** from information provided by **p**. Normally, the questions and corresponding candidate options would vary according to the passages, which coincides with our expectation of modelling different measurements for different evaluations and inspections.

In 2019, Jin et al. (Jin et al., 2019) proposed a multi-stage multi-task training strategy and achieved state-of-the-art performance in many MMRC datasets with coarse-tuning on an NLI dataset and fine-tuning on supplement and target datasets in the form of multi-tasking with newly proposed multi-step attention network classifier. Later in 2020, Zhang et al. (Zhang et al., 2020) and Zhu et al. (Zhu et al., 2020) proposed DCMN+ and DUMA respectively, both focusing on modelling the relationship between passages and question-option pairs with extra cross-attention mechanisms. This methodology is brought to the ultimate attainment by Zhang (Zhang and Yamana, 2022) with HRCA+, which applied attention mechanism to model all 9 types of the correspondence relationship between three components.

The conventional methods or models for MMRC tasks take only the triple of passage, question and one of the options as a sequence and most of the previous efforts focus on the enhancement of modelling the triple. It could cause critical problems when there exist references between options (e.g.

[1]https://www.wysa.com/

'none of the other options are correct', 'all of the options are correct', etc.), as shown in Table 1. More importantly, when options are in the form of scoring, the models need information about intervals and gaps (e.g. In grades 1 to 5 and 1 to 10, the same grade 5 has different meanings) which could be recognized only with the co-occurrence of all options.

Although Ran et al. noticed this problem and proposed an Option Comparison Network (Ran et al., 2019), which takes the correlation between options based on token-level embeddings into account after retrieving each permutation respectively, this feature level aggregation still overlooked the potential direct semantic relation between options.

For the above issue, we believe there is a solution which is feeding the model all components of one MMRC task entry including passage, question, and all candidate options at once and using its encoding ability to directly encode the full context of such task. Meanwhile, this might require extra effort for the model to recognize different components and thus model the relationships between them.

3 Methods and Model

3.1 Tags for Full Context Input

There are three categories of components in MMRC tasks: passages, questions, and options, while former methods allow only one component of each category in one sequence. Moreover, they use $[CLS]$ token and $[SEP]$ token to mark the boundary of different components and retrieve either the embedding of $[CLS]$ token or mean pooling of the sequence as features for further processing.

Formally, the representations of these special tokens are contextualized and they jointly create an explicit distribution pattern for the input sequence. This inspires us that additional prior information could be added to the sequence by inserting particular special tokens in corresponding positions.

We expect that the model can identify different components and establish the semantic-based relationship between them according to their categories, especially when all components are fed together as one sequence. Hence we need identifiers to mark not only the boundary but the category of components in the input sequence for there exists an indefinite number of components from the options category, determined by the datasets. On account of the above-mentioned issues, we design learnable extra tag tokens for indicating the model to recognize the components.

We modify the modelling to be in the form of $P(O^i|P, Q, O^1, ..., O^N)$, where $i = 1, 2, ..., N$ is the index of N candidate options for corresponding passage P and question Q, by allowing the model to read all components at once to construct a full context with different start and end tags identifying every component in it. It is worth mentioning that the tags for all options are identical to ensure the impact of tags on each option is equivalent and it does not require adjustment on the number of introduced extra tag tokens according to the number of options.

Associating with the application of special tokens, including $[CLS]$ for classification tasks, and $[MASK]$ for cloze tasks, we can infer that the intrinsic information of the token itself could be neglected by the network when it is trifling in downstream tasks while the task-relevant context feature would be reinforced. Therefore, we can consider the output embedding of each token to be a characterization of its context, rather than just its own representation.

Similar to the usage of $[CLS]$ and $[MASK]$ tokens, we assume that the embedding of tags can represent the task-relevant information of each component in the input sequence and thus retrieve them as component features in full context for the downstream classifier. The models would be trained in a supervised manner which makes use of the labels as supervision information.

Our goal is to create appropriate representations for components according to the context. With extra knowledge about the boundaries and categories introduced by our tagging method, we expect the model to have different concerns about the content that appears in different components and aggregate the task-related features of them into the embeddings of tags.

3.2 Model Architecture

In addition to the above methods, we use neural networks to achieve our modelling goal. As shown in Figure 1, our model consists of three parts: encoder network, fuser network, and reasoning network.

The Encoder network is originally a pre-trained language model, defined as f_{ENC}. It would become a tag-oriented feature extractor along with the training when we only retrieve the embeddings of tags. These embeddings carry the task-relevant features at the semantic level in the co-occurrences

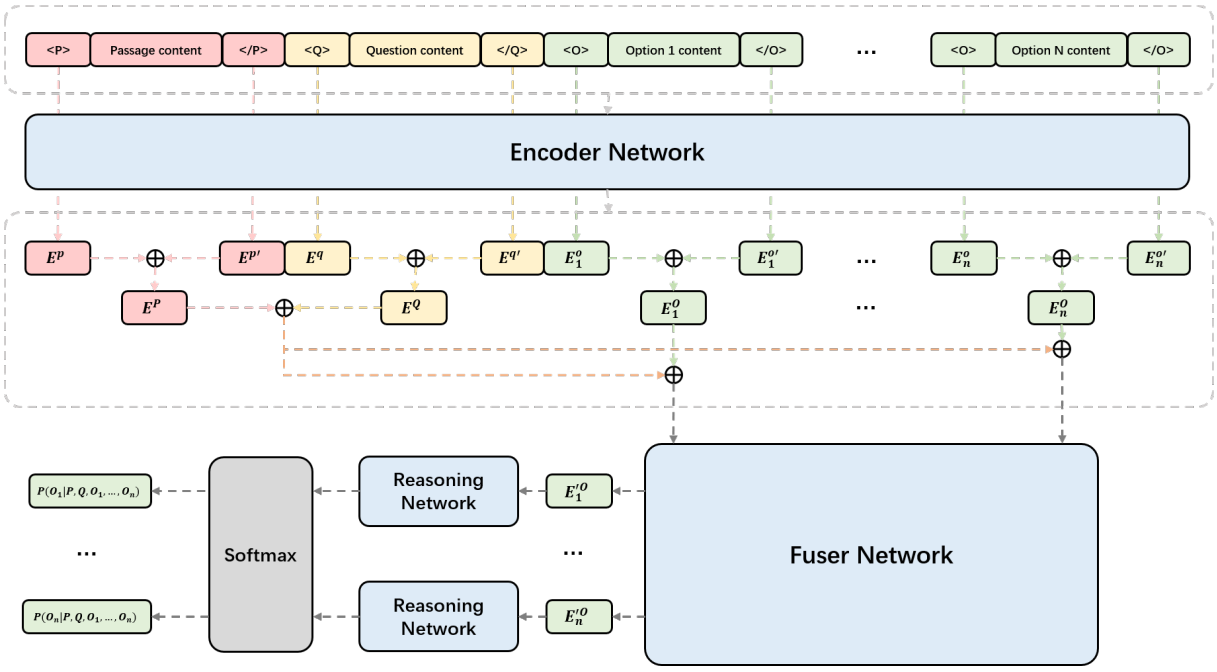

Figure 1: Model architecture. The encoder network is a pre-trained language model, the fuser network is a transformer encoder and the reasoning network is a multi-layer perceptron.

of full context components, which is enabled by the contextualized representation capability of the pre-trained language model.

By feeding the tagged full context sequence $S = concate(p, q, o_1, ..., o_n)$ into the encoder network, we can obtain the corresponding contextual embedding of every token. Considering the purpose of tag-oriented modelling, we retrieve only the embedding of tags as component features for further processing. Based on our previous discussion, we adopt the representation of tags as features for different components and take the mean value of the tag embeddings on both sides to align the features of each dimension.

On top of the pre-trained language model, we insert another transformer encoder network to model the relationship between all options at the feature level, where we take the embeddings of E^P, E^Q and E^O_i to reconstruct the representation of every option by adding their embeddings. These embeddings are the result of the same transformation thus they are in the same subspace having a common basis and it is plausible to assume they are additive.

In particular, we removed the position embedding in this module since the option sets are in fact permutation-invariant and the transformer encoder is position-independent if position encoding is avoided. We use it to strengthen the discrimination between options and directly formulate a new representation in consideration of the whole option set for each of them.

Based on the structure of the transformer encoder, we can assume that this network can create representations of each option by attending to the representation of other options in the option set. Henceforth, we refer to this module as a fuser network while it fuses the information of different options.

Afterwards, we generate logit values for options as a confidence level with a multi-layer perceptron and apply softmax smoothing on them to obtain the probability distribution. Finally, we can regard the probability distribution of options are conditioned to all components in the corresponding MMRC task including passage, question, and all options on the semantic level, noted as $P(o_i|p, q, o_1, ..., o_n)$.

4 Experiment

4.1 Datasets

Datasets with our required characteristics are rare even in English, a high-resource language. We filtered two datasets: DREAM (Sun et al., 2019), a multiple-choice dialogue-based reading comprehension examination dataset and COSMOS-QA (Huang et al., 2019), a commonsense-based reading comprehension dataset with option semantic references, to include the required formal characteristics. In addition, we entail a more general dataset

Datasets	RACE	COSMOS-QA	DREAM
# of train	87,866	25,262	6,116
# of validation	4,887	2,985	2,040
# of test	4,934	6,963	2,041
Avg. P length	321.9	70.3	85.9
Avg. Q length	10.0	10.6	8.6
Avg. O length	5.3	8.1	5.3
Language Characteristics	English Ordinary	English Reference	English Dialogue

Table 3: Statistics of datasets involved.

Method	RACE	COSMOS	DREAM
Single-Tasking			
baseline	79.44	73.79	73.33
+FCC and Tagging	78.36	74.15	71.42
+Fuser Network (Full Methods)	78.56	74.69	72.75
Joint-Training			
Full Methods	78.68	75.33	81.80

Table 4: Result of the XLM-RoBERTa experiments. The number in this table represents the accuracy in percentage achieved on datasets with specific settings.

collected from English examinations in China designed for middle school and high school students, RACE (Lai et al., 2017), to expand the sample capacity and thus improve the generalization ability in MMRC tasks.

Statistics of these datasets are shown in Table 3. Examples are provided in Appendix A. Our methods for full context components mainly focus on modelling the relevance of options, hence the performance of models on COSMOS-QA would be followed with the most interest.

4.2 Experiment Settings

In all experiments, we set the maximum sequence length to 512 tokens including passage, question, all options, and extra tags to match the settings of selected pre-trained models. We truncate the passage which exceeds the length limit into segments and each of them is attached with corresponding question and options as input. The extracted features of passage, question, and options from different segments would be averaged.

We apply the *AdamW* optimizer and adopt the warming up and linear learning rate decay strategy in our training. Besides, our models are trained with automatic mixed precision provided by *PyTorch*, and we modify some hyperparameters of the optimizer accordingly to prevent value overflow.

We load only the pre-trained weights for the encoder network from *Transformers*[2] and initialize the weights of the fuser network and reasoning network for every training. For evaluation, we use accuracy as the indicator to measure the performances of our methods and models for all tasks, as common in MMRC tasks.

4.3 Results

We apply our methods step by step on a pre-trained multilingual language model, XLM-RoBERTa, to assess their impact on all selected datasets respectively as the experiment results shown in Table 4. Note that in unified tuning, the model is jointly post-trained on all intermediate datasets and then fine-tuned on each target dataset respectively.

Taking full context components as input directly improves the model performances on COSMOS-QA compared to the baseline method by 0.36% point. Semantic- and feature-level characteristics are both taken into account when a fuser network is added and it improves the performance compared to simply converting the input format as full context components on all selected datasets.

For full methods single-tasking, the improvement of accuracy on COSMOS-QA by 0.90% point is also substantial, while it impairs the performances compared to the baseline method on RACE by 0.88% point and DREAM by 1.76% point.

We notice when DREAM is jointly trained with RACE and COSMOS-QA as an intermediate task, it achieves the most salient accuracy improvement compared to the single-tasking. For RACE and COSMOS-QA, their performances are also improved in joint training.

To assess the impact of the encoder network and compare our methods to previous research, we also experiment with our methods on the different encoders, which have different pre-training strategies and architecture. From the results shown in Table 5, we can see another particular situation that our methods impair the performances of all selected datasets on BERT-Large. While on RoBERTa-Large and ALBERT-xxLarge, they can improve the performance of RACE and COSMOS-QA.

The performance variation on BERT, RoBERTa, and XLM-RoBERTa reflects the impact of pre-training tasks and data on our methods since all these three encoders have identical architecture (24 layers, 16 attention heads, and 1024 hidden dimen-

[2]https://huggingface.co/

Model	RACE	COSMOS	DREAM
Baselines			
BERT-large	72.0	67.1	66.8
RoBERTa-large	83.2	80.6	85.0
ALBERT-xxlarge	86.5	82.3	88.5
Our Methods			
BERT-large	68.8(**-3.2**)	63.2(**-3.9**)	55.1(**-11.7**)
RoBERTa-large	84.1(**+0.9**)	83.9(**+3.3**)	84.0(**-1.0**)
ALBERT-xxlarge	87.2(**+0.7**)	86.0(**+3.7**)	87.9(**-0.6**)
Human	94.5	94.0	95.5

Table 5: Accuracy on RACE, COSMOS-QA, and DREAM. Baseline performances on RACE are from the previous research papers of language models (Lan et al., 2019; Liu et al., 2019). COSMOS-QA baselines are provided by its original paper (Huang et al., 2019) and Tian et al. (Tian et al., 2020). DREAM baselines are from Jin et al. (Jin et al., 2019)

sions) but different sizes of the vocabulary (30k on BERT, 50k on RoBERTa, and 250k on XLM-RoBERTa). From the overall trend represented by experiments of English encoders, the higher the training quality, the more suitable the model is for our method.

The results of current experiments suggest that in most cases our method can improve the performance of COSMOS-QA, which has obvious option relevance, and it is affected by the pre-training tasks and data of encoders. Additionally, the performance of RACE, COSMOS-QA and DREAM are improved compared to applying our methods in single-tasking when they are jointly trained, which is most outstanding on DREAM.

5 Discussion

Although we see the influence of full context components in previous experiment results, we cannot conclude that the improvement and deterioration of performances are caused by introducing option relations. In this section, we discuss the phenomenon that occurred in previous experiments and try to give explanations for them with extra experiments.

5.1 Visualization of Inferences

To analyze the modelling of options relations, we apply Integrated Gradients(IG) to our fine-tuned model and input samples for analyzing the contribution of tokens to the likelihood of options. This is accomplished by *Captum*[4] library based on *PyTorch* and we use its default visualization scheme. In our implementation, we select the embedding of $[PAD]$ token as baseline.

An example is shown in Figure 2. In this example, the total attribution score reveals the overall propensity of all components towards the corresponding option and the correct option, option D in this case, has the highest score. When we look into the word importance, which is the contribution of each token, we notice not only the tokens in the passage and question but also the tokens in other options are supplementing the result.

From the attribution of options A, B, and C, we can see that except for themselves, tokens in other options generally provide a negative contribution. As for correct option D, tokens in the passage, question, and other options would mostly contribute positively. This could be a piece of evidence for proving taking full context components could be beneficial to modelling option relevance where all options are contributing to the answer formation.

5.2 Investigation of The Fuser Network

We can expect the depth of the fuser network to have a certain influence on the result, while the relationship between depth and effect is uncertain. To further investigate its influence, we explore the relationship between its depth and impact on performances. The performance change on the COSMOS-QA dataset caused by its depth is shown in Figure 3.

As we can see from the result, there is no clear pattern in either accuracy or standard deviation for deepening the fuser network. This result is frustrating since it could not give much information about the setting of it and the depth of it would remain a hyperparameter which needs to be adjusted according to the dataset.

The only thing we can be sure of is that adding a fuser network can indeed improve the model performance under full context components input. From a practical perspective, adding a single-layer fuser network ensure improved performance.

5.3 Reduction of Memory Occupancy

In forward propagation, the memory occupancy of all intermediate computation results could be regarded as linear to the length of the input token sequence. For a sample with n options and the length of the passage, question, and each option is p_l, q_l, and o_l, we can have the total memory occupancy caused by this sample is linear to $n * (p_l + q_l + o_l)$ when the baseline method is applied.

[4] https://captum.ai/

Figure 2: Attribution for all options of a sample from COSMOS-QA. Tokens marked in green indicate positive and in red indicate opposite for the option. We use coloured boxes to indicate the attribution target option and their corresponding text.

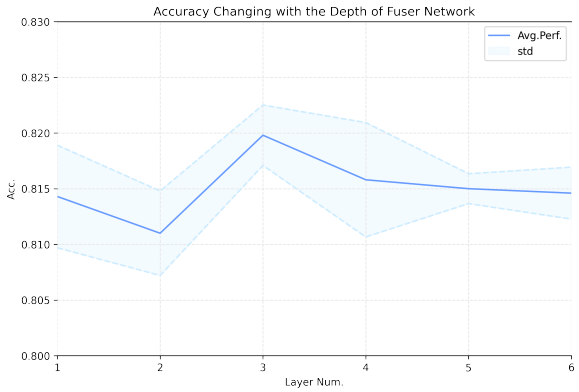

Figure 3: Accuracy on COSMOS-QA validation set with different depth of Fuser Network.

As for full context components input, the memory consumption would be reduced to linear to $p_l + q_l + n * o_l$. We roughly apply the average length of all components to this equation and we can have a cursory estimation of diminution: 74% for RACE, 68% for COSMOS-QA, 63%for DREAM, 72% for C^3, 74% for SweQUAD-MC, and 73% for MuSeRC.

This means that taking the full context components as input could significantly reduce memory occupancy, especially when the samples have long passage text. It would be meaningful for expanding the batch size with limited memory capacity while having long text input. As a result, with full context components input format, the RoBERTa-Large model could be trained on the COSMOS-QA dataset with batch size 16 on a single RTX3090 GPU without any other memory trick.

6 Conclusions

In this work, we analyze the data processing requirements in the mental health detection scenario, especially the adaptability in multiple inspections and evaluations. For the above analytical demands, we propose a solution in the form of an MMRC task and introduce new methods, including inserting extra tags and applying a transformer encoder network to fuse information of options set, for identifying components of inputs to enable the model to attentively recognize the inspections and corresponding measurement with explicit semantics.

As a result, our methods can be applied to MMRC tasks and improve over the performances of prior methods on datasets with similar characteristics to our target mental health detection data on pre-trained language models in most cases. It indicates that our methods help the models understand the relations among all the components of input, which can be regarded as beneficial to the modelling of option relevance.

Due to the absence of open-source mental health detection datasets in the expectation of our needs, all our experiments are conducted on available MMRC datasets. Although our results cannot be directly evaluated as valid for mental health detection, observations based on experimental phenomena in-

dicate that they can meet some of the requirements for such tasks.

Ethics Statement

Acknowledgements

References

Hayda Almeida, Antoine Briand, and Marie-Jean Meurs. 2017. Detecting early risk of depression from social media user-generated content. In *CLEF (working notes)*.

Sheshadri Iyengar Raghavan Bhagyashree, Kiran Nagaraj, Martin Prince, Caroline HD Fall, and Murali Krishna. 2018. Diagnosis of dementia by machine learning methods in epidemiological studies: a pilot exploratory study from south india. *Social psychiatry and psychiatric epidemiology*, 53(1):77–86.

M Hamilton. 1959. Hamilton anxiety scale. *Group*, 1(4):10–1037.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. *CoRR*, abs/1909.00277.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-Tür. 2019. MMM: multi-stage multi-task learning for multi-choice reading comprehension. *CoRR*, abs/1910.00458.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

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. *CoRR*, abs/1909.11942.

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. *CoRR*, abs/1907.11692.

Tom L Osborn, Christine M Wasanga, and David M Ndetei. 2022. Transforming mental health for all.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option comparison network for multiple-choice reading comprehension. *CoRR*, abs/1903.03033.

Jessica Ross, Thomas Neylan, Michael Weiner, Linda Chao, Kristin Samuelson, and Ida Sim. 2015. Towards constructing a new taxonomy for psychiatry using self-reported symptoms. In *MEDINFO 2015: eHealth-enabled Health*, pages 736–740. IOS Press.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *CoRR*, abs/1902.00164.

Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yantao Jia, and Zhicheng Sheng. 2020. Scene restoring for narrative machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3063–3073, Online. Association for Computational Linguistics.

Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. 2015. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and psychosomatics*, 84(3):167–176.

Tung Tran and Ramakanth Kavuluru. 2017. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. Dcmn+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9563–9570.

Yuxiang Zhang and Hayato Yamana. 2022. Hrca+: Advanced multiple-choice machine reading comprehension method. *LREC*.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Dual multi-head co-attention for multi-choice reading comprehension. *CoRR*, abs/2001.09415.

A Example Appendix

This is a section in the appendix.