

Answer Generation through Unified Memories over Multiple Passages

Makoto Nakatsuji, Sohei Okui

NTT Resonant Inc.

Granparktower, 3-4-1 Shibaura, Minato-ku, Tokyo 108-0023, Japan
nakatsuji.makoto@gmail.com, okui@nttr.co.jp

Abstract

Machine reading comprehension methods that generate answers by referring to multiple passages for a question have gained much attention in AI and NLP communities. The current methods, however, do not investigate the relationships among multiple passages in the answer generation process, even though topics correlated among the passages may be answer candidates. Our method, called neural answer Generation through Unified Memories over Multiple Passages (GUM-MP), solves this problem as follows. First, it determines which tokens in the passages are matched to the question. In particular, it investigates matches between tokens in positive passages, which are assigned to the question, and those in negative passages, which are not related to the question. Next, it determines which tokens in the passage are matched to other passages assigned to the same question and at the same time it investigates the topics in which they are matched. Finally, it encodes the token sequences with the above two matching results into unified memories in the passage encoders and learns the answer sequence by using an encoder-decoder with a multiple-pointer-generator mechanism. As a result, GUM-MP can generate answers by pointing to important tokens present across passages. Evaluations indicate that GUM-MP generates much more accurate results than the current models do.

1 Introduction

Machine Reading Comprehension (MRC) methods [Nguyen *et al.*, 2016; Rajpurkar *et al.*, 2016] that empower computers with the ability to read and comprehend knowledge and then answer questions from textual data have made rapid progress in recent years. Most methods try to answer a question by extracting exact text spans from the passages retrieved by search engines [Wang *et al.*, 2018], while a few try to generate answers by copying word tokens from passages in decoding answer tokens [Song *et al.*, 2018; Nishida *et al.*, 2019]. The passages are usually related to the question; thus, the descriptions among the passages are often related to each other. The current methods,

however, do not analyze the relationships among passages in the answer generation process. Thus, they may become confused by several different but related answer descriptions or unrelated descriptions in multiple passages assigned to the question. This lowers the accuracy of the generated answers [Jia and Liang, 2017; Wang *et al.*, 2018].

Table 1 lists examples from the MS-MARCO dataset [Nguyen *et al.*, 2016], which we used in our evaluation, to explain the problem. This table contains the question, passages prepared for this question, and the answer given by human editors. The phrases in bold font include the answer candidates to the question “what is the largest spider in the world?”. They are described across multiple passages, and some describe different spiders. There are also descriptions that are unrelated to the question, for example, about how humans feel about spiders or the characteristics of spiders. The presence of several different answer descriptions or unrelated ones tend to confuse the current answer-generation methods, and this lowers their accuracy.

To solve this problem, we propose neural answer Generation through Unified Memories over Multiple Passages (GUM-MP). This model has question and passage encoders, Multi-Perspective Memories (MPMs) [Song *et al.*, 2018], Unified Memories (UMs), and an answer decoder with a multiple-pointer-generator mechanism (see Fig. 1). It is founded upon two main ideas:

(1) GUM-MP learns which tokens in the passages are truly important for the question by utilizing positive passages that are prepared for the question and negative passages that are not related to the question. In the passage encoders, it receives a question and positive/negative passages on which it performs passage understanding by matching the question embedding with each token embedded in the passages from multiple perspectives. Then it encodes those information into MPMs. In particular, it investigates the difference between matches computed for positive passages and those for negative passages to determine the important tokens in the passages. This avoids confusion caused by descriptions that are not directly related to the question. For example, a phrase like “according to the guinness book of world records” in Table 1 can appear in passages that answer different questions (i.e. negative passages for the current question). GUM-MP can filter out the tokens in this phrase in generating an answer.

Question	What is the largest spider in the world?
Passage 1	Top 10 largest spiders in the world! Some people scare usual spiders to death, while some find these little pests pretty harmless and not disgusting at all. but there are some monsters that may give creeps even to the bravest and the most skeptical.
Passage 2	According to the guinness book of world records, the world’s largest spider is the goliath birdeater native to south america. Scientists say the world’s largest spider, the goliath birdeater, can grow to be the size of a puppy and have legs spanning up to a foot, according to video from geobeats.
Passage 3	The giant huntsman spider is a species of huntsman spider, a family of large, fast spiders that actively hunt down prey. It is considered the world’s largest spider by leg span, which can reach up to 1 foot (30 centimeters).
Answer	The giant huntsman is the largest spider in the world.

Table 1: Example entry in the MS-MARCO dataset.

(2) GUM-MP computes the match between each token embedding in MPM for each passage and the embedding composed from the rest of passages assigned to the question. First, it picks up the target passage. Next, it encodes the sequences of embeddings in the rest of passages into a fixed-dimensional latent semantic space, Passages Alignment Memory (PAM). PAM thus holds the semantically related or unrelated topics described in the passages together with the question context. Then, it computes the match between each embedding in the target MPM and the PAM. Finally, it encodes the embedding sequence in the target MPM with the above matching results into the UM. GUM-MP builds UMs for all passages by changing the target passage. As a result, for example, in Table 1, GUM-MP can distinguish the topics of large spiders, that of human feelings, and that of the spider characteristics by referring to UMs in generating an answer.

Finally, GUM-MP computes the vocabulary and attention distributions for multiple passages by applying encoder-decoder with a multiple-pointer-generator mechanism, wherein the ordinary pointer-generator mechanism [See *et al.*, 2017] is extended to handle multiple passages. As a result, GUM-MP can generate answers by pointing to different descriptions across multiple passages and comprehensively assess which tokens are important or not.

We used the MS-MARCO dataset and a community-QA dataset of a Japanese QA service, Oshiete goo, in our evaluations since they provide answers with multiple passages assigned to questions. The results show that GUM-MP outperforms existing state-of-the-art methods of answer generation.

2 Related Work

Most MRC methods aim to answer a question with exact text spans taken from evidence passages [Yu *et al.*, 2018; Rajpurkar *et al.*, 2016; Yang *et al.*, 2015; Joshi *et al.*, 2017]. Several studies on the MS-MARCO dataset [Nguyen *et al.*, 2016; Song *et al.*, 2018; Tan *et al.*, 2018] define the task as answering a question using information from multiple passages. Among them, S-Net [Tan *et al.*, 2018] developed an extraction-then-synthesis framework to synthesize answers from the extracted results. MPQG [Song *et al.*, 2018] performs question understanding by matching the question with a passage from multiple perspectives and encodes the matching results into MPM. It then generates answers by using an attention-based LSTM with a pointer-generator mechanism. However, it can not handle multiple passages for a question or investigate the relationships among passages when it

generates an answer. Several models based on Transformer [Vaswani *et al.*, 2017] or BERT [Devlin *et al.*, 2018] have recently been proposed in the MRC area [Nogueira *et al.*, 2019; Liu *et al.*, 2018; Shao *et al.*, 2019; Hu *et al.*, 2018]. In particular, [Nishida *et al.*, 2019] is for generating answers on the MS-MARCO dataset. These methods, however, do not utilize positive/negative passages to examine which word tokens are important or analyze relationships among passages, to improve their accuracy.

Regarding studies that compute the mutual attention among documents, [Hao *et al.*, 2017] examined cross attention between the question and the answer. Co-attention models [Xiong *et al.*, 2017; Zhong *et al.*, 2019] as well use co-dependent representations of the question and the passage in order to focus on relevant parts of both. They, however, do not compute the mutual attentions among passages and only focus on the attentions between question and passages. V-Net [Wang *et al.*, 2018] extracts text spans as answer candidates from passages and then verifies whether they are related or not from their content representations. It selects the answer from among the candidates, but does not generate answers.

The neural answer selection method [Tan *et al.*, 2016] achieves high selection accuracy by improving the matching strategy through the use of positive and negative answers for the questions, where the negative answers are randomly chosen from the entire answer space except for the positive answers. There are, however, no answer-generation methods that utilize negative answers to improve answer accuracy.

3 Preliminary

Here, we explain the encoding mechanism of MPM used in MPQG, since we base our ideas on its framework.

The model takes two components as input: a passage and a question. The passage is a sequence $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_i, \dots, \mathbf{p}_{N_p})$ of length N_p , and the question is a sequence $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_i, \dots, \mathbf{q}_{N_q})$ of length N_q . The model generates the output sequence $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{N_x})$ of length N_x word by word. Here, \mathbf{p}_i (or $\mathbf{q}_i, \mathbf{x}_i$) denotes a one-of- D embedding of the i -th word in a sequence \mathbf{P} (or \mathbf{Q}, \mathbf{X}) of length N_p (or N_q, N_x).

The model follows the encoder-decoder framework. The encoder matches each time step of the passage against all time steps of the question from multiple perspectives and encodes the matching result into the MPM. The decoder generates the output sequence one word at a time based on the MPM. It is almost the same as the normal pointer-generator mechanism [See *et al.*, 2017]; thus, we will omit its explanation.

MPQG uses a BiLSTM encoder, which encodes the question in both directions to better capture the overall meaning of the question. It processes in both directions, $\{\mathbf{q}_1, \dots, \mathbf{q}_{N_q}\}$ and $\{\mathbf{q}_{N_q}, \dots, \mathbf{q}_1\}$, sequentially. At time step t , the encoder updates the hidden state by $\mathbf{h}_t^q = [\mathbf{h}_t^{q,f}, \mathbf{h}_t^{q,b}]^T$, where $\mathbf{h}_t^{q,f} = f(\mathbf{q}_{t-1}, \mathbf{h}_{t-1}^{q,f})$ and $\mathbf{h}_t^{q,b} = f(\mathbf{q}_{t+1}, \mathbf{h}_{t+1}^{q,b})$. $f(\cdot)$ is an LSTM unit. $\mathbf{h}_t^{q,f}$ and $\mathbf{h}_t^{q,b}$ are hidden states output by the forward LSTM and backward LSTM, respectively. MPQG then applies a max-pooling layer to all hidden vectors yielded by the question sequence to extract the most salient signal for each word. As a result, it generates a fixed-sized distributed vector representation of the question, \mathbf{o}^q .

Next, MPQG computes the matching vector \mathbf{m}_i by using a function f_m to match two vectors, \mathbf{o}^q and each forward (or backward) hidden vector, $\mathbf{h}^{p,f}$ (or $\mathbf{h}^{p,b}$), output from the passage. In particular, it uses a multi-perspective cosine matching function defined as: $m_{i,z} = f_m(\mathbf{h}_i^{p,f}, \mathbf{o}^q; \mathbf{w}_z) = \cos(\mathbf{h}_i^{p,f} \circ \mathbf{w}_z, \mathbf{o}^q \circ \mathbf{w}_z)$, where the matrix $\mathbf{W} \in \mathbb{R}^{Z \times D}$ is a learnable parameter of a multi-perspective weight, Z is the number of perspectives, the z -th row vector $\mathbf{w}_z \in \mathbf{W}$ represents the weighting vector associated with the z -th perspective, and \circ is the element-wise multiplication operation. The final matching vector for each time step of the passage is updated by concatenating the matching results of the forward and backward operations. MPQG employs another BiLSTM layer on top of the matching layer to smooth the matching results. Finally, it concatenates the hidden vector of the passage \mathbf{h}_i^p and matching vector \mathbf{m}_i to form the hidden vector $\mathbf{h}_i \in \mathbf{H}$ (the length is $2(D+Z)$) in the MPM of the passage, which contains both the passage information and matching information. The MPM for the question is encoded in the same way.

4 Model

Given an input question sequence \mathbf{Q} and K passage sequences (e.g., the k -th passage is denoted as $\mathbf{P}^k = \{\mathbf{p}_1^k, \dots, \mathbf{p}_i^k, \dots, \mathbf{p}_{N_k}^k\}$), GUM-MP outputs an answer sequence $\mathbf{A} = \{\mathbf{a}_1, \dots, \mathbf{a}_t, \dots, \mathbf{a}_{N_a}\}$. Please see Fig. 1 also.

4.1 Passage and Question Encoders

The passage encoder takes a set of passage sequences $\{\mathbf{P}^1, \dots, \mathbf{P}^k, \dots, \mathbf{P}^K\}$ and a question sequence \mathbf{Q} as inputs and encodes the MPMs of the passages (e.g. the MPM of the k -th passage is denoted as $\mathbf{H}^k = \{\mathbf{h}_1^k, \dots, \mathbf{h}_i^k, \dots, \mathbf{h}_{N_k}^k\}$). Word embeddings in inputs are concatenations of Glove embeddings [Pennington *et al.*, 2014] with Bert ones. The computation of \mathbf{H}^k follows the MPQG approach explained in the previous section; however, GUM-MP improves on MPQG by computing the match in more detail by introducing a matching tensor \mathcal{M}^k for passage k . Let us explain how to compute the matching tensor. Each entry in \mathcal{M}^k stores the matching score $m_{i,j,z}^k$ in the z -th perspective between the question vector \mathbf{o}^q and the i -th hidden vector \mathbf{h}_i^k (we regard passages prepared to the question to be positive passages) and the j -th hidden vector $\mathbf{h}_j^{k,-}$ for negative passages (there is one negative passage for each positive one randomly chosen from passages, which are not assigned to the current question, in the

training dataset). It is computed as follows:

$$m_{i,j,z}^k = f_m(\mathbf{h}_i^k, \mathbf{o}^q; \mathbf{w}_z) - f_m(\mathbf{h}_j^{k,-}, \mathbf{o}^q; \mathbf{w}_z). \quad (1)$$

GUM-MP then computes the i -th Z -dimensional matching vector \mathbf{m}_i^k in \mathbf{H}^k as follows:

$$\mathbf{m}_i^k = \sum_{j=1}^{N_k^-} (\alpha_{i,j}^k \mathbf{m}_{i,j}^k) \quad \text{s.t.}$$

$$\alpha_{i,j}^k = \frac{\exp(e_{i,j}^k)}{\sum_{j=1}^{N_k^-} \exp(e_{i,j}^k)}, \quad e_{i,j}^k = \mathbf{w}^+ \cdot \mathbf{o}^k + \mathbf{w}^- \cdot \mathbf{o}^{k,-} + \mathbf{w}^m \cdot \mathbf{m}_{i,j}^k$$

where \mathbf{w}^+ , \mathbf{w}^- , and \mathbf{w}^m are learnable parameters. N_k^- denotes the length of the negative passages prepared for the k -th passage. $\mathbf{m}_{i,j}^k$ is a Z -dimensional matching vector whose elements are multi-perspective matching scores with the i -th token in the positive passage k and the j -th token in the negative passage (see Eq. (1)). \mathbf{o}^k and $\mathbf{o}^{k,-}$ are positive and negative passage vectors, respectively, computed in the same way as \mathbf{o}^q . The computed matching vector \mathbf{m}_i^k considers the margin difference between positive and negative passages for the i -th token in the k -th passage and thus is more concise than the matching vector yielded by MPQG.

The question encoder also encodes the MPM of the question, $\mathbf{H}^q = \{\mathbf{h}_1^q, \dots, \mathbf{h}_i^q, \dots, \mathbf{h}_{N_q}^q\}$. The computation of \mathbf{H}^q follows the computation of \mathbf{H}^k except that it switches the roles of the question and passages. One difference is that there are K $\mathbf{H}^{q,k}$ s, since there are K passages. GUM-MP averages those $\mathbf{H}^{q,k}$ s to compute a single MPM for the question, \mathbf{H}^q and thereby reduce the complexity of the computation.

4.2 Unified Memories

GUM-MP computes the correlations between each hidden vector in the passage and the embedding of the rest of passages. This is because correlated topics among passages tend to include important topics for the question and thus may be possible answers.

First, GUM-MP picks up the i -th MPM. Next, it encodes the sequences of hidden vectors in the rest of MPMs (i.e. $\{\mathbf{H}^1, \dots, \mathbf{H}^{i-1}, \mathbf{H}^{i+1}, \dots, \mathbf{H}^K\}$, whose size is $\sum_{k:k \neq i}^K N_k \times 2(D+Z)$), into $2(D+Z) \times L$ latent semantics, i.e. Passage Alignment Memory (PAM) for the i -th passage (we denote this as \mathbf{PA}^i). We say that this memory is ‘‘aligned’’ since the hidden vectors in the rest of the MPMs are aligned through the shared weighting matrix \mathbf{W}^p (the size is $\sum_{k:k \neq i}^K N_k \times L$) as:

$$\mathbf{PA}^i = \{\mathbf{H}^1, \dots, \mathbf{H}^{i-1}, \mathbf{H}^{i+1}, \dots, \mathbf{H}^K\}^T \cdot \mathbf{W}^p.$$

Then, GUM-MP computes the UM that unifies the information of the token embeddings with the matching results from the question context as well as those about the topics described among the passages. It computes the j -th hidden vector in the i -th UM, \mathbf{u}_j^i , by concatenating the j -th hidden vector in the i -th MPM, \mathbf{h}_j^i , with the inner product of \mathbf{h}_j^i and \mathbf{PA}^i as: $\mathbf{u}_j^i = [\mathbf{h}_j^i, (\mathbf{h}_j^i \cdot \mathbf{PA}^i)]$. The length is $2(D+Z) + L$. Here, the inner product of \mathbf{h}_j^i and \mathbf{PA}^i includes information

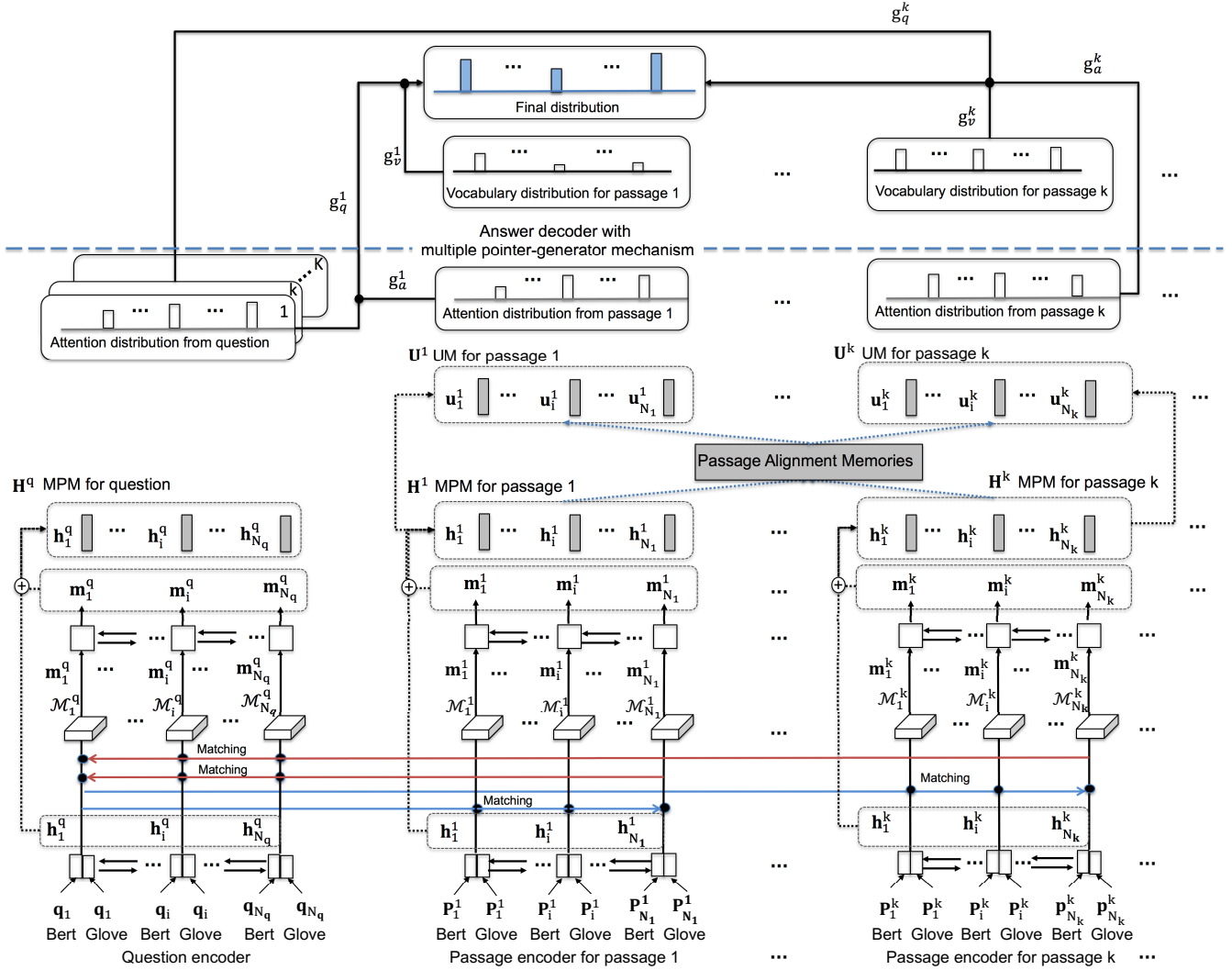


Figure 1: Overview of GUM-MP.

on which tokens in the passage are matched to the other passages assigned to the same question.

Thus, GUM-MP can point the important tokens considering the correlated topics among passages stored in UM in the decoder, which we will describe next.

4.3 Decoder

The decoder is based on attention-based LSTMs with a multiple-pointer-generator mechanism.

Our multiple-pointer-generator mechanism generates the vocabulary distribution, attention distribution for the question, and attention distribution for the passage independently for each passage. It then aligns these distributions across multiple passages (see Eq. (2) described later). This is different from the approach that first concatenates the passages into a single merged passage and then applies the pointer-generator mechanism to the merged passage [Nishida *et al.*, 2019].

GUM-MP requires six different inputs for generating the t -th answer word a_t : (1) K UMs for passages (e.g., the

k -th UM is denoted as $\mathbf{U}^k = \{\mathbf{u}_1^k, \dots, \mathbf{u}_i^k, \dots, \mathbf{u}_{N_k}^k\}$, where each vector $\mathbf{u}_i^k \in \mathbf{U}^k$ is aligned with the i -th word in the k -th passage); (2) the MPM for the question, $\mathbf{H}^q = \{\mathbf{h}_1^q, \dots, \mathbf{h}_i^q, \dots, \mathbf{h}_{N_q}^q\}$, where each vector $\mathbf{h}_i^q \in \mathbf{H}^q$ is aligned with the i -th word in the question; (3) K previous hidden states of the LSTM model, \mathbf{s}_{t-1}^k ; (4) the embedding of the previously generated word, \mathbf{a}_{t-1} ; (5) K previous context vectors, e.g. \mathbf{c}_{t-1}^k , which are computed using the attention mechanism, with \mathbf{U}^k being the attentional memory; (6) K previous context vectors, $\mathbf{c}_{t-1}^{q,k}$, which are computed using the attention mechanism, with \mathbf{H}^q being the attentional memory. At $t = 1$, we initialize \mathbf{s}_0 , \mathbf{c}_0^k , and $\mathbf{c}_0^{q,k}$ as zero vectors and set \mathbf{a}_0 to be the embedding of the token “<S>”.

For each time step t and each passage k , the decoder first feeds the concatenation of the previous word embedding, \mathbf{a}_{t-1} , and context vectors, \mathbf{c}_{t-1}^k and $\mathbf{c}_{t-1}^{q,k}$, into the LSTM model to update the hidden state: $\mathbf{s}_t^k = f(\mathbf{s}_{t-1}^k, [\mathbf{a}_{t-1}, \mathbf{c}_{t-1}^k, \mathbf{c}_{t-1}^{q,k}])$.

Metric	w/o Neg	w/o UM	UM(10)	UM(30)	UM(50)
BLEU-1	0.491	0.484	0.503	0.501	0.514
ROUGE-L	0.557	0.544	0.569	0.568	0.563

Table 2: Ablation study of GUM-MP (MS-MARCO).

Metric	Trans	MPQG	S-Net	V-Net	GUM-MP
BLEU-1	0.060	0.342	0.364	0.407	0.503
ROUGE-L	0.062	0.451	0.383	0.405	0.569

Table 3: Performance comparison (MS-MARCO).

Next, the new context vectors, i.e., \mathbf{c}_t^k and $\mathbf{c}_t^{k,q}$, the attention distribution $\alpha_{t,i}^k$ for each time step for the k -th passage, and the attention distribution $\alpha_{t,i}^{k,q}$ for each time step for the question with the k -th passage are computed as follows:

$$\mathbf{c}_t^k = \sum_{i=1}^{N_k} \alpha_{t,i}^k \mathbf{u}_i^k, \quad \mathbf{c}_t^{k,q} = \sum_{i=1}^{N_q} \alpha_{t,i}^{k,q} \mathbf{h}_i^q, \quad \text{s.t.}$$

$$\alpha_{t,i}^k = \frac{\exp(e_{t,i}^k)}{\sum_{i=1}^{N_k} \exp(e_{t,i}^k)}, \quad e_{t,i}^k = \tanh(\mathbf{w}_h^k \mathbf{u}_i^k + \mathbf{w}_s^k \mathbf{s}_t^k + b_e^k),$$

$$\alpha_{t,i}^{k,q} = \frac{\exp(e_{t,i}^{k,q})}{\sum_{i=1}^{N_q} \exp(e_{t,i}^{k,q})}, \quad e_{t,i}^{k,q} = \tanh(\mathbf{w}_h^{k,q} \mathbf{h}_i^q + \mathbf{w}_s^{k,q} \mathbf{s}_t^k + b_e^{k,q}).$$

$\mathbf{w}_h^k, \mathbf{w}_s^k, \mathbf{w}_h^{k,q}, \mathbf{w}_s^{k,q}, b_e^k$ and $b_e^{k,q}$ are learnable parameters.

Then, the output probability distribution over the vocabulary of words in the current state is computed for passage k :

$$\mathbf{V}_{vocab}^k = \text{softmax}(\mathbf{W}^k [\mathbf{s}_t^k, \mathbf{c}_t^k, \mathbf{c}_t^{q,k}] + \mathbf{b}_v^k).$$

\mathbf{W}^k and \mathbf{b}_v^k are learnable parameters. The number of rows in \mathbf{W}^k represents the number of words in the vocabulary.

GUM-MP then utilizes the multiple-pointer-generator mechanism to compute the final vocabulary distribution to determine the t -th answer word a_t . It first computes the vocabulary distribution computed for passage k by interpolation between three probability distributions, $\mathbf{V}_{vocab}^k, \mathbf{P}^k_{attn}$, and \mathbf{Q}^k_{attn} . Here, \mathbf{P}^k_{attn} and \mathbf{Q}^k_{attn} are computed on the basis of the attention distributions $\alpha_{t,i}^k$ and $\alpha_{t,i}^{k,q}$. It then integrates the vocabulary distributions computed for each passage to compute the final vocabulary distribution as follows:

$$\mathbf{V}_{final} = \sum_k (g_v^k \mathbf{V}_{vocab}^k + g_a^k \mathbf{P}^k_{attn} + g_q^k \mathbf{Q}^k_{attn}) \text{ s.t. } (2)$$

$$g_v^k, g_a^k, g_q^k = \text{softmax}(\mathbf{W}^g [\mathbf{s}_t^k, \mathbf{c}_t^k, \mathbf{c}_t^{q,k}] + \mathbf{b}_g)$$

where \mathbf{W}^g and \mathbf{b}_g are learnable parameters.

Our multiple-pointer-generator naively checks the distributions generated for each passage. With the UM, it determines which tokens in the passages are important or not for answer generation. This improves the generation accuracy.

4.4 Training

GUM-MP trains the model by optimizing the log-likelihood of the gold-standard output sequence \mathbf{A} with the cross-entropy loss (θ represents the trainable model parameters):

Metric	w/o Neg	w/o UM	UM(5)	UM(10)	UM(30)
BLEU-1	0.125	0.129	0.309	0.283	0.321
ROUGE-L	0.224	0.222	0.253	0.248	0.265

Table 4: Ablation study of GUM-MP (Oshiete-goo).

Metric	Trans	MPQG	S-Net	V-Net	GUM-MP
BLEU-1	0.041	0.232	0.247	0.246	0.321
ROUGE-L	0.088	0.251	0.244	0.249	0.265

Table 5: Performance comparison (Oshiete-goo).

$$\mathcal{L} = - \sum_{t=1}^{N_a} \log p(\mathbf{a}_t | \mathbf{Q}, \mathbf{a}_1, \dots, \mathbf{a}_{t-1}; \theta).$$

5 Evaluation

This section evaluates GUM-MP in detail.

5.1 Compared Methods

We compared the performance of the following five methods: (1) *Trans* is a Transformer [Vaswani *et al.*, 2017] that is used for answer generation. It receives questions, not passages, as input; (2) *S-Net* [Tan *et al.*, 2018], (3) *MPQG* [Song *et al.*, 2018], and (4) *V-Net* [Wang *et al.*, 2018]: these are explained in Section 2, though we applied our multiple-pointer-generator mechanism to *MPQG* to make it handle multiple passages. (5) *GUM-MP* is our proposal.

5.2 Datasets

We used the following two datasets:

MS-MARCO The questions are user queries issued to the Bing search engine, and approximately ten passages and one answer are assigned to each question. Among the official datasets in the MS-MARCO project, we chose the natural language generation dataset, since our focus is answer generation rather than answer extraction from passages. Human editors in the MS-MARCO project reviewed the answers to the questions and rewrote them as well-formed ones so that the answers would make sense even without the context of the question or the retrieved passages. We pre-trained a Glove model and also fine-tuned a publicly available Bert-based model [Devlin *et al.*, 2018] by using this dataset. We then randomly extracted one-tenth of the full dataset provided by the MS-MARCO project. The training set contained 16,500 questions and the test set contained 2,500 questions. The questions, passages, and answers had on average 6, 68, and 16 words, respectively.

Oshiete-goo This dataset focused on the ‘‘love advice’’ category of the Japanese QA community, Oshiete-goo [Nakatsuji, 2019; Nakatsuji and Okui, 2020]. It has 771,956 answer documents to 189,511 questions. We pre-trained the word embeddings by using a Glove model on this dataset. We did not use the Bert model since it does not improve the accuracy much. Then, human editors abstractly summarized 24,661 answer documents assigned to 5,202 questions into 10,032 summarized answers¹. Since the topics in several of

¹This summarized answer dataset is used in the actual AI love advice service : ‘‘https://oshiete.goo.ne.jp/ai’’.

Question	Largest lake of USA?
Passage 1	The largest lake (by surface area) in the United States is lake michigan with an area of 45410 square miles.
Passage 2	Iliamna lake is the largest lake in alaska and the second largest freshwater lake contained wholly within the United States (after lake michigan).
Passage 3	Superior is the largest lake that's partly in the United States at 31,820 square miles. The largest lake entirely contained in the United States is lake michigan , 22,400 square miles.
<i>MPQG</i>	Lake is the largest lake of lake.
<i>V-Net</i>	Is the largest lake that's partly in the United States at 31,820 square miles.
<i>GUM-MP</i>	Lake michigan is the largest lake of the United States.
Answer	The largest lake of United States of America is lake michigan.

Table 6: Example of answers output by the methods.

the answer documents assigned to the question overlap, the number of summarized answers is smaller than the number of original answer documents. Then, we randomly chose one-tenth of the questions as the test dataset. The rest was used as the training dataset. The questions, answer documents (hereafter, we call them passages), and summarized answers (hereafter, we call them answers) had on average 270, 208, and 48 words, respectively. There were 58,955 word tokens and on average 4.72 passages for a question.

5.3 Methodology and Parameter Setup

To measure performance, we used BLEU-1 [Papineni *et al.*, 2002] and ROUGE-L [Lin, 2004], which are useful for measuring the fluency of generated texts. For example, MS-MARCO project officially uses those metrics in their evaluations.

In testing the model, for both datasets, the negative passages were selected from passages that were not assigned to the current question. In the Oshiete-goo dataset, there were no passages assigned to newly submitted questions. Thus, we used the answer selection method [Tan *et al.*, 2016] to learn passage selection using the training dataset. Then, we selected three positive passages per question from among all the passages in the training dataset in the Oshiete-goo dataset. This setting is natural for many applications.

We set the word embedding size to 300 and the batch size to 32. The decoder vocabulary was restricted to 5,000 according to the frequency for the MS-MARCO dataset. We did not restrict the attention vocabularies. The decoder vocabulary was not restricted for the Oshiete-goo dataset. Each question, passage, and answer were truncated to 50, 130, and 50 words for the MS-MARCO dataset (300, 300, and 50 words for the Oshiete-goo one). The epoch count was 30, the learning rate was 0.0005, Z in MPM was 5, and the beam size was 20.

5.4 Results

Table 2 and Table 4 summarize the ablation study for the MS-MARCO dataset and for Oshiete-goo dataset. They compare several methods, each of which lacks one function of *GUM-MP*; *w/o Neg* lacks the “matching tensor” in the question and passage encoders. *w/o UM* lacks the “Unified Memories”.

UM(L) is our method *GUM-MP* with L , the length of row in PAM in Section 4.2.

The results indicate that all of the above functions are useful for improving the accuracy of the generated answers for both datasets: *GUM-MP* is better than *w/o Neg*. This means that the matching tensor is useful for extracting good tokens from the answer or question passages when generating the answers. *GUM-MP* is also superior to *w/o UM*. This is because *GUM-MP* utilizes the UMs that include the analysis of topics described in the passages for generating an answer. Especially, for the Oshiete-goo dataset, *GUM-MP* is much better than *w/o UM* regardless of the size of L . This is because this dataset focuses on “love advice” whereas the MS-MARCO dataset has very diverse QA pairs; thus, the PAM well aligns the topics described in multiple passages for the question in the Oshiete-goo dataset.

Table 3 and Table 5 summarize the results of all of the compared methods on the MS-MARCO dataset and on the Oshiete-goo dataset. Here, we present *UM(10)* as *GUM-MP* in the MS-MARCO dataset and *UM(30)* as *GUM-MP* in the Oshiete-goo dataset since they are the best accurate results when changing L . *S-Net* and *MPQG* generated better results than *Trans* on both datasets, since they can make use of passage information to extract tokens important for making the answers, while *Trans* cannot. Looking at the generated answers, it is clear that *MPQG* tends to extract tokens from multiple passages while generating common phrases from the training vocabulary. In contrast, *S-Net* tends to extract whole sentences from the passages. *V-Net* tends to select good answer candidate spans from the passages; however, it fails to generate answers that are edited or changed from the sentences in the passages. Finally, *GUM-MP* is superior to *V-Net* on both datasets, as it can identify the important word tokens and at the same time avoid including redundant or noisy phrases across the passages in the generated answers.

5.5 Meaningful Results

Table 6 presents examples of answers output by *MPQG*, *V-Net*, and *GUM-MP* on the MS-MARCO dataset. *MPQG* mistakenly extracted a token since it could not consider the correlated topics such as “lake michigan” within the passages. *V-Net* extracted the exact text span taken from a passage that includes redundant information for the question. *GUM-MP* points to the tokens (e.g. lake michigan) that are described across multiple passages that match the current question. As a result, it accurately generates the answer.

6 Conclusion

We proposed the neural answer Generation model through Unified Memories over Multiple Passages (*GUM-MP*). *GUM-MP* uses positive-negative passage analysis for passage understanding following the question context. It also performs inter-relationship analysis among multiple passages. It thus can identify which tokens in the passages are truly important for generating an answer. Evaluations showed that *GUM-MP* is consistently superior to state-of-the-art answer generation methods. We will apply our ideas to a Transformer-based encoder-decoder model.

References

- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [Hao *et al.*, 2017] Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proc. ACL'17*, pages 221–231, 2017.
- [Hu *et al.*, 2018] Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Ming Zhou. Read + verify: Machine reading comprehension with unanswerable questions. *CoRR*, abs/1808.05759, 2018.
- [Jia and Liang, 2017] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proc. EMNLP'17*, pages 2021–2031, 2017.
- [Joshi *et al.*, 2017] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551, 2017.
- [Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: In: Proc. ACL-04 Workshop*, pages 74–81, 2004.
- [Liu *et al.*, 2018] Xiaodong Liu, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for natural language inference. *CoRR*, abs/1804.07888, 2018.
- [Nakatsuji and Okui, 2020] Makoto Nakatsuji and Sohei Okui. Conclusion-supplement answer generation for non-factoid questions. In *Proc. AAAI'20*, 2020.
- [Nakatsuji, 2019] Makoto Nakatsuji. Can ai generate love advice?: Toward neural answer generation for non-factoid questions. *arXiv, cs.CL*, 1912.10163, 2019.
- [Nguyen *et al.*, 2016] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with NIPS 2016*, 2016.
- [Nishida *et al.*, 2019] Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. Multi-style generative reading comprehension. *CoRR*, abs/1901.02262, 2019.
- [Nogueira *et al.*, 2019] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *CoRR*, abs/1904.08375, 2019.
- [Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL'02*, pages 311–318, 2002.
- [Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proc. EMNLP'14*, 2014.
- [Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.
- [See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proc. ACL'17*, pages 1073–1083, 2017.
- [Shao *et al.*, 2019] Taihua Shao, Yupu Guo, Zepeng Hao, and Honghui Chen. Transformer-based neural network for answer selection in question answering. *IEEE Access*, PP:1–1, 02 2019.
- [Song *et al.*, 2018] Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. Leveraging context information for natural question generation. In *Proc. NAACL-HLT'18*, pages 569–574, 2018.
- [Tan *et al.*, 2016] Ming Tan, Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. Improved representation learning for question answer matching. In *Proc. ACL'16*, pages 464–473, 2016.
- [Tan *et al.*, 2018] Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. S-net: From answer extraction to answer synthesis for machine reading comprehension. In *Proc. AAAI'18*, pages 5940–5947, 2018.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS'17*, pages 5998–6008, 2017.
- [Wang *et al.*, 2018] Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proc. ACL'18*, pages 1918–1927, 2018.
- [Xiong *et al.*, 2017] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *Proc. ICLR'17*, 2017.
- [Yang *et al.*, 2015] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proc. EMNLP'15*, pages 2013–2018, 2015.
- [Yu *et al.*, 2018] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proc. ICLR'18*, 2018.
- [Zhong *et al.*, 2019] Victor Zhong, Caiming Xiong, Nitish Keskar, and Richard Socher. Coarse-grain fine-grain coattention network for multi-evidence question answering. In *Proc. ICLR'19*, 2019.