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.
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 ).

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

<table>
<thead>
<tr>
<th>Question</th>
<th>What is the largest spider in the world?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage 1</td>
<td>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.</td>
</tr>
<tr>
<td>Passage 2</td>
<td>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.</td>
</tr>
<tr>
<td>Passage 3</td>
<td>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 ).</td>
</tr>
<tr>
<td>Answer</td>
<td>The giant huntsman is the largest spider in the world.</td>
</tr>
</tbody>
</table>

(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-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 \( P = (p_1, \ldots, p_N_p) \) of length \( N_p \), and the question is a sequence \( Q = (q_1, \ldots, q_N_q) \) of length \( N_q \). The model generates the output sequence \( X = (x_1, \ldots, x_N_x) \) of length \( N_x \) word by word. Here, \( p_i \) (or \( q_i, x_i \)) denotes a one-of-\( D \) embedding of the \( i \)-th word in a sequence \( P \) (or \( Q, X \)) 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, \{q_1, \cdots, q_{N_q}\} and \{q_{N_q+1}, \cdots, q_1\}, sequentially. At time step \(t\), the encoder updates the hidden state by \(h_i^q = [h_i^{q,f}, h_i^{q,b}]^T\), where \(h_i^{q,f} = f(q_{t-1}, h_{i-1}^{q,f})\) and \(h_i^{q,b} = f(q_{t+1}, h_{i+1}^{q,b})\). 

4.1 Passage and Question Encoders

\[ m_{i,j}^k = f_m(h_i^k, o^q; w_e) - f_m(h_{i-1}^k, o^q; w_e). \]  

(1)

GUM-MP then computes the \(i\)-th \(Z\)-dimensional matching vector \(m_i^k\) in \(H^k\) as follows:

\[ m_i^k = \sum_{j=1}^{N_p} (\alpha_{i,j}^k m_{i,j}) \text{ s.t.} \]

\[ \alpha_{i,j}^k = \frac{\exp(e_{i,j}^k)}{\sum_{j=1}^{N_p} \exp(e_{i,j}^k)}, \]

where \(w^+, w^-, \) and \(w^m\) are learnable parameters. \(N_p^-\) denotes the length of the negative passages prepared for the \(k\)-th passage. \(m_{i,j}^k\) 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)). \(o^q\) and \(o^-\) are positive and negative passage vectors, respectively, computed in the same way as \(o^q\). The computed matching vector \(m_i^k\) 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, \(H^q = \{h_1^q, \cdots, h_{N_q}^q\}\). The computation of \(H^q\) follows the computation of \(H^k\) except that it switches the roles of the question and passages. One difference is that there are \(K H^k\) for \(K\) passages. GUM-MP averages those \(H^q\)\'s to compute a single MPM for the question, \(H^q\) 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. \(\{H_1, \cdots, H_i^{-1}, H_i^{+1}, \cdots, H_K\}\), whose size is \(\sum_{k: k \neq i} N_k \times 2(D + Z)\)), into \(2(D + Z) \times L\) latent semantics, i.e. Passage Alignment Memory (PAM) for the \(i\)-th passage (we denote this as \(P_{ai}\)). We say that this memory is “aligned” since the hidden vectors in the rest of the MPMs are aligned through the shared weighting matrix \(W^p\) (the size is \(\sum_{K} K N_k \times L)\) as:

\[ P_{ai} = \{H_1, \cdots, H_i^{-1}, H_i^{+1}, \cdots, H_K\}^T \cdot W^p. \]

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, \(u_{ij}\), by concatenating the \(j\)-th hidden vector in the \(i\)-th MPM, \(h_{ij}\), with the inner product of \(h_{ij}\) and \(P_{ai}\) as:

\[ u_{ij} = [h_{ij}^q, (h_{ij}^q \cdot P_{ai})]. \]

The length is \(2(D + Z) + L\). Here, the inner product of \(h_{ij}\) and \(P_{ai}\) includes information.
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) $K$ UM$s$ for passages (e.g., the $k$-th UM is denoted as $U^k = \{u^k_1, \ldots, u^k_{N_k}\}$, where each vector $u^k_i \in U^k$ is aligned with the $i$-th word in the $k$-th passage); (2) the MPM for the question, $H^q = \{h^q_1, \ldots, h^q_{N_q}\}$, where each vector $h^q_i \in H^q$ is aligned with the $i$-th word in the question; (3) $K$ previous hidden states of the LSTM model, $s^k_{t-1}$; (4) the embedding of the previously generated word, $a_{t-1}$; (5) $K$ previous context vectors, e.g. $c^q_{t-1}$, which are computed using the attention mechanism, with $U^k$ being the attentional memory; (6) $K$ previous context vectors, $c^q_{t-1}$, which are computed using the attention mechanism, with $H^q$ being the attentional memory. At $t = 1$, we initialize $s_0$, $c^q_0$, and $c^{q,k}_0$ as zero vectors and set $a_0$ 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, $a_{t-1}$, and context vectors, $c^k_{t-1}$ and $c^{q,k}_{t-1}$, into the LSTM model to update the hidden state: $s^k_t = f(s^k_{t-1}, [a_{t-1}, c^k_{t-1}, c^{q,k}_{t-1}])$. 

![Figure 1: Overview of GUM-MP.](image-url)
Next, the new context vectors, i.e., $c_i^k$ and $c_{i}^{k,q}$, the attention distribution $\alpha_{t,i}^{k}$ for each step for the $k$-th passage, and the attention distribution $\alpha_{t,i}^{k,q}$ for each step for the question with the $k$-th passage are computed as follows:

$$e_{t,i}^k = \sum_{i=1}^{N_k} \alpha_{t,i}^k u_i^k, \quad e_{t,i}^{k,q} = \sum_{i=1}^{N_k} \alpha_{t,i}^{k,q} h_i^q,$$

$$\alpha_{t,i}^{k} = \frac{\exp(e_{t,i}^k)}{\sum_{i=1}^{N_k} \exp(e_{t,i}^k)}, \quad e_{t,i}^{k,q} = \tanh(w_h^k u_i^k + w_s^k s_i^k + b_c^k),$$

where $w_h^k, w_s^k, w_{h}^{k,q}, w_{s}^{k,q}, b_c^k$ and $b_{c}^{k,q}$ are learnable parameters.

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

$$V_{vocabs}^k = \text{softmax}(W^k[s_i^t, c_i^t, c_{i}^{q,k}] + b_c^k).$$

$W^k$ and $b_c^k$ are learnable parameters. The number of rows in $W^k$ represents the number of words in $W^k$.

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, $V_{vocabs}^k$, $P_{attn}^k$, and $Q_{attn}^k$. Here, $P_{attn}^k$ and $Q_{attn}^k$ 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:

$$V_{final} = \sum_k g_k^k V_{vocabs}^k + g_{a}^k P_{attn}^k + g_{q}^k Q_{attn}^k \text{ s.t. (2)}$$

where $g_k^k, g_a^k, g_q^k$ are softmax($W^q[s_i^t, c_i^t, c_{i}^{q,k}] + b_q$) and $W^q$ and $b_q$ 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 $A$ with the cross-entropy loss ($\theta$ represents the trainable model parameters):
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


