Unsupervised Domain Adaptation on Question-Answering System with Conversation Data

Amalia Istiqlali Adiba **Takeshi Homma** **Yasuhiro Sogawa**
Hitachi, Ltd.
Kokubunji, Tokyo, Japan
{amalia.adiba.dw, takeshi.homma.ps,
yasuhiro.sogawa.tp}@hitachi.com

Abstract

Machine reading comprehension (MRC) is a task for question answering that finds answers to questions from documents of knowledge. Most studies on the domain adaptation of MRC require documents describing knowledge of the target domain. However, it is sometimes difficult to prepare such documents. The goal of this study was to transfer an MRC model to another domain without documents in an unsupervised manner. Therefore, unlike previous studies, we propose a domain-adaptation framework of MRC under the assumption that the only available data in the target domain are human conversations between a *user* asking questions and an *expert* answering the questions. The framework consists of three processes: (1) training an MRC model on the source domain, (2) converting conversations into documents using document generation (DG), a task we developed for retrieving important information from several human conversations and converting it to an abstractive document text, and (3) transferring the MRC model to the target domain with unsupervised domain adaptation. To the best of our knowledge, our research is the first to use conversation data to train MRC models in an unsupervised manner. We show that the MRC model successfully obtains question-answering ability from conversations in the target domain.

1 Introduction

Conversation agents such as Siri, as well as search engines, such as Google, have been increasing the scope of user questions in which they can provide direct answers to questions that can be extracted from web pages. Providing answers directly from a structured text is often referred to as machine reading comprehension (MRC). Benefiting from deep learning technology, MRC is a question-answering (QA) task that has been extensively studied (Hermann et al., 2015; Qiu et al., 2019). MRC is used to find an answer position in a document to answer a given question. A number of large corpora have played a critical role in advancing MRC research (Rajpurkar et al., 2016; Trischler et al., 2017; Bajaj et al., 2016; Zhang et al., 2018). Many MRC studies have focused on developing new model structures by introducing a new end-to-end neural network model to obtain state-of-the-art performance (Huang et al., 2018, 2019; Shen et al., 2017; Seo et al., 2017; Xiong et al., 2017). However, these state-of-the-art models were evaluated in one domain. In fact, it has been proven that the generalization capabilities of MRC models do not perform well on different datasets (Yogatama et al., 2019).

Unsupervised domain adaptation is an approach to cope with transferring knowledge from a source domain to a different unlabeled target domain (Pan and Yang, 2010). To provide labels for a new domain dataset, question generation is commonly used to create synthetic data consisting of question-answer pairs from documents of the target domain (Rus et al., 2010), so that an MRC model can be trained with both data from the source domain and syntactic data from the target domain (Yue et al., 2021; Lee et al., 2020; Puri et al., 2020; Shakeri et al., 2020; Cao et al., 2020; Wang et al., 2019).

One critical issue in unsupervised domain adaptation for MRC is that previous studies assumed that the input documents must be available in the target domain. There are also many types of information (not limited to the document) in a real-word scenario. To apply MRC to such information, it is necessary to convert the information into documents. However, this conversion is not an easy task.

Let us take a case in customer service support. Human operators in a customer service usually refer to "manual documents" containing necessary knowledge to answer customer questions. The manual usually has limited information. Thus, when there is no information to answer questions, the operators pass the call to a supervisor, and the su-

Proceedings of the SIGdial 2022 Conference, pages 432–441
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

pervisor continues to talk with the customer to answer the question. This procedure is called "escalation". The frequency of escalation is not trivial. Moreover, the number of supervisors is usually few, thus it is necessary to reduce escalations. It is obvious that conversations between supervisors and customers have plenty of information that is not included in the document. If we add new information to the document based on supervisor-customer conversations, the MRC task can answer more varied questions.

In domain adaptation for MRC, in which the conversation between an *expert* and *user* is the only available data in the target domain, has become a new challenge. The user asks questions and the expert answers the questions. To address this challenge, we propose a framework of domain adaptation of MRC. This framework consists of three processes: (1) training an MRC model on the source domain, (2) converting conversations into documents using document generation (DG), which is a task we developed for retrieving important information from several human conversations and converting it to an abstractive document text, and (3) transferring the MRC model to the target domain with unsupervised domain adaptation, which consists of two stages; self-training and discriminative learning.

Our contributions are summarized as follows:

- We propose a framework of unsupervised domain adaptation of MRC in which the only available data are unlabeled human conversations in the target domain.
- We evaluated MRC models with four different domain data.

2 Related Work

2.1 Machine Reading Comprehension (MRC)

With the wide use of deep learning, significant progress has been achieved on many natural-language-processing tasks including MRC. Hermann et al. (2015) proposed an MRC model using bidirectional long short-term memory to capture the context of documents. The idea has become the foundation of many MRC models. It is a challenging task how to make a machine imitate a human to understand the document and be able to answer questions. A large dataset has played a critical role in progressing MRC research. Rajpurkar et al. (2016) released SQuAD (the Stanford Question Answering Dataset), which contains more than 100,000 sets of a question, answer, and document. After that, the contributions of MRC can be grouped into four categories: developing new model structures, creating new datasets, multi-task learning, and introducing a new evaluation method (Baradaran et al., 2022). Many MRC studies have focused on developing model structures by introducing an end-to-end neural network model to obtain state-of-the-art performance (Huang et al., 2018, 2019; Shen et al., 2017; Seo et al., 2017; Xiong et al., 2017). In a different direction, other papers have focused on creating new datasets (Feng et al., 2020; Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020). The main trend in these papers was to create datasets considering more complex phenomena, i.e., a query is formed by multiple turns and a document has structural elements. Some papers addressed methods to evaluate whether the system acquires a "true" comprehension capability (Jia and Liang, 2017; Wang and Bansal, 2018). To test true comprehension capability, for instance, QA performance was measured when documents were made distracting by inserting adversarial noisy sentences.

2.2 Document Generation (DG)

MRC returns no answer for irrelevant questions to the documents. Self-learning of the MRC model from human conversations is a new challenge. To achieve this, converting human conversations into documents is necessary. We call this task document generation (DG). DG is a similar task to conversation summarization, which focuses on simply extracting important context from conversations (Li et al., 2019). Unlike conversation summarization, however, DG is aimed to use for a specific application, i.e., customer service. Therefore, it is necessary to extract useful information for the application from conversation contexts, e.g., the topics of customer queries and solution the operator provides. Document summarization is divided into two types of methods: extractive and abstractive. Extractive methods select key sentences from original documents (Knight and Marcu, 2000), while abstractive methods highlight key phrases and compactly rewrite them (Gehrmann et al., 2018; Maynez et al., 2020; Chen and Bansal, 2018). DG should ensure the correctness of key facts. For example, when an operator asked, "Were the plates lost or stolen?", and a customer said,

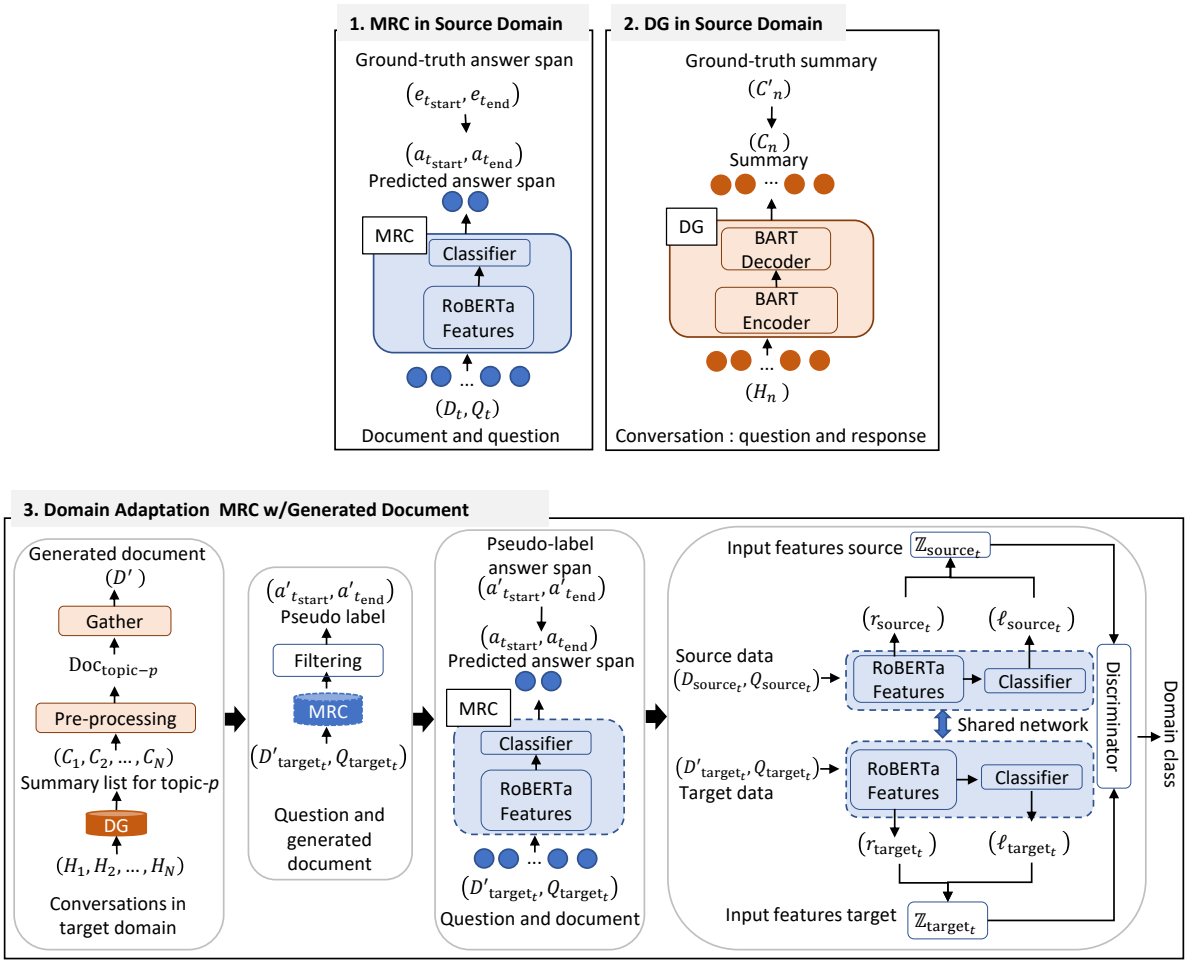

Figure 1: Proposed domain-adaptation framework that includes our developed DG for MRC. Parameters of modules in dash boxes are updated during domain adaptation.

"No", then the operator's response is "You will not be eligible for a refund". In this case, a key sentence that should be included in the document is a sentence such as "You are eligible for a refund if the plates are lost/stolen or destroyed".

2.3 Domain Adaptation

One of the hot topics in MRC is developing simultaneous learning of multiple tasks (transfer learning) (Ruder et al., 2019) and transferring the learned MRC model from one domain to another. This is a promising task for obtaining better results, especially in a data-poor setting. The task is referred to as domain adaptation, which can be divided into two types of methods; supervised, and unsupervised. With supervised methods, the model is trained, where the label is available in the target domain (Kratzwald and Feuerriegel, 2019). The aim of supervised domain adaptation in MRC is to enlarge the number of domains the learned model can cope with. With unsupervised methods, no labeled information is available in the target domain. Cao et al. proposed an unsupervised domain-adaptation method on reading comprehension (Cao et al., 2020). They first trained the MRC model in a source domain by fine-tuning a Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019), then in the adaptation stage, the fine-tuned model is used to generate synthetic question-answer pairs in the target-domain documents, and the synthetic pairs are used in self-training. Their method worked with an assumption that questions and documents are available in the target domain. Wang et al. proposed a similar method (Wang et al., 2019). The difference is that they used a question generator to extract questions from documents in the target domain. Although their method showed promising results, current MRC cannot handle irrelevant questions that have no information in the given document.

Our proposed framework is on unsupervised domain adaptation tasks. Unlike the above-mentioned studies, we used human conversations, which are the only available data in the target domain as the input.

3 Proposed Framework

The main objective of our research is to develop an MRC technique with which the MRC model can be automatically updated on the basis of human conversations. To achieve this, we add DG to the MRC pipeline. The role of DG is to convert human conversations to documents. The generated documents are then added to the training data of the MRC model.

Let us assume that we have two different types of domain data: source domain that has conversations and corresponding documents and target domain that has only conversations. If we have an MRC model trained with source domain data, our goal is to update the model to cover target-domain questions. However, the target domain has no document related to the conversations. Thus, the MRC model should be trained with the only available conversation data in the target domain. As shown in Fig. 1, our framework consists of the following three processes.

1. Training an MRC model with answer spans for given questions and corresponding documents data in the source domain.

2. Converting conversations to documents by DG through model training with source-domain data. Given human conversation as an input, the MRC model returns a summary of the conversation.

3. Transferring the MRC model to the target domain with unsupervised domain adaptation. There are two stages; self-learning to train the MRC model with synthetic data and discriminative learning to learn the feature distribution between source and target domains. Thus, the model can provide the answers of questions from both source and target domains.

3.1 Machine Reading Comprehension in Source Domain

Let $M_{\text{source}} = (D, Q, A)$ denote an MRC dataset in the source data, where D, Q, and A represent documents, questions, and answer span for the questions, respectively. A question contains not only the *user*'s question but also the dialogue history between the *user* and *expert*. An MRC model $\mathscr{M}$ takes documents $D = (d_1, d_2, ..., d_{T_{\text{source}}})$ and questions $Q = (q_1, q_2, ..., q_{T_{\text{source}}})$ as input, where T_{source} is the amount of data in the source domain. The model is trained to predict the correct answer spans:

$$A = ([e_{1_{\text{start}}}, e_{1_{\text{end}}}], ..., [e_{T_{\text{source}_{\text{start}}}}, e_{T_{\text{source}_{\text{end}}}}]) \quad (1)$$

We use Transformer models to implement the MRC model in the source domain. The Transformer encoder is used to contextually represent the question along with the document. Question q_t and document d_t are passed to the Transformer encoder to create contextual representations of the input. To obtain the starting and ending indices of the answer, the encoder output is sent to a linear layer to be converted into logits corresponding to the probabilities of being the start index ($a_{t_{\text{start}}}$) and end index ($a_{t_{\text{end}}}$) of the answer span.

The $a_{t_{\text{start}}}$ and $a_{t_{\text{end}}}$ are optimized by minimizing the following loss function:

$$\mathcal{L} = \frac{1}{2}(CEL(e_{t_{\text{start}}}, a_{t_{\text{start}}}) + CEL(e_{t_{\text{end}}}, a_{t_{\text{end}}})), \quad (2)$$

where CEL is the cross-entropy loss function, and $e_{t_{\text{start}}}$ and $e_{t_{\text{end}}}$ are the labels at token number t for the answer start and end indices $a_{t_{\text{start}}}$ and $a_{t_{\text{end}}}$, respectively.

3.2 Document Generation

Given an input dialogue between a *user* and *expert*, the goal of DG is to produce a multi-sentence summary that captures the highlights of the dialogue. Let N be the total number of dialogues consisting of a conversation about topic-p. By giving a dialogue context H, the goal is to generate the summary C of the dialogue. The n-th dialogue has a list of utterances $H_n = (h_1, h_2, ..., h_L)$, where h_l is the l-th utterance in the dialogue and L is the number of utterances. Each utterance contains a sequence of tokens $h_l = (x_{l,\text{role}}, x_{l,1}, x_{l,2}, ..., x_{l,n_l})$, where $x_{l,j}$ is the j-th token in h_l and n_l is the number of tokens in the l-th role's utterance. At the beginning of the sequence, we add a special token $x_{l,\text{role}}$, which represents the role of the speaker, i.e., $x_{l,\text{role}} \in (\textit{user}, \textit{expert})$. The n-th output from DG is also a sequence of word tokens $C_n = (c_1, c_2, ..., c_K)$, where K is the number of tokens. Note that the highlight in C_n is just some

of the information in topic-p. To gather all information and generate a document of topic-p, we collect the highlights from all conversations, remove duplicate sentences, then put the highlights together as a completed document.

The ground truth summary C'_n is created by combining all the correct answer of the *user*'s question for corresponding input dialogue H_n, where *user*'s questions $\in Q$. Given a *user*'s question q and its answer's spans ($e_{\text{start}}, e_{\text{end}}$), the correct answer sentence w is taken from the original document. Thus, if there are y number of utterances in input dialogue H_n in which the role is *user*, the ground truth summary is a series of consecutive sentences $w_1, w_2, ..., w_y$.

We use a Transformer (Vaswani et al., 2017) model built with a seq2seq model combining an encoder with a decoder. Studies have shown that if the model is first trained on a large corpus, it will learn the distribution of that corpus vocabulary (Gururangan et al., 2020). Motivated by this, we experimented with pre-training on different out-of-domain datasets, such as the news-based CNN/Daily Mail corpus (Nallapati et al., 2016) and conversation-based SAMSum corpus (Gliwa et al., 2019), and continued to fine-tune the model on our experimental dataset in the source domain. We first trained the model for the summarization task on the large CNN/Daily Mail corpus. The reason we first train the model with this corpus is that the corpus has high quality contexts and summaries that can enable the model to learn a structured document. However, our main focus is not on summarizing an article but to generate highlights from conversation data. Thus, we continue fine-tuning the model for the summarization task with a conversation-based corpus, such as SAMSum, to obtain more auxiliary vocabulary.

The trained DG model in the procedure above will be used to generate documents in the target domain by giving conversations in the target domain.

3.3 Unsupervised Domain Adaptation in Target Domain

The unsupervised domain adaptation in our framework consists of two stages; self-learning and discriminative learning.

In the self-learning stage, we have to generate pseudo-label samples. The MRC model $\mathscr{M}$ described in Section 3.1 is used to provide pseudo-labels to unlabeled documents in the target domain generated with DG trained in Section 3.2. However, because the predicted output consists of false answers, we have to choose reliable pseudo-labels. Thus, the underlying assumption in this stage is we only take the samples having high-confidence predictions. Retraining the model using high-confidence samples will further improve its performance (Saito et al., 2017). Despite the fact that the distribution of vocabulary is different between source and target domains, both domains may have similar characteristics. Thus, some samples with high-confidence scores will be similar to or the same as correct answer spans in the target domain. To provide pseudo-labels, we first gather a set of answer spans that have the top n_{best} answer-span probabilities $\hat{p}_t$. The $\hat{p}_t$ is calculated using a softmax function applied to the sums of start index logits $a'_{t_{\text{start}}}$ and end index logits $a'_{t_{\text{end}}}$. We assign a pseudo-label to q_t if the following two conditions are satisfied. First, $\hat{p}_t$ should exceed the threshold parameter (th_{prob}), which we set in the experiment. The second requirement is that the span should not be an empty text. After the pseudo-labeled training set (M'_{target}) is composed, $a_{t_{\text{start}}}$ and $a_{t_{\text{end}}}$ are updated on the basis of the loss in Eq. (1), except we replace $e_{t_{\text{start}}}$ and $e_{t_{\text{end}}}$ with $a'_{t_{\text{start}}}$ and $a'_{t_{\text{end}}}$, respec-

Algorithm 1 Domain adaptation of MRC with DG. $\mathscr{M}$ is MRC source model, D' is generated document in target domain derived from DG model, and $iter_{\text{DA}}$ is training-epoch number for domain adaptation.

Input: $M_{\text{source}} = \{(D_t, Q_t)\}_{t=1}^{T_{\text{source}}}$, ▷ Source data
$M_{\text{target}} = \{(D'_t, Q_t)\}_{t=1}^{T_{\text{target}}}$, ▷ Target data
$\mathscr{M}$
Output: Optimal model $\mathscr{M}$ in the target domain

$M'_{\text{target}} = \emptyset$
for $j \leftarrow 1$ to $iter_{\text{DA}}$ **do**
  for $t \leftarrow 1$ to T_{target} **do** ▷ Pseudo-labeled generation
    Use $\mathscr{M}$ to predict the pseudo-labels $a'_{t_{\text{start}}}$ and $a'_{t_{\text{end}}}$ for (D'_t, Q_t) and obtain probability $\hat{p}_t$
    if $\hat{p}_t \geq th_{\text{prob}}$ **and** $D'_t(a'_{t_{\text{start}}}, a'_{t_{\text{end}}}) \neq$ empty text **then**
      if $Q_t \notin M'_{\text{target}}$ **then**
        Put $(D'_t, Q_t, a'_{t_{\text{start}}}, a'_{t_{\text{end}}})$ into M'_{target}
      end if
    end if
  end for
  for mini-batch $\mathscr{b}$ **in** M'_{target} **do** ▷ Self training
    Train $\mathscr{M}$ with $\mathscr{b}$
  end for
  for mini-batch $\mathscr{b}_{\text{target}}$ **in** M'_{target} **and** $\mathscr{b}_{\text{source}}$ **in** M_{source} **do** ▷ Discriminative learning
    Train $\mathscr{M}$ and $\mathscr{D}$ with $\mathscr{b}_{\text{target}}$, $\mathscr{b}_{\text{source}}$, and domain labels
  end for
end for

tively. In each epoch during adaptation training, pseudo-labeled samples are updated using the last model. An additional sample t will be added if the t-sample did not exist in the last pseudo-labeled samples.

The discriminative-learning stage is used for the MRC model to learn the difference in the feature distribution between source and target domains. We combine the following two representation outputs from both source and target samples: (1) the last hidden state $r \in \mathbb{R}^{\jmath \times \hbar}$, which is the output of the last layer of the MRC model, and (2) concatenation of start-logits and end-logits, which outputs ℓ with dimension $\jmath \times 2$. Note that $\jmath$ and $\hbar$ are the maximum input sequence length and hidden state dimension, respectively. We set $\jmath = 2 \times \hbar$. The input feature $\mathbb{Z}$ is calculated with the following process:

$$\mathbb{Z} = \ell \odot avg_{\text{col}}(r), \tag{3}$$

where avg_{col} means the average along columns, which returns a vector in $\in \mathbb{R}^{\hbar}$, $\odot$ is an element-wise product, $\ell \in \mathbb{R}^{\hbar}$, and $\mathbb{Z} \in \mathbb{R}^{\hbar}$. Discriminator $\mathcal{D}$ takes $\mathbb{Z}$ as input and computes the probability using a neural network consisting of three linear layers, in which the final layer outputs a one-dimensional value that shows the output probability.

The loss function is the binary cross-entropy loss,

$$\mathcal{L}_{\text{dsc}} = -\left(u \log(\hat{u}) + (1-u)\log(1-\hat{u})\right), \tag{4}$$

where $\hat{u}$ is the probability output from $\mathcal{D}$, and $u \in \{0, 1\}$ is the ground-truth label; 0 for the source domain and 1 for the target domain. The entire procedure of domain adaptation is shown in Algorithm 1.

4 Experiments

4.1 Dataset

In our experiments, we used the Doc2dial (Feng et al., 2020) dataset consisting of about 4,800 annotated conversations with an average of 14 turns per conversation. The utterances are grounded in over 480 documents from four domains of public government service websites in the U.S.: Social security administration (**ssa**), Department of Motor Vehicles (**dmv**), Federal Student Aid (**studentaid**), Veteran's Affairs (**va**).

In the training process of the DG model, as we mentioned in Section 3.2, we first trained the model on the large CNN/Daily Mail corpus (Nallapati et al., 2016). This corpus is based on the news articles taken from the CNN and Daily Mail websites. It includes various subjects such as travel and business. It also contains about 300,000 articles written by journalists at CNN and the Daily Mail. We continued to fine-tune the model in the conversation-based SAMSum corpus (Gliwa et al., 2019). The SAMSum corpus is an English dataset consisting of about 15,000 natural conversations in various scenes of real life such as chatting, meeting arrangements, and political discussion. We finally fine-tuned the model in the Doc2dial dataset.

4.2 Hyper-parameters

We implemented the QA from HuggingFace Transformers (Wolf et al., 2019) with a pre-trained model as the encoder and fine-tuned it on the Doc2dial dataset during training. We used two different pre-trained models as the MRC models in the source domain: BERT (Devlin et al., 2019), and Robustly Optimized BERT Approach (RoBERTa) (Liu et al., 2019). Since the grounded document is often longer than the maximum input sequence length for the QA model, we followed a previous study (Feng et al., 2020) to truncate the documents in windows with a stride. We set the stride to 128 tokens, the number of epochs to 5 with cross entropy as the loss function, and the learning rate to 3×10^{-5}. The batch size was set to 15, and the maximum distance between starting (a_{start}) and ending (a_{end}) indices of answers was set to 50.

In the training process of the DG model, we used BART$_{large}$, which includes 12 Transformer layers in the encoder and decoder [1]. We set the number of epochs to 5 with cross entropy as the loss function and set the learning rate to 3×10^{-5}. We set the batch size to 15 and maximum length of input sequences to 1024.

In the unsupervised domain-adaptation process, we set the learning rate to 3×10^{-5} in the self-training stage and 5×10^{-5} in the discriminative-learning stage. We set the same parameters as in MRC in source-domain training for the maximum distance between starting and ending and number of epochs ($iter_{\text{DA}}$). The batch size was set to 5. The input dimension of the first layer in the discriminator network ($\hbar$) was 1024, and the maximum sequence ($\jmath$) was 512. We used a rectified linear unit as the activation function in the first two

[1] https://huggingface.co/facebook/bart-large

layers. The threshold (th_{prob}) was set to 0.5, and n_{best} was 20.

All parameters were determined on the basis of the best ROUGE-1 score for training the DG model and F1 score for MRC models (source and domain adaptation) on the validation dataset in the experiments.

5 Results and Discussion

5.1 MRC in Source Domain

	Specific Domain				All Domains
	ssa	**dmv**	**studentaid**	**va**	
BERT	61.29	53.88	50.99	68.77	62.83
RoBERTa	**64.93**	**63.13**	**62.86**	**73.01**	**70.30**
Baseline	-	-	-	-	65.30

Table 1: The F1 scores [%] for MRC in source domain on Doc2dial validation set. The baseline score is reported in (Feng et al., 2020).

For the MRC source training, we compared the F1 score results (shown in Table 1) between two different language models: BERT and RoBERTa. We trained and evaluated the MRC model with a specific domain, and with all the domain data. With BERT, which is the same Transformer model as the baseline, we obtained an F1 score of 62.83%. This result is lower than the reported baseline (Feng et al., 2020) of 65.30%. However, with RoBERTa, we obtained a higher score of 70.30%. Therefore, we used RoBERTa to train unsupervised domain adaptation of the MRC model in the target domain.

5.2 Document Generation

Dataset	Evaluation Metrics [%]	
	ROUGE-1	ROUGE-L
Doc2dial	67.02	44.06
Doc2dial + CNN/Daily Mail	69.74	45.26
Doc2dial + CNN/Daily Mail + SAMSum	**69.94**	**45.71**

Table 2: DG results on Doc2dial validation set during further pre-training on different QA datasets. We trained with the BART model.

The performances of DG are listed in Table 2. Experiments with BART on the validation set showed that fine-tuning on different datasets is beneficial. Pre-training on more structural corpora, such as CNN/Daily Mail, is more useful than directly fine-tuning BART into the Doc2dial dataset. Furthermore, training the model using SAMSum, which contains conversational data and is more similar to Doc2dial, further improved the performance. An example of the generated document results is shown in Table 3. When we increase the conversation, new information will be added to the generated document. Current generated documents gather all information that is based on the conversation. Thus, the output results will significantly differ compared with the original document in the target domain, especially for the information structure. The information order depends on the given conversation order.

Conversation in VA [Veterans' Affairs] claim topic		
U	:	how do you check your VA claim or appeal status?
E	:	find out how to check the status of a VA claim or appeal online
U	:	can I use the tool?
E	:	do you have one of the following accounts? A Premium My HealtheVet account a Premium DS Logon account used for eBenefits and milConnect , or one you can create here on VA.gov verified ID.me?
U	:	yes
E	:	ok you just log into one of those
Generated document for VA claim topic		
Log in to start finding out how to check the status of a VA claim or appeal online. Use this tool if you have one of the following accounts: A Premium DS Logon account used for eBenefits and milConnect or Verified ID.me.		
Ground truth in original document for VA claim topic		
Check your VA claim or appeal status. Find out how to check the status of a VA claim or appeal online. To use this tool, you'll need to have one of these free accounts: A Premium My HealtheVet account or A Premium DS Logon account used for eBenefits and milConnect, or one you can create here on VA.gov verified ID.me account.		

Table 3: Given conversation between *expert* (E) and *user* (U), DG returns the generated document in the corresponding topic. We used the Doc2dial + CNN/Daily Mail + SAMSum model.

5.3 MRC with Domain Adaptation

5.3.1 With Original Target Documents

Domain (*source* to *target*)	w/o DA	with DA
studentaid to **va**	50.62	**53.27**
studentaid to **ssa**	49.38	**55.12**
studentaid to **dmv**	54.93	**60.19**

Table 4: The F1 scores [%] for MRC without DG when using **studentaid** data as source domain. DA refers to domain adaptation.

We conducted a domain-adaptation test with original target documents to verify the effective-

Domain			studentaid		va		ssa		dmv	
			ROUGE-L	F1	ROUGE-L	F1	ROUGE-L	F1	ROUGE-L	F1
studentaid	w/o DA				53.02	50.62	47.62	49.38	52.04	54.93
	with DA	w/o DSC			53.58	51.63	48.79	50.58	51.55	55.29
		with DSC			**54.26**	**52.45**	**49.70**	**51.02**	**52.17**	**55.61**
va	w/o DA		54.60	52.54			48.97	51.46	53.92	57.85
	with DA	w/o DSC	54.44	52.05			48.72	50.07	52.82	56.24
		with DSC	**55.76**	**53.68**			**49.13**	**51.64**	**53.99**	**57.90**
ssa	w/o DA		51.86	47.60	53.90	51.79			52.02	54.01
	with DA	w/o DSC	51.56	48.52	54.15	53.03			53.14	56.28
		with DSC	**52.48**	**50.12**	**55.05**	**53.12**			**53.32**	**56.95**
dmv	w/o DA		54.66	52.76	53.66	52.06	52.25	56.73		
	with DA	w/o DSC	53.71	52.35	53.71	52.22	51.64	55.99		
		with DSC	**55.68**	**55.07**	**53.77**	**53.08**	**53.33**	**57.86**		

Table 5: MRC results with the document generation (DG). DA refers to domain adaptation and DSC refers to discriminative learning.

ness of a domain-adaptation stage. We first trained an MRC model in the source domain with **studentaid** data. The next procedure was the same as that shown in Algorithm 1, except we used the original document D in the target domain. We set three domain-adaptation dataset pairs, which were **studentaid** to **va**, **studentaid** to **ssa**, and **studentaid** to **dmv**. As shown in Table 4, the F1 scores of the model trained without/with domain adaptation (DA) were 50.62/53.27, 49.38/55.12, and 54.93/60.19% for **studentaid** to **va**, **studentaid** to **ssa**, and **studentaid** to **dmv**, respectively. Thus, the model trained with DA (our framework) outperformed the model trained without DA.

5.3.2 With Generated Target Documents by DG

Finally, we conducted an experiment for our main task, in which the model is trained with unsupervised DA and with DG. The results for each domain are listed in Table 5. We tested under three conditions: the model trained without DA, model trained with DA and without the discriminative-learning stage, and model trained with both DA and discriminative-learning stage. The results indicate that for the model trained with DA, self-learning alone (without discriminative stage) was not strong enough to outperform the model trained without the DA model. We observed that the number of generated pseudo-labeled sets (M'_{target}) remained almost the same in each epoch, such as in **studentaid** to **dmv**. Consequently, the model trained with DA but without the discriminative-learning stage performed worse than the model trained without DA. For **ssa** to **dmv**, the number of generated pseudo-label sets increased during the training process. Thus, the model trained with DA but without the discriminative-learning stage outperformed the model without DA. Despite 1 or 2% improvement, as we add the discriminative stage to the DA-model training, the model trained with both DA and the discriminative-learning stage outperformed the model trained without DA in all datasets. Even with unstructured documents and without labels in the target domain, we proved that our framework can be used to adapt the model from conversation data.

6 Conclusion

We proposed a framework of unsupervised domain adaptation of MRC in which the only available data are unlabeled human conversations in the target domain. DG, which is a task in the framework, converts a given conversation into a document including conversational context. We also tackled a new challenge of conducting domain adaptation from the source domain with a structured document to a new domain with an unstructured document. We showed that only self-learning does not always improve accuracy. However, discriminative learning with self-learning successfully improved conversational-based MRC domain adaptation.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading com-

prehension systems. *Natural Language Engineering*, pages 1–50.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA – Accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314.

Yu Cao, Meng Fang, Baosheng Yu, and Joey Tianyi Zhou. 2020. Unsupervised domain adaptation on reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7480–7487.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. FusionNet: Fusing via fully-aware attention with application to machine comprehension. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization – Step one: Sentence compression. In *Proc. AAAI Conference on Artificial Intelligence*, pages 703–710.

Bernhard Kratzwald and Stefan Feuerriegel. 2019. Putting question-answering systems into practice: Transfer learning for efficient domain customization. *ACM Transactions on Management Information Systems (TMIS)*, 9(4):1–20.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

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. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gùlçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A survey on neural machine reading comprehension. *arXiv preprint arXiv:1906.03824*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. Adversarial domain adaptation for machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2510–2520.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome T. Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9575–9593.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.