Mitigating Data Scarceness through Data Synthesis, Augmentation and Curriculum for Abstractive Summarization

Ahmed Magooda and Diane Litman
University of Pittsburgh
Pittsburgh, PA, USA
{aem132, dlitman}@pitt.edu

Abstract

This paper explores three simple data manipulation techniques (synthesis, augmentation, curriculum) for improving abstractive summarization models without the need for any additional data. We introduce a method of data synthesis with paraphrasing, a data augmentation technique with sample mixing, and curriculum learning with two new difficulty metrics based on specificity and abstractiveness. We conduct experiments to show that these three techniques can help improve abstractive summarization across two summarization models and two different small datasets. Furthermore, we show that these techniques can improve performance when applied in isolation and when combined.

1 Introduction

Training complex neural models usually requires large amounts of data. However, data annotation still poses a challenge for many domains. Thus, much research focuses on data manipulation (e.g., synthesis, augmentation) and additional ways to handle data differently during training. Prior work on the synthesis of textual data has focused on back translation (Parida and Motlicek, 2019; Wang et al., 2018; Sennrich et al., 2016) and word replacement (Wang and Yang, 2015; Zhang et al., 2015). We, on the other hand, propose a different approach to data synthesis through paraphrasing. However, synthesis involves data manipulation on the input level, which might expose the model to grammatically or logically incorrect input. Thus, we explore a second approach to data manipulation based on augmentation rather than synthesis. Augmentation aims to move data manipulation from the input side to any part of the model. This, in turn, can help the model be more resilient to over-fitting. Third, we explore using data more efficiently by integrating curriculum learning into the training process. Curriculum learning reorders training samples based on external criteria, which can help train the model gradually and more efficiently without the need for any external data. We also introduce new difficulty metrics based on specificity and abstractiveness for curriculum construction. Finally, we explore combining multiple techniques (synthesis and curriculum) to overcome the data scarcity issue. Thus, our contribution is threefold: 1) We introduce a simple approach for data synthesis through paraphrasing. 2) We use data augmentation by sample mixing to move augmentation into the model. 3) We integrate a curriculum into the training process and introduce two new difficulty metrics.

2 Related Work

Abstractive summarization for low resource data. Prior proposed methods for tackling domains with scarce data have included finetuning pre-trained models (Bajaj et al., 2021; Yu et al., 2021; Magooda and Litman, 2020) such as BART (Lewis et al., 2020) or using few-shot learning (Bražinskas et al., 2020; Sarkhel et al., 2020). Our work differs in several aspects. First, our work doesn’t focus on improving a certain summarization model; in contrast, we focus on using data efficiently, which can be applied to various models. Second, we focus on techniques that can improve the training process without additional data, e.g., synthesis, augmentation, and curriculum learning.

Data synthesis and augmentation. Data synthesis for text summarization is underexplored, with only a few approaches such as back-generation (Parida and Motlicek, 2019) and template-based summary re-writing (Magooda and Litman, 2020). We propose doing data synthesis by paraphrasing, which is simpler than the back-translation and template methods. While combining synthesis with paraphrasing has been studied in other contexts (Wang et al., 2015; Iyyer et al., 2018), our work differs in both goals and techniques. Wang et al. (2015) proposed synthesizing data, then crowd-
sourcing paraphrases to train semantic parsers, while Iyyer et al. (2018) synthesized data to train a paraphrasing model. Our work, to our knowledge, is the first to use a strong language model finetuned for paraphrasing to synthesize data for text summarization. Finally, for data augmentation, we base our work on the MixText approach (Chen et al., 2020). While the original MixText model is used for classification-based tasks, we introduce a variation for generative tasks (called MixGen) and use it for abstractive summarization.

Curriculum learning. Curriculum learning aims to improve the training procedure with the same amount of data. It has been applied in NLP (Sachan and Xing, 2016, 2018; Tay et al., 2019; Xu et al., 2020; Wang et al., 2020) for machine comprehension, question generation, reading comprehension, NLU and machine translation, respectively. We build on the approach introduced in (Xu et al., 2020); however, the core differences are both the downstream tasks (classification versus abstractive summarization) and the difficulty metrics. In contrast to the only other summarization work that we know of, Kano et al. (2021) focus on large datasets, while we focus on low resource domains. We also introduce two different difficulty metrics (ROUGE and specificity).

3 Summarization Datasets

CourseMirror (CM) is a student reflections dataset that has been used in prior work to study extractive (Luo and Litman, 2015) and abstractive (Magooda and Litman, 2020) summarization. The dataset consists of documents and summaries from four courses. Table 1 summarizes the dataset in terms of the number of documents (# docs) and average reflections per document (# refs). We compiled all courses into one dataset (named CM ALL), then split the documents into training, validation, and test sets (80%, 10%, 10%, respectively) by sampling equally from all courses.

Amazon/Yelp (A/Y) is another small dataset, now consisting of opinions (refer to the appendix for examples) (Bražinskas et al., 2020). The dataset contains customer reviews from Amazon (He and McAuley, 2016) and Yelp. The data contains 160 products/businesses split into training, validation and test sets as shown in Table 1. Each of the products/businesses contains a set of 8 reviews.

4 Proposed Model

4.1 Synthesis via paraphrasing with GPT-2

Influenced by work in style transfer (Krishna et al., 2020), we propose synthesizing new human summaries by using paraphrasing to generate other potential summaries that are paraphrases of the original human summary. We use the paraphraser trained by Krishna et al. (2020). They finetuned a large GPT-2 language model with data from PARANMT-50M (Wieting and Gimpel, 2018) to direct the model into generating diverse paraphrases that they later used for style transfer.

4.2 Augmentation with sample mixing

MixText is a data augmentation approach based on mixing two input samples by weight summing the features corresponding to the two samples at any level of the model (specific layer of the encoder, after encoder, etc.) using $\lambda$. The model is then expected to produce a probability distribution over the available classes, similar to a $\lambda$ weighted sum of the two samples’ gold predictions. We train the model using KL divergence between the predicted distribution and the expected one. We adapt the approach of Chen et al. (2020) for text generation tasks by modifying the decoding process and loss calculation; we call our approach MixGEN. Like the original MixText, we use two input samples and pass them to the encoder. We pass the samples up to a specific layer, then the two hidden states are summed together weighted differently using the $\lambda$ parameter. On the decoder side, first, we construct the expected values using the following:

$$
for \ i \in \min(L_1, L_2) \quad PD_i = [P_1, P_2, P_3, ..., P_v],
$$

$$
for \ j \in [1, v] \quad 
\begin{cases} 
P_j = \lambda, & \text{if } S_1[i] = j \\
P_j = 1 - \lambda, & \text{if } S_2[i] = j \\
P_j = 0, & \text{otherwise}
\end{cases}
$$

where $v$ = vocab size, and $L_1, L_2$ are the human summary of input sample 1 and input sample 2.

<table>
<thead>
<tr>
<th>Data</th>
<th># docs</th>
<th># refs</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM ALL</td>
<td>368</td>
<td>44</td>
<td>294</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>Amazon/Yelp</td>
<td>160</td>
<td>8</td>
<td>58</td>
<td>42</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 1: Dataset summary.
respectively. \( PD_i \) is the probability distribution expected for token \( i \). \( S_1[i], S_2[i] \) are the \( i \)th token of the first sample’s and second sample’s human summary, respectively. In simple wording, during decoding, we expect the output probability distribution across the vocabulary of the decoder at token position \( i \) to have two high values, one with value \( \lambda \) at vocabulary token corresponding to the \( i \)th token of the first sample’s human summary, and another value of \( 1-\lambda \) at vocabulary token corresponding to the \( i \)th token of the second sample’s human summary. This should continue as long as \( i \) is less than or equal to both human summaries’ length. Once \( i \) is greater than the minimum summaries’ length, then the expected distribution would only correspond to the longer summary. Finally, the text generation on the decoder side is auto-regressive, thus, expected token is passed at the end of each generation step. However, if we pass the argmax of the expected distribution, then we will end up always passing the token corresponding to the sample with higher weight (Alpha) vs. (1-Alpha). Thus, we randomly sample from the two input samples based on their weights using the following equation.

\[
\begin{align*}
\text{for } i & \in [1, L_{\text{min}}], \quad P_i \sim U(0,1) \\
\text{for } i & \in [1, L_{\text{min}}] \quad \left\{ \begin{array}{l}
P_i \leq \lambda, \quad T_i \text{ from } S_1 \\
P_i > \lambda, \quad T_i \text{ from } S_2 \\
\end{array} \right.
\end{align*}
\]

where \( S_1 \) and \( S_2 \) are the first and second sample respectively, \( L_{\text{min}} \) is minimum length of both \( S_1 \) and \( S_2 \). \( P_i \) is the sampling probability and \( U(0,1) \) is a uniform distribution.

### 4.3 Curriculum learning (Cur.)

Curriculum learning aims to help the model training process by introducing easier samples first followed by more difficult ones according to a particular difficulty metric. We use the curriculum construction approach introduced in Xu et al. (2020). In this approach, we split data into \( N \) buckets based on a difficulty metric. We then train the model in a difficulty incremental setting. In this work, we use two different curriculum difficulty metrics.

**Specificity (S)** measures how specific or vague a piece of text is. We argue that the more specific a piece of text is, the more complicated it can get. For example, text like (Nothing, Everything is Easy, etc.) are not specific and easy for the model to learn and vice versa. We feed the model less specific pieces of text first during training, then introduce the more specific ones as training progresses. Specificity is calculated (Appendix) on the reflection/review level, so we use the average values of the whole set of reflections/reviews as the document value. For example: for a training sample of an input document \( D \) consists of \( N \) independent reflections/reviews \([r_1, r_2, r_3, \ldots, r_n]\), we calculate the specificity value for the sample as follows:

\[
D_s = \sum_{i=1}^{N} \frac{S(r_i)}{N}
\]

where \( S(r_i) \) is specificity value of the \( i \)'s reflection/review.

**ROUGE (R)** is the standard metric for evaluating summarization performance. Thus, we decided to use ROUGE scores as a difficulty metric. For a training sample, we calculate different ROUGE scores between the input document \( D \) and its corresponding human summary \( S \), then use average of \( R_1, R_2, RL \) as the difficulty metric. According to (Liu et al., 2018), the higher the ROUGE score, the less abstractive the summary is compared to the input, and vice versa. We argue that the more abstractive samples are harder to learn.

### 5 Experiments

#### 5.1 Parameters

**Baselines:** To our knowledge, in prior work there is no data synthesis technique used for summarization except back generation (Parida and Motlicek, 2019) and template synthesis (Magooda and Litman, 2020). Thus, we developed two synthesis baselines (shuffle; shuffle + mask). We generated 10 samples for each of the original training samples for both baselines by randomly shuffling the reflections/reviews. Additionally, for the shuffle-mask baseline, we randomly mask 50% of the reflection/reviews 50% of the time.

**Paraphrasing with GPT-2:** We generate \( N \) synthetic samples for each original sample by generating \( N \) paraphrases of the human summary and shuffle the input reflections. We varied \( N \) between \([5, 10]\) to monitor the effect of synthetic data size.

**MixGEN:** We integrate MixGEN by combining each sample with \( N \) other samples during training. We used \( N=3 \) for our experiments. Moreover, we use mixing probability \( \alpha=0.75 \) as specified by the original code implementation\(^3\).

\(^3\)https://github.com/GT-SALT/MixText
Curriculum learning: In the curriculum learning experiments, we use a specificity prediction model that consists of a DistilBERT (Sanh et al., 2019) encoder with a logistic regression classification layer (Appendix). We normalize the whole training data values between 1 and \( N \), where \( N \) is the number of buckets to split the data. We use \( N=10 \). Similarly, we normalize the average ROUGE value to also be between 1 and \( N \).

5.2 Model Training

In all of our experiments, we use the BERTSum\(^4\) model proposed by Liu and Lapata (2019). We used the same parameters in the original code (Appendix). We conducted experiments on CM and A/Y datasets using proposed methods in a regular training and in a (pretraining \( \rightarrow \) fine-tuning) setting, where we perform pretraining with synthesized data and fine-tuning using original data.

6 Results

<table>
<thead>
<tr>
<th>CM</th>
<th>Pretraining</th>
<th>finetuning</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pretraining</td>
<td>Original</td>
<td>36.34</td>
<td>11.39</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shuff.</td>
<td>38.57</td>
<td>11.72</td>
<td>26.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shuff.+mask</td>
<td>37.07</td>
<td>11.52</td>
<td>26.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cur.(S)</td>
<td>36.88</td>
<td>12.41</td>
<td>27.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cur.(R)</td>
<td>37.01</td>
<td>12.13</td>
<td>27.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mix(n=3)</td>
<td>36.87</td>
<td>11.98</td>
<td>26.57</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>With synthetic data pretraining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synth.(n=5)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A/Y</th>
<th>Pretraining</th>
<th>finetuning</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pretraining</td>
<td>Original</td>
<td>27.71</td>
<td>3.83</td>
<td>17.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shuff.</td>
<td>28.34</td>
<td>4.04</td>
<td>17.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shuff.+mask</td>
<td>28.01</td>
<td>4.21</td>
<td>17.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cur.(S)</td>
<td>28.09</td>
<td>4.28</td>
<td>17.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cur.(R)</td>
<td>\textbf{28.56}</td>
<td>4.33</td>
<td>\textbf{18.18}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mix(n=3)</td>
<td>27.85</td>
<td>3.95</td>
<td>17.89</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>With synthetic data pretraining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synth.(n=5)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Synth.(n=10)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 2: ROUGE results of BERTSum model with different augmentation techniques on CM data.

Table 3: ROUGE results of BERTSum model with different augmentation techniques on A/Y data.

we see that we can get a performance gain across all ROUGE scores across both datasets by mixing training samples compared to normal training with a single sample. Similarly, we can see that providing synthetic data with the proposed paraphrasing approach can help outperform both using original data as well as baselines with (41.14, 14.24, 26.98) compared to (36.34, 11.39, 26) and (38.57, 11.72, 26.94) for original and shuffle baseline respectively on CM, and (28.49, 4.54, 18.08) compared to (27.71, 3.83, 17.83) and (28.34, 4.04, 17.74) on A/Y. Additionally, we can see that increasing the synthetic data size helps to improve the model performance across all ROUGE scores for both CM and A/Y datasets (N=5 vs. N=10).

Now moving to curriculum learning, we can see that integrating a curriculum to reorder training data differently using any of the two proposed difficulty metrics can lead to consistent improvements across all ROUGE scores for both CM and A/Y datasets. Additionally, we can see that curriculum can improve scores compared to the two augmentation baselines across all ROUGE scores except R1 for CM data. On the other hand, we don’t see consistent ROUGE score improvement when using curriculum for fine-tuning after pretraining with synthetic data. We hypothesize that this behavior might be due to performing the pretraining phase without curriculum integration, unlike the fine-tuning phase. We plan to conduct experiments with a curriculum integrated into both pretraining and fine-tuning to validate our hypothesis. Furthermore, while both curriculum difficulty metrics (i.e., Specificity and ROUGE) introduced improvement compared to training with no curriculum, we

\(^4\)Easy to use and one of the SOTA summarization models. https://github.com/nlpyang/PreSumm
didn’t observe any consistent improvement pattern in using one metric over the other.

7 Conclusion and Future work

In this work, we showed that we could mitigate the effect of data scarcity in different datasets (i.e., CourseMirror and Amazon/Yelp) for abstractive summarization using three simple data manipulation techniques. We showed that synthesizing data with paraphrasing to use for pretraining can boost the model performance across all ROUGE scores for different datasets. Additionally, we showed that mixing samples for training can also push the model to be more resilient to overfitting and improve its performance. Finally, we showed that reordering training samples through curriculum, using the proposed difficulty metrics (i.e., Specificity, and ROUGE) would help improve all ROUGE scores across different datasets without the need for any additional data (either true or synthetic). In the future, we plan to try more \( N \) values for synthesis and MixGen. Additionally, we plan to investigate other curriculum difficulty metrics. We plan to use BART model as one of the SOTA models for abstractive summarization. Finally, we are doing additional experiments on multitask learning, and we plan to combine both techniques in one framework targeting low resource domains.

Acknowledgements

The research reported here was supported, in whole or in part, by the institute of Education Sciences, U.S. Department of Education, through Grant R305A180477 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the institute or the U.S. Department of Education. We like to thank the Pitt PETAL group and the anonymous reviewers for advice in improving this paper.

References


A BERTSum training parameters

We train the model for 200K steps using a batch size of 140. We use 20K steps for BERT warmup, 10K steps for decoder warmup, and a max position of 512. We use 4 Nvidia Quadro RTX 5000 GPUs. We use the checkpoint with highest ROUGE score on validation set for testing.

B MixGen Model

Figure 1 shows both the original MixText model and the modified MixText for generative tasks (MixGen).

C Specificity Model

C.1 Model

CourseMirror data is also annotated for specificity. The data contains human annotations for around 7000 reflections\(^5\) using the scheme introduced in (Fan et al., 2017). Table 4 shows the score distribution for CourseMirror specificity dataset. We use the data to train a specificity prediction model. We use the model to predict the specificity values for both CourseMirror and Amazon/Yelp datasets. The specificity prediction model (figure 2) uses DistilBERT encoder to produce reflection embedding, the embeddings are then used as features to train a logistic regression classifier. To keep the number of tuned parameters to minimum, the DistilBERT weights are frozen during the training process. The embeddings are used as fixed features, and all the training is performed on the logistic classifier side.

\[^5\]https://petal-cs-pitt.github.io/data.html

---

Table 4: CourseMirror Specificity dataset score distribution.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1354</td>
<td>2035</td>
<td>2377</td>
<td>1058</td>
</tr>
</tbody>
</table>
D Data samples

D.1 CourseMirror (CM)
Table 5 shows an example of CM sample from CS course.

D.2 Amazon/Yelp
Table 6 shows an example of sample from Amazon/Yelp data.
Point of Interest (POI): Describe what you found most interesting in today’s class.

Student Reflection Document

- the dynamic bag
- I found the creation of the Bag to be the most interesting.
- Learning about bags was very interesting.
- Dr. Ramirez cleared up my understanding of how they should work.
- I was really interested in learning about an entirely new data structure, the Bag.
- I’m also noticing that as these classes get farther along, there is more focus on real world factors that determine strength of code like speed
- The bag concept was cool how basically acts like a bag in real life with its usefulness.
- Bags as a data type and how flexible they are.
- Discussing the Assignment 1
- I found the examples and drawings the teacher drew on the whiteboard the most interesting.
- Abstraction, though seemingly intimidating is kind of just giving programmers a break right?
- We’re given so many more abilities and operations without having to know exactly how to code that.
- That being said, while I understand the applications being explained to me, it’s hard to just manifest that on my own.
- Learning about resizing Bags dynamically
- The discussion of the underlying methods of ADTs such as bags was most interesting
- the implementation of an array bag
- Order does not matter when using a bag.
- It is important to keep all of the values in an array together.
- To do this, you should move an existing element into the vacant spot.
- Looking at ADT’s from both perspectives
- Information held in bags is not in any particular order
- different ways to implement the bag
- Thinking about a more general idea of coding with ADTs and starting to dig into data structures more specifically.
- Code examples of key concepts/methods is always helpful.
- I thought it was a good thing to go through the implementation of both the add( ) and remove( ) methods of the Bag ADT
- Today we were talking about a certain type of ADT called a bag.
- We talked about certain ways that we would implement the methods and certain special cases that we as programmers have to be aware of.
- If you were removing items from ADT bag, you can simply shift the bottom or last item and put it in the place where you removed an item.
- This is because, in bags, order does not matter.
- Learning about managing arrays in a data structure
- The bag ADT and how it is implemented

Reference Abstractive Summary

Students were interested in ADT Bag, and also its array implementation. Many recognized that it should be resizable, and that the underlying array organization should support that. Others saw that order does not matter in bags. Some thought methods that the bag provides were interesting.

Table 5: Sample data from the CourseMirror CS course.
Reviews

This pendant is so unique!! The design is beautiful and the bail is a ring instead of the typical bail which gives it a nice touch!! All the corners are smooth and my daughter loves it - looks great on her. I cannot say anything about the chain because used our own chain. :) Satisfied.

It look perfect in a women's neck!! Great gift, I thought for the price it was going to look cheap, but I was far wrong. It look great. Spect great reward from your woman when you give this to her; D

The prettiest sterling silver piece I own now. I get so many compliments on this necklace. I bought it for myself from my hubby for Valentine’s Day. Why not? When people ask where I got it, I simply say from my loving hubby. And he is off the hook as to what to get me. Win + win.

I love hearts and I love 'love' :) I do not have any negative feedback, the necklace is perfect and the charm is perfect. I just thought it would have been slightly bigger. Overall, I love my new heart necklace.

When I received the package, I was surprised and amazed because the necklace is so elegant, beautiful and the same as the picture shown here. I really love this necklace. It has a unique pendant designed. I will recommend it to someone to order it now...

Item is nice. Not a great quality item, but right for the price. Charm was larger than I expected (I expected small and elegant, but it was large and almost costume jewelry like). I think it is a good necklace, just not what I expected,

I got this as a present for my GF on Valentines day. She loves it and wears it every day! It’s not cheap looking and it hasn’t broken yet. The chain hasn’t broken either even though it is very thin. Strongly recommend it!

Over all service has been great the only problem, I ordered a purple Mickey Mouse case for iPhone 4S they sent a black, n I felt it was to much trouble n such a small item to send back so needless to say its put back in a drawer somewhere

Abstractive Summary

This silver chain and pendant are elegant and unique. The necklace is very well made, making it a great buy for the cost, and is of high enough quality to be worn every day. The necklace looks beautiful when worn bringing many compliments. Overall, it is highly recommended.

Table 6: Sample data from the Amazon/Yelp data.