Abstract Text Summarization: A Low Resource Challenge

Shantipriya Parida Petr Motlicek
Idiap Research Institute
Rue Marconi 19, 1920 Martigny, Switzerland
{shantipriya.parida,petr.motlicek}@idiap.ch

Abstract

Text summarization is considered as a challenging task in the NLP community. The availability of datasets for the task of multilingual text summarization is rare, and such datasets are difficult to construct. In this work, we build an abstract text summarizer for the German language text using the state-of-the-art “Transformer” model. We propose an iterative data augmentation approach which uses synthetic data along with the real summarization data for the German language. To generate synthetic data, the Common Crawl (German) dataset is exploited, which covers different domains. The synthetic data is effective for the low resource condition and is particularly helpful for our multilingual scenario where availability of summarizing data is still a challenging issue. The data are also useful in deep learning scenarios where the neural models require a large amount of training data for utilization of its capacity. The obtained summarization performance is measured in terms of ROUGE and BLEU score. We achieve an absolute improvement of +1.5 and +16.0 in ROUGE1 F1 (R1_F1) on the development and test sets, respectively, compared to the system which does not rely on data augmentation.

1 Introduction

Automatic text summarization is considered as a challenging task because while summarizing a piece of text, we read it entirely to develop our understanding to prepare highlighting its main points. Due to the lack of human knowledge and language processing abilities in computers, automatic text summarization is a major non-trivial task (Allahyari et al., 2017).

Two major approaches for automatic summarization are: extractive and abstractive. The extractive summarization approach produces summaries by choosing a subset of sentences in the original text. The abstract text summarization approach aims to shorten the long text into a human-readable form that contains the most important fact from the original text (Allahyari et al., 2017; Kryściński et al., 2018).

The deep learning-based neural attention model when applying to abstract text summarization performs well compared to standard learning-based approaches (Rush et al., 2015). Abstract text summarization using the attentional encoder-decoder recurrent neural network approach shows a state-of-the-art performance and sets a baseline model (Nallapati et al., 2016). Further improvements are introduced to the baseline model by using the pointer generator network and coverage mechanism using reinforcement learning based training procedure (See et al., 2017; Paulus et al., 2017).

There is an inherent limitation to natural language processing tasks such as text summarization for resource-poor and morphological complex languages owing to a shortage of quality linguistic data available (Kurniawan and Louvan, 2018). The use of synthetic data along with the real data is one of the popular approaches followed in machine translation domain for the low resource conditions to improve the translation quality (Bojar and Tamchyna, 2011; Hoang et al., 2018; China-Ríos et al., 2017). The iterative back-translation (e.g. training back-translation systems multiple times) were also found effective in machine translation (Hoang et al., 2018). We explore similar approaches in our experiments for the text summarization task.

The organizations of this paper is as follows: Section 1 describes related work on abstract text summarization. Section 2 explains the techniques followed in our work. Section 3 describes the dataset used in our experiment. Section 4 explains the experimental settings: models and their parameters. Section 5 provides evaluation results with
analysis and discussion. Section 6 provides conclusion to the paper.

2 Method Description

Across all experiments performed in this paper, we have used the Transformer model as implemented in OpenNMT-py\(^1\) (Vaswani et al., 2018; See et al., 2017). The Transformer model is based on encoder/decoder architecture. In context to summarize, it takes text as input and provides its summary.

We use synthetic data as shown in Figure 1 to increase the size of the training data.

![Figure 1: Generation of synthetic data using a reverse system.](image)

Figure 1: Generation of synthetic data using a reverse system. To generate synthetic data, first, a system in the reverse direction (i.e. source as summary and target as text) is trained and then used to generate text for the given summary. Then both the real and synthetic data acts as input to the final system.

3 Dataset Description

We use German wiki data (spread across different domain) collected from the SwissText 2019\(^2\) (real data) and Common Crawl\(^3\) data (synthetic data) in our experiment. The statistics of all the datasets are shown in Table 1.

### Table 1: Statistics of the experimental data which include the number of texts and their summaries.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Text</th>
<th>#Summ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train_Real(SwissText)</td>
<td>90K</td>
<td>90K</td>
</tr>
<tr>
<td>Train_RealSynth(Swiss+CC)</td>
<td>190K</td>
<td>190K</td>
</tr>
<tr>
<td>Train_RealSynthRegen(Swiss+CC)</td>
<td>190K</td>
<td>190K</td>
</tr>
<tr>
<td>Dev(SwissText)</td>
<td>5K</td>
<td>5K</td>
</tr>
<tr>
<td>Test(SwissText)</td>
<td>5K</td>
<td>5K</td>
</tr>
</tbody>
</table>

The experiments performed over these datasets are described in Section 4.3 (denoted as S1 experimental setup).

3.2 Common Crawl dataset used as synthetic data

The data crawled from the Internet (Common Crawl) used to prepare synthetic data to boost the training. The steps followed to create the synthetic dataset as follows:

Step 1: **Build vocab:** We create vocabulary using SwissText based on the occurrence of the most frequent (top N) German words.

Step 2: **Sentence selection:** The sentences from the Common Crawl data are selected with respect to the vocabulary based on the threshold we provide (e.g. a sentence has 10 words and the threshold is 10% (0.1)). For a sentence to be selected, at least 1 out of 10 words should be in the vocabulary.

Step 3: **Filtering:** Select random sentences (e.g. 100K) from the selected Common Crawl data in the previous step.

Step 4: **Generate summary:** The 100K data obtained from the previous step are used as a summary and required to generate corresponding text. We use the reverse trained model where we provide the summary as source and target as text. This results in the text as well as the corresponding summary as additional data to be utilized along with real data (SwissText).

Eventually, the 190K dataset is created (denote as Train_RealSynth) as a combination of 90K SwissText train data (real) and 100K synthetic data. This dataset is used in the experimental setup S2 (described in details in Section 4.3).

4 Experimental Setup

This section describes our experiments conducted for the text summarization task.
Table 2: Evaluation results of our models on development (dev) and testing (test) sets. The automatic evaluation scores in terms of Rouge (R1,F1, R2,F1, RL,F1) and BLEU for the output summaries are shown in the table.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Dataset</th>
<th>R1_F1</th>
<th>R2_F1</th>
<th>RL_F1</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Dev</td>
<td>43.9</td>
<td>28.5</td>
<td>46.3</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>39.7</td>
<td>22.9</td>
<td>42.2</td>
<td>9.0</td>
</tr>
<tr>
<td>S2</td>
<td>Dev</td>
<td>45.4</td>
<td>29.8</td>
<td>47.4</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>55.7</td>
<td>41.8</td>
<td>57.6</td>
<td>20.8</td>
</tr>
<tr>
<td>S3</td>
<td>Dev</td>
<td>44.3</td>
<td>28.5</td>
<td>46.4</td>
<td>13.1</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>40.0</td>
<td>23.0</td>
<td>42.3</td>
<td>9.4</td>
</tr>
</tbody>
</table>

4.1 Prepossessing
The preprocess step involves preprocessing the dataset such that source and target are aligned and use the same dictionary. Additionally, we truncate the source length at 400 tokens and the target length at 100 tokens to expedite training (See et al., 2017).

4.2 Model Parameters
The Transformer model is implemented in OpenNMT-py. To train the model, we use a single GPU. To fit the model to the GPU cluster, a batch size equal to 4,096 is selected for training. The validation batch size is set to 8. We use an initial learning rate of 2, drop out of 0.2 and 8,000 warm-up steps. Decoding uses a beam size of 10 and we did not set any minimum length of output summary.

4.3 Model Setup
We use 3 settings: (i) real data (we set this as the baseline in our experiment), (ii) real data and synthetic data, and (iii) real and regenerated synthetic data for the summarization task, described as follows:

1. **S1: Transformer model using Train_Real data**
   In this setup, we use the “Train_Real” data for training the Transformer model.

2. **S2: Transformer Model using Train_RealSynth data**
   In this setup, we use the “Train_RealSynth” data for training the Transformer model. As the balance between real and synthetic data is an important factor, we maintain a 1:1 ratio (e.g. 1 (real) :1 (synthetic)) for our experiment (Sennrich et al., 2016).

3. **S3: Transformer Model using Train_RealSynthRegen data**
   We propose an iterative approach to improve the quality of synthetic summaries. In this setup, after training a system with (real+synthetic) data, it is used to regenerate synthetic data for the final system. As a result, the input data to the final system is a combination of real and regenerated synthetic data as shown in Figure 2.

4.4 Training Procedure
The copying mechanism is applied during training. It allows the summarizer to fall back and copy the source text when encounters < unk > tokens by referencing to the softmax of the multiplication between attention scores of the output with the attention scores of the source (See et al., 2017). The systems are trained for 300K iterations.

5 Evaluation and Discussion
We evaluate the results for every 10,000 iterations on the dev and test set. The automatic evaluation results based on the dev and test set are shown in Table 2 with sample summaries in Table 3. To evaluate the proposed algorithms, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score, which is a popular metric for text summarization task, and has several variants like ROUGE-N, and ROUGE-L, which measure the overlap of n-grams between the system and reference summary (LIN, 2004). We use ROUGE_1 F1 (R1_F1), ROUGE_2 F1 (R2_F1), and ROUGE_L F1 (RL_F1) for scoring the generated summary. In addition, we also use the SacreBLEU evaluation metric (Post, 2018).

Figure 3 presents the learning curves for the models (S1 and S2) on the development set. It can be seen that there is a variance (e.g. word

https://github.com/mjpost/sacreBLEU
Figure 2: Regeneration of synthetic data. After training a system with real+synthetic data (Reverse System2 above), used to create synthetic summarization data for the final system.

Figure 3: Learning curves in terms of Rouge1 F1 (R1 F1) Score on dev set.

selection, summary length) for model S2 generated summary as compared with model S1. During manual verification, we found that the summaries generated without a minimum length constraint appear better compared to summaries with minimum length constraint. Although we do not explicitly specify a minimum length parameter for generating summaries for the models, the average length of words generated by model S2 (e.g. 41.42 words) is longer than the model S1 (e.g. 39.81 words). Some data (e.g. name, year) were found inconsistent during a comparison of the generated summary with respect to the reference. There is a variance in summaries generated by model S3 as compared to S2 and S1. In terms of Rouge score model S3 outperforms model S1 but perform worse than model S2 (see Table 2).

6 Conclusion

In this paper, we highlighted the implementation of synthetic data for the abstract text summarization task under low resource condition, which helps improving the text summarization system in terms of automatic evaluation metrics. As the next step, we plan to investigate: i) synthetic summarization data, and ii) applying transfer learning on text summarization for the multilingual low resource data set with little or no ground truth summaries (Keneshloo et al., 2018).
Acknowledgments

The work is supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: “SM2: Extracting Semantic Meaning from Spoken Material” funding application no. 29814.1 IP-ICT.

References


