Extractive-Abstractive Cross Lingual Summarization

Prishita Kadam
Purdue University
pkadam@purdue.edu

Sarayu Vyakaranam
Purdue University
svyakara@purdue.edu

Abstract

This note presents a novel approach to cross-lingual summarization tasks- extraction and abstraction combined. We explore approaching the summarization task using plain abstraction or a combination of extraction and abstraction, followed by which the summaries are translated into two languages (Chinese and German). We also explore two approaches- a few-shot approach and a fine-tuned approach. Furthermore, we experiment with the best-performing dialogue-to-extractive-summary ratio for the given data set. We conclude that (i) combination of extractive and abstractive summarization results in the better performance than plain abstractive summarization (ii) the optimal dialogue-to-extractive-summary ratio for extraction was found to be 0.5 (iii) the few-shot model showed most efficient results, as compared to the fine-tuned model. We present loss curves along with comparative performance metrics to conclude that the few-shot model with extraction (SBERT with summarization ratio- 0.5) and abstraction (flan-t5-small) worked best for the cross-lingual data set.

1 Introduction

Cross-lingual summarization is the task of summarizing text in one language and translating it to another language (Zhu et al., 2019). It can be thought of as a two-stage problem with summarization followed by translation. The summarization task is typically performed by two methods. An extractive approach computes a subset of the input text to generate a summary. Different ratios of the original text can be used to obtain the summary. In contrast, the abstractive method reorganizes the language in the input text and can also add additional new phrases if necessary (Zhu, 2021).

Problem Statement: We propose to develop a two phase summarization model with extractive and abstractive, and observe if the model performed good summarizations and translations on dialogue data. This approach is different from other approaches to summarization and translation because instead of inputting an entire dialogue directly for summarization, we extract the relevant information inputting this to an abstractive model. We hypothesize that this combined model will produce more accurate summaries. We run a plain abstraction summarization model as a baseline to compare with the extractive-abstractive model with and observe any benefit of combining the two methods. For the combined model, the extraction phase is performed using an optimal number of clusters generated by the elbow method. We simultaneously experiment with the fraction of sentences present in the final summary to obtain the optimal ratio of sentences in the summary. For the abstraction phase, we train two models, t5-small and flan-t5-small for the fine-tuned and few shot approaches respectively Consequently, we asses both the model's performance using ROUGE scores. Finally, we perform the translation task to convert the summaries in English language to German and Chinese using a pre-trained model.

We approach the cross-lingual summarization task with an interesting twist of combining two summarization techniques to observe if it performs better than a baseline model on plain. The data inputted to the model is english dialogue data.

A risk of our proposed approach for cross-lingual summarization is that during the extraction phase of the combined model, if the optimal clustered are not computed accurately it may lead to loss of information and hence, bad or meaningless summarization. We experimented with different dialogue-to-extractive-summary ratios and elaborate on this in Section 3.2.

2 Methods

Approach: The first step was to break down the cross-lingual task into two sub-tasks, namely summarization and translation. For the summarization task, we explored two possible model architectures (i) abstractive and (ii) extractive followed by abstractive. We justify that the exploration of the combined model because extraction performed before abstraction would provide the most relevant context from the entire dialogue to the abstractive model. This would ensure that the abstractive model would train more efficiently and produce better predictions that plain abstractive where the input was the entire dialogue rather than the extracted dialogue.

Challenges:

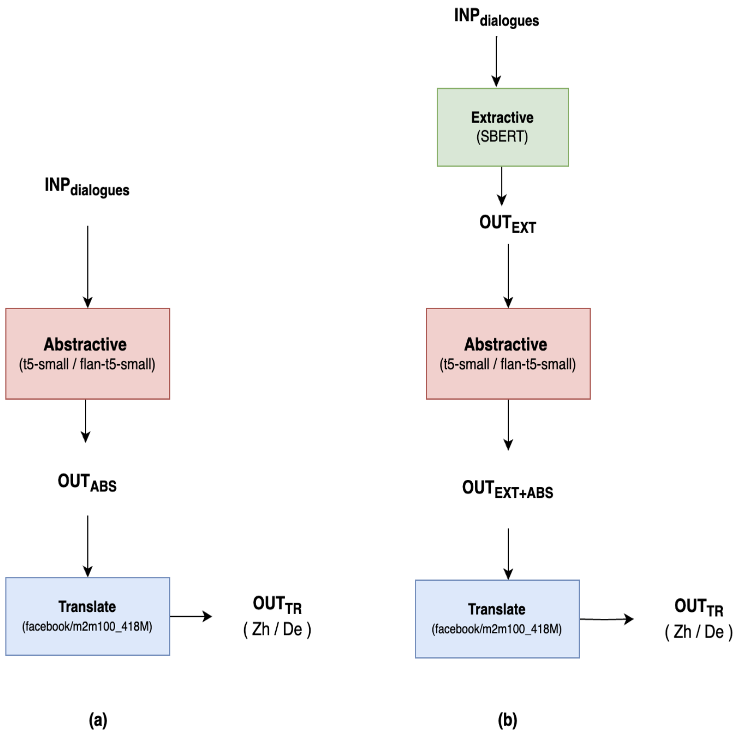

Figure 1: (a) Baseline Model: Plain abstractive model, (b) Combined model: extractive and abstractive model

- Dataset: The data set of dialogues we utilized had less sample (train: 20000, Validation: 10000). We struggled with fine tuning the parameters as they were few data points.

- Finding the optimal number of clusters proved to be an important factor in determining an accurate summary. This was because the clusters would influence the relevant components of the dialogue being extractied.

For extractive summarization, inorder to obtain the extractive summary we first tried to use SBERT and use K-Means. This proved to be inefficient because there is a different K-value (optimal number of clusters) for each dialogue. Hence we used *SBertSummarizer* which would calculate the optimal k value.

3 Experiments

3.1 Experimental Setup

3.1.1 Dataset

The dataset that we used to conduct our experiments is *GEM/xmediasum* (Wang et al., 2022) which is a cross-lingual dialogue summarization dataset with 40K English dialogues and their summaries in Chinese and German. The dataset was created by manually translating the English summaries of MediaSum to both Chinese and German.

3.1.2 Baselines

For our project as we wanted to show that if you use the combined output of an extractive summarizer and abstractive summarizer you would get a better summary and its translation than only using an abstractive summarizer. Therefore, we chose a model that would give us an abstractive summary to be our baseline. The model that we used for finetuning was *t5-small* and the model that we used for few shot approach is *flan-t5-small*.

3.1.3 Implementation Details

We implemented the proposed approach- extraction and abstraction combined, following which the summaries were translated. For this architecture we trained the model using two approaches- fine tuning and few shot. We compared this architecture with the baseline architecture of plain abstractive summarization, followed by translation. Figre 1 specifies the baseline and proposed architecture.

For extractive summarization in the proposed architecture, we experimented with three different ratios of dialogue-to-summary values, namely 0.25, 0.5, and 0.75.

3.2 Results

Analysis of fine-tuned approach: The first experiment performed was for the proposed architecture trained using a fine-tuned model. Table 1 shows that the combination of extractive and abstractive summarization performs better than the baseline plain abstractive model, with higher ROUGE scores. We can infer that using an extractive summarization before abstraction helps input relevant sentences to the abstractive model resulting in better summaries.

Summary Type	**R-1**			**R-2**			**R-L**		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
Abs	0.252	0.288	0.258	0.087	0.103	0.089	0.225	0.258	0.231
Ext+Abs	0.299	0.315	0.296	0.119	0.127	0.118	0.269	0.284	0.267

Table 1: Rouge Scores for model trained using *finetuning* approach

Ratio	**R-1**			**R-2**			**R-L**		
(Ext+Abs)	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
0.25	0.351	0.644	0.438	0.246	0.435	0.308	0.337	0.616	0.420
0.50	**0.467**	**0.671**	**0.533**	**0.360**	**0.524**	**0.412**	**0.454**	**0.652**	**0.519**
0.75	0.442	0.608	0.495	0.308	0.438	0.348	0.423	0.582	0.474

Table 2: Rouge Scores for model trained using the *few shot* approach with different sentence to summary ratios

Summary	**R-1**			**R-2**			**R-L**		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
Abs	0.410	0.557	0.457	0.266	0.377	0.300	0.389	0.530	0.434
Ext+Abs (r=0.5)	**0.467**	**0.671**	**0.533**	**0.360**	**0.524**	**0.412**	**0.454**	**0.652**	**0.519**

Table 3: Rouge Scores for model trained using *few shot* approach

Approach	**R-1**			**R-2**			**R-L**		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
Finetuning	0.299	0.315	0.296	0.119	0.127	0.118	0.269	0.284	0.267
Few Shot	**0.467**	**0.671**	**0.533**	**0.360**	**0.524**	**0.412**	**0.454**	**0.652**	**0.519**

Table 4: Rouge Scores for best combination in both approaches

Optimal dialogue-to-extractive-summary ratio: The second experiment was performed to evaluate the optimal dialogue-to-extractive-summary ratio. Table 2 shows that the best summaries were obtained when this ratio was set to 0.5, as compared to 0.25 and 0.75. We argue that this happens because a low value for the ratio removes important contexts from the dialogue which results in a poor summary; consequently, a high ratio value implies that the summary contains many irrelevant parts leading to redundant and poor summaries.

Analysis of Few-shot approach: Table 3 shows a comparison of ROUGE scores for few shot approach. We focus on training both the architectures (baseline and proposed) using a fine-tuning model, and as seen from Table 3, the extraction process before abstraction results in better performance (the optimal value of extraction ratio was chosen in the training experiment).

Analysis of extractive and abstractive architecture: Table 4 shows the results for the proposed architecture run using both approaches, fine-tuned and few shot. As seen from Table 4, the few shot model has superior performance.

4 Discussion

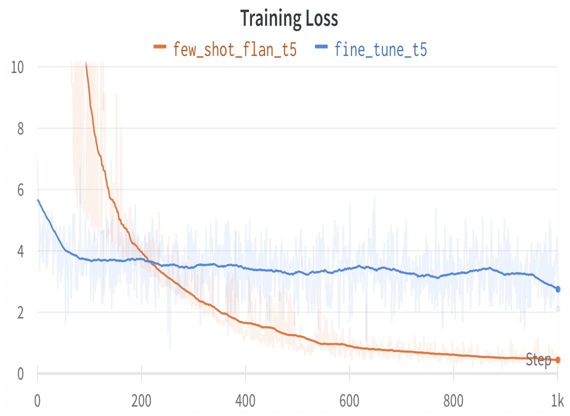

Figure 2: Training loss for first 1000 steps

- *Learnings*:
  - Adding an additional extractive step before performing abstraction on the dialogue summarization and translation task resulted in superior summaries being generated.
  - The utilization of sentence-BERT rather than BERT, ensuring that there is better generation of summarization as SBERT is better at finding the semantic text similarity.
  - The few shot model performs better than the fine-tuning approach, especially in a case where there is scarcity of data
- Future Extensions: Given more time, we would mod-

ify the architecture to obtain an end-to-end model which would result in superior performance. We would also train on much larger data set if we had access to more computational power.

References

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chenguang Zhu. 2021. Chapter8-applications and future of machine reading comprehension,machine reading comprehension elsevier.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization.