Automatic Slides Generation for Scholarly Papers: A Fine-Grained Dataset and Baselines (Student Abstract)

Sheng Xu, Xiaojun Wan

Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{sheng.xu, wanxiaojun}@pku.edu.cn

Abstract

Slides are broadly used to present the research works and there are several studies on the problem of automatic slides generation. However, the lack of dataset hinders further research. In this paper, we construct a benchmark dataset for the problem of slides generation from scholarly papers. The dataset is fine-grained and consists of aligned pairs of single slide and specific region of a paper. Then we deploy several baseline models and conduct preliminary experiments. The results show that this task is challenging and awaits more exploration. The dataset and code will be released.

Introduction

Nowadays, many researchers rely on slides to present their work in an effective and expressive way. However, it takes much effort to make slides from scratch. There are several prior studies on this problem, but automatic slides generation is still far from reach. Most slides have hierarchical structure. The bullet points of slides are placed in different levels according to their importance. Readers will focus on first-level (L1) bullet points at first sight. Then they can concentrate on the point they are interested in and further refer to its second-level (L2) points for detailed information. The hierarchical structure makes the automatic generation of slides difficult, because the model needs to generate content and structure simultaneously. For brevity, the structure of an L1 point and its corresponding L2 points is denoted as a cluster.

Most recently, Sun et al. (2021) propose a dataset for this problem. To tackle this task, they come up with a model, which needs users to input the slide title and then retrieve related sentences to form the slide. To our knowledge, this is the only dataset that is publicly available by now. However, they ignore the hierarchical structure of slides and regard the generation of slide text as a common text generation task.

In this paper, we construct a large benchmark dataset for this problem. The hierarchical structure of slides is retained. Previous works mainly focus on generating whole slides given a specific paper. However, it is more than challenging for the model to encode the long content of paper and then generate the whole slides. Thus we make our dataset fine-grained, which explicitly aligns each of the slide to the regions of its corresponding paper, based on similarity score. Then we implement several baseline models to tackle this task. We further design several evaluation metrics based on ROUGE (Lin 2004). Experimental results show the feasibility of our proposed models and the difficulty of this task.

Dataset

We crawl the homepages of researchers to get paper-slides data pairs. Most of documents are stored in PDF format. We use automatic processing tools to extract text content and manually correct the errors.

Then we try to align each slide to specific region of the paper. For each slide, we calculate its similarity score to each paragraph of the paper. We set a threshold of similarity score. Each slide is aligned to at most k paragraphs most similar to it if the similarity score is above the threshold. Some slides are filtered out. Here we set $k = 3$. By this way, we get fine-grained data pairs.

Finally, we acquire a dataset containing 12267 data samples. We randomly shuffle the samples and split 8/1/1 for training, validation and test. We also notice that about 56.37% of the bullet points are first-level and about 35.69% of them are second-level, adding up to about 92.06%. Since the first-level and the second-level points make up of most of the slides, we abandon third-level and subsequent lower-level bullet points and only focus on L1 and L2 points.

Preliminary Experiment

For preliminary experiment, we implement the following baseline models:

TextRank: TextRank is an unsupervised algorithm to calculate the importance of the sentences. The most important sentences are selected as L1 points until length limitation is reached.

Phrase-Based: We implement the phrase-based models proposed in (Wang, Wan, and Du 2017). Their models generate slides in an extractive manner and take phrases as the minimum unit. Specifically, they extract phrases from the paper. Then a model is trained to evaluate the importance of the phrases. Another model is used to estimate the hierarchical relationship of two phrases. If the output of the relationship model is larger than threshold, the model will regard

Metric	Hierarchy			Align			Penalty		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
TextRank	9.89	2.09	8.65	7.77	1.55	6.73	8.20	1.63	7.19
Saliency-First	9.62	2.04	9.21	7.40	1.69	7.08	7.78	1.51	7.46
Relation-First	9.77	2.10	9.35	7.62	1.72	7.26	7.97	1.59	7.63
Transformer	5.78	1.34	5.21	3.54	0.87	3.20	5.76	1.34	5.20
Two-Stage	**13.45**	**4.11**	**12.73**	**11.60**	**3.72**	**11.02**	**10.15**	**3.41**	**9.65**
Oracle	20.53	6.25	18.19	17.26	5.02	15.09	20.53	6.25	18.19

Table 1: ROUGE scores of different models over different metrics.

one phrase to be L1 point and the other to be its L2 point. Finally, they deploy two greedy search algorithms, namely **Salieny-First** and **Relation-First** to attain the slides. The former attaches more importance to the first model and vice versa. We make the improvement by replacing their random forrset classifiers with LSTM based models.

Transformer: We also try to generate a slide in a traditional seq2seq manner. The model takes the sentences of paper as input. The target sequence is acquired by concatenating the bullet points with special tag token $\langle L1 \rangle$ or $\langle L2 \rangle$ depending on their levels. We further employ copy mechanism to handle the Out-Of-Vocabulary problem.

Two-Stage: It is another abstractive summary model. The language styles of L1 and L2 points are different. The expressions in L1 points tend to be general while L2 points are more detailed and specific. Thus we deploy two modules to generate different parts. Given the text from paper, the first module will generate L1 points of the slide. After this, the second module takes paper text, together with L1 point as input, then generates corresponding L2 points. On the other side, the content of L1 and L2 points are semantically related, so we share the parameters of several components between these two modules. The model is based on transformer and also makes use of copy mechanism.

Oracle: in addition, we implement extractive oracle models. Given a specific evaluation metric, it extracts sentences from the source input to maximize the ROUGE-1 score. The performance of oracle can be considered the upper bound of extractive summary models.

We design several automatic evaluation metrics based on ROUGE score. Given a generated slide and a gold slide, we first evaluate the matching degree of the clusters. As for cluster, we calculate the ROUGE score of L1 point and L2 points respectively, then use their weighted average to evaluate the matching degree. We attach different importance to L1 and L2 points, and assign different weight to them. After getting scores of clusters, we use them to finally get the ROUGE score of slides. We generally follow this idea, and make some adjustment to get different metrics, namely: (1) **Hierarchy**, each cluster of generated slide will be aligned to the one of gold slide who has the largest ROUGE score, and the final score of the slides is the average score of clusters. (2) **Align**, which adds the constraint that each cluster from the reference slide can only be aligned at most once. (3) **Penalty**, which is similar to Hierarchy, but adds penalty term for slides whose number of clusters exceeds gold slides. Please refer to complementary material for more details.

Experimental Results

The performance of different models is demonstrated in table 1. Of all baseline models, Two-Stage performs the best. It takes the hierarchical structure of slides into consideration and adopts neural network based models. The transformer model performs not quite well, mostly because it neglects the hierarchical structure of slides and try to generate slides in a seq2seq way. It seems that the model performs badly on learning the sequence representation of slides. TextRank performs close to Saliency-First and Relation-First, although it is a simple extractive model and ignores the hierarchical structure of slides. The perform of Saliency-First and Relation-First is close, while the latter performs slightly better. Our further analysis indicates that these two models perform not good enough on the metrics mainly because the bullet points in their generated slides are much too short. Thus they score relatively low. It seems that phrase is much too fine-grained for the bullet points of slides.

Oracle is the upper bound performance that could be achieved with extractive methods. It outperforms all other baseline models by a large margin, although some models are abstractive. It shows the difficulty of this task and that there is still much space for improvement.

Conclusion

In this paper, we construct a fine-grained benchmark dataset for the problem of automatic slides generation from academic papers. The dataset emphasizes the hierarchical structure, which is a key feature of slides. We conduct experiments on the dataset with several baseline models. The results verify the feasibility of summarization models. However, the performance is still not satisfying enough by now. The problem still awaits further exploration.

References

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Sun, E.; Hou, Y.; Wang, D.; Zhang, Y.; and Wang, N. X. 2021. D2S: Document-to-Slide Generation Via Query-Based Text Summarization. *arXiv preprint arXiv:2105.03664*.

Wang, S.; Wan, X.; and Du, S. 2017. Phrase-based presentation slides generation for academic papers. In *Thirty-First AAAI Conference on Artificial Intelligence*.