Automated Presentation Generation System using Design Documents

Author
Sneha Bandi
Kirthika Gurumurthy
Tejas Kashinath
Surya Roshan Mugada
Rhythm Girdhar

Group 40
snehaban@usc.edu
kgurumur@usc.edu
tkashina@usc.edu
mugada@usc.edu
rgirdhar@usc.edu

Abstract

The purpose of design documents is to outline the goals, specifications, and requirements to design a framework or product. It ensures that the whole team working on a specific product is on the same page during the product development cycle, that no information is lost, that there is no lack of ambiguity, coherent insights are derived and the design and development process is efficient. In addition, we can also keep track of other metrics, such as time spent, issues, and bugs during the implementation, which will give us valuable insights into tracking the progress of the project. The aim of this project is to automate the process of generating presentations from design documents, which can increase the operational efficiency of analyzing the goals, requirements, etc. of large amounts of data from the reports, and also makes it easier to present the key information in the design document in a concise, succinct format.

1 Project Domain and Goals

The goal of this project is to automate the process of generating presentations from design documents. We aim to summarize the text in the design documents and present them in concise readable format in the presentation. Another aspect the project aims to tackle is automated FAQ generation from the corpus provided, to easily answer any common questions asked. The generated presentation would include features like sectional summarising (Goals, methodology, challenges, technology), visualizations, conclusion and FAQs. The intended users are developers and project managers, who can be significantly benefited from using the presentations not only to save time by obtaining the extracted requirements, goals, and other sections from the design documents corpus but also can utilize the presentation generated for collaboration with other team members and managers. As design documents are primarily written in human language, natural language processing poses a good solution to this problem.

2 Related Work

Existing works based off of our complete, proposed application include - generating presentations from general documents and generating presentations from research papers (Qiu and Xiong, 2019). Our proposed application is intended to generate clear, cogent presentations specific to software design documents which will include extracting specific sections as described in the previous section, generating visuals and flow charts for applicable components - such as flow charts for methodology, gantt charts for timeline etc. Our proposal intends to select information which are relevant to the cause of the project containing the design document which includes the business needs for such a project, any profits, timeline and budgeting while also extracting the in-depth technical information regarding the project. A novel addition in our application is to include commonly asked questions and their corresponding answers for ease of understanding during collaboration and another source to deliver information which would not be added to the presentation. Current related works (Wein and Briggs, 2021) use extractive text summarization (Akay and Kim, 2021) in research articles, whereas our proposal intends to use both extractive as well as abstractive methods. To summarize/ extract the relevant information in each slide we intend to explore specific methods for certain sections - extractive Q/A methods for sections such as requirements, goals etc., abstractive Q/A methods for sections such as methodology, design etc.

3 Datasets

For extractive summarization and Q/A for sections such as requirements, goals etc (and FAQ generation):

The dataset being used are primarily a combination of 2 datasets. The first dataset used is the corpus of Natural Questions and Answers [1] , collected by Google AI which consists of 323000 examples which are the annotated questions and answers from wikipedia pages. The second dataset is the TechQA dataset [2] by IBM which has 800000 technical documents that address specific technical questions and their answers (Kwiatkowski et al., 2019). The proposal intends to apply standard preprocessing in addition to combining the two datasets to get a common format.(Duan et al., 2017)

For the Abstractive information extraction and Q/A: The dataset considered is the ELI5 dataset [3] which consists of user asked questions and answers obtained from various users replicating the abstractive question and answering from 3 popular sub-reddits covering general, science and historical questions

4 Technical Challenge

The main principle involved would be to design a Question Answer framework comprising each of the sections (goals, requirements, methodology, timeline issues, bugs) phrased as the questions. Given a context, the question must be answered succinctly by either using extractive or abstractive methods. Sections such as requirements, goals, etc. that must be reproduced as is can utilize the extractive method of Q/A while sections such as methodology, issues, etc. can be presented using the abstractive methodology of Q/A. For extractive question answering, a neural network such as BERT (Devlin et al., 2019) can be trained on the dataset (which can be generated manually containing contexts/documents, questions, and answers (references in the Datasets section can be used) to perform extractive Q/A with the required input. For sections requiring abstractive Q/A deep neural networks can be trained as encoders and decoders to map input text to an output sequence. Appropriate preprocessing can be carried out before feeding the input to these models - such as TD-IDF, POS tagging, etc. The task would essentially be to perform transfer learning for the domain of design documentation.

While text summarization has been well researched, the task of automatically generating slides in domains of software is relatively new. Our intended approach is to implement a question based summarization approach. Abstractive summarization is challenging and benefits more from domain knowledge and language patterns to generate outputs.

One of the limitations of the proposed application is data collection and/or generation. The model needs to be trained on design documents and retrieving specific design documents could be challenging as they could be protected documents. Due to a lack of data in the same domain, we propose implementing a transfer learning approach.

Another technically challenging aspect of our application is FAQ generation. Questions need to be framed appropriately for each section with relevant technical answers.

The evaluation we are performing is two-fold: How coherent the model is, i.e, how well the language model has learned to produce sentences that are accurate and readable and how well the model represents the document that we are condensing into a presentation.

For the first part, we propose the use of BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Both metrics compare how well the model generated reference compares to the human generated reference. BLEU is focused on precision and ROUGE is focused on recall. In essence, we can also calculate a single F-1 score for our model.

In addition to metric driven evaluation, we can also add a human evaluation component, similar to A/B testing, wherein we present a machine generated sample of our output alongside a human generated sample, and ask a human to assign a score or a preference to the results.

If this application was put into use it would be highly beneficial to managers and developers to automate the summarization of certain key sections such as requirements from a huge corpus of design documentation and use the output-generated presentation as an aid for collaboration in projects.

[1] https://ai.google.com/research/NaturalQuestions

[2] https://leaderboard.techqa.us-east.containers.appdomain.cloud/index.hbs

[3] https://huggingface.co/datasets/eli5

5 Division of Work

The project highlights two important sections - Text summarization and FAQ generation. As a team we plan on dividing the tasks equally with Kirthika, Roshan and Rhythm working on Text summarization, and Sneha and Tejas working on FAQ generation. Internally the individual tasks for achieving these goals are laid such that steps like data acquisition, preprocessing, modelling, training, evaluation and testing are evenly divided between all team members, to ensure fair understanding and learning. We also plan on reviewing each others work and giving feedback to each other.

References

Haluk Akay and Sang-Gook Kim. 2021. Extracting functional requirements from design documentation using machine learning. *Procedia CIRP*, 100:31–36. 31st CIRP Design Conference 2021 (CIRP Design 2021).

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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and M. Zhou. 2017. Question generation for question answering. In *EMNLP*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jiazuo Qiu and Deyi Xiong. 2019. Generating highly relevant questions. *CoRR*, abs/1910.03401.

Shira Wein and Paul Briggs. 2021. A fully automated approach to requirement extraction from design documents. In *2021 IEEE Aerospace Conference (50100)*, pages 1–7.