

Patent Document Summarization with Contextual Embeddings

*Are you interested in making a big impact with your thesis?
Work with us on an innovative approach for patent summarization.*

Patents drive innovations by enabling international organizations to protect their inventions from a legal perspective. Consequently, these documents are important resources that describe inventions. Due to the rapid growth of the number of available patent documents, manually analyzing such data is beyond human capabilities. Therefore, text summarization as an initial step to enable more efficient processing of the documents became a necessity. There exist several summarization models and they seem to perform well. However, most of them focus on conventional documents, e.g., news articles. In addition, the existing summarization models are either abstractive or extractive [1,2]. While the extractive summarization models aim to extract the most descriptive sentences from a given document, the abstractive summarization approaches generate phrases or sentences that may not appear in the original document.

In contrast to the existing approaches, this thesis aims to design a patent summarization model which has the ability to generate summaries that are similar to the Derwent World Patents Index (DWPI)'s patent abstracts. These patent abstracts are manually written by several patent domain experts and help users to understand the inventions and their key technical details without going through the entire documents [3]. In order to avoid such a costly task, the goal of the thesis is to propose a fully automated patent summarization approach that uses contextual embedding models (e.g., BERT), and applies both extractive and abstractive summarization techniques to generate summaries like the Derwent abstracts. The model will be trained with the patent documents for which Derwent abstracts are available. The generated summaries can be utilized as a starting point for different tasks, e.g., patent classification as well.

This thesis will be supervised by **Prof. Dr. Harald Sack, Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe.**

[1] <https://arxiv.org/pdf/1909.03186.pdf>

[2] <https://arxiv.org/pdf/2004.08795.pdf>

[3] <https://www.sciencedirect.com/science/article/abs/pii/S1474034619306007>

Which prerequisites should you have?

- Good programming skills in Python
- Interest in Natural Language Processing
- Interest in Deep Learning technologies

Contact person:

Rima Türker

rima.tuerker@fiz-karlsruhe.de