**T5 Model Documentation**

**Table of Contents**

[1. Introduction: 2](#_Toc154519969)

[2. T5 model Structure 2](#_Toc154519970)

[3. Applications 3](#_Toc154519971)

[3.1 Text Summarization 3](#_Toc154519972)

[3.2 Language Translation 3](#_Toc154519973)

[3.3 Text Classification: Textual Entailment 3](#_Toc154519974)

[3.4 Linguistic Acceptability 3](#_Toc154519975)

[3.5 Sentence Similarity 3](#_Toc154519976)

[4. T5 Pretrained dataset 4](#_Toc154519977)

[4.1 Scope 4](#_Toc154519978)

[4.2 Format 4](#_Toc154519979)

[4.3 Use in NLP 5](#_Toc154519980)

[5. T5 Model Variants 5](#_Toc154519981)

# 1. Introduction:

The Text-to-Text Transfer Transformer, commonly known as T5, is a powerful natural language processing (NLP) model. The T5 Transformer Model was introduced in 2020 by the Google AI team and stands for Text-To-Text Transfer Transformer (5 Ts, or, in our case, T5). The main problem T5 addresses is the lack of systematic studies comparing best practices in the field of NLP. Most of the current SOTA models are derived from the Transformer architecture. The transformer was introduced in the legendary paper “Attention Is All You Need” by Vaswani et al. and had two main architectural blocks, namely Encoders and Decoders. This documentation aims to provide a comprehensive understanding of the T5 model, its architecture, applications, and how to use it for various tasks.

# 2. T5 model Structure

The T5 transformer model works by using the same standard encoder-decoder structure as standard transformer models. It consists of **12-pair** blocks of encoder-decoder. Each block contains self-attention, a feed-forward network, and optional encoder-decoder attention.

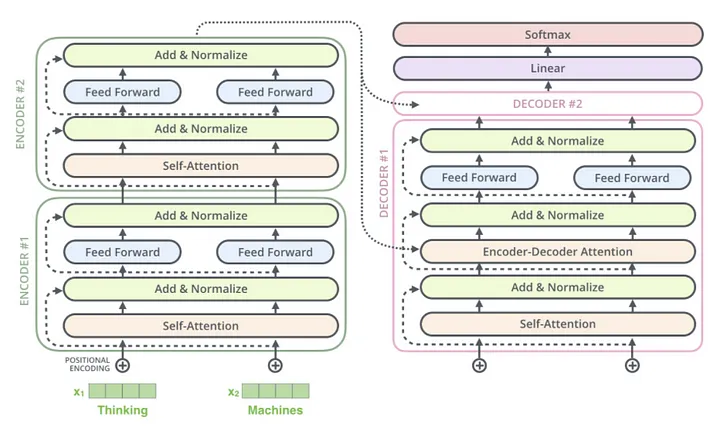


Figure 1. Transformer Structure

# 3. Applications

T5 has demonstrated exceptional performance across various applications. T5 (Text-to-Text Transfer Transformer) model is a versatile natural language processing (NLP) model that has been applied across various tasks. Here are some notable applications of the T5 model:

## 3.1 Text Summarization

Text Summarization is the NLP task in which a model, given a long text sequence, produces a summarized version of the input.

For summarizing we need to add the “summarize:” prefix to the input sequence.

## 3.2 Language Translation

Language Translation is the NLP task in which a model, given a text in one language, produces translated version of the same text in another language. The T5 model was trained on the C4 dataset, which contains the following languages: English, German, French, and Romanian.

Using T5, we can translate between these languages. Below we are going to translate from English to French. For translation, we need to add the “translate English to French: ” prefix to the input sequence. Our prompt: “translate English to French: You should definitely watch 'One Piece', it is so good, you will love the comic book.”

## 3.3 Text Classification: Textual Entailment

Textual Entailment is a NLP task in which a model is given two sentences, one being the premise and the other being hypothesis. Based on these two sentences, the output is classified into three classes: entailment, contradiction, and neutral. For textual entailment, we need to add “mnli premise: ” and “hypothesis: ” to the sentence pairs. Our prompt: “mnli premise: I love One Piece. hypothesis: My feelings towards One Piece are filled with love.”

## 3.4 Linguistic Acceptability

Linguistic Acceptability is an NLP task in which a model, given a text prompt checks if the sentence is grammatically correct. For linguistic acceptability, we need to add “cola sentence: ” to the sentence. COLA is the dataset that contains sentences mapped to their acceptability.

Our prompt: “cola sentence: Luffy is a great pirate.”

## 3.5 Sentence Similarity

Sentence similarity is an NLP task in which a model, given two sentence pairs would rate their similarity on a 1 to 5 scale. The output is considered a string value and is found to be incremented by 0.2. This means we can consider this as a text classification task with 21 classes: 1.0, 1.2, … 5.0. For sentence similarity, we need to add “stsb sentence 1: ” and “sentence 2: ” to the sentence pairs.

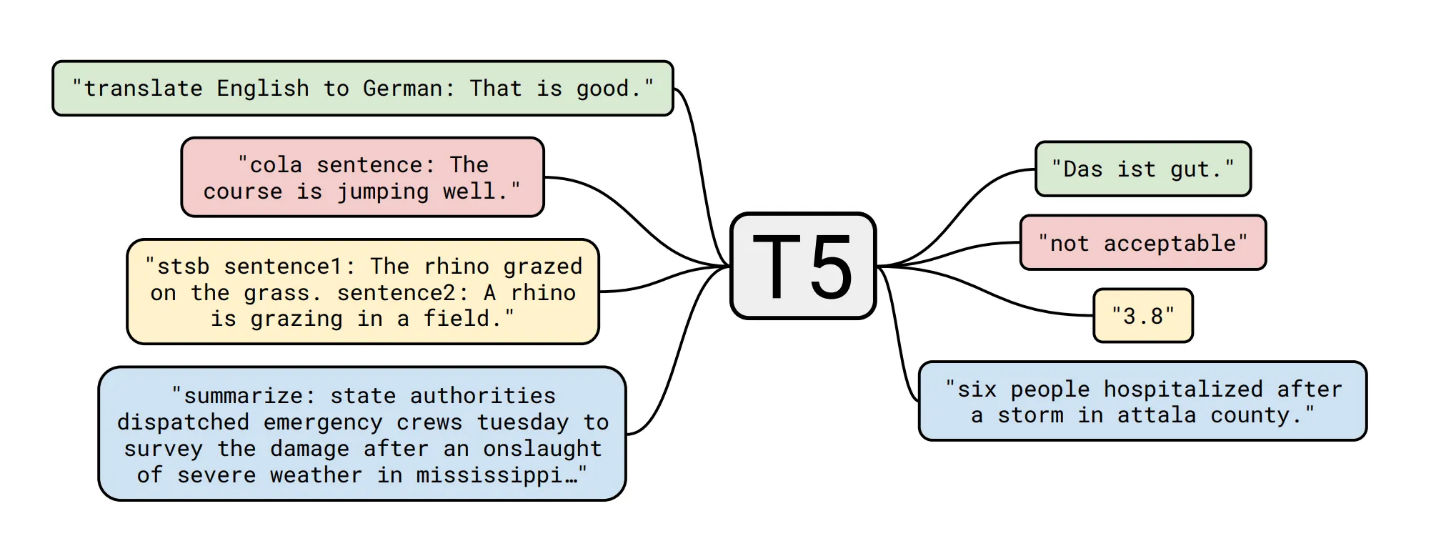


Figure 2. Applications of T5 Model

# 4. T5 Pretrained dataset

T5 uses common crawl web extracted text. The authors apply some pretty simple heuristic filtering. T5 removes any lines that didn’t end in a terminal punctuation mark. It also removes line with the word java-script and any pages that had a curly bracket (since it often appears in code). It deduplicates the dataset by taking a sliding window of 3 sentence chunks and deduplicated it so that only one of them appeared the dataset. For example, above 3 pages, the last paragraph on the middle page is removed since the same content appears on the first page. It ends up with **750 gigabytes** of cleanish English text.

The **C4 (Common Crawl Corpus)** dataset is a large-scale web corpus that contains web pages collected by the Common Crawl initiative. The Common Crawl project is dedicated to crawling and archiving web pages to make the web's content freely accessible to the public and facilitate research in various domains, including natural language processing (NLP) and machine learning.

Here are some key points about the C4 dataset:

## 4.1 Scope

The C4 dataset includes a diverse set of web pages from across the internet. It is designed to be comprehensive and cover a wide range of topics and languages.

## 4.2 Format

The data is typically provided in a raw format, and researchers may preprocess and extract the content based on their specific needs. It may include HTML content, metadata, and other information.

## 4.3 Use in NLP

Researchers and practitioners often use the C4 dataset for various NLP tasks, such as training language models, pre-training models for transfer learning, and large-scale web text analysis.

# 5. T5 Model Variants

There are 5 T5 variants with varying parameters and model sizes.

1. **Base:** Comparable to that of BERT-base. It is a baseline model with **222 million** parameters.
2. **Small:** It is a scaled-down version of the Base model. It only has **60 million** parameters with only 6 layers of encoders and decoders.
3. **Large:** Scaled-up version of the base with **770 million** parameters.
4. **3B:** Scaled up version of the base with **3 billion** parameters.
5. **11B:** Scaled up version of the base with **11 billion** parameters.

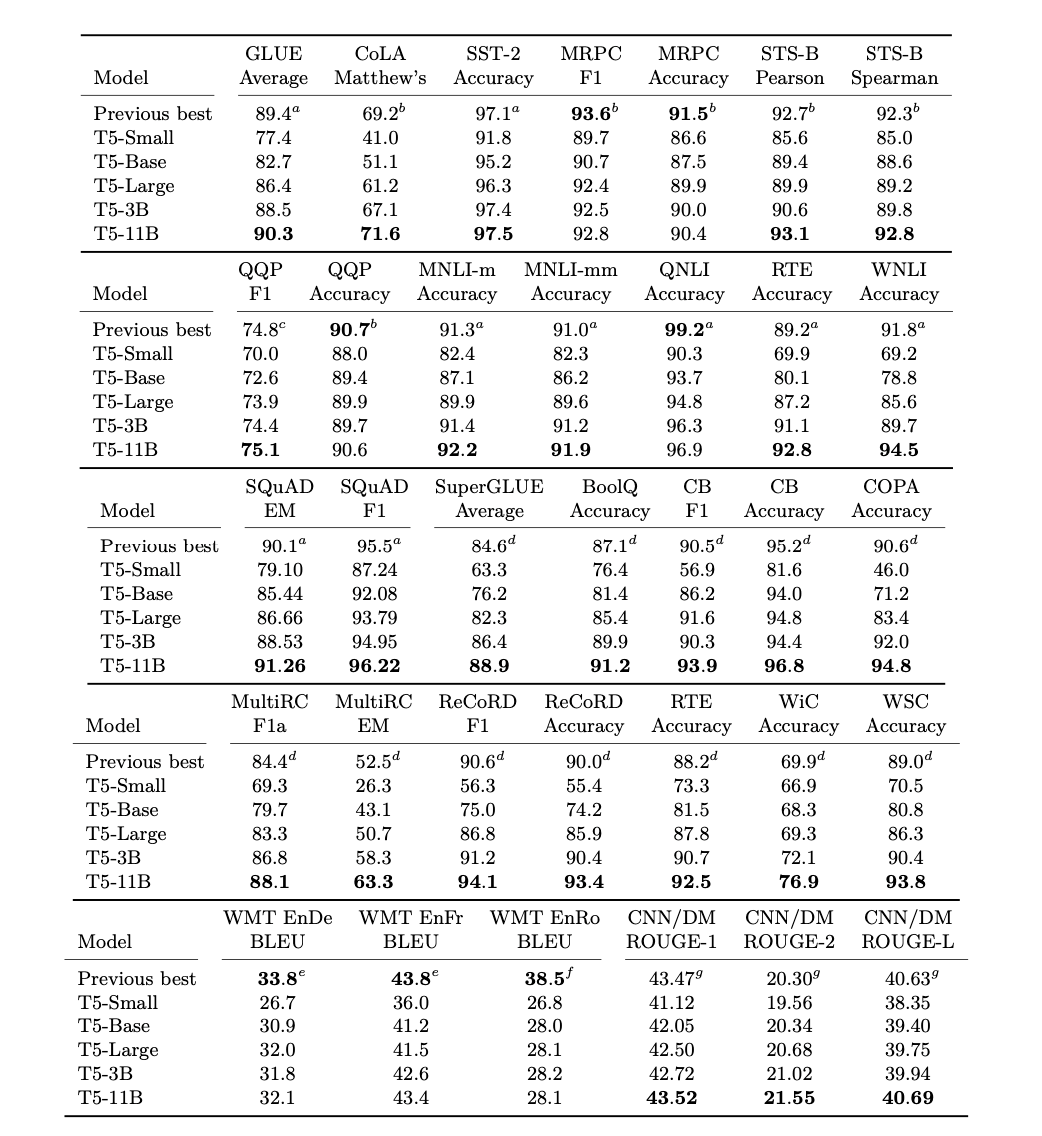


Figure 3. Performance of T5 Model Variants