

Lecture:

Natural Language Processing

Chapter 01: Intro

Tim Metzler, Jörn Hees

2024-04-09



**Hochschule
Bonn-Rhein-Sieg**
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

What is NLP?

What is NLP?

focus in this lecture

“**Natural language processing (NLP)** is an interdisciplinary subfield of **computer science and information retrieval**. It is primarily concerned with giving computers the ability to support and manipulate **human language**. It involves processing **natural language datasets**, such as **text corpora** or **speech corpora**, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "**understanding**" the contents of documents, including the contextual nuances of the language within them. To this end, natural language processing often **borrowes ideas from theoretical linguistics**. The technology can then accurately **extract information** and **insights** contained in the documents as well as **categorize** and **organize** the documents themselves.”

Source: https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=1215529997

What are common NLP tasks?

- Text Classification / Sentiment Analysis / Moderation Systems
- Summarization
- Text Generation / Autocomplete / Recommendation
- Assistant systems
- Translation
- Search / Retrieval / QA
- Speech to Text
- Entity Recognition (linking to Knowledge Bases)

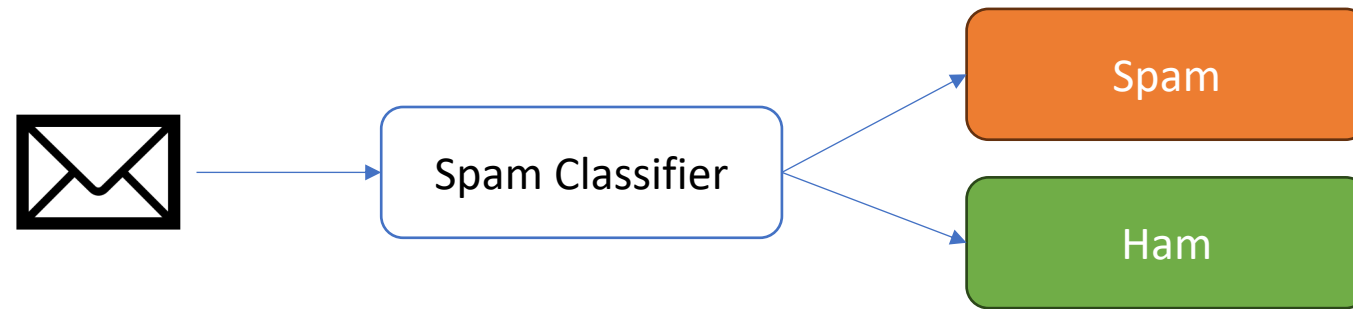
What are common challenges in NLP?

- Ambiguities / Homonyms
- Computation
- Speech 2 Text losses: Informal Speech / Filler Words, Utterances
- Vectorization / Representation (flexible input lengths, vocab sizes...)
- Typos
- Dataset sizes
- Languages, Character Sets, Writing styles, Accents
- Hallucinations
- Explainability
- Biases in datasets / need for diverse datasets

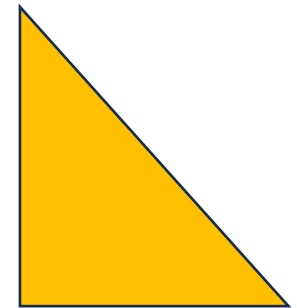
Application Areas, Tasks & Examples

Text classification

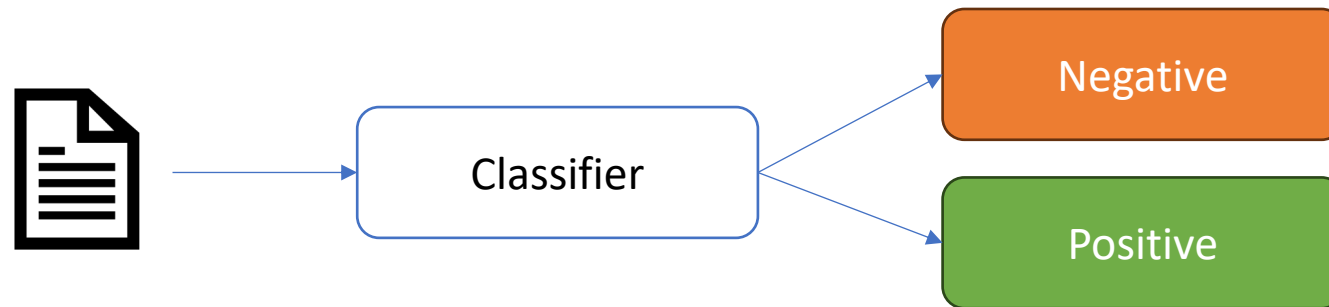
- Spam classification



Often also as
scoring /
regression variant



- Text Sentiment Analysis



Text classification: Sentiment Analysis

Hosted inference API ⓘ

Text Classification

Examples ▾

This was the most horrible experience ever!

Compute

Computation time on cpu: 0.049 s

1 star

0.969

2 stars

3 stars

4 stars

5 stars

</> JSON Output

Hosted inference API ⓘ

Text Classification

Examples ▾

The food was mediocre at best. The staff was friendly though.

Compute

Computation time on cpu: 0.151 s

1 star

0.046

2 stars

0.385

3 stars

0.522

4 stars

0.044

5 stars

0.003

</> JSON Output

Maximize

Hosted inference API ⓘ

Text Classification

Examples ▾

This movie was great. I watched it 3 times already!

Compute

Computation time on cpu: 0.058 s

1 star

0.004

2 stars

0.003

3 stars

0.013

4 stars

0.106

5 stars

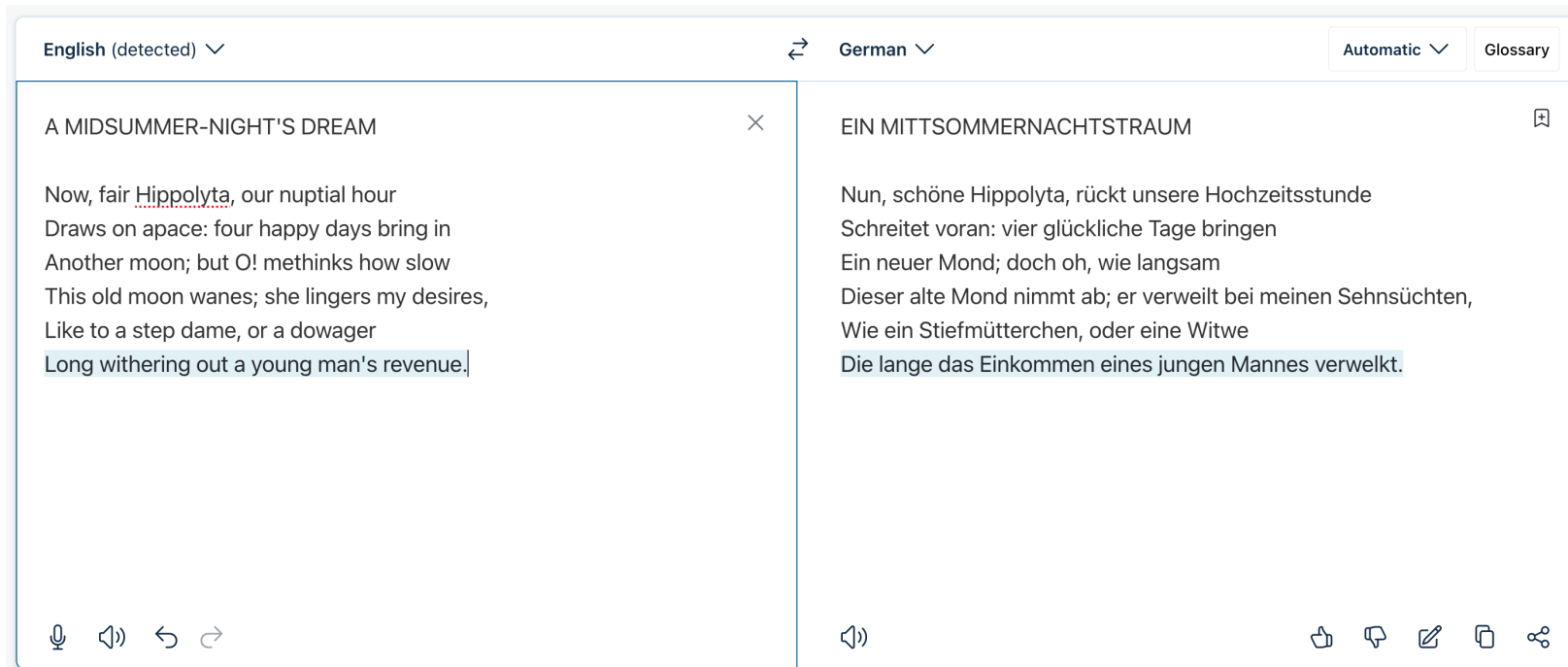
0.874

</> JSON Output

Maximize

Machine Translation (MT)

- Google Translate (translate.google.com)
- DeepL (www.deepl.com)



Keyword Extraction

- Extract the most important phrases (keywords, key phrases) from a document
- Token classification

⚡ Hosted inference API ⓘ

🔍 Token Classification

Examples ▾

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Compute

Computation time on cpu: 0.081 s

Natural language processing LABEL_1 (LABEL_0 NLP LABEL_1) is a subfield of linguistics, computer science, and LABEL_0 artificial intelligence LABEL_1 concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of LABEL_0 natural language LABEL_1 data. LABEL_0

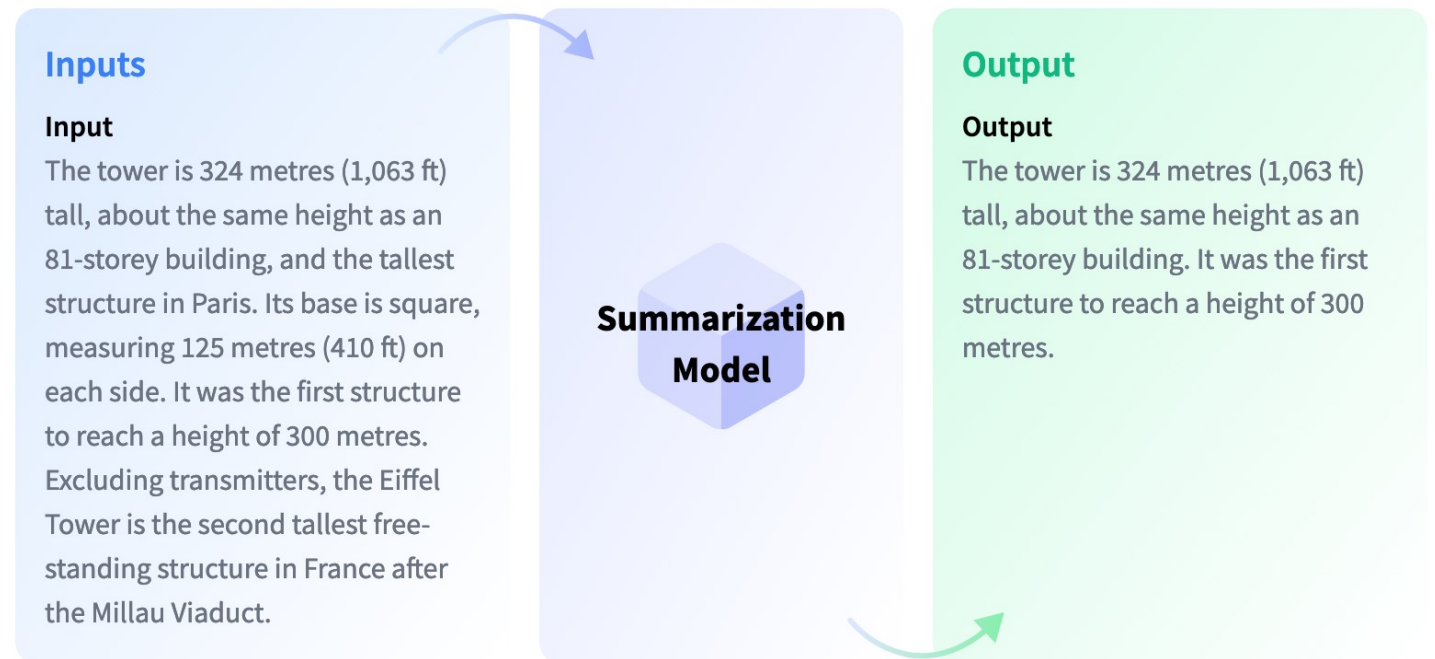
</> JSON Output

🖼 Maximize

Source: <https://huggingface.co/jasminejwebb/KeywordIdentifier>

Text Summarization

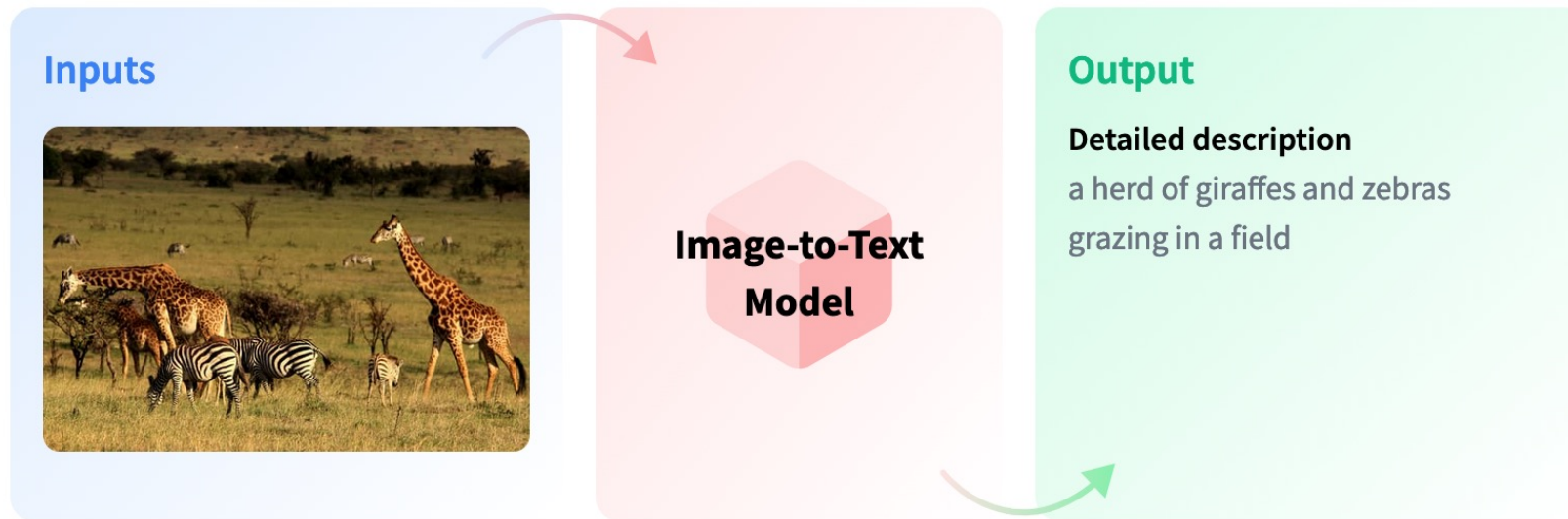
- Produce a shorter version
- Preserve important info



Source: <https://huggingface.co/tasks/summarization>

Image Captioning

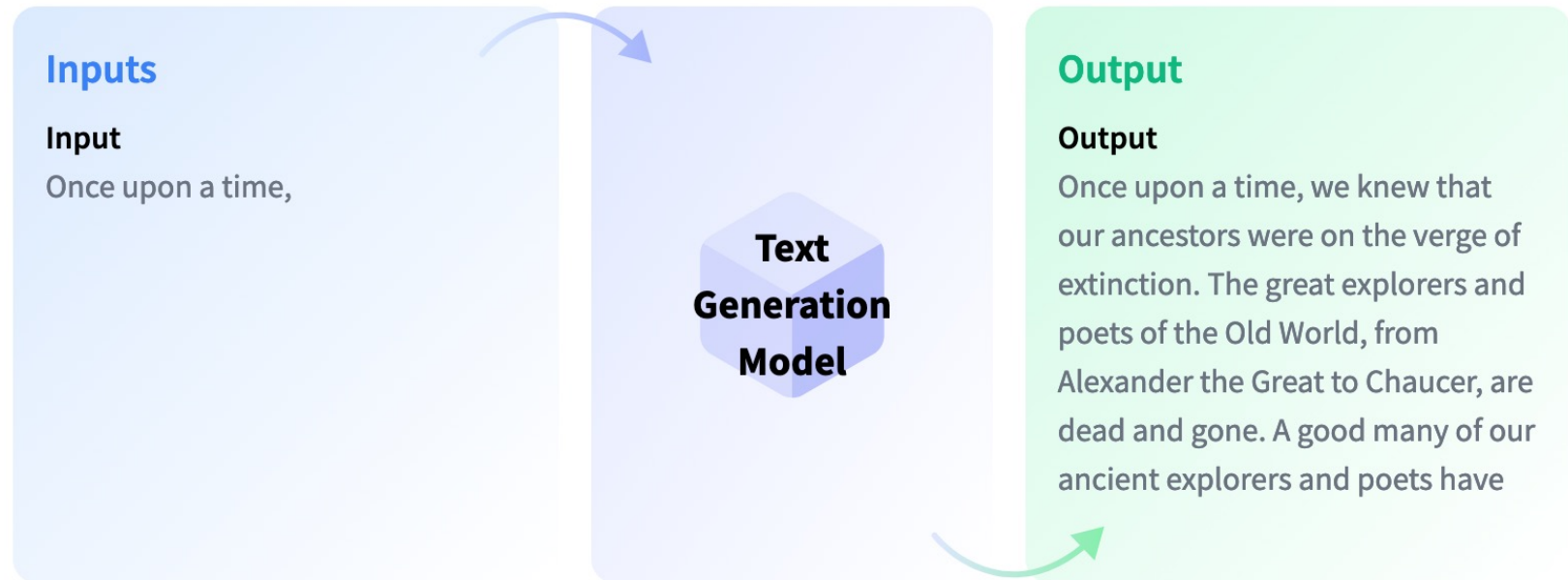
- Describe contents of an image
- Generate a title for an image



Source: <https://huggingface.co/tasks/image-to-text>

Text Generation

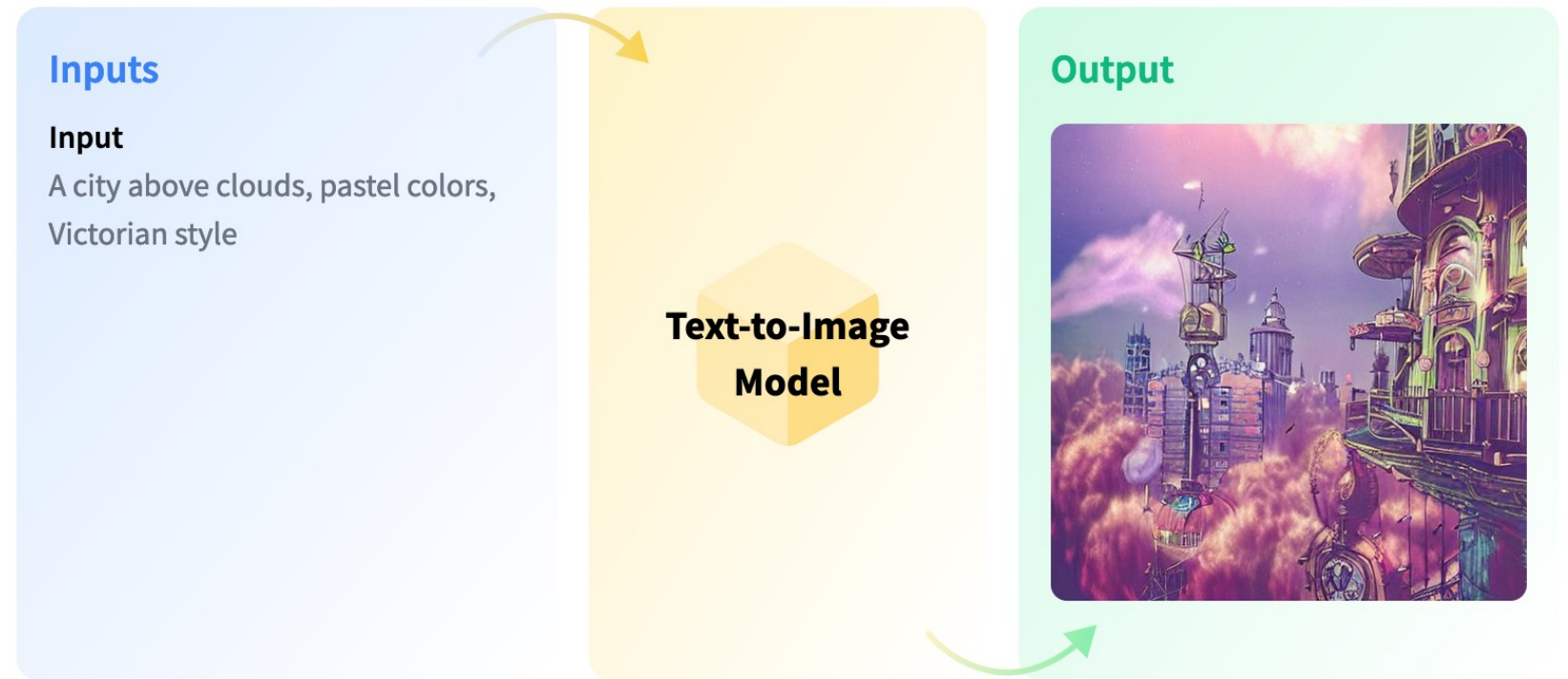
- Text continuations
- Can also be conditioned
 - Context
 - Topic
 - Contents
 - Questions
 - Language
 - ...



Source: <https://huggingface.co/tasks/text-generation>

Image Generation from Text (T2I)

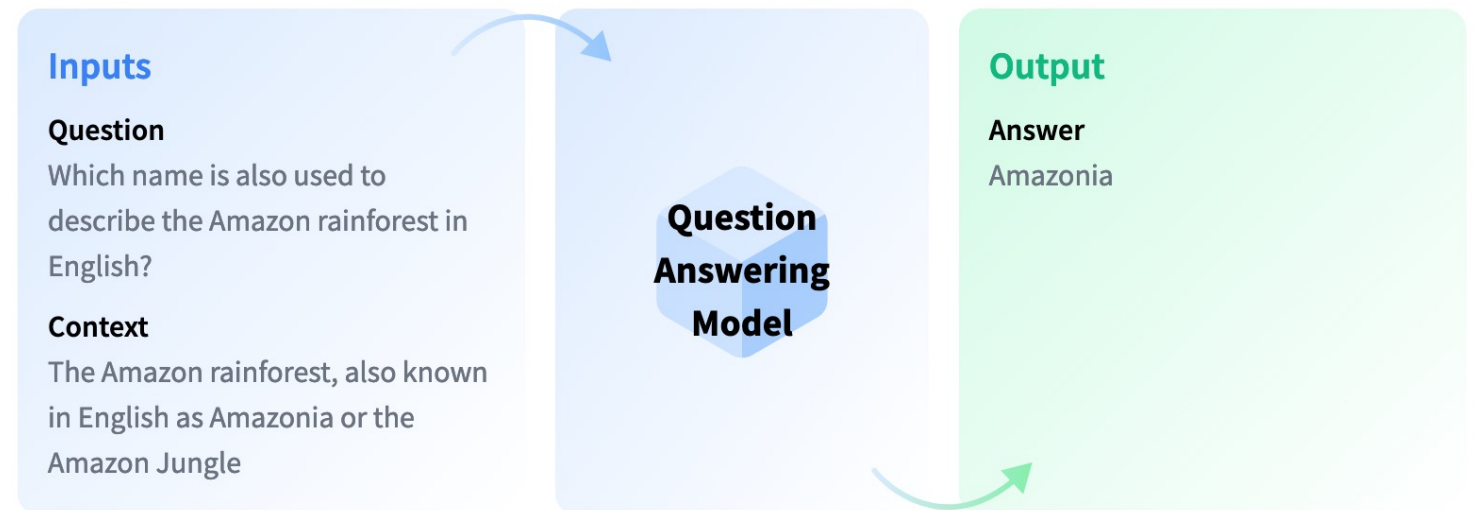
- Text to Image
- Also interactive



Source: <https://huggingface.co/tasks/text-to-image>

Question Answering

- Ask questions about text
- Get answers
- Variants:
 - Relevant passage given
 - Corpus based
 - General purpose model



Source: <https://huggingface.co/tasks/question-answering>

“Chat Bots”

- IRC / Discord
 - Bot Users / Chat integration
 - Often keyword / rule based
- Intent Recognition Systems:
 - Customer support (the annoying things on websites / phone hotlines)
 - (Air Canada Incident! <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>)
 - (To some degree after Speech 2 Text) Alexa, Cortana, Google Assistant, Siri
- Chat assistant / conversational AI systems
 - ChatGPT, Gemini (Bard), Copilot, Claude, ...
 - General Task Interfaces

Next

Text Processing