Lecture:
Natural Language Processing

Chapter 01: Intro

Tim Metzler, Jörn Hees

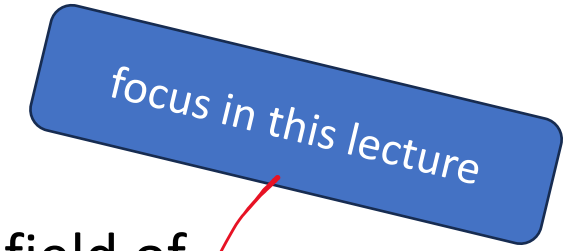
2024-04-09

Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

What is NLP?

What is NLP?



“**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 **borrows 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?

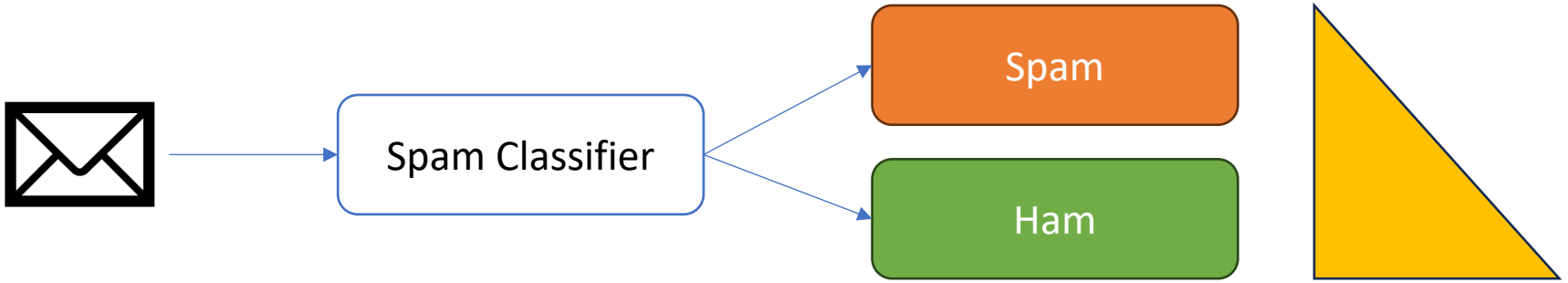
- Ambiguities / Homonyms
- Computation
- Speech 2 Text losses: Informal Speech / Filler Words, Utterances
- Vectorization / Representation (flexible input lengths, vocab sizes…)
- Typos
- Dataset sizes
- Languages, Character Sets, Writing styles, Accents
- Hallucinations
- Explainability
- Biases in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

Application Areas, Tasks & Examples

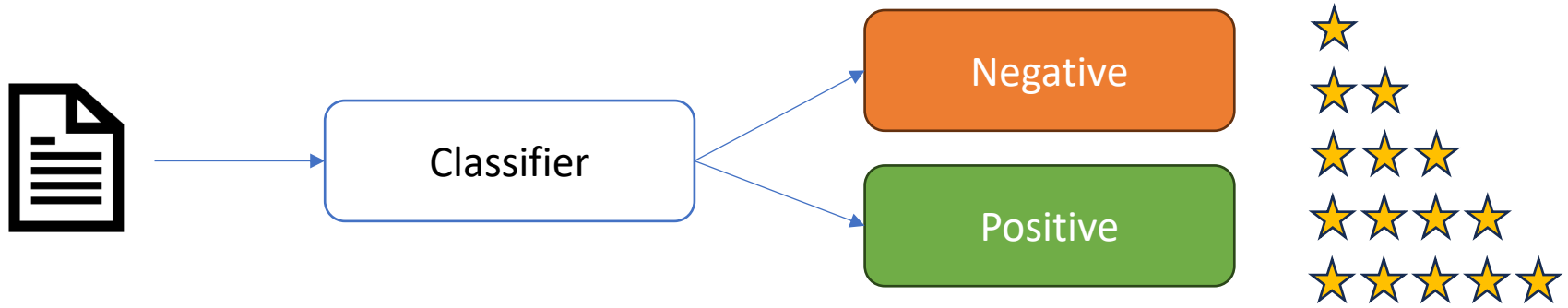
Text classification

Often also as scoring / regression variant

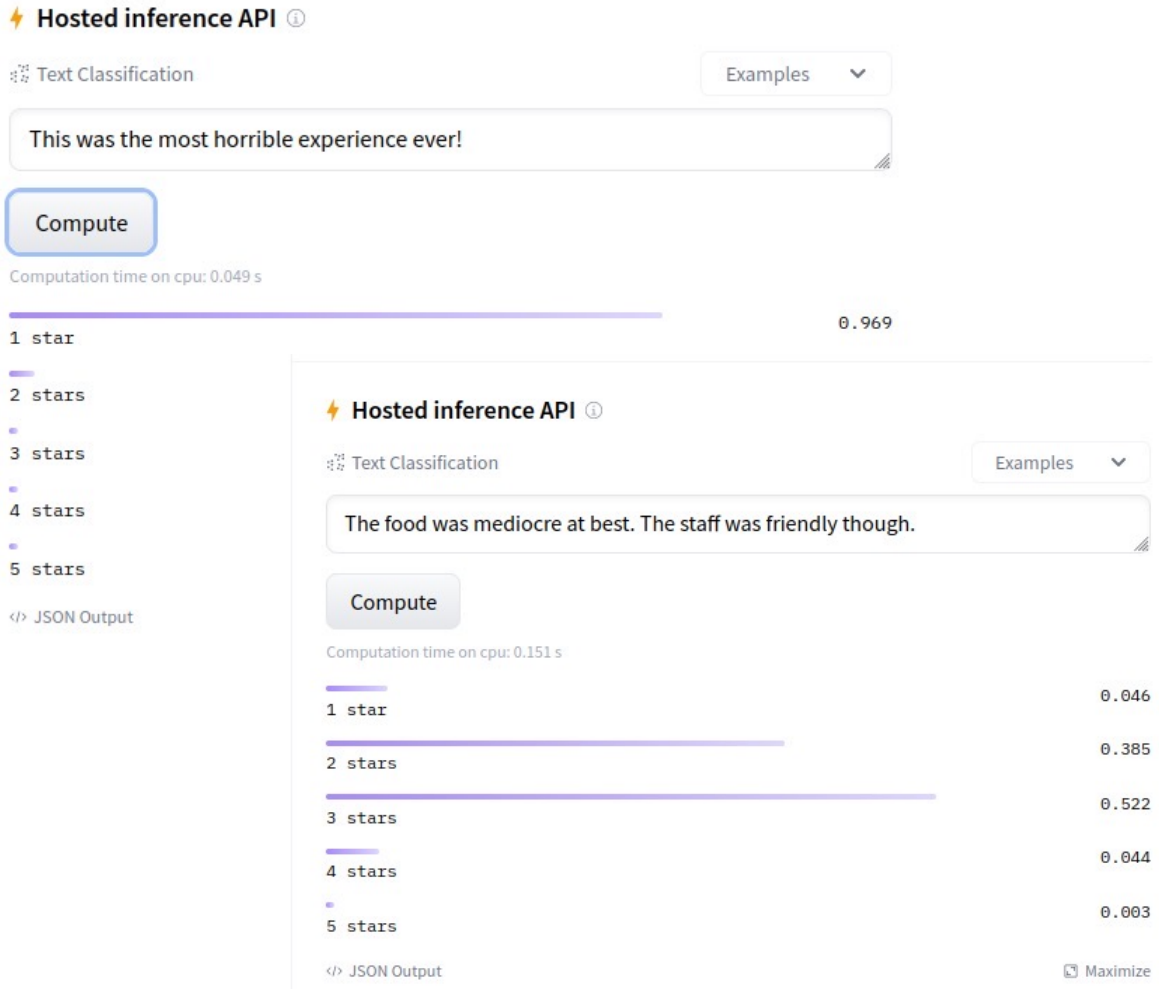
- Spam classification

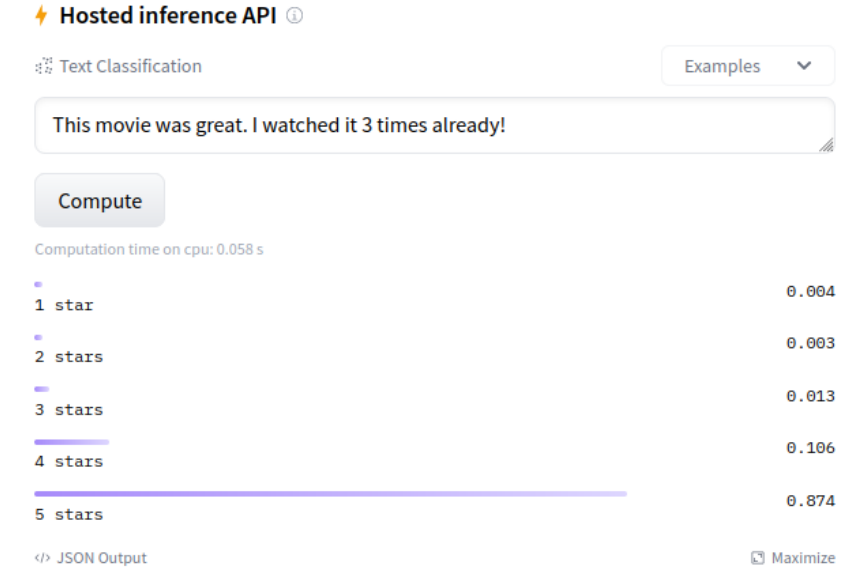


- Text Sentiment Analysis



Text classification: Sentiment Analysis

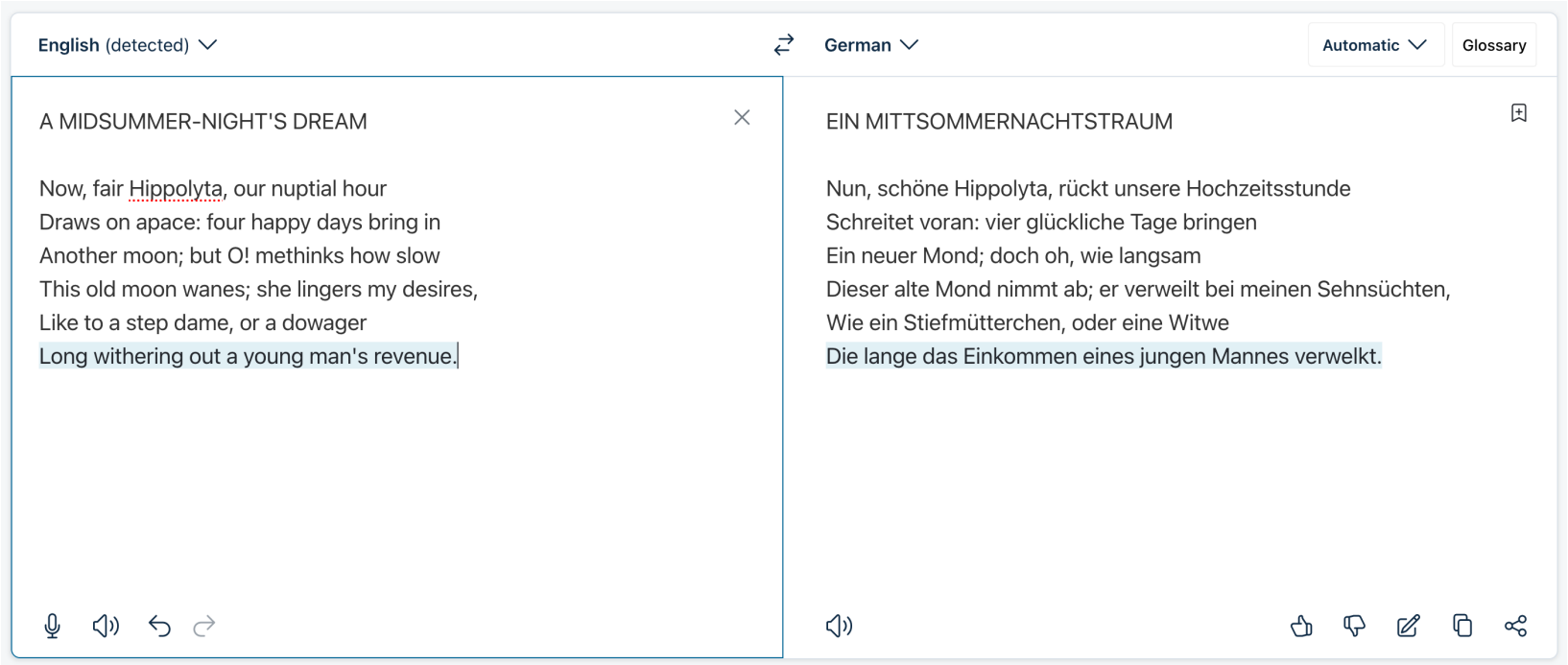




Source: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

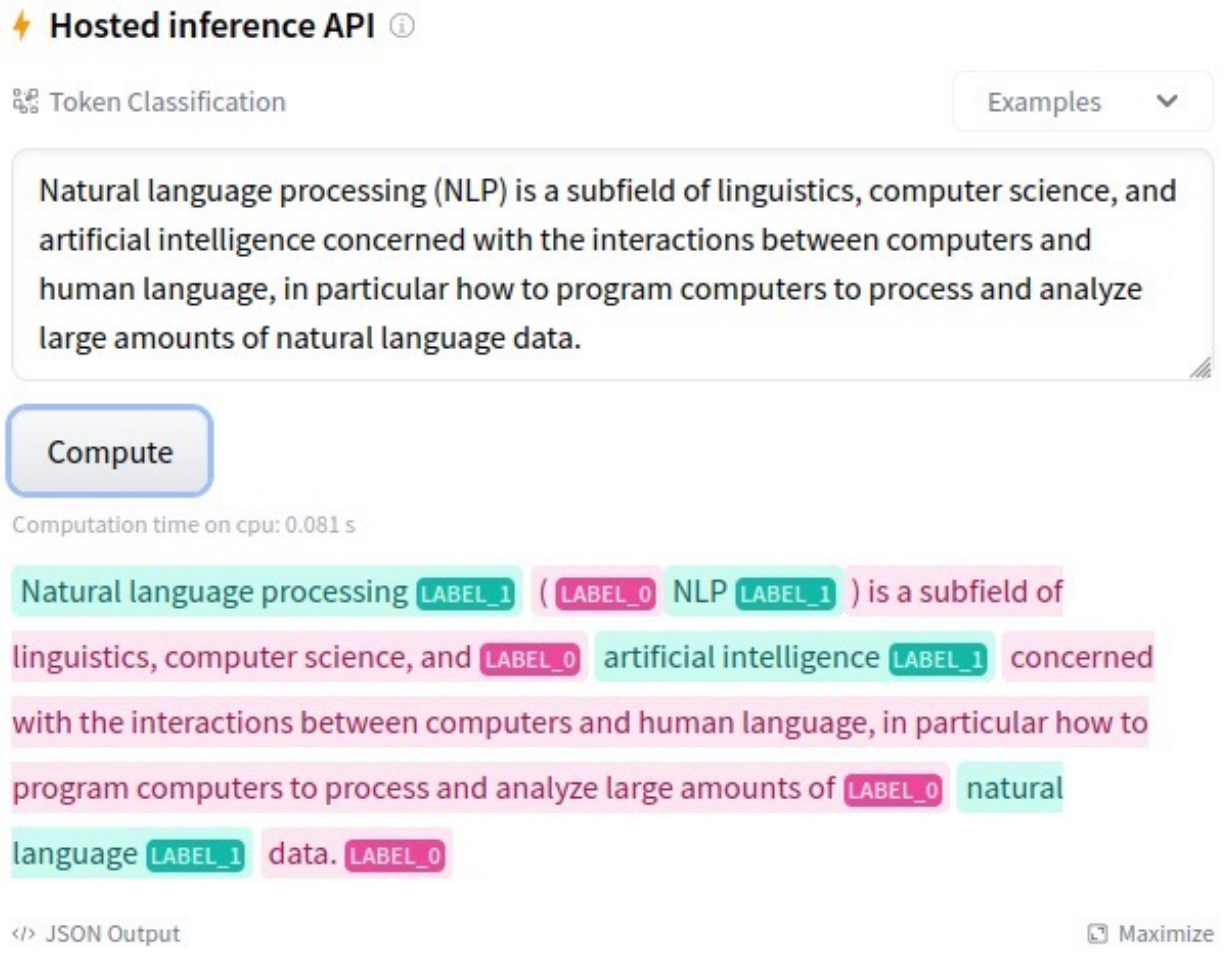
Machine Translation (MT)

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



Keyword Extraction

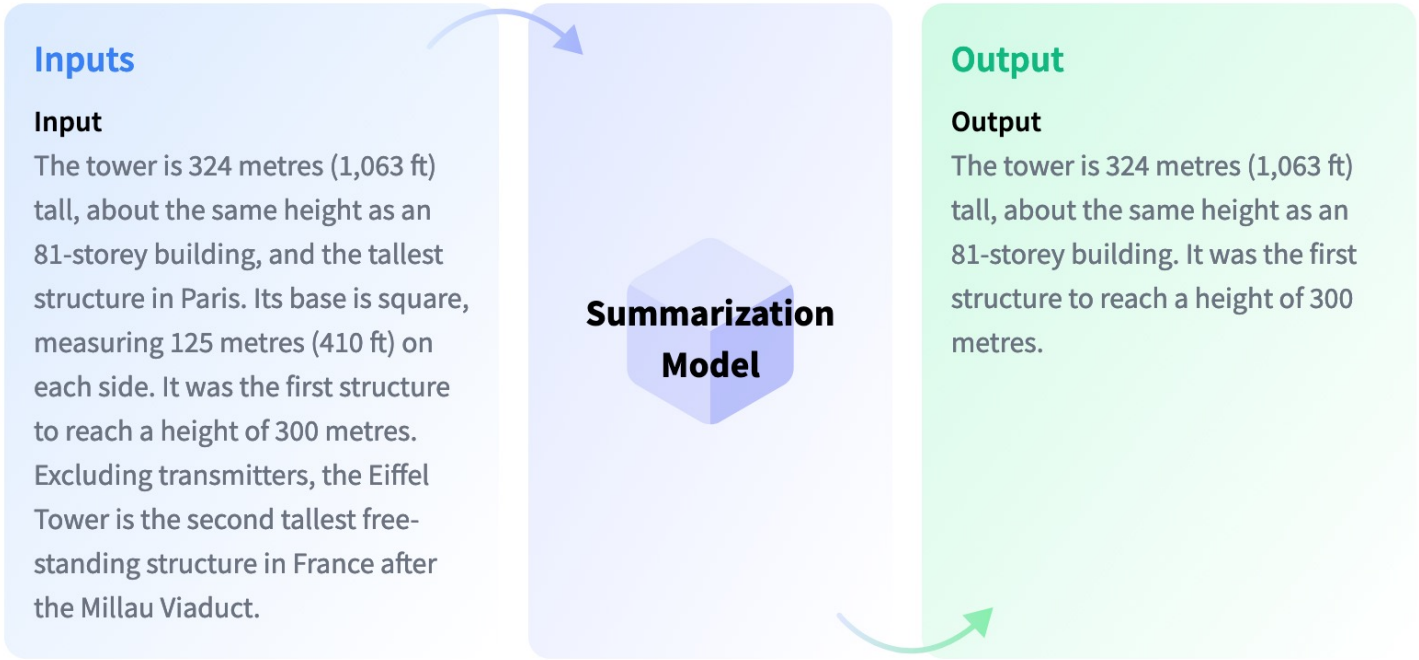
- Extract the most important phrases (keywords, key phrases) from a document
- Token classification



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

Text Summarization

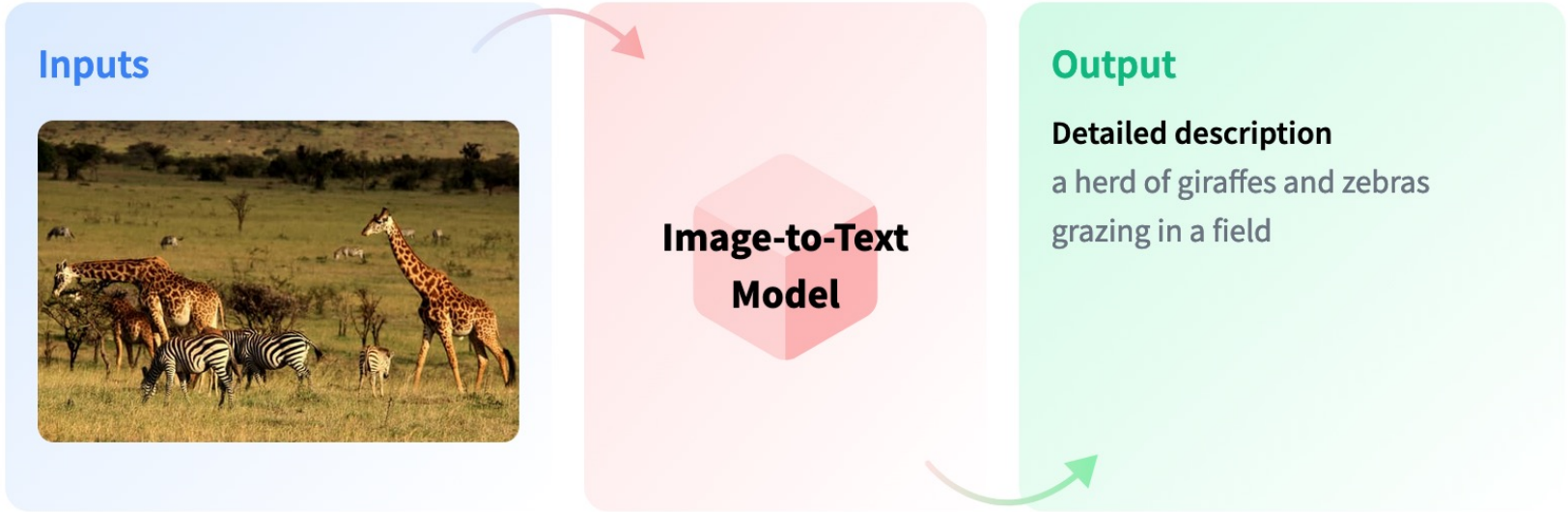
- Produce a shorter version
- Preserve important info



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

Image Captioning

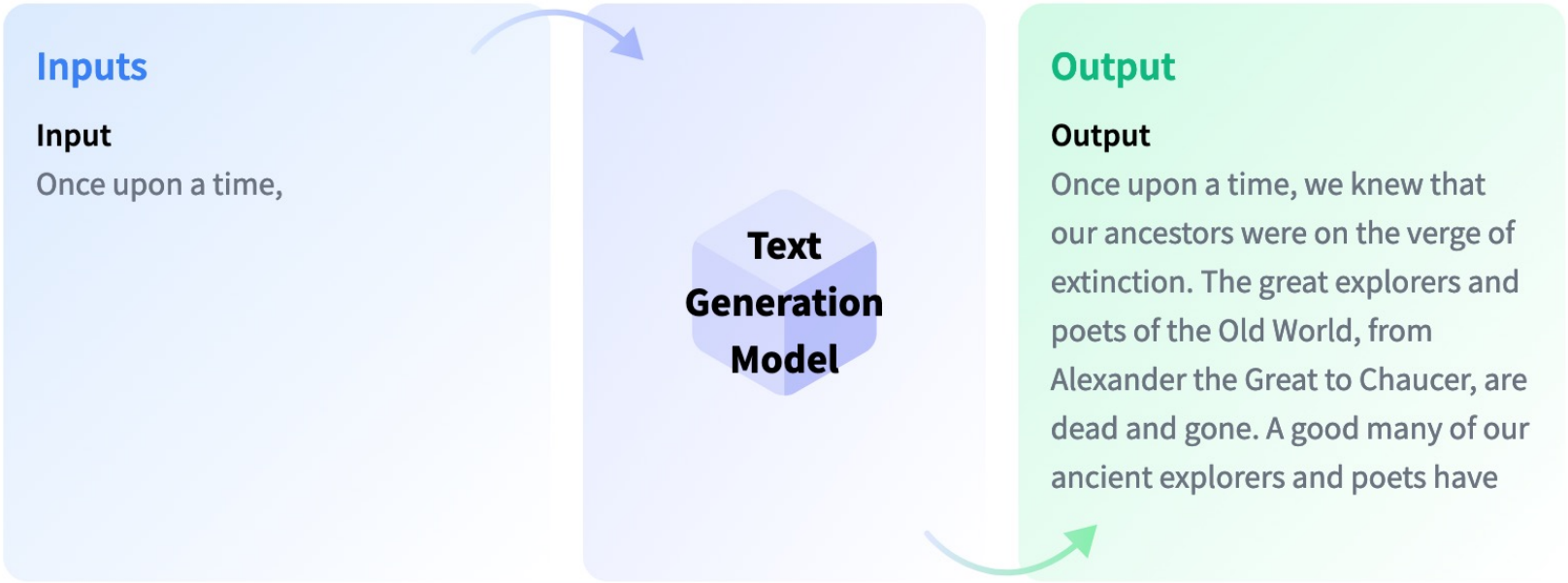
- Describe contents of an image
- Generate a title for an image



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

Text Generation

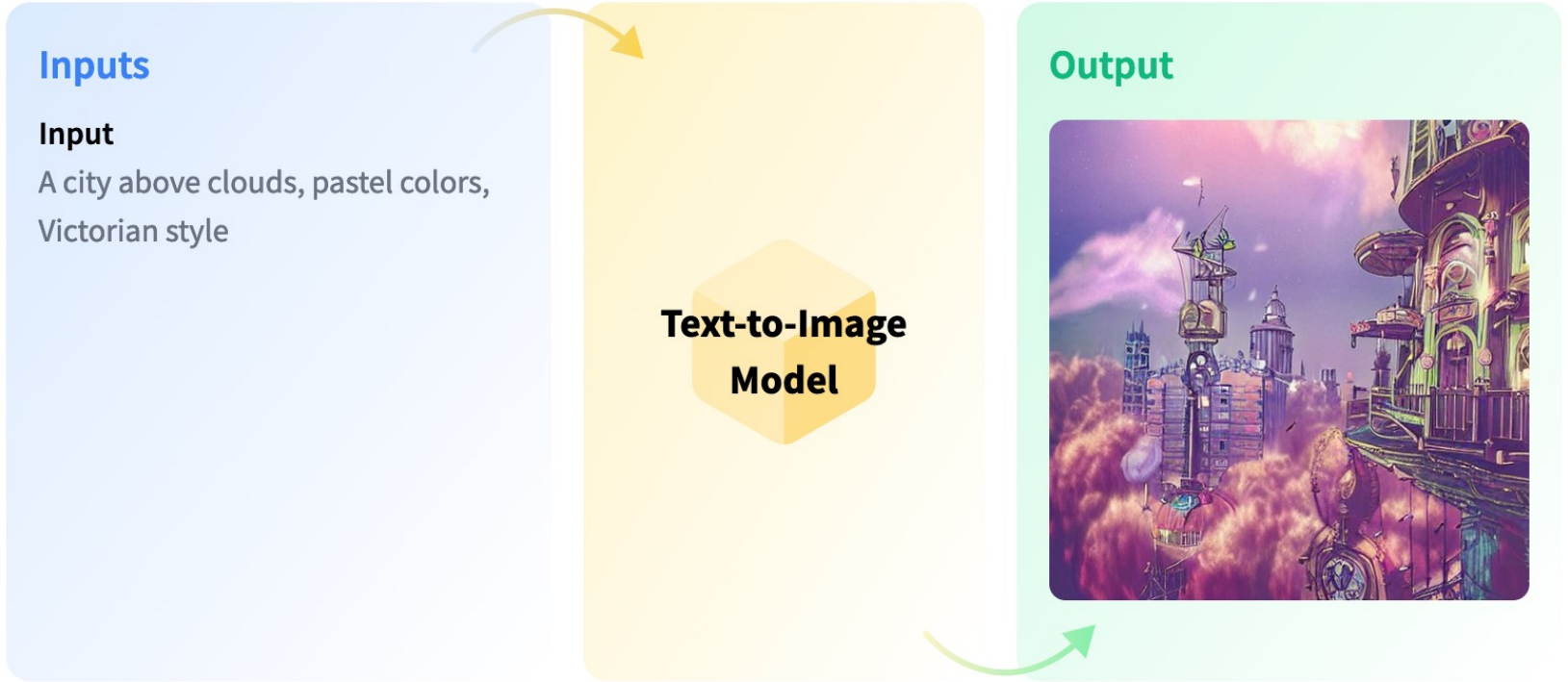
- Text continuations
- Can also be conditioned
  - Context
  - Topic
  - Contents
  - Questions
  - Language
  - …



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

Image Generation from Text (T2I)

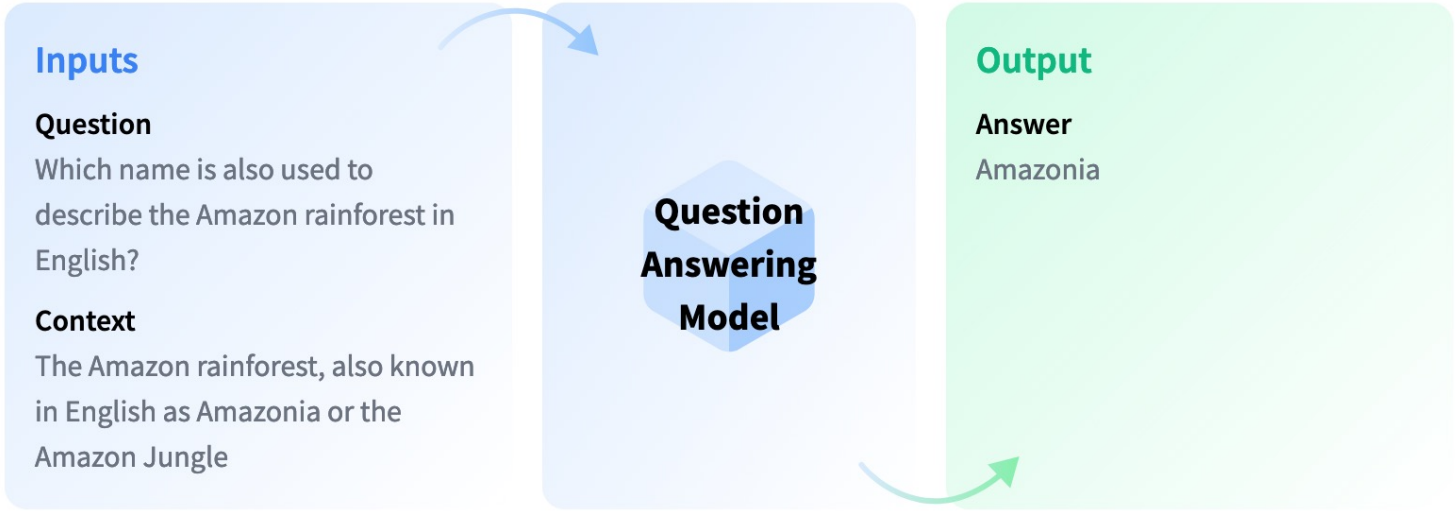
- Text to Image
- Also interactive



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