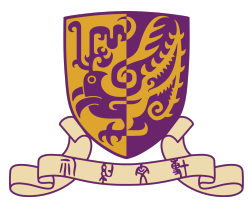
LING3401 Linguistics and Information Technology

Tutorial: Machine Translation, Speech Synthesis, and Multimodality

Yige Chen

The Chinese University of Hong Kong

April 9, 2025



Machine Translation

- Translates text or speech from one language to another

Google Translate

- https://translate.google.com
- Supports over 100 languages
- Uses neural machine translation (NMT) based on deep learning
  - This means it is not based on large language models
- Pros: fast, consistent, and reliable translations
  - Output is stable: same input gives nearly the same output every time
- Cons: limited understanding of nuance, genre, and style
  - Cannot be easily adapted for specific domains

Machine Translation with LLMs

- Large language models can perform translation without explicit training
- Use prompts to translate between languages
- Can handle code-switching and multilingual input
- Allows customization based on genre, tone, or speaker persona
  - For example, you can ask the model to translate as if spoken by a historical figure or in a formal tone

Have a Try!

- The Kingdom of Pigland has just received several communications from its close neighbor, the Republic of Duckland – a news article, a presidential address, and an interview. Pigland's officials want to translate these texts into various languages so that all pigs, including those who don't speak English, can understand them. Can you help?
  - If you speak a language other than English, you may ask Google Translate and LLMs to translate the texts into another language (e.g., Cantonese, Mandarin, Classical Chinese, etc.)
  - Otherwise, I do have a version of the texts prepared in Korean
  - Try including a specific genre, style, or persona if you are using LLMs
- After translating, please compare your versions and submit the one you believe best preserves the context to the Royal Pig Translations Committee.

Multimodality

- Combines text, audio, image, and/or video in one model
- Models can describe images, generate videos, speak texts, etc.
- Supports tasks like visual question answering and narration
- Helps bridge communication across sensory modalities

Speech Synthesis

- Converts written text into spoken audio
  - Also known as Text-To-Speech (TTS) as long as the input is textual
- Early systems used concatenative synthesis
- Modern systems use deep learning for natural speech

Text-To-Speech/Audio Tools

- Google Cloud TTS provides a wide range of voices and languages
  - https://cloud.google.com/text-to-speech
- GPT-SoVITS allows few-shot (1 minute) or zero-shot (5 seconds) voice cloning for personalized TTS
  - https://github.com/RVC-Boss/GPT-SoVITS
  - I believe one of the groups will present this tool, so I won't go into detail here to avoid spoilers!
- Suno can generate songs from text prompts (e.g., lyrics)
  - https://suno.com
- Some TTS models are also available on Poe

Have a Try!

- If you have access to a TTS or audio generation tool, try generating speech from one of the three sample texts
- Feel free to play the audio out using your speaker – I don't mind

About Image

- The recent GPT-4o image generation excels at generating images given textual prompts, with or without image input
  - Unfortunately, OpenAI's service is not yet available in Hong Kong
  - You may want to wait for it to be available on Poe (not sure if that'll happen)
- Otherwise, Poe is linked to some other good image generation models, such as DALL·E 3
- Additionally, some multimodal LLMs that output text can take images as input, such as GPT-4o (not the image generation one, just the regular GPT-4o)

Have a Try!

- Ask an LLM to generate a prompt for image generation based on a news article
- If you have access to an image generation model, try using the prompt to create an image representing the news
  - Try including a specific style or visual theme in your prompt
- If you don't have access, feel free to submit your prompt – I'll try generating the image for some of you

About Video

- To be honest, I am not very familiar with video generation from text
- But there are definitely multimodal LLMs that support it
- You may want to explore video generation models on Poe yourself
- They typically take textual prompts as instruction, and sometimes an image as input

Integrated in a Pipeline

- Some companies provide a pipeline of services integrated into an easy-to-use interface
- These are often branded as personal AI assistants
- For instance, Google's Gemini
  - Unfortunately, it is not available in Hong Kong
- Alternatively, if you are a Chinese speaker, you may have heard of ByteDance's Doubao

Miscellaneous

- Please do not hesitate to ask questions
- We enjoy feedback from you, so please let us know if you feel there's anything we could have done better