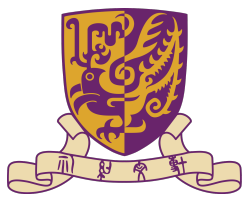
LING3401 Linguistics and Information Technology

Tutorial: Demonstrating NLP tasks and applications

Yige Chen

The Chinese University of Hong Kong

January 8, 2025



Before We Start...

- This is the tutorial session of **LING3401 Linguistics and Information Technology**
- You are now at **Lee Shau Kee Building 302**
- Make sure you are in the right tutorial session

About Me

- Yige Chen
- Third-year PhD student in Linguistics at CUHK
- Research interests: Natural language processing
  - Information extraction in economics, finance and business
  - Incorporating linguistic knowledge into NLP models and tasks
  - Effects of languages and NLP on economic decision-making
  - The interplay of conversational agents (multi-agent systems)

About Me

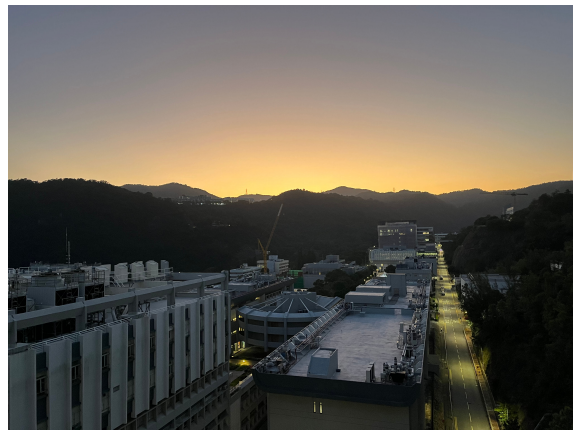
- Education
  - M.S., Computational Linguistics, University of Washington
  - M.Phil., Theoretical and Applied Linguistics, University of Cambridge
  - B.S., Economics, Linguistics, Mathematics, Asian Studies & Cert., Computer Sciences, University of Wisconsin–Madison


UW-Madison, Cambridge, and UW Seattle; I took those photos

About Me

- Experience
  - Teaching Assistant, at Chinese University of Hong Kong
  - Research Assistant, at HKUST Business School
  - Financial NLP Intern, at China Merchants Bank
- Languages: Wu Chinese, English, Mandarin Chinese, Korean
  - My Cantonese is very poor. If you have to talk to me in the language, please speak as slowly as possible. Thanks for your understanding.





CMB Shanghai, HKUST Business School, and CUHK; I took those photos

Contact

- Email: yigechen@link.cuhk.edu.hk
- Office hours
  - Time: Tuesday 12:00 – 14:00
  - Location: G19, Leung Kau Kui Bldg. or via Zoom
    - Meeting on Zoom is feasible if you let me know beforehand, i.e., send me an email and I'll set up the Zoom meeting
  - If you visit G19 and do not find me there during my office hours, please shoot me an email, and I'll be back in 5 minutes
  - If the time slot above does not fit in your schedule...
    1. Send me an email and (hopefully) we can schedule a meeting then
    2. You may also want to attend the office hours of Prof. Feng
- Website: https://yigechen.com

Course Materials

- Blackboard: https://blackboard.cuhk.edu.hk/
- Tutorial materials will also be posted at https://yigechen.com/teaching/ling3401/sp25
- Textbook I would recommend:
  - Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed. draft). 2024.
  - https://web.stanford.edu/~jurafsky/slp3/

Academic Honesty

- Any kind of plagiarism is prohibited
- But we will allow the use of LLMs, AI tools, etc. to some extent due to the nature of this course
  - With explicit acknowledgment and proper citation
  - Prohibited for midterm and final exams
- Please refer to https://www.cuhk.edu.hk/policy/academichonesty/ or ask us if you are unsure

Academic Accommodations

- Please contact Special Educational Needs (SEN) Service if you need special accommodations, facilities or arrangements due to physical or mental illnesses
- Please refer to https://www2.osa.cuhk.edu.hk/sens/en-GB/

Basic Text and Document Analysis

- Using web interfaces such as https://voyant-tools.org/
- Document frequency (of a word): total occurrence of a word in the document/corpus/text
- Vocabulary: count of unique words in the document/corpus/text
- Average Words Per Sentence
- Type/Token Ratio (TTR): the number of types divided by the number of tokens (words)
- ...

Visualizing Parsed Trees

- Check Universal Dependencies and its visualization tools!
  - https://universaldependencies.org/conllu_viewer.html
- We will visualize two already parsed sentences, one in English and the other in Cantonese

Downstream NLP Applications

- Some examples are:
  - Spam detection: determining whether a message (e.g., email, text) is spam or not
  - Hate and toxic speech detection: detecting harmful, toxic, or abusive language
  - Financial news sentiment analysis: analyzing financial news articles, headlines, or reports to determine their sentiment (positive, negative, or neutral)
  - Summarization: generating short summaries of long documents, articles, or emails.
  - Machine translation: translating text from one language into another
- We will use:
  1. Fine-tuned models dedicated to the aforementioned tasks, and
  2. Large language models (LLMs)

About Hugging Face

- A platform that allows users to share models and datasets and showcase their work
  - And their transformers library (which you will get to know later)!
- As a result, this is where you can find and download language models pre-trained or fine-tuned by others
  - For instance, the recently trending DeepSeek-V3 at https://huggingface.co/deepseek-ai/DeepSeek-V3 (but trust me, this model is way too large for you to download and use)

Accessing Large Language Models

- Through AI companies/developers such as OpenAI, Anthropic, etc.
  - But some may not be accessible in Hong Kong
- Through "middlewares", such as Poe, Microsoft Azure
  - If you want an interface, go with Poe (`https://poe.com`)
  - If you want an API (will introduce later during the semester), go with Microsoft Azure (`https://azure.microsoft.com`)
- For all the above, you may need to subscribe to their services

NLP Application: Spam Detection

- So let's get started!
- A fine-tuned model from Hugging Face: `mrm8488/bert-tiny-finetuned-sms-spam-detection` (https://huggingface.co/mrm8488/bert-tiny-finetuned-sms-spam-detection)
- And how about LLMs?

NLP Application: Toxic speech detection

- A fine-tuned model from Hugging Face: `unitary/toxic-bert` (https://huggingface.co/unitary/toxic-bert)
- And how about LLMs?

NLP Application: Financial news sentiment analysis

- A fine-tuned model from Hugging Face: `mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis` (https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis)
- And how about LLMs?

NLP Application: Summarization

- A fine-tuned model from Hugging Face: `facebook/bart-large-cnn` (https://huggingface.co/facebook/bart-large-cnn)
- And how about LLMs?

NLP Application: Machine Translation

- Let's compare Google Translate and LLMs!
- Google Translate: https://translate.google.com/

Miscellaneous

- Midterm: February 26, in class
- Final Exam: April 16, in class
- 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
- It would be great if you'd bring your laptop to the class every week