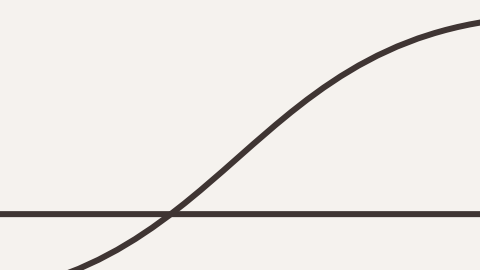




Simplified-Traditional Chinese Translation App

Emily Ho
Capstone Project



Overview

- Introduction and background
- Fine-tuning
 - Data and model
 - Prompt engineering
- App use & recorded demo

Background: Simplified vs Traditional

- Simplified Chinese: Used in Mainland China, Malaysia, Singapore, etc
- Traditional Chinese: Used in Taiwan, Hong Kong, Macau
- Differences:
 - Character form (e.g., “**体**” vs “**體**”) (*body*)
 - Vocabulary choices (e.g., “**酒店**” vs “**飯店**”) (*hotel*)
 - Some syntax or usage preferences
 - 有(*have*) + VP

Why Is Conversion Difficult?

- One-to-many character mapping: “面” → “面” or “麵”
 - 表面 → 表面 (*surface*)
 - 面包 → 麵包 (*bread*)
- Ambiguity in vocabulary: “质量” could be “質量” (*mass*) or “品質” (*quality*)
- Slangs
- Context matters: Not a simple word-for-word task

Previous Approaches

- Rule-based
 - Dictionary lookup
 - limited flexibility
- Statistical MT (SMT):
 - A Simplified-Traditional Chinese Character Conversion Model Based on **Log-Linear Models** by Xiamen University
 - Focused on character & lexical conversion

Previous Approaches

- Rule-based
 - Dictionary lookup
 - limited flexibility
- Statistical MT (SMT):
 - A Simplified-Traditional Chinese Character Conversion Model Based on **Log-Linear Models** by Xiamen University
 - Focused on character & lexical conversion

Sentence?

Previous Approaches

- Rule-based
 - Dictionary lookup
 - limited flexibility
- Statistical MT (SMT):
 - A Simplified-Traditional Chinese Character Conversion Model Based on **Log-Linear Models** by Xiamen University
 - Focused on character & lexical conversion

Sentence?

Neural MT (NMT)?

Fine-tuning a Transformer

- GPT-4o mini
- Data
 - Open Parallel Corpora (OPUS)
 - Subtitles (10 million) & Localization documents (100k)
 - Sentence-level alignment
 - Pre-processing
 - Longer sentences
 - 10k pairs

Experiments

- Chinese vs. English system prompts
 - Overall performance is similar across both languages
 - Chinese prompts tend to perform better in word choices
- Zero-shot vs. Few-shot prompting
 - Few-shot prompting provides more accurate translations

App Flow

Choose a translation style

Enter a sentence or upload a txt file

Edit the output if needed & save to the history

Download the output file if needed



Recorded Demo



Conclusion & Future Work

- Next steps:
 - Explore more models
 - Expand dataset
 - Style-aware conversion

Questions?