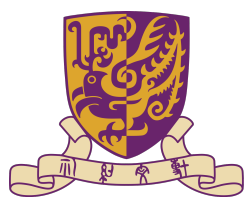
LING3401 Linguistics and Information Technology

Tutorial: Text Classification and Information Extraction

Yige Chen

The Chinese University of Hong Kong

April 2, 2025



Text Classification

- Assigns a label to an entire piece of text (sentence, document, etc.)
- One label per input
- Example:
  - Sentence: "I love this movie!"
  - Label: **Positive**
- Common applications:
  - Sentiment analysis
  - Spam detection

Sequence Labeling

- Sequence labeling is a task where a label is assigned to each token in a sequence
- The BIO scheme is commonly used
  - B (Begin) indicates the start of an entity
  - I (Inside) indicates a continuation of the entity
  - O (Outside) indicates no entity
- Example: Named entity recognition (NER)
  - Sentence: "John lives in New York"
  - Labels: B-PER O O B-LOC I-LOC

Information Extraction

- Information Extraction (IE) is about finding structured information from unstructured text
- IE includes tasks like:
  - Named Entity Recognition (NER) – identifying names of people, places, etc.
  - Relation Extraction – finding relationships between entities
- Example: "Barack Obama was born in Hawaii"
  - Entities: Barack Obama (Person), Hawaii (Location)
  - Relation: bornIn(Barack Obama, Hawaii)

Text Classification: Sentiment Analysis

- Classifies a sentence/document as Positive, Negative, or Neutral
- Used in reviews, social media, customer feedback
- Example: "I hate this product." $\rightarrow$ **Negative**

Text Classification: Spam Detection

- Classifies messages as **Spam** or **Not Spam**
- Used in email and messaging apps
- Example: "You've won a free iPhone!" $\rightarrow$ **Spam**

Information Extraction: Named Entity Recognition

- Finds names of people, places, organizations, etc.
- Example: "Barack Obama visited Paris." $\rightarrow$ **Barack Obama = PER**, **Paris = LOC**

Information Extraction: Relation Extraction

- Identifies relationships between entities
- Example: "Barack Obama was born in Hawaii" $\rightarrow$ **Relation: bornIn(Barack Obama, Hawaii)**

Text Classification using LLMs

- LLMs can understand natural language prompts and classify text without fine-tuning
- Two prompt strategies:
  - Zero-shot: Provide only task description and input
  - In-context learning: Show a few examples before asking

Information Extraction using LLMs

- LLMs can extract structured information from text using natural language prompts
- Very useful when you don't have task-specific models!
- Either zero-shot or in-context learning

Suggestions for using LLMs

- Process one sentence at a time for clarity and consistency
  - Avoid sending long paragraphs unless doing document-level tasks
  - LLMs may hallucinate or lose focus over long inputs
- Clearly define the task and desired output format
  - Specify whether you want JSON, bullet points, or plain text
- Use few-shot examples if the task is complex or ambiguous
  - Show the model how to respond by giving 1–3 examples
  - Helps reduce misinterpretation of instructions

Example: NER (zero-shot)

- You are an expert in named entity recognition (NER).

  Your task is to extract all named entities from the sentence provided below.

  Please output the results as a list of entities, where each line contains: <entity text>: <entity type>

  Sentence: "Barack Obama was born in Hawaii and worked at the White House."

  Output:

Example: NER (in-context learning)

- You are an expert in named entity recognition (NER).

  Your task is to extract all named entities from the sentence provided below.

  Please output the results as a list of entities, where each line contains: <entity text>: <entity type>

  Examples:

  Sentence:
  "Tim Cook is the CEO of Apple and lives in California."
  Output:
  Tim Cook: Person
  Apple: Organization
  California: Location

  Now extract entities from the following sentence:

  Sentence: "Barack Obama was born in Hawaii and worked at the White House."

  Output:

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