SUDO CODE - Week 3: Basic ML Techniques for NLP

Doan Ngoc Mai

September 13, 2025

Abstract

This survey introduces key Natural Language Processing (NLP) tasks, highlighting their definitions, examples, and relevance to Vietnamese. We discuss which tasks can be addressed with Large Language Models (LLMs) and prompt engineering, and which are most useful for chatbot development. Finally, we review common libraries and Vietnamese resources, aiming to connect theory, tools, and real-world applications.

1 Fundamental NLP Tasks

Based on recent surveys and educational resources **nlplanet2022**; Bhatti et al., 2024; Doan et al., 2024, core NLP tasks can be categorized as follows:

- **Text Classification:** Assigning predefined categories to text, e.g., sentiment analysis or topic classification. Example: determining whether a product review is "positive" or "negative". This task is highly popular in Vietnamese NLP due to applications in social media monitoring and customer feedback analysis.
- **Sequence Labeling:** Assigning labels to tokens in a sequence, including Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and word segmentation. For Vietnamese, word segmentation is crucial due to its whitespace-ambiguous writing system.
- **Text Generation:** Producing coherent text given input, e.g., machine translation, summarization, or dialogue generation. Example: translating English sentences into Vietnamese.
- **Information Retrieval and Question Answering:** Extracting or retrieving relevant information from documents. Example: a QA system answering "Thủ đô Việt Nam là gì?" with "Hà Nội".
- **Speech and Multimodal NLP:** Includes speech recognition, speech-to-text, and multimodal tasks (text + image/video). For Vietnamese, speech-to-text (ASR) is an active research area for education and accessibility.

2 Popular NLP Tasks for Vietnamese

Among the wide range of NLP tasks, certain problems have received more attention in the Vietnamese context due to the unique characteristics of the language and the practical needs of local applications. The most widely researched and deployed tasks include:

- **Sentiment Analysis and Text Classification.** These tasks aim to classify text into predefined categories such as positive/negative/neutral opinions or topical classes (e.g., politics, sports, technology). For example, the review *"Sản phẩm này rất tốt, tôi sẽ mua lại"* ("This product is very good, I will buy it again") should be classified as **positive**, while *"Dịch vụ quá tệ, tôi sẽ không quay lại nữa"* ("The service was terrible, I will not come back again") belongs to the **negative** class. This task is widely used in monitoring public opinion on platforms like Facebook, Zalo, or Shopee reviews. Benchmark datasets include UIT-VSFC (Vietnamese Students' Feedback Corpus) and VLSP sentiment analysis challenges.

- **Vietnamese Word Segmentation.** Unlike English, where whitespace separates words, Vietnamese uses spaces between syllables. For instance, the phrase *"học sinh giỏi"* ("excellent student") consists of three syllables but should be segmented into two words: `[học_sinh, giỏi]`. Incorrect segmentation (e.g., `[học, sinh, giỏi]`) can break meaning and harm downstream tasks like POS tagging or translation. Another example: *"sinh viên"* (student) must be treated as a single word, not two separate tokens. Popular segmentation tools include `underthesea`, `VnCoreNLP`, and `pyvi`.

- **Part-of-Speech (POS) Tagging and Named Entity Recognition (NER).** POS tagging assigns grammatical roles to words. Example: in the sentence *"Tôi đang học tại Đại học Quốc gia TP.HCM"* ("I am studying at Vietnam National University Ho Chi Minh City"), POS tags may be: `[Tôi/PRON, đang/ADV, học/VERB, tại/ADP, Đại_học_Quốc_gia_TP.HCM/PROPN]`.

  NER, on the other hand, identifies entities. In the same sentence, *"Đại học Quốc gia TP.HCM"* should be recognized as an **Organization**, while *"TP.HCM"* is a **Location**. Example queries:

  – "Ai là Chủ tịch nước Việt Nam?" → `[Chủ tịch nước/ROLE, Việt Nam/LOCATION]`.
  – "Nguyễn Nhật Ánh là tác giả nổi tiếng." → `[Nguyễn Nhật Ánh/PERSON]`.

  Benchmark datasets are provided by the VLSP shared tasks.

- **Machine Translation (English–Vietnamese and vice versa).** Translation is crucial for bilingual communication, education, and cross-border business. For example:

  – English → Vietnamese: *"How are you today?"* → *"Hôm nay bạn thế nào?"*
  – Vietnamese → English: *"Tôi đang học xử lý ngôn ngữ tự nhiên"* → *"I am studying natural language processing"*

  Challenges include handling tone markers, compound expressions, and cultural idioms. For instance, the idiomatic phrase *"nước đến chân mới nhảy"* should be translated as *"leaving things until the last minute"*, not literally as *"jump only when water reaches the feet"*. Transformer-based models such as mBART, MarianMT, and PhoMT achieve strong performance for this task.

3 LLM Applications in NLP Tasks

Large Language Models (LLMs) offer versatile ways to solve or assist many NLP tasks, especially via prompt engineering, fine-tuning, or instruction learning. Below are more detailed descriptions, examples, and findings (including for Vietnamese) from recent studies.

- **Text Classification (Sentiment, Topic, Intent)**

– *Prompting examples:*

```
Prompt: "Đánh giá cảm xúc của câu sau: 'Sản phẩm này rất tốt, tôi sẽ mua lại'
- Positive / Negative / Neutral?"

Prompt: "Classify the topic: 'Thị trường chứng khoán hôm nay tăng mạnh'
- Finance / Politics / Health / Others?"
```

– **Zero-shot / Few-shot:** Many studies show that LLMs (e.g. GPT, Llama-based, SeaLLM) can do sentiment analysis with few examples in the prompt, sometimes close to finetuned models, especially when labeled data is sparse. For Vietnamese, Thin et al. (2024) demonstrated this on multiple classification benchmarks.

– **Fine-tuning:** Full or parameter-efficient tuning (e.g., LoRA) can significantly boost accuracy on domain-specific text. For example, finetuning PhoBERT or LLaMA on Vietnamese e-commerce reviews improved F1 by $> 10\%$ compared to zero-shot prompting.

- **Named Entity Recognition (NER) and Information Extraction**

  – *Prompting examples:*

```
Prompt: "Trong đoạn văn sau, hãy tìm tên người, địa điểm và tổ chức:
'Thủ tướng Nguyễn Xuân Phúc đến thăm Thành phố Hồ Chí Minh'."

Prompt: "Extract the entities: Person, Location, Organization from:
'Công ty VinFast đặt trụ sở tại Hà Nội'."
```

  – **Findings for Vietnamese:** D. Van Thin et al. (2024) tested GPT, LLaMA-3 8B and SeaLLM v3 7B with few-shot prompting for NER and relation extraction. Results were competitive but still behind specialized supervised models in boundary accuracy.

  – **Instruction-tuning:** By aligning the LLM with NER-specific instructions (e.g., "always return entities in JSON"), performance and consistency improved.

  – **Knowledge Distillation:** Smaller models (e.g., PhoBERT) distilled from LLM outputs can perform NER more efficiently on resource-constrained systems.

- **Summarization and Dialogue / Response Generation**

  – Summarization: Both extractive and abstractive styles are possible. Prompts can specify length, style, or domain.

```
Prompt: "Tóm tắt đoạn văn sau trong 2 câu:
'Việt Nam đã ghi nhận tăng trưởng kinh tế toàn diện trong quý vừa qua...' "

Prompt: "Summarize the following technical article for a non-expert audience."
```

  – Dialogue Generation: Natural conversational style, context maintenance, multi-turn dialogue.

  – **Vietnamese examples:** BARTpho (Nguyen et al. 2022) fine-tuned on VNDS summarization dataset achieved higher ROUGE than mBART. For dialogue, GPT-style LLMs with instruction-tuning show strong generalization.

- **Machine Translation (MT)**
  - *Prompting examples:*

```
Prompt: "Translate the following English sentence into Vietnamese:
'Climate change is a global issue affecting us all.'"

Prompt: "Dịch sau đây từ tiếng Việt sang tiếng Anh:
'Tôi yêu thích học xử lý ngôn ngữ tự nhiên.'"
```

  - **Vietnamese MT resources:** PhoMT (3.02M sentence pairs) benchmarks show that mBART and MarianMT achieve strong BLEU scores.
  - **LLMs vs specialized models:** GPT and LLaMA-3 handle idiomatic translation well with context-aware prompting, but specialized MT systems (e.g., Google Translate tuned on Vietnamese) remain stronger on domain-specific data.
  - **Adapter tuning:** Lightweight adapters trained on parallel corpora can adapt a multilingual LLM to Vietnamese MT at much lower cost than full fine-tuning.
- **Machine Reading Comprehension (MRC)**
  - Example: Given a passage, ask a question such as: "Trong bài báo này, nguyên nhân nào được nêu lên?" $\rightarrow$ LLM extracts the answer.
  - **Vietnamese study:** Hoang et al. (2025) fine-tuned LLaMA-3 (8B) and Gemma (7B) on ViMMRC. They outperformed BERT-based baselines by > 5 EM (exact match).
  - **Techniques:** Few-shot prompting, chain-of-thought prompting for reasoning, and supervised fine-tuning on QA pairs.
  - **Retrieval-augmented generation (RAG):** Incorporating external passages from a Vietnamese knowledge base (e.g., VnExpress news) improves factual accuracy and reduces hallucinations.

4 NLP Tasks for Chatbot Development

For building an effective chatbot system, several NLP tasks play a central role:

- **Intent Recognition (Text Classification):** Determines the user's goal behind an utterance (e.g., booking a ticket, checking the weather, or asking for product information). Accurate intent detection is crucial for routing the conversation to the correct action or response module.
- **Named Entity Recognition (NER):** Identifies specific entities in user input such as names, dates, times, locations, or product names. For example, in the query *"Đặt vé đi Hà Nội ngày mai"*, the chatbot extracts *Hà Nội* (Location) and *ngày mai* (Date).
- **Dialogue Generation:** Produces natural and context-aware responses, enabling the chatbot to maintain coherent multi-turn conversations. This may involve rule-based templates, retrieval-based methods, or generative approaches using LLMs.
- **Question Answering and Retrieval:** Provides factual or knowledge-based answers by retrieving information from a knowledge base or external sources. For instance, when asked *"Thủ đô của Việt Nam là gì?"*, the system should return *"Hà Nội"*. Integration with retrieval-augmented generation (RAG) can improve accuracy and reduce hallucinations.

5 Tools and Resources for Vietnamese NLP

To support the above tasks (LLM-based solutions and chatbot development), several libraries and datasets are particularly relevant:

- **Libraries:**
  - `underthesea` Anh, 2018: word segmentation, POS tagging, NER, and sentiment analysis — useful for preprocessing and intent/entity recognition.
  - `VnCoreNLP`: fast and accurate segmentation + NER, often used as baseline for Vietnamese chatbots.
  - Hugging Face `transformers`: models such as PhoBERT (classification, sentiment), BARTpho (summarization, dialogue), ViT5 (QA, generation).
- **Datasets:**
  - **Text classification:** UIT-VSFC, UIT-ViCTSD, Kaggle Vietnamese classification dataset.
  - **NER and entity extraction:** VLSP datasets, BKTreebank.
  - **Dialogue and QA:** UIT-ViSQ (QA), ViMMRC (MRC), VNDS (summarization).
  - **Machine translation:** PhoMT corpus, useful for bilingual chatbot support.
- **Community resources:** NLP Progress (Vietnamese), Awesome Vietnamese NLP.

References

Anh, V. (2018). Underthesea: Vietnamese nlp toolkit [Accessed: 2025-09-11]. https://github.com/undertheseanlp/underthesea

Bhatti, R., et al. (2024). Prompt engineering for large language models for different nlp tasks.

Doan, V. T., et al. (2024). Investigating recent large language models for vietnamese nlp.