FINAL PROJECT

1. Thông tin chung

Môn học	NLP501 - Natural Language Processing
Trọng số	50% tổng điểm
Hình thức	Cá nhân hoặc Nhóm (1-3 members)
Thời gian	Trước buổi 10
Ngôn ngữ	Python (TensorFlow/PyTorch)

2. Timeline

Mốc	Thời gian	Hoạt động
Nộp bài	Trước buổi 10 (1-3 ngày)	Submit code + báo cáo
Trình bày	Buổi 10	Present và demo dự án

3. Các lựa chọn đề tài

Sinh viên tham khảo trong đề tài sau hoặc tự đề xuất đề tài phù hợp với nội dung môn học:

Option A: End-to-End Chatbot

Xây dựng chatbot cho một domain cụ thể với khả năng hội thoại nhiều lượt.

Mô tả

Phát triển một chatbot có khả năng hiểu và trả lời câu hỏi trong một lĩnh vực cụ thể (customer service, FAQ bot, booking assistant, etc.). Chatbot phải xử lý được multi-turn conversations và duy trì context.

Yêu cầu:

1. Triển khai Seq2Seq architecture với Attention mechanism
2. Xử lý multi-turn dialogue với conversation history
3. Intent classification để hiểu ý định người dùng
4. Entity extraction để trích xuất thông tin quan trọng
5. Response generation với beam search hoặc sampling
6. Simple UI/Interface cho demo (Gradio, Streamlit, hoặc CLI)

Dataset gợi ý

- Cornell Movie Dialogs Corpus
- DailyDialog Dataset
- MultiWOZ (task-oriented dialogues)
- Tự tạo dataset cho domain cụ thể

Evaluation Metrics

- BLEU score cho response quality

- Human evaluation (coherence, relevance, fluency)
- Intent accuracy (nếu có intent classification)
- Task completion rate (cho task-oriented chatbot)

Option B: Machine Translation System

Xây dựng hệ thống dịch máy neural cho một cặp ngôn ngữ.

Mô tả

Phát triển Neural Machine Translation (NMT) system để dịch giữa hai ngôn ngữ (khuyến khích Vietnamese-English hoặc English-Vietnamese). Hệ thống phải sử dụng encoder-decoder architecture với attention.

Yêu cầu:

1. Encoder-Decoder architecture với LSTM/GRU hoặc Transformer
2. Attention mechanism (Bahdanau hoặc Luong attention)
3. Subword tokenization (BPE hoặc SentencePiece)
4. Beam search decoding với adjustable beam width
5. Handling of unknown words (UNK tokens)
6. Interactive translation interface

Dataset gợi ý

- IWSLT (TED Talks translation)
- WMT datasets
- Tatoeba (sentence pairs)
- PhoMT (Vietnamese-English parallel corpus)
- OpenSubtitles

Evaluation Metrics

- BLEU score (corpus-level và sentence-level)
- METEOR score
- Human evaluation (adequacy, fluency)
- Attention visualization

Option C: Text Summarization System

Xây dựng hệ thống tóm tắt văn bản tự động.

Mô tả

Phát triển hệ thống có khả năng tóm tắt văn bản dài thành bản tóm tắt ngắn gọn. Có thể chọn extractive summarization (chọn câu quan trọng) hoặc abstractive summarization (sinh câu mới).

Yêu cầu:

1. Chọn một approach: Extractive HOẶC Abstractive (hoặc cả hai)

2. Extractive: sentence scoring, sentence selection, redundancy removal
3. Abstractive: Seq2Seq với attention, copy mechanism (optional)
4. Xử lý văn bản dài (truncation, chunking strategies)
5. Length control cho output summary
6. Web interface để demo với input documents

Dataset gợi ý

- CNN/DailyMail Dataset
- XSum (BBC articles)
- Multi-News (multi-document summarization)
- Vietnamese news dataset (tự thu thập)
- arXiv/PubMed (scientific papers)

Evaluation Metrics

- ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L)
- BLEU score
- Human evaluation (informativeness, coherence, conciseness)
- Compression ratio

Option D: Question Answering System

Xây dựng hệ thống trả lời câu hỏi dựa trên văn bản.

Mô tả

Phát triển hệ thống QA có khả năng trả lời câu hỏi dựa trên một corpus văn bản cho trước. Hệ thống bao gồm document retrieval và answer extraction/generation.

Yêu cầu:

1. Document Retriever: TF-IDF, BM25, hoặc dense retrieval
2. Reader/Extractor: span extraction hoặc answer generation
3. Xử lý multiple documents (ranking, re-ranking)
4. Handling no-answer cases
5. Confidence scoring cho answers
6. Simple Q&A interface cho demo

Dataset gợi ý

- SQuAD 2.0 (Stanford Question Answering Dataset)
- Natural Questions (Google)
- TriviaQA
- HotpotQA (multi-hop reasoning)
- Vietnamese QA dataset (UIT-ViQuAD)

Evaluation Metrics

- Exact Match (EM)

- F1 score (token-level overlap)
- Mean Reciprocal Rank (MRR) cho retrieval
- Human evaluation (correctness, completeness)

4. Sản phẩm nộp

Tất cả các sản phẩm cần được nộp trước buổi 10 (1-3 ngày):

4.1. Source Code

1. GitHub repository (public hoặc private với access cho giảng viên)
2. Cấu trúc thư mục rõ ràng: /src, /data, /models, /notebooks, /docs
3. Requirements.txt hoặc environment.yml
4. README.md với hướng dẫn cài đặt và chạy
5. Training scripts và inference scripts tách biệt

4.2. Trained Models

- Model weights (upload Google Drive/Hugging Face nếu file lớn)
- Model configuration files
- Tokenizer/Vocabulary files

4.3. Report (8-10 trang)

- Abstract: Tóm tắt project
- Introduction: Mô tả bài toán và motivation
- Related Work: Tổng quan các phương pháp liên quan
- Methodology: Chi tiết architecture và approach
- Experiments: Dataset, training setup, hyperparameters
- Results: Kết quả với tables và figures
- Analysis: Phân tích lỗi, case studies
- Conclusion: Kết luận và future work

4.4. Presentation Slides

- 10-12 slides cho phần trình bày (10 phút)
- Bao gồm demo live

5. Tiêu chí đánh giá

Tiêu chí	Tỷ trọng	Điểm	Mô tả
Accuracy & Correctness	25%	**2.5**	Mô hình hoạt động đúng, kết quả chính xác
Creativity & Problem-Solving	20%	**2.0**	Có sáng tạo, giải quyết vấn đề hiệu quả
Completeness (Code + Docs)	20%	**2.0**	Code đầy đủ, documentation chi tiết
Presentation Skills	15%	**1.5**	Trình bày rõ ràng, demo mượt mà
Theoretical Application	15%	**1.5**	Áp dụng kiến thức lý thuyết đúng đắn
Timeliness & Requirements	5%	**0.5**	Nộp đúng hạn, đúng yêu cầu

Tiêu chí	Tỷ trọng	Điểm	Mô tả
TỔNG CỘNG	100%	10.0	Quy về 50% tổng điểm môn học