Bài tập lớn
Xử lý ngôn ngữ tự nhiên
(NLP Summer 2024)

Nội dung:

Sinh viên cần nghiên cứu tìm hiểu, cài đặt một số thuật toán đã học, chạy thực nghiệm, phân tích kết quả, đánh giá phương pháp.

Có thể tự viết chương trình hoặc cải tiến chương trình có sẵn. Hiểu và nắm được kỹ thuật, phương pháp áp dụng, cần chỉ rõ các ưu nhược điểm của chương trình hiện có, sau đó đề xuất và cài đặt chương trình. Chú ý việc phân tích các cách tiếp cận liên quan đến vấn đề cài đặt và đánh giá các cách tiếp cận này.

Bài tập lớn làm theo nhóm 2-4 sinh viên. Báo cáo cần chỉ rõ công việc của mỗi thành viên trong nhóm, có đầy đủ các mục đặt vấn đề, các cách tiếp cận để giải quyết vấn đề, phân tích thiết kế cách tiếp cận đề xuất, thử nghiệm và đánh giá hệ thống, kết luận, tài liệu tham khảo. Báo cáo trong khoảng 20-40 trang.

Một số mã nguồn chương trình hoặc chương trình chạy có sẵn trên web: *theo tài liệu file đăng ký và phân công đề tài.*

Một số đề tài gợi ý:

1. Tìm hiểu mô hình Roberta và ứng dụng cho bài toán nhận dạng thực thể tên riêng.

   https://wandb.ai/madhana/Named_Entity_Recognition/reports/A-Beginner-s-Guide-to-Named-Entity-Recognition-NER---VmlldzozNjE2MzI1#what-is-named-entity-recognition-(ner)

2. Tìm hiểu bài toán tóm tắt văn bản tiếng Việt.

ViMs Dataset: CLC-HCMUS/ViMs-Dataset - 300 Cụm văn bản tiếng Việt dùng cho tóm tắt đa văn bản *by Nghiêm Quốc Minh (2016)*

Vietnamese MDS: lupanh/VietnameseMDS - 200 Cụm văn bản tiếng Việt dùng cho tóm tắt đa văn bản *by TM Vu (2013)*

Abstractive Text Summarization: VietAI/vit5-large-vietnews-summarization · Hugging Face

https://aclanthology.org/2022.naacl-srw.18.pdf

3. Tìm hiểu mô hình transformer và ứng dụng trong xử lý NNTN.

   https://d2l.ai/chapter_attention-mechanisms-and-transformers/index.html

   https://www.tensorflow.org/text/tutorials/transformer

   https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

   ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation | VietAI Research

4. Tìm hiểu Rasa chatbot và ứng dụng trong việc xây dựng module phân tích feedback người dùng

- https://rasa.com/docs/rasa/nlu-training-data/
- Introduction to Rasa Open Source & Rasa Pro

5. Mô hình nhúng từ word embedding trong việc hỗ trợ xử lý các bài toán NLP
6. Mô hình ngôn ngữ và bài toán thêm dấu câu trong tiếng Việt
7. Tìm hiểu về Question Answering và ứng dụng trong xây dựng hệ thống hỏi đáp

- https://github.com/facebookresearch/DrQA
- https://github.com/zaghaghi/drqa-webui
- https://github.com/thviet79/QA

8. Dịch máy NMT cho các cặp ngôn ngữ Việt-Anh, Việt-Khmer, Việt-Lào, Việt-Trung

- https://github.com/thviet79/KC4.0_MultilingualNMT
- http://210.245.53.88:1190/
- MTet: Multi-domain Translation for English and Vietnamese | VietAI Research
- A guidance for training MTet: Multi-domain Translation for English-Vietnamese - Colab (google.com)

9. Tìm hiểu xây dựng phân tích cú pháp phụ thuộc cho tiếng Việt

- https://www.aclweb.org/anthology/D14-1082.pdf
- https://text.xemtailieu.com/tai-lieu/phan-tich-cu-phap-phu-thuoc-tieng-viet-1125738.html
- https://courses.engr.illinois.edu/cs546/sp2020/Slides/Lecture17.pdf

10. Sử dụng mô hình BERT trong các ứng dụng xử lý NNTN
11. Viết chương trình tách từ. Đánh giá độ chính xác của hệ thống. Phân tích kết quả, đề xuất cải tiến chương trình.

    VnCoreNLP: https://arxiv.org/pdf/1801.01331v1

    A Vietnamese Text Processing Toolkit: https://github.com/phuonglh/vn.vitk

    Vietnamese NLP toolkit: Tokenizer, Sentence detector, POS tagger, Phrase chunker: https://vlsp.org.vn/node/70

12. Xây dựng công cụ gióng hàng văn bản cho các cặp ngôn ngữ Việt-Anh, Việt-Khmer, Việt-Lào, Việt-Trung.

    KC4Align: Improving Sentence Alignment method for Low-resource Language Pairs - ACL Anthology

Sinh viên có thể gửi các đề xuất đề tài trước 28/7. Đề xuất dài 1 trang, bao gồm:

- Mô tả vấn đề nghiên cứu

- Cách tiếp cận để giải quyết

- Một số tài liệu tham khảo

Đề xuất gửi về thviet@vnu.edu.vn

Ghi chú:

- Thời hạn deadline nộp bài: trước buổi cuối tuần kết thúc (tuần số 06) (trước ngày 09/08/2024)
- Nội dung nộp: Thư mục nén chứa code chương trình và dữ liệu dùng để chạy (có thể gồm file readme mô tả …)
- *Sinh viên có thể dùng github để upload toàn bộ chương trình, dữ liệu, readme và gửi link github.*