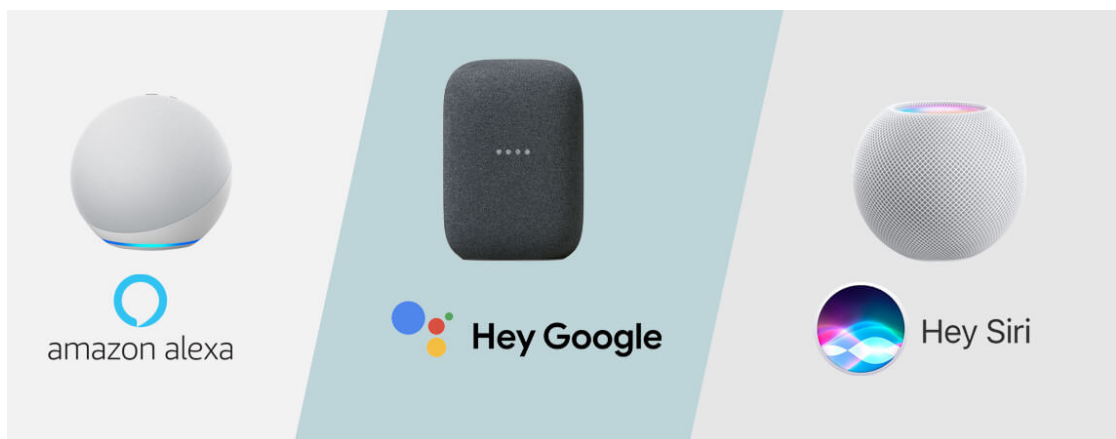


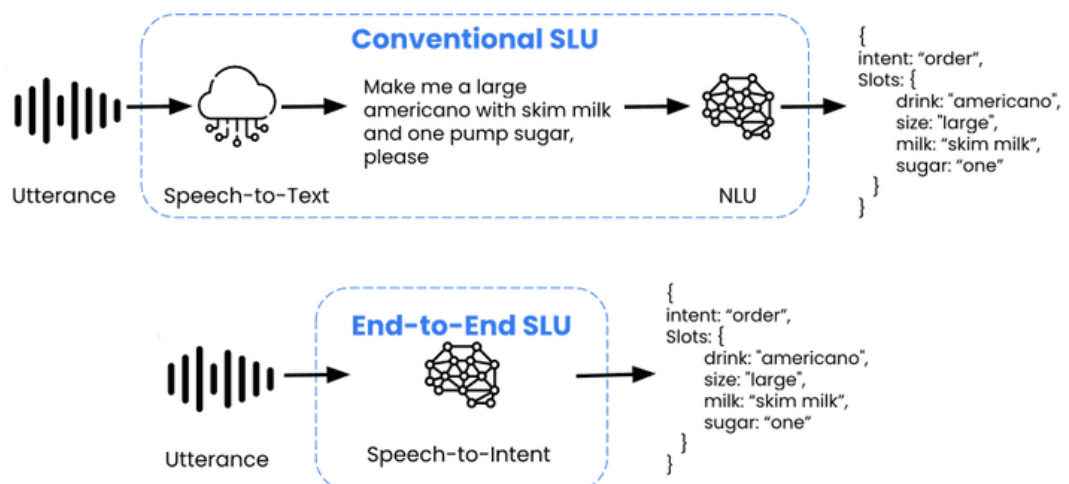
# SolCT Hackathon 2023: Vietnamese Spoken Language Understanding

## 1. Chủ đề:

- Cuộc thi Vietnamese Spoken Language Understanding tập trung vào giải quyết bài toán “Hiểu ngôn ngữ dạng nói tiếng Việt trong nhà thông minh”
- Spoken Language Understanding là bài toán nhận dạng ý định (intent) và thực thể (entity) của câu nói đầu vào. Hiện nay, các hệ thống SLU đã được các công ty lớn như Apple, Google, Alexa áp dụng rộng rãi trong việc triển khai các sản phẩm như nhà thông minh, trợ lý ảo:



- Hiện nay, có 2 hướng chính để giải quyết bài toán SLU, như thể hiện ở hình sau đây:



### 1.1. Conventional SLU (Pipeline SLU)

- Hệ thống gồm 2 mô hình chính: Speech-to-Text và NLU.

- Với câu nói đầu vào (input utterance), mô hình Speech-to-Text được sử dụng để trích xuất nội dung văn bản của câu nói.
- Nội dung văn bản được đưa qua mô hình Natural Language Understanding (NLU) để xác định ý định (intent) và thực thể (entity).

### 1.2. End-to-End SLU

- Hệ thống chỉ bao gồm 1 mô hình duy nhất
- Với câu nói đầu vào, mô hình End-to-End SLU trực tiếp đưa ra ý định và thực thể của câu nói.

## 2. Nhiệm vụ:

- Nội dung của cuộc thi là hiểu được ý định của câu nói và trích xuất được các thực thể có trong câu nói trong chủ đề nhà thông minh - smart home.
- Bằng cách nghiên cứu và huấn luyện các mô hình đặc thù cho bài toán, các đội thi sẽ đưa ra kết quả đầu ra là ý định của câu nói (intent) và nhãn của các thực thể xuất hiện trong câu nói đó (entity).

## 3. Dữ liệu:

Bộ dữ liệu VN-SLU bao gồm khoảng 11,000 file âm thanh ghi âm câu nói với tổng thời lượng khoảng 12 tiếng.

- Dữ liệu được cung cấp bởi ban tổ chức gồm 3 tập như sau:
  - Training data: là tập dữ liệu có gán nhãn, dùng để huấn luyện mô hình. Tập này gồm khoảng 8,000 file âm thanh.
  - Public test: là tập dữ liệu dùng để các đội thi phát triển mô hình và để BTC sử dụng sơ khảo.
  - Private test: là tập dữ liệu khó hơn và được sử dụng để đánh giá kết quả chung kết.
  - Cả 3 tập dữ liệu sẽ không trùng nhau
- Đầu vào mô hình là file âm thanh chưa được gán nhãn. Tập nhãn là các file định dạng .jsonl. Mỗi dòng của tệp nhãn chứa các thông tin bao gồm: id, sentence, intent, sentence\_annotated của một file âm thanh.
- Với mỗi tên file âm thanh là {id}.wav, đội thi cần xuất ra kết quả tương ứng và lưu vào file kết quả với định dạng .jsonl sau đó nén thành file .zip và nộp lên hệ thống.

## 4. Tiêu chí đánh giá.

- Tiêu chí đánh giá là trung bình cộng của 2 chỉ số Intent-F1 và SLU-F1. Trung bình cộng càng cao sẽ càng chính xác.

$$SLU\_score = \frac{(Intent\_F1 + SLU\_F1)}{2}$$

### 4.1. Cách tính Intent-F1:

- True Positive (TP): Số lượng intent dự đoán chính xác
- False Positive (FP): Số lượng intent dự đoán bởi hệ thống nhưng không có trong groundtruth
- False Negative (FN): Số lượng intent có trong groundtruth nhưng không được hệ thống dự đoán

#### Công thức tính Intent-F1:

$$Intent\_F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Với:  $Precision = \frac{TP}{TP + FP}$  và  $Recall = \frac{TP}{TP + FN}$

### 4.2. Cách tính SLU-F1:

- True Positive (TP): Số slot value dự đoán đúng
- False Positive (FP): Số lượng slot value hệ thống dự đoán nhưng không có trong groundtruth
- False Negative (FN): Số lượng slot value có trong groundtruth nhưng không được hệ thống dự đoán

#### Công thức tính SLU-F1 như sau:

$$SLU\_F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Với:  $Precision = \frac{TP}{TP + FP}$  và  $Recall = \frac{TP}{TP + FN}$

## 5. Định dạng dữ liệu:

File kết quả là file .jsonl với định dạng tên file là Prediction.jsonl.

Mỗi dòng của file kết quả sẽ biểu diễn intent và entities của một file âm thanh tương ứng. Định dạng như sau:

```
{“intent”: ..., “entities”: [{“type”: ..., “filler”: ...}, {“type”: ..., “filler”: ...},...], “file”: ...}
```

Trong đó:

- intent: là ý định của câu nói
- entities: là tập hợp các giá trị xuất hiện trong câu nói với "type" là loại giá trị và "filler" là giá trị đó.
- File là tên file của câu nói

Ví dụ: {"intent": "Mở thiết bị", "entities": [{"type": "command", "filler": "mở"}, {"type": "device", "filler": "điều hòa"}], "file": "6485cf1b850351e0dce4d289.wav"}

Sau khi ghi xong file .jsonl thì nén lại vào file .zip với định dạng tên file là Submission.zip và nộp lên hệ thống. File kết quả trong file .zip không được nằm trong thư mục cha nào.

## 6. Quy định:

- Các đội thi cần tuân thủ các quy định chung của Hackathon Soict 2023.
- Các đội thi chỉ được phép sử dụng bộ training data để huấn luyện mô hình. Không được phép can thiệp hay sử dụng public test hay private test trong quá trình huấn luyện dưới bất kỳ hình thức nào.
- Các bộ dữ liệu không có tiếng nói (vd. dữ liệu tiếng ồn, tiếng vang...) có thể được sử dụng trong quá trình huấn luyện và phải được chia sẻ với các đội thi khác.
- Các đội thi được phép sử dụng pretrained model nhưng không được sử dụng các pretrained model đã được huấn luyện cho các bài toán SLU và NLU (trên cả tiếng Việt và các ngôn ngữ khác).
- Các pretrained model phải là các pretrained công khai (vd. PhoBERT), có thể được sử dụng bởi các đội thi khác.
- Để vượt qua vòng sơ khảo, các đội cần phải nộp code kèm hướng dẫn chạy, đồng thời kết quả trên public test của đội thi phải reproduce được.

## 7. Trích dẫn và tài liệu tham khảo:

<https://aclanthology.org/2020.emnlp-main.588.pdf>

## **8. Thành viên ban tổ chức:**

- TS. Nguyễn Thị Thu Trang
- Trần Đăng Tuyền
- Lê Trọng Khánh
- Ngô Thị Thu Huyền
- Nguyễn Duy Khánh
- Phạm Việt Thành