

KỊCH BẢN BUỔI THUYẾT TRÌNH

I. Mở đầu:

a. Giới thiệu:

i. Giới thiệu nhóm:

- Nhóm: *STT nhóm, thành viên nhóm.*
- Lời nói đầu: Quan tâm hỏi đến có bạn/anh/chị nào trong lớp theo định hướng NLP hay không? Nếu có thì xin phép thưa trước với các bạn/anh/chị đang theo định hướng này thì mong mọi người bỏ qua những sai sót của chúng mình khi nói về chủ đề mà bọn mình lần đầu tìm hiểu,. Nếu có gì sai sót mong các bạn/anh/chị hãy góp ý ở cuối giờ để tránh bài thuyết trình bị ngắt đoạn và không thể tiếp tục... Nếu không thì cũng xin thưa trước với các bạn/anh/chị chúng mình lần đầu tiên tìm hiểu về chủ đề mới này và khó có thể tránh khỏi sai sót về kiến thức. Nếu có thì ở cuối giờ mình cũng xin nhận nghe ý kiến từ các bạn/anh/chị và thầy để chúng mình kịp thời chỉnh sửa và đưa đến các bạn những kiến thức đúng.
- Đề tài: *Word-lattices.*
- *Quá trình làm việc nhóm.*

ii. Giới thiệu chi tiết đề tài:

- Ứng dụng của “**Mô hình ngôn ngữ**” (**Language Model**):

*“Trước khi bắt vào chủ đề chính thì chúng ta sẽ nghiêng sang hướng nói về các công nghệ có liên quan đến thuật ngữ “**Mô hình ngôn ngữ**” một chút. Vì sao lại nói về nó thì ở cuối phần sau khi chúng ta nói về các công nghệ mình sẽ cho các bạn biết lý do, các bạn có thể xem phần này là bước đệm để vào chủ đề chính “**Word-lattices**”, theo một hướng tiếp cận top-down để có thể một cách nào đó dễ dàng tiếp cận với “**Mô hình ngôn ngữ**” và rồi đến với “**Word-lattices**”.”*

- **Speech recognition (Nhận diện giọng nói)** (Nguyễn Văn Hoàng): Phần này sẽ dùng phần nhận diện giọng nói của *Google Translate* để dẫn dắt vấn đề dù nói với ngữ điệu đặc trưng, không rõ ràng nhưng phần nhận diện giọng nói của *Google* có thể bắt được và cho ra từ ngữ chính xác/gần chính xác của người nói (Phần này cần chú ý hơn về phần các từ đồng âm mà khác mặt chữ để cho thấy việc ứng dụng mô hình

ngôn ngữ để suy diễn ra các từ chính xác và hợp lý). Một số ứng dụng tách giọng nói của người ra khỏi tạp âm: *Lalal.ai, Adobe Audition, EasySplitter,...*

"Như mọi người có thể thấy về công nghệ nhận diện giọng nói xuất hiện rất nhiều trong cuộc sống hiện nay. Nói đến Google dịch là một ví dụ điển hình mà các bạn hay sử dụng nhiều trong việc các bạn bắt giọng nói của một ai đó để phiên dịch sang ngôn ngữ khác, một cái hay là nếu như các âm thanh khác không phải giọng nói con người thì sẽ ít được bắt vào và dịch. Hay chức năng tìm kiếm bằng giọng nói cũng bắt giọng các bạn và từ đó tìm kiếm. Ngoài ra còn một ứng dụng mà các bạn hay thấy đó là tách giọng hát ra khỏi nền nhạc,..."

- **Suggestion in messengers (Mẫu tin nhắn gợi ý):**

Dùng *Google Chat* hoặc *Gmail* để minh họa cho việc dựa vào những nội dung nhắn, *Google* có thể đưa ra những gợi ý đáp trả gần chính xác cho cuộc hội thoại.

"Một công nghệ khác ta có thể thấy thông qua việc nhắn tin trên điện thoại (đưa ra các từ gợi ý – word recommend (predict word)), các đoạn mẫu tin nhắn trả lời gợi ý, các mẫu mail reply gợi ý,..."

- **Spelling correction (Sửa lỗi chính tả):** Tương tự có thể dùng *Gmail, Word* để minh họa cho việc khi viết sai một từ nào đó, *Gmail, Word* sẽ thông báo cho chúng ta nên sửa lại, một ví dụ cao hơn là dùng *Grammarly* (<https://tinyurl.com/44ptv9hv>, chỉ nên cho xem sơ trang web vì đầu trang web có một video demo *Grammarly* gợi cú pháp câu hay và hợp lý hơn so với bản gốc).

Lời nói có thể bám vào phía trên để nói, không cần soạn ra thêm.

- **Handwriting Recognition.**

- **Machine Translating.**

- **,...**

"Tất cả ứng dụng trên đều có liên quan ít nhiều đến "Mô hình ngôn ngữ". Các bạn hẳn là đang thắc mắc vì sao chúng ta nói nhiều đến "Mô hình ngôn ngữ" đến vậy? Nó có liên quan gì đến "Word-lattices"? (Đến đây mình giải thích ngắn gọn việc "Word-lattices" là một mạng từ tổ chức thành một "Đồ thị có hướng không chu trình" (Directed Acyclic Graph (DAG)) sử dụng kết quả của "Mô hình ngôn ngữ" để hình thành nên "Word-lattices")"

- Tóm tắt sơ bộ nội dung bài thuyết trình:

- **Ngữ liệu (Corpus)?**

- **N-gram**

- Khái niệm, công thức, ý nghĩa.

- Sơ lược về **Markov Assumption** và cơ sở vì sao có thể dùng **Markov Assumption**. Ví dụ đơn giản giải thích.
- Ví dụ trên một đoạn **ngữ liệu (Corpus)**.
- Biểu diễn kết quả của N-gram thành một ma trận liên kết.
- Sơ bộ về vấn đề của N-gram gặp phải (**Dữ liệu thưa (Sparse Data)**).
- Khắc phục vấn đề **Sparse Data** của N-gram thông qua các kỹ thuật **Làm mịn (Smoothing)** (Ở mức độ giới thiệu, chỉ giới thiệu 2 kỹ thuật đơn giản là **thêm-1** và **thêm-alpha**).
- **Word-lattices**
 - Khái niệm, đặc điểm, công dụng.
 - Tổ chức dưới dạng nào? Sử dụng cấu trúc dữ liệu nào để lưu trữ. (Đồ thị - Ma trận liên kết).
 - Sử dụng kết quả của **N-gram** để đưa lên các cạnh của đồ thị thành trọng số cạnh của “**Word-lattices**”.
- Demo thử từng giai đoạn
 - Đọc ngữ liệu và tách các từ
 - Dùng **Bigram** train để cho ra ma trận liên kết.
 - Cấu tạo đồ thị.
 - Duyệt thử đồ thị với ứng “**Word-lattices**” cho bài toán “**Word Prediction**”.
- Thảo luận thêm về kết quả của **N-gram**.

II. Vấn đề chính:

a. Ngữ liệu?

b. N-gram:

- i. Nói về khái niệm của “**Mô hình ngôn ngữ**”.
- ii. Đề cập đến **N-gram**
 1. Khái niệm
 2. Ý nghĩa
 3. Công thức
- iii. Sơ lược về **Markov Assumption** và cơ sở vì sao có thể dùng **Markov Assumption**. Ví dụ đơn giản giải thích.

iv. Ví dụ

v. Minh họa kết quả của **N-gram** thành ma trận liên kết

vi. Vấn đề “**Sparse Data**” mà **N-gram** gặp phải và cách khắc phục

c. Word-lattices

i. Thảo luận về **Word-lattices**

1. Khái niệm

2. Công dụng

ii. Thảo luận về việc dùng cấu trúc dữ liệu nào để lưu trữ

iii. Vận dụng kết quả của **N-gram** để sinh ra **Word-lattices**.

d. Demo

e. Thảo luận ngoài thêm về kết quả của N-gram. (Sử dụng tổng các log của xác suất thay vì tích của các xác suất để cho ra dự đoán độ tin cậy của câu so với thực tế).

III. Phần kết:

a. Giải đáp các câu hỏi liên quan.

b. Cảm ơn mọi người đã lắng nghe.

c. Kết thúc.