Guideline gán nhãn cho bộ dữ liệu ATIS tiếng Việt

1. Bộ dữ liệu ATIS

Giới thiệu

Airline travel information system(ATIS) là bộ dữ liệu tiếng anh về các đoạn hội thoại giữa người dùng và người chăm sóc khách hàng của Microsoft. Mục đích chính-intent của các câu trong đoạn hội thoại được chia thành 16 intent: như atis_flight, atis_airfare, atis_airline,… các nhãn này vì được gán theo từng câu nên khi dịch ra tiếng Việt không cần phải gán lại. Một câu có thể có nhiều hơn 2 nhãn intent.

Ví dụ: show flight and prices kansas city to chicago on next wednesday arriving in chicago by 7 pm

Câu này sẽ có 2 nhãn intent là atis_flight và atis_airfaire.

Ngoài ra nó còn có nhãn IOB: Inside, Outside, Begin:

O O O O O B-fromloc.city_name I-fromloc.city_name O B-toloc.city_name O B-depart_date.date_relative B-depart_date.day_name O O B-toloc.city_name B-arrive_time.time_relative B-arrive_time.time I-arrive_time.time

Khi dịch sang tiếng việt bằng Google dịch: hiển thị chuyến bay và giá thành phố kansas đến Chicago vào ngày thứ tư tiếp theo đến Chicago trước 7 giờ tối

Nhãn intent của câu tiếng Việt vẫn giữ nguyên, tuy nhiên nhãn IOB đã thay đổi do trật tự từ trong câu tiếng Việt thay đổi. Guideline này hướng dẫn cách sửa lại câu tiếng việt cho hoàn chỉnh và gán lại nhãn IOB cho câu Tiếng Việt dựa vào IOB của câu Tiếng Anh.

Các bước gán nhãn IOB cho bộ dữ liệu ATIS tiếng Việt

1. Sửa lại các câu Tiếng Việt cho đúng nghĩa, tự nhiên.
2. Dùng label-studio gán nhãn IOB cho câu Tiếng Việt

2. Gán nhãn cho data

1. Sửa lại câu Tiếng Việt

Bước cần làm đầu tiên là sửa lại câu tiếng Việt nếu cần. Các từ cần sửa như:

- Tên riêng người, tên địa danh, tên hãng hàng không không được dịch ra tiếng việt mà để y tiếng Anh.
    - Ví dụ sau không cần sửa:
        - En: i would like to find a flight from charlotte to las vegas that makes a stop in st. louis
        - Vi: tôi muốn tìm một chuyến bay từ charlotte đến las vegas dừng ở st. louis
    - Ví dụ sau cần sửa lại:
        - En: on april first i need a flight going from phoenix to san diego
        - Vi: vào tháng tư đầu tiên tôi cần một chuyến bay đi từ phượng hoàng đến san diego
        - Sửa lại: vào tháng tư tôi cần một chuyến bay đi từ Phoenix đến San Diego

    - Cách để biết một từ có phải tên riêng tiếng Anh không: dùng từ đó tra Google
- Một số câu tiếng Việt không được dịch mà giống như câu tiếng Anh, cần phải dịch lại câu đó bằng Google dịch.
- Những câu Tiếng Việt nào đọc xong không rõ nghĩa hoặc vô lý thì phải dịch lại. Tỉ lệ của các câu này rất thấp.
    - Ví dụ câu sao cần dịch tay lại:
        - En: what is airline wn
        - Vi: hãng hàng không là gì
        - Sửa lại: câu hỏi này hỏi về từ viết tắt, sửa lại: airline wn là gì
- Các từ trong tiếng Anh có một từ thuộc một nhãn nào đó, được gán B-tên-nhãn nhưng khi dịch ra tiếng Việt tạo ra 2 từ cùng mang 1 nhãn đó thì phải gán B-tên-nhãn cho từ đầu và I-tên-nhãn cho các từ tiếp theo của cụm từ đó.
    - Ví dụ
        - En: what is the earliest breakfast flight from philadelphia to fort worth
        - IOB:O O O B-flight_mod B-meal_description O O B-fromloc.city_name O B-toloc.city_name I-toloc.city_name
        - Vi: chuyến bay ăn sáng sớm nhất từ Philadelphia đến Fort Worth giá bao nhiêu
        - Từ breakfast trong câu tiếng Anh được gán B-meal_description, nhưng dịch sang tiếng Việt tạo ra 2 từ "ăn sáng" nên phải gán là B-meal_description I-meal_description
        - Từ earliest tương tự, cần gán lại là B-flight_mod I-flight_mode.
- Các câu tiếng Việt dịch ra không tự nhiên cần được dịch lại:
    - En: i'm interested in a flight from pittsburgh to atlanta
    - Vi: Tôi quan tâm đến một chuyến bay từ pittsburgh đến atlanta
    - Sửa: tôi muốn biết/xem thông tin/ một chuyến bay từ pittsburgh đến atlanta
- Các câu dịch có cụm từ "một điểm dừng" cần sửa lại thành "quá cảnh ở …"
    - En: i'm looking for a flight from oakland to denver with a stopover in dallas fort worth
    - Vi: tôi đang tìm một chuyến bay từ Oakland đến Đan Mạch với một điểm dừng ở pháo đài dallas đáng giá
    - Sửa: Tôi đang tìm một chuyến bay từ Oakland đến denver quá cảnh ở dallas fort worth
- Các câu hỏi về từ viết tắt thường không được dịch tốt, mất đi từ viết tắt nên cần được sửa lại
    - En: what does the abbreviation co mean
    - Vi: từ viết tắt nghĩa là gì
    - Sửa: từ viết tắt co nghĩa là gì
- Nonstop dịch thành "không quá cảnh", hoặc đi thẳng
    - En: i need a flight from atlanta to baltimore nonstop arriving at 7 pm please
    - Vi: Tôi cần một chuyến bay từ Atlanta đến Baltimore không quá cảnh, đến nơi lúc 7 giờ tối
    -

2. Gán nhãn IOB

Các nhãn IOB được chia theo các slot, mỗi slot là một từ mang một thông tin nào đó mà mô hình cần phải tìm ra nó khi đưa cho nó một câu nói tự nhiên.

Ví dụ 1:

- En: what does ua mean
- IOB: O O O B-airline_code O
- Vi: us có nghĩa là gì
- Nhãn IOB này được gán theo câu tiếng Anh, O là từ không quan tâm tới, B-airline_code là nhãn bắt đầu của airline_code. Số lượng nhãn bằng với số các từ trong câu cần gán.

Ví dụ 2:

- En: show me all nonstop flights from st. petersburg to charlotte
- IOB: O O O O B-flight_stop O O B-fromloc.city_name I-fromloc.city_name O B-toloc.city_name
- Vi: chỉ cho tôi tất cả các chuyến bay thẳng từ st. petersburg đến charlotte
- Trong câu này vì st. petersburg gồm có 2 từ cùng mang ý nghĩa là: nơi khởi hành, nên nó được gán nhãn là: B-fromloc.city_name I-fromloc.city_name.

Khi gán IOB cho câu Tiếng Việt, chỉ cần tham khảo nhãn từ câu Tiếng Anh tương ứng và gán cho câu TV. Chỉ gán các nhãn không phải O.

Gán bằng label-studio được hướng dẫn cách cài đặt và dùng riêng.

Xử lý các trường hợp nhập nhằng:

1. Tên các tháng, thứ trong tiếng Anh chỉ có một từ, nhưng khi dịch ra tiếng Việt tạo ra 2 từ. June -> tháng 6, Monday -> thứ 2,.. khi gán nhãn gán nhãn B cho từ đầu và I cho các từ còn lại tương ứng, Tháng \B-month_name 6\I-month_name
2. Các từ ghép về ngày : ngày 5, ngày 6… gán nhãn B cho từ ngày và I cho từ tiếp theo
3.