week-7-assignment-note

Tìm hiểu về EVBCorpus

(Reference: EVBCorpus_EnVnNEguide_v1.1.pdf)

Named Entity		là đối tượng trong thế giới thực — ví dụ như con người, địa điểm, hoặc tổ chức — được đề cập bằng tên riêng, biệt danh, chữ viết tắt hoặc tên gọi tắt. ví dụ: Mt. Fuji → Núi Fuji the Kremlin → Điện Kremlin
Entity Types	6 loại entity chính	
	PERSON (PER)	Tên người thật, tên người đã khuất, nhân vật hư cấu, hoặc các vị thần trong tôn giáo đều được gắn nhãn là PERSON Phó tổng thống [Cheney]PER ghé thăm nơi này Chủ tịch [Microsoft]ORG [Bill Gates]PER phát biểu rằng...
	ORGANIZATION (ORG)	tên gọi của công ty, cơ quan chính phủ, trường học, đội thể thao, bệnh viện, khách sạn, bảo tàng, ... có cấu trúc tổ chức rõ ràng Công ty thể thao [Bridgestone]ORG thu lợi nhuận từ các sản phẩm xa xỉ Cổ phiếu [NASDAQ]ORG [Bộ Ngoại giao]ORG
	LOCATION (LOC)	tên các địa điểm địa lý hoặc chính trị như thành phố, quốc gia, tỉnh, đường phố, sông, hồ, núi, khu vực hư cấu, hoặc công kiến trúc nổi tiếng Người dân thường tụ tập ở [Nhà thờ Đức Bà]LOC vào mỗi dịp lễ lớn [Sông Sài Gòn]LOC [Thành phố Hồ Chí Minh]LOC

	TIME (TIM)	**khoảng thời gian cụ thể như giờ, phút, ngày, tháng, năm** **- Absolute Time:** [twelve o'clock noon]TIM [5 p.m. EST]TIM [January 1990]TIM **- Relative Time: "hôm nay", "tuần trước", "năm tới"** [tối hôm qua]TIM [sáng ngày mai]TIM
	PERCENTAGE (PCT)	**các giá trị phần trăm** khoảng [5%]PCT hơn [55%]PCT
	MONEY (MON)	**các giá trị tiền tệ** xấp xỉ [20 triệu New Pesos]MON [5 triệu đồng]MON hơn [90,000 đô-la]MON
Các trường hợp khó		**Nhiều thực thể cùng xuất hiện: Nếu một câu đề cập đến nhiều thực thể, từng thực thể sẽ được gắn nhãn riêng biệt** [Trung Quốc]LOC và [Hàn Quốc]LOC ký thỏa thuận thương mại
		Thực thể lồng nhau: Nếu một thực thể chứa trong nó một thực thể khác, chỉ gắn nhãn cho thực thể lớn hơn [Bảo tàng Guggenheim]ORG, không tách riêng [Guggenheim]ORG
		Thực thể làm tính từ: Trong tiếng Anh, các thực thể có thể xuất hiện dưới dạng tính từ và vẫn được gắn nhãn, nhưng quy tắc này không áp dụng cho tiếng Việt. Tuy nhiên, khi dịch ra nó đại diện cho 1 danh từ "[American]LOC companies", "công ty [Mỹ]LOC"
		Sự kiện: như "Thế vận hội Mùa đông" ko được gắn nhãn Ủy ban Olympic → gắn nhãn [Ủy ban Olympic]ORG

	Thế vận hội Mùa đông → không gắn nhãn
	Sản phẩm và hiện vật: Các tên sản phẩm, hiện vật không được gắn nhãn Taurus là mẫu xe mới nhất → không gắn nhãn
Trường hợp không chắc chắn	sử dụng nhãn "No_Annotation" cho đến khi được giải quyết sau
Tóm tắt 1. 6 loại chính của named entity: PERSON, ORGANIZATION, LOCATION, TIME, PERCENTAGE, và MONEY. 2. Quy tắc chi tiết giúp quyết định khi nào và làm thế nào để gắn nhãn chính xác cho từng loại thực thể, và xử lý các trường hợp khó hoặc không rõ ràng như thế nào. 3. Những gì không được gắn nhãn: tên sự kiện, sản phẩm, hiện vật, và các thực thể không mang tính xác định.	

Thông tin về dataset:

EVBCorpus: chứa các cặp song ngữ Anh-Việt,gồm các văn bản từ sách, luật, bài báo, và phụ đề phim.		
EVBCorpus_EVBNews_v2.0.rar		chứa dữ liệu song ngữ Anh-Việt, các câu được ghép cặp và liên kết với nhau thông qua ánh xạ từ vựng
Ví dụ: **N0001.sgml**		
thẻ doc	<doc id='N0001'>	
thẻ title	<title>What is a Fenqing ?</title>	
thẻ <corpus>, thẻ <author>. thẻ <citation>, thẻ <release>		datasource, thông tin liên hệ với tác giả,...
thẻ <spair>	<spair id='1'> <s id='en1'>What is a Fenqing ?</s> <s id='vn1'>Fenqing là gì ?</s> 1-3;2-2;4-1; </spair>	mỗi đoạn dịch song ngữ en-vi là mỗi cặp spair

thẻ <a>	1-3;2-2;4-1;	ánh xạ giữa các từ trong hai câu. vd 1-3: từ thứ nhất của câu tiếng Anh tương ứng với từ thứ ba của câu tiếng Việt

References:

https://github.com/bangoc123/transformer/tree/master
https://www.tensorflow.org/community?authuser=0#training