Tóm tắt báo cáo

Đồ án "Xây dựng hệ thống tìm tập từ tiếng Anh tốt nhất biểu diễn nghĩa của đoạn mô tả tiếng Anh" giúp khắc phục được vấn đề suy nghĩ bằng tiếng mẹ đẻ khi sử dụng tiếng Anh của những người không phải người Anh bản xứ. Chức năng chính của hệ thống là khi người dùng muốn sử dụng một từ nào đó mà người dùng không biết trong tiếng Anh, nhưng biết nghĩa của nó, biết mô tả nó như thế nào, thì người dùng có thể mô tả từ đó và hệ thống sẽ đưa ra tập từ biểu diễn nghĩa tốt nhất của đoạn mô tả đó, từ đây người dùng có thể xem nghĩa của những từ trong tập từ hệ thống đưa ra và chọn ra từ nào phù hợp với ngữ cảnh mà người dùng muốn nhất. Người dùng có thể đánh giá từ nào trong tập từ phù hợp với đoạn mô tả, hệ thống sẽ dựa vào đánh giá của người dùng để đưa ra kết quả tốt hơn trong các lần tìm kiếm sau.

Đồ án đi sâu vào tìm hiểu mạng neural nhân tạo, các thuật toán, mô hình bên trong công cụ word2vec, một công cụ do Google nghiên cứu, chuyển mỗi từ thành một vector, có ý nghĩa rất lớn trong lĩnh vực xử lý ngôn ngữ tự nhiên và đã đạt được những kết quả rất ấn tượng.

Khi hệ thống nhận được mô tả của người dùng, nó sẽ tính độ tương đồng bằng khoảng cách cosin của câu mô tả và nghĩa của từ trong từ điển (ở đây hệ thống dùng từ điển Oxford vì nghĩa chính xác, nghĩa được giải thích dễ hiểu). Từ đó lấy tập từ trong từ điển có độ tương đồng gần nhất với câu mô tả. Dựa vào word2vec thì chỉ tính được độ tương đồng của hai từ. Nên muốn tính độ tương đồng của hai câu. Tôi đề xuất hai phương pháp, một là tính ra vector của câu bằng cách tính trung bình các vector từ trong câu, hai là tính độ tương đồng cosine của mỗi cặp từ, một từ của câu này với các từ trong câu kia, lấy giá trị max, sau đó tính trung bình cộng các giá trị max này.

Để tính toán độ chính xác của mô hình đề xuất, tôi tiến hành thử nghiệm trên các bộ dữ liệu word2vec train sẵn khác nhau (với số chiều của vector khác nhau 100, 200, 300, word2vec model khác nhau như CBOW, Skip-gram). Bộ dữ liệu để thử nghiệm được xây dựng bởi người dùng thực. Thử nghiệm cho thấy hệ thống xử lý đưa ra kết quả khá chính xác từ 70 đến 89%. Và tốc độ xử lý trung bình tầm 1s đến 10s tùy từng phương pháp tính độ tương đồng của câu được đề xuất ở trên.