**Trần Tấn Quý**

**Phạm Hoàng Việt**

**Natural Language Processing (NLP)**

Trong đề tài Khóa luận tốt nghiệp của nhóm, nhóm sẽ sử dụng NLP nhằm xử lý ngôn ngữ tự nhiên do người dùng nhập vào khi đặt câu hỏi trên diễn đàn, từ đó gợi ý cho người dùng các tag cụ thể khi đặt vấn đề (VD: hệ thống phân tích thấy người dùng đang đặt câu hỏi về ngôn ngữ C# sẽ gợi ý tag C#)

**Định nghĩa về NLP**

- Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh trong lĩnh vực Trí tuệ nhân tạo (AI), tập trung chủ yếu vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người. Mục tiêu chính của NLP là giúp cho máy tính hiểu và thực hiện được các việc liên quan đến ngôn ngữ của con người một cách hiệu quả hơn, nâng cao hiệu quả xử lý văn bản cũng như lời nói.

- NLP được xây dựng dựa trên ngôn ngữ học phức tạp, các nguyên lý thống kê và neural network algorithms. Các NLP program có khả năng đọc và hiểu văn bản với tốc độ cao mà không quan trọng kích thước của văn bản có lớn đến cỡ nào. Một NLP program có khả năng xử lý được tất cả các văn bản và từ đó đưa ra được những thông tin có giá trị quan trọng (được gọi là knowledge) cho các doanh nghiệp tương tác nhiều với dữ liệu. VD: thông tin về khách hàng trong lĩnh vực marketing, sales, CRM, social media,...

**Những khó khăn trong xử lý ngôn ngữ tự nhiên**

- Ngôn ngữ tự nhiên là thứ xuất phát từ cảm xúc của con người, do đó chúng không hề có một quy tắc nhất định hay tuân thủ theo các logic nào về mặt cú pháp, ngữ nghĩa. Bản thân của ngôn ngữ không mang một cấu trúc nhất định và điều đó khiến cho việc xử lý ngôn ngữ trên máy tính trở nên khó khăn hơn.

- Ngôn ngữ tự nhiên đôi lúc cần phải có một ngữ cảnh nhất định để truyền tải thông điệp.

- Ngôn ngữ tự nhiên có thể gây nhầm lẫn (từ đồng nghĩa, từ nhiều nghĩa,...)

**Những bài toán cơ bản trong NLP**

- Xử lý ngôn ngữ tự nhiên bao gồm Natural Language Understanding (NLU) và Natural Language Generation (NLG).

- NLU bao gồm các bước chính:

+ Phân tích hình vị: là việc nhận biết, phân tích và miêu tả cấu trúc của những hình vị trong ngôn ngữ như từ gốc, phụ tố, từ loại,...

+ Phân tích cú pháp: phân tích một chuỗi các biểu tượng ở dạng ngôn ngữ tự nhiên/máy tính dựa theo các grammar như context-free grammar, combinatory categorial grammar, dependency grammar

+ Phân tích ngữ nghĩa: bao gồm 2 mức độ: **ngữ nghĩa từ vựng** biểu hiện các ý nghĩa của những từ trong một chuỗi và phân biệt nghĩa của từ; **ngữ nghĩa thành phần** liên quan đến cách thức các từ liên kết với nhau và hình thành ra nghĩa rộng hơn.

+ Phân tích diễn ngôn: phân tích về mối quan hệ giữa ngôn ngữ và ngữ cảnh mà ngôn ngữ được sử dụng (context-of-use)

- NLG là việc tạo ra các ngôn ngữ tự nhiên từ máy, NLG đóng vai trò quan trọng trong các ứng dụng xử lý ngôn ngữ tự nhiên như: tương tác người-máy, dịch thuật máy, và tóm tắt văn bản tự động.

**Các ứng dụng của NLP**

**-** Truy xuất thông tin: truy xuất, tìm ra các thông tin, tài liệu không có cấu trúc từ các nguồn tổng hợp lớn. VD: Google Search.

- Trích lọc thông tin: trích lọc ra các thông tin cụ thể chính xác mà người dùng cần.

- Trả lời câu hỏi: tự động trả lời câu hỏi của con người bằng việc truy xuất thông tin từ các nguồn tài liệu lớn.

- Tóm tắt văn bản tự động: thu gọn văn bản đầu vào và đưa ra một một bản tóm tắt ngắn gọn với những nội dung quan trọng nhất trong bản gốc.

- Dịch máy (machine translation): sử dụng máy tính để tự động hóa việc dịch thuật từ ngôn ngữ sang ngôn ngữ

Bài viết có tham khảo từ nguồn: <https://techinsight.com.vn/xu-li-ngon-ngu-tu-nhien-nhung-dieu-can-biet/>