Xử lý ngôn ngữ tự nhiên

**Xử lý ngôn ngữ tự nhiên** (*natural language processing* - NLP) là một nhánh của [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) tập trung vào các ứng dụng trên [ngôn ngữ](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF) của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của [tư duy](https://vi.wikipedia.org/wiki/T%C6%B0_duy) và [giao tiếp](https://vi.wikipedia.org/wiki/Giao_ti%E1%BA%BFp).

**Các bước xử lý**

* **Phân tích hình thái** - Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong [tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong [tiếng Việt](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Vi%E1%BB%87t), dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ. Cùng với các ngôn ngữ như [tiếng Trung](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Trung_Qu%E1%BB%91c), [tiếng Hàn](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Tri%E1%BB%81u_Ti%C3%AAn),[tiếng Nhật](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Nh%E1%BA%ADt), [phân tách từ](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_t%E1%BB%AB_v%E1%BB%B1ng) trong tiếng Việt là một công việc không hề đơn giản.
* **Phân tích cú pháp** - Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các luật văn phạm.
* **Phân tích ngữ nghĩa** - Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ [phân tích cú pháp](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_c%C3%BA_ph%C3%A1p).
* **Tích hợp văn bản** - Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.
* **Phân tích thực nghĩa** - Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.

Tuy nhiên, ranh giới giữa 5 bước xử lý này cũng rất mong manh. Chúng có thể được tiến hành từng bước một, hoặc tiến hành cùng lúc - tùy thuộc vào giải thuật và ngữ cảnh cụ thể.

**Các bài toán và ứng dụng**

* [Nhận dạng chữ viết](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt): Có hai kiểu nhận dạng, thứ nhất là nhận dạng chữ in, ví dụ nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc của [Microsoft Word](https://vi.wikipedia.org/wiki/Microsoft_Word) chẳng hạn. Phức tạp hơn là [nhận dạng chữ viết tay](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt_tay), có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng và thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).
* [Nhận dạng tiếng nói](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i): Nhận dạng tiếng nói rồi chuyển chúng thành văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.
* [Tổng hợp tiếng nói](https://vi.wikipedia.org/wiki/T%E1%BB%95ng_h%E1%BB%A3p_gi%E1%BB%8Dng_n%C3%B3i): Từ một văn bản tự động tổng hợp thành tiếng nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang [web](https://vi.wikipedia.org/wiki/World_Wide_Web), nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, tổng hợp tiếng nói là sự trợ giúp tốt cho [người khiếm thị](https://vi.wikipedia.org/wiki/Ng%C6%B0%E1%BB%9Di_khi%E1%BA%BFm_th%E1%BB%8B), nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa [robot](https://vi.wikipedia.org/wiki/Robot) với người.
* [Dịch tự động](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng) (*machine translate*): Như tên gọi đây là chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Một phần mềm điển hình về tiếng Việt của chương trình này là [Evtrans](https://vi.wikipedia.org/w/index.php?title=Evtrans&action=edit&redlink=1) của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web vdict.com mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Tháng 10 năm 2008 có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và [Google](https://vi.wikipedia.org/wiki/Google), một thời gian sau đó Xalo\_vn cũng đưa ra dịch vụ tương tự.
* [Tìm kiếm thông tin](https://vi.wikipedia.org/wiki/T%C3%ACm_ki%E1%BA%BFm_th%C3%B4ng_tin) (*information retrieval*): Đặt câu hỏi và chương trình tự tìm ra nội dung phù hợp nhất. Thông tin ngày càng đầy lên theo cấp số nhân, đặc biệt với sự trợ giúp của [internet](https://vi.wikipedia.org/wiki/Internet) việc tiếp cận thông tin trở lên dễ dàng hơn bao giờ hết. Việc khó khăn lúc này là tìm đúng nhất thông tin mình cần giữa bề bộn tri thức và đặc biệt thông tin đó phải đáng tin cậy. Các [máy tìm kiếm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) dựa trên giao diện web như Google hay [Yahoo](https://vi.wikipedia.org/wiki/Yahoo!) hiện nay chỉ phân tích nội dung rất đơn giản dựa trên tần suất của từ khoá và thứ hạng của trang và một số tiêu chí đánh giá khác để đưa ra kết luận, kết quả là rất nhiều tìm kiếm không nhận được câu trả lời phù hợp, thậm chí bị dẫn tới một liên kết không liên quan gì do thủ thuật đánh lừa của các trang web nhằm giới thiệu sản phẩm (có tên tiếng Anh là [SEO](https://vi.wikipedia.org/wiki/T%E1%BB%91i_%C6%B0u_h%C3%B3a_c%C3%B4ng_c%E1%BB%A5_t%C3%ACm_ki%E1%BA%BFm) viết tắt của từ *search engine optimization*). Thực tế cho đến bây giờ chưa có máy tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người trừ trang www.ask.com được đánh giá là "hiểu" được những câu hỏi có cấu trúc ở dạng đơn giản nhất. Mới đây cộng đồng mạng đang xôn xao về trang [Wolfram Alpha](https://vi.wikipedia.org/wiki/Wolfram_Alpha), được hứa hẹn là có khả năng [hiểu ngôn ngữ tự nhiên](https://vi.wikipedia.org/wiki/Hi%E1%BB%83u_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn) của con người và đưa ra câu trả lời chính xác[[1]](https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn#cite_note-1). Lĩnh vực này hứa hẹn tạo ra bước nhảy trong cách thức tiếp nhận tri thức của cả cộng đồng.
* [Tóm tắt văn bản](https://vi.wikipedia.org/wiki/T%C3%B3m_t%E1%BA%AFt_t%E1%BB%B1_%C4%91%E1%BB%99ng): Từ một văn bản dài tóm tắt thành một văn bản ngắn hơn theo mong muốn nhưng vẫn chứa những nội dung thiết yếu nhất.
* [Khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u) (*data mining*) và [phát hiện tri thức](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A1t_hi%E1%BB%87n_tri_th%E1%BB%A9c&action=edit&redlink=1): Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình [học tập](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_t%E1%BA%ADp), khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển. Ở mức độ đơn giản khi kết hợp với [máy tìm kiếm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) nó cho phép đặt câu hỏi để từ đó công cụ tự tìm ra câu trả lời dựa trên các thông tin trên web mặc cho việc trước đó có câu trả lời lưu trên web hay không (giống như trang Yahoo! hỏi và đáp, nơi chuyên đặt các câu hỏi để người khác trả lời), nói một cách nôm na là nó đã biết xử lý dữ liệu để trả lời câu hỏi của người sử dụng, thay vì máy móc đáp trả những gì chỉ có sẵn trong bộ nhớ.
* [Sửa lỗi chính tả](https://vi.wikipedia.org/wiki/S%E1%BB%ADa_l%E1%BB%97i_ch%C3%ADnh_t%E1%BA%A3): Là quá trình phát hiện và sửa các lỗi chính tả trong một đoạn văn bản. Phần mềm sửa lỗi chính tả thường được tích hợp trong các ứng dụng văn phòng như [Microsoft Word](https://vi.wikipedia.org/wiki/Microsoft_Word) hay [Google Docs](https://vi.wikipedia.org/wiki/Google_Docs) để hỗ trợ người dùng trong việc soạn thảo văn bản. Sửa lỗi chính tả có thể được áp dụng cho nhiều ngôn ngữ khác nhau, bao gồm tiếng Việt.