1. Từ điển

Con người khi đọc và viết đều phải sử dụng những kiến thức về ngôn ngữ. Trong NLP cũng vậy, máy tính cũng cần những kiến thức về ngôn ngữ như thế, ví dụ như từ điển, ngữ pháp, câu văn ví dụ, …

1.1 Từ điển đơn ngữ (TĐĐN)

Từ điển đơn ngữ là 1 tài nguyên rất quan trọng và cần thiết trong NLP. Những thông tin lấy từ TĐĐN được xử dụng nhiều nhất trong phân tích hình thái (phân tích từ/cụm từ) và phân tích ý nghĩa của từ.

Sau đây là định nghĩa về từ điển trên Wikipedia :

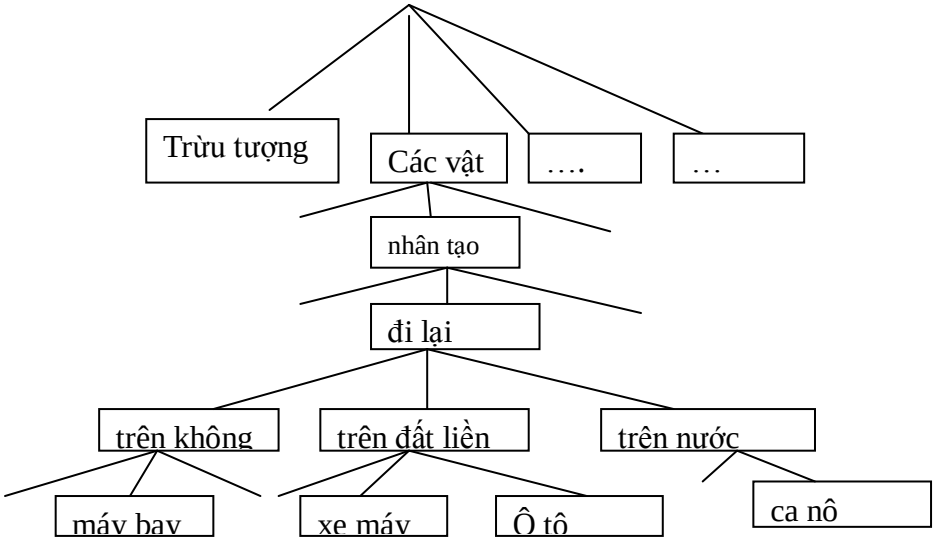
Bộ sách cho danh sách các từ (được sắp xếp theo thứ tự ABC), thường thường giải thích ý nghĩa, từ nguyên, cách dùng, dịch, cách phát âm, và thường kèm theo các ví dụ về cách sử dụng từ đó.

Ngoài ra, bạn có thể thao khảo thêm về cuốn từ điển tiếng Việt do Viện Ngôn ngữ học biên soạn. Nhưng từ điển dành cho NLP có đôi chút khác biệt và phức tạp hơn. Ví dụ, để phục vụ cho quá trình phân tích ý nghĩa, từ điển phải cung cấp được các thông tin như : các từ gần nghĩa, ví dụ thông dụng, … Hiện tại đang có đề tài VLSP do giáo sư Hồ Quốc Bảo chủ trì có cung cấp 1 từ điển dành cho NLP. (Mặc dù đã liên lạc bằng email tới địa chỉ được ghi trên website, nhưng sau gần 1 tháng tôi vẫn chưa nhận được hồi âm nên chưa thể đánh giá được về chất lượng của cuốn từ điển này, nhưng đề tài VLSP được thực hiện bởi những chuyên gia hàng đầu của Việt Nam về NLP nên tôi rất kì vọng vào chất lượng của nó).

1.2 Từ điển thesaurus

Từ điển thesaurus là 1 dạng từ điển được sắp xếp theo các tầng khái niệm (tầng ý nghĩa) được biểu diễn dưới dạng cây nhiều tầng. Thesaurus cung cấp cho NLP 1 thông tin rất quan trọng là "**độ giống nhau**" giữa các từ.

Sau đây là 1 ví dụ đơn giản (mang tính hình dung) về thesaurus.



Thesaurus được biểu diễn dưới dạng cây nên chúng ta sẽ sử dụng 1 số khái niệm của cây như độ sâu, nút cha chung gần nhất, …
Độ giống nhau của 2 từ được tính bởi công thức :

$$sim(w_i, w_j) = \frac{d_c \times 2}{d_i + d_j}$$

Trong đó, w_i, w_j là 2 từ, d_i, d_j là độ sâu của 2 từ trên cây theraurus, d_c là độ sâu của nút cha chung gần nhất. Dễ dàng nhận thấy 0<=sim(*,*)<=1.
Với cây theraurus được ví dụ ở trên, ta có :
Sim(“xe máy”,”ô tô”) = 4*2/(5+5)=0.8
Sim(“xe máy”,”máy bay”)=3*2/(5+5)=0.6

Hiện tại, tôi chưa thấy có 1 dự án nào về xây dựng từ điển Thesaurus nào cho tiếng Việt. Đối với tiếng Anh và tiếng Nhật, bạn có thể tham khảo ở đây :

- Thesaurus của Roget
- Wordnet tiếng Anh với 150.000 từ.
- Từ ngữ tiếng Nhật (nihongo-goi-taikei) với khoảng 300.000 từ.
- Wordnet tiếng Nhật với 90.000 từ.