**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**Báo cáo đồ án**

**XÂY DỰNG KHO NGỮ LIỆU SONG NGỮ ANH – VIỆT**

Môn: Nhập Xử lý Ngôn ngữ tự nhiên

**Sinh viên thực hiện:**

Đặng Đỗ Huỳnh Như - 18120219

Nguyễn Dương Binl - 18120292

**Giáo viên hướng dẫn:**

TS. Đinh Điền

ThS. Lương An Vinh

ThS.

Mục lục

[I. GIỚI THIỆU VỀ KHO NGỮ LIỆU SONG NGỮ 3](#_Toc59325335)

[1. Ngữ liệu 3](#_Toc59325336)

[2. Kho ngữ liệu song ngữ 3](#_Toc59325337)

[3. Ứng dụng của kho ngữ liệu song ngữ 3](#_Toc59325338)

[II. CÁC BƯỚC XÂY DỰNG KHO NGỮ LIỆU SONG NGỮ 3](#_Toc59325339)

[1. Thu thập ngữ liệu 3](#_Toc59325340)

[2. Gióng hàng cho câu 4](#_Toc59325341)

[3. Lưu kết quả 4](#_Toc59325342)

[III. THAM KHẢO 4](#_Toc59325343)

# I. GIỚI THIỆU VỀ KHO NGỮ LIỆU SONG NGỮ

## 1. Ngữ liệu

Trong việc nghiên cứu, giảng dạy ngôn ngữ, ta cần thống kê, so sánh, đối chiếu để tìm ra các quy luật của ngôn ngữ, quy luật chuyển ngữ, các điểm tương đồng và dị biệt ở các bình diện khác nhau, các cấp độ khác nhau giữa các ngôn ngữ. Nhưng để thống kê, so sánh, đối chiếu như trên, ta cần phải có cứ liệu của các ngôn ngữ mà ta đang cần so sánh, ta gọi đó là “ngữ liệu” (corpus). Ngữ liệu là những “dữ liệu, cứ liệu của ngôn ngữ”, tức là những chứng cứ thực tế sử dụng ngôn ngữ.

## 2. Kho ngữ liệu song ngữ

Tập hợp các mảnh ngôn ngữ được chọn lựa và sắp xếp theo một số tiêu chí ngôn ngữ học rõ ràng và được gióng hàng của hai ngôn ngữ.

Trong ngữ liệu song ngữ, các bản dịch tương ứng của mỗi ngôn ngữ phải được đặt song song với nhau hay còn được gọi là gióng hàng với nhau (alignment).

## 3. Ứng dụng của kho ngữ liệu song ngữ

Ngữ liệu nói chung và ngữ liệu song ngữ Anh-Việt nói riêng sẽ giúp chúng ta rất nhiều trong vô vàn các ứng dụng khác nhau. Từ lĩnh vực Ngôn ngữ học so sánh đối chiếu, cho đến việc giảng dạy ngoại ngữ nói chung, môn biên dịch nói riêng, ngữ liệu song ngữ Anh-Việt sẽ giúp cho người học tự “nghiệm” ra các quy luật chuyển ngữ mà các cách tiếp cận truyền thống không thể bao quát hết được. Đặc biệt trong công tác dịch thuật chuyên nghiệp, kho ngữ liệu Anh – Việt này sẽ làm giảm đáng kể công sức dịch thuật thủ công. Nếu kho ngữ liệu này được tiếp tục cập nhật thì hiệu quả khai thác càng tăng gấp bội. Ngoài ra, cách thức này có thể áp dụng cho các cặp ngôn ngữ khác.

# II. CÁC BƯỚC XÂY DỰNG KHO NGỮ LIỆU SONG NGỮ

Dù là kho ngữ liệu đơn ngữ hay song ngữ, cả hai đều phải trải qua những công việc chung, như: thu thập ngữ liệu, chuẩn hoá ngữ liệu và gán nhãn ngữ liệu. Riêng phần gán nhãn ngữ liệu với kho ngữ liệu song ngữ chúng ta phải giải quyết thêm bài toán gióng câu, gióng hàng. Từ kho ngữ liệu song ngữ Anh – Việt chúng ta có thể áp dụng để xây dựng kho ngữ liệu đơn ngữ hay cho ngữ liệu song ngữ khác. Ở đây chúng ta thực hiện xây dựng kho ngữ liệu song ngữ không gán nhãn nên chỉ dừng lại ở việc gióng hàng để biết câu nào là bản dịch của câu nào thôi.

## 1. Thu thập ngữ liệu

Có nhiều nguồn tài liệu mà chúng ta có thể thu thập được bao gồm tài liệu sách vở và tài liệu điện tử. Nhưng để cho thuận tiện thì đồ án này ưu tiên khai thác và sử dụng nguồn tài liệu điện tử vì dễ khai thác và xử lí.

Tiêu chí để thu thập ngữ liệu là ngữ liệu phải đúng ngữ pháp, được nhiều người chấp nhận và sử dụng. Về cách dịch thì phải dịch sát nghĩa, không được dịch thoát ý, dịch tóm lượt hay dịch theo kiểu giải thích, diễn giải. Chọn các mẫu ngữ liệu trong lĩnh vực khoa học kỹ thuật và các câu thông thường, không chọn trong lĩnh vực văn học vì độ đa dạng từ trong các tác phẩm văn học là rất lớn, sử dụng những từ ngữ hoa mỹ ít được sử dụng trong thực tế.

Đồ án này sử dụng nguồn ngữ liệu từ từ điển song ngữ Anh – Việt của trang web từ điển Glosbe, các bài viết song ngữ trên trang web của tổ chức UNICEF tại Việt Nam.

Ngữ liệu sau khi thu thập ít nhiều cũng có những sai sót hoặc những kí tự đặc biệt nên cần chỉnh sửa lại.

## 2. Gióng hàng cho câu

Quan sát thấy ngữ liệu song ngữ thu thập được đã có sự tương đồng về văn bản và đoạn văn, do đó ta chỉ cần tiến hành gióng hàng câu cho ngữ liệu thu thập được.

Các bước gióng hàng câu:

* Tách đoạn thành các câu:

Ta sử dụng thư viện spaCy để tách câu trong tiếng Anh và thư viện pyvi để tách câu trong tiếng Việt.

* Tìm cặp câu có khả năng là cặp câu song ngữ trong 2 đoạn tiếng Anh và tiếng Việt tương ứng:

Thứ tự các câu trong 2 đoạn song ngữ tiếng Anh và tiếng Việt không phải lúc nào cũng là tương ứng nhau, vì có những câu tiếng Anh được dịch thành 2 – 3 câu tiếng Việt tương ứng, và ngược lại. Do đó ta cần so khớp 1 câu tiếng Anh với 5 câu tiếng Việt có vị trí gần nó trong đoạn tiếng Việt song ngữ. Vị trí của 5 câu này là: câu có vị trí tương ứng với nó, 2 câu trước và 2 câu sau vị trí đó.

* Tiền xử lý câu:

Trong câu tiếng Anh, có các từ chỉ mang nghĩa ngữ pháp như: the, a, an, am, is, are, … Tương tự, trong tiếng Việt cũng có các từ chỉ mang nghĩa ngữ pháp như: bị, được, đã, đang, sẽ, cái, bức, quyển, … Do đó ta cần loại bỏ những từ này khỏi câu. Để làm được điều này, nhóm em đã thống kê và tổng hợp các từ thuộc dạng trên vào trong file e\_stopwords.txt và v\_stopwords.txt, hàm Process\_E\_Sentence và Process\_V\_Sentence sẽ đọc danh sách các từ trong 2 file này và loại nó ra khỏi câu.

Trong câu tiếng Anh, các từ như: learned, learning, best, … không có trong từ điển, do đó để tra nghĩa của từ ta cần đưa nó về dạng lemma. Ta sử dụng thư viện spaCy với lệnh .lemma\_ để thực hiện việc này. Tuy nhiên, khi sử dụng spaCy, những đại từ quan hệ như he, she, it, … sẽ bị chuyển thành -PRON-, do đó ta cần bổ sung các từ này và nghĩa của chúng vào trong 2 file stopwords ở trên.

* So sánh độ dài câu:

Nếu 2 câu có số lượng từ chênh lệch quá lớn thì khó thể nào là cặp song ngữ của nhau. Do đó để tránh mất nhiều thời gian và công sức, ta tiến hành tính tỉ lệ độ dài của 2 câu input trước, nếu tỉ lệ độ dài của 2 câu lớn hơn hoặc bằng 1 ngưỡng alpha nào đó, thì ta tiến hành bước xét độ tương đồng về ngữ nghĩa giữa chúng, ngược lại ta có thể nhanh chóng kết luận được đây không phải là cặp câu song ngữ của nhau.

Ngưỡng alpha mà nhóm quyết định chọn là: alpha = 0.7

Sử dụng thư viện spaCy để tách từ tiếng Anh và thư viện pyvi để tách từ tiếng Việt trong câu.

* So sánh độ tương đồng ngữ nghĩa giữa 2 câu:

Vì nhận thấy việc tra từ điển Anh – Việt là dễ dàng hơn nên ta sẽ thực hiện tra từng từ trong câu tiếng Anh để tạo ra 1 câu dịch gồm nghĩa tiếng Việt của các từ tiếng Anh đó, sau đó so sánh độ tương đồng giữa câu dịch đó với câu tiếng Việt song ngữ được đưa vào.

Sử dụng file từ điển Anh – Việt tên anhviet.txt, tiến hành các bước tiền xử lý file này để thu được một dictionary trong python.

Việc ánh xạ 1-1 giữa 1 từ tiếng Anh và 1 từ tiếng Việt với nghĩa tương ứng là rất khó. Ví dụ:

Theo từ điển, từ student có nghĩa là học sinh, sinh viên đại học.

Tuy nhiên trong câu tiếng Việt tương ứng chỉ có từ sinh viên, ta nhận thấy 2 từ sinh viên và sinh viên đại học đều có nghĩa như nhau, nhưng chương trình lại không thể nhận biết được điều đó vì về hình thức 2 từ này khác nhau.

Để tránh việc nhập nhằng như trên, chúng em đã đưa ra ý tưởng:

* Với mỗi từ tiếng Anh, khi tra từ điển và thu được các từ tiếng Việt có nghĩa tương ứng, trong các từ dịch tiếng Việt thu được, tìm từ nào có nhiều hình vị trùng với các hình vị trong câu tiếng Việt input nhất, sau đó thêm từ đó vào câu dịch.
* Tìm tỉ lệ giao nhau giữa các hình vị giữa câu dịch thu được với câu tiếng Việt input, nếu tỉ lệ này lớn hơn hoặc bằng một ngưỡng beta nào đó thì ta đưa ra kết luận 2 câu input là bản dịch song ngữ của nhau
* Qua quan sát, nhóm em nhận thấy ngưỡng beta = 0.5 là hợp lí.

## 3. Lưu kết quả

Lưu cặp câu song ngữ đã gióng hàng vào file có định dạng .txt.

Cấu trúc lưu trữ:

[thứ tự câu \_ loại ngôn ngữ] câu

Vd: [1\_e] Draw a line from A to B.

[1\_v] Vẽ một đường từ A đến B.

# III. THAM KHẢO

Giáo trình Xử lí Ngôn ngữ tự nhiên, Đinh Điền, 2006

Trang web Trung tâm ngôn ngữ học tính toán: [www.clc.hcmus.edu.vn](http://www.clc.hcmus.edu.vn)

Trang từ điển song ngữ: www.glosbe.com

Trang web của UNICEF tại Việt Nam: www.unicef.org/vietnam