# V. Xây dựng bộ dữ liệu Vietnamese-Laos

## 5.1. Xác định mục tiêu và phạm vi của bộ dữ liệu

Việc xây dựng bộ dữ liệu song ngữ Việt - Lào đóng vai trò quan trọng trong các ứng dụng xử lý ngôn ngữ tự nhiên (NLP), đặc biệt trong lĩnh vực dịch máy, nhận dạng giọng nói và phân tích ngữ nghĩa. Để đảm bảo bộ dữ liệu có tính ứng dụng cao và phù hợp với nhu cầu thực tế, nghiên cứu này xác định rõ mục tiêu và phạm vi như sau:

**1. Mục tiêu**

Bộ dữ liệu được thiết kế nhằm phục vụ các tác vụ chính trong NLP. Cụ thể, trong lĩnh vực dịch máy, bộ dữ liệu cung cấp nguồn tài nguyên huấn luyện cho các mô hình dịch tự động giữa tiếng Việt và tiếng Lào. Bên cạnh đó, nó hỗ trợ xây dựng hệ thống nhận dạng giọng nói và chuyển đổi giọng nói thành văn bản, góp phần nâng cao hiệu suất của các ứng dụng liên quan. Ngoài ra, bộ dữ liệu còn đóng vai trò quan trọng trong phân tích ngữ nghĩa, giúp cải thiện độ chính xác của các mô hình xử lý văn bản song ngữ.

**2. Quy mô và phân bố dữ liệu**

Bộ dữ liệu được phát triển với khoảng 2000 cặp câu song ngữ, nhằm đảm bảo độ bao phủ rộng và đa dạng về ngữ cảnh. Các câu trong bộ dữ liệu có mức độ phức tạp khác nhau, từ các cấu trúc câu đơn giản đến các câu có cú pháp phức tạp hơn. Sự phân bố này giúp tối ưu hóa khả năng học của mô hình NLP đối với cả hai ngôn ngữ.

**3. Lĩnh vực nội dung**

Nội dung của bộ dữ liệu tập trung vào các lĩnh vực có tính ứng dụng cao, đặc biệt là báo chí và truyền thông. Cụ thể, bộ dữ liệu bao gồm các văn bản thuộc nhiều chuyên mục như tin tức chính trị, kinh tế và xã hội, qua đó cung cấp ngữ liệu phong phú và đa dạng cho các ứng dụng NLP trong thực tế.

**4. Đặc điểm ngôn ngữ**

Bộ dữ liệu chủ yếu sử dụng văn phong chính quy nhằm đảm bảo tính chính xác và phù hợp với các mô hình NLP. Đồng thời, chỉ bao gồm các dạng ngôn ngữ chuẩn của tiếng Việt và tiếng Lào, không đề cập đến phương ngữ địa phương, nhằm duy trì sự nhất quán trong quá trình huấn luyện mô hình.

Bộ dữ liệu này được kỳ vọng sẽ trở thành một nguồn tài nguyên quan trọng cho các nghiên cứu và ứng dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên, đặc biệt là trong dịch thuật, nhận dạng giọng nói và phân tích ngữ nghĩa giữa hai ngôn ngữ Việt - Lào.

# 5.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu đóng vai trò quan trọng trong việc xây dựng bộ dữ liệu song ngữ Việt - Lào nhằm đảm bảo tính nhất quán, chất lượng và khả năng ứng dụng. Quá trình này đối mặt với nhiều thách thức, bao gồm việc văn bản chứa các câu dài hoặc không đúng định dạng, sự xuất hiện của ngôn ngữ khác ngoài tiếng Việt và tiếng Lào, lỗi chính tả hoặc định dạng không thống nhất, cũng như sự chênh lệch lớn về độ dài giữa hai ngôn ngữ, có thể ảnh hưởng đến hiệu quả huấn luyện mô hình.

Để giải quyết những vấn đề này, quá trình tiền xử lý bao gồm nhiều bước. Trước tiên, cấu trúc câu được chuẩn hóa bằng cách chỉ bao gồm các cặp câu đơn lẻ thay vì đoạn văn dài, đồng thời loại bỏ hoặc tách riêng những câu chứa nhiều mệnh đề không phù hợp. Tiếp theo, các câu chứa từ hoặc cụm từ thuộc ngôn ngữ khác ngoài tiếng Việt và tiếng Lào, cũng như các câu có ký tự đặc biệt hoặc lỗi mã hóa, đều được loại bỏ để duy trì chất lượng dữ liệu.

Bên cạnh đó, các chuyên gia ngôn ngữ tiến hành kiểm tra và sửa lỗi chính tả nhằm đảm bảo độ chính xác của văn bản. Để duy trì tính đồng nghĩa giữa các cặp câu song ngữ, các chuyên gia cũng thực hiện quá trình xác minh để đảm bảo rằng cả hai câu trong một cặp đều truyền tải cùng một ý nghĩa. Ngoài ra, văn bản được chuẩn hóa về dấu câu, khoảng trắng và mã hóa ký tự nhằm đảm bảo tính đồng nhất và phù hợp với các hệ thống NLP. Cuối cùng, để cân bằng giữa tính tự nhiên của câu và tính khả dụng của bộ dữ liệu trong huấn luyện mô hình, sự chênh lệch độ dài giữa hai ngôn ngữ được giới hạn ở mức tối đa 5 từ.

Những bước tiền xử lý này giúp nâng cao chất lượng dữ liệu, tối ưu hóa khả năng áp dụng vào các mô hình NLP, đồng thời đảm bảo độ chính xác và tính nhất quán trong xử lý ngôn ngữ song ngữ Việt - Lào.

# 5.3. Chọn lọc các cặp câu

Sau khi thực hiện tiền xử lý và áp dụng các quy tắc chuẩn hóa ngữ liệu, việc chọn lọc các cặp câu đóng vai trò quan trọng trong việc đảm bảo chất lượng của bộ dữ liệu. Quá trình này được thực hiện với sự cân nhắc của các chuyên gia ngôn ngữ nhằm tạo ra một tập dữ liệu có phân bố điểm tương đồng hợp lý, giúp tối ưu hóa việc huấn luyện các mô hình NLP.

Việc lựa chọn cặp câu dựa trên một số tiêu chí quan trọng. Trước tiên, cặp câu phải có mức độ tương đồng về ngữ nghĩa theo thang điểm từ 0 đến 4, đảm bảo tính chính xác trong việc đánh giá mức độ đồng nghĩa. Đồng thời, để tránh sự chênh lệch trong dữ liệu, bộ dữ liệu phải có sự phân bố đồng đều giữa các mức điểm, tránh tình trạng bị lệch về một nhóm điểm cụ thể. Ngoài ra, các cặp câu được chọn lọc cần đảm bảo đúng cấu trúc ngữ pháp, không có lỗi chính tả và có độ dài phù hợp để tránh mất cân bằng trong tập dữ liệu.

Một vấn đề quan trọng phát sinh khi chọn ngẫu nhiên các cặp câu từ tập dữ liệu ban đầu là hơn 99% các cặp câu sẽ có điểm số 0 hoặc 1. Điều này có thể gây ra sự thiên lệch trong quá trình huấn luyện, khiến mô hình học máy kém hiệu quả khi xử lý các mức độ tương đồng cao hơn. Đồng thời, sự phân bố không đồng đều này làm thiếu hụt dữ liệu ở các mức điểm 2, 3 và 4, trong khi đây là những mức điểm quan trọng giúp mô hình nhận diện được sự khác biệt tinh tế trong ý nghĩa của các câu.

Để giải quyết vấn đề này, phương pháp chọn lọc có kiểm soát được áp dụng nhằm đảm bảo số lượng cặp câu ở mỗi mức điểm đồng nghĩa tương đối đồng đều. Ngoài ra, một tập hợp mẫu được chuyên gia ngôn ngữ đánh giá lại nhằm xác minh rằng các cặp câu thực sự phản ánh đúng mức độ tương đồng theo thang điểm đã đặt ra. Nhờ những điều chỉnh này, bộ dữ liệu đạt được sự cân bằng tốt hơn và có tính đa dạng cao hơn, từ đó phục vụ hiệu quả cho các ứng dụng NLP như dịch máy và phân tích ngữ nghĩa giữa tiếng Việt và tiếng Lào.

## 5.4 Quy trình chuẩn bị bộ dữ liệu Việt-Lào.

Trong nghiên cứu này, chúng tôi sử dụng một bộ dữ liệu ban đầu gồm 7.000 cặp câu tiếng Việt và tiếng Lào, được thu thập từ một website tin tức song ngữ. Các cặp câu trong tập dữ liệu gốc là bản dịch trực tiếp của nhau, tạo nền tảng cho quá trình xây dựng một tập dữ liệu có cấu trúc chặt chẽ và phù hợp cho các ứng dụng NLP.

Mục tiêu của nghiên cứu là xây dựng một tập dữ liệu gồm 2.000 cặp câu có phân bố mức độ tương đồng đồng đều, bao quát từ cặp câu đồng nghĩa hoàn toàn đến cặp câu không liên quan. Điều này giúp đảm bảo tính đa dạng trong dữ liệu, từ đó cải thiện khả năng tổng quát hóa của các mô hình NLP khi xử lý ngôn ngữ song ngữ Việt - Lào.

**1. Các bước chuẩn bị dữ liệu**

**Bước 1: Chọn lọc câu đồng nghĩa chất lượng cao**

Quá trình lựa chọn dữ liệu ban đầu được thực hiện một cách cẩn trọng. Các chuyên gia ngôn ngữ đã đánh giá và chọn lọc 700 cặp câu có chất lượng tốt nhất từ tập dữ liệu gốc, đảm bảo các tiêu chí sau:

* Độ chính xác về mặt ngữ nghĩa và tính đồng nghĩa cao giữa hai câu.
* Cấu trúc ngữ pháp chính xác và không chứa lỗi chính tả.

**Bước 2: Biến đổi và tạo tập dữ liệu đa dạng**

Để mở rộng tính đa dạng của bộ dữ liệu, chúng tôi tiến hành biến đổi câu tiếng Việt dựa trên 700 cặp câu đã chọn, tạo ra các mức độ tương đồng khác nhau. Cụ thể:

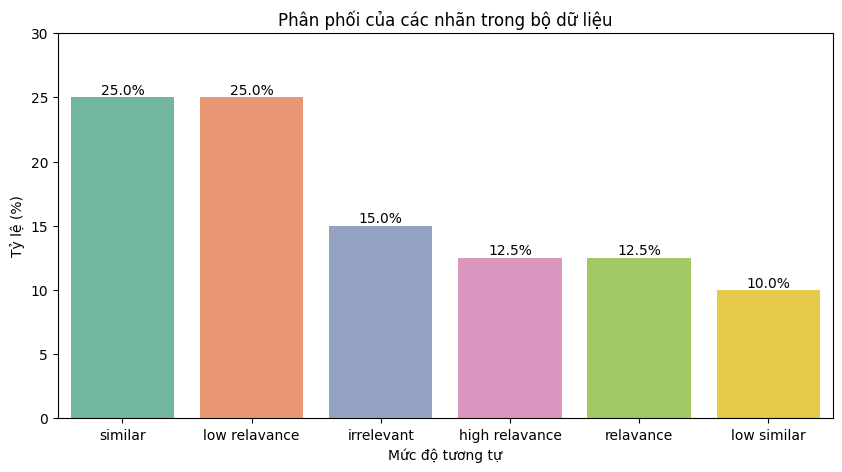
* **Tương đồng thấp:** Câu tiếng Việt được điều chỉnh một phần, giữ lại ý nghĩa chung nhưng có sự khác biệt đáng kể về cách diễn đạt.
* **Liên quan cao:** Câu tiếng Việt và tiếng Lào vẫn duy trì mối liên hệ chặt chẽ về ngữ nghĩa, nhưng có một số khác biệt nhỏ.
* **Liên quan thấp:** Cặp câu có một số yếu tố liên quan, nhưng nội dung không hoàn toàn giống nhau.
* **Trái nghĩa:** Một trong hai câu có nghĩa trái ngược với câu còn lại.
* **Không liên quan:** Cặp câu không có bất kỳ sự liên kết nào về ngữ nghĩa.

**Bước 3: Kiểm tra và tối ưu hóa tập dữ liệu**

Sau khi hoàn tất quá trình tạo tập dữ liệu, các chuyên gia tiếp tục đánh giá lại toàn bộ cặp câu nhằm đảm bảo:

* Không có lỗi ngữ pháp, chính tả hoặc sai lệch về ngữ nghĩa.
* Sự phân bố điểm số đồng đều, tạo ra một bộ dữ liệu cân bằng và phù hợp cho huấn luyện mô hình.

Với quy trình này, bộ dữ liệu được kỳ vọng sẽ cung cấp một nguồn tài nguyên chất lượng cao, phục vụ hiệu quả cho các nghiên cứu về dịch máy, phân tích ngữ nghĩa và các ứng dụng NLP khác.



Biểu đồ : phân phối nhãn trong tập dữ liệu

## 5.5. Quá trình chú thích

Quá trình chú thích dữ liệu là yếu tố thiết yếu trong việc đảm bảo tính chính xác và chất lượng của bộ dữ liệu. Để tối ưu hóa cả tốc độ và chất lượng, chúng tôi đã tiến hành quá trình này với sự tham gia của nhiều chuyên gia ngôn ngữ.

**1. Các bước Chú thích Cặp câu**

**Bước 1: Gán Điểm Mức độ Tương đồng**  
Mỗi cặp câu được chuyên gia thực hiện chú thích và gán điểm từ 0 đến 5, phản ánh mức độ tương đồng về ngữ nghĩa. Thang điểm được quy định như sau:

* 0 – Không liên quan
* 1 – Liên quan thấp
* 2 – Liên quan
* 2.5 – Liên quan cao
* 3 – Gần tương đương
* 4 – Tương đương

**Bước 2: Kiểm tra và Hiệu chỉnh**  
Mặc dù đã phát triển các tiêu chí và quy trình tiêu chuẩn, vẫn có khả năng xảy ra lỗi chính tả hoặc lỗi ngữ nghĩa trong các cặp câu. Do đó, sau khi gán điểm, các chuyên gia sẽ tiến hành kiểm tra và hiệu chỉnh các cặp câu, nếu cần thiết, nhằm đảm bảo rằng dữ liệu đạt tiêu chuẩn chính xác tối đa.

**2. Hướng dẫn Chú thích**

Để bảo đảm tính nhất quán giữa các chuyên gia, chúng tôi đã xây dựng một bảng hướng dẫn chi tiết mô tả quy trình gán điểm và các tiêu chí đánh giá. Tất cả các chuyên gia tham gia đều bắt buộc phải tuân thủ hướng dẫn này để duy trì tính đồng nhất và độ tin cậy của bộ dữ liệu.

Nhờ vào những biện pháp này, quá trình chú thích dữ liệu không chỉ nâng cao độ chính xác mà còn củng cố tính hợp lệ trong nghiên cứu và ứng dụng thực tiễn.

|  |  |  |
| --- | --- | --- |
| score | Định nghĩa | Nhãn |
| 4 | Hai câu là bản dịch của nhau | Tương đương |
| 3 | Hai câu gần như là bản dịch của nhau, nhưng có một vài chi tiết được được thêm vào. | Gần tương đương |
| 2.5 | Hai câu cùng chủ đề, có chi tiết giống nhau nhưng cách diễn đạt đã thay đổi | Liên quan cao |
| 2 | Hai câu cùng chủ đề nhưng các chi tiết quan trọng bị thay đổi về bản tính chất của câu. | Liên quan |
| 1 | Hai câu có một số chi tiết giống nhau | Liên quan thấp |
| 0 | Hai câu hoàn toàn khác biệt. | Không liên quan |

## 5.6. Tạo bộ dữ liệu

Để xây dựng một bộ dữ liệu phản ánh sự phân bố đồng nhất của các mức độ tương đồng, chúng tôi đã trích xuất 2.000 cặp câu từ tập dữ liệu ban đầu gồm 7.000 cặp câu song ngữ Việt-Lào thu thập từ các bài báo song ngữ.

**1. Quy trình Xây dựng Bộ dữ liệu**

**Bước 1: Chọn lọc Dữ liệu**  
Từ 7.000 cặp câu ban đầu, chúng tôi đã thực hiện chọn lọc 2.000 cặp câu, đảm bảo rằng bộ dữ liệu phản ánh sự đa dạng về các mức độ tương đồng. Các mức độ tương đồng được xác định bao gồm:

* Tương đương
* Gần tương đương
* Liên quan cao
* Liên quan
* Liên quan thấp
* Không liên quan

**Bước 2: Biến đổi và Tạo cặp câu**  
Phần câu tiếng Việt trong bộ dữ liệu gốc đã được các nhà ngôn ngữ học chỉnh sửa hoặc thay thế nhằm tạo ra sự đa dạng về các mức độ tương đồng. Các chuyên gia ngôn ngữ này thực hiện việc dịch lại câu tiếng Việt sang tiếng Lào, đảm bảo rằng các câu mới đạt tiêu chuẩn chất lượng cao về ngữ nghĩa.

**Bước 3: Chú thích Mức độ Tương đồng**  
Trong bước cuối cùng, mỗi cặp câu được gán hai loại nhãn:

* **Cột 3**: Điểm số đánh giá mức độ tương đồng (từ 0 đến 5), phản ánh mức độ liên quan giữa hai câu.
* **Cột 4**: Phân loại mức độ tương đồng ngữ nghĩa, bao gồm các nhãn như đồng nghĩa, liên quan một phần, không liên quan, v.v.

Quá trình xây dựng bộ dữ liệu này nhằm tăng cường tính chính xác và độ tin cậy trong việc phân tích ngữ nghĩa, đồng thời phục vụ cho các nghiên cứu và ứng dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên.

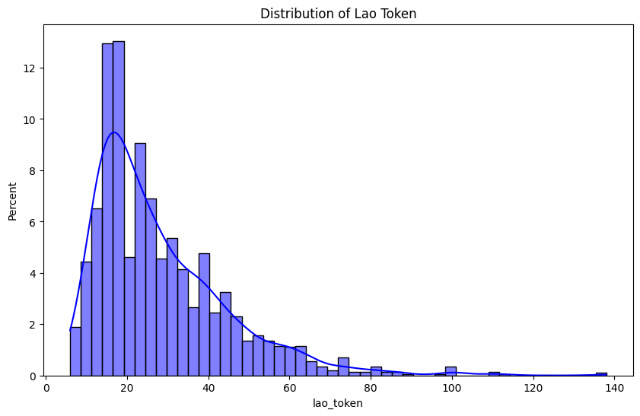
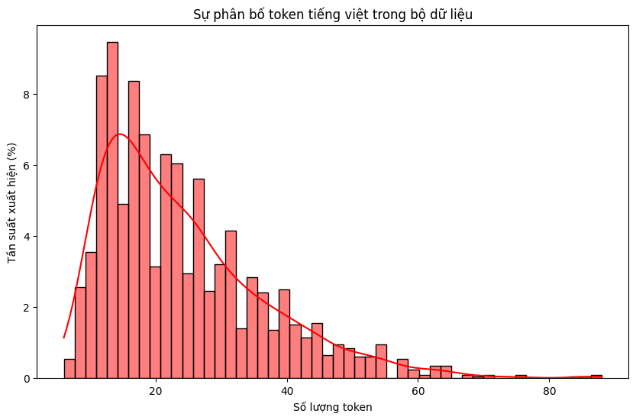
2. Cấu trúc Tập Dữ liệu

Dưới đây là một số ví dụ cấu trúc của tập dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| Laos | Vietnamese | Điểm tương đồng | Mức độ tương đồng |
| ໂດຍລັດຖະບານສະຫະລັດອາເມລິກາທີ່ໃຫ້ຂ່າວແລະຂໍ້ມູນຂ່າວສານແກ່ຫຼາຍກວ່າ 22 ປະເທດອາຣັບ. | Con chuột này có sự biến đổi khác mà biểu thị cho các loại mù loà ở con người . | 4 | Similar |
| ລາວເຊື່ອວ່າລັດຖະບານສາມາດສ້າງຂື້ນໂດຍປະຊາຊົນ, ແລະພວກເຮົາເອີ້ນພວກເຂົາວ່າພວກແຮັກເກີ້ພົນລະເມືອງ | Họ không tin rằng chính phủ có thể được xây dựng bởi nhân dân, và chúng tôi không gọi họ là những người lãnh đạo thành phố. | 3 | Low similar |
| ພວກເຮົາຕ້ອງຄິດຄືນໃໝ່ເຖິງຫຼັກທຳພື້ນຖານ ທີ່ເຮົາອີງໃສ່ເພື່ອສຶກສາລູກຫຼານຂອງເຮົາ. | Việc xem xét các nguyên tắc này sẽ giúp chúng ta tạo ra nền giáo dục giải quyết những thách thức của thế kỷ 21 và thúc đẩy phát triển bền vững. | 2.5 | High relavance |
| ເປັນຄັ້ງ ທຳ ອິດໃນປະຫວັດສາດຂອງດາວເຄາະນີ້, ພວກເຮົາສາມາດອອກແບບສິ່ງມີຊີວິດໂດຍກົງ. | Lần đầu tiên trong lịch sử thế giới này, chúng ta có khả năng thiết kế động vật trực tiếp. | 2 | Relavance |
| ສິ່ງນັ້ນໄດ້ເຮັດໃຫ້ຂ້ອຍສົນໃຈ, ເພາະວ່າຂ້ອຍໄດ້ຢູ່ກັບກ້ອງຖ່າຍຮູບຕະຫຼອດຊີວິດ. | Điều đó thực sự hấp dẫn tôi, bởi suốt cả cuộc đời, tôi đã gắn bó với chiếc máy quay và dành phần lớn thời gian để khám phá thế giới qua ống kính của nó. | 1 | Low relavance |
| ນີ້ແມ່ນຄັ້ງທີ 3 ທີ່ຫວຽດນາມ ໄດ້ເປັນເຈົ້າພາບຈັດຕັ້ງ Vesak, ພາຍຫຼັງສອງຄັ້ງຈັດຕັ້ງຢ່າງ ສຳເລັດຜົນໃນ ປີ 2008 ແລະ 2014. | Đây là lần thứ ba Việt Nam đăng cai Vesak, sau hai lần tổ chức thành công vào các năm 2008 và 2014. | 0 | Irrelevant |

## 5.7. Thống kê bộ dữ liệu

1. Phân tích phân phối độ dài câu trong tập dữ liệu song ngữ Việt - Lào



Trong quá trình xây dựng tập dữ liệu dịch song ngữ Việt - Lào, việc phân tích độ dài câu đóng vai trò quan trọng trong việc hiểu rõ đặc điểm ngữ liệu và tối ưu hóa mô hình xử lý ngôn ngữ tự nhiên (NLP). Hình X và Hình Y minh họa phân phối số lượng token (từ hoặc đơn vị ngôn ngữ) trong các câu tiếng Lào và tiếng Việt.

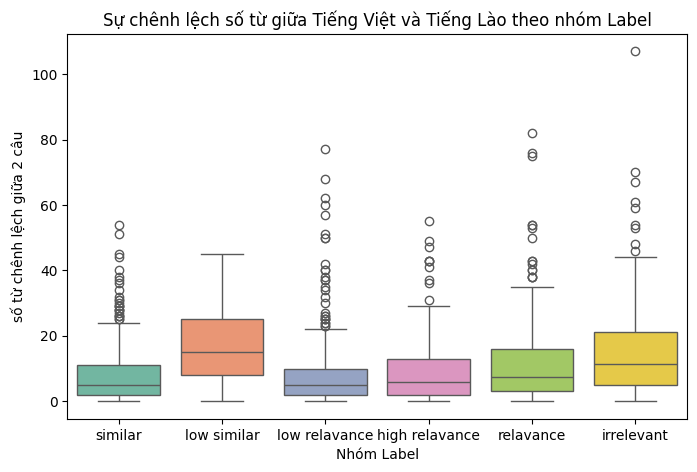
Kết quả phân tích cho thấy phân phối độ dài câu trong cả hai ngôn ngữ đều có dạng lệch phải, nghĩa là phần lớn các câu trong tập dữ liệu có độ dài tương đối ngắn, trong khi số lượng câu dài ít hơn đáng kể. Điều này phù hợp với đặc điểm chung của các ngôn ngữ tự nhiên, trong đó các câu ngắn thường xuất hiện với tần suất cao hơn.

Đáng chú ý, số lượng token trung bình trong câu tiếng Lào có xu hướng lớn hơn so với tiếng Việt. Sự chênh lệch này có thể xuất phát từ sự khác biệt về cấu trúc ngữ pháp cũng như cách biểu đạt ý nghĩa của hai ngôn ngữ. Trong khi tiếng Việt có xu hướng sử dụng câu ngắn và linh hoạt về trật tự từ, tiếng Lào có thể yêu cầu nhiều từ hơn để diễn đạt cùng một nội dung.

Sự khác biệt về độ dài câu giữa hai ngôn ngữ có thể ảnh hưởng đến hiệu quả của các mô hình dịch tự động. Các phương pháp như padding hoặc attention trong mạng nơ-ron cần được điều chỉnh để xử lý sự chênh lệch này, nhằm đảm bảo mô hình có thể học được các quy tắc dịch chính xác mà không bị ảnh hưởng bởi sự khác biệt về độ dài câu.

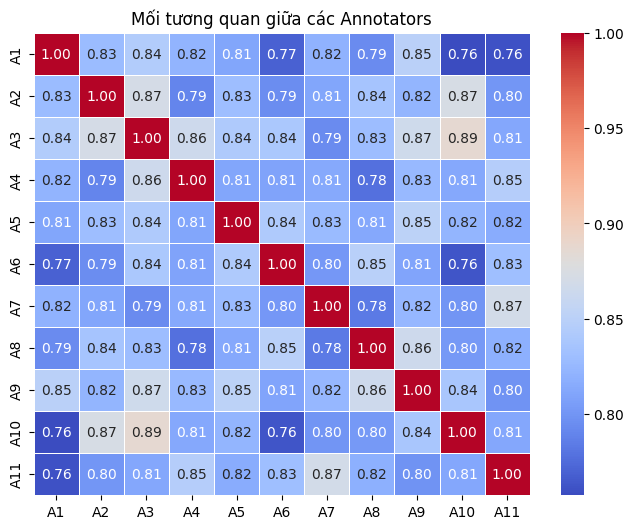
Việc phân tích phân phối token không chỉ giúp hiểu rõ hơn về tập dữ liệu mà còn hỗ trợ trong việc tối ưu hóa quá trình huấn luyện mô hình. Chẳng hạn, thông tin này có thể được sử dụng để thiết lập giới hạn độ dài câu, giúp cải thiện tốc độ huấn luyện và tránh hiện tượng mất thông tin do cắt ngắn câu quá mức. Ngoài ra, dữ liệu này cũng có thể hỗ trợ trong việc phát triển các kỹ thuật tiền xử lý, như điều chỉnh độ dài câu để cân bằng dữ liệu giữa hai ngôn ngữ, qua đó nâng cao chất lượng dịch tự động.

1. Sự chênh lệch số từ giữa cặp câu Việt - Lào theo từng mức độ tương đồng



1. Phân tích sai lệch giữa annotators

a.Phân tính mức độ tương quan giữa các annotators (spearman)



b.Tính toán sự nhất quán giữa annotators (Fleiss' Kappa)

