# V. Xây dựng bộ dữ liệu Vietnamese-Laos

# 5.1. Tiền xử lý dữ liệu

Một trong những thách thức lớn nhất mà người ta thường gặp phải trong các bộ sưu tập văn bản quy mô lớn như các nguồn tài liệu học thuật, và các nền tảng thông tin trực tuyến, là sự tồn tại của những vấn đề phát sinh do thiếu các bước xử lý ban đầu cho các câu văn. Những văn bản và câu chữ đa dạng trong các bộ dữ liệu này cần phải được phân tích kỹ lưỡng, và để có thể ứng dụng hiệu quả, những yếu điểm tồn tại cần phải được khắc phục. Tiếp theo chúng tôi sẽ trình bày rõ ràng về các bước xử lý đã được thực hiện để tạo ra một bộ dữ liệu.

1. Kiểm tra cấu trúc câu: Để đảm bảo tính đồng nghĩa, các cặp dữ liệu phải là một câu duy nhất. Bộ dữ liệu gốc có nhiều cặp bao gồm đoạn văn với nhiều câu, do đó chúng tôi chỉ giữ lại các cặp câu đơn.

2. Trong tập dữ liệu, một số câu, một số trong đó (ví dụ, một hoặc nhiều từ) là bằng các ngôn ngữ khác, đã bị loại bỏ để các mô hình muốn được sản xuất trong tương lai sử dụng tập này có chất lượng mong muốn.

3. Các câu được các chuyên gia kiểm tra sửa lỗi chính tả và được chỉnh sửa nếu có lỗi chính tả.

4. Xác minh đồng nghĩa: Chúng tôi đã được các chuyên gia ngôn ngữ hỗ trợ để kiểm tra tính đồng nghĩa của các câu, nhằm đảm bảo mỗi cặp câu thực sự tương đương về ngữ nghĩa.

5. Các ký tự của các từ trong các câu đã được kiểm tra bằng công cụ điều chỉnh định dạng.

6. Sự khác biệt về chiều dài câu có thể chênh lệch lên tới 5 từ.

# 5.2. Quá trình chọn lọc và tạo bộ dữ liệu

Sau khi xử lý trước các câu và áp dụng các quy tắc liên quan, các cặp câu đã được chọn bởi sự cân nhắc của chuyên gia từ tập dữ liệu sao cho phân phối điểm tương đồng trong tập dữ liệu được tạo ra là mong muốn và từ tất cả các điểm số (từ 0 đến 5) có một cặp câu phù hợp. Nếu chúng ta chọn ngẫu nhiên các cặp câu từ một tập hợp các câu đã được xử lý trước, hơn 99% trong số đó có điểm số là 0 hoặc 1 cuối cùng, và thực tế là hầu hết các cặp câu này có sự tương đồng về nghĩa rất ít và phân phối mức độ tương đồng của chúng không đồng đều.

Quy trình chuẩn bị bộ dữ liệu Việt-Lào.

Trong nghiên cứu này, chúng tôi đã sử dụng một bộ dữ liệu sẵn có với khoảng 7.000 cặp câu tiếng Việt và tiếng Lào, trong đó mỗi cặp câu là bản dịch của nhau. Mục tiêu của chúng tôi là chọn lọc 700 cặp câu đồng nghĩa có chất lượng tốt nhất cho việc tạo ra các cặp mang nhiều mức độ tương đồng khác nhau.

Đầu tiên các chuyên gia chọn lọc được 700 cặp câu tương đồng có chất lượng tốt nhất. Tiếp theo, chúng tôi đã lựa chọn ra 400 câu tiếng Lào từ 700 cặp câu đồng nghĩa này để tạo ra cặp dữ liệu tương đồng thấp, liên quan cao, liên quan thấp, trái nghĩa và không liên quan.

Sau khi chuyên gia đã biến đổi sang các câu tiếng Việt, chúng tôi đã ghép nó với câu tiếng Lào gốc để tạo thành các cặp câu mới, bao gồm câu tiếng Lào gốc và câu tiếng Việt.

|  |  |  |
| --- | --- | --- |
| Laos | Vietnamese | 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 . | 0 |
| 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. | 1 |
| 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 |
| ເປັນຄັ້ງ ທຳ ອິດໃນປະຫວັດສາດຂອງດາວເຄາະນີ້, ພວກເຮົາສາມາດອອກແບບສິ່ງມີຊີວິດໂດຍກົງ. | Lần đầu tiên trong lịch sử thế giới này, chúng ta có khả năng thiết kế động vật trực tiếp. | 3 |
| 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. | 4 |
| 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. | 5 |

Bảng 1 : Mẫu câu tiếng Viet-Lao có sự tương đồng về mặt ngữ nghĩa

Trong quá trình này, việc giữ nguyên các câu tiếng Lào là rất quan trọng, trong khi chỉ biến đổi câu tiếng Việt. Chúng tôi đã giữ lại các câu tiếng Lào chất lượng đã được chọn lọc từ 700 cặp câu đồng nghĩa đó để xây dựng các độ tương đồng khác nhau với phần câu tiếng Việt được biến đổi.

Để đạt được một câu tiếng Việt chất lượng do chuyên gia soạn thảo, các chuyên gia đã phải chọc lọc phân tích rất kĩ lượng và mất nhiều chi phí để cho ra được cặp câu chất lượng.

Với quy trình này, chúng tôi kỳ vọng sẽ tạo ra một bộ dữ liệu đồng nghĩa phong phú và đa dạng, đáp ứng nhu cầu nghiên cứu trong lĩnh vực ngôn ngữ và ngữ nghĩa học.

# 5.3. quá trình chú thích

Trong quá trình sản xuất tập dữ liệu , các chuyên gia khác nhau được sử dụng để chú thích. Việc để các chuyên gia chú thích giúp nâng cao chất lượng của tập dữ liệu sản xuất và cũng tăng tốc độ sản xuất của nó. Trong việc tạo ra tập dữ liệu tương đồng ngữ nghĩa.

Quá trình chú thích cho các cặp câu bao gồm hai phần. Trong phần đầu tiên, mỗi người chú thích có thể xem các cặp câu liên quan gán điểm từ 0 đến 5 cho mỗi cặp câu.

Phân thứ hai, bởi vì, bên cạnh các đánh giá và xử lý trước được thực hiện, các cặp câu vẫn có thể bị sai chính tả hoặc có các lỗi ngữ nghĩa khác, nên sẽ được chỉnh sửa và đánh giá tiếp. Để chú thích cho mỗi cặp câu bởi các chuyên gia, hướng dẫn dưới đây được mô tả trong bảng sẽ được sử dụng.

|  |  |  |
| --- | --- | --- |
| score | Định nghĩa | Nhãn |
| 5 | Hai câu hoàn toàn tương đương, vì chúng có nghĩa giống nhau. | Tương đồng |
| 4 | Hai câu hầu như tương đương, nhưng có một vài chi tiết không quan trọng khác nhau. | Tương đồng thấp |
| 3 | Hai câu tương đối tương đương, nhưng một số thông tin quan trọng khác nhau hoặc bị thiếu. | Liên quan cao |
| 2 | Hai câu không tương đương, nhưng chia sẻ một số chi tiết. | Liên quan thấp |
| 1 | Hai câu trái nghĩa, nhưng đều nói về cùng một chủ đề. | Trái nghĩa |
| 0 | Hai câu hoàn toàn khác biệt. | Không liên quan |

# 5.4. Tạo bộ dữ liệu

Để xây dựng một tập dữ liệu phản ánh sự phân bố đồng nhất của các khoảng điểm tương đồng, một tập dữ liệu nhỏ hơn bao gồm 2000 cặp câu đã được lấy ra và mẫu từ hơn 7000 cặp câu Viet-Lao đã được chọn lọc từ nhiều bài báo. Phần câu tiếng Việt của bộ sưu tập này đã được dịch sang tiếng Lào bởi các nhà ngôn ngữ học thông thạo và đã thay thế câu tiếng Việt. Do đó, một tập dữ liệu đã được hình thành, trong đó một bên của các cặp câu là Lào và bên kia là các câu tiếng Việt và trong cột thứ ba, mức độ tương đồng giữa hai câu tiếng Việt tiếng Lào được chỉ ra bên trên.

5.5. Thống kê bộ dữ liệu