**5. Xây dựng tập kiểm tra Vietnamese-Lao (ViLoSTS)**

Bộ dữ liệu đánh giá đóng vai trò then chốt trong việc đo lường hiệu quả và khả năng tổng quát hóa của các mô hình xử lý ngôn ngữ tự nhiên (NLP), đặc biệt trong bài toán đo cross-lingual Semantic Textual Similarity. Tuy nhiên, đối với các cặp ngôn ngữ ít tài nguyên như tiếng Việt và tiếng Lào, hiện chưa tồn tại một bộ dữ liệu đánh giá chuẩn hóa nào được công bố để đánh giá hiệu quả của các mô hình STS song ngữ. Điều này này gây cản trở đáng kể cho quá trình nghiên cứu, so sánh các phương pháp STS song ngữ Việt – Lào.

Nhằm giải quyết vấn đề trên, chúng tôi giới thiệu một bộ dữ liệu được thiết kế đặc biệt cho mục đích đánh giá các mô hình STS trên cặp ngôn ngữ Vietnamese–Lao, bộ dữ liệu có tên Vietnamese–Lao Semantic Textual Similarity Dataset (ViLoSTS). ViLoSTS được xây dựng với mục tiêu đánh giá các mô hình học biểu diễn ngữ nghĩa giữa hai ngôn ngữ có sự khác biệt đáng kể về hệ thống ngữ pháp và chữ viết, từ đó đóng góp vào kho tài nguyên ngôn ngữ cho các ngôn ngữ ít được nghiên cứu. ViLoSTS được kỳ vọng sẽ trở thành nguồn dữ liệu chuẩn mực, hỗ trợ hiệu quả cho việc đánh giá và so sánh các mô hình STS trong bối cảnh xử lý ngôn ngữ tự nhiên song ngữ Vietnamese–Lao – một hướng nghiên cứu còn nhiều tiềm năng nhưng chưa được khai thác đúng mức.

Quy trình xây dựng ViLoSTS được triển khai một cách có hệ thống và kiểm soát chất lượng nghiêm ngặt, bao gồm bốn bước chính:

1. Lựa chọn các cặp câu song ngữ Việt – Lào có chất lượng tốt, theo các domain

2. Biến đổi các cặp câu nhằm tạo ra nhiều mức độ tương đồng ngữ nghĩa khác nhau;

3. Gán nhãn similarity thủ công cho từng cặp câu theo thang điểm định lượng

4. Đánh giá độ nhất quán giữa các annotator;

5. Tổng hợp và phân tích thống kê bộ dữ liệu hoàn chỉnh.

Kết quả là một tập dữ liệu kiểm tra có phân bố similarity hợp lý, đa dạng về ngữ cảnh, và đạt độ tin cậy cao trong đánh giá.

**5.1 Lựa chọn và tạo cặp câu có độ tương đồng đa dạng**

Dựa trên nguồn ngữ liệu các cặp câu song ngữ đã được xây dựng, chúng tôi chọn lựa các cặp câu với các domain khác nhau như Tin tức, Sức khỏe, Giáo dục, … để tạo ra một bộ dữ liệu kiểm tra có khả năng phản ánh đa dạng các mức độ tương đồng ngữ nghĩa. Mục tiêu chính là xây dựng một bộ dữ liệu đánh giá chất lượng cao, có khả năng phân biệt hiệu quả các mức độ tương đồng từ hoàn toàn không liên quan đến tương đương ngữ nghĩa tuyệt đối, qua đó phục vụ cho việc kiểm định các mô hình STS trong bối cảnh song ngữ Vietnamese–Lao.

Trước tiên, chúng tôi lựa chọn 700 cặp câu theo các domain khác nhau, các cặp câu đảm bảo sự tương ứng rõ ràng về mặt nội dung và cấu trúc giữa tiếng Việt và tiếng Lào. Các cặp này đóng vai trò như nền tảng ban đầu để tạo thành các biến thể ngữ nghĩa khác nhau. Trong quá trình tạo biến thể, các câu tiếng Việt được chỉnh sửa theo nhiều hướng, trong khi câu tiếng Lào được giữ nguyên để đảm bảo tính chính xác.

Chúng tôi áp dụng các chiến lược biến đổi có hệ thống nhằm tạo ra các cấp độ tương đồng rõ rệt, bao gồm:

|  |  |  |
| --- | --- | --- |
| Chiến lược | Mô tả | Ví dụ |
| Biến đổi cú pháp | Thay đổi cấu trúc câu như chủ động ↔ bị động, trực tiếp ↔ gián tiếp mà không làm thay đổi ý nghĩa cốt lõi. | *Chính phủ ban hành chính sách mới → Chính sách mới được chính phủ ban hành* |
| Điều chỉnh sắc thái biểu đạt | Thay đổi cường độ, mức độ biểu cảm hoặc tính đánh giá của từ ngữ để tạo ra sự khác biệt tinh tế trong ngữ nghĩa. | *Hôm nay trời rất nóng* *→* *Hôm nay trời hơi nóng* |
| Bổ sung yếu tố gây nhiễu hoặc lược bỏ thông tin | Thêm hoặc bớt các chi tiết phụ nhằm tạo ra sự thay đổi trong mức độ liên quan hoặc phạm vi thông tin giữa hai câu. | *Sinh viên cần chuẩn bị bài và mang theo tài liệu tham khảo → Sinh viên cần chuẩn bị bài* |
| Tạo cặp không liên quan | |  | | --- | |  |  |  | | --- | | Ghép hai câu từ các lĩnh vực, chủ đề hoàn toàn khác nhau nhằm tạo ra mức độ tương đồng ngữ nghĩa bằng 0. | | *Trường tổ chức hội thảo khoa học → Giá cà phê ở Lào tăng mạnh trong tháng qua* |

Bảng 5.1: các chiến lược biến đổi ngữ nghĩa trong quá trình xây dựng tập kiểm tra ViLoSTS

## 5.2. Gán nhãn tương đồng ngữ nghĩa và đánh giá độ tin cậy

Sau khi hoàn tất quá trình lựa chọn các cặp câu đại diện, bước tiếp theo trong quy trình là gán nhãn mức độ tương đồng ngữ nghĩa cho từng cặp song ngữ và đánh giá độ nhất quán giữa các annotator nhằm đảm bảo độ tin cậy của bộ dữ liệu.

Mỗi cặp câu trong tập dữ liệu được gán một nhãn điểm phản ánh mức độ tương đồng ngữ nghĩa, theo thang điểm từ 0 đến 4. Các mức điểm được định nghĩa như sau:

|  |  |  |
| --- | --- | --- |
| Điểm số | Mức độ tương đồng | Diễn giải |
| 4 | Tương đương | Hai câu là bản dịch sát nghĩa |
| 3 | Gần tương đương | Câu có cùng nội dung chính, khác tiểu tiết |
| 2.5 | Liên quan cao | Cùng chủ đề, chia sẻ nhiều yếu tố ngữ nghĩa |
| 2 | Liên quan | Có điểm tương đồng nhưng khác thông tin chính |
| 1 | Liên quan thấp | Chia sẻ vài vài chi tiết lẻ, không trùng ý chính |
| 0 | Không liên quan | Hai câu hoàn toàn không có liên kết ngữ nghĩa |

Bảng 5.2 : bảng gán nhãn mức độ tương đồng trong tập test

Để đảm bảo tính nhất quán, nhóm nghiên cứu đã xây dựng bộ hướng dẫn gán nhãn chi tiết, bao gồm định nghĩa, ví dụ minh họa và quy tắc xử lý các tình huống mơ hồ. Tất cả annotator được huấn luyện theo bộ hướng dẫn này trước khi tham gia gán nhãn chính thức.

Toàn bộ 2.000 cặp câu sau khi được tạo và biến đổi từ giai đoạn 5.1 đều được gán nhãn thủ công bởi ba annotator có chuyên môn ngôn ngữ. Mỗi cặp câu được đánh giá độc lập, không tiết lộ điểm số của các annotator khác để đảm bảo tính khách quan.

Sau khi hoàn tất quá trình gán nhãn, bộ dữ liệu được kiểm tra độ đồng thuận giữa các annotator bằng các chỉ số thống kê chuẩn trong lĩnh vực NLP:

* Spearman’s ρ (rho): đo tương quan thứ hạng giữa các annotator – phù hợp với dữ liệu dạng thứ tự.
* Fleiss’ Kappa (κ): đo mức độ đồng thuận khi có nhiều hơn hai người đánh giá.
* Krippendorff’s Alpha :chỉ số linh hoạt, áp dụng cho nhiều loại dữ liệu (nominal, ordinal, interval), đặc biệt phù hợp với dữ liệu thứ tự và không yêu cầu tất cả annotator phải gán nhãn cho mọi mẫu. Alpha cho phép xử lý trường hợp thiếu nhãn và phản ánh độ đồng thuận có trọng số dựa trên mức sai khác giữa các annotator.

Trong trường hợp phát hiện sự không nhất quán đáng kể (ví dụ: chênh lệch từ 2 điểm trở lên), các cặp câu được đưa vào vòng thảo luận lại và gán nhãn đồng thuận. Mục tiêu là đảm bảo mọi điểm số cuối cùng phản ánh đúng mức độ tương đồng về mặt ngữ nghĩa và có thể được dùng để huấn luyện các mô hình học máy một cách ổn định.

Thông qua quy trình này, tập dữ liệu đầu ra không chỉ có nhãn chất lượng cao mà còn đạt độ tin cậy đáng kể, giúp đảm bảo tính tái lập và giá trị học thuật cho các nghiên cứu kế tiếp.

## 5.3. Tổng hợp dữ liệu và phân tích thống kê

Sau quá trình gán nhãn và đánh giá độ tin cậy giữa các annotator, tập dữ liệu Vietnamese–Lao được tổng hợp thành phiên bản hoàn chỉnh gồm 2.000 cặp câu song ngữ. Dữ liệu được tổ chức dưới dạng bảng, với bốn trường chính: câu tiếng Việt, câu tiếng Lào, điểm tương đồng (score) và nhãn mức độ tương đồng (label).

Bảng dưới đây trình bày các thông số chính của tập dữ liệu cuối:

|  |  |
| --- | --- |
| Thông tin | Giá trị |
| Tổng số cặp câu ban đầu | 7000 |
| Số cặp câu được chọn để xây dựng | 700 |
| Số cặp câu sau khi xây dựng | 2000 |
| Trung bình độ dài câu (tiếng Việt) | 33.15 |
| Trung bình độ dài câu (tiếng Lào) | 27.57 |
| Số annotator tham gia gán nhãn | 3 chuyên gia ngôn ngữ |

Các chỉ số thống kê đánh giá độ nhất quán giữa các annotator được trình bày như sau:

|  |  |  |
| --- | --- | --- |
| Chỉ số | Giá trị | Diễn giải |
| Spearman’s trung bình | 0.848 | Tương quan thứ hạng rất cao |
| Fleiss’ Kappa | 0.406 | Đồng thuận trung bình |
| Krippendorff’s | 0.846 | Đồng thuận gần như tuyệt đối (almost perfect) |

Để đánh giá mức độ nhất quán trong quá trình gán nhãn, nghiên cứu sử dụng ba chỉ số phổ biến: Spearman’s rho, Fleiss’ Kappa, và Krippendorff’s Alpha. Mỗi chỉ số phản ánh một khía cạnh khác nhau về mức độ đồng thuận giữa các annotator.

Spearman’s rho được tính giữa các cặp annotator nhằm đo lường mức tương quan thứ hạng giữa các điểm gán nhãn. Trung bình Spearman của ba cặp annotator đạt 0.848, phản ánh mức tương quan thứ hạng rất cao. Điều này cho thấy các annotator có xu hướng xếp hạng mức độ tương đồng ngữ nghĩa khá thống nhất, dù có thể chênh lệch nhẹ về điểm số cụ thể.

Fleiss’ Kappa, được tính trên toàn bộ tập nhãn rời rạc (0–4, bao gồm cả 2.5), đạt giá trị 0.406. Theo thang phân loại của Landis & Koch (1977), đây là mức độ đồng thuận trung bình (moderate agreement). Kết quả này là hợp lý vì Fleiss’ Kappa không xét đến khoảng cách thứ tự giữa các nhãn, do đó có thể đánh giá thấp các chênh lệch nhỏ như giữa điểm 2 và 2.5.

Krippendorff’s Alpha (ordinal) đạt giá trị 0.846, cho thấy mức đồng thuận gần như hoàn hảo giữa các annotator. Đây là chỉ số được đánh giá là phù hợp nhất trong bối cảnh gán nhãn theo thang điểm thứ tự, đặc biệt khi dữ liệu có thể có nhãn thiếu hoặc phân bố không đều.

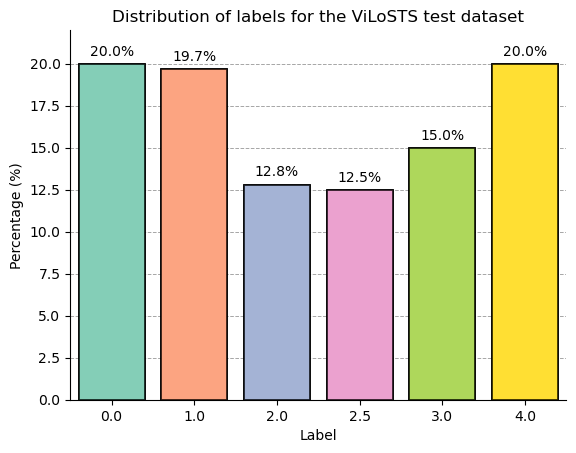
Các giá trị trên cho thấy bộ dữ liệu có độ tin cậy cao và sự đồng thuận tốt giữa các annotator, đặc biệt trong bối cảnh đánh giá ngữ nghĩa mang tính chủ quan.

### 5.4 Phân bố mức độ tương đồng

Biểu đồ phân bố tỷ lệ các mức độ tương đồng ngữ nghĩa trong tập dữ liệu kiểm tra Vietnamese–Lao (ViLoSTS). Sau quá trình điều chỉnh, tổng số 2.000 cặp câu được gán nhãn đã đạt được phân bố tương đối cân bằng giữa các mức điểm, đảm bảo tính đại diện và đa dạng ngữ nghĩa.

Cụ thể, các mức điểm 0 (không liên quan) và 4 (tương đương ngữ nghĩa tuyệt đối) đều chiếm tỷ lệ cao, lần lượt là 20.0%, thể hiện rõ hai cực của thang đo STS. Các mức điểm 1 (liên quan thấp) và 3 (gần tương đương) cũng chiếm tỷ lệ đáng kể (19.7% và 15.0%), góp phần tạo nên sự đa dạng về mức độ tương đồng. Trong khi đó, các mức trung gian 2 (liên quan vừa phải) và 2.5 (liên quan cao) chiếm khoảng 12.8% và 12.5%, phản ánh thách thức của annotator trong việc phân loại các cặp câu có độ tương đồng mơ hồ hoặc không rõ ràng.

Biểu đồ cho thấy bộ dữ liệu kiểm tra đã đạt được sự phân bố nhãn gần như đồng đều giữa các mức điểm, đồng thời vẫn duy trì được các đặc điểm ngữ nghĩa đa dạng và thực tế của ngữ liệu song ngữ. Cấu trúc nhãn như vậy giúp đảm bảo tính khách quan trong đánh giá và so sánh mô hình, đặc biệt trong các nghiên cứu liên quan đến biểu diễn ngữ nghĩa cho cặp ngôn ngữ ít tài nguyên như Vietnamese–Lao.



Hình 5.4: Biểu đồ phân bố mức độ tương đồng ngữ nghĩa trong tập test ViLoSTS

## 5.5 Kết luận phần xây dựng tập test

Việc xây dựng tập dữ liệu kiểm tra ViLoSTS đóng vai trò then chốt trong việc đảm bảo khả năng đánh giá khách quan và toàn diện các mô hình đo lường mức độ tương đồng ngữ nghĩa cho cặp ngôn ngữ Vietnamese–Lao. Từ nguồn ngữ liệu báo chí song ngữ được chuẩn hóa, chúng tôi đã áp dụng các chiến lược biến đổi ngữ nghĩa có hệ thống để tạo ra các cặp câu có mức độ tương đồng đa dạng, đồng thời đảm bảo chất lượng gán nhãn thông qua quy trình đánh giá độ tin cậy liên chủ thể nghiêm ngặt.

Tập dữ liệu kiểm tra thu được có sự phân bố nhãn hợp lý, phản ánh được đầy đủ các mức độ tương đồng từ không liên quan đến tương đương tuyệt đối, đồng thời thể hiện được sự đa dạng về ngữ pháp, ngữ nghĩa và ngữ cảnh. ViLoSTS không chỉ đóng vai trò như một bộ dữ liệu tham chiếu tiêu chuẩn để đánh giá các mô hình STS song ngữ, mà còn mở ra cơ hội nghiên cứu sâu hơn về biểu diễn ngữ nghĩa trong điều kiện tài nguyên hạn chế.

Chúng tôi kỳ vọng rằng tập kiểm tra ViLoSTS sẽ đóng góp giá trị thiết thực cho cộng đồng nghiên cứu NLP, đặc biệt trong việc phát triển và chuẩn hóa các phương pháp học ngữ nghĩa cho các ngôn ngữ ít được chú trọng như tiếng Việt và tiếng Lào.