**A Benchmark for Vietnamese-Lao Semantic Textual Similarity: Dataset and Model Evaluation**

**ABSTRACT** — Semantic textual similarity (STS) plays a crucial role in natural language processing (NLP), enabling various applications such as machine translation, information retrieval, and question answering. Cross-lingual STS (CL-STS) aims to measure the semantic similarity between text pairs in different languages. However, research on CL-STS for low-resource languages remains limited, and no publicly available benchmark exists for evaluating Vietnamese-Lao semantic similarity. In this paper, we introduce the first benchmark dataset for Vietnamese-Lao STS, which provides high-quality sentence pairs annotated with similarity scores. Additionally, we propose a novel approach that combines lexical-based and distributed-based semantic similarity methods to improve CL-STS performance for Vietnamese and Lao. Experimental results demonstrate that our proposed model achieves strong performance on the Vietnamese-Lao STS task, highlighting the effectiveness of our approach and the importance of a dedicated benchmark for this low-resource language pair.

**Từ khóa**— Xử lý ngôn ngữ tự nhiên, độ tương tự ngữ nghĩa, mạng từ, từ nhúng.

**A Benchmark for Vietnamese-Lao Semantic Textual Similarity: Dataset and Model Evaluation**

**Tóm tắt**— Độ tương đồng ngữ nghĩa văn bản (STS) đóng vai trò quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), cho phép nhiều ứng dụng khác nhau như dịch máy, truy xuất thông tin và trả lời câu hỏi. STS liên ngôn ngữ (CL-STS) nhằm mục đích đo độ tương đồng ngữ nghĩa giữa các cặp văn bản ở các ngôn ngữ khác nhau. Tuy nhiên, nghiên cứu về CL-STS cho các ngôn ngữ có ít tài nguyên vẫn còn hạn chế và không có chuẩn mực nào có sẵn công khai để đánh giá độ tương đồng ngữ nghĩa tiếng Việt-Lào. Trong bài báo này, chúng tôi giới thiệu bộ dữ liệu chuẩn đầu tiên cho STS tiếng Việt-Lào, cung cấp các cặp câu chất lượng cao được chú thích bằng điểm tương đồng. Ngoài ra, chúng tôi đề xuất một phương pháp mới kết hợp các phương pháp độ tương đồng ngữ nghĩa dựa trên từ vựng và dựa trên phân tán để cải thiện hiệu suất CL-STS cho tiếng Việt và tiếng Lào. Kết quả thử nghiệm chứng minh rằng mô hình đề xuất của chúng tôi đạt được hiệu suất cao trong nhiệm vụ STS tiếng Việt-Lào, làm nổi bật hiệu quả của phương pháp tiếp cận của chúng tôi và tầm quan trọng của một chuẩn mực chuyên dụng cho cặp ngôn ngữ có ít tài nguyên này.