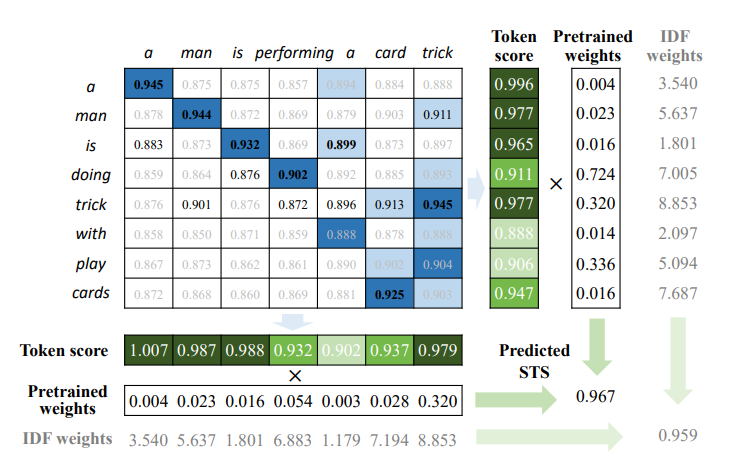
**Going Beyond Sentence Embeddings: A Token-Level Matching Algorithm for Calculating Semantic Textual Similarity**

* Mục tiêu: Bài nghiên cứu đề xuất một thuật toán so khớp cấp độ token để tính toán Semantic Textual Similarity (STS) - Độ tương đồng ngữ nghĩa văn bản, khắc phục hạn chế của phương pháp embedding câu truyền thống.
* Phương pháp:
  + Thay vì mã hóa câu thành embedding, thuật toán tính toán ma trận tương đồng giữa các token của hai câu đầu vào.
  + Sau đó, một hàm tính điểm so khớp token được áp dụng để đo lường khả năng hai token từ hai câu có thể so khớp với nhau, dựa trên điểm tương đồng và tính duy nhất. Trong đó, Token Score được tính toán dựa trên hai yế tố chính:

Significance cao cho thấy token đó có khả năng cao là bản dịch hoặc từ đồng nghĩa với một token trong câu kia.

Uniqueness cao cho thấy sự tương ứng giữa hai token đó rõ ràng, không bị nhầm lẫn với các token khác.

* + Cuối cùng, điểm STS được tính bằng cách lấy trung bình trọng số của tất cả các điểm so khớp token, với trọng số được học từ một tập dữ liệu văn bản lớn.
* Kết quả:
  + Thử nghiệm trên 7 bộ dữ liệu STS cho thấy phương pháp này cải thiện hiệu suất của hầu hết các mô hình ngôn ngữ hiện có, đặc biệt là những mô hình hoạt động kém.
  + Cải thiện lên đến 12.7% về hệ số tương quan Spearman.
  + Phương pháp cũng được chứng minh là có khả năng giải thích cao và hiệu quả về thời gian tính toán.



Điểm nổi bật của phương pháp này là họ đã có riêng một bộ trọng số khác biệt với bộ trọng số IDF, chính bộ trọng số này đã cải thiện đáng kể kết quả.

**Để so sánh và đánh giá, cần thiết phải chuyển đổi một trong hai thang điểm trên về cùng một khoảng giá trị.** Thông thường, có hai cách tiếp cận:

* Chuyển đổi điểm mô hình (0-1) sang thang điểm bộ dữ liệu (1-5): Có thể sử dụng công thức đơn giản như linear scaling (tỷ lệ tuyến tính) để ánh xạ điểm số từ khoảng 0-1 sang 1-5.

**Hệ số tương quan Spearman:**

Hệ số tương quan Spearman được sử dụng trong bài báo không yêu cầu hai biến phải có cùng khoảng giá trị. Nó chỉ quan tâm đến thứ tự xếp hạng của các cặp câu.

Ví dụ: Nếu mô hình dự đoán cặp câu A tương đồng hơn cặp câu B, và nhãn trong bộ dữ liệu cũng thể hiện điều tương tự (cho dù điểm số cụ thể là bao nhiêu), thì hệ số tương quan sẽ phản ánh sự tương đồng đó.

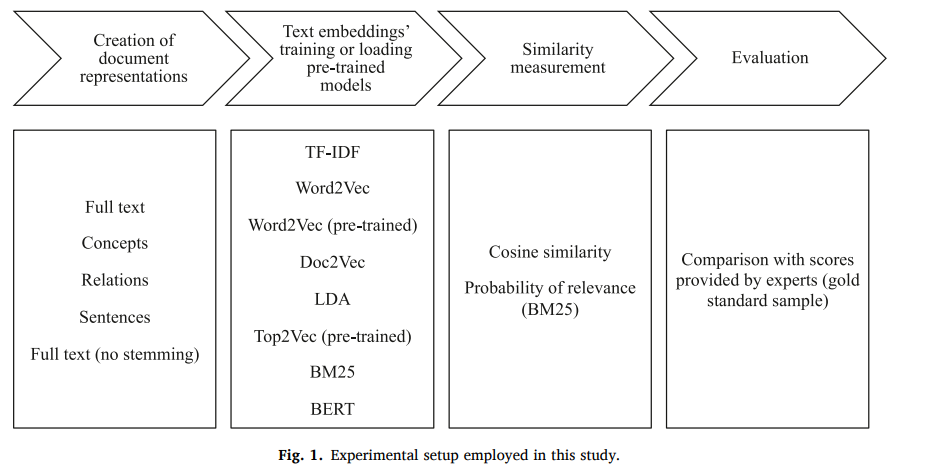
Các gợi mở rút ra:

Đối với tiếng việt thường có các từ ghép như đại\_học, vậy áp dụng ra sao với ma trận phía trên

Có thể áp dụng cách tạo trọng số từ các văn bản tương tự như bài nghiên cứu.

**Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court**

Bài viết đánh giá hiệu quả của các kỹ thuật Machine Learning (ML) / LLM trong việc xác định các tiền lệ pháp lý trong luật hành chính Brazil (Đưa ra các văn bản pháp lý có độ trùng lặp cao)

****

* Mô hình Word2Vec hiệu quả nhất: Mô hình Word2Vec, sử dụng biểu diễn văn bản chi tiết ở cấp độ từ, vượt trội hơn các mô hình truyền thống như TF-IDF và thậm chí cả các mô hình nâng cao như BERT trong việc đánh giá mức độ liên quan của tài liệu pháp lý.
* Nhu cầu tinh chỉnh theo lĩnh vực cụ thể: Mặc dù BERT là một công cụ mạnh mẽ trong Xử lý Ngôn ngữ Tự nhiên, nhưng nó không nhất thiết phải vượt trội hơn các mô hình truyền thống trong bối cảnh đánh giá mức độ liên quan của tài liệu pháp lý, ngay cả khi đã được huấn luyện trước trên ngữ liệu pháp lý. Điều này nhấn mạnh tầm quan trọng của việc tinh chỉnh theo lĩnh vực cụ thể để phù hợp với ngữ nghĩa độc đáo của văn bản pháp lý.
* Đào tạo lại thường xuyên: Bản chất động của luật pháp và cả các loại văn bản sử dụng ngôn ngữ nói chung đòi hỏi phải đào tạo lại thường xuyên các mô hình ML để đảm bảo chúng duy trì hiệu quả.

**Semantic similarity prediction is better than other semantic similarity measures**

Bài nghiên cứu lập luận rằng dự đoán trực tiếp độ tương đồng ngữ nghĩa bằng mô hình được tinh chỉnh tốt hơn so với các phương pháp đo lường độ tương đồng ngữ nghĩa khác, chẳng hạn như dựa trên nhúng (ví dụ: BERTScore, S-BERT) hoặc dựa trên n-gram (ví dụ: BLEU)

Chuẩn hóa tài liệu:

Gồm 600 câu với điểm số từ 0 tới 4

cần các cách xử lý thông thường nếu như cho vào học máy như lower case, loại bỏ stopword

Cần có các cách để làm giàu dữ liệu hơn

Nếu có thu thập các bộ dữ liệu khác thì cần chuẩn hóa lại thang điểm.