1. Giới thiệu

* Bài báo này mô tả một bộ dữ liệu mới bao gồm các câu với các chú thích về tương đồng ngữ nghĩa. Bộ dữ liệu này bao gồm các cặp câu trong ngôn ngữ tiếng Việt, trong đó mỗi cặp được gán một điểm tương đồng. Mục đích của bộ dữ liệu này là để huấn luyện và đánh giá các hệ thống dự đoán tương đồng ngữ nghĩa của câu.
* Bài báo tập trung vào phần kiểm thử của bộ dữ liệu. Đối với phần này, mỗi cặp câu được chú thích độ tương đồng độc lập bởi 9 nhà chú thích. Do đó, chúng ta lưu trữ giá trị trung bình của các chú thích độc lập làm điểm cuối cùng.
* Chúng tôi chú thích các điểm tương đồng của các cặp huấn luyện có và không có ngữ cảnh.
* Chúng tôi quyết định hợp tác với sinh viên báo chí để tạo ra các chú thích tốt hơn, vì họ thường có kỹ năng tốt trong việc xử lý dữ liệu văn bản. Để cải thiện chất lượng chú thích hơn nữa, các nhà báo chuyên nghiệp đã giám sát sinh viên thực hiện chú thích. Chúng tôi tin rằng sự đóng góp của các chú thích có kỹ năng cao sẽ nâng cao chất lượng của bộ dữ liệu.
* Bộ dữ liệu cuối cùng bao gồm 138,556 chú thích của con người. Việc tạo ra nó đòi hỏi một lượng lớn công việc chú thích của con người - 485 người chú thích và thời gian đã dành để tạo ra bộ dữ liệu là khoảng 1,017 giờ lao động của con người.

2.Công việc liên quan

* Những bộ dữ liệu này bao gồm các cặp câu lấy từ bài báo, diễn đàn, tiêu đề tin tức, mô tả hình ảnh và video, được đánh nhãn với điểm tương đồng từ 0 đến 6. Mục tiêu là đánh giá cách khoảng cách cosine giữa hai câu tương quan với một điểm tương đồng được đánh nhãn bởi con người thông qua các tương quan Pearson và Spearman.

3.Nguồn dữ liệu

* Dữ liệu văn bản gốc cho bộ dữ liệu được giới thiệu được lấy từ báo Dân Trí. Dân Trí cung cấp dịch vụ đầy đủ thông tin bao gồm cả hình ảnh, tin tức ngắn gọn nhanh chóng, bản báo cáo, quan sát các sự kiện kéo dài trong nhiều lĩnh vực.
* Các sự kiện thông thường được kết luận bằng một bản tóm tắt được tạo ra bởi con người. Chúng tôi sử dụng các tóm tắt và báo cáo được nhóm lại theo các sự kiện trong bộ dữ liệu của chúng tôi.

4. Quá trình thu thập chú thích từ con người

* Để có được dữ liệu phù hợp cho việc huấn luyện mô hình, chúng tôi cần kết hợp các câu từ các báo cáo và các bản tóm tắt.
* Chúng tôi đã xây dựng một ứng dụng web để thu thập chú thích từ con người về độ tương đồng giữa các cặp câu nhằm hỗ trợ quá trình chú thích. Những người chú thích được yêu cầu cung cấp hai thông tin cơ bản:

1)Tương đồng ngữ nghĩa văn bản không phụ thuộc ngữ cảnh,

2)Tương đồng ngữ nghĩa văn bản phụ thuộc ngữ cảnh.

* Nhân Sự: Trong công việc này, chúng tôi hợp tác với hai nhóm sinh viên báo chí. Nhóm 1 (272 người) đã chú thích dữ liệu trong vòng đầu tiên. Họ chọn các cặp câu và chú thích độ tương đồng - những dữ liệu này được sử dụng cho phần huấn luyện của bộ dữ liệu. Nhóm 2 (229 người) tham gia vào vòng thứ hai để tạo ra phần kiểm thử của bộ dữ liệu.

4.1 Vòng 1 (R1)

* Trong vòng chấm điểm đầu tiên (R1), chúng tôi yêu cầu các người chú thích chọn ba câu (An; Bn; Cn) từ các báo cáo cho mỗi câu (Sn) trong bản tóm tắt. Chúng tôi hướng dẫn họ chọn câu tương tự nhất (A), câu khác biệt nhất (C), và một cái ở giữa (B), nhằm tạo ra một bộ dữ liệu cân bằng hơn .Chúng tôi yêu cầu người dùng sử dụng thanh trượt để chú thích mức độ tương đồng trên một thang điểm từ 0-6

0.Hoàn toàn khác biệt

1.Một chút liên quan đến chủ đề

2.Được liên quan đến chủ đề

3.Giống nhau 1 phần

4.Tương tự nhau

5.Gần như giống nhau

6.Giống nhau hoàn toàn

* Related (liên quan) : 2 câu nói về cùng 1 chủ đề: sức khỏe, bác sĩ,...
* Similar (giống nhau) : 2 câu giống nhau về nghĩa và nói về cùng 1 chủ đề

4.2 Vòng 2 (R2)

* Chúng tôi tạo ra phần kiểm thử của bộ dữ liệu trong vòng chấm điểm thứ hai (R2). Chúng tôi yêu cầu người chú thích gán số tương đồng cho các cặp câu (SA, SB, SC) được tạo ra trong R1 trên cùng thang điểm (0-6). Chi tiết như sau
* Khoảng hai tháng trước khi bắt đầu chấm điểm chính, chúng tôi thực hiện một giai đoạn chú thích sơ bộ ngắn ngắn để thu thập một số dữ liệu cho việc nghiên cứu sự đồng thuận giữa các người chú thích trong tương lai.
* Chúng tôi đã thêm một quy tắc nghiêm ngặt vào giao diện người dùng, yêu cầu một đợt nghỉ 72 giờ giữa các chú thích không phụ thuộc ngữ cảnh và chú thích phụ thuộc ngữ cảnh.
* Khối Điều Chỉnh (R2-CAL): Trong giai đoạn chú thích sơ bộ ngắn, để tăng độ đồng thuận giữa các người chú thích thì chúng tôi lấy ra sáu cặp trong tổng số các cặp cần chú thích và đưa cho những người chú thích để họ chú thích các cặp được 1 cách thủ công bao phủ từ 0->6. Qua kết quả chú thích ta thấy có được độ đồng thuận và độ tin cậy của người chú thích.
* Khối Chính R2 (R2-MAIN): Trong R2 Main người chú thích thực hiện chú thích mỗi cặp câu 9 lần để tăng độ tin cậy và loại bỏ độ chênh lệch STS tiềm ẩn có thể phát sinh, chúng tôi chọn kích thước cho khối chính là 50 cặp để phù hợp với thời gian của mỗi người. Do đó mỗi người chú thích chính xác 50 cặp trước tiên là không có ngữ cảnh sau đó là có ngữ cảnh.
* R2-INTRA: được thực hiện sau khi hoàn thành khối chính (R2-MAIN) và yêu cầu mỗi người chú thích chú thích thêm năm cặp câu mà họ đã chú thích trong giai đoạn chấm điểm dự bị (R2-CAL). Thông qua việc so sánh các chú thích trên những cặp này, bạn có thể đo lường sự đồng thuận giữa các chú thích và đánh giá mức độ tin cậy của họ. Điều này giúp kiểm tra tính nhất quán và độ tin cậy của các chú thích trong quá trình đánh giá tương đồng ngữ nghĩa (STS) của các cặp câu.

5.Thống kê tập dữ liệu

* Bộ dữ liệu cuối cùng được chia thành phần huấn luyện và phần kiểm thử. Phần huấn luyện chứa 116,956 mẫu. Trong bộ dữ liệu kiểm thử, chúng tôi quyết định chú thích 1200 cặp cho cả hai biến thể không phụ thuộc và phụ thuộc ngữ cảnh. Chúng tôi đã thiết kế quy trình để có đến chín chú thích cho mỗi cặp. Tuy nhiên, chỉ có 21 trong số 1200 cặp có ít hơn bảy chú thích.
* Chúng tôi ước lượng thời gian đã dành để tạo ra bộ dữ liệu huấn luyện (R1) là 876 giờ công của con người (269 người chú thích, trung bình mỗi người khoảng 3.26 giờ). Những người chú thích đã tạo ra tập kiểm thử trong 141 giờ công của con người (216 người chú thích, trung bình mỗi người khoảng 40 phút).
* Thống nhất giữa người chú thích: Để khám phá giới hạn của sự đồng thuận giữa con người, chúng tôi tính toán một số độ đo phù hợp cho mục đích này.
* Trong giai đoạn chuẩn bị ngắn (R2-CAL), chúng tôi lưu trữ mười chú thích mà người dùng đã thực hiện. Sau đó, chúng tôi cho phép họ thực hiện lại chú thích trên những cặp đó sau một thời gian. Mỗi người dùng đã chú thích mười cặp câu (năm cho cả không phụ thuộc và phụ thuộc ngữ cảnh). Bằng cách này, chúng tôi có thể đánh giá sự đồng thuận giữa các chú thích của cùng một người chú thích trên cùng một cặp câu sau khoảng thời gian
* Điểm số được tạo ra bởi tất cả người dùng trong cả hai giai đoạn ngữ cảnh được nối lại thành hai vector (duy trì thứ tự của chúng). Chúng tôi tính toán độ tương quan Pearson và Spearman giữa hai vector này để đo lường mức độ tương quan giữa điểm số trước đó và điểm số của cùng một cặp và cùng một người chú thích được chú thích sau này
* Chúng tôi sử dụng các hệ số tương quan cùng với Độ lỗi bình phương trung bình (MSE) và Độ lỗi bình phương trung bình căn (RMSE) để chỉ ra mức độ khác biệt trung bình giữa các chú thích của từng người chú thích
* MSE: (*x*¯:giá trị trung bình điểm STS cho từng cặp câu, µ: trung bình STS thực sự, *σ:* độ lệch chuẩn của các điểm STS cho từng cặp, n: số lượng cặp chú thích) . Tính σ *= x*¯- µ, làm tương tự với các cặp còn lại ta tính được σ cho từng cặp. Bình phương kết quả của từng độ lệch chuẩn (σ*)*.Sau đó tính tổng của tất cả các kết quả bình phương này và sau đó chia cho n để tính được mức độ khác biệt trung bình giữa các chú thích

6. Thử nghiệm ban đầu

* Để thiết lập một điểm cơ bản cho bộ dữ liệu mới, chúng tôi sử dụng mô hình cơ bản và mạnh mẽ từ wordtovec (Mikolov et al., 2013), cùng với các mô hình hiện đại dựa trên kiến trúc BERT.Các mô hình dựa trên BERT có thể xử lý các câu theo hai cách: một bộ mã hóa Cross-attention trên cả hai câu cùng một lúc; và một bộ mã hóa Two Tower/Siamese cho từng câu một cách riêng lẻ với một đo lường sự tương đồng ở đỉnh.
* **WordToVec**: tạo ra những vectơ nhúng cho mỗi từ trong câu. Chúng ta sử dụng trung bình của các vectơ này để tạo ra biểu diễn cho toàn bộ câu.Để đánh giá sự tương đồng giữa hai câu, chúng tôi áp dụng độ tương đồng cosin lên các vectơ câu được tạo ra từ Word2Vec. Điều này tạo ra một điểm số tương đồng, đo lường mức độ giống nhau giữa các câu.
* **Two tower mode**: Đầu tiên, chúng tôi sử dụng mô hình giống Bert để tạo ra các biểu diễn đầu ra từ phần pooler cho mỗi câu độc lập.Các biểu diễn này được xếp chồng lên nhau, tạo thành một ngăn xếp. Mỗi tầng của ngăn xếp đại diện cho một cấp độ biểu diễn ngữ cảnh cao hơn.Sau đó, chúng tôi áp dụng đo lường tương đồng cosin giữa các biểu diễn của hai câu để đánh giá mức độ tương đồng giữa chúng.Cuối cùng, chúng tôi điều chỉnh mô hình trên bộ dữ liệu huấn luyện để tối ưu hóa đo lường tương đồng và đảm bảo mô hình hiệu quả trên tập dữ liệu mới.
* **Cross-Attention Model**: Pooler output là biểu diễn cuối cùng của câu, được tạo ra bằng cách áp dụng lớp Fully Connected lên token đại diện của câu (thường là token [CLS]).Chúng tôi sử dụng một lớp chiếu có kích thước 200 và kích hoạt RELU (Rectified Linear Unit) để chuyển đổi biểu diễn từ pooler thành một không gian đặc trưng có chiều sâu hơn.Sau đó, chúng tôi sử dụng một neuron duy nhất với kích hoạt tuyến tính để thu được đo lường tương đồng giữa hai câu. Điều này tạo ra một số liệu tương tự để đánh giá mức độ tương đồng giữa các câu.Cuối cùng, chúng tôi điều chỉnh mô hình trên bộ dữ liệu huấn luyện để cải thiện hiệu suất và đảm bảo rằng mô hình đáp ứng tốt trên các dữ liệu mới.

7 Lọc Dữ Liệu Tập Dữ Liệu

* Trong quá trình xử lý sau cùng, chúng tôi làm sạch phần huấn luyện của tập dữ liệu bằng cách so sánh thống kê với phần kiểm thử. Chúng tôi giả định rằng phần kiểm thử chứa các chú thích đáng tin cậy do giai đoạn khám phá và một vòng lặp thứ hai được thiết kế cẩn thận. Các điểm số cuối cùng được tính bằng cách lấy trung bình của chín số, làm cho phần kiểm thử trở nên mạnh mẽ hơn.
* **Lọc Dữ Liệu Liên Quan Đến Sự Thay Đổi Ngữ Nghĩa**: Chúng tôi đã thiết kế một quy trình để thu thập 2 loại dữ liệu là không có sự phụ thuộc đặc biệt vào ngữ cảnh và có sự phụ thuộc vào ngữ cảnh, để cho phép các nhà nghiên cứu trong tương lai có thể xem xét vai trò của ngữ cảnh trong ngữ nghĩa. Chúng tôi không thực hiện bất kỳ bước lọc đặc biệt nào trước khi trình bày các câu cho người chú thích, do đó sự thay đổi ngữ nghĩa theo ngữ cảnh sẽ xuất hiện trong dữ liệu thu thập được.
* Chúng tôi quyết định đo lường sự quan trọng của ngữ cảnh bằng cách thực hiện một loạt các kiểm định t-tests (Đây là kiểm định thống kê được sử dụng để so sánh giá trị trung bình giữa nhóm ko phụ thuộc vào ngữ cảnh và phụ thuộc vào ngữ cảnh và xác định xem sự khác biệt đó có đáng kể hay không). Mỗi kiểm định t-tests được thực hiện cho mỗi cặp câu riêng biệt, tức là từng phần tử trong dữ liệu. Trung bình của giá trị tương đồng ngữ nghĩa (STS) được đo lường cho mỗi cặp câu. Dữ liệu mẫu gồm chín giá trị STS được thu thập cho mỗi cặp câu trong mỗi điều kiện (không phụ thuộc vào ngữ cảnh và phụ thuộc vào ngữ cảnh). Chúng tôi giả định rằng dữ liệu mẫu của chúng tôi có phân phối xấp xỉ chuẩn, một điều kiện quan trọng để áp dụng kiểm định t-tests
* Tóm lại đối với mỗi cặp câu (của các khối chính), chúng tôi có chín điểm tương đồng ngữ nghĩa (STS) cho một phiên bản không phụ thuộc vào ngữ cảnh và chín điểm STS cho một phiên bản có ngữ cảnh của cặp câu đó. Sau đó, chúng tôi thực hiện một kiểm định t-tests hai phía để kiểm tra sự tương đồng về giá trị trung bình tương ứng. Giả thuyết là giá trị trung bình của điểm STS bằng nhau (µ1=µ2) . Ta sẽ tính giá trị t và *df* để có thể so sánh với bảng phân phối t hoặc sử dụng phần mềm thống kê để xác định giá trị p tương ứng. Nếu p > 0.05 thì khẳng định không có sự khác biệt giữa 2 nhóm (không phụ thuộc vào ngữ cảnh và phụ thuộc vào ngữ cảnh) nghĩa là ngữ cảnh không quan trọng đối với ngữ nghĩa của tôi và ngược lại
* Lọc đường chéo của Bộ dữ liệu Huấn luyện: Sau các chú thích, chúng tôi tính toán một số thống kê bổ sung về dữ liệu đã thu thập. Chúng tôi phát hiện có một số trường hợp ranh giới trong giai đoạn đầu tiên. Chúng tôi thực hiện phân tích tần suất của mỗi đánh giá (0–6) cho ba lớp được đề xuất (xanh, cam, đỏ)
* Có không nhiều lý do có thể dẫn đến sự chệch lệch dữ liệu theo chiều ngược lại. Một lần nữa, nếu chúng ta loại bỏ lỗi hệ thống của con người, điều này không có khả năng xảy ra, thì lý do duy nhất cho sự chệch hệ thống như vậy (có thể quan sát được cũng trên lớp được đề xuất màu cam) - những người chú thích không có lựa chọn khác. Nguồn duy nhất có lý cho sự chệch này là sự xuất hiện của thông tin hoàn toàn mới trong bản tóm tắt hoặc câu chưa được sử dụng trong báo cáo gốc.
* Việc thu thập dữ liệu trong vòng đầu tiên có thể mang lại sự chệch hệ thống vào dữ liệu đào tạo - con người đánh dấu các cặp khác nhau như những chiếc lều màu xanh để chọn số tương đồng cao hơn.

8. Định dạng dữ liệu

* Chúng tôi trình bày dữ liệu đã thu thập trong các tập tin văn bản có sẵn trên trang web của chúng tôi và trên github. Họ cung cấp thông tin về số lượng trung bình các đánh giá hoặc nhận xét cho mỗi cặp dữ liệu cụ thể. Họ cũng liệt kê tất cả các nhận xét chi tiết được thực hiện trong các tập tin này, có thể để cho người dùng có cái nhìn tổng quan về dữ liệu và các đánh giá liên quan. Điều này giúp trong việc hiểu và đánh giá chất lượng của dữ liệu cũng như kết quả của các thí nghiệm hoặc mô hình dự đoán dựa trên dữ liệu này.
* Chúng tôi cũng trình bày tập dữ liệu kiểm tra đã được lọc theo độ quan trọng của sự thay đổi giữa các nhận xét không phụ thuộc vào ngữ cảnh và những nhận xét phụ thuộc vào ngữ cảnh được gắn nhãn như testsig.tsv. Các mẫu dữ liệu được lọc dựa trên mức độ tin cậy 0.95 của sự khác biệt đáng kể giữa các nhận xét không phụ thuộc vào ngữ cảnh và các nhận xét phụ thuộc vào ngữ cảnh trong tập tin này.

9. Thảo luận

* Vai trò của Ngữ cảnh: Mục tiêu chính là tạo ra một tập dữ liệu tiếng Séc mới cho sự tương đồng văn bản ngữ nghĩa với các nhận xét không phụ thuộc vào ngữ cảnh và phụ thuộc vào ngữ cảnh được căn chỉnh và đánh giá tầm quan trọng của ngữ cảnh trong các văn bản thông thường.Trong lĩnh vực hẹp của tin tức và các tóm tắt của chúng, chúng tôi quan sát được 90 mẫu từ tổng số 1200 (7.5%) bị dịch chuyển đáng kể. Chúng tôi thu thập các mẫu này vào các tập dữ liệu riêng biệt để các nhà nghiên cứu trong tương lai có thể sử dụng các mô hình hiện đại thể hiện rõ lợi ích của ngữ cảnh
* Vai trò của Việc Lọc Đường Chéo: Động lực ban đầu của chúng tôi để lọc dữ liệu huấn luyện đến từ các nhận xét mâu thuẫn được thu thập trong vòng đầu tiên (R1). Việc đo đạc sự tương quan giữa các nhận xét của R1 và R2 xác nhận nghi ngờ của chúng tôi - sự tương quan tăng lên khi loại bỏ các mẫu mâu thuẫn. Việc lọc như vậy có thể hỗ trợ một số phương pháp
* Các thí nghiệm ban đầu cho thấy rằng các mô hình tiêu biểu hiện nay vượt trội so với con người. Một người gán nhãn ngẫu nhiên đạt được 0.832 điểm của hệ số tương quan Pearson và 0.777 điểm của hệ số tương quan Spearman với tập dữ liệu kiểm tra. Tuy nhiên, một mô hình tiên tiến với khả năng chú ý chéo vượt qua một con người ngẫu nhiên với 0.9189 điểm của hệ số tương quan Pearson và 0.8929 điểm của hệ số tương quan Spearman

10.Kết luận

* Chúng tôi kết luận bài báo của mình với một tổng hợp về các đặc điểm độc đáo của tập dữ liệu được giới thiệu. Kích thước lớn của tập dữ liệu (138.556 cặp câu được gán nhãn) cho phép huấn luyện và đánh giá mô hình ngữ nghĩa một cách mạnh mẽ. Phần kiểm tra của tập dữ liệu chứa các nhận xét dựa trên sự đồng thuận của chín người gán nhãn. Hơn nữa, chúng tôi đã thực hiện một phân tích chi tiết về các nhận xét kết quả và lọc ra những nhận xét không đáng tin cậy. Chúng tôi tính toán giới hạn dưới lý thuyết của sai số bình phương trung bình (MSE) là khoảng 0.1731. Con số này đáng kể thấp hơn (tốt hơn) so với hiệu suất của một người gán nhãn ngẫu nhiên. Do đó, phần kiểm tra cho phép đánh giá các mô hình hoạt động tốt. Cuối cùng, chúng tôi chỉ ra rằng tập dữ liệu của chúng tôi hỗ trợ việc huấn luyện các mô hình hoạt động tốt cho việc so sánh ngữ nghĩa của các câu. Chúng tôi cung cấp tập dữ liệu của mình và các mô hình cho sự tương đồng ngữ nghĩa có sẵn cho mục đích nghiên cứu