Bài báo viết về đo độ tượng đồng ngữ nghĩa của văn bản.

# Tóm tắt

Bài báo này mô tả một tập dữ liệu mới bao gồm các câu với hai loại chú thích độ tương đồng ngữ nghĩa khác nhau; có và không có ngữ cảnh xung quanh. Dữ liệu xuất phát từ lĩnh vực báo chí trong tiếng Séc. Tập dữ liệu cuối cùng chứa 138,556 chú thích của con người được phân chia thành các tập huấn luyện và kiểm tra. Tổng cộng, có 485 sinh viên ngành báo chí tham gia vào quá trình tạo ra nó. Để tăng độ tin cậy của tập kiểm tra, chúng tôi tính toán các chú thích cuối cùng như một trung bình của 9 điểm số chú thích cá nhân. Chúng tôi đánh giá chất lượng tập dữ liệu bằng cách đo lường sự đồng thuận giữa các nhà chú thích và trong nội bộ các nhà chú thích. Ngoài các con số đồng thuận, chúng tôi cung cấp thống kê chi tiết của tập dữ liệu đã thu thập. Chúng tôi kết thúc bài báo với một thí nghiệm cơ bản về việc xây dựng một hệ thống để dự đoán độ tương đồng ngữ nghĩa của các câu. Với số lượng chú thích huấn luyện khổng lồ (116,956), mô hình này vượt trội hơn hẳn một nhà chú thích trung bình (0.92 so với 0.86 của hệ số tương quan Pearson).

# Giới thiệu

Bài báo này mô tả một tập dữ liệu mới gồm các câu với các chú thích về độ tương đồng ngữ nghĩa. Tập dữ liệu bao gồm những cặp câu trong tiếng Séc, mỗi cặp được gán một điểm số tương đồng. Mục đích của tập dữ liệu là để huấn luyện và đánh giá các hệ thống dự đoán độ tương đồng ngữ nghĩa của các câu.

Hiện tại, lĩnh vực NLP ngày càng phụ thuộc vào các mô hình không giám sát hoặc tự giám sát. Tuy nhiên, vẫn cần những tập dữ liệu được chú thích tốt cho việc điều chỉnh hoặc kiểm tra mô hình. Chúng tôi chú trọng nhiều hơn đến phần kiểm tra của tập dữ liệu. Đối với phần này, mỗi cặp câu được chú thích độc lập bởi chín nhà chú thích. Do đó, chúng tôi lưu lại giá trị trung bình của các chú thích độc lập làm điểm số cuối cùng.

Với phần huấn luyện, chúng tôi muốn bao phủ càng nhiều dữ liệu đa dạng càng tốt. Vì vậy, chúng tôi chỉ sử dụng một chú thích cho mỗi cặp câu. Các mô hình NLP hiện tại đang ngày càng có khả năng xử lý các câu trong ngữ cảnh của chúng. Do đó, chúng tôi bao gồm ngữ cảnh của các câu khi tạo ra các cặp huấn luyện. Cuối cùng, chúng tôi chú thích các điểm số tương đồng có và không có ngữ cảnh.

Chúng tôi đã quyết định hợp tác với các sinh viên ngành báo chí để tạo ra những chú thích tốt hơn, vì họ thường có kỹ năng tốt trong việc xử lý dữ liệu văn bản. Để cải thiện các chú thích này, các phóng viên chuyên nghiệp đã giám sát các nhà chú thích sinh viên. Chúng tôi tin rằng sự đóng góp của những nhà chú thích có kỹ năng sẽ nâng cao chất lượng của tập dữ liệu.

Tập dữ liệu cuối cùng chứa 138,556 chú thích của con người. Việc tạo ra nó đòi hỏi một khối lượng công việc chú thích rất lớn – 485 nhà chú thích và thời gian dành cho việc tạo ra tập dữ liệu này khoảng 1,017 giờ công.

# Nghiên cứu liên quan

Về tiếng Anh, có rất nhiều tập dữ liệu cho độ tương đồng ngữ nghĩa (STS). Các tập dữ liệu thường được sử dụng nhất đến từ cuộc thi SemEval. Đã có sáu cuộc thi về STS kể từ năm 2012. (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017). Những tập dữ liệu này bao gồm các cặp câu được lấy từ bài báo, thảo luận trên diễn đàn, tiêu đề, và mô tả hình ảnh cũng như video, được gán điểm tương đồng từ 0 đến 5. Mục tiêu là để đánh giá xem khoảng cách cosin giữa hai câu tương quan như thế nào với điểm tương đồng được gán bởi con người thông qua các hệ số tương quan Pearson và Spearman. Các tập dữ liệu từ tất cả các cuộc thi STS của SemEval được tập hợp trong tập dữ liệu SentEval (Conneau và Kiela, 2018).

Còn về tiếng Séc, một tập dữ liệu STS cũng tồn tại. Svoboda và Brychcín (2018) đã tạo ra tập dữ liệu này bằng cách dịch các dữ liệu STS tiếng Anh từ SemEval sang tiếng Séc. Sau đó, họ đã chú thích các câu mà họ đã dịch. Nhược điểm lớn nhất của tập dữ liệu này là kích thước nhỏ (1425 cặp câu) so với tập dữ liệu được trình bày trong bài báo này.

# Nguồn dữ liệu

Dữ liệu văn bản thô cho tập dữ liệu được giới thiệu đến từ Cơ quan Tin tức Séc (CNA). CNA cung cấp dịch vụ toàn diện cho các nhà báo Séc, bao gồm hình ảnh, tin tức ngắn gọn, báo cáo, và theo dõi các sự kiện dài hạn trong nhiều lĩnh vực. Nhân viên của CNA công bố tin tức thông qua một quy trình chuẩn hóa nội bộ. Mỗi sự kiện mà CNA theo dõi được gọi là một sự cố. Mỗi sự cố đều đi kèm với tin tức từng phần, cập nhật sự cố theo thời gian (chúng tôi gọi những tin tức này là báo cáo trong bài viết này). Các sự cố thường được kết thúc bằng một bản tóm tắt do con người thực hiện. Chúng tôi sử dụng các bản tóm tắt và báo cáo được nhóm theo sự cố trong tập dữ liệu của mình.

Cơ sở dữ liệu gốc của các bản tóm tắt và báo cáo là tài sản riêng của CNA. Để công bố một phần dữ liệu gốc, chúng tôi đã đồng ý xây dựng tập dữ liệu sao cho không thể tái tạo hoàn toàn dữ liệu gốc. Chúng tôi chỉ công bố những câu rời rạc từ các bản tóm tắt và báo cáo với các ngữ cảnh hạn chế. Chúng tôi loại bỏ tất cả các mối quan hệ giữa các báo cáo.

# Quá trình thu thập các chú thích của con người

Dữ liệu văn bản để chú thích đến từ sự hợp tác của chúng tôi với CNA. Chúng tôi cố gắng giúp các nhà báo tự động phân tích các báo cáo liên quan khi tạo ra các bản tóm tắt. Mục tiêu của chúng tôi là huấn luyện một mô hình để biên soạn các bản tóm tắt từ các báo cáo gốc – hoặc ít nhất, giúp các nhà báo viết tóm tắt. Để có được dữ liệu liên quan cho việc huấn luyện mô hình, chúng tôi cần ghép cặp các câu từ báo cáo và tóm tắt.

Chúng tôi đã xây dựng một ứng dụng web để thu thập các chú thích của con người về sự tương đồng của các cặp câu nhằm hỗ trợ quá trình chú thích. Câu đầu tiên của cặp câu thuộc về bản tóm tắt, và câu còn lại thuộc về các báo cáo gốc (Xem Hình 1). Những người chú thích được yêu cầu cung cấp hai thông tin cơ bản:

1. sự tương đồng văn bản ngữ nghĩa không phụ thuộc vào ngữ cảnh,

2. sự tương đồng văn bản ngữ nghĩa phụ thuộc vào ngữ cảnh.

Các câu được lấy từ các báo cáo và các bản tóm tắt của chúng; do đó, chúng tôi có lý do để kỳ vọng rằng sẽ có những câu liên quan và tương đồng về ngữ nghĩa xuất hiện.

Về nguồn nhân lực, trong công việc này, chúng tôi đã hợp tác với hai nhóm sinh viên báo chí. Nhóm 1 (272 người) đã chú thích dữ liệu trong vòng đầu tiên. Họ đã chọn các cặp câu và chú thích sự tương đồng – dữ liệu này được sử dụng cho phần huấn luyện của tập dữ liệu. Nhóm 2 (229 người) tham gia vào vòng thứ hai để tạo ra phần kiểm tra của tập dữ liệu. Chúng tôi sẽ thảo luận chi tiết về quy trình trong các phần sau.

# Vòng thứ nhất – R1

Trong vòng đầu tiên (R1), chúng tôi đã yêu cầu những người chú thích chọn ba câu (An, Bn, Cn) từ các báo cáo cho mỗi câu (Sn) trong bản tóm tắt. Chúng tôi đã hướng dẫn họ chọn câu tương đồng nhất (A), câu ít tương đồng nhất (C), và một câu ở giữa (B), nhằm tạo ra một tập dữ liệu cân bằng hơn (Hình 2). Để hỗ trợ những người chú thích, chúng tôi đã làm nổi bật các từ từ các câu tóm tắt trong tất cả các báo cáo (Hình 1a). Chúng tôi đã yêu cầu người dùng sử dụng thanh trượt để chú thích mức độ tương đồng trên thang điểm từ 0 đến 6. Các mức độ được gán nhãn để giúp những người chú thích dễ dàng diễn giải hơn. Các nhãn là: 0: Hoàn toàn khác biệt, 1: Hơi liên quan về chủ đề, 2: Liên quan về chủ đề, 3: Tương đồng một phần, 4: Tương đồng, 5: Gần như giống hệt, 6: Giống hệt.

# Vòng thứ hai – R2

Chúng tôi đã tạo ra phần kiểm tra của tập dữ liệu trong vòng thứ hai (R2). Chúng tôi yêu cầu các người chú thích gán số độ tương đồng cho các cặp câu (SA, SB, SC) được tạo ra trong R1 trên cùng một thang điểm (0-6). Các cặp câu được chọn ngẫu nhiên nhưng không độc lập với nhau, vì chúng được thiết kế đặc biệt để kết hợp lại thành tập hợp đầy đủ các câu gốc (S, A, B, C) trong khi không có giao nhau (Xem Hình 3). Những người chú thích đã xử lý các chuỗi câu riêng lẻ trong R2 theo cùng một thứ tự như trong R1. Khoảng hai tháng trước khi bắt đầu chú thích chính, chúng tôi đã tiến hành một giai đoạn chú thích sơ bộ ngắn để thu thập 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 dừng lại 72 giờ giữa các ghi chú không có ngữ cảnh và có ngữ cảnh. Ở cuối cả hai giai đoạn, không có ngữ cảnh và có ngữ cảnh, chúng tôi đã thêm các khối đặc biệt để đo sự đồng thuận giữa các người chú thích. Để nhìn tổng quan về vòng thứ hai, xem Hình 5.

Khối hiệu chuẩn (R2-CAL) Trong giai đoạn chú thích sơ bộ ngắn, chúng tôi nhận thấy sự giảm sút trong sự đồng thuận giữa các người chú thích ở giai đoạn đầu công việc của họ (xem Hình 4b). Chúng tôi ước tính số cặp câu dẫn đầu cần thiết cho việc hiệu chuẩn người dùng thành công là 6. Do đó, chúng tôi đã thêm sáu cặp câu hiệu chuẩn trước giai đoạn chú thích chính. Các cặp này được chọn bằng tay sao cho điểm số STS của chúng bao phủ toàn bộ thang điểm chú thích.

Khối chính R2 (R2-MAIN) Trong khối R2 chính, các người chú thích đã làm việc với các cặp câu cấu thành phần kiểm tra của tập dữ liệu. Để có độ chính xác tốt hơn trong việc chú thích, chúng tôi quyết định sử dụng chín lần chú thích cho một cặp câu để loại bỏ độ nhiễu. Các mẫu đã được xáo trộn ngẫu nhiên để loại bỏ bất kỳ độ thiên lệch STS tiềm năng nào từ việc lặp lại một nhóm người chú thích có trực giác ngữ nghĩa tương tự.

Chúng tôi đã chọn kích thước khối chính là 50 để phù hợp với thời gian dự kiến tiêu tốn bằng giờ công. Do đó, mỗi người chú thích đã ghi chú chính xác 50 cặp STS, trước tiên là không có ngữ cảnh và sau đó bao gồm ngữ cảnh.

Khối giữa các người chú thích (R2-INTRA) Sau khối chính, mỗi người chú thích được yêu cầu ghi chú thêm năm cặp câu nữa, mà họ đã ghi chú trong giai đoạn chú thích sơ bộ ngắn (Xem Hình 6). Dữ liệu này được sử dụng để đo sự đồng thuận giữa các người chú thích.

# Thống kê tập dữ liệu

Tập dữ liệu cuối cùng được chia thành hai phần: phần huấn luyện và phần kiểm tra. Phần huấn luyện chứa 116.956 mẫu. Trong tập dữ liệu kiểm tra, chúng tôi quyết định chú thích 1.200 cặp cho cả hai biến thể không có ngữ cảnh và có ngữ cảnh. Chúng tôi thiết kế quy trình để mỗi cặp có chín lần chú thích. Tuy nhiên, trái với nỗ lực của chúng tôi, tập dữ liệu có ngữ cảnh nhỏ hơn khoảng 6%. Tuy nhiên, chỉ có 21 trong số 1.200 câu có ít hơn bảy lần chú thích.

Chúng tôi ước tính thời gian dành cho việc tạo tập dữ liệu huấn luyện (R1) là 876 giờ công (269 người chú thích, trung bình mỗi người khoảng 3,26 giờ). Các người chú thích đã tạo ra trong tập dữ liệu kiểm tra trong 141 giờ công (216 người chú thích, mỗi người trung bình khoảng 40 phút).

Thỏa thuận giữa các người chú thích Để khám phá giới hạn của sự thỏa thuận giữa con người, chúng tôi đã tính toán một số chỉ số phù hợp cho mục đích này.

Chúng tôi lưu trữ mười lần chú thích do người dùng thực hiện trong một giai đoạn chú thích sơ bộ ngắn, mà chúng tôi cho phép người dùng chú thích lại sau đó. Vì mỗi người dùng đã chú thích mười cặp (năm cặp cho cả hai loại không có ngữ cảnh và có ngữ cảnh), chúng tôi có thể đánh giá sự đồng thuận trong một người chú thích. Các điểm số do tất cả người dùng trong hai giai đoạn ngữ cảnh sản xuất được nối lại thành hai vectơ (giữ nguyên thứ tự của chúng). Chúng tôi tính toán hệ số tương quan Pearson và Spearman giữa hai vectơ này để định lượng mức độ tương quan của các điểm số trước đó với các điểm số của cùng một cặp và cùng một người đã chú thích sau này. Chúng tôi sử dụng hệ số tương quan cùng với lỗi bình phương trung bình (MSE) và lỗi căn bậc hai trung bình (RMSE) để cho thấy độ khác biệt trung bình giữa các lần chú thích của từng người chú thích.

Đối với sự đồng thuận giữa các người chú thích, chúng tôi đã lấy cảm hứng từ (Agirre et al., 2014), nơi các tác giả tính toán hệ số tương quan Pearson giữa các chú thích của từng người dùng và trung bình các chú thích của những người dùng khác (trên các cặp câu tương ứng), sau đó trung bình các hệ số tương quan riêng lẻ. Ngoài hệ số tương quan Pearson, chúng tôi cũng quyết định tính toán hệ số tương quan Spearman trung bình, RMSE trung bình và MSE trung bình; một lần nữa, để đánh giá xem mức độ khác biệt của những điểm số này như thế nào. Chúng tôi trình bày các chỉ số đồng thuận trong Bảng 1. Bây giờ hãy tưởng tượng rằng chúng tôi có một oracle cho STS, một máy lý thuyết luôn trả về STS đúng cho bất kỳ cặp câu nào. Điều này rất hữu ích trong việc phân tích hiệu suất của nó trên tập dữ liệu của chúng tôi, vì nó lý thuyết cung cấp một giới hạn cho chất lượng của bất kỳ mô hình học máy nào được đánh giá trên tập dữ liệu của chúng tôi. Tất nhiên, tập dữ liệu của chúng tôi không chứa các giá trị STS đúng; các quy trình chú thích của con người thường rất ồn ào. Do đó, vì chúng tôi không thể tìm thấy giới hạn lý thuyết cho các hệ số tương quan, chúng tôi chỉ tính toán và đưa ra một giới hạn dưới cho MSE.

Giới hạn dưới lý thuyết cho MSE Trong bối cảnh của một cặp câu đơn, giả sử rằng các điểm số STS mà các người chú thích của chúng tôi cung cấp trong R2 có phân phối gần như bình thường, thì biến ngẫu nhiên x¯−µ σ/√n (trong đó x¯ là trung bình mẫu STS, µ là trung bình STS thực, σ là độ lệch chuẩn mẫu STS và n là số lượng kết quả STS đã trung bình) có phân phối t của Student với 8 bậc tự do (vì chúng tôi có 9 người chú thích cho mỗi ví dụ). Khi quy mô biến ngẫu nhiên này bằng một yếu tố √σ/n, chúng tôi thu được phân phối của x¯−µ, với phương sai là MSE giữa một oracle cho STS (luôn trả về trung bình thực µ) và trung bình STS tương ứng x¯, mà nằm trong tập dữ liệu của chúng tôi. Do Var[aX] = a²Var[X] và phương sai của phân phối của Student với 8 bậc tự do là 1.¯3, chúng tôi ước lượng giới hạn dưới cho MSE là trung bình của 1.¯3 σ²/n cho tất cả các cặp câu, mà khoảng 0.1731.

Sự tương quan giữa các điểm số STS gốc và các tập dữ liệu thử nghiệm Để hoàn thiện, chúng tôi trình bày các hệ số tương quan của các điểm số STS giữa các điểm số được ghi lại trong R1 và các tập dữ liệu thử nghiệm không ngữ cảnh và có ngữ cảnh tương ứng (xem dòng đầu tiên trong Bảng 2). Chúng tôi thấy rằng cả hai tập dữ liệu thử nghiệm đều tương quan mạnh với các điểm số R1 gốc. Đặc biệt, sự tương quan giữa các điểm số STS không ngữ cảnh và có ngữ cảnh được thu thập trong R2 (và được tổng hợp vào các tập dữ liệu thử nghiệm) là rất cao. Kết quả phân tích của chúng tôi về sự tương đương của các điểm số STS không ngữ cảnh và có ngữ cảnh (được hình dung trong ...) cho thấy chỉ có 7.5% của tập dữ liệu thử nghiệm (90 cặp) là khác biệt đáng kể so với tương ứng ngữ cảnh của chúng.

Ảnh hưởng của ngữ cảnh đến STS Hình 8 cho thấy một điểm số STS của một cặp câu tăng hoặc giảm bao nhiêu sau khi ngữ cảnh được giới thiệu cho các người chú thích ở vòng hai (R2). Trong Hình 8a, chúng ta thấy rằng trung bình của sự tăng là lớn hơn không (chênh lệch chính xác là +0.26 của điểm số STS). Phân phối ban đầu không vượt qua bất kỳ bài kiểm tra tính bình thường nào (do tính nhiễm sắc cực của nó). Tuy nhiên, nếu chúng ta coi các sự gia tăng STS khoảng 1.5 là những điểm ngoài, thì có thể an toàn khi coi phân phối các gia tăng này gần như là phân phối bình thường với µ = 0.268 và σ = 0.504. Các bước rõ ràng trong Hình 8c được gây ra bởi phân phối không liên tục của các trung bình của 9 số nguyên giữa 0 và 6.

# Các thí nghiệm ban đầu

Để thiết lập một cơ sở cho tập dữ liệu mới, chúng tôi sử dụng mô hình word2vec nổi tiếng và mạnh mẽ (Mikolov et al., 2013) làm cơ sở, cùng với các mô hình hiện đại dựa trên kiến trúc BERT. Vì tập dữ liệu được trình bày có chứa dữ liệu bằng tiếng Séc, chúng tôi chọn các mô hình đã được huấn luyện trước trên ngôn ngữ Séc như Czert (Sido et al., 2021) và SlavicBert (Arkhipov et al., 2019). Các mô hình dựa trên BERT có thể xử lý câu theo hai cách: một là bộ mã hóa Cross-attention cho cả hai câu cùng một lúc; và hai là bộ mã hóa Two Tower/Siamese cho từng câu riêng biệt với một chỉ số độ tương đồng ở trên cùng.

Mô hình Word2Vec Chúng tôi sử dụng trung bình không trọng số của các nhúng word2vec để mã hóa các câu. Chúng tôi tính toán điểm số tương đồng bằng cách áp dụng độ tương đồng cosine trên các vectơ câu thu được.

Mô hình Two Tower Chúng tôi sử dụng một ngăn xếp dựa trên đầu ra pooler từ các mô hình giống BERT để mã hóa mỗi câu một cách độc lập. Sau đó, chúng tôi thêm chỉ số độ tương đồng cosine lên trên và điều chỉnh mô hình trên tập dữ liệu huấn luyện.

Mô hình Cross-Attention Đối với bộ mã hóa cross-attention, chúng tôi sử dụng đầu ra pooler với một lớp chiếu kích thước 200 kèm theo kích hoạt RELU, tiếp theo là một nơ-ron duy nhất với kích hoạt tuyến tính để có được chỉ số tương đồng. Một lần nữa, chúng tôi điều chỉnh mô hình trên tập dữ liệu huấn luyện. Các kết quả được trình bày trong Bảng 3 và thảo luận trong phần 9.

# Lọc tập dữ liệu

Trong quá trình xử lý sau, chúng tôi làm sạch phần dữ liệu huấn luyện bằng cách so sánh thống kê với phần kiểm tra. Chúng tôi giả định rằng phần kiểm tra chứa các chú thích đáng tin cậy nhờ vào giai đoạn khám phá và một vòng thứ hai được thiết kế cẩn thận. Các điểm số cuối cùng được tính bằng cách trung bình chín số, giúp phần kiểm tra trở nên vững chắc hơn.

Lọc Liên Quan Đến Sự Chuyển Nghĩa

Chúng tôi thiết kế quy trình thu thập dữ liệu đã được căn chỉnh không phụ thuộc vào ngữ cảnh và phụ thuộc vào ngữ cảnh để cho phép các nhà nghiên cứu trong tương lai xem xét vai trò của ngữ cảnh trong ngữ nghĩa. Chúng tôi không thực hiện bất kỳ lọc trước đặc biệt nào đối với các câu được trình bày cho những người chú thích, vì vậy sự phân bố tự nhiên (không thiên lệch) của sự chuyển nghĩa theo ngữ cảnh nên xuất hiện trong dữ liệu thu thập. Chúng tôi quyết định định lượng tầm quan trọng của ngữ cảnh bằng cách thực hiện một loạt các bài kiểm tra t, cụ thể là để kiểm tra sự khác biệt có ý nghĩa giữa các trung bình của các khối chính không phụ thuộc ngữ cảnh và phụ thuộc ngữ cảnh theo từng yếu tố (= một bài kiểm tra cho mỗi cặp câu). Chúng tôi giả định rằng các trung bình của STS với kích thước chín mẫu cho mỗi cặp câu đơn và sự hiện diện ngữ cảnh khoảng phân bố theo phân phối chuẩn. Để tóm tắt - cho mỗi cặp câu (của các khối chính), chúng tôi có chín điểm số STS cho phiên bản không phụ thuộc ngữ cảnh và chín điểm số STS cho phiên bản phụ thuộc ngữ cảnh của cặp câu. Sau đó, chúng tôi thực hiện một bài kiểm tra t hai phía để so sánh tính tương đương của các trung bình tương ứng. Giả thuyết không của bài kiểm tra này là các trung bình điểm số STS là bằng nhau; nói cách khác, ngữ cảnh bổ sung là không quan trọng trong lĩnh vực STS. Mỗi bài kiểm tra như vậy tạo ra một giá trị p nhất định, tức là, mức độ ý nghĩa lớn nhất mà tại đó giả thuyết không bị bác bỏ. Hàm r(x) được định nghĩa là tỷ lệ của các cặp câu mà chúng tôi bác bỏ giả thuyết không liên quan đến ngữ cảnh, nếu mức độ ý nghĩa của chúng tôi là x, tức là, CDF của phân phối các giá trị p (Xem Hình 9).

Chúng tôi nhận thấy rằng giá trị p là 0.05 cho 7.50% các câu có ghi chú ngữ cảnh khác nhau và giá trị p là 0.01 cho 1.83%. Chúng tôi tạo ra một tập dữ liệu mới từ các câu có ghi chú ngữ cảnh khác biệt rõ rệt tại giá trị p là 0.05. Chúng tôi gọi các biến thể của tập dữ liệu này là test-sig free và test-sig dep cho ghi chú không phụ thuộc ngữ cảnh và phụ thuộc ngữ cảnh tương ứng. Bảng 4 trình bày thêm các thống kê (trung bình điểm ghi chú, phương sai, số lượng) của các tập dữ liệu.

Lọc Chéo Dữ Liệu Huấn Luyện Sau Khi Ghi Chú, chúng tôi tính toán một số thống kê bổ sung cho dữ liệu đã thu thập. Chúng tôi phát hiện ra rằng có một số trường hợp biên giới trong giai đoạn đầu. Chúng tôi đã thực hiện phân tích tần suất cho mỗi đánh giá (0–6) cho ba lớp đề xuất (xanh, cam, đỏ) (Hình 10). Do kết quả mong đợi, chúng tôi đã ngạc nhiên bởi một số mẫu được ghi chú không theo sơ đồ này. Có một số mẫu nhỏ, nhưng không bằng không, được chọn cho lớp xanh (nên giống nhau) nhưng lại được đánh giá với số điểm tương đồng thấp, và ngược lại, những mẫu đỏ (không nên giống nhau) lại được đánh giá với số điểm tương đồng cao.

Chúng tôi cho rằng điều này có thể do một số lý do. Nếu chúng tôi bỏ qua những sai sót của con người, thì nguyên nhân có khả năng nhất là dữ liệu được hiển thị cho người ghi chú không thể được ghi chú theo cách khác. Có thể chỉ có một sự khác biệt nhỏ giữa các báo cáo và tóm tắt, vì vậy người ghi chú không thể tìm thấy câu nào quá khác biệt, và họ đã buộc phải chọn một câu tương tự ngay cả cho lớp đỏ. Tuy nhiên, chúng tôi tin rằng những ví dụ không bình thường như vậy chỉ xuất hiện trong số ít người ghi chú.

Không có nhiều lý do khả thi cho việc dữ liệu bị thiên lệch theo hướng ngược lại. Một lần nữa, nếu chúng ta bỏ qua sai sót của con người, điều này khó xảy ra, thì lý do duy nhất cho sự thiên lệch hệ thống như vậy (cũng quan sát thấy ở lớp cam) là người ghi chú không có lựa chọn nào khác. Nguồn duy nhất hợp lý của sự thiên lệch này là sự xuất hiện của thông tin hoàn toàn mới trong tóm tắt hoặc các câu không được sử dụng trong báo cáo gốc. Chúng tôi đã kiểm tra khả năng này với nhà báo - Họ gọi đó là bối cảnh, và lý do sử dụng nó là để đặt tóm tắt vào một số bối cảnh (thường là trước đó). Và, đương nhiên, họ thường diễn đạt lại các câu gốc và lồng ghép các thông tin gốc trong khi tạo ra tóm tắt.

Việc thu thập dữ liệu trong vòng đầu tiên có thể mang lại sự thiên lệch hệ thống trong dữ liệu huấn luyện - con người đánh dấu các cặp khác nhau là xanh lá để chọn số điểm tương đồng cao hơn. Chúng tôi quyết định điều tra giả thuyết này bằng cách lọc sự thiên lệch hệ thống có thể có trong dữ liệu huấn luyện bằng cách chỉ sử dụng hàng xóm gần nhất có sự chênh lệch điểm số là 1 và 2 của lớp màu dự định (N1, N2), trong đó số là chênh lệch tối đa của điểm số so với giá trị mong đợi (xanh lá = 6, cam = 3, đỏ = 0). Để biết thống kê và kết quả của các thí nghiệm cơ bản, xem Bảng 2, 5 và Hình 7.

Bảng 2 cũng cho thấy tác động đáng kể mà các phương pháp lọc dữ liệu R1 của chúng tôi có đến mối tương quan STS giữa các dữ liệu R1 và dữ liệu kiểm tra tương ứng. Chúng tôi có thể thấy sự cải thiện nhẹ về mối tương quan sau khi lọc sử dụng N2, và một cải thiện nhỏ nữa sau khi lọc sử dụng N1. Để chứng minh rằng việc cải thiện mối tương quan không phải do thu hẹp kích thước giao nhau, chúng tôi lấy mẫu một bộ dữ liệu kiểm tra cắt bớt, tính toán mối tương quan giữa nó và phần tương ứng của toàn bộ bộ dữ liệu R1, và trung bình các kết quả. Sự khác biệt đáng chú ý là giữa N1 và N1-ngẫu nhiên, cho thấy rằng chỉ đơn giản loại bỏ ngẫu nhiên 287 phần tử từ bộ dữ liệu kiểm tra là không đủ để cải thiện mối tương quan.

Các thống kê bổ sung (trung bình, phương sai và kích thước) của các biến thể dữ liệu có sẵn trong bảng 4.

# Định dạng dữ liệu

Do những lý do được mô tả trong Phần 3, chúng tôi có thể phát hành bộ dữ liệu chứa các câu với bối cảnh hạn chế. Chúng tôi chỉ thực hiện các thí nghiệm ban đầu với những tệp này. Do đó, các nhà nghiên cứu tương lai có thể sử dụng cùng một dữ liệu như chúng tôi đã làm. Thật không may, chúng tôi không thể phát hành cơ sở dữ liệu gốc với dữ liệu thô được thu thập trong quá trình ghi chú. Chúng tôi cố gắng mang đến cho người đọc cái nhìn tốt nhất về toàn bộ quá trình, đặc điểm của dữ liệu gốc và chất lượng của tập dữ liệu đã được ghi chú cuối cùng. Chúng tôi công bố tất cả các phiên bản có và không có việc lọc đã được thực hiện.

Chúng tôi trình bày dữ liệu đã thu thập trong các tệp văn bản có sẵn trên trang web của chúng tôi và trên github. Số lượng ghi chú trung bình cho cùng một cặp và tổng hợp tất cả các ghi chú được trình bày trong các tệp phần kiểm tra.

Chúng tôi cũng trình bày bộ dữ liệu kiểm tra đã được lọc theo mức độ quan trọng của sự thay đổi giữa các ghi chú không có bối cảnh và có bối cảnh, được gán nhãn là testsig.tsv. Các mẫu dữ liệu được lọc dựa trên mức độ tin cậy 0,95 về sự khác biệt đáng kể giữa các ghi chú không có bối cảnh và có bối cảnh trong tệp này.

Các phần kiểm tra không có bối cảnh và có bối cảnh thì tự nhiên được căn chỉnh. Tuy nhiên, phần không có bối cảnh lớn hơn khoảng 6% so với phần có bối cảnh.

Phần dữ liệu huấn luyện của bộ dữ liệu gồm hai câu với ghi chú của người dùng được thực hiện trong R1. Phần kiểm tra chứa hai câu theo sau là giá trị STS trung bình từ R2, tất cả các ghi chú gốc được thu thập trong R2 và giá trị từ R1.

Câu chính được bao quanh bằng các dấu <sent></sent> trong các biến thể tập dữ liệu có bối cảnh.

# Thảo luận

Vai trò của Bối Cảnh Mục tiêu chính là tạo ra một bộ dữ liệu mới của Séc về sự tương đồng ngữ nghĩa văn bản với các ghi chú không có bối cảnh và có bối cảnh được căn chỉnh, cũng như đánh giá tầm quan trọng của bối cảnh trong các văn bản thông thường. Như đã chỉ ra, bối cảnh ảnh hưởng đáng kể đến một tập hợp con của dữ liệu đã thu thập. Trong lĩnh vực hẹp của tin tức và tóm tắt của chúng, chúng tôi quan sát thấy 90 mẫu từ 1200 (7,5%) bị thay đổi đáng kể. Chúng tôi tập hợp những mẫu này thành các bộ dữ liệu riêng biệt để các nhà nghiên cứu trong tương lai có thể sử dụng các mô hình tiên tiến nhận thức bối cảnh để thể hiện lợi ích của chúng.

Vai trò của Lọc Đường Chéo Động lực ban đầu của chúng tôi để lọc dữ liệu huấn luyện đến từ những ghi chú mâu thuẫn được thu thập trong vòng đầu tiên. Đo lường mối tương quan giữa ghi chú R1 và R2 xác nhận nghi ngờ của chúng tôi - mối tương quan tăng khi loại bỏ các mẫu mâu thuẫn. Việc lọc như vậy có thể giúp một số phương pháp; tuy nhiên, chúng tôi xác nhận một nguyên lý đã được biết đến chung là các chỉ số đánh giá cao hơn trên các phiên bản không được lọc - các mô hình sâu có thể hưởng lợi từ các bộ dữ liệu lớn hơn mặc dù có tiếng ồn hơn là từ các phiên bản sạch hơn và nhỏ hơn. Tuy nhiên, chúng tôi đã quyết định công khai chia sẻ cả hai bộ dữ liệu đã lọc (N2, N1).

Phương pháp và phương sai Chúng tôi đánh giá phân phối các điểm số trong các phần chia của bộ dữ liệu và các phiên bản khác nhau của chúng. Các thống kê được tóm tắt trong Bảng 4. Chúng tôi có thể thấy rằng trung bình của bộ dữ liệu kiểm tra có bối cảnh cao hơn đáng kể so với trung bình của bộ dữ liệu không có bối cảnh. Sự quan sát này có lý vì với một bối cảnh lớn hơn, nhiều thông tin hơn có mặt trong văn bản, và có nhiều khả năng cho một số sự chồng chéo chủ đề. Tiếp theo, bộ dữ liệu kiểm tra có trung bình cao hơn và phương sai thấp hơn. Điều này là do việc lấy trung bình điểm số của chín người ghi chú trong bộ dữ liệu kiểm tra. Theo định lý giới hạn trung tâm, phương sai của trung bình thấp hơn và trung bình gần hơn với trung tâm của khoảng.

Các Thí Nghiệm Ban Đầu Các thí nghiệm ban đầu chỉ ra rằng các mô hình vani tốt nhất trong ngành vượt trội hơn một người ngẫu nhiên một cách dễ dàng. Một người ghi chú ngẫu nhiên đạt được hệ số tương quan Pearson là 0.832 và hệ số tương quan Spearman là 0.777 với bộ kiểm tra của bộ dữ liệu. Tuy nhiên, mô hình chú ý chéo tiên tiến hơn một người ngẫu nhiên với 0.9189 hệ số tương quan Pearson và 0.8929 hệ số tương quan Spearman. Chúng tôi tin rằng sự khác biệt lớn như vậy đến từ khả năng của một mô hình máy tính trong việc nắm bắt sự đồng thuận của một nhóm người lớn. Chúng tôi có thể đo lường hiệu suất cao hơn hiệu suất của từng người ghi chú vì phần kiểm tra của bộ dữ liệu là một trung bình của 9 ghi chú.

# Kết luận

Chúng tôi kết thúc bài báo của mình bằng cách tóm tắt những đặc điểm nổi bật của bộ dữ liệu được giới thiệu. Kích thước lớn của bộ dữ liệu (138,556 cặp câu được chú thích) cho phép đào tạo và đánh giá các mô hình ngữ nghĩa một cách mạnh mẽ. Bộ dữ liệu này nằm trong số những tài nguyên đào tạo không phải tiếng Anh lớn nhất để học ngữ nghĩa của một ngôn ngữ. Phần kiểm tra của bộ dữ liệu chứa các chú thích dựa trên sự đồng thuận của chín người chú thích. Hơn nữa, chúng tôi đã thực hiện một phân tích chi tiết về các chú thích kết quả và lọc ra những chú thích không đáng tin cậy. Chúng tôi tính toán giới hạn lý thuyết của MSE khoảng 0.1731. Số này thấp hơn (tốt hơn) nhiều so với hiệu suất của một người chú thích ngẫu nhiên. Do đó, phần kiểm tra cho phép đánh giá các mô hình hoạt động tốt. Cuối cùng, chúng tôi chỉ ra rằng bộ dữ liệu của chúng tôi hỗ trợ đào tạo các mô hình hiệu quả cho sự tương đồng ngữ nghĩa của các câu. Mô hình chú ý chéo của chúng tôi vượt trội hơn đáng kể so với một người chú thích trung bình. Chúng tôi cung cấp bộ dữ liệu và các mô hình cho sự tương đồng ngữ nghĩa công khai để phục vụ cho mục đích nghiên cứu.

Cảm ơn

Công việc này một phần được hỗ trợ bởi Cơ quan Công nghệ Cộng hòa Séc trong khuôn khổ Chương trình ETA – dự án TL02000288. Tài nguyên tính toán được cung cấp bởi dự án "e-Infrastruktura CZ" (e-INFRA LM2018140) trong khuôn khổ chương trình Dự án Cơ sở hạ tầng Nghiên cứu, Phát triển và Đổi mới Lớn.