Impoving Information Extraction from Visually Rich Documents using Visual Span Representations

(Cải thiện việc trích xuất thông tin từ các tài liệu trực quan phong phú bằng cách sử dụng biểu diễn khoảng cách trực quan­)

Các từ dịch:

Visual features: tính năng trực quan

Visual element: phần tử trực quan

1. Tóm tắt

Cùng với nội dung văn bản, các tính năng trực quan đóng một vai trò thiết yếu trong ngữ nghĩa của các tài liệu giàu tính trực quan. Các tác vụ trích xuất thông tin (IE) hoạt động kém trên các tài liệu này nếu các dấu hiệu trực quan này không được tính đến. Trong bài báo này, chúng tôi trình bày Artemis - một phương pháp IE dựa trên máy học, nhận thức trực quan cho các tài liệu đa dạng về hình ảnh. Artemis đại diện cho một khoảng trực quan trong tài liệu bằng cách cùng mã hóa ngữ cảnh văn bản và hình ảnh của nó cho các tác vụ IE. Đóng góp chính của chúng tôi là gấp hai lần. Đầu tiên, chúng tôi phát triển một mô hình học sâu xác định ranh giới ngữ cảnh cục bộ của một khoảng trực quan với sự gắn nhãn tối thiểu của con người. Thứ hai, chúng tôi mô tả một mạng nơ-ron sâu mã hóa ngữ cảnh đa phương thức của khoảng trực quan thành một vectơ có độ dài cố định bằng cách tính đến các đặc điểm cụ thể về văn bản và bố cục của nó. Nó xác định (các) khoảng trực quan chứa một thực thể được đặt tên bằng cách tận dụng biểu diễn đã học này, sau đó là một nhiệm vụ suy luận. Chúng tôi đánh giá Artemis trên bốn bộ dữ liệu không đồng nhất từ ​​các miền khác nhau qua một bộ các tác vụ khai thác thông tin. Kết quả cho thấy nó hoạt động tốt hơn các phương pháp dựa trên văn bản hiện đại lên tới 17 điểm trong điểm F1

1. Giới thiệu

Một số lượng đáng kể các tài liệu chúng ta gặp hàng ngày có tính chất trực quan phong phú. Cùng với các dấu hiệu ngôn ngữ, chúng sử dụng một số đặc điểm hình ảnh rõ ràng (ví dụ: định vị tương đối, cỡ chữ, màu chữ) và ẩn (ví dụ: khoảng cách âm, cân bằng khoảng trắng) để tăng cường hoặc làm nổi bật thông tin mà chúng phổ biến. Cho dù gọi món từ thực đơn nhà hàng, so sánh tài sản trong tờ rơi bất động sản hay tìm kiếm sự kiện trên áp phích, thông tin chứa trong các tài liệu này thường là đặc biệt, tức là chúng không thể được truy xuất từ ​​cơ sở dữ liệu được lập chỉ mục một cách khá dễ dàng. Các phương pháp dựa trên văn bản không giải quyết được các thách thức của việc trích xuất thông tin (IE) trong các tình huống này, vì chúng không xem xét sự đóng góp của các tính năng trực quan vào ngữ nghĩa của các tài liệu này, do đó khiến người dùng không có nhiều lựa chọn hơn là phân tích cú pháp toàn bộ tài liệu. trích xuất thông tin liên quan. Điều gì làm cho nhiệm vụ tự động hóa việc khai thác khó khăn hơn là bản chất không đồng nhất của các tài liệu này. Các tài liệu trong thế giới thực có thể đa dạng về bố cục, định dạng và nội dung. Những thách thức này tích tụ để thúc đẩy nhu cầu về một phương pháp tổng quát cho IE từ các tài liệu trực quan phong phú. Trước khi chúng tôi trình bày phương pháp của mình, để giải thích rõ hơn những đóng góp của chúng tôi, chúng tôi chứng minh những hạn chế của phương pháp chỉ văn bản cho IE từ các tài liệu trực quan phong phú trong một ví dụ bên dưới. Ví dụ 1.1: Alice, một nhà phân tích nghiên cứu chứng nghiện thuốc lá, muốn thực hiện một nghiên cứu dài hạn về mức độ ảnh hưởng của việc sử dụng thuốc lá trên báo in. Cô đã thu thập một số bài báo từ các tờ báo và tạp chí quốc gia về chủ đề này được xuất bản trong ba thập kỷ qua. Các tài liệu trong bộ sưu tập của Alice rất phong phú về mặt hình ảnh và không đồng nhất về bản chất, tức là chúng có bố cục, định dạng và nội dung đa dạng. Mặc dù hầu hết các bài báo được xuất bản gần đây đều ở dạng kỹ thuật số và có sẵn ở định dạng HTML, các bài báo cũ hơn chỉ có sẵn dưới dạng bản sao được quét có độ phân giải cao. Alice muốn trích xuất ba thực thể được đặt tên từ mỗi tài liệu trong bộ sưu tập của cô ấy: 𝑁 = {‘Tên sản phẩm thuốc lá’, ‘Tên nhà sản xuất’, ‘Tác dụng phụ được liệt kê’}. Đối với mỗi thực thể được đặt tên 𝑛𝑖 𝜖 𝑁, 𝑖 = {1, 2, 3}, cô ấy muốn xác định vị trí mà thực thể được đặt tên xuất hiện trong tài liệu, phiên âm nó và lưu trữ nó trong một tệp CSV. Một phương pháp IE dựa trên văn bản, trong các tình huống như vậy, thường bắt đầu với việc làm sạch tài liệu trước. Điều này bao gồm cong vênh phối cảnh, chỉnh lệch và làm mờ. Sau đó, tài liệu được phiên âm, văn bản của nó được chuẩn hóa và các từ dừng được xóa. Cuối cùng, các thực thể được đặt tên được trích xuất bằng cách xác định các khoảng văn bản có chứa các mẫu cú pháp hoặc ngữ nghĩa cụ thể đại diện cho một thực thể được đặt tên 𝑛𝑖 ∈ 𝑁. Thách thức: Mặc dù hợp lý đối với các tài liệu văn bản thuần túy, nhưng việc tuân theo cách tiếp cận tương tự mà bộ sưu tập của Alice có thể gặp khó khăn. Thứ nhất, nếu không biết trước bố cục tài liệu, các trình trích xuất dựa trên văn bản không phải lúc nào cũng dẫn đến hiệu suất trích xuất đáng tin cậy [51]. Các lỗi được đưa ra trong quá trình nhận dạng ký tự quang học và tuần tự hóa văn bản được phiên mã dẫn đến phân tích ngữ nghĩa không chính xác, ảnh hưởng đến tác vụ IE hạ lưu. Hơn nữa, sự đóng góp của các đặc điểm trực quan (xem Hình 1) về ngữ nghĩa chẳng hạn như vị trí tương đối và sự giống nhau về kích thước phông chữ thường không được ghi lại trong phiên mã. Điều này thúc đẩy nhu cầu mã hóa khoảng trực quan (được định nghĩa trong Phần 3) trong tài liệu bằng cách tính đến cả phương thức văn bản và hình ảnh của nó. Thứ hai, một phương pháp tổng quát để trích xuất thông tin phải mạnh mẽ [48] hướng tới các bố cục và định dạng đa dạng (ví dụ: XML, HTML). Nó cũng phải đủ linh hoạt để được mở rộng cho các tác vụ trích xuất khác nhau trên các bộ dữ liệu khác nhau. Giả thuyết của chúng tôi: Các công trình gần đây đã cho thấy tầm quan trọng của việc kết hợp thông tin ngữ cảnh về hiệu suất của các phương pháp IE dựa trên văn bản. Từ các đặc điểm cú pháp / ngữ nghĩa được thiết kế cẩn thận [29, 58, 60] đến mạng nơ-ron sâu [25], các màn trình diễn hiện đại đã được báo cáo bằng cách mã hóa các tín hiệu ngữ cảnh từ cấp độ cụm từ, cấp độ câu và cấp độ đoạn văn trừu tượng của một tài liệu văn bản. Chúng tôi cho rằng việc kết hợp các dấu hiệu ngữ cảnh tương tự cũng có thể tăng hiệu suất trích xuất từ ​​các tài liệu trực quan phong phú. Ai cũng biết rằng hệ thống nhận thức của con người tích hợp thông tin ngữ cảnh từ nhiều phương thức [43] để hiểu dữ liệu trong thế giới thực. Chúng tôi đưa ra giả thuyết rằng việc bắt chước hành vi này để mã hóa thông tin ngữ cảnh đa phương thức có thể giúp giải quyết một số thách thức mà các phương pháp IE dựa trên văn bản đã đề cập trước đây. Ví dụ: khoảng thời gian trực quan xuất hiện gần đầu một bài báo có thể là trường ngày tháng. Do đó, nó có nhiều khả năng được phiên âm là "11 tháng 11" hơn là "ll Gov". Chúng tôi lập luận rằng việc biểu diễn một khoảng trực quan bằng cách mã hóa các dấu hiệu ngữ cảnh như vậy có thể nâng cao hiệu suất của các phương pháp IE trên các tài liệu như vậy. Thật không may, trong khi các mô hình học sâu như mạng nơ-ron lặp lại [1] có hiệu quả trong việc mô hình hóa thông tin ngữ cảnh cho các tài liệu văn bản, chúng sẽ bị thiếu hụt nếu các dấu hiệu ngữ cảnh trải dài trên cả phương thức văn bản và hình ảnh [23]. Theo Doan et al. [9], chúng tôi xác định nhiệm vụ IE được đề cập trong bài báo này là nhiệm vụ truy xuất khóa-giá trị trên tài liệu đầu vào. Mỗi khóa đại diện cho một thực thể được đặt tên sẽ được trích xuất và giá trị của nó biểu thị khoảng trực quan nhỏ nhất có chứa đề cập thực sự (được định nghĩa trong Phần 3) của thực thể được đặt tên đó.

Đóng góp và Tổng quan: Chúng tôi trình bày Artemis - một phương pháp tổng quát để trích xuất thông tin từ các tài liệu không đồng nhất, giàu hình ảnh trong bài báo này. Nó lấy một tài liệu 𝐷 và danh sách các kiểu thực thể được đặt tên 𝑁 làm đầu vào và trả về khoảng trực quan nhỏ nhất trong tài liệu có chứa một thực thể được đặt tên 𝑛𝑖 ∈ 𝑁, ∀𝑖. Artemis hoạt động theo hai giai đoạn.

Đầu tiên, nó xác định một tập hợp các nhịp trực quan ứng viên trong 𝐷 có khả năng chứa các thực thể được đặt tên là 𝑛𝑖 ∈ 𝑁, ∀𝑖.

Sau đó, nó xác định các khoảng trực quan nhỏ nhất trong các ứng cử viên này có chứa các thực thể được đặt tên trong giai đoạn thứ hai của quy trình làm việc của nó. Chúng tôi biểu diễn từng khoảng thời gian trực quan bằng cách sử dụng hai vectơ ngữ cảnh có độ dài cố định trong quy trình làm việc của chúng tôi. Để mã hóa thông tin ngữ cảnh cục bộ của nó, chúng tôi mở rộng mạng bộ nhớ ngắn hạn dài hạn (LSTM) hai chiều (chi tiết trong Phần 4.3) - tiêu chuẩn học sâu thực tế trong xử lý ngôn ngữ tự nhiên [33]. Nó mã hóa cả thuộc tính trực quan và văn bản của một khoảng trực quan từ ranh giới ngữ cảnh cục bộ của nó (được định nghĩa trong Phần 3). Để xác định ranh giới ngữ cảnh cục bộ của khoảng trực quan, chúng tôi phát triển mô hình học máy [12] dựa trên mạng nơron đối nghịch (chi tiết trong Phần 4.2) kết hợp kiến ​​thức miền cụ thể để xác định các dấu hiệu trực quan hoạt động như ranh giới khái niệm cục bộ . Tổng quan về quy trình trích xuất của chúng tôi được trình bày trong Hình 2. Artemis xây dựng dựa trên các tác phẩm hiện có không kết hợp các dấu hiệu ngữ cảnh dành riêng cho miền [51] hoặc dựa vào các tính năng định dạng / bố cục cụ thể [4, 28, 59] để xác định tên các thực thể trong một tài liệu trực quan phong phú. Trái ngược với những nỗ lực này, chúng tôi không tận dụng bất kỳ tính năng định dạng cụ thể nào hoặc thừa nhận kiến ​​thức rõ ràng về bố cục tài liệu ở bất kỳ giai đoạn nào trong quy trình làm việc của chúng tôi. Thông qua các thử nghiệm toàn diện trên bốn tập dữ liệu tên miền chéo, chúng tôi cho thấy rằng: (a) các mô hình học sâu hiện có được thiết kế riêng cho việc trích xuất thông tin dựa trên văn bản, ví dụ, mạng bộ nhớ ngắn hạn dài (LSTM) [1] phải vật lộn để nắm bắt tính đa phương thức của các tài liệu trực quan, và (b) nhập thông tin miền cụ thể để tìm hiểu cách biểu diễn khoảng cách trực quan giúp tăng hiệu suất trích xuất đầu cuối. Những đóng góp chính của chúng tôi như sau: • Chúng tôi mô tả Artemis, một phương pháp IE nhận biết trực quan dành cho các tài liệu có hình ảnh phong phú không đồng nhất • Chúng tôi đề xuất cách trình bày theo ngữ cảnh cho khoảng trực quan trong tài liệu. Để xác định ranh giới ngữ cảnh của khoảng trực quan trong khi kết hợp kiến ​​thức theo miền cụ thể, chúng tôi phát triển mô hình học máy dựa trên mạng nơ-ron đối nghịch. • Thông qua đánh giá toàn diện trên bốn tập dữ liệu không đồng nhất cho các tác vụ riêng biệt, chúng tôi cho thấy rằng IE sử dụng biểu diễn nhịp được ngữ cảnh hóa của chúng tôi hoạt động tốt hơn các đường cơ sở chỉ có văn bản Chúng tôi chính thức hóa nhiệm vụ IE được đề cập trong bài báo này trong Phần 2 và mô tả phương pháp của chúng tôi trong Phần 4. Chúng tôi đánh giá phương pháp trên bốn nhiệm vụ IE riêng biệt và thảo luận về những phát hiện của chúng tôi trong Phần 5.

1. Artermis framwork

Artemis hoạt động theo hai giai đoạn.

Trong giai đoạn đầu, nó xác định một tập hợp các khoảng hình ảnh ứng viên cho mọi thực thể được đặt tên sẽ được trích xuất. Mỗi nhịp ứng cử viên này sau đó được tìm kiếm để xác định các nhịp hình ảnh nhỏ nhất bên trong chúng có chứa một thực thể được đặt tên trong giai đoạn thứ hai. Yếu tố quan trọng của cả hai giai đoạn là một mô hình dữ liệu đại diện cho các tài liệu không đồng nhất theo một cách có nguyên tắc. Chúng tôi mô tả nó dưới đây

* 1. Atermis Data Model
* 1 tài liệu ( document) sẽ được biểu diễn dưới dạng một tuple (V, T)

V : tập hợp các atomic element trong tài liệu

T: visual organization

* + 1. Atomic element
* Atomic element đề cập đến phần tử trực quan nhỏ nhất trong tài liệu
* Có 2 loại atomic element trong 1 tài liệu: text element, image element
* Text element: Nó là phần tử hình ảnh nhỏ nhất trong tài liệu có thuộc tính text-thuộc tính. Giả sử một hệ tọa độ hình chữ nhật với điểm gốc ở góc trên cùng bên trái, chúng ta có thể biểu diễn phần tử văn bản 𝑎𝑡 dưới dạng bộ tuple lồng nhau (𝑡𝑒𝑥𝑡-𝑑𝑎𝑡𝑎, 𝑥, 𝑦, 𝑤, ℎ). Ở đây, 𝑡𝑒𝑥𝑡-𝑑𝑎𝑡𝑎 đại diện cho phiên âm của khoảng hình ảnh được bao phủ bởi 𝑎𝑡, ℎ & 𝑤 biểu thị chiều cao và chiều rộng của hộp giới hạn nhỏ nhất bao quanh 𝑎𝑡. 𝑥 & 𝑦 đại diện cho tọa độ của góc trên bên trái của hộp giới hạn. Chúng tôi phiên âm một khoảng thời gian trực quan bằng cách sử dụng Tesseract [54], một công cụ phiên mã nguồn mở phổ biến. Chúng tôi coi mỗi từ là một phần tử văn bản trong mô hình dữ liệu của mình.

B. Phần tử hình ảnh: Nó biểu thị một thuộc tính hình ảnh trong tài liệu. Chúng tôi biểu diễn một phần tử hình ảnh 𝑎𝑖 dưới dạng một bộ tuple lồng nhau (𝑖𝑚𝑎𝑔𝑒- 𝑑𝑎𝑡𝑎, 𝑥, 𝑦, 𝑤, ℎ). Ở đây, 𝑖𝑚𝑎𝑔𝑒-𝑑𝑎𝑡𝑎 biểu thị bản đồ pixel, ℎ & 𝑤 biểu thị chiều cao và chiều rộng, và 𝑥 & 𝑦 biểu thị tọa độ của góc trên cùng bên trái của hộp giới hạn nhỏ nhất bao quanh 𝑎t

4.1.2 Visual organization ( tổ chức trực quan)

Chúng tôi đại diện cho tổ chức trực quan của một tài liệu bằng cách sử dụng cấu trúc giống như cây (𝑇). Mỗi nút trong 𝑇 đại diện cho một khoảng trực quan ở các cấp độ khác nhau của hệ thống phân cấp bố cục. Theo định dạng đặc tả hOCR của Breuel et al. [2], chúng tôi xác định𝑇 với năm cấp độ phân cấp. Mỗi cấp độ tương ứng với một cấp độ trừu tượng cụ thể. Chúng tôi xác định hệ thống phân cấp như sau: mỗi tài liệu được tạo thành từ nhiều cột, mỗi cột được chia thành một số đoạn văn, mỗi đoạn văn được chia thành các dòng văn bản và mỗi dòng văn bản bao gồm nhiều từ. Chúng tôi đại diện cho 𝑇 bằng cách tận dụng hệ thống phân cấp này. Một nút (𝑣1) là con của một nút khác (𝑣2) trong𝑇 nếu khoảng trực quan được đại diện bởi 𝑣1 được bao bởi khoảng được đại diện bởi 𝑣2. Các nút lá của 𝑇 đại diện cho các từ riêng lẻ, trong khi nút gốc đại diện cho vùng trực quan được bao phủ bởi trang đơn của tài liệu. Chúng tôi sử dụng thuật toán phân đoạn trang mã nguồn mở [54] để xây dựng cây bố cục của tài liệu.1 Mô hình dữ liệu của chúng tôi giúp biểu diễn các tài liệu có bố cục và định dạng đa dạng theo cách có nguyên tắc. Tận dụng cấu trúc đã trình bày, chúng tôi xây dựng cây bố cục của một tài liệu từ hình ảnh được kết xuất của nó thay vì sử dụng cấu trúc cây DOM vì nó giúp chúng tôi mở rộng phương pháp của mình sang các định dạng tài liệu không hỗ trợ đặc tả cây DOM ở trên, chúng tôi có thể biểu diễn bất kỳ hình ảnh nào span trong tài liệu dưới dạng một tuple lồng nhau (𝑣, 𝑡), trong đó 𝑣 ⊂ 𝑉 biểu thị tập hợp các phần tử nguyên tử xuất hiện trong khoảng hiển thị và 𝑡 biểu thị cây con nhỏ nhất của cây bố cục tài liệu 𝑇 chứa tất cả các phần tử nguyên tử trong 𝑣.

* 1. Tạo khoảng thời gian trực quan ứng viên

Artemis dựa vào một số chức năng được giám sát yếu để giới thiệu kiến ​​thức về miền cụ thể nhằm xác định các khoảng hình ảnh ứng viên nơi một thực thể được đặt tên có thể xuất hiện. Chúng tôi triển khai các hàm này dưới dạng các hàm Python chấp nhận một thực thể được đặt tên, danh sách các đề cập và vị trí chúng xuất hiện trong kho dữ liệu đào tạo và tài liệu kiểm tra 𝐷. Mỗi hàm trả về một tập hợp các khoảng hình ảnh ứng viên trong 𝐷. Các hàm này có thể bao gồm từ các biểu thức chính quy đơn giản, các thư viện có sẵn công khai [13, 14] đến các hàm phức tạp tính đến các tín hiệu trên cả phương thức trực quan và văn bản. Chúng tôi cung cấp ví dụ về một số hàm được sử dụng trong các thử nghiệm của chúng tôi bên dưới.