A Span Extraction Approach for Information Extraction on Visually-Rich Documents

1. Abstract ( tóm tắt)

IE for VRD: đã đạt được hiệu suất SOTA gần đây nhờ sự thích ứng của các mô hình ngôn ngữ dựa trên Transformer, điều này cho thấy tiềm năng to lớn của các phương pháp đào tạo trước. Trong bài báo này, chúng tôi trình bày một cách tiếp cận mới để cải thiện khả năng đào tạo trước mô hình ngôn ngữ trên VRD. Đầu tiên, chúng tôi giới thiệu một mô hình IE dựa trên truy vấn mới sử dụng tính năng trích xuất khoảng thay vì sử dụng phương pháp ghi nhãn trình tự phổ biến. Thứ hai, để mở rộng công thức trích xuất nhịp, chúng tôi đề xuất một nhiệm vụ đào tạo mới tập trung vào việc mô hình hóa các mối quan hệ giữa các thực thể ngữ nghĩa trong một tài liệu. Tác vụ này cho phép các nhịp mục tiêu được trích xuất đệ quy và có thể được sử dụng để đào tạo trước mô hình hoặc như một tác vụ xuôi dòng của IE. Đánh giá trên ba tập dữ liệu của các tài liệu kinh doanh phổ biến (hóa đơn, biên lai) cho thấy rằng phương pháp đề xuất của chúng tôi đạt được những cải tiến đáng kể so với các mô hình hiện có. Phương pháp này cũng cung cấp một cơ chế để tích lũy kiến ​​thức từ nhiều tác vụ IE.

1. Giới thiệu

Trích xuất thông tin (IE) cho các tài liệu trực quan (VRD) khác với các tài liệu văn bản thuần túy ở nhiều khía cạnh. VRD chứa nội dung văn bản thưa thớt hơn, có cấu trúc hình ảnh được xác định chặt chẽ và bố cục phức tạp không có trong tài liệu văn bản thuần túy. Do những đặc điểm này, các tác phẩm hiện có trên IE cho VRD hầu hết sử dụng kỹ thuật Computer Vision [1] và / hoặc dựa trên đồ thị [9,6,11]. Gần đây, BERT cung cấp nhiều quan điểm hơn về VRD từ quan điểm Xử lý ngôn ngữ tự nhiên [13,12,14,4,3,15]

LayoutLM [13] nổi lên như một cách tiếp cận đơn giản nhưng tiềm năng, trong đó kiến ​​trúc BERT được mở rộng bằng cách bổ sung nhúng vị trí 2D và nhúng trực quan vào nhúng đầu vào của nó [2]. Mô hình được đào tạo bởi mục tiêu Tạo mô hình ngôn ngữ hình ảnh có mặt nạ (MVLM) [13], đây cũng là sự điều chỉnh của nhúng vị trí và nhúng trực quan vào mục tiêu Mô hình hóa ngôn ngữ được che giấu [2] của BERT. Gần đây hơn, LayoutLMv2 [12] đã được giới thiệu tập trung hơn vào việc tận dụng tốt hơn thông tin vị trí và hình ảnh 2D. Ngoài ra còn có thêm hai mục tiêu đào tạo trước mới được thiết kế để mô hình hóa mối tương quan giữa hình ảnh và văn bản.

Mặc dù LayoutLM rất hiệu quả để lập mô hình dữ liệu đa phương thức như trực quan phong phú tài liệu, nó mang một vài hạn chế về tính thực tiễn. Đầu tiên, đào tạo trước một mô hình ngôn ngữ yêu cầu một lượng lớn dữ liệu và không giống như các tài liệu văn bản thuần túy,dữ liệu tài liệu trực quan phong phú không có sẵn với số lượng lớn như vậy,

đặc biệt là đối với các ngôn ngữ ít tài nguyên (ví dụ: tiếng Nhật). Thứ hai, các phương pháp trước đây

chủ yếu sử dụng nhãn trình tự cho IE, nhưng chúng tôi tranh luận rằng phương pháp này có thể

không hoạt động tốt đối với các đơn vị ngắn hạn do sự mất cân bằng của các lớp học và đào tạo

hàm mất mát [10,5]. Thứ ba, ghi nhãn trình tự yêu cầu một tập hợp cố định và được xác định trước

của các thẻ / lớp thực thể cho mỗi tập dữ liệu, điều này cản trở việc áp dụng

cùng một mô hình IE cho nhiều bộ dữ liệu

Để giải quyết các vấn đề nói trên, chúng tôi đề xuất một phương pháp mới sử dụng khai thác khoảng thay vì sử dụng nhãn trình tự. Phương pháp kết hợp

Trích xuất nhịp dựa trên QA và trích xuất đa giá trị bằng cách sử dụng đệ quy mới

lược đồ dự đoán quan hệ. Chúng tôi cũng giới thiệu một quy trình đào tạo trước để mở rộng LayoutLM sang cả bộ dữ liệu tiếng Anh và không phải tiếng Anh bằng cách sử dụng ngữ nghĩa các thực thể. Vì truy vấn được nhúng độc lập với ngữ cảnh nên phương pháp này có thể được áp dụng trên nhiều bộ dữ liệu, do đó cho phép tích lũy kiến thức liền mạch thông qua một công thức chung.

1. IE as Span Extraction

Cách tiếp cận phổ biến nhất để trích xuất thông tin từ các tài liệu giàu hình ảnh là ghi nhãn theo trình tự gán nhãn cho mỗi mã thông báo của tài liệu [13,12,14,4,3,15]. Mặc dù giải pháp này cho kết quả đầy hứa hẹn, chúng tôi tranh luận rằng nhãn trình tự có thể không hoạt động tốt khi một số loại thực thể có chỉ một lượng nhỏ mẫu hoặc khi các thực thể mục tiêu có khoảng cách ngắn văn bản trong các tài liệu dày đặc. Lý do có thể là sự mất cân bằng của các nhãn khi đào tạo các mô hình IE. Hơn nữa, ghi nhãn trình tự yêu cầu một định nghĩa rõ ràng của các lớp mã thông báo, làm giảm khả năng sử dụng lại của mô hình trên các tác vụ khác.

Để vượt qua thách thức này, chúng tôi thực hiện theo cách tiếp cận chiết tách khoảng, còn được gọi là Trả lời câu hỏi chiết xuất (QA) trong NLP [2]. Khác với việc gắn nhãn theo trình tự gán nhãn cho mọi mã thông báo dựa trên việc nhúng của chúng, các mô hình QA dự đoán vị trí bắt đầu và vị trí kết thúc của câu trả lời tương ứng cho mỗi truy vấn. Đối với vấn đề IE, một truy vấn đại diện cho một trường / thẻ bắt buộc mà chúng tôi muốn trích xuất từ ngữ cảnh tài liệu gốc. Mỗi truy vấn được biểu diễn dưới dạng một vectơ nhúng có thể học được và được giữ riêng biệt với bộ mã hóa mô hình ngôn ngữ.

Để tính toán tổn thất khai thác khoảng thời gian, hoạt động sof tmax được áp dụng trên tất cả các mã thông báo trong ngữ cảnh thay vì trên các lớp mã thông báo cho mỗi mã thông báo như trong gắn thẻ theo trình tự. Sự mất mát softmax trên tất cả các mã thông báo trình tự loại bỏ vấn đề mất cân bằng lớp đối với các thực thể trả lời ngắn hơn. Sự tách biệt giữa nhúng câu hỏi (truy vấn) và nhúng ngữ cảnh cũng mở ra tiềm năng tích lũy và sử dụng lại thông tin trên các tập dữ liệu khác nhau.

1. Pre-trainig Objectives with Recursive Span Extraction
   1. LayoutLM for Low-resource Languages

Phần này mô tả một số phương pháp hiệu quả để chuyển LayoutLM sang các ngôn ngữ tài nguyên thấp, ví dụ: Tiếng Nhật. Đào tạo trước một mô hình ngôn ngữ từ đầu với mục tiêu MLM thường yêu cầu hàng triệu dữ liệu và có thể mất nhiều thời gian để đào tạo. Mặc dù lượng dữ liệu như vậy có thể được lấy ở dạng văn bản thuần túy, nhưng dữ liệu trực quan phong phú không tự nhiên tồn tại với số lượng lớn do thiếu chú thích OCR cấp độ từ. Vì các trọng số được đào tạo trước chỉ có sẵn cho tiếng Anh, nên việc áp dụng LayoutLM cho các ngôn ngữ khác vẫn còn là một câu hỏi mở. Để khắc phục điều này, chúng tôi đề xuất một quy trình đào tạo trước được chuyển giao đơn giản cho LayoutLM trước tiên tận dụng các trọng số được đào tạo trước có sẵn của BERT và sau đó chuyển biểu diễn theo ngữ cảnh của nó sang LayoutLM.

Overcoming data shortage: Khắc phục tình trạng thiếu dữ liệu. Để chuyển mô hình LayoutLM từ tiếng Anh (ngôn ngữ nguồn) sang ngôn ngữ khác (ngôn ngữ đích), chúng ta thực hiện theo quy trình sau. Đầu tiên, lớp nhúng từ của mô hình được khởi tạo từ BERT được đào tạo trước của ngôn ngữ đích và các lớp còn lại (bao gồm các lớp nhúng vị trí, lớp bộ mã hóa và đầu MVLM) được khởi tạo từ LayoutLM có sẵn công khai cho tiếng Anh. Sau đó, mô hình được đào tạo với mục tiêu MVLM sử dụng tập dữ liệu nhỏ hơn nhiều (khoảng 17.000 mẫu) với 100 kỷ nguyên. Mô hình cuối cùng có thể được tinh chỉnh cho các tác vụ xuôi dòng như ghi nhãn trình tự, phân loại tài liệu hoặc trả lời câu hỏi.

Lớp nhúng từ ánh xạ các mã thông báo văn bản thuần túy vào một không gian tiềm ẩn có ý nghĩa về mặt ngữ nghĩa, trong khi không gian tiềm ẩn có thể tương tự, bản thân quá trình ánh xạ hoàn toàn khác biệt với ngôn ngữ. Chúng tôi lập luận rằng thông tin không gian không phụ thuộc vào ngôn ngữ và có thể hữu ích giữa các ngôn ngữ có thứ tự đọc như nhau. Tương tự, các lớp mã hóa được lấy từ các trọng số của LayoutLM để tận dụng khả năng thu hút sự chú ý của nó từ cả đầu vào ngữ nghĩa và vị trí. Các thử nghiệm của chúng tôi trên tập dữ liệu nội bộ cho thấy rằng cài đặt đào tạo trước được chuyển giao này mang lại kết quả tốt hơn đáng kể so với việc không sử dụng trọng số nhúng từ BERT. Ngoài ra, tập dữ liệu của chúng tôi với 17.000 mẫu không đủ để đào tạo trước LayoutLM từ đầu, điều này sẽ đưa ra một so sánh không công bằng giữa việc sử dụng quy trình đào tạo trước được chuyển giao được đề xuất của chúng tôi và đào tạo từ đầu. Do đó, chúng tôi dự định xác minh cách tiếp cận của mình trên tập dữ liệu XFUN được giới thiệu gần đây [14].

Overcoming inadequate annotations: Khắc phục những chú thích không đầy đủ. VRD thường tồn tại ở định dạng hình ảnh mà không có bất kỳ chú thích OCR nào. Do đó, để giải quyết tình trạng thiếu chú thích OCR cấp độ từ, chúng tôi đã sử dụng trình đọc OCR nội bộ để trích xuất chú thích cấp độ dòng. Sau đó, chú thích cấp độ từ được ước lượng bằng cách chia hộp giới hạn của mỗi dòng tương ứng với độ dài của các từ trong dòng đó.

* 1. Span Extraction Pre-training

Để giải quyết hạn chế của yêu cầu dữ liệu đào tạo khi đào tạo trước mô hình LayoutLM, chúng tôi trình bày một mục tiêu đào tạo trước mới sử dụng lại các chú thích của các thực thể hiện có từ các tác vụ IE hạ lưu. Nhiệm vụ đào tạo trước của chúng tôi dựa trên công thức trích xuất nhịp đệ quy mở rộng phương pháp QA truyền thống để trích xuất nhiều nhịp câu trả lời từ một truy vấn duy nhất.

Recursive Span Extraction Khai thác khoảng đệ quy. Trái ngược với gắn nhãn trình tự, công thức trích xuất khoảng cách của chúng tôi cho phép truy vấn độc lập linh hoạt (trường) với ngữ cảnh, cho phép tích lũy kiến ​​thức từ nhiều tập dữ liệu bằng cách liên tục cập nhật bộ sưu tập nhúng câu hỏi. Tuy nhiên, một hạn chế của mô hình trích xuất nhịp được mô tả trong Hình 1 là mô hình không thể trích xuất nhiều hơn một câu trả lời (thực thể) cho mỗi truy vấn, điều này được cho là rất quan trọng trong nhiều trường hợp sử dụng IE. Do đó, chúng tôi đề xuất một chiến lược để mở rộng khả năng của mô hình bằng cách sử dụng cơ chế giải mã liên kết đệ quy được mô tả trong Hình 2.

Hình 3 Dự đoán liên kết đệ quy được chứng minh trên một ví dụ nhận [8]. (mũi tên xanh lục và xanh lam biểu thị vị trí bắt đầu và kết thúc mục tiêu của khoảng trả lời tương ứng với mã thông báo truy vấn (hộp giới hạn màu xanh lá cây). Mã thông báo truy vấn là mã thông báo đầu tiên bắt đầu mỗi phân đoạn văn bản.

Khoảng câu trả lời đầu tiên được trích xuất bởi biểu diễn vectơ T của truy vấn và biểu diễn ẩn H của ngữ cảnh (Phương trình 2). Sau đó, biểu diễn ẩn tương ứng với mã thông báo bắt đầu của câu trả lời thứ i được sử dụng làm vectơ truy vấn để trích xuất đệ quy câu trả lời (i + 1) -th (Phương trình 3). Quy trình giải mã đệ quy này dừng nếu si = ei = 0 hoặc khi si = sj, ei = ej với j <i. Bằng cách sử dụng phương pháp này, chúng tôi có thể giải mã tất cả các khoảng câu trả lời thuộc cùng một thẻ truy vấn / thực thể trong ngữ cảnh tài liệu

Pre-training with semantic entities : Đào tạo trước với các thực thể ngữ nghĩa. Nhiệm vụ trích xuất nhịp đệ quy tạo thành một cấu trúc quan hệ chung giữa các thực thể, điều này củng cố ngôn ngữ khả năng mô hình hóa thông tin ngữ cảnh và không gian trong các tài liệu trực quan phong phú. Mô hình phải suy ra các liên kết liên tục giữa các mục trong cùng một trường thường được biểu diễn bằng các liên kết dọc hoặc ngang theo cấu trúc bảng. Có thể xem hình ảnh trực quan mẫu của tài liệu và công thức liên kết liên tục tương ứng trong Hình 3. Bằng cách buộc mô hình xây dựng lại mối quan hệ giữa các thực thể liên tục thuộc cùng một truy vấn, chúng tôi giả thuyết rằng mô hình có thể tạo ra hiệu suất tốt hơn khi ổn -tuned trên các tác vụ IE xuôi dòng. Kết quả thử nghiệm trong Phần 4 xác thực giả thuyết của chúng tôi. Kiến trúc dựa trên truy vấn được đề xuất và giải mã khoảng cách đệ quy cho phép một công thức IE chung độc lập với các thẻ truy vấn cụ thể của từng loại tài liệu. Nhớ lại rằng nhúng truy vấn được xây dựng riêng biệt với mô hình chính được đào tạo trước, bộ mã hóa LayoutLM sẽ tập trung vào việc mô hình hóa ngữ cảnh tài liệu và tương tác giữa các thực thể của nó. Chúng tôi có thể đào tạo trước mô hình LayoutLM một cách liền mạch trên nhiều bộ dữ liệu IE với các trường khóa-giá trị khác nhau. Vì vậy, chúng tôi đặt tên mục tiêu đào tạo trước này là đào tạo trước ngữ nghĩa trích xuất Span, cho phép chúng tôi đào tạo trước LayoutLM với một số tập dữ liệu có kích thước nhỏ được IE chú thích bằng các ngôn ngữ tài nguyên thấp, thay vì sử dụng dữ liệu tài liệu quy mô lớn như [ 12].

Training and inference: Mô hình LayoutLM dựa trên tiếng Anh bắt đầu từ trọng lượng được đào tạo trước của [13] được cung cấp cùng với bài báo, được gọi là ENLayoutLM-base. Chúng tôi đã sử dụng phiên bản LayoutLM-base từ giấy gốc (thông số 113M) do giới hạn phần cứng. Một mô hình dựa trên Nhật Bản khác (JP-LayoutLM-base) đã được đào tạo trước bằng cách sử dụng quy trình được mô tả trong Phần 3.1 với cấu hình tương tự như LayoutLM-base, sử dụng tác vụ MVLM trên dữ liệu nội bộ của chúng tôi. Mô hình này được thiết kế để sử dụng với các bộ dữ liệu của Nhật Bản. Để tiến hành các thử nghiệm của chúng tôi, trước tiên chúng tôi đào tạo trước mô hình cơ sở JP-LayoutLM trên tập dữ liệu BizDocs của Nhật Bản đào tạo trước với các mục tiêu khai thác khoảng cách được đề xuất. Sau đó, chúng tôi thực hiện tinh chỉnh trên hai bộ dữ liệu IE: hóa đơn Nhật Bản và CORD với trọng lượng được đào tạo trước tương ứng là JP-LayoutLM-base và EN-LayoutLMbase. Chúng tôi so sánh kết quả của việc khai thác khoảng cách với việc ghi nhãn trình tự trên các tác vụ IE hạ lưu này. Ngoài ra, chúng tôi đo lường kết quả tinh chỉnh với mô hình cơ sở LayoutLM mà không cần đào tạo trước dựa trên nhịp để chứng minh hiệu quả của phương pháp được đề xuất. Đối với chi tiết siêu tham số, hai mô hình JP-LayoutLM-base và ENLayoutLM-base có chung đường trục với kích thước ẩn là 768. Trong mô hình Nhật Bản, chúng tôi áp dụng lớp nhúng và từ vựng từ BERT cl- được đào tạo trước của Nhật Bản. tohoku 3 chứa 32.000 từ phụ. Chúng tôi cũng đặt độ dài chuỗi đầu vào là 512 trong tất cả các mô hình đã thử nghiệm. Tỷ lệ học tập và phương pháp tối ưu hóa được giữ nguyên lần lượt là 5e - 5 và trình tối ưu hóa Adam [13]. Về chức năng mất mát, với định dạng QA, tổn thất entropy chéo (CE) tiêu chuẩn được sử dụng cho các chỉ số đầu và cuối từ tất cả các mã thông báo tài liệu. Với mô hình gắn thẻ trình tự, tổn thất CE được sử dụng cho toàn bộ đầu ra phân loại trình tự của tài liệu.

Các thước đo đánh giá. Chúng tôi sử dụng điểm F1 cấp thực thể phổ biến (vi mô) để đánh giá, được sử dụng phổ biến trong các nghiên cứu trước đây [13,12,4,3]. Thực thể 3 https://github.com/cl-tohoku/bert-japanese 8 Tuan-Anh et. al. Điểm F1 (macro) được sử dụng làm số liệu thứ hai để biểu thị kết quả tổng thể trên tất cả các trường khóa-giá trị, vì nó nhấn mạnh hiệu suất mô hình trên một số trường / thẻ hiếm. Tình trạng này thường xuất hiện trong các cơ sở công nghiệp, nơi việc thu thập dữ liệu rất tốn kém đối với một số lĩnh vực không thường xuyên.

* 1. Kết quả và thảo luận

CORD. Chúng tôi báo cáo kết quả về thử nghiệm và phát triển các tập hợp con của CORD trong Bảng 3 và Bảng 2 tương ứng. Phương pháp khai thác khoảng cải thiện một cách nhất quán hiệu suất của IE so với phương pháp ghi nhãn trình tự hiện có. Chúng ta có thể quan sát thấy sự cải thiện 1% - 3% trong cả hai chỉ số, đây là kết quả đáng kể dựa trên kết quả cơ bản cạnh tranh của mô hình LayoutLM mặc định. Thật thú vị, so với kết quả từ LayoutLMv2 [12], mô hình của chúng tôi hoạt động tốt hơn cả LayoutLM-large và LayoutLM-base-v2, mặc dù chúng tôi chỉ sử dụng phiên bản nhẹ hơn của mô hình. Điều này chứng tỏ hiệu quả của công thức trích xuất nhịp của chúng tôi cho các tác vụ IE hạ lưu. Lưu ý rằng công thức trích xuất nhịp độc lập với kiến ​​trúc mô hình, do đó chúng ta có thể mong đợi sự cải thiện tương tự khi bắt đầu từ các mô hình phức tạp hơn được đào tạo trước. Chúng tôi để thử nghiệm này là công việc trong tương lai của chúng tôi. Một phát hiện khác từ kết quả là macro điểm F1 được tăng lên một lượng đáng kể cùng với điểm cấp thực thể (vi mô) trong cả hai tập hợp con. Điều này minh họa rằng phương pháp khai thác theo khoảng cách của chúng tôi giúp cải thiện hiệu suất trên tất cả các trường / thẻ, không chỉ các thẻ thông thường có nhiều chú thích của thực thể hơn.

Visualization. Đầu ra của mô hình trên tập dữ liệu CORD (thử nghiệm) có thể được hình dung tại Hình 4. Mô hình có thể suy diễn thành công các mối quan hệ không gian giữa các phần tử tài liệu khác nhau để tạo thành đầu ra cuối cùng của IE. Chúng ta có thể quan sát một số cấu trúc không đều của bảng (do điều kiện máy ảnh chụp) và có thể xử lý chính xác. Nó thể hiện khả năng của mô hình để hiểu cấu trúc bố cục chung của các tài liệu đầu vào.

Hình dung. Đầu ra của mô hình trên tập dữ liệu CORD (thử nghiệm) có thể được hình dung tại Hình 4. Mô hình có thể suy diễn thành công các mối quan hệ không gian giữa các phần tử tài liệu khác nhau để tạo thành đầu ra cuối cùng của IE. Chúng ta có thể quan sát một số cấu trúc không đều của bảng (do điều kiện máy ảnh chụp) và có thể xử lý chính xác. Nó thể hiện khả năng của mô hình để hiểu cấu trúc bố cục chung của các tài liệu đầu vào.

1. Kết luận

Bài báo này giới thiệu một công thức trích xuất nhịp mới cho nhiệm vụ trích xuất thông tin trên các tài liệu trực quan. Một chương trình đào tạo trước dựa trên chú thích các thực thể ngữ nghĩa từ dữ liệu IE cũng được trình bày. Phương pháp được đề xuất cho thấy kết quả đầy hứa hẹn trên hai bộ dữ liệu IE của các tài liệu kinh doanh. Phương pháp của chúng tôi luôn đạt được cải thiện 1% -3% so với phương pháp gắn nhãn trình tự, ngay cả với đường trục LayoutLM nhỏ hơn. Hơn nữa, nó có thể được áp dụng cho hầu hết các mô hình được đào tạo trước hiện có cho IE và mở ra một hướng mới để liên tục cải thiện hiệu suất của các mô hình được đào tạo trước thông qua tinh chỉnh các tác vụ xuôi dòng. Đối với công việc trong tương lai, chúng tôi muốn xem xét việc tích hợp cơ chế trích xuất nhịp của chúng tôi với mô hình LayoutLMv2 mới nhất. Chúng tôi cũng có kế hoạch mở rộng nhiệm vụ đào tạo trước mô hình với các quan hệ tài liệu đa dạng hơn như: cấu trúc bảng, quan hệ tiêu đề-đoạn văn để thúc đẩy hơn nữa hiệu suất của LM được đào tạo trước trên VRDs