Nhận dạng thực thể được đặt tên bằng tiếng Bồ Đào Nha bằng BERT-CRF

Fabio Souza 1,3, Rodrigo Nogueira2 , Roberto Lotufo1,3

1Đại học Campinas

f116735@dac.unicamp.br, lotufo@dca.fee.unicamp.br

2Đại học New York

rodrigonogueira@nyu.edu

3NeuralMind Inteligencia Nhân tạo

{fabiosouza, roberto}@neuralmind.ai

trừu tượng

Những tiến bộ gần đây trong biểu diễn ngôn ngữ bằng cách sử dụng mạng thần kinh đã làm cho nó khả thi để chuyển giao các trạng thái bên trong đã học của một đào tạo mô hình cho các tác vụ xử lý ngôn ngữ tự nhiên xuôi dòng, chẳng hạn như nhận dạng thực thể được đặt tên (NER) và trả lời câu hỏi. Nó có đã được chứng minh rằng đòn bẩy của đào tạo trước các mô hình ngôn ngữ cải thiện hiệu suất tổng thể trong nhiều nhiệm vụ và rất có lợi khi dữ liệu được dán nhãn là khan hiếm. Trong công việc này, chúng tôi đào tạo các mô hình BERT của Bồ Đào Nha và triển khai kiến trúc BERT-CRF cho NER nhiệm vụ trên ngôn ngữ Bồ Đào Nha, kết hợp khả năng chuyển giao của BERT với dự đoán có cấu trúc của CRF. chúng tôi khám phá chiến lược đào tạo dựa trên tính năng và tinh chỉnh cho mô hình BERT. Phương pháp tiếp cận ap tinh chỉnh của chúng tôi đạt được kết quả tiên tiến nhất trên bộ dữ liệu HAREM I, cải thiện điểm số F1 bằng 1 điểm trong kịch bản chọn lọc (5 NE các lớp) và bằng 4 điểm trên tổng kịch bản (10 lớp ĐB).

1. Giới thiệu

Nhận dạng đối tượng được đặt tên (NER) là nhiệm vụ xác định các đoạn văn bản có đề cập đến các đối tượng được đặt tên (NE) và phân loại chúng thành các danh mục được xác định trước, chẳng hạn như người, tổ chức, địa điểm, hoặc bất kỳ các lớp quan tâm. Mặc dù đơn giản về mặt khái niệm , NER không phải là một nhiệm vụ dễ dàng. Thể loại của một thực thể được đặt tên phụ thuộc nhiều vào ngữ nghĩa văn bản và ngữ cảnh xung quanh nó. Hơn nữa, có có nhiều định nghĩa về thực thể được đặt tên và tiêu chí đánh giá, giới thiệu các biến chứng đánh giá (Marrero và cộng sự, 2013).

Các hệ thống NER tiên tiến nhất hiện nay sử dụng cấu trúc thần kinh đã được đào tạo trước về nhiệm vụ mô hình hóa ngôn ngữ. Ví dụ về các mô hình như vậy là ELMo (Peters và cộng sự, 2018), OpenAI GPT (Rad ford và cộng sự, 2018), BERT (Devlin và cộng sự, 2018), XL Net (Yang và cộng sự, 2019), RoBERTa (Liu và cộng sự, 2019), Albert (Lan và cộng sự, 2019) và T5 (Raffel và cộng sự, 2019).

Người ta đã chứng minh rằng việc đào tạo trước khi mô hình hóa ngôn ngữ cải thiện đáng kể hiệu suất của nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên và cũng giảm lượng dữ liệu được dán nhãn cần thiết cho việc học tập được giám sát (Howard và Ruder, 2018; Peters và cộng sự, 2018).

Việc áp dụng những kỹ thuật gần đây này vào ngôn ngữ Bồ Đào Nha có thể rất có giá trị, do rằng tài nguyên chú thích là khan hiếm, nhưng không được gắn nhãn dữ liệu văn bản rất phong phú. Trong công việc này, chúng tôi đánh giá một số kiến trúc thần kinh bằng cách sử dụng BERT (Hai chiều Biểu diễn bộ mã hóa từ Transformers) mô hình hóa tác vụ NER bằng tiếng Bồ Đào Nha và so sánh chiến lược đào tạo dựa trên tính năng và tinh chỉnh . Đây là công trình đầu tiên sử dụng mô hình BERT đến nhiệm vụ NER bằng tiếng Bồ Đào Nha. Chúng tôi cũng thảo luận về các biến chứng chính mà chúng ta gặp phải khi sử dụng bộ dữ liệu. Với ý nghĩ đó, chúng tôi mong muốn tạo điều kiện thuận lợi cho khả năng tái sản xuất của công việc này bằng cách cung cấp công khai các mô hình và triển khai của chúng tôi.1 2

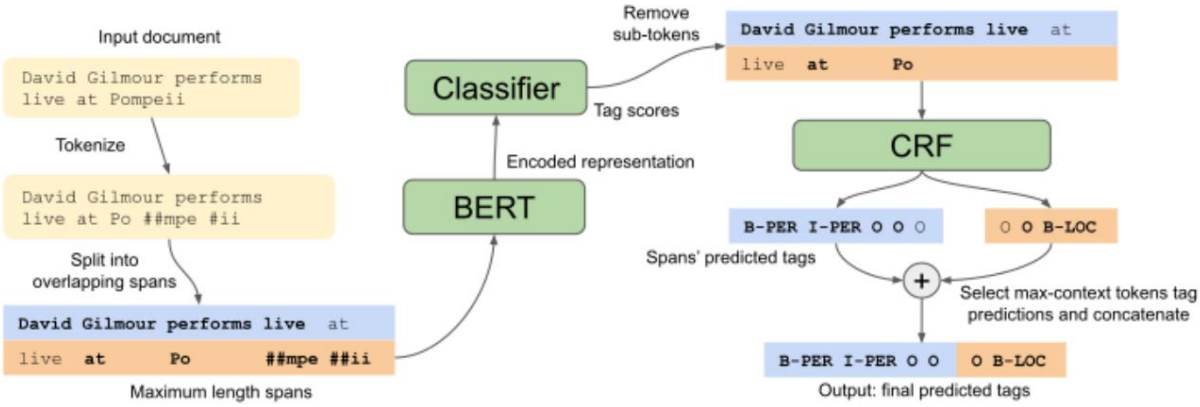
2 công việc liên quan

Các hệ thống NER có thể dựa trên các quy tắc thủ công hoặc các phương pháp học máy. Đối với người Bồ Đào Nha ngôn ngữ, các công trình trước đây đã khám phá các kỹ thuật máy học và một số công trình áp dụng các mô hình hoạt động của mạng nơ-ron. do Amaral và Vieira (2014) đã tạo ra một mô hình CRF sử dụng 15 tính năng được trích xuất từ các từ trung tâm và xung quanh. (Pirovani và Oliveira, 2018) đã kết hợp mô hình CRF với Local Ngữ pháp, theo cách tiếp cận tương tự.

Bắt đầu với Collobert et al. (2011), các hệ thống NER làm việc theo mạng nơ-ron đã trở nên phổ biến do

1Mã sẽ là có sẵn tại https://gist.github.com/fabiocapsouza/62c98576d1c826894be2b3ae0993ef53.

Các mô hình 2BERT có sẵn tại https://github.com/thần kinh-ai/tiếng Bồ Đào Nha-bert.



Hình 1: Minh họa phương pháp đề xuất. Với một tài liệu đầu vào, văn bản được mã hóa bằng WordPiece (Wu et al., 2016) và tài liệu được mã hóa được chia thành các khoảng chồng chéo có độ dài tối đa bằng cách sử dụng một bước xác định (ví dụ với một bước là 3). Mã thông báo ngữ cảnh tối đa của mỗi khoảng được đánh dấu in đậm. Các khoảng được đưa vào BERT và sau đó vào lớp phân loại, tạo ra một chuỗi điểm thẻ cho mỗi khoảng. Các mục nhập mã thông báo phụ (bắt đầu bằng ##) được xóa khỏi các khoảng và các mã thông báo còn lại được chuyển đến lớp CRF. Các mã bối cảnh tối đa được chọn và nối để tạo thành các thẻ dự đoán cuối cùng.

yêu cầu kỹ thuật tính năng tối thiểu, góp phần tăng tính độc lập của miền (Ya dav và Bethard, 2018). Mô hình CharWNN (Santos và Guimaraes, 2015) đã mở rộng công việc của Collobert et al. (2011) bằng cách sử dụng lớp chập để trích xuất các đặc trưng cấp độ ký tự từ mỗi từ. Các tính năng này được kết hợp với

nhúng từ được đào tạo trước và sau đó được sử dụng để thực hiện phân loại tuần tự.

Mô hình CharWNN (Santos và Guimaraes, 2015) đã mở rộng công việc của Collobert et al. (2011) bằng cách sử dụng lớp tích chập để trích xuất các tính năng cấp độ ký tự từ mỗi từ. Các

Kiến trúc LSTM-CRF (Lample et al., 2016) đã được sử dụng phổ biến trong tác vụ NER (Castro et al., 2018; de Araujo et al., 2018; Fernandes et al., 2018). Mô hình bao gồm hai mạng LSTM hai chiều trích xuất và kết hợp

các tính năng cấp độ ký tự và cấp độ từ. Sau đó, lớp CRF sẽ thực hiện phân loại tuần tự .

Các công trình gần đây đã khám phá các nhúng theo ngữ cảnh được trích xuất từ các mô hình ngôn ngữ kết hợp với kiến trúc LSTM-CRF. Santos và cộng sự. (2019b,a) sử dụng Flair Embeddings (Akbik et al., 2018) để trích xuất các từ nhúng theo ngữ cảnh từ một LM cấp độ ký tự hai chiều được đào tạo trên

kho ngữ liệu tiếng Bồ Đào Nha. Các phần nhúng này được nối với các phần nhúng từ được đào tạo trước và được cung cấp cho mô hình BiLSTM-CRF. Castro et al. (2019) sử dụng các nhúng ELMo là sự kết hợp của các tính năng cấp độ ký tự được trích xuất bởi tích chập mạng thần kinh và các trạng thái ẩn của từng lớp của LM hai chiều (biLM) bao gồm mô hình BiL STM.

3 người mẫu

Trong phần này, chúng tôi mô tả kiến trúc mô hình và các thủ tục đào tạo và đánh giá cho NER.

3.1 BERT-CRF cho NER

Kiến trúc mô hình bao gồm một mô hình BERT với bộ phân loại cấp mã thông báo ở trên cùng, sau đó là CRF chuỗi tuyến tính. Đối với chuỗi đầu vào gồm n mã thông báo, BERT xuất ra chuỗi mã thông báo được mã hóa với kích thước ẩn H. Mô hình phân loại chiếu biểu diễn được mã hóa của mỗi mã thông báo không gian thẻ ^{h} thành R^{K}, trong đó K là số, tức là R thẻ và phụ thuộc vào số lượng lớp và sơ đồ gắn thẻ. Điểm số đầu ra P_{i1}^{K} của mô hình phân loại sau đó được đưa đến lớp CRF, có tham số là ma trận chuyển tiếp thẻ $A \in R^{K+2\times K+2}$. Ma trận A sao cho $A_{i,j}$ đại diện cho điểm chuyển đổi từ thẻ i sang thẻ j. A bao gồm 2 trạng thái bổ sung: bắt đầu và kết thúc chuỗi.

Theo mô tả của Lample et al. (2016), đối với chuỗi đầu vào X = (x1, ..., xn) và chuỗi dự đoán thẻ y = (y1, ..., yn), $y_i \in \{1, ..., K\}$, thì điểm của chuỗi được định nghĩa là

$$s(X, y) = \sum_{i=0}^{N} A_{i, y_{i+1}} + \sum_{i=1}^{N} P_{i, y_i},$$

trong đó y0 và yn+1 là thẻ bắt đầu và thẻ kết thúc. Mô hình được đào tạo để tối đa hóa xác suất đăng nhập của chuỗi thẻ chính xác:

$$\log(p(y|X)) = s(X, y) \quad \log \sum_{\tilde{y} \in YX} e^{s(X,\tilde{y})} \tag{1}$$

trong đó YX là tất cả các chuỗi thẻ có thể. Tổng kết trong phương trình. 1 được tính bằng lập trình động. Trong quá trình đánh giá, trình tự có khả năng nhất thu được bằng cách giải mã Viterbi. Theo ing Devlin et al. (2018), chúng tôi chỉ tính toán các dự đoán và tổn thất cho mã thông báo phụ đầu tiên của mỗi mã thông báo.

3.2 Phương pháp tiếp cận dựa trên tính năng và Tinh chỉnh

Chúng tôi thử nghiệm hai cách tiếp cận ap học chuyển tiếp: dựa trên tính năng và tinh chỉnh. Đối với cách tiếp cận dựa trên tính năng, các trọng số của mô hình BERT được giữ cố định và chỉ có mô hình phân loại và lớp CRF được đào tạo. Mô hình phân loại bao gồm BiLSTM 1 lớp với dLSTM kích thước ẩn theo sau bởi lớp Tuyến tính. Thay vì chỉ sử dụng lớp đại diện ẩn cuối cùng của BERT, chúng tôi tính tổng 4 lớp cuối cùng, theo Devlin et al. (2018).

Kiến trúc kết quả giống với mô hình LSTM CRF, Lample et al. (2016) nhưng với nhúng BERT.

Đối với phương pháp tinh chỉnh, bộ phân loại là một lớp tuyến tính và tất cả các trọng số, bao gồm cả BERT, được cập nhật chung trong quá trình đào tạo. Đối với cả hai phương pháp ap, các mô hình không có lớp CRF cũng được đánh giá. Trong trường hợp này, chúng được tối ưu hóa bằng cách giảm thiểu tối đa tổn thất entropy chéo.

3.3 Ngữ cảnh tài liệu và ngữ cảnh tối đa sự đánh giá

Để tận dụng ngữ cảnh dài hơn khi đặt biểu diễn mã thông báo từ BERT, chúng tôi sử dụng ngữ cảnh tài liệu cho các ví dụ đầu vào thay vì ngữ cảnh câu. Theo cách tiếp cận của Devlin et al. (2018) trên tập dữ liệu SQuAD, các bài kiểm tra dài hơn S mã thông báo được chia thành các khoảng thời gian có độ dài lên đến S bằng cách sử dụng một dải D mã thông báo. Mỗi khoảng được sử dụng như một ví dụ riêng biệt trong quá trình đào tạo.

Tuy nhiên, trong quá trình đánh giá, một mã thông báo Ti có thể xuất hiện trong $\overline{\text{ĐN}}^{S}$ = nhiều nhịp sj và do đó có thể có tối đa N dự đoán thẻ riêng biệt yi,j . Mỗi dự đoán cuối cùng của ken được lấy từ khoảng mà mã thông báo gần vị trí trung tâm hơn, nghĩa là khoảng mà nó có nhiều thông tin theo ngữ cảnh nhất. Hình 1 minh họa quy trình đánh giá.

4 thí nghiệm

Trong phần này, chúng tôi trình bày các thiết lập thử nghiệm cho đào tạo trước BERT và đào tạo NER. Chúng tôi trình bày các bộ dữ liệu được sử dụng, các thiết lập đào tạo và siêu tham số.

4.1 Đào tạo trước BERT

Chúng tôi đào tạo các mô hình BERT của Bồ Đào Nha cho hai kích thước mô hình được xác định trong Devlin et al. (2018): BERT Cơ sở và BERT Lớn. Độ dài câu tối đa được đặt thành S = 512 mã thông báo. Chúng tôi chỉ đào tạo các mô hình theo trường hợp vì cách viết hoa có liên quan đến NER (Castro et al., 2018).

4.1.1 Tạo từ vựng Một từ vựng tiếng

Bồ Đào Nha có vỏ bọc gồm 30 nghìn đơn vị từ phụ được tạo bằng cách sử dụng SentencePiece (Kudo và Richardson, 2018) với thuật toán BPE và 200 nghìn bài viết ngẫu nhiên trên Wikipedia tiếng Bồ Đào Nha, sau đó được chuyển đổi sang định dạng WordPiece. Chi tiết về chuyển đổi Câu văn thành Đoạn văn có thể tìm thấy trong Phụ lục A.1.

4.1.2 Dữ liệu trước khi đào

tạo Đối với dữ liệu trước khi đào tạo, chúng tôi sử dụng kho văn bản brWaC (Wagner Filho và cộng sự, 2018), chứa 2,68 tỷ mã thông báo từ 3,53 triệu tài liệu và được kho ngữ liệu tiếng Bồ Đào Nha mở lớn nhất cho đến nay. Ngoài kích thước của nó, brWaC bao gồm toàn bộ tài liệu và phương pháp của nó đảm bảo tính đa dạng miền và chất lượng nội dung cao, đây là những tính năng mong muốn cho đào tạo trước BERT.

Chúng tôi chỉ sử dụng nội dung tài liệu (bỏ qua tiêu đề) và chúng tôi áp dụng một bước xử lý hậu kỳ duy nhất trên dữ liệu để xóa mojibakes3 và các thẻ HTML còn sót lại bằng thư viện ftfy (Speer, 2019). Kho dữ liệu được xử lý cuối cùng có 17,5 GB văn bản thô.

4.1.3 Thiết lập trước khi đào

tạo Các chuỗi đầu vào trước khi đào tạo được tạo với các tham số mặc định và sử dụng mặt nạ toàn bộ công việc (nếu một từ bao gồm nhiều đơn vị từ con bị che khuất, tất cả các đơn vị từ con của nó sẽ bị che khuất và phải được dự đoán bằng Ngôn ngữ che dấu Nhiệm vụ mô hình hóa). Các mô hình được đào tạo trong 1.000.000 bước. Chúng tôi sử dụng tỷ lệ học tập là 1e-4, tỷ lệ học tập

3Mojibake là một loại lỗi văn bản xảy ra khi các chuỗi được giải mã bằng mã hóa ký tự không chính xác. Đối với A˜ §A˜ £o", từ "codificac¸ao"˜ trở thành "codifica ví dụ khi được mã hóa bằng UTF-8 và được giải mã bằng ISO-8859-1.

khởi động trong 10.000 bước đầu tiên, sau đó là sự phân rã tuyến tính của tốc độ học tập.

Đối với các mô hình BERT Base, các trọng số được khởi tạo với điểm kiểm tra của BERT Base đa ngôn ngữ. Chúng tôi sử dụng kích thước lô là 128 và chuỗi 512 mã thông báo trong toàn bộ quá trình đào tạo. Quá trình đào tạo này mất 4 ngày trên phiên bản TPUv3-8 và thực hiện khoảng 8 kỷ nguyên đối với dữ liệu đào tạo.

Đối với BERT Lớn, các trọng số được khởi tạo với điểm kiểm tra là BERT Lớn bằng tiếng Anh. Vì nó là một mô hình lớn hơn với thời gian đào tạo dài hơn, chúng tôi làm theo hướng dẫn của Devlin et al. (2018) và sử dụng chuỗi 128 mã thông báo theo lô kích thước 256 cho 900.000 bước đầu tiên, sau đó sử dụng chuỗi 512 mã thông báo và lô kích thước 128 cho 100.000 bước cuối cùng. Quá trình đào tạo này mất 7 ngày trên phiên bản TPUv3-8 và thực hiện khoảng 6 kỷ nguyên đối với dữ liệu đào tạo.

Lưu ý rằng khi tính toán số lượng kỷ nguyên, chúng tôi đang xem xét hệ số trùng lặp là 10 khi tạo các bài kiểm tra đầu vào . Điều này có nghĩa là dưới 10 kỷ nguyên, cùng một câu được nhìn thấy với cặp mặt nạ và câu khác nhau trong mỗi kỷ nguyên, hiệu quả tương đương với việc tạo ví dụ động.

4.2 Thí nghiệm NER

4.2.1 Bộ dữ liệu NER

tập dữ liệu	Mã thông báo tài liệu	Thực thể (chọn lọc/tổng
HAREM đầu tiên	129	95585 cộng) 4151 /
miniHAREM	128	64853 5017 3018 / 3642

Bảng 1: Thống kê tập dữ liệu và mã thông báo cho tập đoàn HAREM I. Cột Mã thông báo đề cập đến mã thông báo khoảng trắng và dấu chấm câu. Cột Thực thể bao gồm hai tình huống đã xác định.

Bộ dữ liệu phổ biến để đào tạo và đánh giá NER tiếng Bồ Đào Nha là Bộ sưu tập vàng HAREM (GC) (Santos et al., 2006; Freitas et al., 2010).
Chúng tôi sử dụng các GC của các cuộc thi đánh giá HAREM đầu tiên, được chia thành hai tập hợp con: HAREM đầu tiên và MiniHAREM. Mỗi GC chứa

các thực thể được đặt tên được chú thích thủ công của 10 lớp: Vị trí, Người, Tổ chức, Giá trị, Ngày, Tiêu đề, Điều, Sự kiện, Trừu tượng và Khác.

Theo dõi Santos và Guimaraes (2015) và Castro et al. (2018), chúng tôi sử dụng HAREM đầu tiên làm tập huấn luyện và MiniHAREM làm tập kiểm tra. Các thử nghiệm được tiến hành trên hai kịch bản: kịch bản Chọn lọc, với 5 lớp thực thể (Người, Tổ chức, Vị trí, Giá trị và Ngày) và kịch bản Tổng cộng, xem xét tất cả 10 lớp. bảng 1 chứa một số thống kê tập dữ liệu.

4.2.2 Tiền xử lý HAREM Bộ dữ liệu

HAREM được chú thích có tính đến sự mơ hồ và không xác định trong văn bản, chẳng hạn như sự mơ hồ trong câu. Bằng cách này, một số phân đoạn văn bản chứa các thẻ <ALT> kèm theo nhiều giải pháp nhận dạng thực thể được đặt tên thay thế. Ngoài ra, nhiều danh mục có thể được gán cho một thực thể được đặt tên duy nhất.

Để lập mô hình NER như một vấn đề gắn thẻ trình tự, chúng ta phải chọn một sự thật duy nhất cho từng phân đoạn và/hoặc thực thể không xác định. Để giải quyết từng thẻ <ALT> trong tập dữ liệu, cách tiếp cận của chúng tôi là chọn biến gốc chứa số lượng thực thể được đặt tên cao nhất. Trong trường hợp hòa, cái đầu tiên được chọn. Để giải quyết từng thực thể được đặt tên được gán nhiều lớp, chúng ta chỉ cần chọn lớp hợp lệ đầu tiên cho kịch bản. Tập lệnh tiền xử lý tập dữ liệu có sẵn trên GitHub[4] và Phụ lục A.2 chứa một ví dụ.

4.2.3 Thiết lập thử nghiệm NER Để

đào tạo NER, chúng tôi sử dụng 3 mô hình BERT riêng biệt: BERT-Base đa ngôn ngữ, BERT-Base 5 tiếng Bồ Đào Nha và BERT-Large tiếng Bồ Đào Nha. Chúng tôi sử dụng sơ đồ gắn thẻ IOB2 và một bước D = 128 mã thông báo để chia các ví dụ đầu vào thành các khoảng.

Các tham số mô hình được chia thành hai nhóm với tốc độ học khác nhau: 5e-5 cho mô hình BERT và 1e-3 cho phần còn lại. Số lượng kỷ nguyên là 100 đối với BERT-LSTM, 50 đối với BERT-LSTM-CRF và BERT và 15 đối với BERT-CRF. Số lượng kỷ nguyên được tìm thấy bằng cách sử dụng bộ phát triển được đánh giá cao bằng 10% của bộ đào tạo HAREM đầu tiên. Chúng tôi sử dụng lô kích thước 16 và trình tối ưu hóa Adam tùy chỉnh của Devlin et al. (2018) với mức giảm trọng lượng là 0,01. Tương tự như trước khi đào tạo, chúng tôi sử dụng khởi động tốc độ học tập cho 10% bước đầu tiên và phân rã tuyến tính của tốc độ học tập cho các bước còn lại.

Để giải quyết sự mất cân bằng của lớp, chúng tôi khởi tạo thuật ngữ sai lệch của thẻ “O” trong lớp tuyến tính của bộ phân loại với giá trị là 6 để thúc đẩy sự ổn định tốt hơn trong quá trình đào tạo sớm (Lin et al., 2017). Chúng tôi cũng sử dụng trọng số 0,01 cho các tổn thất thẻ "O" khi không sử dụng lớp CRF.

Đối với cách tiếp cận dựa trên tính năng, chúng tôi sử dụng biLSTM có 1 lớp và kích thước ẩn dLSTM = 100 đơn vị cho mỗi hướng.

[4] https://github.com/fabiocapsouza/harem tiền xử-lý 5Có tại https://github.com/google-research/bert

Ngành kiến trúc	Kịch bản tổng thể			Kịch bản chọn lọc		
	Prec.	Ghi âm	F1	Prec.	Ghi âm	F1
CharWNN (Santos và Guimaraes, 2015)	67,16	63,74	65,41	73,98	68,68	71,23
LSTM-CRF (Castro và cộng sự, 2018)	72,78	68,03	70,33	78,26	74,39	76,27
BiLSTM-CRF+FlairBBP (Santos và cộng sự, 2019a)	74,91	74,37	74,64	83,38	81,17	82,26
ML-BERTBASE-LSTM †	69,68	69,51	69,59	75,59	77,13	76,35
ML-BERTBASE-LSTM-CRF †	74,70	69,74	72,14	80,66	75,06	77,76
ML-BERTBASE	72,97	73,78	73,37	77,35	79,16	78,25
ML-BERTBASE-CRF	74,82	73,49	74,15	80,10	78,78	79,44
PT-BERTBASE-LSTM †	75,00	73,61	74,30	79,88	80,29	80,09
PT-BERTBASE-LSTM-CRF †	78,33	73,23	75,69	84,58	78,72	81,66
PT-BERTBASE	78.36	77.62	77.98	83.22	82.85	83.03
PT-BERTBASE -CRF	78,60	76,89	77,73	83,89	81,50	82,68
PT-BERTLARGE-LSTM †	72.96	72.05	72.50	78.13	78.93	78.53
PT-BERTLARGE-LSTM-CRF †	77,45	72,43	74,86	83,08	77,83	80,37
PT-BERTLARGE	78,45	77,40	77,92	83,45	83,15	83,30
PT-BERTLARGE-CRF	80,08	77,31	78,67	84,82	81,72	83,24

Bảng 2: So sánh kết quả Điểm chính xác, Thu hồi và Điểm F1 trên tập kiểm tra (MiniHAREM). Tất cả các chỉ số đều được tính toán bằng tập lệnh đánh giá CoNLL 2003. Các giá trị in đậm cho biết kết quả SOTA (nhiều kết quả được in đậm nếu chênh lệch trong khoảng tin cậy bootstrap 95%). Các giá trị được báo cáo là giá trị trung bình của nhiều lần chạy với hạt ngẫu nhiên khác nhau. †: phương pháp tiếp cận dựa trên tính năng.

Khi đánh giá, chúng tôi đưa ra dự đoán hợp lệ bằng cách xóa tất cả các chuyển đổi thẻ không hợp lệ cho IOB2 lược đồ, chẳng hạn như các thẻ "I-" xuất hiện ngay sau "O" thẻ hoặc sau thẻ "I-" của một lớp khác. Cái này bước hậu xử lý đánh đổi việc thu hồi để có thể độ chính xác cao hơn.

5 kết quả

Các kết quả chính của thí nghiệm của chúng tôi được trình bày trong Bảng 2. Chúng tôi so sánh hiệu suất của chúng tôi mô hình trên hai kịch bản (tổng thể và chọn lọc). Tất cả các số liệu được tính toán bằng cách sử dụng tập lệnh đánh giá CoNLL 2003,6 bao gồm một vi mô cấp thực thể. Điểm F1 chỉ xem xét các trận đấu chính xác.

Mô hình BERT-CRF tiếng Bồ Đào Nha được đề xuất của chúng tôi thực hiện công nghệ tiên tiến nhất trước đây (BiLSTM CRF+FlairBBP), cải thiện điểm số F1 khoảng 1 điểm cho kịch bản chọn lọc và 4 điểm cho kịch bản kịch bản tổng thể. Thật thú vị, nhúng Flair vượt trội so với các mô hình BERT trên NER tiếng Anh (Akbik và cộng sự, 2018). So với kiến trúc LSTM-CRF không có nhúng theo ngữ cảnh, mô hình của chúng tôi vượt trội hơn 8,3 và 7,0 điểm tuyệt đối trên điểm số F1 trên các kịch bản tổng thể và chọn lọc, tương ứng.

Chúng tôi cũng loại bỏ lớp CRF để đánh giá sự đóng góp. BERT Bồ Đào Nha (PT-BERT-BASE và PT-BERT-LARGE) cũng vượt trội so với trước đó hoạt động, ngay cả khi không thực thi tuần tự phân loại được cung cấp bởi lớp CRF. người mẫu với CRF cải thiện hoặc thực hiện tương tự như của nó các biến thể đơn giản hơn khi so sánh tổng thể F1 điểm số. Chúng tôi lưu ý rằng trong hầu hết các trường hợp, chúng hiển thị cao hơn điểm chính xác nhưng thu hồi thấp hơn.

Trong khi các mô hình BERTLARGE của Bồ Đào Nha là hiệu suất cao nhất trong cả hai trường hợp, chúng tôi quan sát thấy rằng họ bị suy giảm hiệu suất khi được sử dụng theo cách tiếp cận dựa trên tính năng, hoạt động kém hơn các biến thể nhỏ hơn của chúng nhưng vẫn tốt hơn hơn BERT đa ngôn ngữ. Ngoài ra, nó có thể được thấy rằng các mô hình BERTLARGE không mang lại nhiều cải thiện kịch bản chọn lọc khi so sánh với các mô hình BERTBASE . Chúng tôi đưa ra giả thuyết rằng đó là do kích thước nhỏ của bộ dữ liệu NER.

Các mô hình của cách tiếp cận dựa trên tính năng trên mỗi biểu mẫu kém hơn đáng kể so với các mô hình của phương pháp tinh chỉnh. Khoảng cách hiệu suất là được tìm thấy là cao hơn nhiều so với các giá trị được báo cáo cho NER về ngôn ngữ tiếng Anh (Peters và cộng sự, 2019).

Bước xử lý hậu kỳ để lọc ra những nội dung không hợp lệ chuyển đổi cho sơ đồ IOB2 làm tăng F1- điểm trung bình là 1,9 và 1,2 điểm cho phương pháp tiếp cận dựa trên tính năng và tinh chỉnh, tương ứng. Bước này làm giảm 0,4 điểm khi thu hồi, nhưng tăng độ chính xác lên 3,5 điểm, Trung bình.

6https://www.clips.uantwerpen.be/conll2002/ner/bin/conlleval.txt

6 Kết luận

Chúng tôi giới thiệu một công nghệ tiên tiến nhất trên HAREM I tập đoàn bằng các mô hình BERT Bồ Đào Nha trước khi đào tạo trên một kho văn bản lớn chưa được gắn nhãn và tinh chỉnh một mô hình BERT-CRF trong nhiệm vụ NER của Bồ Đào Nha. Mô hình đề xuất của chúng tôi vượt trội so với trạng thái trước đó nghệ thuật (BiLSTM-CRF+FlairBBP), mặc dù nó đã được đào tạo trước trên ít dữ liệu hơn nhiều. xem xét các vấn đề liên quan đến quyết định tiền xử lý và tập dữ liệu ảnh hưởng đến khả năng tương thích đánh giá, chúng tôi đưa ra đặc biệt chú ý đến khả năng tái tạo kết quả của chúng tôi và chúng tôi cung cấp mã và mô hình của mình một cách công khai . Chúng tôi hy vọng rằng bằng cách phát hành tiếng Bồ Đào Nha của chúng tôi Các mô hình BERT, những mô hình khác sẽ có thể đánh giá chuẩn và cải thiện hiệu suất của nhiều NLP khác nhiệm vụ bằng tiếng Bồ Đào Nha. Các thử nghiệm với gần đây hơn và các mô hình hiệu quả, chẳng hạn như RoBERTa và T5, là còn lại cho các công việc trong tương lai.

7 lời cảm ơn

R Lotufo thừa nhận sự hỗ trợ của chính phủ ian Brazil thông qua giới thiệu học bổng CNPq. 310828/2018-0.

Người giới thiệu

Alan Akbik, Duncan Blythe và Roland Vollgraf. 2018. Nhúng chuỗi theo ngữ cảnh cho chuỗi dán nhãn. Trong COLING 2018, Hội nghị quốc tế lần thứ 27 về Ngôn ngữ học tính toán, trang 1638–1649.

Daniela Oliveira F do Amaral và Renata Vieira. 2014. Nerp-crf: Một công cụ để nhận dạng thực thể được đặt tên sử dụng các trường ngẫu nhiên có điều kiện. Linguamática, 6(1):41–49.

Pedro Henrique Luz de Araujo, Teofilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, và Paulo Bermejo. 2018. Lener-br: A bộ dữ liệu để nhận dạng thực thể được đặt tên ở Brazil văn bản pháp luật. Trong Hội nghị quốc tế về xử lý tính toán của tiếng Bồ Đào Nha, trang 313–323. lò xo.

Pedro Castro, Nadia Felix và Anderson Soares. 2019. Biểu diễn theo ngữ cảnh và bán giám sát nhận dạng thực thể được đặt tên cho ngôn ngữ Bồ Đào Nha.

Pedro Vitor Quinta de Castro, Nadia F elix Felipe da Silva và Anderson da Silva Soares. 2018. Nhận dạng thực thể có tên tiếng Bồ Đào Nha bằng cách sử dụng lstm-crf. TRONG Xử lý tính toán của ngôn ngữ Lan Bồ Đào Nha, trang 83–92, Chăm. Springer quốc tế xuất bản.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu và Pavel Kuksa. 2011. Xử lý ngôn ngữ tự nhiên (gần như) từ cào. Tạp chí nghiên cứu máy học, 12(tháng 8):2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, và Kristina Toutanova. 2018. Bert: Đào tạo trước của máy biến áp hai chiều sâu cho ngôn ngữ hiểu biết. Kho lưu trữ nghiên cứu máy tính, arXiv:1810.04805.

Ivo Fernandes, Henrique Lopes Cardoso, và Euge nio Oliveira. 2018. Áp dụng mạng lưới thần kinh sâu hoạt động để nhận dạng thực thể được đặt tên bằng tiếng Bồ Đào Nha văn bản. Năm 2018 Hội nghị quốc tế lần thứ năm về phân tích, quản lý và bảo mật mạng xã hội (SNAMS), trang 284–289. IEEE.

Claudia Freitas, Paula Carvalho, Hugo Gonc¸alo Oliveira, Cristina Mota và Diana Santos. 2010. Hậu cung thứ hai: tiến công bang về nghệ thuật nhận dạng thực thể được đặt tên bằng tiếng Bồ Đào Nha. Trong trích dẫn; Trong Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Kỷ yếu của Hội nghị Quốc tế về Tài nguyên Ngôn ngữ và Đánh giá (LREC 2010)(Valletta 17-23 tháng 5 năm 2010) Hiệp hội tài nguyên ngôn ngữ Lan Châu Âu. Ngôn ngữ Châu Âu Hiệp hội tài nguyên

Jeremy Howard và Sebastian Ruder. 2018. Phổ quát tinh chỉnh mô hình ngôn ngữ để phân loại văn bản. TRONG Kỷ yếu Hội nghị thường niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Long Papers), trang 328–339.

Taku Kudo và John Richardson. 2018. Câu đối: Một trình mã hóa và giải mã từ khóa phụ đơn giản và độc lập với ngôn ngữ để xử lý văn bản thần kinh.

Guillaume Lample, Miguel Ballesteros, Sandeep Sub ramanian, Kazuya Kawakami và Chris Dyer. 2016. Kiến trúc thần kinh để nhận dạng thực thể được đặt tên. Kho lưu trữ nghiên cứu máy tính, arXiv:1603.01360. Phiên bản 3.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma và Radu Soricut. 2019. Albert: Một giải pháp nhỏ cho việc học tự giám sát các biểu diễn ngôn ngữ. bản in trước arXiv arXiv:1909.11942.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming Anh ấy và Piotr Dollar. 2017. Mất tiêu điểm để phát hiện đối tượng dày đặc. Trong Kỷ yếu của hội nghị quốc tế IEEE về thị giác máy tính, trang 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, và Veselin Stoyanov. 2019. Roberta: Một phương pháp tiếp cận ap đào tạo trước bert được tối ưu hóa mạnh mẽ. bản in trước arXiv arXiv:1907.11692.

Monica Marrero, Juli an Urbano, Sonia S anchez- ´ , Cuadrado, Jorge Morato, và Juan Miguel Gomez- Berb´ıs. 2013. Công nhận thực thể được đặt tên: ngụy biện, Thách thức và cơ hội. Tiêu chuẩn máy tính & Giao diện, 35(5):482–489.

Diễn viên:Matthew PetersMark NeumannMohit IyyerMatt Người làm vườn, Christopher Clark, Kenton Lee và Luke Zettlemoyer. 2018. Các câu đại diện từ được ngữ cảnh hóa sâu sắc. Trong Kỷ yếu Hội nghị năm 2018 của Hiệp hội Bắc Mỹ cho Ngôn ngữ học tính toán: Ngôn ngữ con người Technologies, Tập 1 (Long Papers), trang 2227–2237.

Matthew E Peters, Sebastian Ruder và Noah A Smith. 2019. Chỉnh hay không chỉnh? thích nghi trước khi đào tạo biểu diễn cho các nhiệm vụ đa dạng. trong Kỷ yếu của Hội thảo lần thứ 4 về Học biểu diễn cho NLP (RepL4NLP-2019), trang 7–14.

Juliana Pirovani và Elias Oliveira. 2018. Tiếng Bồ Đào Nha nhận dạng thực thể được đặt tên bằng cách sử dụng ngẫu nhiên có điều kiện trường và ngữ pháp địa phương. Trong Kỷ yếu của Hội nghị quốc tế lần thứ 11 về nguồn và đánh giá ngôn ngữ (LREC-2018).

Alec Radford, Karthik Narasimhan, Thời gian Salimans, và Ilya Sutskever. 2018. Cải thiện khả năng hiểu ngôn ngữ bằng cách học không giám sát. Kỹ thuật báo cáo, Báo cáo kỹ thuật, OpenAI.

Diễn viên:Colin RaffelNoam ShazeerAdam RobertsKatherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, và Peter J Liu. 2019. Khám phá những giới hạn của việc học chuyển đổi với một cựu chuyển đổi văn bản thành văn bản thống nhất. bản in trước arXiv arXiv:1910.10683.

Cicero Nogueira dos Santos và Victor Guimaraes. 2015. Tăng cường nhận dạng thực thể được đặt tên bằng cách nhúng ký tự trung tính. Nghiên cứu máy tính Kho lưu trữ, arXiv:1505.05008. Phiên bản 2.

Diana Santos, Nuno Seco, Nuno Cardoso và Rui Vilela. 2006. Harem: Một đánh giá ner nâng cao cuộc thi cho người Bồ Đào Nha.

Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini và Renata Vieira. 2019a. Đánh giá tác động của lớp lót em theo ngữ cảnh đối với nhận dạng thực thể được đặt tên theo tiếng Bồ Đào Nha. Trong Hội nghị Brazil lần thứ 8 về Hệ thống thông minh, BRACIS, Bahia, Brazil, ngày 15-18 tháng 10, trang 437–442.

Joaquim Santos, Juliano Terra, Bernardo Scapini Con soli và Renata Vieira. 2019b. Nhúng văn bản đa miền để nhận dạng thực thể được đặt tên. TRONG IberLEF@SEPLN.

Robyn Speer. 2019. ftfy. Zenodo. Phiên bản 5.5.

Jorge Wagner Filho, Rodrigo Wilkens, Marco Idiart, và Aline Villavicencio. 2018. Kho dữ liệu brwac: Một tài nguyên mở mới cho tiếng Bồ Đào Nha ở Brazil.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quốc V Lê, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Máy thần kinh của Google hệ thống dịch thuật: Thu hẹp khoảng cách giữa dịch thuật của con người và máy móc. Nghiên cứu máy tính Kho lưu trữ, arXiv:1609.08144. Phiên bản 2.

Vikas Yadav và Steven Bethard. 2018. Một cuộc khảo sát về những tiến bộ gần đây trong nhận dạng thực thể được đặt tên từ sâu mô hình học tập. Trong Kỷ yếu Hội nghị quốc tế lần thứ 27 về Ngôn ngữ học tính toán, trang 2145–2158.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Car Bonell, Ruslan Salakhutdinov, và Quoc V Le. 2019. Xlnet: Đào tạo trước tự hồi quy tổng quát để hiểu ngôn ngữ. bản in trước arXiv arXiv:1906.08237.

Phụ lục

A.1 Chuyển đổi câu văn thành văn bản

Từ vựng mảnh ghép câu được tạo ra được chuyển đổi thành mảnh ghép từ tuân theo các quy tắc mã thông báo của BERT. Đầu tiên, tất cả các mã thông báo đặc biệt BERT được chèn vào ([CLS], [MASK], [SEP] và [UNK]) và tất cả các ký tự dấu chấm câu của bảng từ vựng đa ngôn ngữ được thêm vào từ vựng tiếng Bồ Đào Nha. Sau đó, vì BERT phân tách văn bản theo khoảng trắng và dấu chấm câu trước khi áp dụng mã thông báo WordPiece trong các đoạn kết quả, mỗi mã thông báo Câu có chứa các ký tự dấu chấm câu được phân tách tại các ký tự này, các dấu chấm câu bị xóa và các đơn vị từ phụ kết quả được thêm vào từ vựng[7]. Cuối cùng, các đơn vị từ phụ không bắt đầu bằng có tiền tố là “##” và ký tự “ ” bị xóa khỏi các mã thông báo còn lại.

A.2 Ví dụ về tiền xử lý tập dữ liệu HAREM

Một chú thích ví dụ về HAREM chứa nhiều giải pháp, ở định dạng XML, là:

```
<ALT><EM CATEG="PER|ORG">Governo de Cavaco Silva</EM>|<EM CATEG="ORG">Chính phủ</EM>

de <EM CATEG="PER" TIPO="CÁ NHÂN">Cavaco Silva</EM></ALT>
```

trong đó <EM> là thẻ dành cho Thực thể được đặt tên (NE) và “|” xác định các giải pháp thay thế. Chú thích này có thể được hiểu như nhau là có chứa các NE sau:

1. 1 NE: Người ”Governo de Cavaco Silva”
2. 1 ĐB: Tổ chức ”Governo de Cavaco Silva”
3. 2 ĐB: Tổ chức “Governo” và Người “Cavaco Silva”

Các quy tắc được mô tả trong 4.2.2 sẽ chọn giải pháp thứ ba trong ví dụ trên.

[7] Tách ở dấu chấm câu ngụ ý không mã thông báo từ phụ nào có thể chứa cả ký tự chấm câu và ký tự không chấm câu.