**Một tập dữ liệu tiếng Việt để đánh giá khả năng đọc hiểu của máy**

Kiet Van Nguyen1,2, Duc-Vu Nguyen1,2, Anh Gia-Tuan Nguyen1,2, Ngan Luu-Thuy Nguyen1,2

1 Đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh, Việt Nam

2 Đại học Quốc gia Việt Nam, Thành phố Hồ Chí Minh, Việt Nam

{kietnv, vund, anhngt, ngannlt}@uit.edu.vn

**Tóm tắt**

Trên thế giới có hơn 97 triệu người nói tiếng Việt như ngôn ngữ mẹ đẻ. Tuy nhiên, có rất ít nghiên cứu về khả năng đọc hiểu máy (MRC) cho tiếng Việt, nhiệm vụ hiểu một văn bản và trả lời các câu hỏi liên quan đến nó. Do thiếu các bộ dữ liệu chuẩn cho tiếng Việt, chúng tôi giới thiệu Tập dữ liệu Hỏi đáp Tiếng Việt (UIT-ViQuAD), một bộ dữ liệu mới cho ngôn ngữ ít tài nguyên như tiếng Việt để đánh giá các mô hình MRC. Bộ dữ liệu này bao gồm hơn 23.000 cặp câu hỏi-trả lời do con người tạo ra dựa trên 5.109 đoạn văn của 174 bài báo tiếng Việt từ Wikipedia. Đặc biệt, chúng tôi đề xuất một quy trình tạo bộ dữ liệu mới cho MRC tiếng Việt. Các phân tích chuyên sâu của chúng tôi minh họa rằng bộ dữ liệu của chúng tôi yêu cầu các khả năng vượt xa lý luận đơn giản như kết hợp từ và yêu cầu suy luận đơn câu và nhiều câu. Bên cạnh đó, chúng tôi tiến hành các thí nghiệm trên các phương pháp MRC tiên tiến cho tiếng Anh và tiếng Trung như các mô hình thử nghiệm đầu tiên trên UIT-ViQuAD. Chúng tôi cũng ước tính hiệu suất của con người trên bộ dữ liệu và so sánh nó với kết quả thử nghiệm của các mô hình học máy mạnh mẽ. Kết quả là, sự khác biệt đáng kể giữa hiệu suất của con người và hiệu suất mô hình tốt nhất trên bộ dữ liệu chỉ ra rằng có thể cải thiện UIT-ViQuAD trong nghiên cứu trong tương lai. Bộ dữ liệu của chúng tôi có sẵn miễn phí trên trang web của chúng tôi để khuyến khích cộng đồng nghiên cứu vượt qua những thách thức trong MRC tiếng Việt.

**1 Giới thiệu**

Đọc hiểu máy (MRC) là một nhiệm vụ hiểu ngôn ngữ tự nhiên đòi hỏi máy tính phải hiểu một văn bản và sau đó trả lời các câu hỏi liên quan đến nó. MRC là một cốt lõi thiết yếu cho một loạt các ứng dụng xử lý ngôn ngữ tự nhiên như công cụ tìm kiếm và các tác nhân thông minh (Alexa, Google Assistant, Siri và Cortana). Để đánh giá các mô hình MRC, các nguồn tiêu chuẩn vàng với các cặp câu hỏi-trả lời dựa trên tài liệu phải được thu thập hoặc tạo ra bởi con người. Xây dựng một bộ dữ liệu chuẩn đóng một vai trò quan trọng trong việc đánh giá các mô hình xử lý ngôn ngữ tự nhiên, đặc biệt là đối với một ngôn ngữ ít tài nguyên như tiếng Việt.

Các nguồn MRC tiêu chuẩn vàng điển hình cho tiếng Anh là bộ dữ liệu MRC trích xuất khoảng (Rajpurkar et al., 2016; Rajpurkar et al., 2018; Trischler et al., 2017), bộ dữ liệu MRC kiểu điền vào chỗ trống (Hermann et al., 2015; Hill et al., 2015; Cui et al., 2016), bộ dữ liệu MRC nhiều lựa chọn (Richardson et al., 2013; Lai et al., 2017) và bộ dữ liệu MRC dựa trên hội thoại (Reddy et al., 2019; Sun et al., 2019). Đối với các ngôn ngữ khác, có bộ dữ liệu tiếng Trung của MRC trích xuất khoảng (Cui et al., 2019b; Duan et al., 2019), bộ dữ liệu tiếng Trung truyền thống của MRC (Shao et al., 2018), bộ dữ liệu DuReader dựa trên nhật ký truy vấn người dùng tiếng Trung (He et al., 2018) và bộ dữ liệu MRC tiếng Hàn (Lim et al., 2019). Do sự phát triển nhanh chóng của các bộ dữ liệu đọc hiểu, nhiều mô hình dựa trên mạng nơ-ron đã được đề xuất và tạo ra một bước tiến đáng kể trong lĩnh vực nghiên cứu này như Match-LSTM (Wang and Jiang, 2016), BiDAF (Seo et al., 2017), R-Net (Wang et al., 2017), DrQA (Chen et al., 2017), FusionNet (Huang et al., 2018), FastQA (Weissenborn et al., 2017) và QANet (Yu et al., 2018). Các mô hình ngôn ngữ được đào tạo trước, BERT (Devlin et al., 2019) và XLM-R (Conneau et al., 2020) gần đây đã trở nên cực kỳ phổ biến và đạt được hiệu suất tiên tiến cho các tác vụ MRC.

Tiếng Việt là một ngôn ngữ có ít tài nguyên để xử lý ngôn ngữ tự nhiên. Bộ dữ liệu cho MRC được giới thiệu bởi (Nguyen et al., 2020) bao gồm 2.783 câu hỏi và câu trả lời trắc nghiệm dựa trên một tập hợp 417 văn bản tiếng Việt được sử dụng để đánh giá kỹ năng đọc hiểu cho học sinh lớp 1 đến lớp 5. Tuy nhiên, bộ dữ liệu này có kích thước tương đối nhỏ để đánh giá các mô hình học sâu cho MRC tiếng Việt. Vì vậy, chúng tôi đặt mục tiêu xây dựng một bộ dữ liệu lớn mới để đánh giá MRC tiếng Việt.

Mặc dù phương pháp học sâu đã vượt qua hiệu suất của con người trên các bộ dữ liệu SQuAD (Rajpurkar et al., 2016) và NewsQA (Trischler et al., 2017), chúng tôi tự hỏi liệu các mô hình tiên tiến này có thể đạt được hiệu suất tương tự trên các bộ dữ liệu của các ngôn ngữ khác hay không. Để tăng cường hơn nữa sự phát triển của MRC, chúng tôi xây dựng một bộ dữ liệu MRC trích xuất khoảng, trong đó câu trả lời cho các câu hỏi luôn là các khoảng từ một văn bản nhất định cho tiếng Việt. Hình 1 cho thấy một số ví dụ về đọc hiểu trích xuất khoảng tiếng Việt. Trong nghiên cứu này, chúng tôi có bốn đóng góp chính được mô tả như sau.

* Chúng tôi tạo ra một bộ dữ liệu chuẩn để đánh giá MRC tiếng Việt: UIT-ViQuAD bao gồm 23.074 cặp câu hỏi-trả lời do con người tạo ra dựa trên 5.109 đoạn văn của 174 bài báo Wikipedia tiếng Việt. Bộ dữ liệu có sẵn miễn phí trên trang web của chúng tôi cho mục đích nghiên cứu.
* Để có được cái nhìn sâu sắc về bộ dữ liệu, chúng tôi phân tích bộ dữ liệu theo các khía cạnh ngôn ngữ khác nhau bao gồm phân tích dựa trên độ dài (độ dài câu hỏi, độ dài câu trả lời và độ dài đoạn văn) và phân tích dựa trên loại (loại câu hỏi, loại câu trả lời và loại lý luận).
* Để đạt được đánh giá MRC đầu tiên trên UIT-ViQuAD, chúng tôi tiến hành các thí nghiệm với các mô hình MRC là tiên tiến cho tiếng Anh và tiếng Trung. Sau đó, chúng tôi so sánh hiệu suất giữa các mô hình máy và con người về các khía cạnh ngôn ngữ khác nhau. Các phân tích chuyên sâu này cung cấp cái nhìn sâu sắc về MRC dựa trên khoảng trong tiếng Việt.
* MRC đa ngôn ngữ (Cui et al., 2019a) là một xu hướng mới trong xử lý ngôn ngữ tự nhiên. Bộ dữ liệu MRC được đề xuất của chúng tôi cho tiếng Việt cũng có thể là một nguồn tài nguyên cho nghiên cứu đa ngôn ngữ cùng với các bộ dữ liệu tương tự khác như SQuAD, CMRC và KorQuAD.

Phần còn lại của bài báo này được cấu trúc như sau. Phần 2 xem xét các bộ dữ liệu hiện có. Phần 3 giới thiệu quy trình tạo bộ dữ liệu của chúng tôi. Các phân tích chuyên sâu về bộ dữ liệu của chúng tôi được trình bày trong Phần 4. Sau đó, Phần 5 trình bày các thí nghiệm và kết quả phân tích của chúng tôi. Cuối cùng, Phần 6 trình bày kết luận và hướng nghiên cứu trong tương lai.

**2 Các bộ dữ liệu hiện có**

Vì chúng tôi đặt mục tiêu xây dựng một bộ dữ liệu MRC dựa trên khoảng cho tiếng Việt, một loạt các bộ dữ liệu MRC trích xuất khoảng gần đây như SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), CMRC (Cui et al., 2019b) và KorQuAD (Lim et al., 2019) được xem xét trong phần này. Các bộ dữ liệu này được mô tả như sau.

* **SQuAD** là một trong những bộ dữ liệu tiếng Anh phổ biến nhất của MRC dựa trên khoảng. Rajpurkar et al. (2016) đã đề xuất SQuAD v1.1 được tạo bởi những người làm việc đám đông trên 536 bài báo Wikipedia với 107.785 cặp câu hỏi-trả lời. SQuAD v2.0 (Rajpurkar et al., 2018) đã được phát hành với việc bổ sung hơn 50.000 câu hỏi không thể trả lời được tạo ra một cách đối nghịch bởi những người làm việc đám đông theo các câu hỏi ban đầu.
* **NewsQA** là một bộ dữ liệu tiếng Anh khác được đề xuất bởi Trischler et al. (2017), bao gồm 119.633 cặp câu hỏi-trả lời được tạo bởi những người làm việc đám đông trên 12.744 bài báo từ CNN news. Bộ dữ liệu này tương tự như SQuAD vì câu trả lời cho mỗi câu hỏi là một đoạn văn bản có độ dài tùy ý trong bài báo tương ứng.
* **CMRC** (Cui et al., 2019b) là một bộ dữ liệu trích xuất khoảng cho MRC tiếng Trung được giới thiệu trong Hội thảo Đánh giá lần thứ hai về Đọc hiểu Máy tiếng Trung năm 2018, bao gồm khoảng 20.000 câu hỏi do con người chú thích trên các bài báo Wikipedia.
* **KorQuAD** (Lim et al., 2019) là một bộ dữ liệu tiếng Hàn cho MRC dựa trên khoảng, bao gồm hơn 70.000 cặp câu hỏi-trả lời do con người tạo ra trên các bài báo Wikipedia tiếng Hàn.

Các bộ dữ liệu này được nghiên cứu trong việc phát triển và đánh giá các mô hình mạng nơ-ron sâu khác nhau trong NLP, chẳng hạn như Match-LSTM (Wang and Jiang, 2016), BiDAF (Seo et al., 2017), R-Net (Wang et al., 2017), DrQA (Chen et al., 2017), FusionNet (Huang et al., 2018), FastQA (Weissenborn et al., 2017) và QANet (Yu et al., 2018). Gần đây nhất, BERT (Devlin et al., 2019) và XLM-R (Conneau et al., 2020), là các mô hình mạnh mẽ được đào tạo trên nhiều ngôn ngữ, đã đạt được hiệu suất tiên tiến trên các bộ dữ liệu MRC.

Cho đến nay, chưa có bất kỳ bộ dữ liệu nào về văn bản Wikipedia tiếng Việt cho nghiên cứu MRC dựa trên khoảng. Như đã đề cập ở trên, các bộ dữ liệu là điểm chuẩn cho nhiệm vụ MRC và có thể được sử dụng để tổ chức một thử thách khuyến khích các nhà nghiên cứu khám phá các mô hình xử lý tốt nhất. Do đó, đây là động lực chính của chúng tôi để tạo ra bộ dữ liệu mới cho MRC tiếng Việt.

**3 Tạo bộ dữ liệu**

Trong phần này, chúng tôi giới thiệu quy trình đề xuất của chúng tôi để tạo bộ dữ liệu MRC cho ngôn ngữ tiếng Việt. Cụ thể, chúng tôi xây dựng bộ dữ liệu UIT-ViQuAD của mình thông qua năm giai đoạn bao gồm tuyển dụng người làm, thu thập đoạn văn, tìm nguồn câu hỏi-trả lời, xác thực và thu thập câu trả lời bổ sung. Các giai đoạn này được mô tả chi tiết như sau.

* **Giai đoạn 1 - Tuyển dụng người làm:** Chất lượng của bộ dữ liệu phụ thuộc vào người làm chất lượng cao và quy trình tạo dữ liệu. Trong phần này, chúng tôi trình bày việc tuyển dụng người làm để tạo bộ dữ liệu của chúng tôi theo một quy trình nghiêm ngặt, bao gồm bốn giai đoạn khác nhau. (1) Mọi người đăng ký trở thành người làm để tạo các cặp câu trả lời-câu hỏi của bộ dữ liệu; (2) Những người được chọn có kiến thức chung xuất sắc và đã vượt qua bài kiểm tra đọc hiểu của chúng tôi; (3) Người làm việc chính thức được đào tạo cẩn thận trên 500 cặp câu hỏi-trả lời và kiểm tra chéo dữ liệu đã tạo của họ để phát hiện các lỗi phổ biến có thể tránh được khi tạo dữ liệu.
* **Giai đoạn 2 - Thu thập đoạn văn:** Tương tự như SQuAD, chúng tôi cũng sử dụng PageRanks nội bộ của Wikipedia của Project Nayuki để lấy một tập hợp 5.000 bài báo tiếng Việt hàng đầu, từ đó chúng tôi chọn ngẫu nhiên 151 bài báo để tạo bộ dữ liệu. Mỗi đoạn văn tương ứng với một đoạn trong một bài báo. Hình ảnh, hình vẽ và bảng bị loại trừ. Chúng tôi cũng xóa các đoạn văn ngắn hơn 300 ký tự hoặc chứa nhiều ký tự và biểu tượng đặc biệt.
* **Giai đoạn 3 - Tìm nguồn câu hỏi-trả lời:** Người làm hiểu từng đoạn văn và sau đó tạo câu hỏi và câu trả lời tương ứng. Trong quá trình tạo câu hỏi và câu trả lời, người làm tuân theo các quy tắc sau: (1) Người làm được yêu cầu tạo ít nhất ba câu hỏi cho mỗi đoạn văn. (2) Người làm được khuyến khích đặt câu hỏi bằng lời của họ. (3) Câu trả lời là các khoảng văn bản trong đoạn văn được sử dụng để trả lời các câu hỏi. (4) Người làm được khuyến khích tạo sự đa dạng về câu hỏi, câu trả lời và lý luận.
* **Giai đoạn 4 - Xác thực câu hỏi và câu trả lời:** Trong giai đoạn này, người làm thực hiện hai giai đoạn phụ khác nhau để kiểm tra lỗi trong các cặp câu hỏi-trả lời bao gồm tự kiểm tra và kiểm tra chéo. Các lỗi được phân loại thành năm loại khác nhau: câu hỏi không rõ ràng, lỗi chính tả, câu trả lời không chính xác, thiếu hoặc thừa thông tin trong câu trả lời và câu trả lời có ranh giới không chính xác. Hai giai đoạn phụ được mô tả như sau.
  + Tự kiểm tra: Người làm tự sửa đổi các cặp câu hỏi-trả lời của họ.
  + Kiểm tra chéo: Người làm kiểm tra chéo các cặp câu hỏi-trả lời của nhau. Nếu họ phát hiện bất kỳ lỗi nào trong bộ dữ liệu, họ thảo luận với nhau để sửa lỗi.
* **Giai đoạn 5 - Thu thập câu trả lời bổ sung:** Để đánh giá chất lượng tạo bộ dữ liệu, đối với bộ dữ liệu phát triển và thử nghiệm, chúng tôi thêm ba câu trả lời nữa cho mỗi câu hỏi bởi những người làm khác nhau ngoài câu trả lời ban đầu. Trong giai đoạn này, người làm không thể nhìn thấy câu trả lời của nhau và họ được khuyến khích đưa ra câu trả lời đa dạng.

**4 Phân tích bộ dữ liệu**

**4.1 Thống kê tổng quan**

Thống kê của tập huấn luyện (Train), phát triển (Dev) và kiểm tra (Test) của bộ dữ liệu của chúng tôi được mô tả trong Bảng 1. Số lượng câu hỏi của UIT-ViQuAD là 23.074. Trong bảng, số lượng bài báo và đoạn văn, độ dài trung bình của câu hỏi và câu trả lời, và kích thước từ vựng cũng được trình bày.

**4.2 Phân tích dựa trên độ dài**

Chúng tôi trình bày thống kê của bộ dữ liệu của chúng tôi theo ba loại độ dài bao gồm độ dài câu hỏi (xem Bảng 2), độ dài câu trả lời (xem Bảng 2) và độ dài đoạn văn (xem Bảng 3). Các câu hỏi có 11-15 từ của bộ dữ liệu chiếm tỷ lệ cao là 45,29%. Các câu trả lời chủ yếu từ 1 đến 10 từ, chiếm 73,68%. Độ dài của các đoạn văn phần lớn là từ 101 đến 200 từ với 73,13%. Các phân tích này cho thấy bộ dữ liệu của chúng tôi có những đặc điểm riêng.

**4.3 Phân tích dựa trên loại**

Trong phần này, chúng tôi phân tích tập Dev theo các loại khác nhau như loại câu hỏi, loại lý luận và loại câu trả lời. Vì tiếng Việt là một ngôn ngữ chủ-động-từ tương tự như tiếng Trung (Nguyen et al., 2018), các loại câu hỏi tiếng Việt trong UIT-ViQuAD theo một cách thức trong CMRC (Cui et al., 2019b). Do đó, chúng tôi cũng chia các câu hỏi thành bảy loại: Ai, Cái gì, Khi nào, Ở đâu, Tại sao, Như thế nào và Khác. Tuy nhiên, trong tiếng Việt, các từ để hỏi rất đa dạng, vì vậy chúng tôi đã yêu cầu người làm chú thích thủ công loại câu hỏi. Hình 3a trình bày sự phân bố các loại câu hỏi trên bộ dữ liệu của chúng tôi. Các câu hỏi Cái gì chiếm tỷ lệ lớn nhất là 49,97%. So với SQuAD, tỷ lệ câu hỏi Cái gì trong bộ dữ liệu của chúng tôi tương tự như tỷ lệ trong SQuAD (53,60%) (Aniol et al., 2019).

Để khám phá độ khó của lý luận cần thiết, chúng tôi tiến hành chú thích của con người cho các mức độ lý luận khác nhau của câu hỏi, được hiển thị trong Hình 3b. Theo Hill et al. (2015) và Nguyen et al. (2020), người làm chú thích thủ công các câu hỏi thành năm loại lý luận khác nhau với thứ tự tăng dần về độ khó: kết hợp từ (WM), diễn giải (PP), lý luận đơn câu (SSR), lý luận đa câu (MSR) và mơ hồ / không đủ (AoI). Bộ dữ liệu của chúng tôi khó hơn SQuAD và NewsQA vì tỷ lệ các loại suy luận (68,29%) trong bộ dữ liệu của chúng tôi cao hơn so với tỷ lệ trong SQuAD (20,5%) và NewsQA (33,90%) (Trischler et al., 2017).

Theo Rajpurkar et al. (2016) và Trischler et al. (2015), chúng tôi phân loại câu trả lời dựa trên các loại ngôn ngữ của chúng như thời gian (N1), số khác (N2), người (E1), địa điểm (E2), thực thể khác (E3), cụm danh từ (P1), cụm động từ (P3), cụm tính từ (P3), cụm giới từ (P4), mệnh đề (P5) và các loại khác (O). Không giống như SQuAD (Rajpurkar et al., 2016) và NewsQA (Hill et al., 2015), thay vì sử dụng các công cụ tự động để chú thích, các loại câu trả lời trên tập Dev của UIT-ViQuAD được chú thích hoàn toàn bởi người làm. Bảng 4 cho thấy sự phân bố các loại câu trả lời dựa trên các cấu trúc cú pháp khác nhau trên tập Dev của bộ dữ liệu của chúng tôi. Các cụm danh từ phổ biến chiếm tỷ lệ lớn nhất trong UIT-ViQuAD, tương tự như thống kê của SQuAD (Rajpurkar et al., 2016) và NewsQA (Trischler et al., 2017). Ngoài ra, cụm động từ (P2) và các thực thể khác (E3) xếp thứ hai và thứ ba về tỷ lệ phần trăm trong bộ dữ liệu của chúng tôi.

**5 Đánh giá thực nghiệm**

Trong phần này, chúng tôi tiến hành các thí nghiệm với các mô hình MRC tiên tiến để đánh giá bộ dữ liệu của chúng tôi. Để đo lường độ khó của bộ dữ liệu, chúng tôi cũng ước tính hiệu suất của con người đối với nhiệm vụ MRC tiếng Việt. Tương tự như các đánh giá trên bộ dữ liệu tiếng Anh và tiếng Trung (Rajpurkar et al., 2016; Cui et al., 2019b), chúng tôi đã sử dụng hai số liệu đánh giá, kết quả khớp chính xác (EM) và điểm F1, để đánh giá hiệu suất của các mô hình MRC trên bộ dữ liệu của chúng tôi.

**5.1 Hiệu suất của con người**

Để đo lường hiệu suất của con người trên tập phát triển và kiểm tra, chúng tôi đã thuê ba nhân viên khác để trả lời độc lập các câu hỏi trên tập kiểm tra và phát triển. Kết quả là, mỗi câu hỏi trong tập phát triển và kiểm tra có bốn câu trả lời, như được mô tả trong Giai đoạn 5 của Mục 3. Không giống như Rajpurkar et al. (2016) và giống như Cui et al. (2019b), để đo lường hiệu suất, chúng tôi sử dụng phương pháp xác thực chéo. Cụ thể, chúng tôi coi câu trả lời đầu tiên là dự đoán của con người và coi các câu trả lời còn lại là sự thật cơ bản. Chúng tôi thu được ba hiệu suất dự đoán của con người bằng cách lặp đi lặp lại coi câu trả lời đầu tiên, thứ hai và thứ ba là dự đoán của con người. Chúng tôi lấy hiệu suất tối đa trên tất cả các câu trả lời đúng cơ bản cho mỗi câu hỏi. Cuối cùng, chúng tôi tính trung bình của bốn kết quả làm hiệu suất cuối cùng của con người trên bộ dữ liệu.

**5.2 Các phương pháp và đường cơ sở được triển khai lại**

Trong bài báo này, chúng tôi đã triển khai lại các mô hình MRC sau đây trên bộ dữ liệu của chúng tôi như được mô tả trong Mục 4.

* **DrQA:** Chen et al. (2017) đã giới thiệu một mô hình dựa trên mạng nơ-ron đơn giản nhưng hiệu quả cho nhiệm vụ MRC. DrQA Reader đạt được hiệu suất tốt trên nhiều bộ dữ liệu MRC (Rajpurkar et al., 2016; Reddy et al., 2019; Labutov et al., 2018). Do đó, chúng tôi triển khai lại phương pháp này vào bộ dữ liệu của chúng tôi làm mô hình cơ sở đầu tiên để so sánh các mô hình trong tương lai.
* **QANet:** QANet được đề xuất bởi Yu et al. (2018) và mô hình này cũng thể hiện hiệu suất tốt trên nhiều bộ dữ liệu MRC (Rajpurkar et al., 2016; Dua et al., 2019). Mô hình này bao gồm nhiều lớp tích chập theo sau là hai thành phần: self-attention và lớp được kết nối đầy đủ, cho cả mã hóa câu hỏi và đoạn văn cũng như một số lớp được xếp chồng lên nhau trước khi dự đoán đầu ra cuối cùng.
* **BERT:** BERT được đề xuất bởi Devlin et al. (2019). Mô hình này là một phương pháp mạnh mẽ để đào tạo trước các biểu diễn ngôn ngữ, đạt được kết quả tiên tiến trên nhiều nhiệm vụ đọc hiểu. Trong bài báo này, chúng tôi đã sử dụng mBERT (Devlin et al., 2019), một mô hình ngôn ngữ đa ngôn ngữ quy mô lớn được đào tạo trước để đánh giá nhiệm vụ MRC tiếng Việt của chúng tôi.
* **XLM-R:** XLM-R được đề xuất bởi Conneau et al. (2020), một phương pháp cực kỳ mạnh mẽ để đào tạo trước các mô hình ngôn ngữ đa ngôn ngữ ở quy mô lớn, dẫn đến những cải thiện đáng kể về hiệu suất cho một loạt các nhiệm vụ chuyển giao đa ngôn ngữ. Mô hình này vượt trội hơn đáng kể so với BERT đa ngôn ngữ (mBERT) trên nhiều điểm chuẩn đa ngôn ngữ, bao gồm XNLI, MLQA và NER. Trong bài báo này, chúng tôi đánh giá XLM-RBase và XLM-RLarge trên bộ dữ liệu của chúng tôi.

**5.3 Thiết lập thử nghiệm**

Chúng tôi sử dụng một GPU NVIDIA Tesla P100 duy nhất thông qua Google Colaboratory để đào tạo tất cả các mô hình MRC trên bộ dữ liệu của chúng tôi. Chúng tôi sử dụng các nhúng từ được đào tạo trước được giới thiệu bởi (Xuan et al., 2019), bao gồm Word2vec, fastText, ELMO và BERTBase cho DrQA và QANet. Bên cạnh đó, chúng tôi đặt kích thước lô = 32 và số epochs = 40 cho cả hai mô hình. Để đánh giá BERT trên bộ dữ liệu của chúng tôi, chúng tôi triển khai mô hình đa ngôn ngữ được đào tạo trước mBERT (Devlin et al., 2019) và các mô hình đa ngôn ngữ được đào tạo trước XLM-R (Conneau et al., 2020) với cấu hình cơ sở được cung cấp bởi HuggingFace. Dựa trên các đặc điểm của bộ dữ liệu, chúng tôi sử dụng độ dài câu trả lời tối đa là 300, độ dài câu hỏi là 64 và độ dài chuỗi đầu vào là 384 cho tất cả các thí nghiệm trên mBERT và XLM-R.

**5.4 Kết quả đánh giá**

Bảng 5 trình bày hiệu suất của các mô hình của chúng tôi cùng với hiệu suất của con người trên tập phát triển và kiểm tra của bộ dữ liệu của chúng tôi. Đối với EM và F1-core, XLM-RLarge vượt trội hơn đáng kể so với các mô hình khác nhưng vẫn thấp hơn nhiều so với hiệu suất của con người. Trên tập kiểm tra, mô hình dự đoán câu trả lời với điểm F1 là 87,02%. Tuy nhiên, kết quả khớp chính xác của mô hình này đạt 68,98%, thấp hơn đáng kể so với điểm F1.