**1. Phân tích đoạn văn (Context Processing)**

Mô hình **BERT** hoặc các biến thể như **PhoBERT** có khả năng hiểu ngữ nghĩa của đoạn văn bằng cách trích xuất các đặc điểm quan trọng từ đoạn văn. Dưới đây là các bước:

**a. Tokenization**

* Đoạn văn được chia nhỏ thành các **token** (các từ hoặc cụm từ).
* Mỗi token được ánh xạ thành một vector số sử dụng bộ tokenizer tương ứng (ví dụ: bert-base-uncased cho tiếng Anh hoặc vinai/phobert-base cho tiếng Việt).

**Ví dụ**:

* Đoạn văn: *"Môi trường là không gian sống của con người."*
* Tokenized: ['Môi', 'trường', 'là', 'không', 'gian', 'sống', 'của', 'con', 'người', '.']

**b. Hiểu ngữ cảnh (Contextual Embeddings)**

* Các token sau đó được đưa qua các lớp của **BERT** để tạo ra **embedding ngữ cảnh**: Các vector số đại diện cho ý nghĩa của từ trong ngữ cảnh.

**Ví dụ**:

* "Môi trường" có embedding khác khi ở đoạn văn về sinh thái so với đoạn văn về hóa học.

**2. Trích xuất thông tin quan trọng (Key Information Extraction)**

Mục tiêu ở đây là tìm ra các đoạn hoặc từ khóa quan trọng trong đoạn văn, có thể được sử dụng làm câu hỏi hoặc đáp án.

**a. Xác định ý chính**

* Sử dụng mô hình QA (Question Answering) của BERT để trả lời các câu hỏi đơn giản từ đoạn văn.
* Câu trả lời thường là một cụm từ hoặc một câu trong đoạn văn.

**Ví dụ**:

* Đoạn văn: *"Môi trường là không gian sống của con người, bao gồm không khí, đất, nước và các sinh vật."*
* Câu hỏi: *"Môi trường bao gồm những yếu tố nào?"*
* Đáp án: *"Không khí, đất, nước và các sinh vật."*

**b. Sinh câu hỏi tự động**

* Từ các đoạn thông tin chính, sinh câu hỏi phù hợp bằng cách đảo cấu trúc câu hoặc sử dụng các công cụ tạo câu hỏi (Question Generation).

**Ví dụ**:

* Đoạn văn: *"Trái đất quay quanh mặt trời với tốc độ trung bình là 29,8 km/s."*
* Câu hỏi sinh ra: *"Trái đất quay quanh mặt trời với tốc độ bao nhiêu?"*

**3. Sinh phương án trả lời (Answer Generation)**

Sau khi đã có câu hỏi và đáp án đúng:

* **Đáp án đúng**: Được trích xuất từ đoạn văn bằng mô hình QA.
* **Đáp án nhiễu (Distractors)**:
  + Sinh ra từ các từ khóa khác trong đoạn văn.
  + Hoặc sử dụng các thuật toán như từ đồng nghĩa, từ trái nghĩa.

**Ví dụ**:

* Đoạn văn: *"Nước chiếm 71% bề mặt Trái đất."*
* Câu hỏi: *"Nước chiếm bao nhiêu phần trăm bề mặt Trái đất?"*
* Đáp án đúng: *71%*
* Đáp án nhiễu: *51%, 61%, 81%*

**4. Hiệu chỉnh (Post-Processing)**

**a. Đảm bảo tính hợp lý của câu hỏi**

* Kiểm tra ngữ pháp và logic của câu hỏi.
* Loại bỏ các câu hỏi mơ hồ hoặc không rõ ràng.

**b. Đánh giá chất lượng câu hỏi**

* Đánh giá chất lượng câu hỏi dựa trên các tiêu chí:
  + Rõ ràng.
  + Có một đáp án đúng duy nhất.
  + Các đáp án nhiễu không dễ đoán.

**5. Hoạt động tổng thể**

**Luồng xử lý**

1. **Đầu vào**: Một đoạn văn.
2. **Xử lý**:
   * Trích xuất ý chính từ đoạn văn.
   * Sinh câu hỏi phù hợp.
   * Sinh đáp án đúng và các đáp án sai.
3. **Đầu ra**: Một câu hỏi trắc nghiệm hoàn chỉnh với các đáp án.

**Sơ đồ hoạt động**

css

Copy code

[Đoạn văn]

↓

[Tách ý chính] - Mô hình QA

↓

[Sinh câu hỏi] - Mô hình Tạo Câu hỏi

↓

[Sinh đáp án] - Đáp án đúng + Đáp án nhiễu

↓

[Hiệu chỉnh]

↓

[Câu hỏi trắc nghiệm]

**Ứng dụng thực tế**

* **Giáo dục**: Tạo nhanh các đề thi từ tài liệu giảng dạy.
* **Luyện thi**: Sinh nhiều dạng câu hỏi cho học sinh luyện tập.
* **E-learning**: Kết hợp vào các hệ thống học liệu trực tuyến.