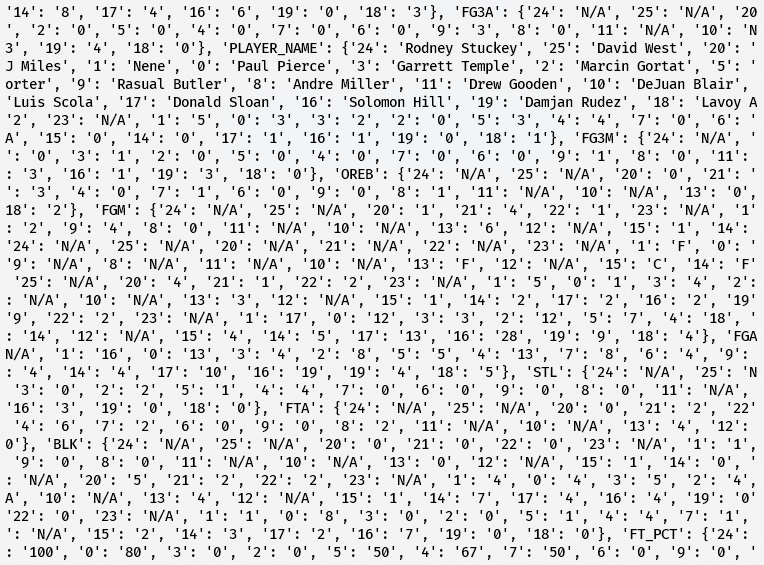
**CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP**

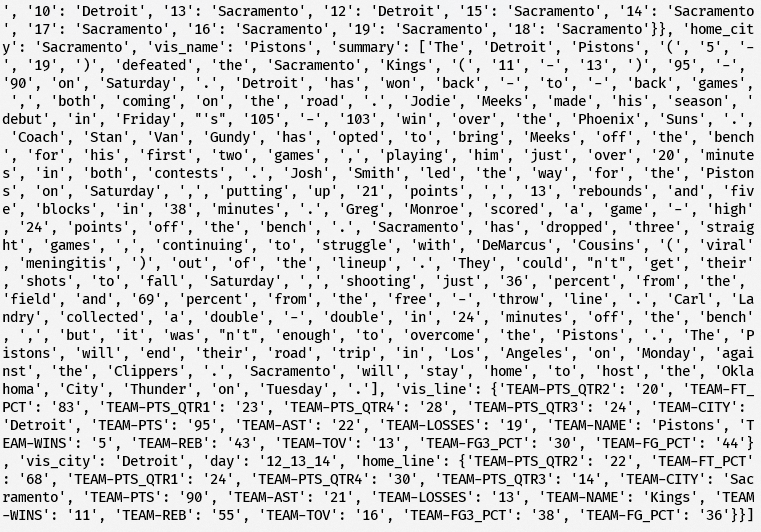
1. **Giới thiệu tập dữ liệu**

Mô hình sử dụng tập dữ liệu Rootowire:

<https://github.com/harvardnlp/boxscore-data.git>

Tập dữ liệu Rotowire là các bản tóm tắt các trận đấu bóng rổ NBA (do con người lập nên) được trình bài một cách hợp lý dựa vào sự canh chỉnh điểm số thành từng dòng tương ứng. Rotowire có 4853 bản tóm tắt riêng biệt, bao gồm các trận đầu NBA được chơi từ 1/1/2014 đến 29/3/2017, trong đó có một số trận đấu có nhiều bản tóm tắt. Các bản tóm tắt đã được chia ngẫu nhiên thành những tệp JSON bao gồm: 3398 tập huấn luyện (training), 727 tập thẩm định (validation) và 728 tập kiểm thử (test).

Hình. Tập dữ liệu huấn luyện (training)

Hình. Tập dữ liệu thẩm định (validation)



Hình. Tập dữ liệu kiểm thử (test)

Trong đó có các trường dữ liệu bao gồm:

* home\_name: tên đội chủ nhà (unicode).
* home\_city: Tên thành phố đội chủ nhà (unicode).
* vis\_name: tên đội khách (unicode).
* vis\_name: tên thành phố của đội khác (unicode).
* day: ngày thi đầu được định dạng %MM\_%DD\_%YY (unicode).
* summary: Mã của tóm tắt trận đầu.
* home-line: Dòng ghi nhận điểm của đội chủ nhà.
* vis-line: Dòng ghi nhận điểm của đội khách.
* box-core: Điểm hộp số.

Đối với trường dữ liệu Line-score (điểm theo dòng) bao gồm các thuộc tính như sau:

* TEAM-NAME: Tên đội bóng (unicode).
* TEAM-CITY: Thành phố cả đội bóng (unicode).
* TEAM-AST: Số lần hỗ trợ của đội bóng (integer as unicode).
* TEAM-FG3\_PCT: Tỷ lệ 3 con trỏ theo nhóm (integer as unicode).
* TEAM-FG\_PCT: Tỷ lệ bàn thắng được thực hiện bởi đội bóng (integer as unicode).
* TEAM-FT\_PCT: Tỷ lệ 3 điểm được thực hiện bởi đội bóng (integer as unicode).
* TEAM\_LOSSES: Đội thua (integer as unicode).
* TEAM-PTS: Tổng điểm của đội bóng (integer as unicode).
* TEAM-PTS\_QTR1: Điểm trong hiệp đầu tiên của đội bóng (integer as unicode).
* TEAM-PTS\_QTR2: Điểm trong hiệp thứ hai của đội bóng (integer as unicode).
* TEAM-PTS\_QTR3: Điểm trong hiệp thứ ba của đội bóng (integer as unicode).
* TEAM-PTS\_QTR4: Điểm trong hiệp thứ tư của đội bóng (integer as unicode).
* TEAM-REB: Tổng điểm Rebounds của đội (integer as unicode).
* TEAM-TOV: Tổng điểm Turnovers (integer as unicode).
* TEAM-WINS: Đội bóng thắng cuộc (integer unicode).

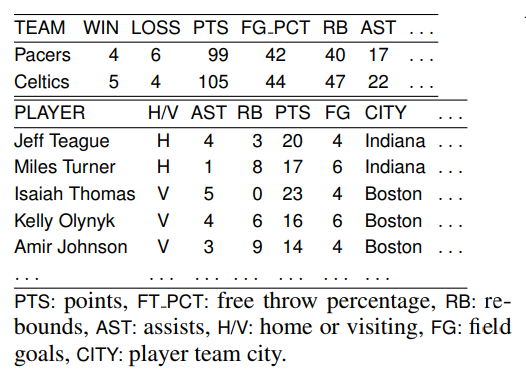
Đối với trường dữ liệu Box-core gồm có các thuộc tính sau:

* AST – cầu thủ hỗ trợ (row\_number -> integer as unicode).
* BLK - cầu thủ vị trí blocks (row\_number -> integer as unicode).
* DREB – cầu thủ ngăn chặn rebounds (row\_number -> integer as unicode).
* FG3A – cầu thủ tấn công 3 điểm (row\_number -> integer as unicode).
* FG3M – cầu thủ tạo 3 điểm (row\_number -> integer as unicode)
* FG3\_PCT – Tỷ lệ cầu thủ tạo 3 điểm (row\_number -> integer as unicode).
* FGA – cầu thủ với mục tiêu tấn công (row\_number -> integer as unicode).
* FGM – cầu thủ với mục tiêu ghi điểm (row\_number -> integer as unicode).
* FG\_PCT – Tỷ lệ cầu thủ ghi điểm (row\_number -> integer as unicode).
* FIRST\_NAME – Tên cầu thủ (row\_number -> unicode).
* FTA – Cầu thủ ném phạt (row\_number -> integer as unicode).
* FTM – Cầu thủ thực hiện ném phạt (row\_number -> integer as unicode).
* FT\_PCT – Tỷ lệ phần trăm cầu thủ ném phạt (row\_number -> integer as unicode).
* MIN – Số phút cầu thủ đã ra sân (row\_number -> integer as unicode).
* OREB – Cầu thủ phản hồi rebounds (row\_number -> integer as unicode).
* PF – Cầu thủ phạm lỗi cá nhân (row\_number -> integer as unicode).
* PLAYER\_NAME – Tên đầy đủ của cầu thủ (row\_number -> integer as unicode).
* PTS – Điểm của cầu thủ (row\_number -> integer as unicode).
* REB – Tổng điểm rebounds của cầu thủ (row\_number -> integer as unicode).
* SECOND\_NAME – Tên lót của cầu thủ (row\_number -> integer as unicode).
* START\_POSITION – Vị trí cầu thủ (row\_number -> unicode).
* STL – Cầu thủ lấy bóng (row\_number -> integer as unicode).
* TEAM\_CITY – Thành phố của đội bóng (row\_number -> unicode).
* TO – Cầu thủ turnovers (row\_number -> integer as unicode).

1. **Thiết kế và cài đặt giải pháp**
   1. **Mô hình xử lý dữ liệu**

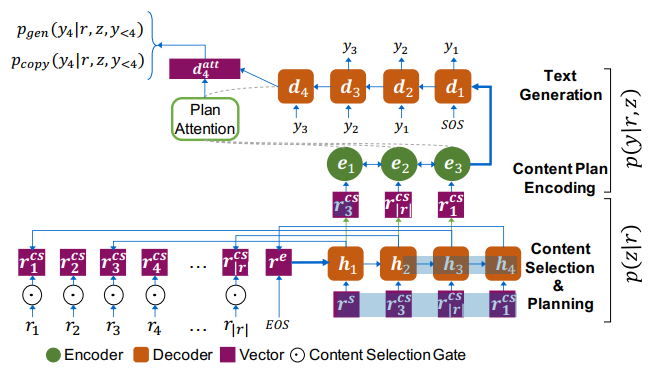
Đầu vào của mô hình là một bảng các bảng ghi (hình ). Mỗi bản ghi sẽ được gọi là rj, được biểu diễn dưới dạng , trong mỗi bản ghi gồm có 4 thuộc tính:

* Loại thuộc tính (rj,1 ; …; LOSS, CITY): là thuộc tính tên từng cột trong bản ghi bao gồm các thuộc tính như WIN, LOSS, PTS…
* Đối tượng (rj,2 ;…; Pacers, Miles Turner): là những thuộc tính liên quan đến đội bóng thi đầu và cầu thủ của đội bóng.
* Dữ liệu (rj,3 ;…; 11; Indiana): là thuộc tính liên quan đến thành phố và tổng ba chỉ số AST, RB, FG của cầu thủ.
* Dữ liệu (rj,4 ;…, H): là thuộc tính chỉ ra là đội khách hay đội nhà (cột H/V).



Hình. Một bản ghi đầu vào

Đầu ra của mô hình là một tài liệu được gọi là y bao gồm các thành phần y = y1 · · · với là độ dài của tài liệu.

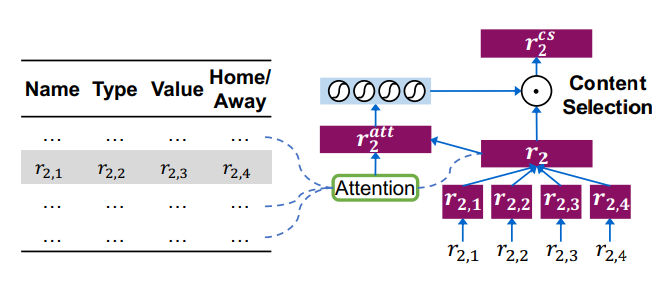
Hình. Mô hình lựa chọn và lập kế hoạch nội dung

Từ dữ liệu đầu vào là các bản ghi đã được lựa chọn nội dung để mã hóa dưới dạng các véc-tơ. Các véc-tơ sẽ được đi qua cổng tuyển chọn nội dung để tìm lọc ra các nội dung hữu ích, tránh sự trùng lập nội dung không cần thiết tạo thành một bản ghi mới. Tiếp theo, bản ghi mới này sẽ được sắp xếp theo trình tự kế hoạch tương ứng với thứ tự các véc-tơ trong bản ghi ban đầu. Đồng thời tại bước này, sử dụng bộ giải mã LSTM để tạo ra mã thông báo tương ứng với các vị trí trong đầu vào. Sau khi đã đồng bộ được thứ tự các véc-tơ, quá trình sẽ giải mã các thông tin văn bản dựa trên mạng nơ-ron tuần hoàn với các đơn vị LSTM, đó là sự đối chiếu của bản ghi với sự dự đoán từ trước. Đồng thời, mô hình cũng tăng cường thêm với bộ giải mã với cơ chế sao chép trực tiếp từ các phần giá trị của bản ghi trong kế hoạch nội dung.

* 1. **Quá trình xử lý dữ liệu**
     1. **Bộ mã hóa bản ghi (Record Encoder)**
* Đầu vào của mô hình là một bảng các bản ghi không có thứ tự, mỗi bản ghi được biểu diễn dưới dạng các tính năng .
* Công việc tiếp theo được thực hiện đó là nhúng các tính năng thành véc-tơ và sử dụng multilayer perceptron để lấy một biểu diễn véc-tơ rj cho mỗi bản ghi theo công thức:

Trong đó “ ; ” là chỉ phép nối véc-tơ, với n là tham số, ReLU là chức năng kích hoạt chỉnh lưu.

* + 1. **Cổng lựa chọn nội dung (Content Selection Gate)**
* Ngữ cảnh của một bản ghi rất quan trọng trong việc xác định giá trị của bản ghi đó có hữu ích hay không so với các bản ghi khác trong bảng. Vì thế mô hình sử dụng cơ chế cổng lựa chọn nội dung nhằm xác định tính hữu dụng của bản ghi trong bảng theo từng ngữ cảnh.
* Bước đầu tiên, mô hình tính điểm chú ý () qua đầu vào bảng, và sử dụng chúng để thu được một véc-tơ lưu ý () cho mỗi bản ghi , theo công thức:

Trong đó , là ma trận tham số, = 1

Hình. Cơ chế lựa chọn nội dung

* Tiếp theo cơ chế lựa cổng lựa chọn nội dung được áp dụng cho các véc-tơ . Đồng thời tìm được bản ghi đại diện mới được xác định dựa theo công thức:

Trong đó ⊙ biểu thị phép nhân khôn ngoan nguyên tố và cổng ∈ kiểm soát lượng thông tin xuất ra từ .

Nói cách khác, mỗi phần tử trong  được cân nhắc bởi phần tử tương ứng của cổng lựa chọn nội dung .

* + 1. **Kế hoạch nội dung (**Content Planning**)**
* Kế hoạch nội dung được xây dựng bằng cách lập bảng đồ văn bản trong phần tóm tắt vào các thực thể trong bảng đầu vào là giá trị và kiểu (mối quan hệ giữa các thực thể).
* Một kế hoạch là một chuỗi các con trỏ với mỗi mục nhập trỏ đến bản ghi đầu vào .



Hình. Đoạn trích của một kế hoạch

* Cho *z = z1 …*  biểu thị trình tự lập kế hoạch nội dung với mỗi đơn vị *zk* trỏ đến một bản ghi đầu vào và xác suất *p(z|r)* được phân rã được tính theo công thức:

Trong đó

* Vì các mã thông báo đầu ra của giai đoạn lập kế hoạch nội dung tương ứng với các vị trí trong trình tự đầu vào nên mô hình sử dụng mạng con trỏ (Pointer Nerwork). Từ quá trình này, mô hình sẽ chú ý mã thông báo trình tự của các dữ liệu đầu vào thay vì phải tạo ra các biểu diễn có trọng số của các dữ liệu đầu vào đã mã hóa.
* Tiếp đến, mô hình được áp dụng bộ giải mã LSTM để tạo mã tương ứng với các vị trí ở đầu vào. Tại bước này, trạng thái đầu tiên của bộ giải mã được khởi tạo bởi công thức là giá trị trung bình của véc-tơ bản ghi.
* Mô hình được xây dựng bằng cách quan tâm đến giá trị các bản ghi đầu vào. Với mỗi bản ghi tại bước *k*, gọi *hk* là trạng thái ẩn của LTSM:

Trong đó xác suất được chuẩn hóa thành 1 và là các tham số. Khi trỏ đến bản ghi , mô hình sử dụng véc-tơ tương ứng làm đầu vào của đơn vị LSTM tiếp theo trong bộ giải mã.

* + 1. **Tạo văn bản (Text Generation)**
* Xác suất xuất ra văn bản *y* có điểu kiện có kế hoạch nội dung *z* và bảng đầu vào *r* được mô hình hóa như sau:

Trong đó , mô hình sử dụng kiến trúc bộ mã hóa và giải mã với cơ chế chú ý để tính .

* Bước đầu tiên của quá trình, mô hình sẽ mã hóa kế hoạch nội dung *z* thành bằng LSTM hai chiều. Bởi vì kế hoạch nội dung là một chuỗi các bản ghi đầu vào, vì thế mô hình cung cấp trực tiếp các véc-tơ bản ghi tương ứng làm đầu vào cho các đơn vị LTSM, chúng chia sẻ bản ghi mã hóa với trạng thái đầu tiên.
* Bộ giải mã văn bản dựa trên mạng nơ-ron tuần hoàn làm việc với các đơn vị LSTM được khởi tạo với các trạng thái ẩn ở bước cuối cùng trong bộ mã hóa. Lúc giải mã bước *t*, đầu vào của các đơn vị LSTM là sự nhúng dự đoán trước đó . Gọi là trạng thái ẩn của đơn vị LTSM thứ *t*. Xác suất dự đoán từ vựng đầu ra được tính bằng công thức:

Trong đó , , ,, là các tham số và là đầu ra kích thước từ vựng.

* Mô hình được tăng cường thêm cho bộ giải mã với cơ chế sao chép, tức là khả năng sao chép các từ trực tiếp từ các phần giá trị của bản ghi trong kế hoạch nội dung, được thử nghiệm với phương pháp joint (Gu et al. 2016) và các phương pháp sao chép có điều kiện (Gulcehre et al. 2016). Cụ thể, với biến cho mỗi bước thời gian để cho biết liệu mã thông báo đã được sao chép () hay chưa được sao chép (). Xác suất tạo ra được tính bằng công thức:

Trong đó bị loại bỏ ra ngoài lề.

* Với phương pháp sao chép joint, xác suất sao chép từ các giá trị bản ghi và tạo từ vựng được chuẩn hóa toàn cầu dựa trên công thức:

Trong đó chỉ ra rằng có thể sao chép từ , .

* Với các phương pháp sao chép có điều kiện, biến được tính như một cổng chuyển đổi, sau đó được sử dụng để lấy xác suất đầu ra:

Trong đó và , , là các tham số.

* + 1. **Huấn luyện và suy luận (Training and Inference)**