**TÓM TẮT**

**A. MÔ TẢ**

* Có đầu vào là phi ngôn ngữ (Reiter và Dale 2000; Gatt và Krahmer 2018) là các dạng khác nhau như cơ sở dữ liệu hồ sơ, bảng tính, cơ sở kiến ​​thức hệ thống chuyên gia, mô phỏng hệ thống vật lý, v.v.
* Bảng 1 sử dụng cơ sở dữ liệu chứa số liệu thống kê về các trò chơi bóng rổ NBA.
* Các phương pháp truyền thống để chuyển dữ liệu thành văn bản (data-to-text) là triển khai mô-đun bao gồm lập kế hoạch nội dung chọn nội dung cụ thể từ một số đầu vào và xác định cấu trúc của văn bản đầu ra), lập kế hoạch câu (xác định cấu trúc và từ vựng nội dung của mỗi câu) và hiện thực hóa bề mặt (chuyển đổi kế hoạch câu thành một chuỗi bề mặt).
* Các hệ thống thần kinh gần đây (Lebret và cộng sự 2016; Mei và cộng sự 2016; Wiseman và cộng sự al. 2017) không mô hình hóa rõ ràng bất kỳ giai đoạn nào trong số này, thay vì họ được đào tạo theo kiểu end-to-end bằng cách sử dụng kiến ​​trúc bộ mã hóa-giải mã rất thành công (Bahdanau et al. 2015) như xương sống của họ nhưng lại gặp khó khăn trong việc nắm bắt cấu trúc dài và tạo ra các tài liệu dài hơn một câu. Đồng thời các kỹ thuật tạo văn bản thần kinh hoạt động kém hơn trong việc lựa chọn nội dung, chúng phải vật lộn để duy trìsự gắn kết liên tâm và nói chung là hợp lý thứ tự của các dữ kiện đã chọn trong văn bản đầu ra.
* Bài báo này mô hình hóa rõ ràng việc lập kế hoạch và lựa chọn nội dung trong kiến ​​trúc chuyển dữ liệu thành văn bản thần kinh. Mô hình của họ học một kế hoạch nội dung từ đầu vào và các điều kiện trên kế hoạch nội dung để tạo tài liệu đầu ra, từ đó tạo ra những lợi thế cho việc tạo ra tài liệu nhiều câu như sau:
  + Nó đại diện cho một tổ chức cấp cao của cấu trúc tài liệu cho phép bộ giải mã tập trung vào các nhiệm vụ dễ dàng hơn lập kế hoạch câu và hiện thực hóa bề mặt.
  + Nó làm cho quá trình tạo dữ liệu thành tài liệu dễ hiểu hơn bằng cách tạo ra một đại diện trung gian.
  + Giảm dư thừa trong đầu ra, vì ít có khả năng kế hoạch nội dung chứa cùng một thông tin ở nhiều nơi.

**B. TIỂU SỬ**

* Công việc trước đây về lập kế hoạch nội dung đã dựa vào các nhà hoạch định (Dale 1988), dựa trên Lý thuyết Cấu trúc Tu từ (Hovy 1993) và lược đồ (McKeown et al. 1997). Nội dung người lập kế hoạch được xác định bằng cách phân tích văn bản mục tiêu và đưa ra các quy tắc thủ công. Duboue và McKeown (2001) nghiên cứu các ràng buộc về thứ tự đối với các kế hoạch nội dung và trong công việc tiếp theo (Duboue và McKeown 2002) học cách lập kế hoạch nội dung từ một kho dữ liệu đầu vào và đầu ra của con người phù hợp. Một số nhà nghiên cứu (Mellish và cộng sự 1998; Karamanis 2004) chọn kế hoạch nội dung theo chức năng xếp hạng.
* Các công việc gần đây hơn tập trung vào các hệ thống end-to-end để thay thế các thành phần riêng lẻ.
* Hầu hết các mô hình đều đơn giản hóa các giả định như tạo ra mà không có bất kỳ lựa chọn hoặc lập kế hoạch nội dung nào (Belz 2008; Wong và Mooney 2007) hoặc lựa chọn nội dung mà không lập kế hoạch (Konstas và Lapata 2012; Angeli và cộng sự.
* Công việc của chúng tôi là gần nhất với các mô hình mạng nơron gần đây tìm hiểu trình tạo từ dữ liệu và tài nguyên văn bản đi kèm.

**C. VẤN ĐỀ XÂY DỰNG**

* Đầu vào của mô hình là một bảng ghi (Bảng 1). Mỗi bản ghi có bốn tính năng bao gồm loại của nó (rj, 1; ví dụ: LOSS, CITY), thực thể (rj, 2; ví dụ: Pacers, Miles Turner), giá trị (rj, 3; ví dụ: 11, Indiana) và liệu một cầu thủ thuộc đội nhà hoặc đội khách (rj, 4; xem cột H / V trong Bảng 1), được biểu diễn dưới dạng .
* Đầu ra y là một tài liệu chứa các từ y = y1 · · · y | y | ở đâu | y | là độ dài tài liệu.
* Kiến trúc tổng thể của mô hình gồm 2 giai đoạn (Hình 2):
  + Lựa chọn nội dung và lập kế hoạch hoạt động dựa trên các bản ghi đầu vào của cơ sở dữ liệu và tạo ra một kế hoạch nội dung xác định những bản ghi nào sẽ được được diễn đạt bằng lời trong tài liệu và theo thứ tự nào.
  + Tạo văn bản tạo ra văn bản đầu ra với kế hoạch nội dung làm đầu vào; ở mỗi bước giải mã, thế hệ mô hình chú ý đến các biểu diễn vector của các bản ghi trong kế hoạch nội dung.

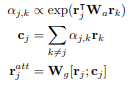
1. **Bộ mã hóa bản ghi:**

* 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 .
* chúng tôi nhúng các tính năng thành vectơ, rồi sử dụng perceptron nhiều lớp để lấy một biểu diễn vectơ cho mỗi bản ghi:

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

1. **Cổng lựa chọn nội dung:**

* Bối cảnh của một bản ghi có thể hữu ích trong việc xác định tầm quan trọng của nó so với các bản ghi khác trong bảng. chúng tôi sử dụng cơ chế cổng lựa chọn nội dung (Hình 3).



Trong đó , là ma trận tham số, = 1.

* Tiếp theo, họ áp dụng cơ chế kiểm soát lựa chọn nội dung đến và biểu diễn bản ghi mới.

1. **Nội dung kế hoạch**

* Văn bản đầu ra tuy dài nhưng vẫn tuân theo cấu trúc kinh điển. Mô hình này học từ kế hoạch nội dung từ dữ liệu đào tạo thông qua cách tiếp cận trích xuất thông tin bằng cách lập bản đồ văn bản trong phần tóm tắt vào các thực thể trong bảng đầu vào, giá trị và kiểu (tức là quan hệ).
* Một kế hoạch là một chuỗi các con trỏ với mỗi mục nhập trỏ đến bản ghi đầu vào.
* Thứ tự trong kế hoạch tương ứng với trình tự mà các thực thể xuất hiện trong phần tóm tắt trò chơi.
* 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, họ sử dụng Mạng con trỏ (Vinyals và cộng sự 2015).
* Chúng tôi sử dụng bộ giải mã LSTM để tạo mã thông báo tương ứng với các vị trí trong đầu vào.

1. **Tạo văn bản**

* Bộ giải mã văn bản cũng dựa trên mạng nơ-ron lặp lại với các đơn vị LSTM. Bộ giải mã được khởi tạo với các trạng thái ẩn của bước cuối cùng trong bộ mã hóa. Lúc giải mã bước t, đầu vào của đơn vị LSTM là sự nhúng củatừ dự đoán trước đó.
* Họ tăng cường thêm 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.

1. **Huấn luyện và suy luận**

* Mô hình của họ được đào tạo để tối đa hóa khả năng ghi nhật ký của vàng3 kế hoạch nội dung cho trước bảng ghi r và sản lượng vàng văn bản đưa ra kế hoạch nội dung và bản ghi bảng.
* Đối với mỗi giai đoạn, chúng tôi sử dụng tìm kiếm chùm để đạt được kết quả tốt nhất.

**D. THIẾT LẬP THỬ NGHIỆM**

* Dữ liệu: đã đào tạo và đánh giá mô hình trên ROTOWIRE, Các bản tóm tắt được viết chuyên nghiệp, tương đối có cấu trúc tốt và dài (trung bình 337 từ). Số loại bản ghi là 39, số lượng bản ghi trung bình là 628, kích thước từ vựng là 11,3K từ và số lượng mã thông báo là 1,6 triệu (tập dữ liệu các trận đấu bóng rổ), đồng thời đã đào tạo trên 3.398 bản tóm tắt, thử nghiệm trên 728 và sử dụng 727 để xác thực.
* Nội dung kế hoạch: Sử dụng Hệ thống IE được giới thiệu trong Wiseman et al. (2017) xác định thực thể ứng cử viên và giá trị xuất hiện trong văn bản, sau đó dự đoán loại (hay còn gọi là quan hệ) của mỗi cặp ứng viên. Đào tạo hệ thống IE trên ROTOWIRE bằng cách xác định các khoảng từ có thể đại diện các thực thể và số. Sau đó, họ xem xét từng cặp số thực thể trong cùng một câu và nếu có ghi lại trong cơ sở dữ liệu với các thực thể và giá trị phù hợp, cặp được chỉ định loại bản ghi tương ứng hoặc cách khác được gắn nhãn "không có" để chỉ ra các cặp không liên quan. Dự đoàn quan hệ bằng cách tập hợp 3 mô hình tích chập và 3 mô hình LSTM hai chiều. Từ đó mô hình đạt độ chính xác 94% và nhớ lại khoảng 80% các mối quan hệ được cấp phép bởi các hồ sơ.
* Cấu hình huấn luyện:
  + Xác nhận các siêu tham số của mô hình trên tập phát triển.
  + Không điều chỉnh kích thước của nhúng từ và các lớp ẩn LSTM và sử dụng một giá trị iá trị 600 được báo cáo trong Wiseman etal. (2017)
  + Sử dụng mạng con trỏ một lớp trong quá trình lập kế hoạch nội dung và LSTM hai lớp trong quá trình tạo văn bản.
  + Đầu vào áp dụng văn bản bộ giải mã với tỷ lệ bỏ học là 0,3 và tỷ lệ học tập ban đầu là 0,15, giảm tốc độ học tập được chọn từ {0,5, 0,97},và kích thước lô là 5.
  + Sử dụng BPTT (Mikolov et al. 2010) và đặt kích thước cắt bớt đến 100 đặt kích thước chùm thành 5 trong khi suy luận.
  + Được triển khai trong OpenNMT-py (Klein et al. 2017).

**E. KẾT QUẢ**

* Đánh giá đầu ra mô hình bằng các chỉ số được xác định trong Wiseman et al. (2017). Ý tưởng là sử dụng một hệ thống IE khá chính xác trên bản tóm tắt vàng và tự động và so sánh xem các mối quan hệ đã xác định sắp xếp hay phân kỳ.
* So sánh mô hình Lập kế hoạch Nội dung Thần kinh của chúng tôi (Viết tắt là NCP) so với hai mô hình bộ mã hóa-giải mã (ED) được trình bày trong Wiseman et al. (2017) với bản sao chung (JC) và bản sao có điều kiện (CC), tương ứng.
* NCP sử dụng kế hoạch nội dung tiên tri (OR) làm đầu vào, tạo ra một tài liệu bao gồm tám câu mẫu: một câu giới thiệu, sáu câu dành riêng cho người chơi và một kết luận kết án.
* NCP + CC đạt được lựa chọn nội dung tốt nhất và điểm sắp xếp nội dung của BLEU và so với hệ thống báo cáo tốt nhất ở Wiseman et al. (2017) đạt được sự cải tiến tuyệt đối về xấp xỉ 12% về mối quan hệ; Nội dung độ chính xác lựa chọn cũng cải thiện 5% và thu hồi 15%, thứ tự nội dung tăng 3% và BLEU tăng 1,5 điểm.
* Kết quả của hệ thống oracle (NCP + OR) cho thấy rằng việc lựa chọn và sắp xếp nội dung thực sự tương quan với chất lượng của kế hoạch nội dung và bất kỳ cải tiến nào trong thành phần lập kế hoạch sẽ mang lại kết quả tốt hơn.
* 84,5% bản ghi trong NCP + CC là không trùng lặp so với Wiseman và cộng sự. (2017) đạt 72,9% cho thấy mô hình của chúng tôi ít lặp lại hơn.
* So với toàn bộ hệ thống (NCP + CC), độ chính xác khi lựa chọn nội dung và thu hồi cao hơn (lần lượt là 4,5% và 2%) cũng như thứ tự nội dung (tăng 1,8%).
* Sử dụng các chỉ số CS và CO để đo lường mức độ văn bản được tạo ra tuân theo nội dung kế hoạch do người lập kế hoạch đưa ra cho thấy NCP + CC tạo ra các bản tóm tắt trò chơi theo sát kế hoạch nội dung: Độ chính xác CS cao hơn 85%, độ thu hồi CS cao hơn 93% và CO cao hơn 84%.
* NCP đạt độ chính xác cao hơn trong tất cả các chỉ số bao gồm tạo mối quan hệ, lựa chọn nội dung, sắp xếp nội dung và BLEU so với Wiseman etal. (2017).
* Đánh giá dựa trên con người: Sử dụng dữ liệu từ Amazon Mechanical Turk (AMT) cho thấy việc lập kế hoạch nội dung cải thiện việc tạo dữ liệu thành văn bản trên các chỉ số và hệ thống (NCP + CC nói chung hoạt động tốt nhất, tuy nhiên ở đó là một khoảng cách đáng kể giữa bản tóm tắt được tạo tự động và bản tóm tắt do con người tạo ra)

***KẾT LUẬN CUỐI CÙNG***:

Mô hình tạo dữ liệu thành văn bản được cải tiến với các mô-đun lập kế hoạch và lựa chọn nội dung tạo ra cải thiện cả về số lượng các dữ kiện liên quan có trong văn bản đầu ra và thứ tự theo đó chúng được trình bày trong đó quan trọng vẫn là cải tiến bổ sung về ngữ pháp và tính ngắn gọn của văn bản được tạo.