Report paper: Enriching Word Vectors with Subword Information

Giáp Ngọc Hiệu

**Tóm tắt nội dung**

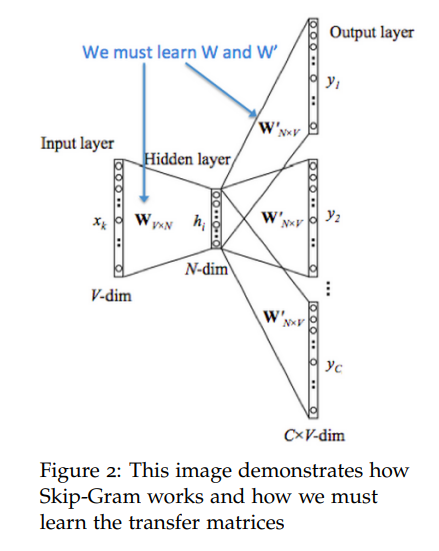
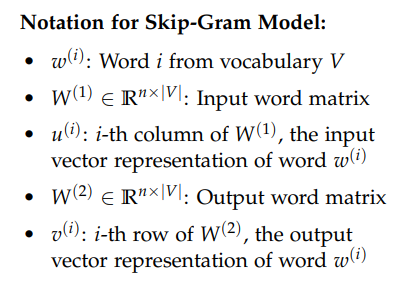
Bài report này là phần tóm tắt lại ý tưởng, nội dung chính, những điều làm được của paper Enriching Word Vectors with Subword Information

1. **Ý tưởng**

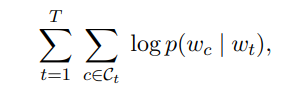
Các biểu diễn từ liên tục, được đào tạo trên các kho dữ liệu lớn không được gắn nhãn rất hữu ích cho nhiều tác vụ xử lý ngôn ngữ tự nhiên. Các mô hình phổ biến học cách biểu diễn như vậy bỏ qua hình thái của từ, bằng cách gán một vectơ riêng biệt cho mỗi từ. Đây là một hạn chế, nhất là đối với những ngôn ngữ có vốn từ vựng lớn, nhiều từ hiếm. Tác giả đề xuất một cách tiếp cận mới dựa trên mô hình skipgram, trong đó mỗi từ được biểu diễn dưới dạng một túi ký tự n-gram. Một vector đại diện cho mỗi ký tự n-gram; các từ được biểu diễn dưới dạng tổng của các vector này.

1. **Phương pháp**
   1. **General model**

Đề xuất của tác giả bắt nguồn từ mô hình continuous skip- gram.

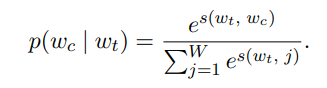
 

Cho trước một bộ từ vựng có kích thước W , trong đó một từ được xác định bởi chỉ mục của nó w ∈ {1, ..., W }, mục tiêu là học cách biểu diễn véc tơ cho mỗi từ w. Tổng quát, với một kho dữ liệu đào tạo lớn được biểu diễn dưới dạng một chuỗi các từ w1, ..., wT, mục tiêu của mô hình skipgram là tối đa hóa:

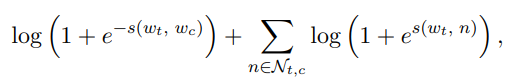


Trong đó là tập các index của các từ xung quanh từ . Xác suất một từ ngữ cảnh  với  cho trước sẽ được tham số hóa bằng cách sử dụng các word vectors. Giả sử ta có hàm  ánh xạ các cặp (từ, ngữ cảnh) thành độ tương quan.

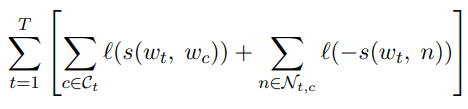
Một lựa chọn có thể để xác định xác suất của một từ ngữ cảnh là softmax:



Tuy nhiên, trong trường hợp này, với một từ , chỉ dự đoán một từ ngữ cảnh . Thay vào đó, vấn đề dự đoán các từ ngữ cảnh có thể được xem như một tập hợp các nhiệm vụ phân loại nhị phân độc lập. Sau đó, mục tiêu là dự đoán độc lập sự hiện diện (hoặc vắng mặt) của các từ ngữ cảnh. Đối với từ ở vị trí t, chúng tôi coi tất cả các từ ngữ cảnh là mẫu tích cực và các mẫu tiêu cực được lấy ngẫu nhiên từ từ điển. Đối với vị trí ngữ cảnh đã chọn c, sử dụng tổn thất logistic nhị phân, ta có negative log-likelihood:



trong đó  là một tập hợp các mẫu phủ định được lấy mẫu từ bộ từ vựng. Bằng cách biểu thị logistic loss function , ta có thể viết lại hàm mục tiêu dưới dạng:



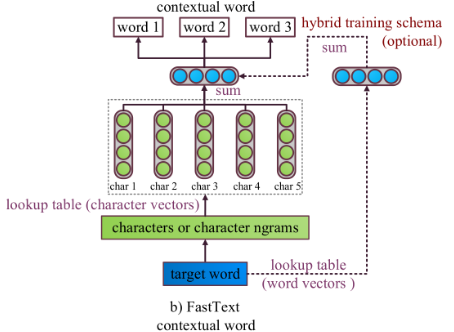
* 1. **Subword model**

Mỗi từ w được biểu diễn dưới dạng một túi ký tự n-gram. Tác giả thêm các ký hiệu ranh giới đặc biệt < và > ở đầu và cuối từ, cho phép phân biệt các tiền tố và hậu tố với các chuỗi ký tự khác. Tác giả cũng bao gồm chính từ w trong tập hợp n-gram của nó, để tìm hiểu cách biểu diễn cho mỗi từ (ngoài ký tự n-gram).

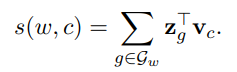
Lấy từ where và n = 3 làm ví dụ, nó sẽ được biểu thị bằng ký tự n-gram:

[<wh, whe, her, ere, re>, <where>]

Lưu ý rằng chuỗi <her> tương ứng với từ her khác với tri-gram her từ where. Trong thực tế, tác giả trích xuất tất cả n-gram cho n lớn hơn hoặc bằng 3 và nhỏ hơn hoặc bằng 6. Đây là một cách tiếp cận rất đơn giản và có thể xem xét các bộ n-gram khác nhau, ví dụ như lấy tất cả các tiền tố và hậu tố.



Giả sử có một từ điển n-gram kích thước G. Cho một từ w, tập n-gram xuất hiện trong w: Gw ⊂ {1, . . . , G}. Gọi  là vector ứng với mỗi n-gram g. Vector một từ bằng tổng các vector n-gram của nó.



Điều này cho phép chia sẻ cách biểu diễn giữa các từ, do đó cho phép học cách biểu diễn đáng tin cậy cho các từ hiếm.

Để giới hạn các yêu cầu bộ nhớ của mô hình, tác giả sử dụng hàm băm ánh xạ

n-gram thành số nguyên từ 1 đến K với . Cuối cùng, một từ được biểu thị bằng chỉ mục của nó trong từ điển từ và tập hợp n-gram được băm mà nó chứa.

* 1. **Optimization**

Tác giả giải quyết vấn đề tối ưu hóa bằng cách sử dụng Stochastic Gradient Descent với Linear Learning Rate decay và hàm mục tiêu negative log likelihood.

* 1. **Implementation details**

- Kích thước vectơ từ: 300

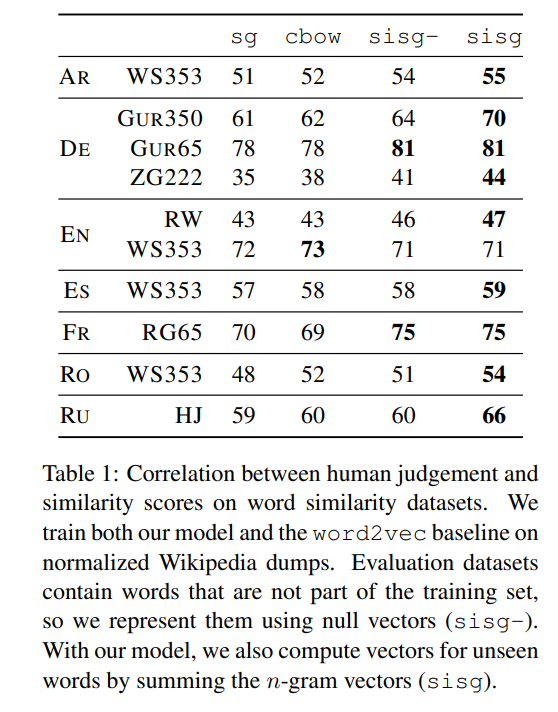
- Kích thước lấy mẫu âm tính: 5 mẫu trên mỗi mẫu dương tính

- Giữ lại những từ xuất hiện ít nhất 5 lần trong tập huấn luyện

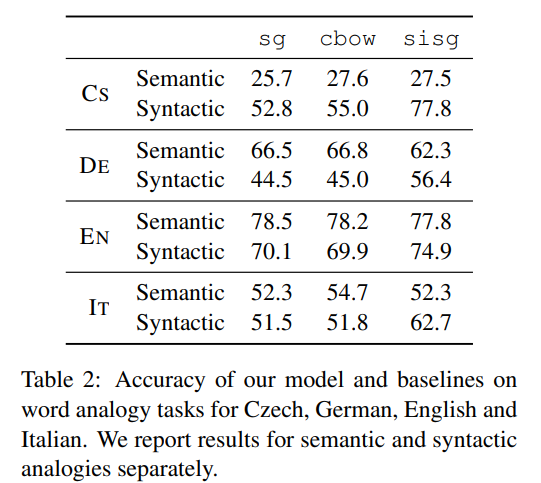
1. **Kết quả**

Datasets: Các tập dữ liệu lấy từ Wikipedia với 9 ngoonn gữ khác nhau: Arabic, Czech, German, English, Spanish, French, Italian, Romanian and Russian.

**Human similarity judgment** (Đánh giá sự tương đồng của con người): Tính hệ số tương quan hạng Spearman giữa đánh giá của con người và độ tương đồng cosine giữa các biểu diễn vector.



**Word analogy tasks:** Độ chính xác tương tự được chia thành Ngữ nghĩa (dựa trên ý nghĩa) và Cú pháp (dựa trên cú pháp và ngữ pháp)

****

1. **Note**

Ưu điểm của phương pháp:

+ Biểu diễn từ dựa trên từ vựng và các thành phần nhỏ hơn (subword) như các n-gram. Điều này giúp nắm bắt được thông tin ngữ nghĩa sâu hơn, đặc biệt đối với các từ không có trong từ điển.

+ Có thể xử lý tốt cho các trường hợp từ hiếm gặp, không có trong dataset. Vì các từ này vẫn có thể được chia thành các ký tự n-gram, chúng có thể chia sẻ các n-gram này với các từ phổ biến.

Nhược điểm:

+ Yêu cầu bộ nhớ cao vì mô hình này tạo ra vector từ từ các ký tự của nó chứ không phải từ các từ.

Ta có thể lấy word embedding của 1 từ bằng cách đưa từ đó vào mô hình. Từ sẽ được chia thành các n-gram với các vector tương ứng trong từ điển các n-gram. Đầu ra của lớp ẩn chính là word embedding.