NER, viết tắt của Nhận dạng thực thể có tên là một nhiệm vụ trong Xử lý ngôn ngữ tự nhiên (NLP) nhằm mục đích xác định và phân loại các thực thể trong văn bản, chẳng hạn như tên người, tổ chức, địa điểm, ngày tháng, v.v.

Phân loại là một phương pháp học máy có giám sát, trong đó mô hình cố gắng dự đoán nhãn chính xác của dữ liệu đầu vào nhất định. Trong phân loại, mô hình được đào tạo đầy đủ bằng cách sử dụng dữ liệu đào tạo, sau đó được đánh giá trên dữ liệu thử nghiệm trước khi được sử dụng để thực hiện dự đoán trên dữ liệu mới chưa thấy.

BERT (Bidirectional Encoder Representations from Transformers): Một mô hình học sâu phổ biến cho các tác vụ phân loại, đặc biệt là đối với NER.

BERT dựa trên kiến ​​trúc Transformer, được đào tạo trước theo cách song hướng để dự đoán một từ trong ngữ cảnh của nó. Nó được sử dụng rộng rãi cho NER do khả năng hiểu ngữ cảnh mạnh mẽ của nó.

- Mã hóa vị trí: Vì Transformer không có các lớp hồi quy hoặc lớp tích chập, nó không tự nhiên biết thứ tự của các token đầu vào. Do đó, cần có một cách để mô hình biết thông tin này, đó là nhiệm vụ của mã hóa vị trí. Sau các lớp embedding, tạo ra các embedding token, chúng ta thêm các vector mã hóa vị trí đại diện cho vị trí của mỗi từ trong câu.

- Lớp chuẩn hóa: Trong kiến trúc của sơ đồ, lớp "Add & Norm" đề cập đến lớp chuẩn hóa. Lớp này đơn giản là chuẩn hóa đầu ra của multi-head attention, cải thiện hiệu quả hội tụ.

- Kết nối dư: Kết nối dư là một khái niệm đơn giản của việc thêm đầu vào của một khối vào đầu ra của nó. Kết nối này cho phép xếp chồng nhiều lớp trong mạng. Trong sơ đồ, kết nối dư được sử dụng sau các khối FFN (Feed-Forward Network) và attention. Trong phần "Add" của "Add & Norm," nó đại diện cho kết nối dư.

- Khối Feed-Forward: Đây là một khối cơ bản, nơi sau khi thực hiện các tính toán trong khối attention ở mỗi lớp, khối tiếp theo là FFN. Bạn có thể hiểu rằng cơ chế attention giúp thu thập thông tin từ các token đầu vào, và FFN xử lý thông tin đó.

Word2vec, viết tắt của "Word To Vector," là một phương pháp giúp máy tính hiểu được mối quan hệ giữa các từ khác nhau bằng cách chuyển chúng thành các vector.

Word2vec không phải là một thuật toán duy nhất mà là một tập hợp các kiến trúc mô hình và các tối ưu hóa có thể được sử dụng để học các vector nhúng từ.

Hai phương pháp cơ bản và phổ biến nhất của Word2vec là CBOW và Skip-Ngram:

* **Continuous bag-of-words (CBOW)**: Dự đoán từ ở giữa dựa trên các từ trong ngữ cảnh xung quanh. Ngữ cảnh bao gồm một vài từ trước và sau từ hiện tại (ở giữa).
* **Continuous skip-gram (Skip-Ngram)**: Dự đoán các từ trong một khoảng nhất định trước và sau từ hiện tại trong cùng câu.

FastText là một phương pháp biểu diễn từ được phát triển bởi phòng nghiên cứu AI của Facebook (FAIR). FastText phát triển dựa trên Word2Vec bằng cách học các vector biểu diễn cho từng từ.

Điểm khác biệt của nó so với các phương pháp cơ bản của Word2vec là khả năng xử lý các ngôn ngữ giàu hình thái học (như tiếng Nhật, tiếng Ấn, tiếng Ả Rập,...) bằng cách coi chúng như các từ con có vector giống hoặc tương tự với từ gốc.