**Fine-tune BERT Model for Named Entity Recognition**

Named Entity Recognition is a major task in Natural Language Processing (NLP) field. It is used to detect the entities in text for further use in the downstream tasks as some text/words are more informative and essential for a given context than others. It is the reason NER is sometimes referred to as Information retrieval, as extracting relevant keywords from the text and classifying them into required classes.

BERT stands for Bidirectional Encoder Representations from Transformers. It is a famous transformer in the field of NLP. This transformer is a pre-trained transformer like the others.  The training for this transformer is performed by deep bidirectional representation from the unlabeled text by jointly conditioning on both left and right context and using the data from English Wikipedia(2500 M words) and wordsBooksCorpus.

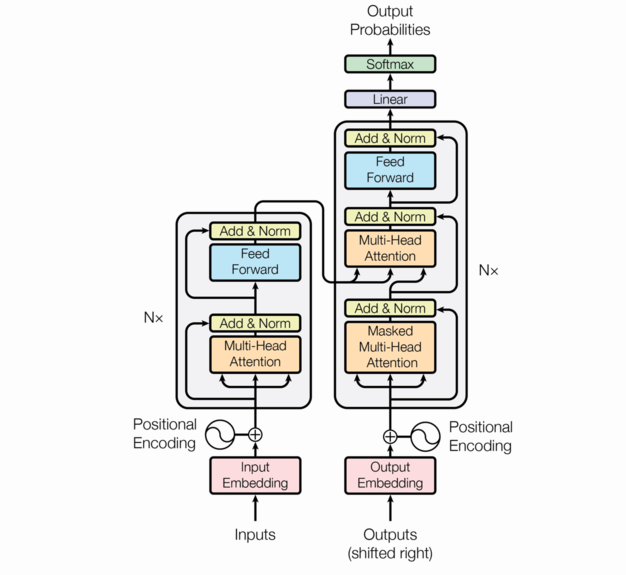
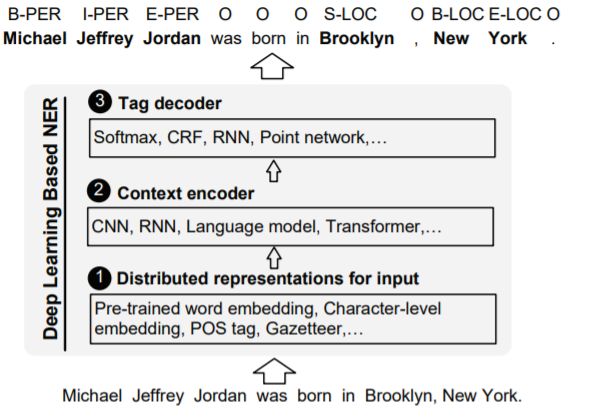


Figure 1: BERT Architecture

Named entity recognition (NER) uses a specific annotation scheme, which is defined (at least for European languages) at the word level. An annotation scheme that is widely used is called IOB-tagging, which stands for Inside-Outside-Beginning. Each tag indicates whether the corresponding word is inside, outside or at the beginning of a specific named entity. The reason this is used is because named entities usually comprise more than 1 word.

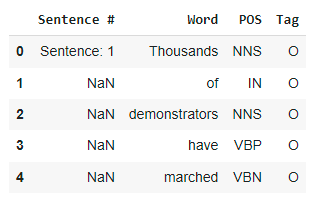
Let's have a look at an example. If you have a sentence like "Barack Obama was born in Hawaï", then the corresponding tags would be [B-PERS, I-PERS, O, O, O, B-GEO]. B-PERS means that the word "Barack" is the beginning of a person, I-PERS means that the word "Obama" is inside a person, "O" means that the word "was" is outside a named entity, and so on. So one typically has as many tags as there are words in a sentence.

So if you want to train a deep learning model for NER, it requires that you have your data in this IOB format (or similar formats such as BILOU). There exist many annotation tools which let you create these kind of annotations automatically (such as Spacy's Prodigy, Tagtog or Doccano). You can also use Spacy's biluo\_tags\_from\_offsets function to convert annotations at the character level to IOB format.

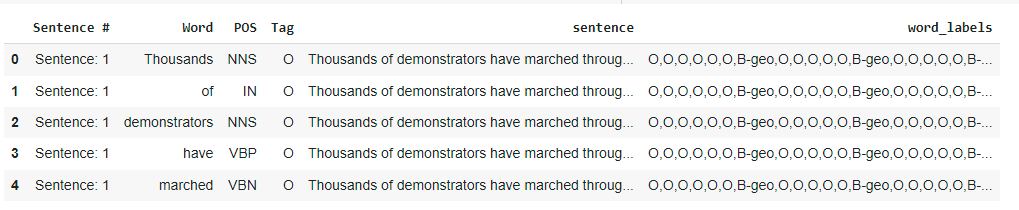


Dataset Example;

Must be labelled like



For each word in a sentence, we have tag



Simply Train a Model

Complete Code along with steps available in this Jupyter Notebook

<https://colab.research.google.com/github/NielsRogge/Transformers-Tutorials/blob/master/BERT/Custom_Named_Entity_Recognition_with_BERT_only_first_wordpiece.ipynb#scrollTo=ahwMsmyG5ZPE>

The Idea is basically doing Transfer Learning, we will take pre-trained BERT NER model (hugging face provide best possible model for this), and will fine tune with our own dataset having customer tags.

Other Resources

<https://www.kaggle.com/datasets/abhinavwalia95/entity-annotated-corpus>

SPACY VERSION NER <https://newscatcherapi.com/blog/train-custom-named-entity-recognition-ner-model-with-spacy-v3>