Part-of-Speech Tagging with Word Embeddings

CS 9875 Final Project

Paul Moore Juhani Dickinson

December 2, 2025

Outline

Part-of-Speech Tagging

Parts of Speech

- Each word in a sentence carries out a syntactic role: denote an object, denote an action, modify an object, etc.
- This role is mapped to the word using a part-of-speech (PoS) tag

Spelled the Same, Used Differently: Syntactic and Semantic Differences

Words can be different parts of speech (do different things) depending on where they are in the sentence and what is around them

1. Noun-Verb: "spot"
   - "Your nose has a spot on it." (N)
   - "Can you spot him?" (V)
2. Noun-Adjective: "red"
   - "The red I like is that one." (N)
   - "The red car has arrived." (A)

Tagging

Given a sentence, assign a PoS tag to each word in the sentence.

Can you spot the large spot on your nose ?

Tagging

Given a sentence, assign a PoS tag to each word in the sentence.

Can	you	spot	the	large	spot	on	your	nose	?
VB	D	VB	D	Adj	N	PP	D	N	.

Tagging

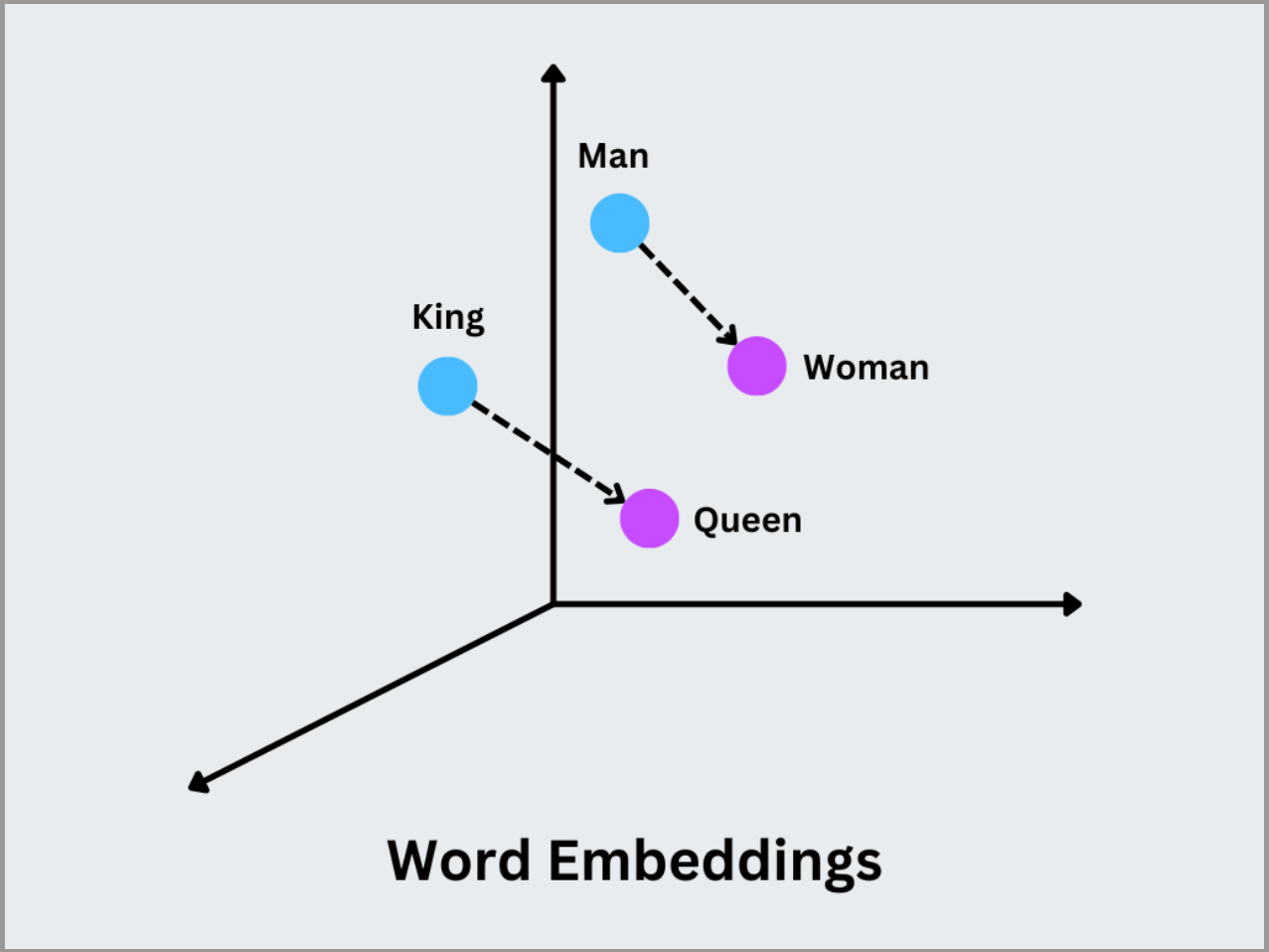
Given a sentence, assign a PoS tag to each word in the sentence.

Can	you	spot	the	large	spot	on	your	nose	?
VB	D	VB	D	Adj	N	PP	D	N	.

What makes PoS tagging different from a standard classification task: context is very important.

Word Embedding Models

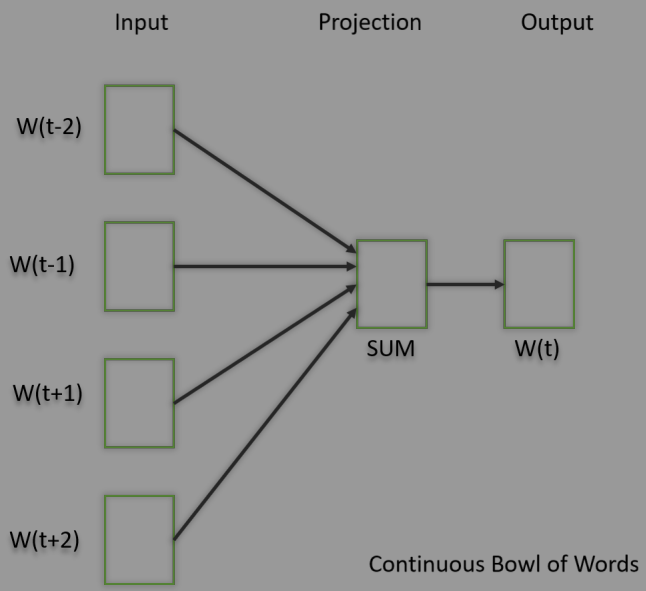
What are word embeddings?



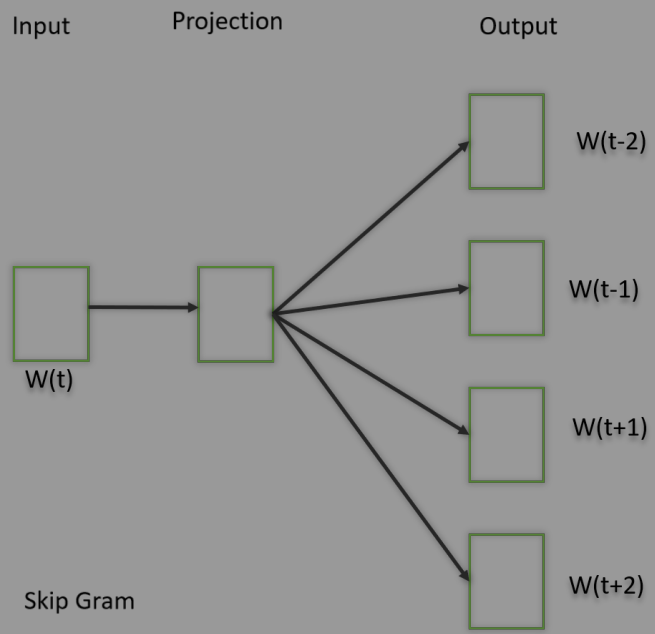
Static Word Embeddings

What are static word embedding models?

Word2Vec



CBOW



Skip Gram

Word2Vec Training Procedures

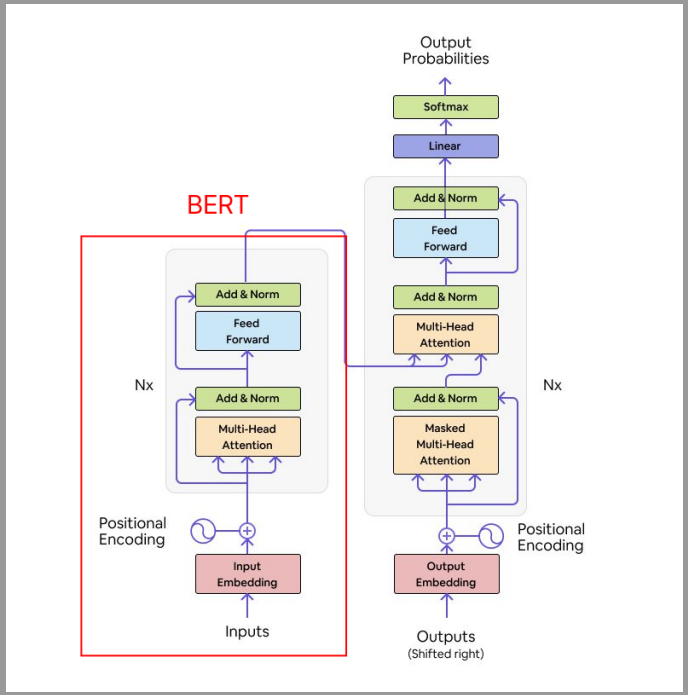
Word2Vec

$$\begin{bmatrix} W_{00} & W_{01} & W_{02} & \dots \\ W_{10} & W_{11} & W_{12} & \dots \\ W_{20} & W_{21} & W_{22} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

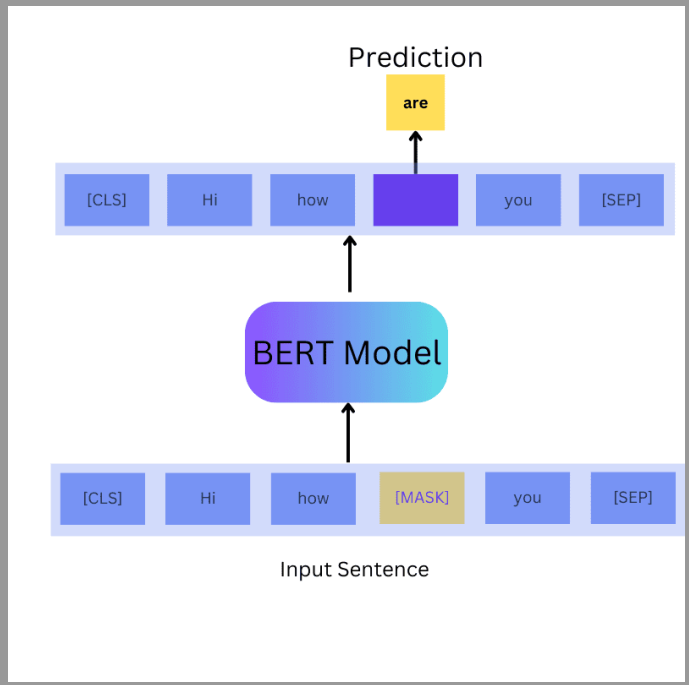
Contextual Word Embeddings

What are contextual word embeddings?

BERT

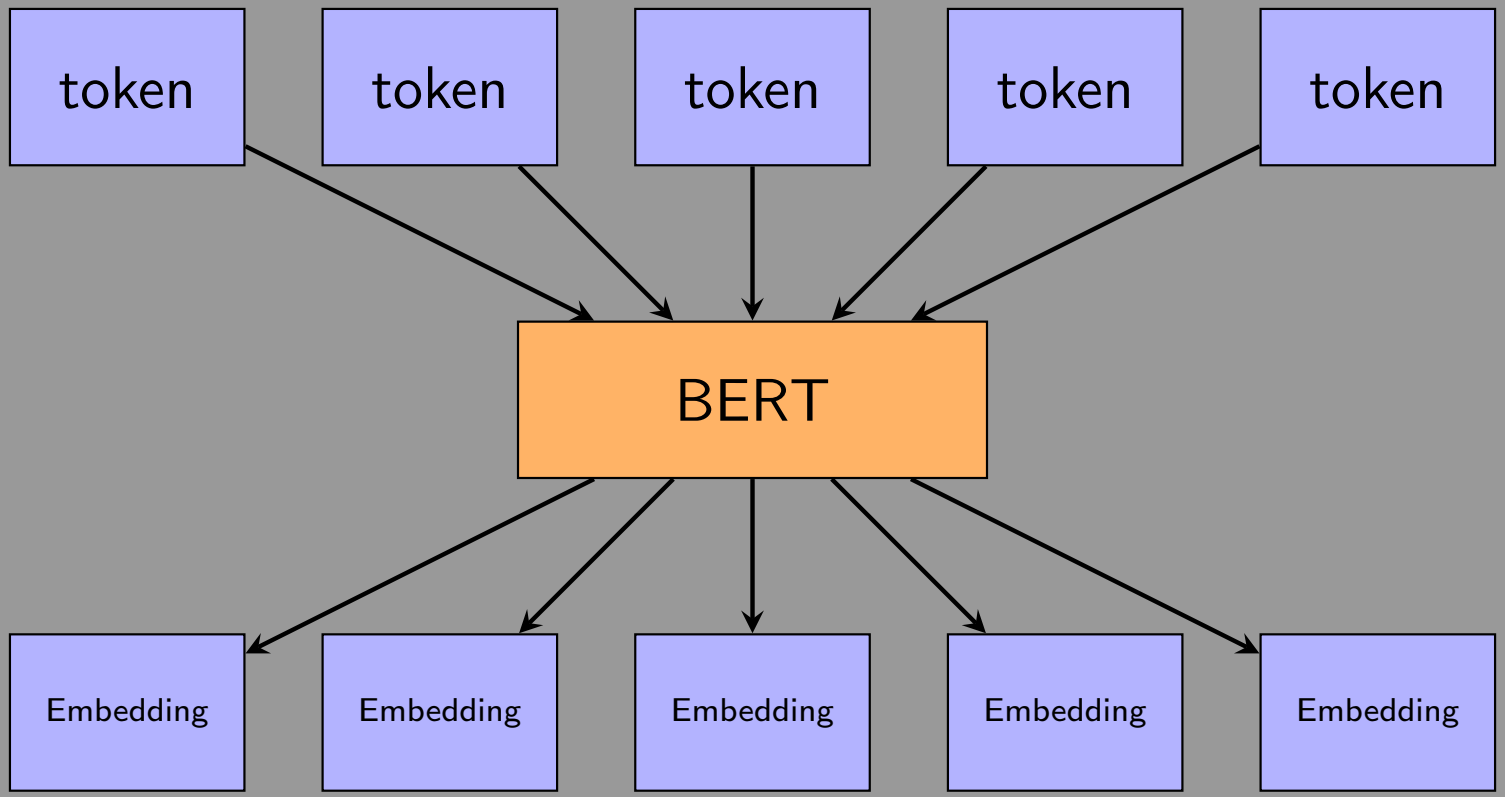


Transformer architecture



MLM

BERT


token
token
token
token
token
BERT
Embedding
Embedding
Embedding
Embedding
Embedding

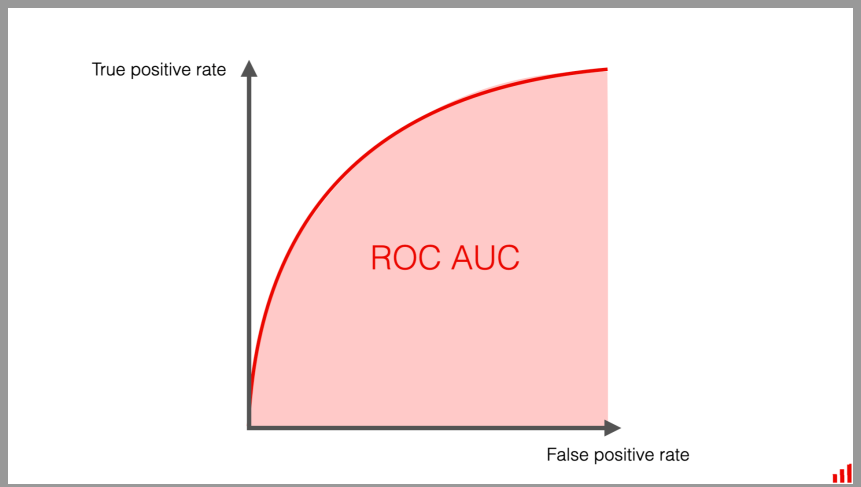
How do we use these embeddings?

How do we use these embeddings?
As input for a downstream model: SVM, Boosting, CNN, etc.

Evaluation and Theoretical Analysis

Evaluation Metrics

We will evaluate each of our models using F1 score and AUC-ROC adapted for this multi-class classification problem.



AUC-ROC

Dataset Used

Brown Corpus:

- 1 million tagged words, validated by humans
- American English
- Provided by the Natural Language ToolKit (NLTK)

Theoretical Analysis

We will conduct a theoretical analysis on the word embedding models, and the downstream classifier models.

Anticipated Results

We anticipate achieving near perfect scores in both evaluation metrics using contextual embeddings from BERT and a simple classification model like an SVM.