Makine Öğrenmesi

NLP: Tokenization, TF-IDF, Word2Vec

Orijinal not: Ders-10.pdf

UYARI Bu belge, el yazısı ders notlarından Claude (Anthropic) yardımıyla dijitale aktarılmıştır. Yardımcı kaynak olarak hazırlanmıştır; hata veya eksik içerebilir. Lütfen orijinal notları ve ders kaynaklarını esas alın.

NOTICE *This document was transcribed from handwritten lecture notes with the help of Claude (Anthropic). It is meant as a supplementary resource and may contain errors or omissions. Always check against the original notes and official course material.*

orj. s. 1

NLP (Natural language processing)

Tokenization

My$_{token}$ name is Mehmed.?

Morfolojik gitmek istersek (şekle göre)

Benim adım Mehmet.

Ben-im ad-ım

↑ kelimelere bölerek word piece

→ Morfolojik gramere göre

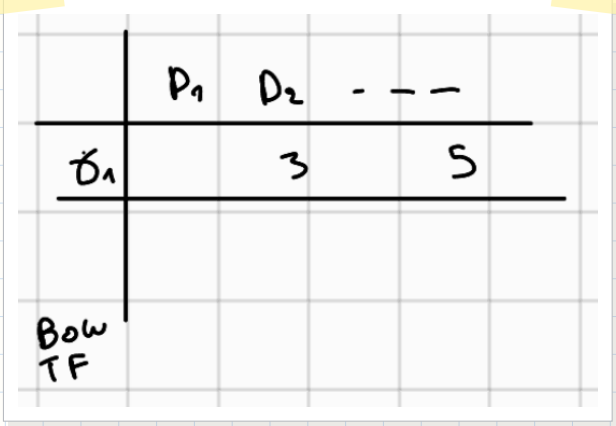
Türkçede 10000 kelime

10 000

ahmed → 2000

one hot encoding $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$

Elimizde dokümanlar var



orijinal çizim, s. 1

TF
Term frequency

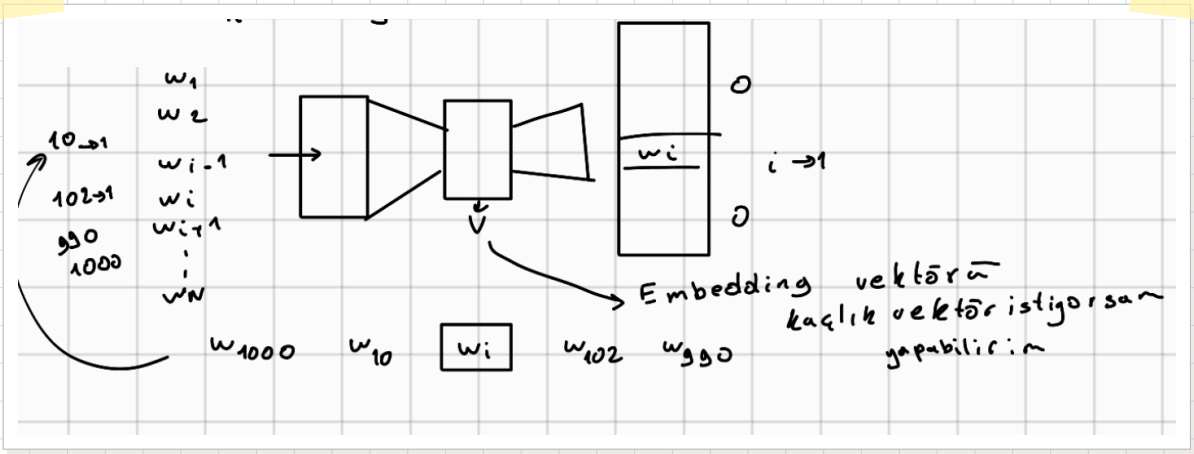
Tüm dokümanlarda iki kelime aynı frekansla geçmemeli

TF-IDF → baskınlığı azaltmak için

Word2Vec

CBOW

Continuous Bag of Words



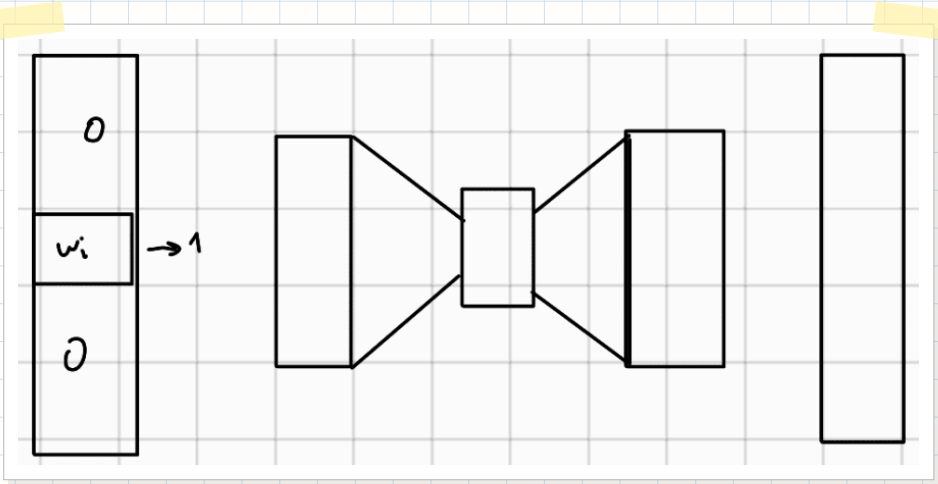
orijinal çizim, s. 1

orj. s. 2

Skip gram / n-gram

2 gram iki komşuluk

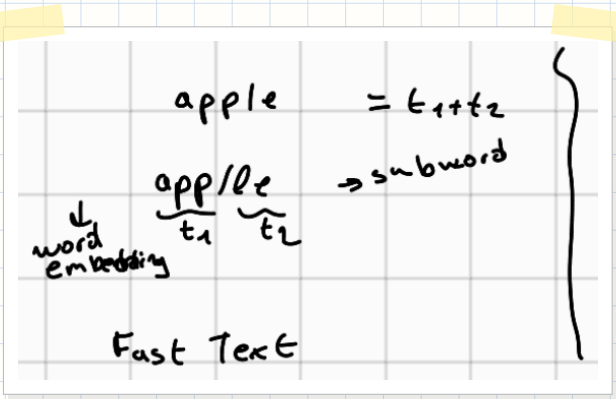
tersten de yapabilirim



orijinal çizim, s. 2

Global vectors for word Representation (Glove)

$$J = \sum_{ij} f(x_{ij}) \left(\underbrace{w_i^T}_{\text{bunu bulmaya çalıştığımız kelimenin ağırlığı}} \tilde{w}_j + b_i + \tilde{b}_j - \underbrace{\log X_{ij}}_{\text{i'inci doküman j'inci kelimenin geçme frekansı}} \right)^2$$



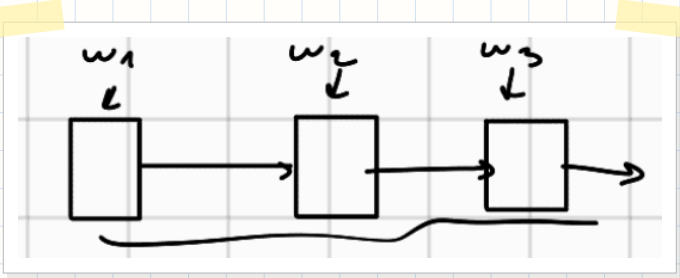
orijinal çizim, s. 2

Static embedding

Covid geldi yeni kelime

Benim adım Mehmet / Seq 2Seq

↙ Supervised



orijinal çizim, s. 2

encoder

RNN → exploding vanishing unutma fazla ?hatırlama

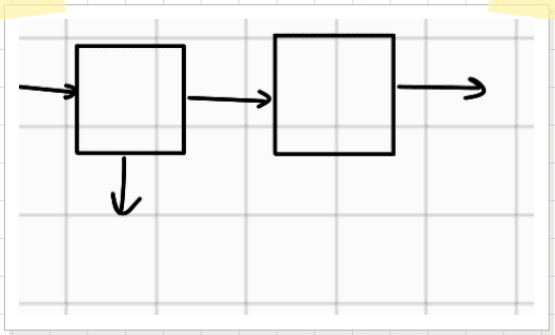
My name is Mehmet

$\tilde{w}_1$ $\tilde{w}_2$ $\tilde{w}_3$ $\tilde{w}_4$

sonraki güncellemede LSTM

Bu metin kısa olmadıktan sonra context'i yakalayamıyor.

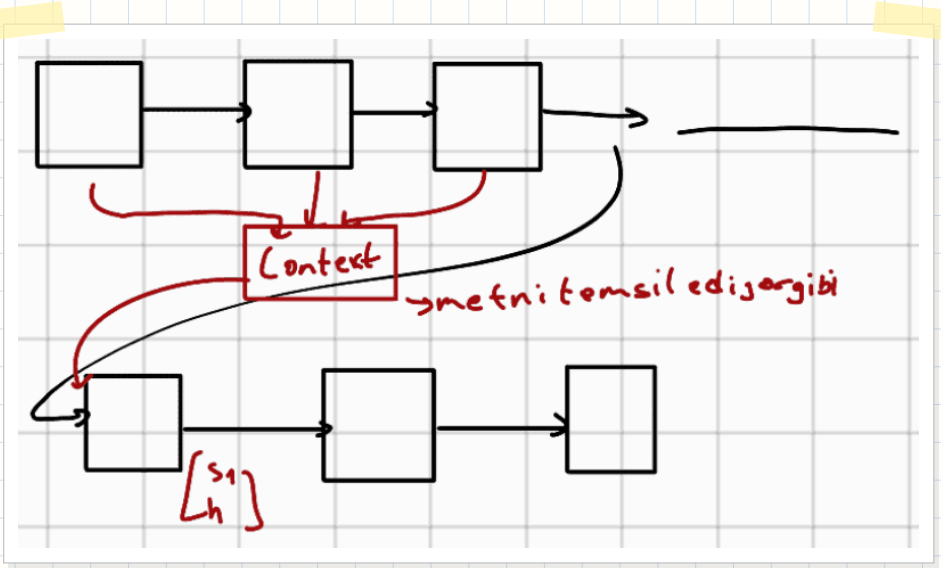
decoder



orijinal çizim, s. 2

orj. s. 3

Attention Mechanism



orijinal çizim, s. 3

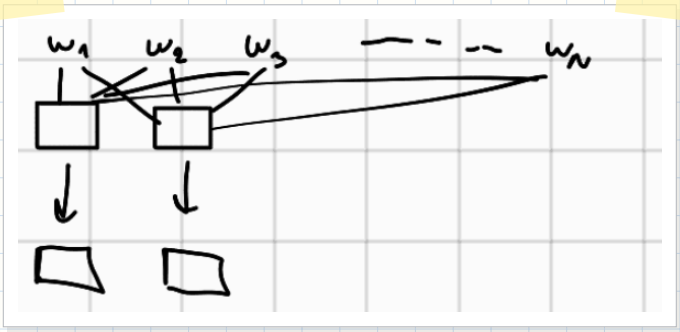
kelimeleri hidden layerle çarpıyoruz

softmax ↓ $w \cdot h$

$[\alpha_1, \alpha_2, \ldots, \alpha_{100}]$

hangi kelime ne kadar önemli

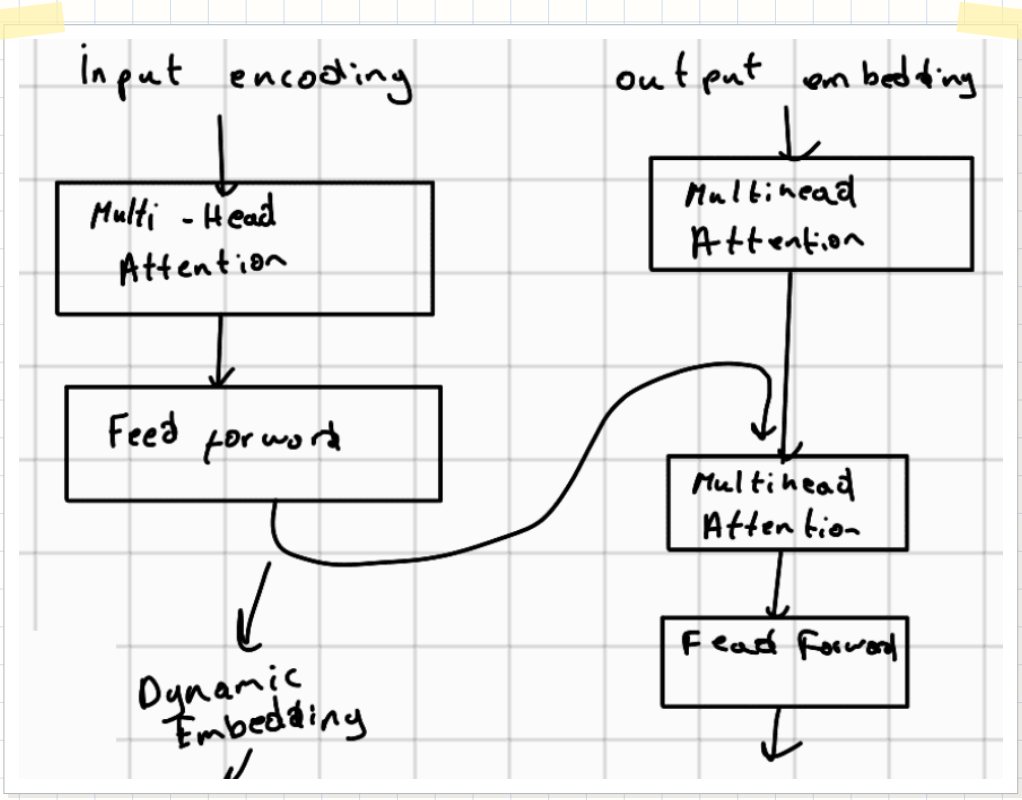
sonra dönmek yerine MLP koyuyoruz



orijinal çizim, s. 3

paralelde çalışabiliyor

Encoding: Decoder: GPT sadece decoder



orijinal çizim, s. 3

ali kopu at
a/li/ /ko/pu/ /at/
arada hataları düzeltmek için token
? sığıyor

2019 BERT: Sadece Encoder , kelimeleri ayırabilirsin

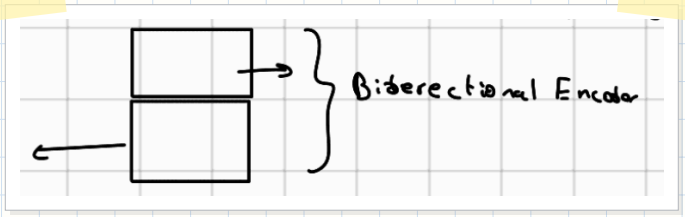
BERT NLP'nin YOLO'su

BERT : Bidirectional Encoder Representations from Transformers

orj. s. 4

ELMO: Embeddings from language model

context mekanizması yok



orijinal çizim, s. 4