Natural Language Processing

Mohammad Amin Samadi
Mohammad Sadegh Akhondzadeh

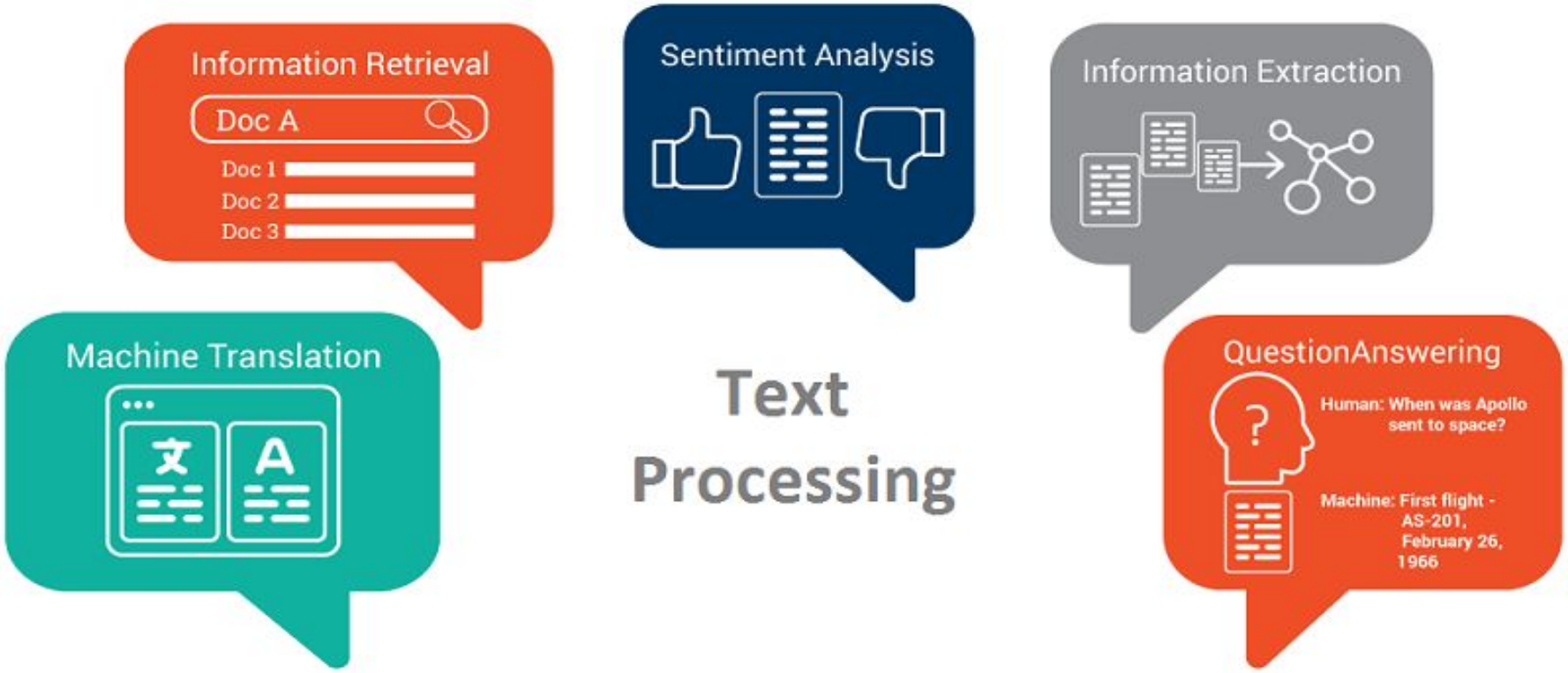
What is Natural Language Processing

- A subfield of linguistics, Artificial Intelligence and Cognitive Sciences
- An interdisciplinary subject
- Aim: To build intelligent computers that can interact with human being like human beings

Why NLP

- Huge amounts of data Internet = at least 20 billions pages.
- Text data: web sites, blogs, tweets.
- Audio data speech.

Applications



Why NLP is HARD

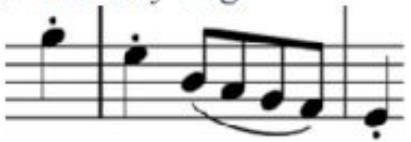
Cross Language Problems

- Ambiguity
- Compression
- Example sentiment analysis

Problem of other languages

- German:Donaudampfschiffahrtsgesellschaft skapitän (5 “words”)
- Chinese: 50,000 different characters.
- Japanese: 3 writing systems

Examples of sequence data

Speech recognition		"The quick brown fox jumped over the lazy dog."
Music generation		
Sentiment classification	"There is nothing to like in this movie."	★☆☆☆☆
DNA sequence analysis	AGCCCCTGTGAGGAACTAG	AGCCCCTGTCAGCAACTAG
Machine translation	Voulez-vous chanter avec moi?	Do you want to sing with me?
Video activity recognition		Running
Name entity recognition	Yesterday, Harry Potter met Hermione Granger.	Yesterday, Harry Potter met Hermione Granger.

Sequence in Language

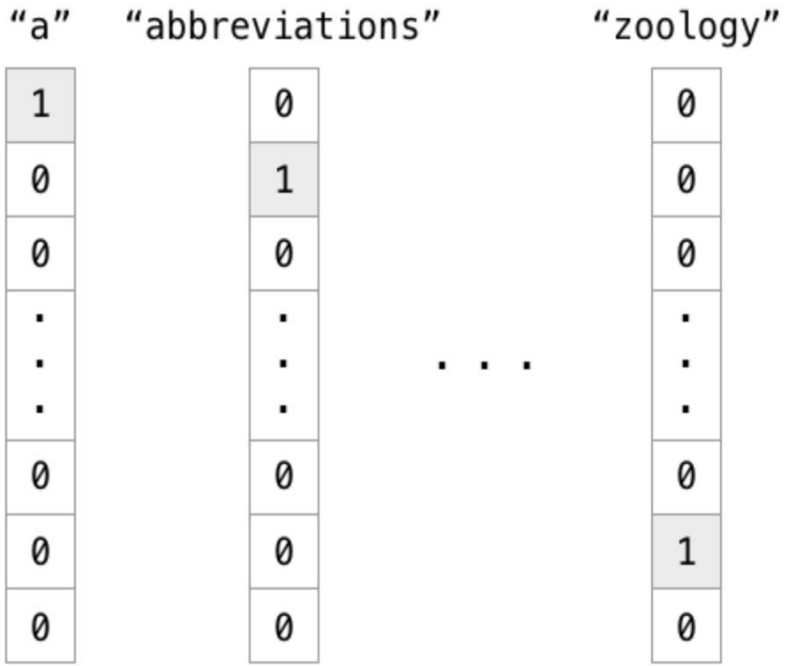
Word Representation

How to express words to a computer?

One-hot Representation

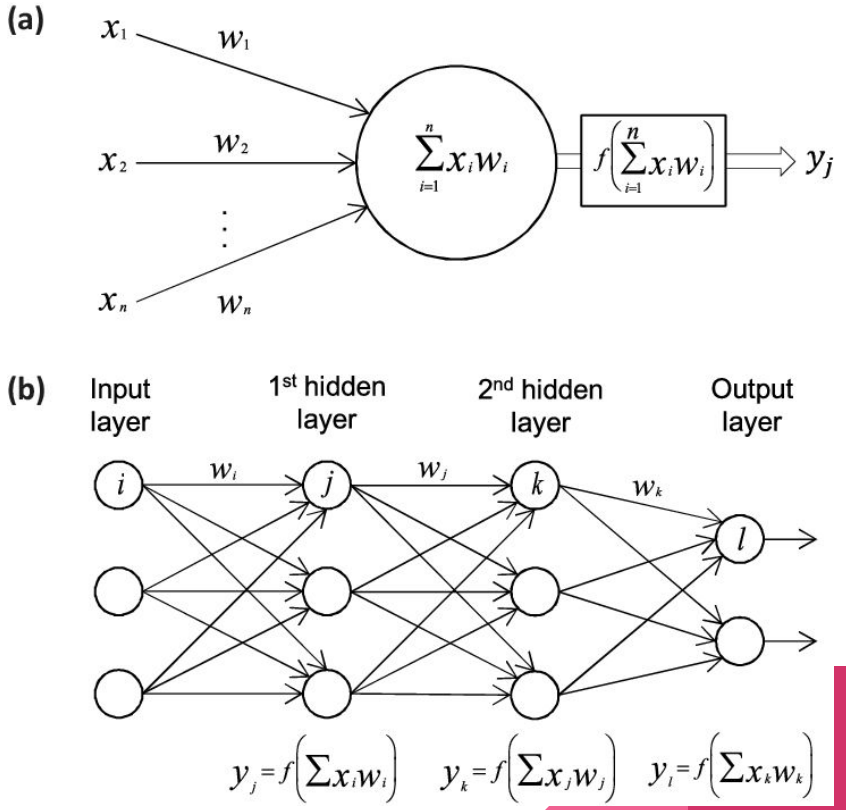
Problems :

- Meaning relations and similarities ignored
- Inefficient Memory consumption
- Vector size dependant on vocabulary size



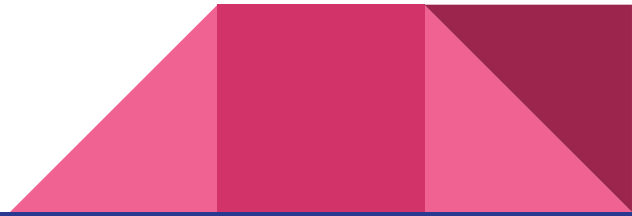
Neural Network

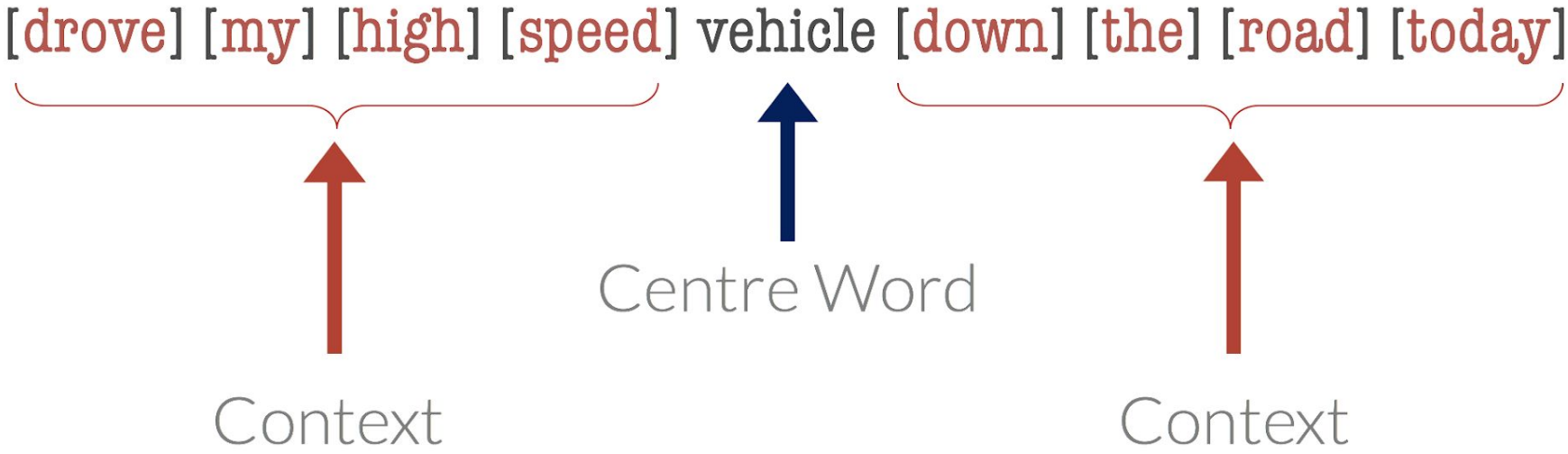
- Based on human's brain and nervous system
- Approximates a function from input to output



Word embedding

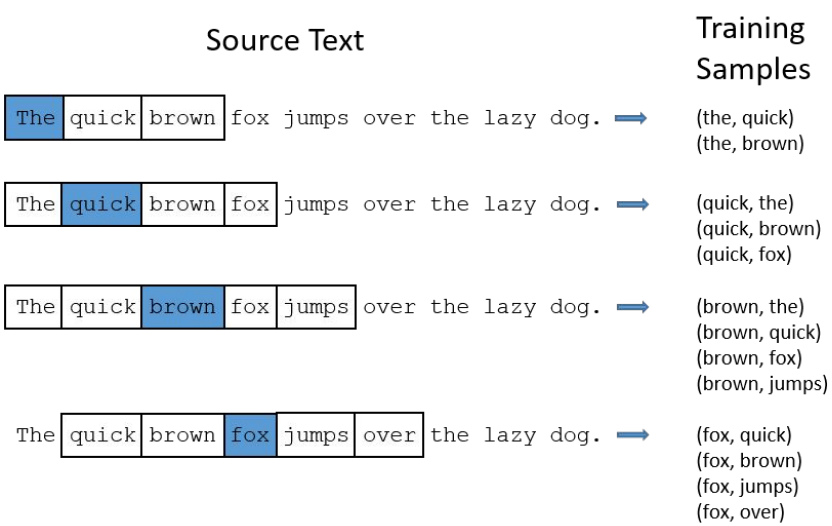
- Featurized representation of words
- Each words is embedded to a vector in a 100 or 200 or … dimension space
- Learned from large text corpus (1-100B words)
  - We can train it
  - Download pre-trained
- Helps us learn a context with fewer examples

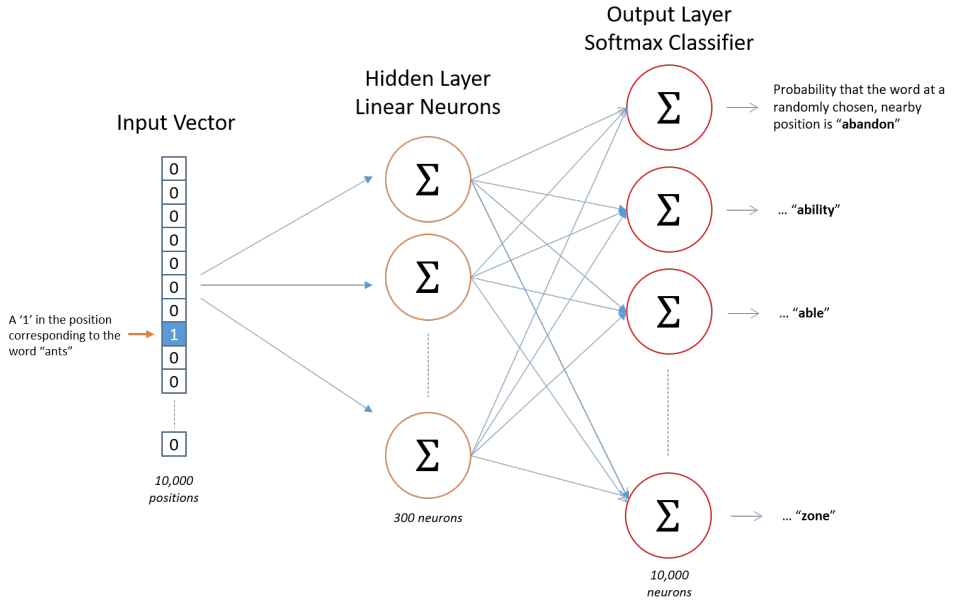



[drove] [my] [high] [speed] vehicle [down] [the] [road] [today]
Centre Word
Context
Context

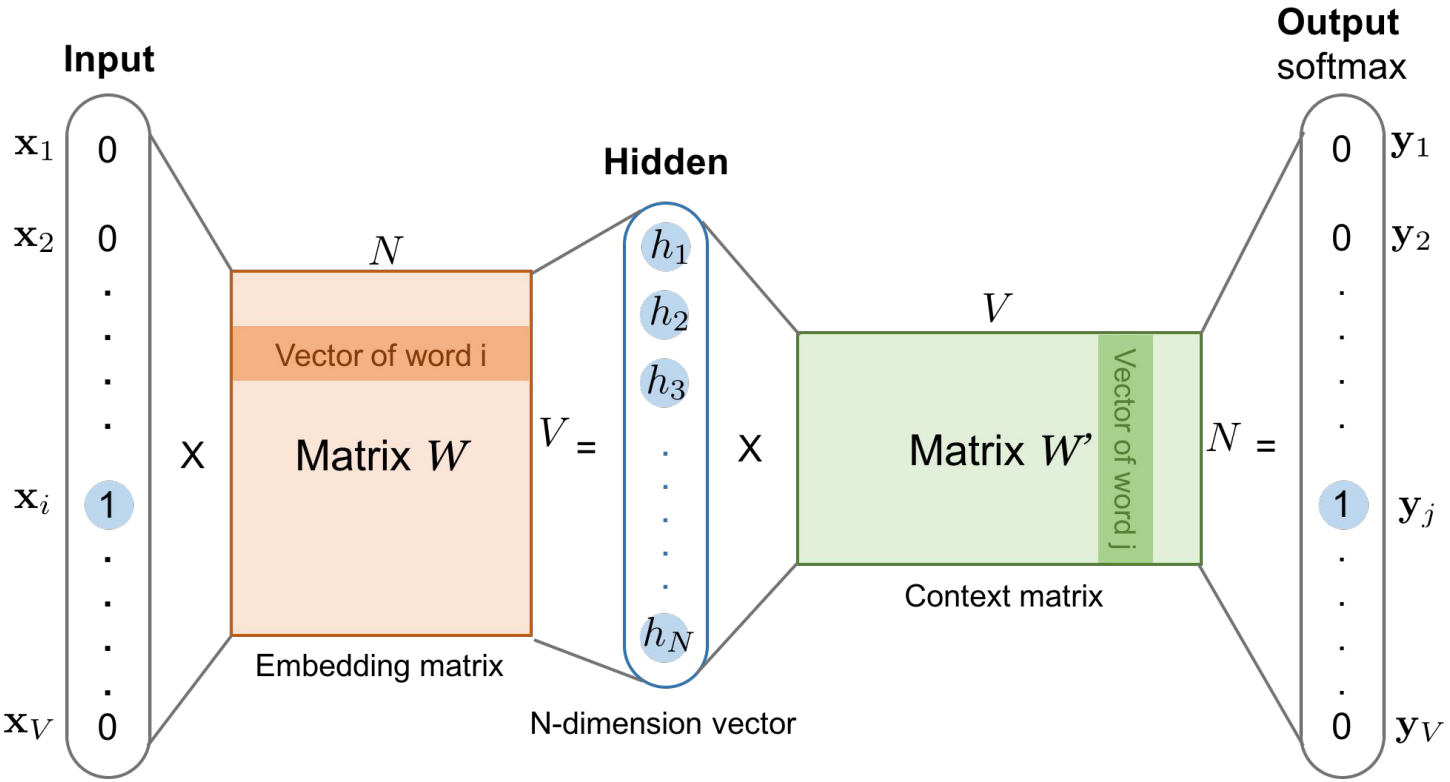
Skip-gram

- Fake task explanation
- Tries to predict the context based on the center word



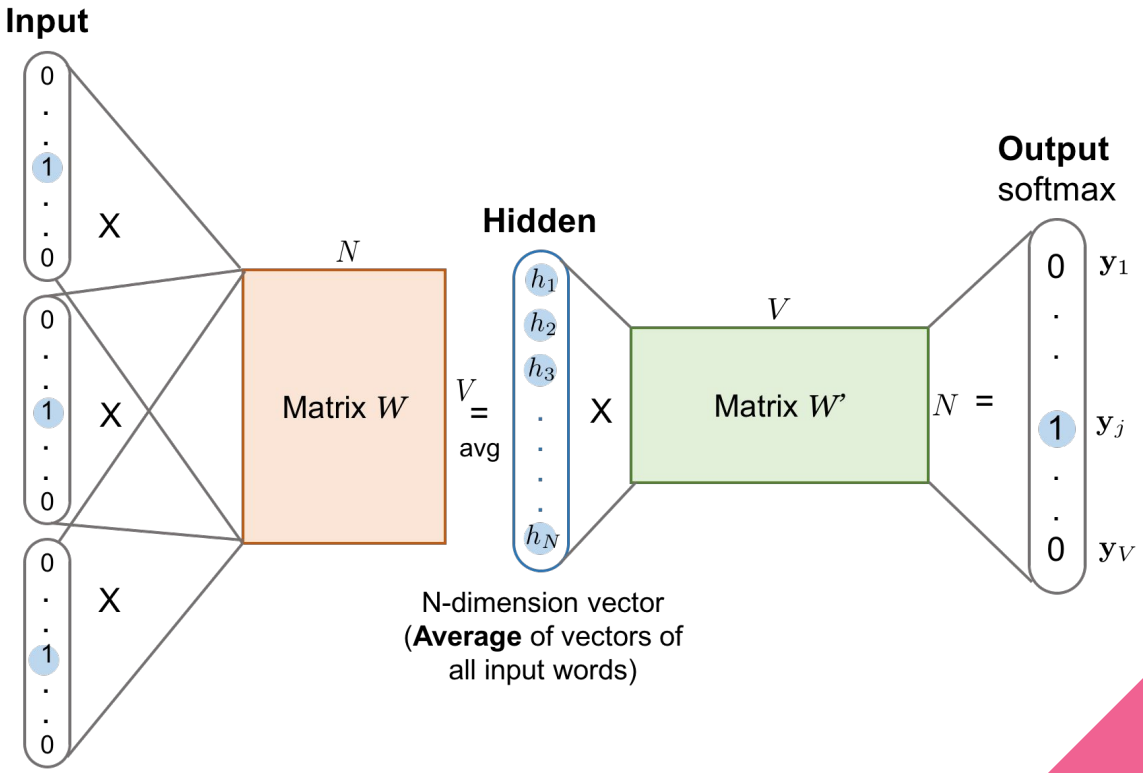


Representation Learning (Skip-gram)



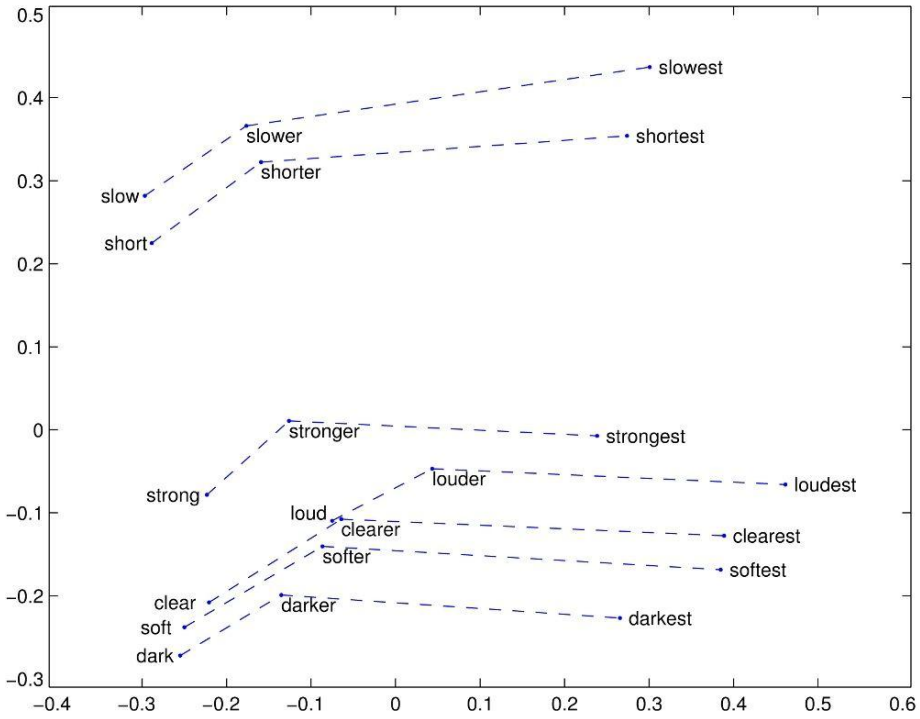
Representation Learning (CBoW)

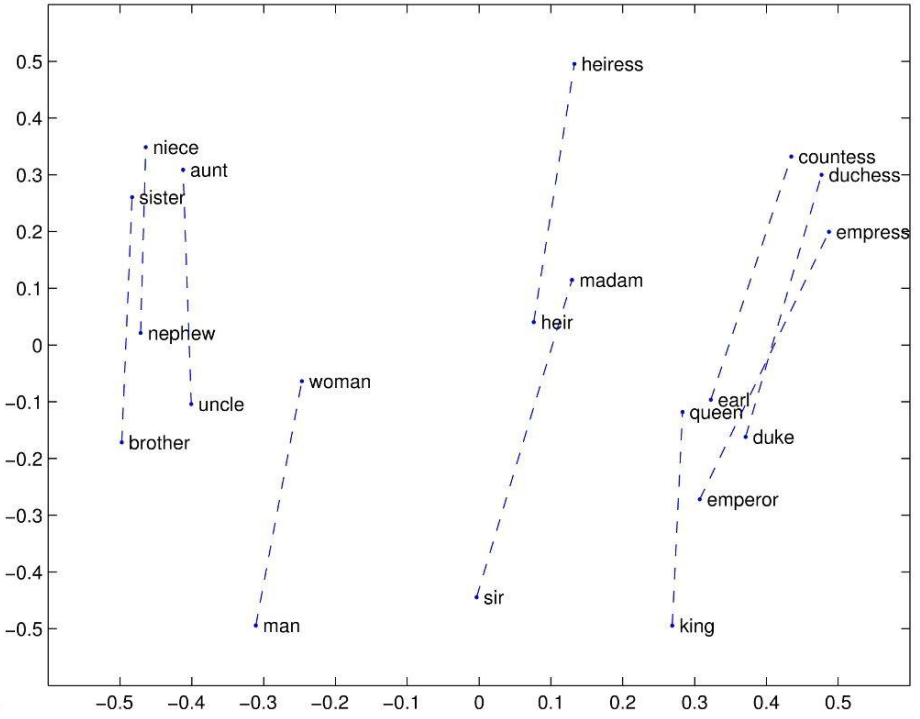
Tries to predict the center word based on the context



Vector offset

can encode semantic and syntactic relations between words





Dimension Reduction