An Overview of Natural Language Processing

From Word2Vec To GPT

Jaihua Yen

March 11, 2023

QNAP

Table of contents

Introduction

Generative AI

What is natural language processing

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on the interaction between computers and humans using natural language. The goal of NLP is to enable computers to understand, interpret, and generate human language.

NLP involves the use of techniques from computer science, linguistics, and machine learning to process, analyze, and generate natural language. Some common applications of NLP include sentiment analysis, language translation, text classification, chatbots, and speech recognition.

NLP is a rapidly evolving field with new developments and advancements being made regularly. As computers become better at processing language, the potential applications for NLP continue to expand, making it an important field of study in both industry and academia.

Applications in NLP

- Natural Language Processing (NLP) is now maturely developed in many tasks. For example, we can classify documents into several caegories for as to search files easily. Moreover, we can generate text by generative AI such as GPT series and ChatGPT.
- In general, NLP is used in the following several scenarios:
  1. Named-entity recognition
  2. Document Classification
  3. Machine Translation
  4. ChatBot
  5. Text Semantic Analysis
  6. Text Generation

TF-IDF

Word2Vec

Issues in Word Representation

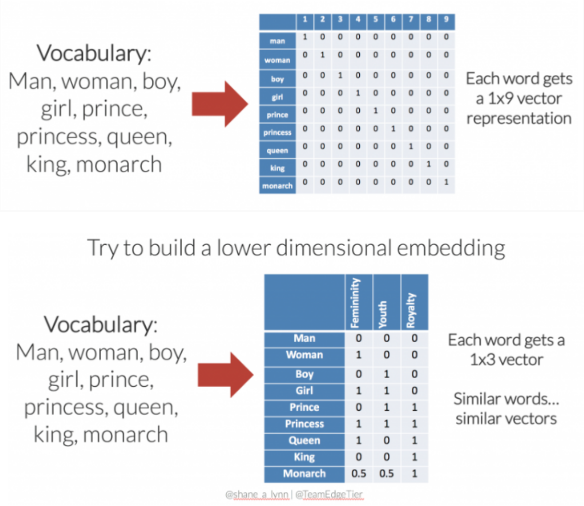
- Words are often represented as one-hot encoding in computer.
- For example, we can set hotel and motel to two different representative as one-hot encoding:

$$v_{motel} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \; v_{hotel} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

Issues in Word Representation

- However, there are several issues when words are represented as one-hot encoding:
    1. We cannot gain information of relation between words.
        - We have no idea how motel and hotel are relate to each other while those two vectors are orthogonal! i.e. $\langle v_{motel},\ v_{hotel} \rangle = 0$
    2. The vector would be very sparse if there exists large amount of unique words.
- That's why we introduce word embedding approach to tackle these problems. [1]

Word Embedding

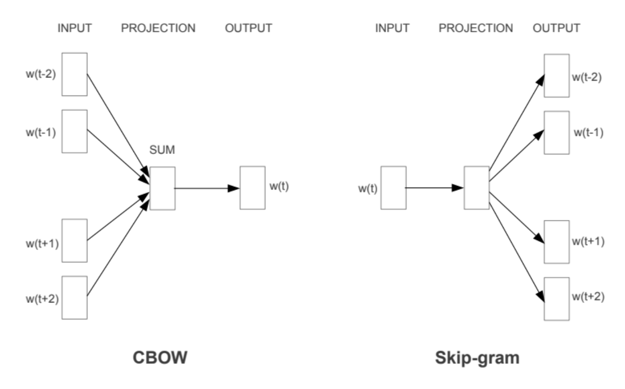


	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

	Femininity	Youth	Royalty
Man	0	0	0
Woman	1	0	0
Boy	0	1	0
Girl	1	1	0
Prince	0	1	1
Princess	1	1	1
Queen	1	0	1
King	0	0	1
Monarch	0.5	0.5	1

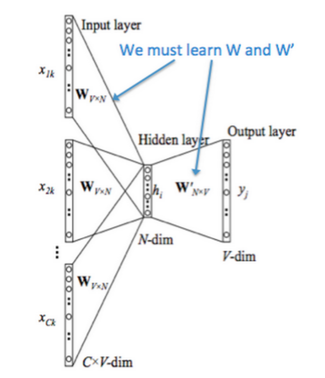
[Image Source]

Word2Vec



[Image Source]

CBOW



[Image Source]

- Use probability $P(y_i|x_{1k}, x_{2k}, ..., x_{Ck})$ to learn the weight matrix W!
- W is used to be the pre-trained model when we transform the words into embedding vectors in the unseen documents.

Transformer

BERT

Conclusion

Questions?

References i

T. Mikolov et al.
Efficient estimation of word representations in vector space.
arXiv preprint arXiv, 1301.3781, 2013.