Lecture - Information Retrieval using LLMs

August 14, 2024

SPEAKER: DR. KUSHAL SHAH
Professor, CS | Sitare University

SUBMITTED BY: SHUCHIKA SHARMA
Roll No: 24901325

INTRODUCTION

Some important terms were discussed by Dr. Kushal Shah. These included -

- NLP - It is a subset of Artificial Intelligence which processes large amounts of human language data. NLP is an end-to-end process between the system and humans, from understanding the information to making decisions while interacting
- NLU - Part of NLP whose objective is to understand natural language. Eg - The algorithm to understand the phrase ‘I live in Jalandhar’
- NLG - Stands for Natural Language Generation to generate new sentences in a grammatically correct fashion. Eg - Chatgpt generates an answer to a question

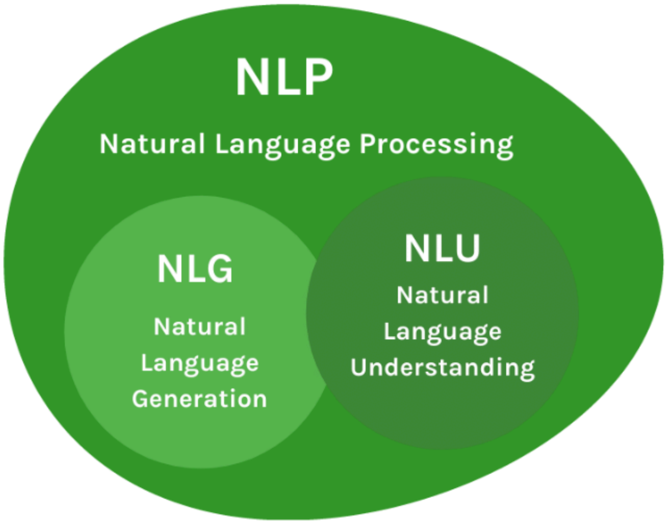

NLP
Natural Language Processing
NLG
Natural Language Generation
NLU
Natural Language Understanding

Figure 1 - Venn diagram for NLP, NLG and NLU

APPLICATIONS OF NLP

1. Information Retrieval - Eg. When you search in Google, it retrieves and displays some links
2. Question Answering - Eg. Chatbots to help answer user questions
3. Machine translation - Eg. Translation systems of French universities help non-french speakers understand the data on their website by converting it to English

4. Text Summarization - Eg. Certain NLP models exist which take a research paper as input and give a text summary.
5. Spam filtering - Eg. Gmail moves certain mails to spam box
6. Named entity recognition
7. Automated reasoning
8. Generative AI

INFORMATION RETRIEVAL

In the above example of Google, instead of 1 million links for our search query, we want to get the best 3 links to help us get the links with the best information. Google algorithm uses classical NLP techniques like document ranking/keyword matching.

EXAMPLE OF INFORMATION RETRIEVAL SYSTEM

Task - A user query regarding ML needs to be processed and answered.

Below is the procedure on how the information retrieval system will work -

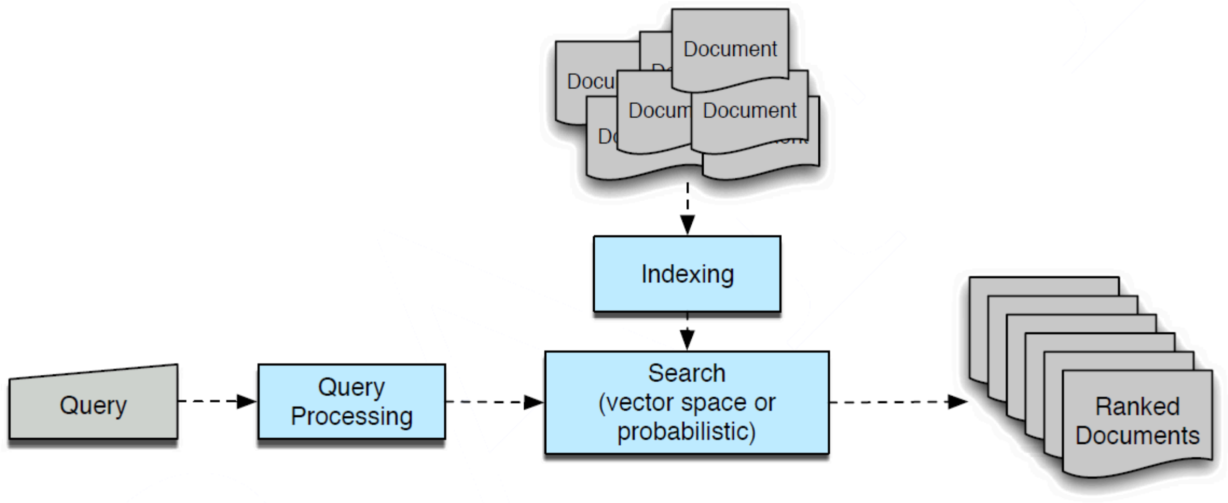


Figure 2 - Working of an Information Retrieval System

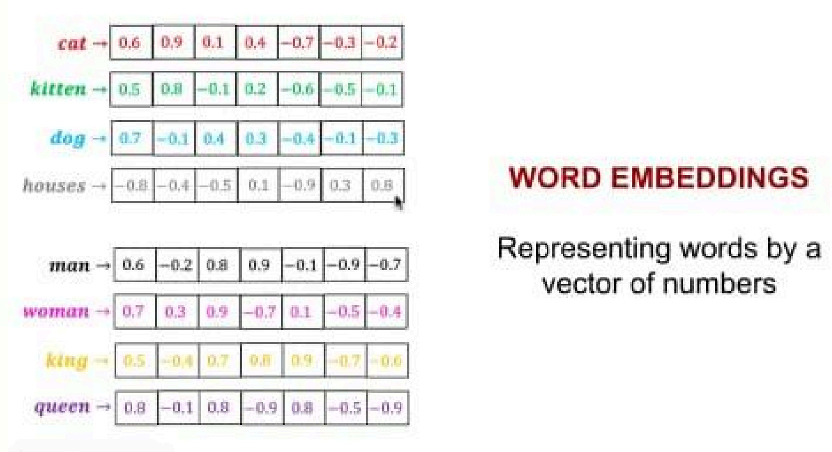
1. Database creation - There is a search space, i.e. a large number of documents in our database. Usually, the PostGres DB is recommended.

    a. Take an ML book and extract the text.
    b. Place the text in a DB in the form of documents

2. Identify which documents in the DB can be potential answers to the user query
3. Convert text to numbers so that the computer can understand it. This can be done in following ways -
    a. Feature Extraction - Used in classical NLP
    b. Embeddings - Numerical representation of a given sentence/word.

EMBEDDINGS

An Embedding is a numerical representation (i.e., vector of numbers) of a given sentence/word. Larger the vector, the better it is for capturing the meaning/essence of a sentence.



cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

WORD EMBEDDINGS

Representing words by a vector of numbers

Figure 3 - Example of word embeddings

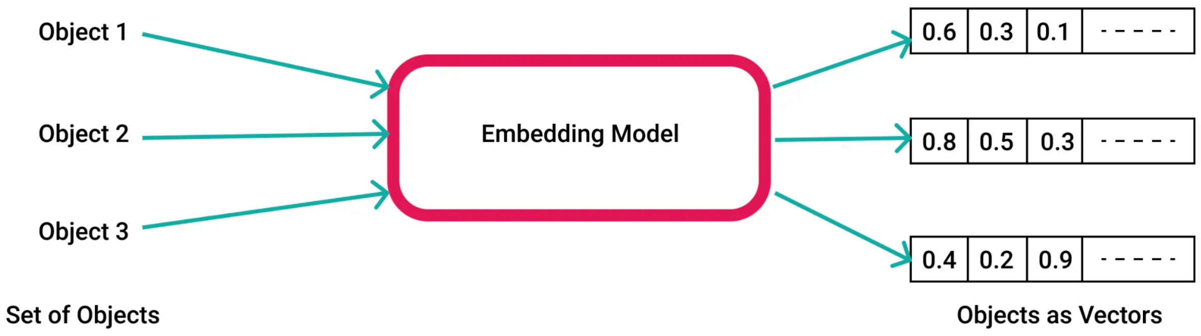


Figure 4 - Embedding Model

Words and sentences which are similar to each other, should have vectors which are closer to each other. This is calculated by calculating their cosine similarity, i.e. calculating the dot product of the two vectors (that is, cos θ).

- If sentences are similar, their dot product is closer to 1
- If sentences are dissimilar, their dot product is far less than 1

So, the complexity of the language gets encoded in the form of numbers.

Challenges -

1. Dot product can be misleading
2. Can take too much time for a large set of documents

METHODS OF GENERATING WORD EMBEDDINGS

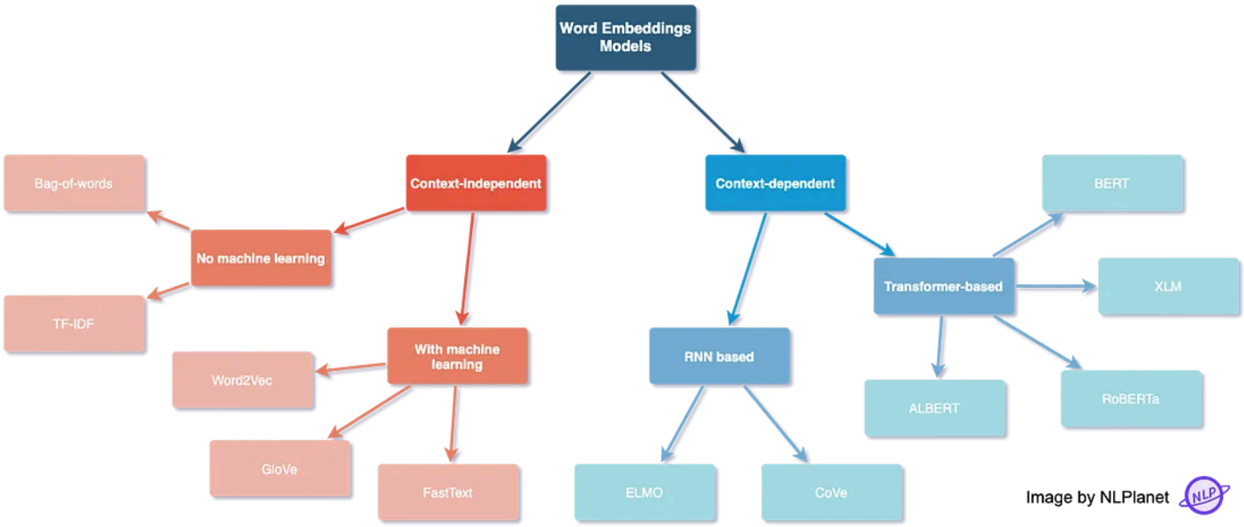


Image by NLPlanet

Figure 5 - Methods of generating Word Embedding Models

- Context Independent Embedding - In Figure 3, if we fix the embedding of the words, it becomes context independent embedding. So, if King or Queen occur in the same sentence, they would have the same meaning irrespective of the context.

- Context Dependent Embedding - The above is problematic because different words can have different meanings in different sentences. Hence, there is a need for a context dependent embedding which is dependent on the context of the surrounding words.

TRANSFORMERS

- Transformers are context dependent embeddings.
- Introduced in 2017 and completely changed the landscape of NLP
- Made of two components -
  - Encoder - Represents the given sentence in the form of vector
  - Decoder - Takes the input vector and decodes it to another sentence
- Application - Language Translation

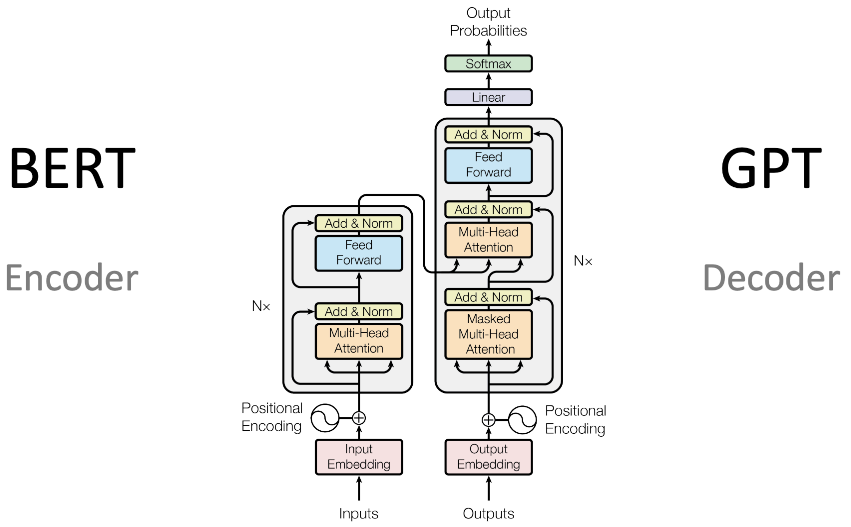


Figure 6 - Transformer

TRAINING

Input - Sentence A (vector corresponding to the query) and Sentence B (vectors generated by our documents)

Procedure -

- Both the sentences are converted into embedding using BERT.

- Then we calculate dot product of these 2 sentences to check how similar they are.
- This helps to find the top ‘n’ documents which are similar to the user query

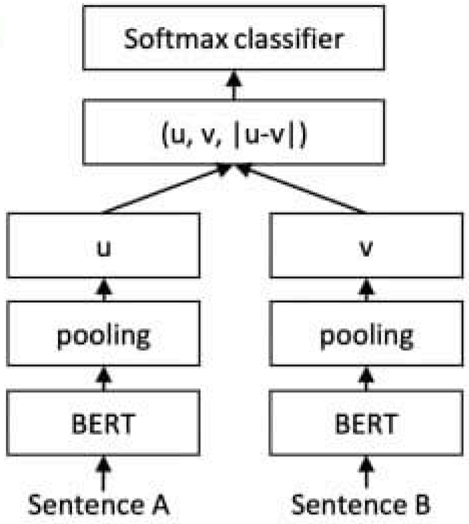


Figure 7 - Training Process