

# Latent Semantic Indexing

John Wong



# Context

---

You want to read through all the feedback to know their topics



# Question

---

How can you do so quickly?



# Answer

---

You can apply Latent Semantic Indexing on all the feedback



# Latent Semantic Indexing (LSI)

---

It is an unsupervised machine learning method to identify topics in the feedback based on the co-occurrence of words in them



# Topic

---

If "data" and "science" appear together frequently, the topic is on "data science"

If "data", "science", "complicated", "statistics", and "mathematics" appear together frequently, the topic is most likely on "the difficulty of learning data science"



# Words and Numbers

---

Machine learning model accepts only numbers, not words, as input

Hence words must be converted to vectors of numbers (ID, count)

This conversion is known as embedding

# Embedding

Embedding is also used in Transformer (Google) and Generative Pre-Trained Transformer (GPT) (OpenAI)

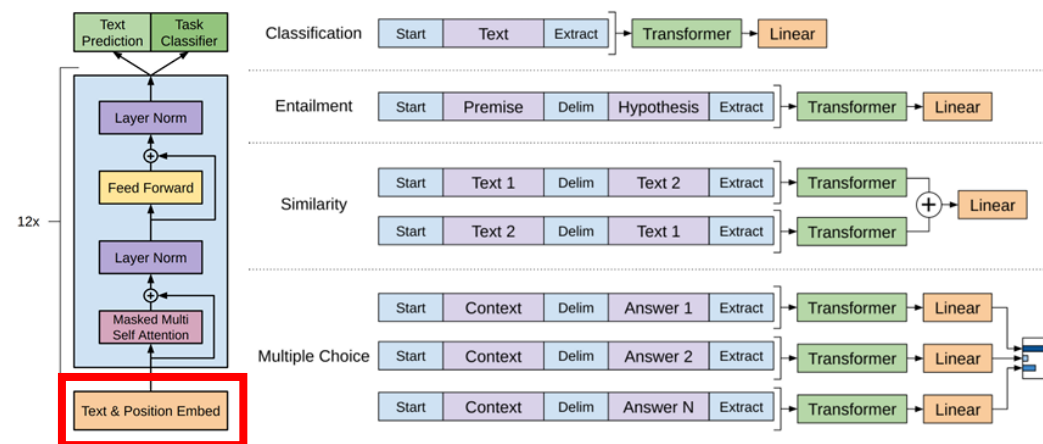


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Source:

[https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf)

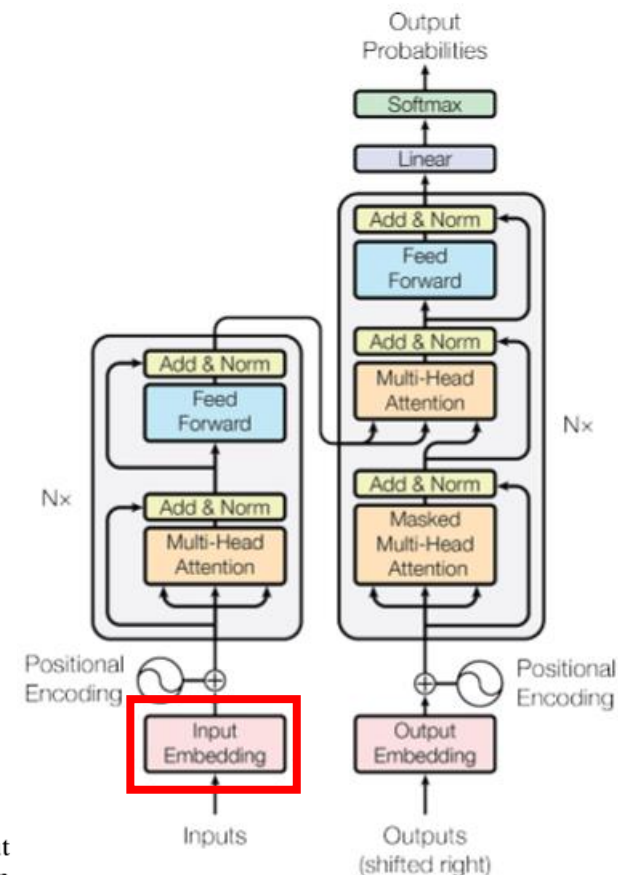


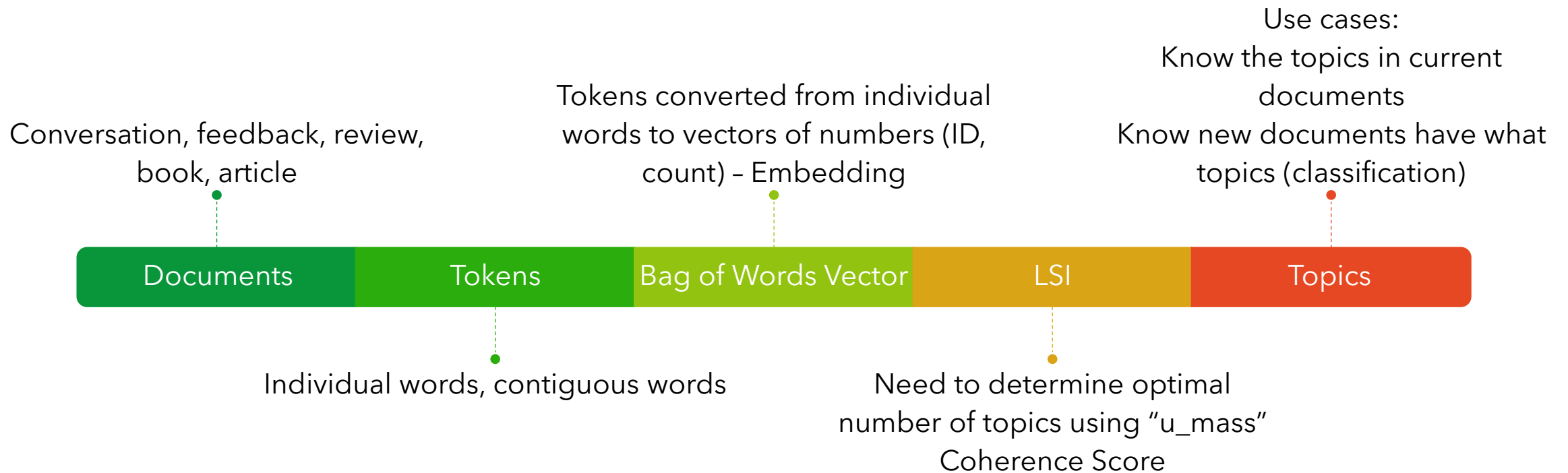
Figure 1: The Transformer - model architecture.

Source:

<https://www.turing.com/kb/guide-on-word-embeddings-in-nlp>



# Data Science Workflow





# The End

---

Hope you find this useful

Visit my GitHub for the codes

[https://github.com/johnwck/my\\_da\\_ds\\_work/tree/master/my\\_projects\\_github\\_pages/latent\\_semantic\\_indexing](https://github.com/johnwck/my_da_ds_work/tree/master/my_projects_github_pages/latent_semantic_indexing)