LECTURER: TAI LE QUY

ARTIFICIAL INTELLIGENCE

TOPIC OUTLINE

UNIT 5

NATURAL LANGUAGE PROCESSING

PART 2

STUDY GOALS

- Identify the typical tasks in NLP.
- Understand how to vectorize data, including
  - Bag-of-Words
  - Neural word vectorization techniques
  - Neural sentence vectorization techniques

EXPLAIN SIMPLY

1. What are the typical tasks in NLP?

2. How does Bag-of-Words work?

3. How can words and sentences be vectorized using neural models?

NLP TASKS

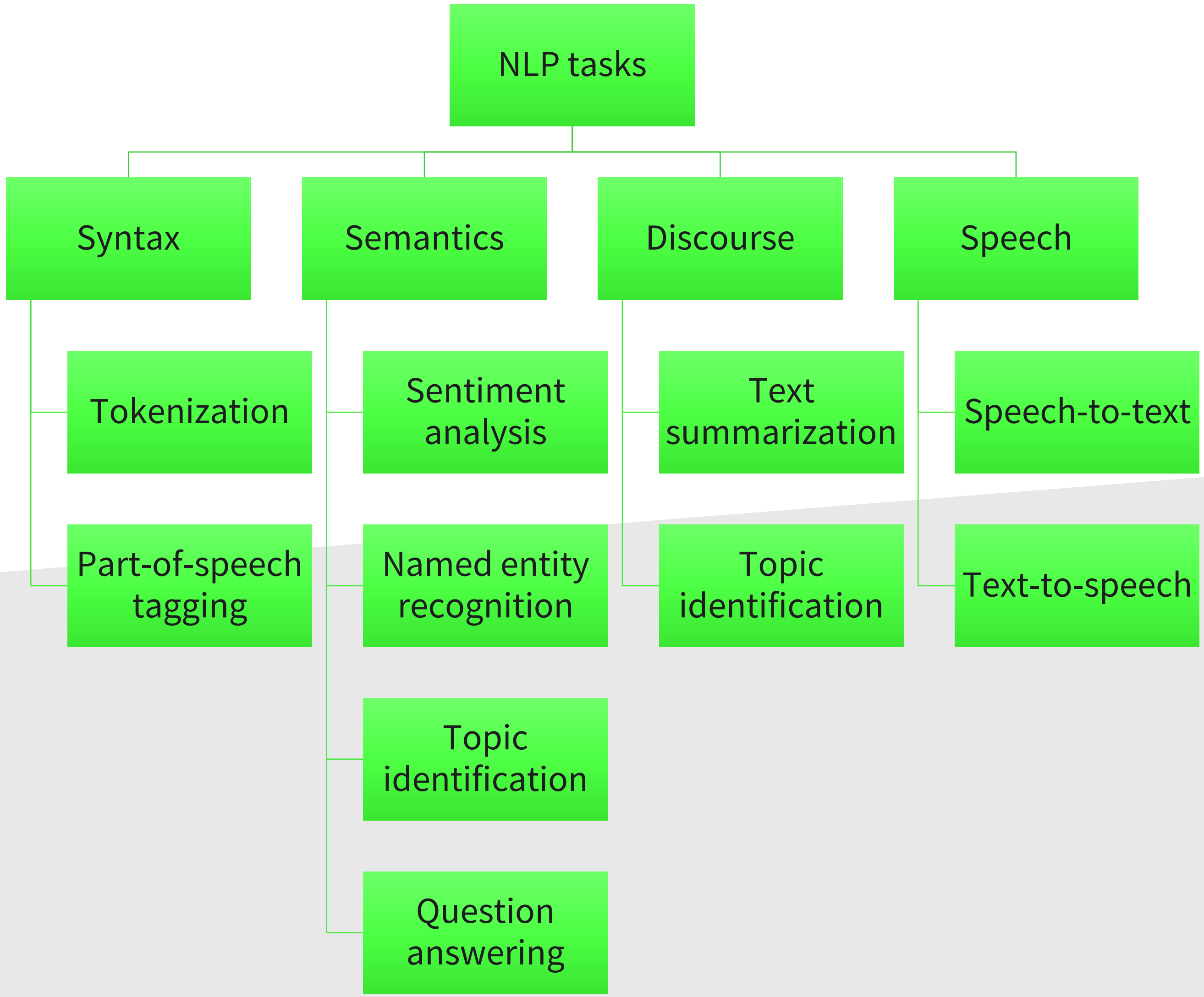


Image Source: Custom Depiction

VECTORIZING DATA – BAG-OF-WORDS (BOW)

Darren loves dogs.

Darren does not like cats.

Cats are not like dogs.

Darren, loves, dogs, does, not, like, cats, are

[2, 1, 2, 1, 2, 2, 2, 1]

Bag-of-words: describes the number of word occurrences in a given text document.

- Limitations of BoW
  - Selection of vocabulary: very carefully
  - Risk of high sparity
  - Loss of meaning

VECTORIZING DATA – WORD VECTORS

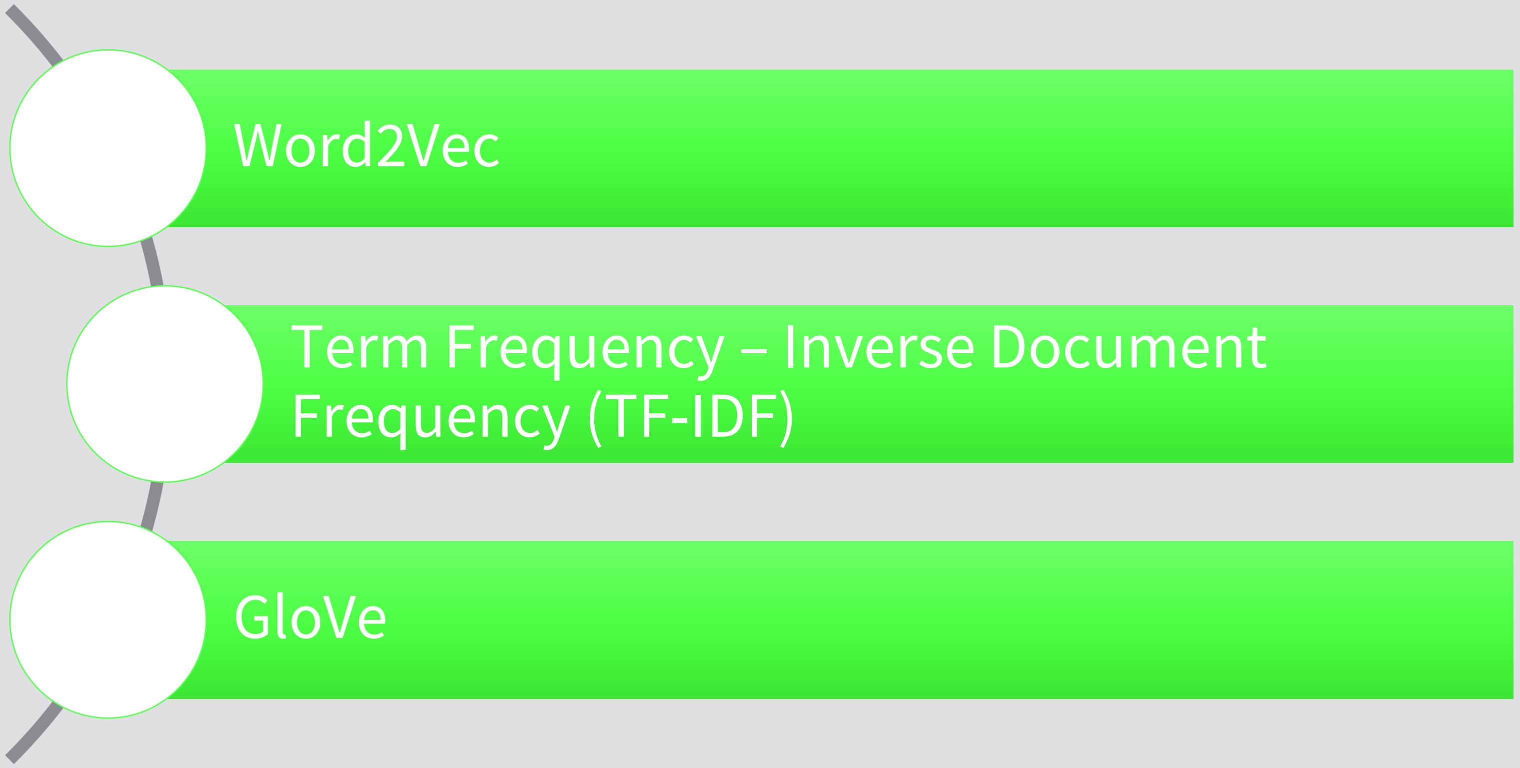


Image Source: Custom Depiction

WORD2VEC

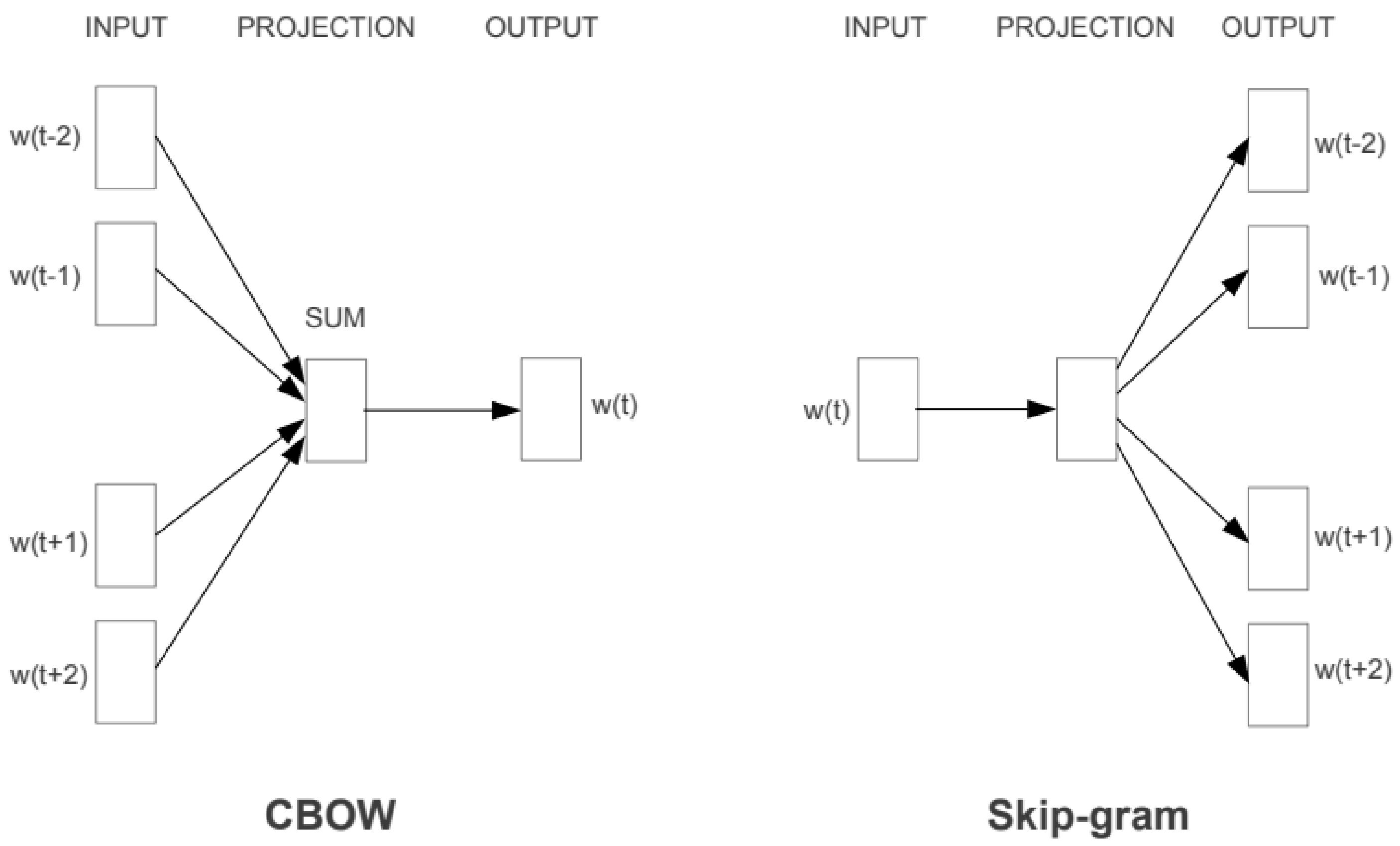


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Source of image: https://arxiv.org/pdf/1301.3781.pdf

WORD2VEC – CBOW VS. SKIP GRAM

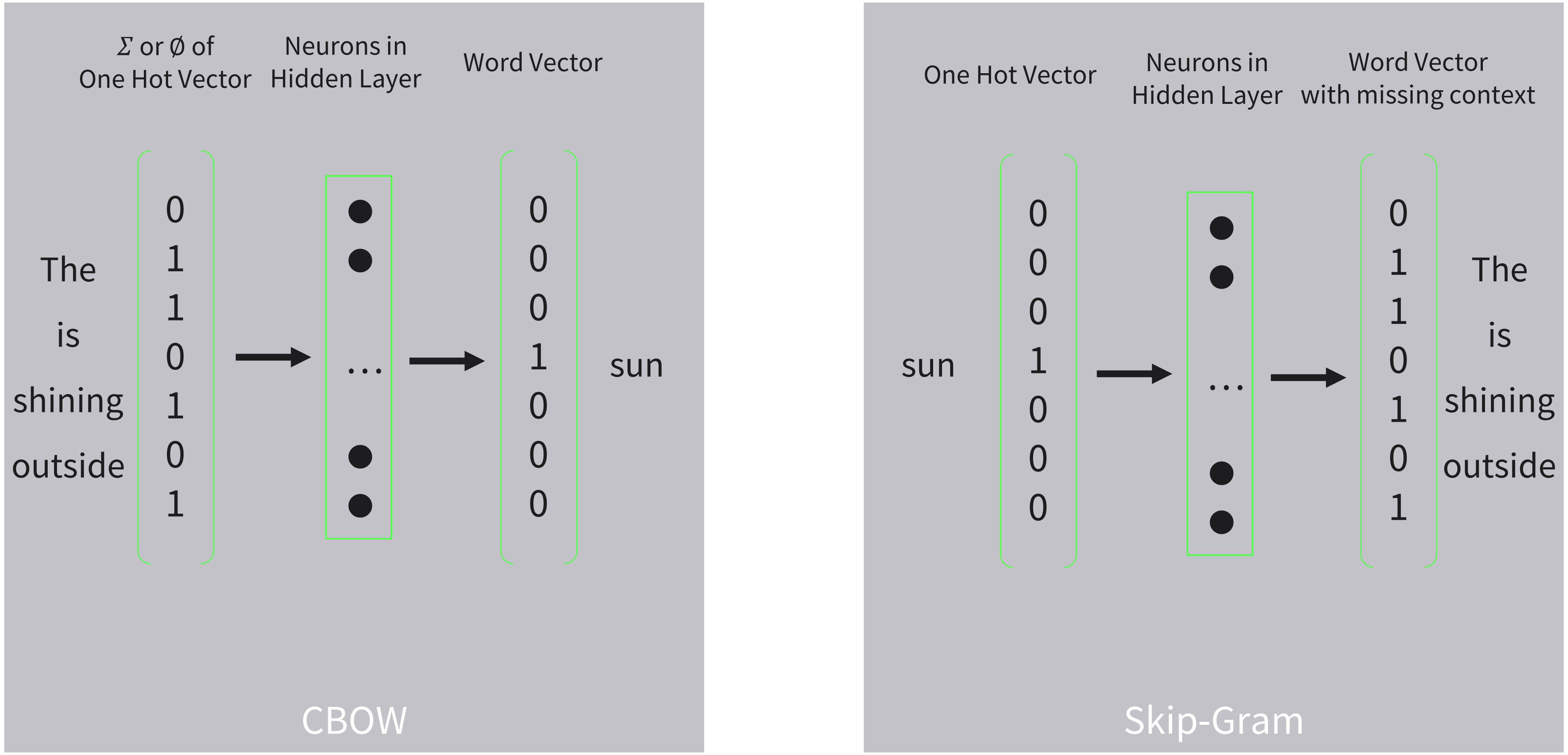


Image Source: Custom Depiction

WORD2VEC

Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot
3	[The **wide** road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road **shimmered** in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot **sun**].	sun, in sun, the sun, hot

Source of the image: https://www.tensorflow.org/tutorials/text/word2vec

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY

1

$$TF(t,d)$$

$$= \frac{number\ of\ occurences\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

2

$$DF(t,d.,D)$$

$$= \frac{number\ of\ documents\ d\ containing\ t}{total\ number\ of\ documents\ D}$$

3

$$IDF(t)$$

$$= \log \frac{1}{DF(t,d,D)}$$

4

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

- GloVe
  - An unsupervised approach based on the counts of words
  - Because skip-gram approach (Word2Vec) does not fully consider the statistical information when it comes to word co-occurrences
  - Combine the skip-gram approach with the benefit of matrix factorization
  - Use a co-occurrence matrix
    - Count each “word” (the rows), and how frequently we see this word in some “context” (the columns)

Darren does not like cats.

	Darren	Does	Not	Like	Cats
Darren	0	1	0	0	0
Does	1	0	1	0	0
Not	0	1	0	1	0
Like	0	0	1	0	1
Cats	0	0	0	1	0

Co-occurence matrix, window size = 1

VECTORIZING DATA – SENTENCE VECTORS

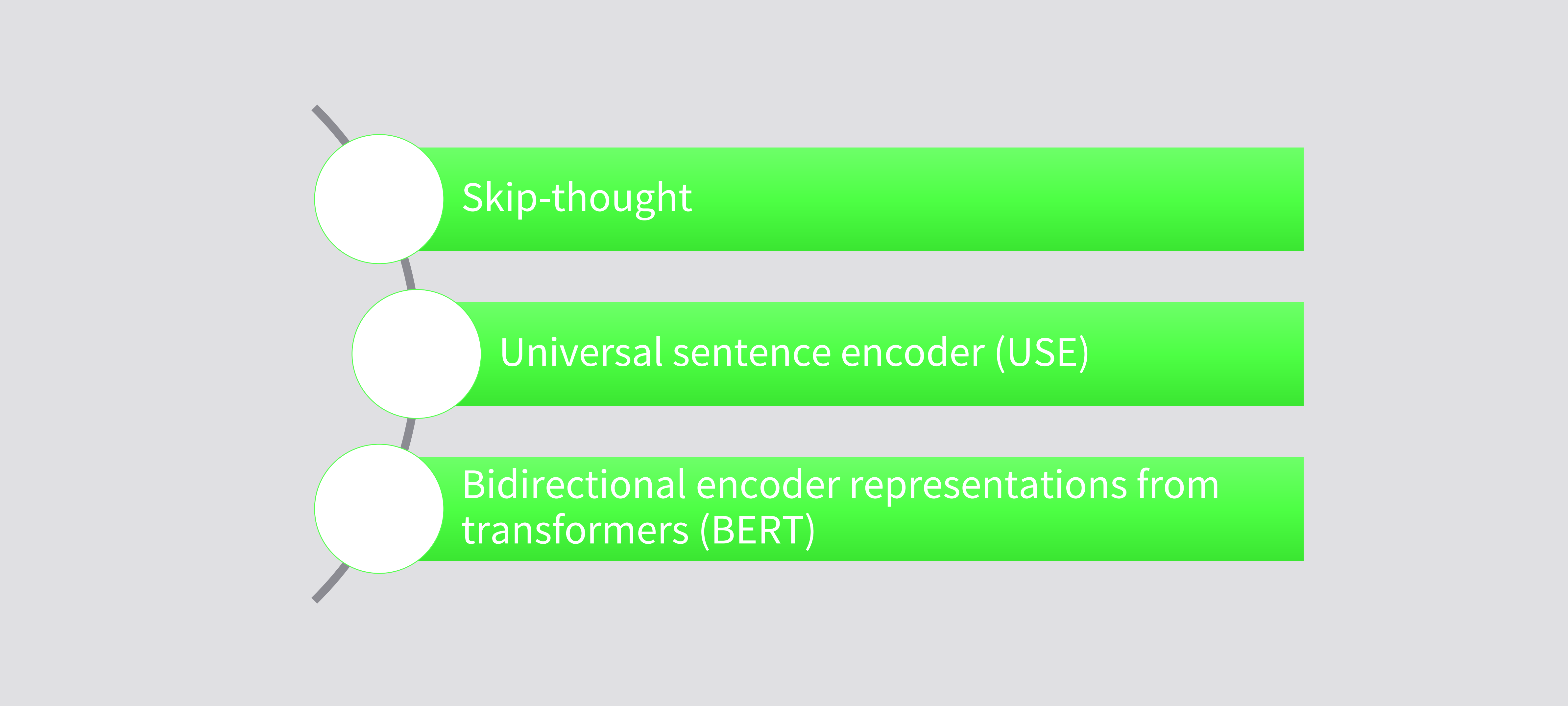
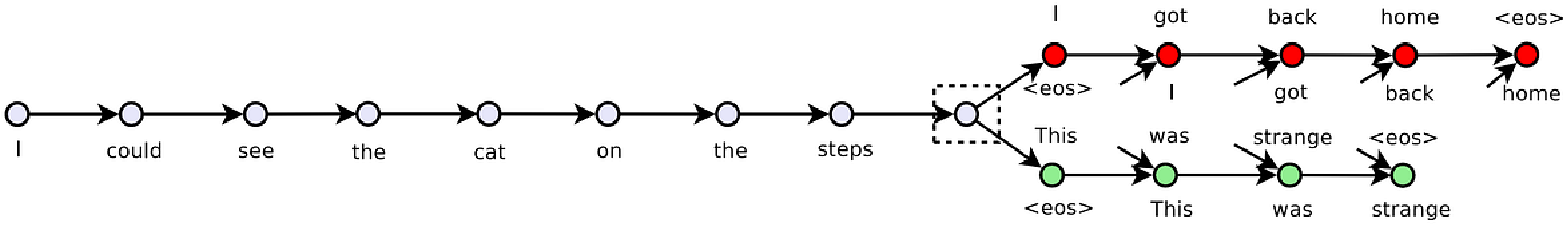


Image Source: Custom Depiction

- Use the concept of skip-gram architecture
- In contrast to Word2Vec
  - Skip-thought analyzes a triple of three consecutive sentences
  - Skip-thought vectors are created using an encoder-decoder model
  - The encoder takes in the training sentence and outputs a vector.
    - The first attempts to predict the previous sentence and the second attempts to predict the next sentence
- Model is pre-trained on the BookCorpus dataset



Source of image: https://arxiv.org/pdf/1506.06726.pdf

- Family of models for sentence embedding, developed by Gooogle Research
- Two variations
  - Trained with Transformer encoder
    - Higher accuracy, computationally more intensive
  - Trained with Deep Averaging Network (DAN)
    - Computationally less expensive and with little lower accuracy

Paper: https://arxiv.org/abs/1803.11175

- Based on the transformer architecture
- Pre-trained on a large text corpus in two combined and unsupervised ways:
  - Masked language model
    - Mask some words in the sentence (~15%)
    - Predict the missing words → understand the context of the words
  - Next sentence prediction
    - Received a pair of two sentences
    - Predict if the first sentence is followed by the second sentence → how a pair of sentences are related

Paper: https://arxiv.org/abs/1810.04805

REVIEW STUDY GOALS

- Identify the typical tasks in NLP.

- Understand how to vectorize data, including
  - Bag-of-Words
  - Neural word vectorization techniques
  - Neural sentence vectorization techniques

SESSION 5

TRANSFER TASK

1. Use the Bag-of-Words (BoW) approach to convert the following sentence into the corresponding vector representation:

John is taller than Mary and Mary is taller than Joe.

Now think about the question “Is John taller than Joe?” and discuss the shortcomings of the BoW approach.

2. In 10 documents, the words **NLP**, **study**, and **cat** have the following frequencies:

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	12	5	0	0	3	2	8	1	0	0
Study	1	0	7	1	0	0	2	0	5	12
Cat	0	12	0	6	8	1	3	10	0	9

Assume, that the D1-D5 contain 20 words. D6-D10 contain 100 words each.

Compute the TF-IDF for each term.

Which document will be returned if somebody wants to study something other than NLP?

Which document contains the most information about cats?

$$TF(t,d) = \frac{number\ of\ occurences\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

$$DF(t,d.,D) = \frac{number\ of\ documents\ d\ containing\ t}{total\ number\ of\ documents\ D}$$

$$IDF(t) = \log\frac{1}{DF(t,d,D)}$$

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

Go back to the GloVe example sentence “Darren does not like cats.” How would the co-occurence matrix change for a window size of 2?

Please present your results.
The results will be discussed in plenary.

1. *John is taller than Mary and Mary is taller than Joe.*

[John, is, taller, than, Mary, and, Joe]

[1, 2, 2, 2, 2, 1, 1]

The question if Joe is taller than John can not be answered, as the structure of the sentence gets lost.

2. Term frequencies

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	0.6	0.25	0	0	0.15	0.02	0.08	0.01	0	0
Study	0.05	0	0.35	0.05	0	0	0.02	0	0.05	0.12
Cat	0	0.6	0	0.3	0.4	0.01	0.03	0.1	0	0.09

2. Document frequencies

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
NLP	12	5	0	0	3	2	8	1	0	0	→ 6
Study	1	0	7	1	0	0	2	0	5	12	→ 6
Cat	0	12	0	6	8	1	3	10	0	9	→ 7

→ DF(NLP, 6, 10) = 0.6 → IDF(NLP) = 0.737

→ DF(Study, 6, 10) = 0.6 → IDF(Study) = 0.737

→ DF(Cat, 6, 10) = 0.7 → IDF(Cat) = 0.515

2. TF-IDF

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	0.442	0.184	0.000	0.000	0.111	0.015	0.059	0.007	0.000	0.000
Study	0.037	0.000	0.258	0.037	0.000	0.000	0.015	0.000	0.037	0.088
Cat	0.000	0.309	0.000	0.154	0.206	0.005	0.015	0.051	0.000	0.046

Studying something other than NLP: D3

Information about cats: D2

Darren does not like cats

	Darren	Does	Not	Like	Cats
Darren	0	1	1	0	0
Does	1	0	1	1	0
Not	1	1	0	1	1
Like	0	1	1	0	1
Cats	0	0	1	1	0

LEARNING CONTROL QUESTIONS

1. Name the four categories of NLP tasks.

2. How is the meaning of a text represented using the BoW model?

3. Name three methods for word vectorization.