Natural Language Processing

Session-3

CLARUSWAY©
WAY TO REINVENT YOURSELF

Word Embedding

NLP with Deep Learning

Word Embedding (Feature Representation)

```
In [6]: model.wv['ankara']

Out[6]: array([ 0.28124622, -0.40046975,  0.4623563 , -0.26006582,  0.4291707 ,
               -0.5458609 , -0.36119318, -0.28867924, -0.60250515,  0.4237076 ,
                0.28818154, -0.17176527,  0.19708447, -0.37333247, -0.5264093 ,
               -0.6427167 ,  0.14568327,  0.17329371, -0.5378837 , -0.3682972 ,
               -0.15588258, -0.20114157,  0.67830926,  0.04082025, -0.19011535,
                0.59075046, -0.6102628 , -0.42992386, -0.12169785, -0.2939379 ,
               -0.13670284, -0.13515596, -0.1399645 , -0.72388405, -0.7582215 ,
                0.10269422,  0.23371245,  0.02973733,  0.10352834,  0.17330961,
                0.31032264, -0.00689159, -0.51990616,  0.4342847 ,  0.24778119,
                0.08977021,  0.37872604,  0.2631365 ,  0.11655401,  0.02951079,
                0.2531644 ,  0.03372194, -0.35848543, -0.21600617,  0.33282214,
                0.27077678,  0.53709596,  0.26062086, -0.24004991,  0.10561307,
                0.36568683, -0.8039388 ,  0.41826522, -0.32496533,  0.21453372,
               -0.1572319 , -0.7251499 ,  0.08052048,  0.08987505,  0.1112271 ,
                0.14825632,  0.58504164,  0.684992  ,  0.00776701,  0.18076658,
               -0.09477395, -0.1137936 ,  0.5623219 , -0.3188122 ,  0.5662095 ,
                0.32072476,  0.17244305, -0.58838886, -0.05057955, -0.47085264,
               -0.17080188, -0.17386362, -0.12876017, -0.02662375,  0.36869773,
               -0.23252094, -0.10350469, -0.09519712, -0.09009365, -0.2726341 ,
               -0.04719334, -0.25645226, -0.03347286,  0.10456851,  0.06934233],
              dtype=float32)
```

NLP with Deep Learning

Word Embedding (Feature Representation)

Word Embeddings

Rome = [0.91, 0.83, 0.17, …, 0.41]

Paris = [0.92, 0.82, 0.17, …, 0.98]

Italy = [0.32, 0.77, 0.67, …, 0.42]

France = [0.33, 0.78, 0.66, …, 0.97]

Word embeddings are numeric vectors that show **semantic relationships** between **words.**

NLP with Deep Learning

Word Embedding (Feature Representation)

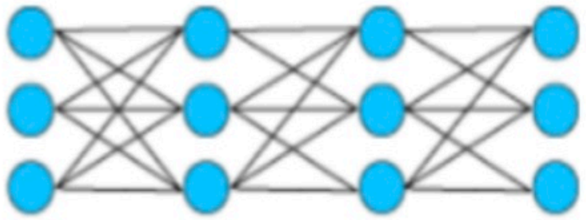
Feature Extraction

Machine Learning

TEXT →  → **CountVectorizer**, **Tfidfvectorizer**

Feature extraction

Deep Learning

TEXT →  → **Word Embedding**, **Sentence Embedding**

Feature extraction

NLP with Deep Learning

Word Embedding (Feature Representation)

For example;

-The teacher made a test to the students in school

-The teacher graded the students

-The teacher assigned a project to the student

-The instructor gave a quiz to the learners at the university

-The instructor made an verbal to the learner

-The learners graduated from the university

-The learners received diplomas

NLP with Deep Learning

Word Embedding (Feature Representation)

Unique Tokens:

-the	-school	-university
-teacher	-graded	-verbal
-made	-assigned	-graduated
-a	-project	-from
-test	-instructor	-received
-to	-quiz	-gave
-students	-learners	-diplomas
-in	-at	
-learner	-student	

Feature Representation:

education, teaching, organisation, article, verb, preposition

NLP with Deep Learning

Word Embedding (Feature Representation)

Unique Tokens / Feature Representations	the	teacher	made	a	test	to	student/s	in	school	graded	assigned	project
education	0.01	-0.95	0.01	0.02	0.92	0.01	-0.95	0.01	0.88	0.85	-0.03	0.85
teaching	0.02	-0.90	0.02	0.01	0.89	0.03	-0.95	0.02	0.90	0.87	-0.02	0.88
organisation	0.5	0.02	0.03	0.45	0.01	0.40	0.01	0.42	0.91	-0.02	0.02	0.01
article	1	0.60	0.01	1	0.02	-0.01	0.48	0.01	0.50	-0.01	0.01	0.10
verb	0.01	0.53	1	0.20	1	0.40	0.49	0.02	0.40	1	1	0.80
preposition	0.02	0.03	-0.02	-0.01	0.02	1	0.01	1	0.02	0.01	0.04	0.02

Unique Tokens / Feature Representations	instructor	quiz	learner/s	at	university	verbal	graduated	from	received	gave	diplomas
education	-0.93	0.91	-0.93	0.01	0.90	0.85	0.83	-0.01	0.03	0.02	0.82
teaching	-0.88	0.92	-0.94	0.02	0.89	0.86	0.87	-0.03	0.04	0.03	0.85
organisation	0.01	-0.02	0.02	0.45	0.92	-0.02	-0.02	0.30	-0.02	-0.02	0.02
article	0.55	0.05	0.60	0.01	0.45	0.01	0.03	0.02	0.01	0.03	-0.03
verb	0.52	0.85	0.51	0.02	0.45	0.70	1	0.01	1	1	-0.42
preposition	0.01	0.03	0.02	1	0.01	0.01	0.02	1	0.02	0.03	0.02

NLP with Deep Learning

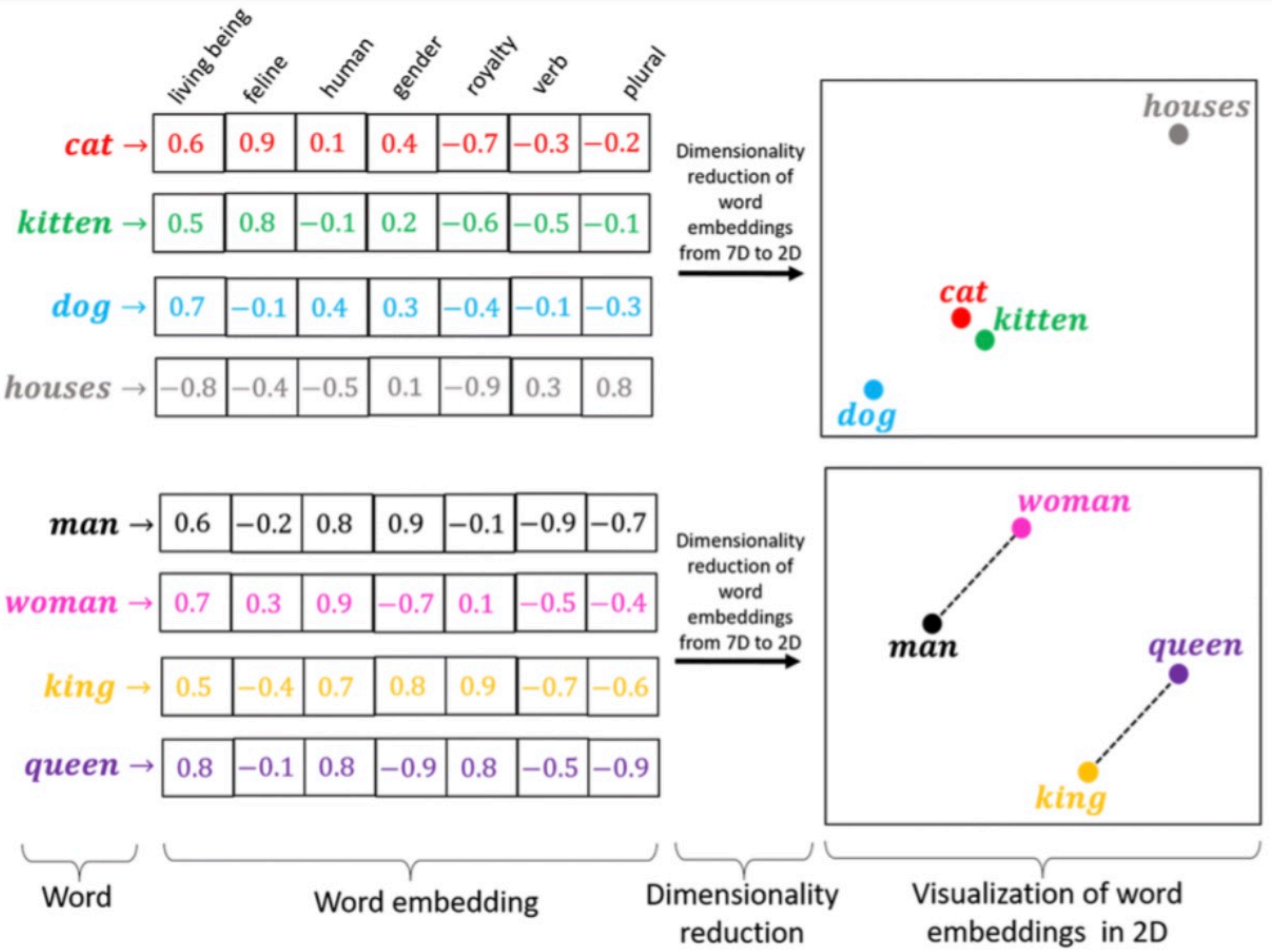
Word Embedding-Feature Representation

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
300 Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮				
size						
cost						
alive						
verb						
	e_{5391}	e_{9853}				

Andrew

NLP with Deep Learning

Word Embedding (Feature Representation)



NLP with Deep Learning

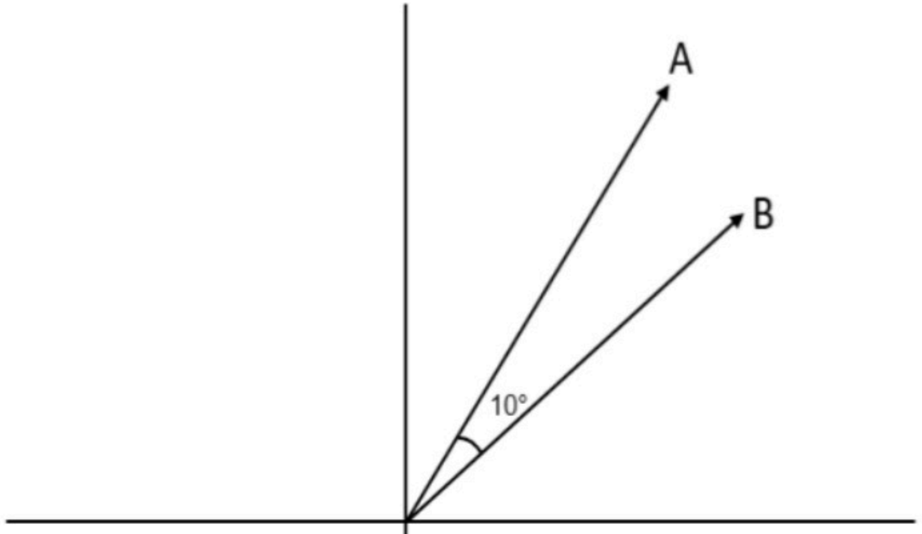
Cosine Similarity

-Cosine similarity is a measurement that quantifies the similarity between two or more vectors.

-The cosine similarity is the cosine of the angle between vectors. The cosine similarity is a value that is bound by a constrained range of -1 and 1.

NLP with Deep Learning

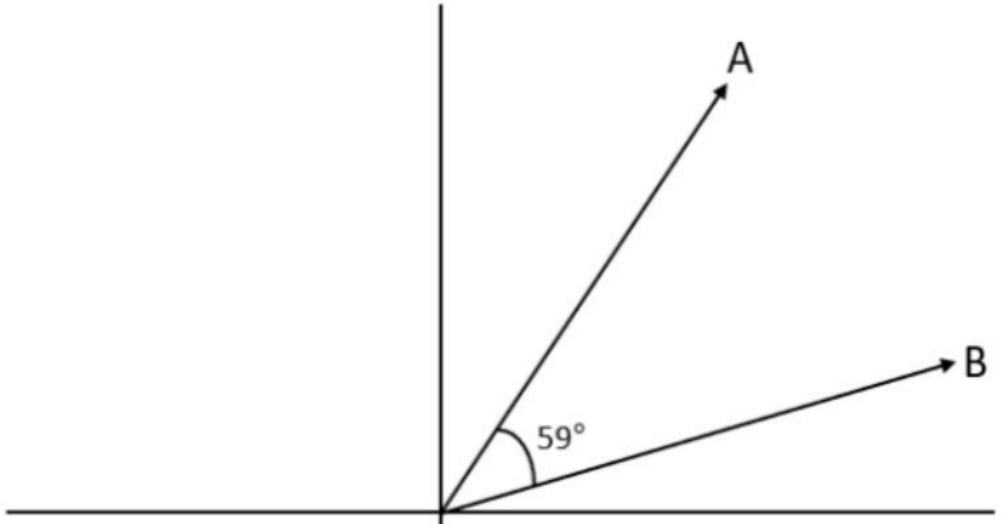
Cosine Similarity



The angle between vector A and B is 10 deg.

Cos(10) = 0.9848...

The angles could be said to be 98% similar



The angle between vector A and B is 59 deg.

Cos(59) = 0.559...

The angles could be said to be 55% similar

NLP with Deep Learning

Word Embedding Algorithms

-**Embedding Layer**

-**Word2Vec**

- Continuous Bag of Words (CBOW Model)
- Skip-Gram Model

-**Global Vectors (Glove)**

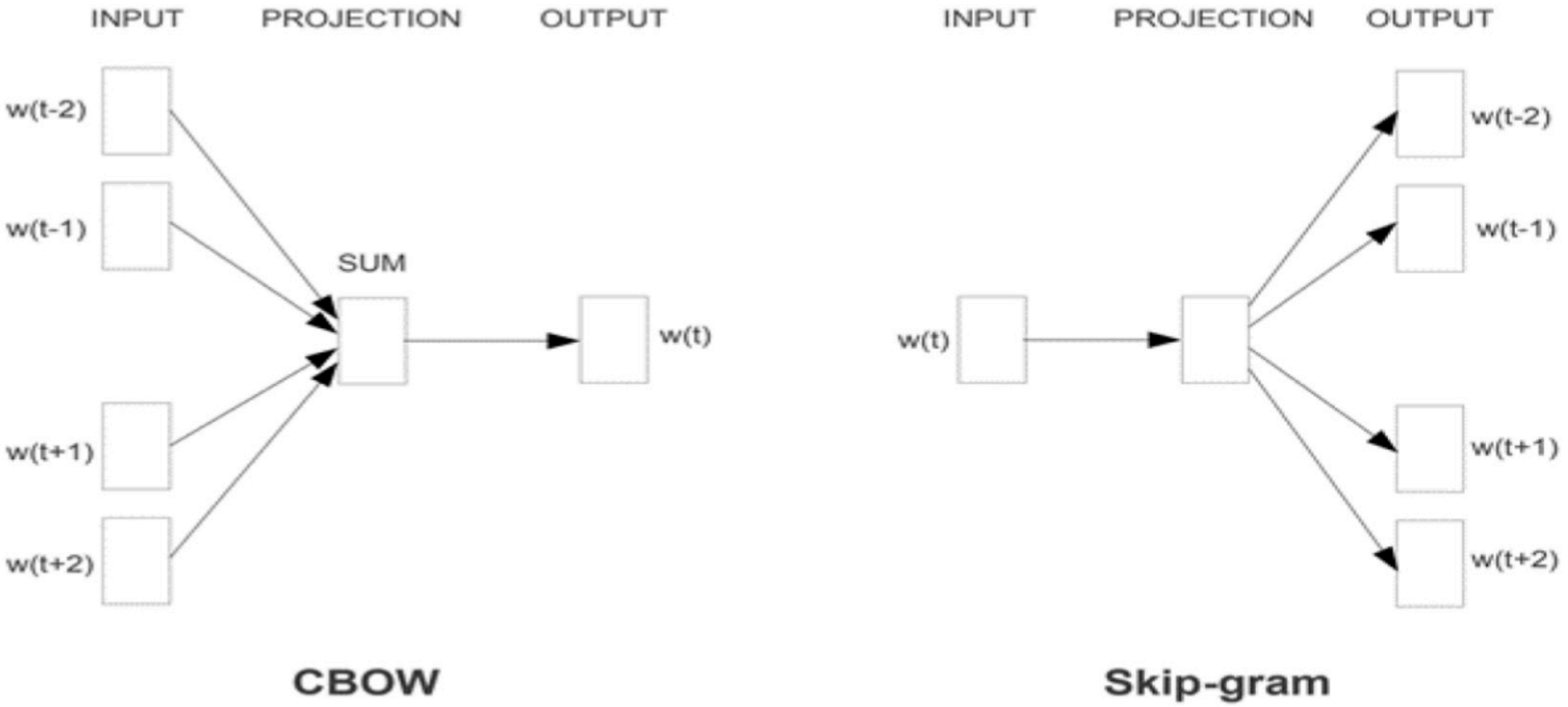
-**Embeddings from Language Models (ELMo)**

-**Bidirectional Encoder Representations from Transformers (BERT)**

-**Generative Pre-trained Transformer (GPT-2/3)**

NLP with Deep Learning

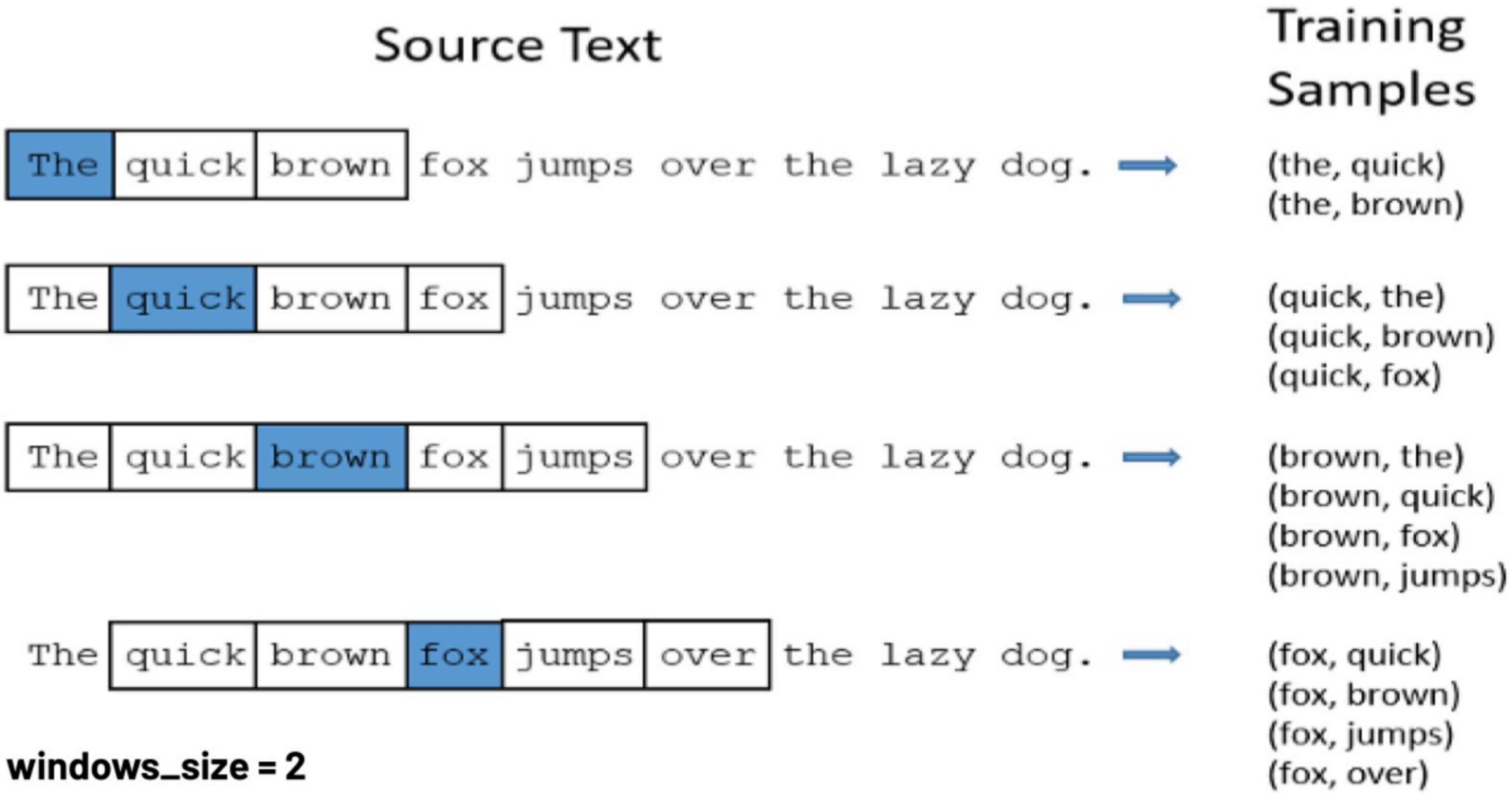
Word2Vec



windows_size = 2

NLP with Deep Learning

Word2Vec-window_size parameter



windows_size = 2

NLP with Deep Learning

ELMO and BERT

-The word2vec and glove models are unfortunately unsuccessful in capturing homophones with different meanings.

-However, the **ELMO** and **BERT** model is quite successful in capturing the semantic differences between words.

I love **the apple** that my mother bought from the market.

I love **the apple,** the biggest company in the world.