

# Ontologies and text mining

# Textual components of ontologies


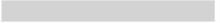
	Neoplasm of the breast
	Breast carcinoma
	— Multifocal breast carcinoma
Breast carcinoma HP:0003002	
<i>The presence of a carcinoma of the breast.</i>	
<b>Synonyms:</b> Breast cancer	
<b>Cross References:</b> MSH:D001943, NCIT:C2916, SNOMEDCT_US:254838004, UMLS:C0678222	

Figure from the disease ontology website

# Learning objectives

- Learn applications of the textual component of ontologies
- Learn text mining basics
- Get familiarized with popular text mining methods and their applications

# Popular applications

- Generate cross references: Manually/Automatically
- Retrieve information from literature in text form
- Generate annotations
- Align/expand ontologies

# How to compare/find AND link text?

- Exactly?
  - Brain tumor
  - Brain tumour
- Approximately?
  - How?

# Exact match

- Dictionaries

## Human Phenotype Ontology

Breast cancer → HP:003002  
Breast Carcinoma → HP:003002

## Disease Ontology

Breast cancer → DOID:1612  
breast tumor → DOID:1612  
malignant neoplasm of breast → DOID:1612  
malignant tumor of breast → DOID:1612

# Exact match

- Dictionaries
- Example of applications:
  - Mapping entities from different sources

## Human Phenotype Ontology

Breast cancer → HP:003002  
Breast Carcinoma → HP:003002

## Disease Ontology

Breast cancer → DOID:1612  
breast tumor → DOID:1612  
malignant neoplasm of breast → DOID:1612  
malignant tumor of breast → DOID:1612

# Exact match

- Dictionaries
- Example of applications:
  - Mapping entities from different sources
  - **Finding mentions in literature and their co-occurrences**

## Human Phenotype Ontology

Breast cancer → HP:003002

Breast Carcinoma → HP:003002

*Breast cancer is more prevalent in females, however, males can also develop breast cancer.*

*Birth control can increase risk of breast cancer in females.*



# Any ideas?

Brain tumor was found in a patient.

There are any forms of brain tumour.

# Approximate comparison of text

- Exclude unimportant information
- Stemming
  - Remove affixes
    - Cancers → cancer
    - Hyperpigmentation → hyperpig
  - Different stems
  - Novel words cannot be stemmed

# Approximate comparison of text

- Exclude unimportant information
- Stemming
  - Remove affixes
    - Cancers → cancer
    - Hyperpigmentation → hyperpig
  - Different stems
  - Novel words cannot be stemmed
- Numerical representation
  - Numerical methods
  - How?

# Numerical representation and analysis of text

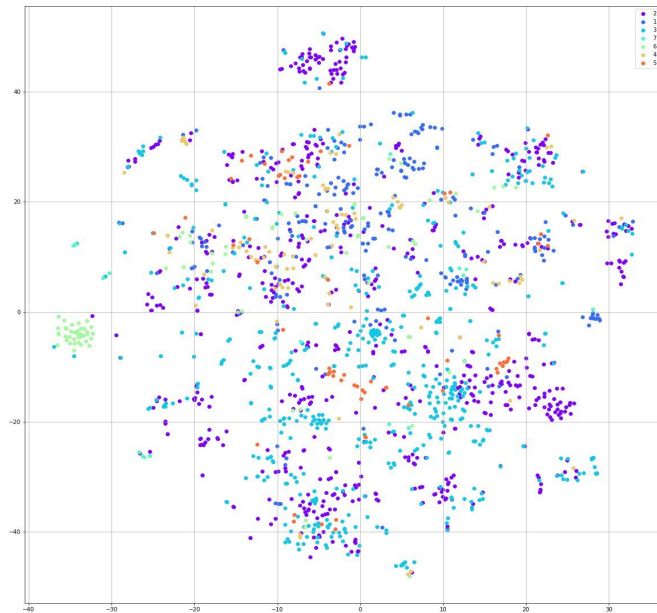
Popular methods:

- Word2Vec
- BERT

Both methods use embeddings to represent text

# What is an embedding?

- An embedding is a representation of a structure in a different space that **preserves** properties of that structure
- This is done by an embedding function  $f$
- $f$  preserves some property (structure-preserving)



# Why embeddings?

- Perform functions on instances that were not possible in their original form
- Represent instances in a compact dimension
- Visualize instances and their relations

# How are embeddings created for text?

- How is text represented numerically?
- Units of text are assigned IDs to create a vocabulary
  - Characters (letters)
  - N-grams (words)
  - Sentences
- How can we make this useful for comparison?

# How are embeddings created for text?

- How is text represented numerically?
- Units of text are assigned IDs to create a vocabulary
  - Characters (letters)
  - Words
  - N-grams
  - Sentences
- How can we make this useful for comparison?



# How are embeddings created for text?

- How is text represented numerically?
- Units of text are assigned IDs to create a vocabulary
  - **Characters (letters)**
  - Words
  - N-grams
  - Sentences
- How can we make this useful for comparison?


be bad → 250214

	0
a	1
b	2
c	3
d	4
e	5
f	6
g	7
h	8
i	9

# How are embeddings created for text?

- How is text represented numerically?
- Units of text are assigned IDs to create a vocabulary
  - Characters (letters)
  - **Words**
  - N-grams
  - Sentences
- How can we make this useful for comparison?

be bad → 102

	0
be	1
bad	2
good	3
well	4

# Word2Vec

- Well-known method
- Generates embeddings that capture co-occurrences based on a corpus
- Embeddings are in the form of  $n$ -dimensional vectors

Breast cancer is more prevalent in females, however, males can also develop breast cancer.

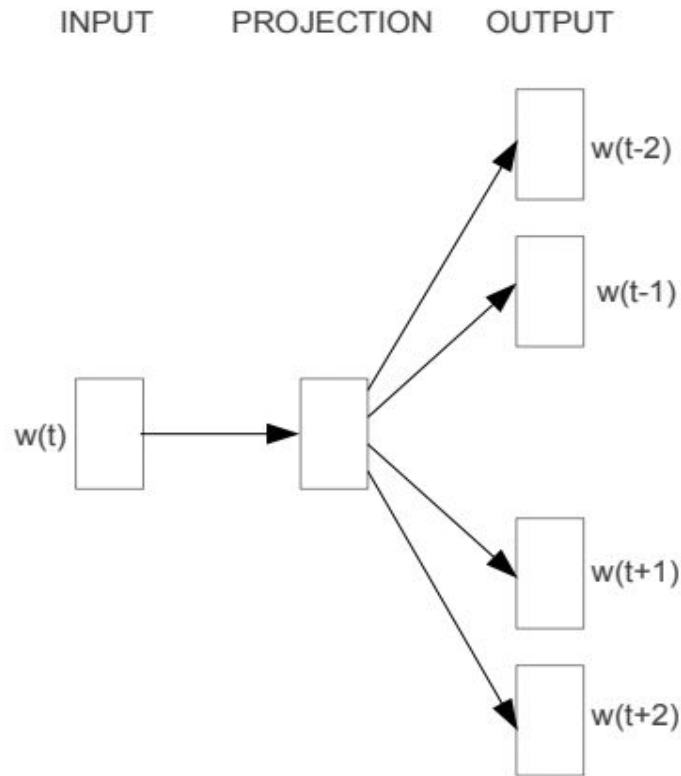
Birth control can increase risk of breast cancer in females.

# Word2Vec

Word2Vec captures co-occurrences

Given a word:

- Capture the words it frequently co-occurred within the given corpus



**Skip-gram**

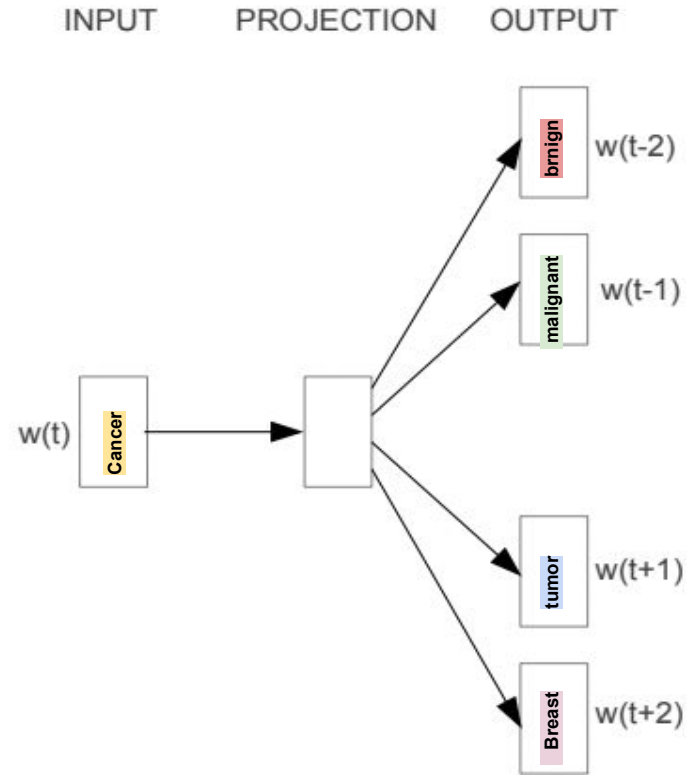
Figure from the original paper: Efficient Estimation of Word Representations in Vector Space, Mikolov et al.

# Word2Vec

Word2Vec captures co-occurrences

Given a word:

- Capture the words it frequently co-occurred within the given corpus
- Minimize the cross-entropy loss



**Skip-gram**

Figure from the original paper: Efficient Estimation of Word Representations in Vector Space, Mikolov et al.

# Word2Vec

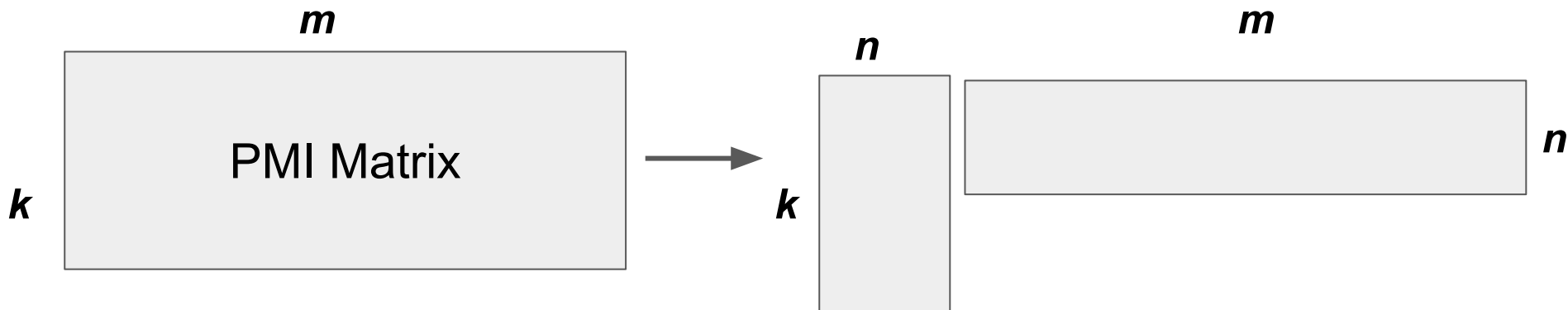
You can think of it as a factorization of a Pointwise Mutual Information (PMI) matrix

	Cancer	Benign	Malignant	Tumor	Breast
Cancer	0	2	1	1	4
Benign	2	0	1	1	2
Malignant	1	1	0	2	2
Tumor	1	1	2	0	3
Breast	4	2	2	3	0

$$\text{pmi}(x; y) \equiv \log \frac{p(x, y)}{p(x)p(y)}$$

# Word2Vec

You can think of it as a factorization of a Pointwise Mutual Information (PMI) matrix



# Words → embeddings

Breast cancer is more prevalent in females, however, males can also develop breast cancer.

...

Birth control can increase risk of breast cancer in females.

**Word2Vec**

[0.50929456, 0.6771953, 0.91371871, 0.48265797, 0.18390237]
[0.9146623, 0.7340195, 0.78049964, 0.54384624, 0.01162719]
[0.22451245, 0.97085067, 0.79003223, 0.74382914, 0.26143969]
[0.11487895, 0.43190008, 0.86119749, 0.96533036, 0.56099287]
[0.77668599, 0.52129723, 0.71529702, 0.82580858, 0.40596435]



# Word2Vec

- Embeddings capture co-occurrences
- Words that appear together frequently have similar vectors
- Language semantics?
- Distance measure can be used:
  - Edit distance
- Limitations?
  - Fixed representations
  - Context
  - Beyond co-occurrences

# Word2Vec applications

- Linking of ontology concept mentions to class IDs
  - Cho, H., Choi, W. & Lee, H. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*

Word2Vec embeddings are generated  
Cosine similarity is used to rank candidate class IDs

# Word2Vec applications

- Linking of ontology concept mentions to class IDs
  - Cho, H., Choi, W. & Lee, H. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*
- Relation extraction between entities from text
  - Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In Proceedings of NAACL-HLT.

Word2Vec embeddings are generated  
Neural convolutional models are trained to predict relations

# Word2Vec applications

- Linking of ontology concept mentions to class IDs
  - Cho, H., Choi, W. & Lee, H. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*
- Relation extraction between entities from text
  - Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In Proceedings of NAACL-HLT.
- Matching concepts between ontologies
  - Liao, J., Huang, Y., Wang, H., Li, M. (2021). Matching Ontologies with Word2Vec Model Based on Cosine Similarity. In: , et al. Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)

Word2Vec embeddings are generated  
Concepts from two ontologies are aligned based on cosine similarity

# Word2Vec shortcomings

Static representations

Context agnostic representations

# BERT

## Bidirectional Encoder Representations from Transformers

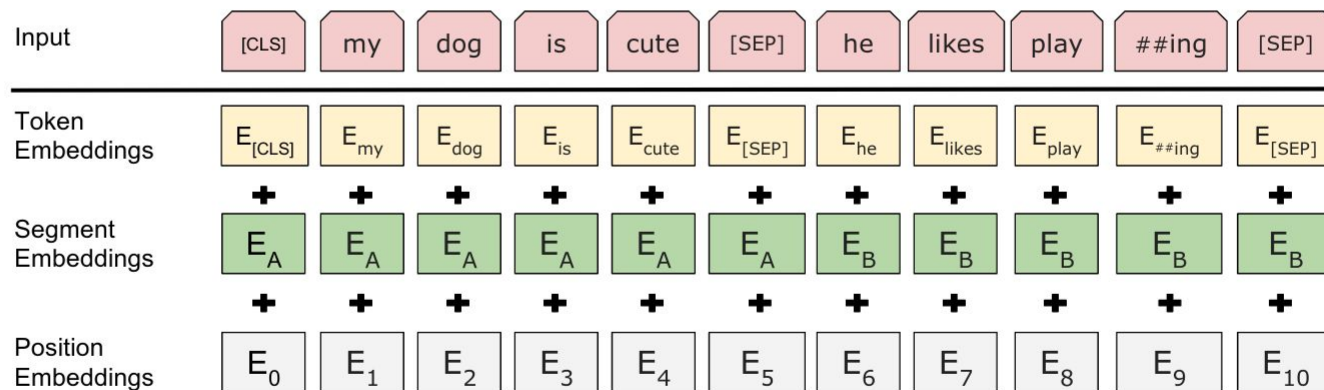
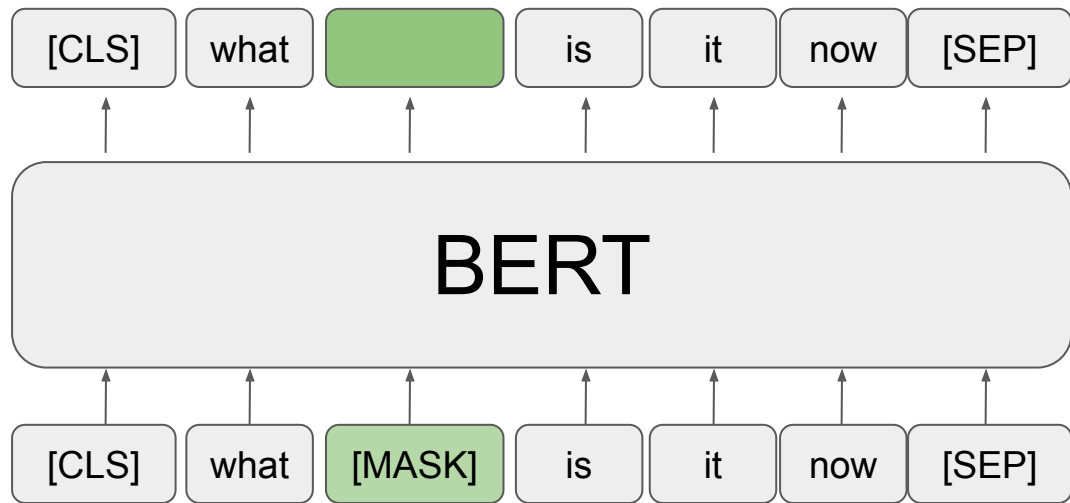


Figure from the original paper: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al.

# BERT

Masked Language Model (MLM)



# BERT

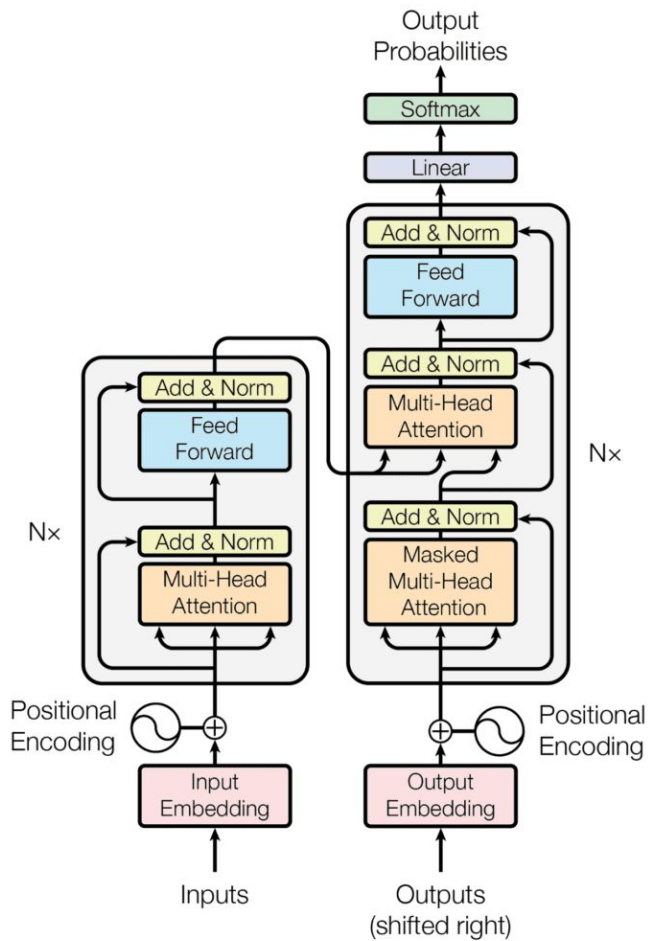


Figure from the original paper: Attention Is All You Need, Vaswani et al.



# BERT

- Dynamic
- Context-aware
- Word/sentence embeddings

# BERT applications

- Finding and linking concept mentions from text to ontology IDs
  - Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, Zhiyong Lu. PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. *Bioinformatics*, Volume 37, Issue 13, 1 July 2021, Pages 1884–1890.

BERT is fine-tuned:

Labels and synonyms → positives

Negatives are randomly sampled from some corpus

# BERT applications

- Finding and linking concept mentions from text to ontology IDs
  - Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, Zhiyong Lu. **PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology**. *Bioinformatics*, Volume 37, Issue 13, 1 July 2021, Pages 1884–1890.
- Relation extraction between entities from text
  - Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**, *Bioinformatics*, Volume 36, Issue 4, 15 February 2020.

BERT is trained on Biomedical corpora  
BERT is then fine-tuned using curated tuples

# BERT applications

- Finding and linking concept mentions from text to ontology IDs
  - Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, Zhiyong Lu. PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. *Bioinformatics*, Volume 37, Issue 13, 1 July 2021, Pages 1884–1890.
- Relation extraction between entities from text
  - Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, Volume 36, Issue 4, 15 February 2020.
- Matching concepts between ontologies
  - He, Y., Chen, J., Antonyrajah, D., & Horrocks, I. (2022). BERTMap: A BERT-based ontology alignment system. *Proceedings of the . AAAI Conference on Artificial Intelligence*

# Take home messages

- Textual components of ontologies can help
  - Extract knowledge from literature
  - Transfer knowledge from one source to another
- Methods to represent text
  - Word2Vec
  - BERT
- Important aspects of text:
  - Word meaning
  - Context