How to build a foundation of AI-based healthcare systems through language models?

Mehdi Bahrami, Ph.D., IEEE Senior Member
Artificial Intelligence Laboratory
Fujitsu Research of America,
Sunnyvale, CA

NLP SUMMIT

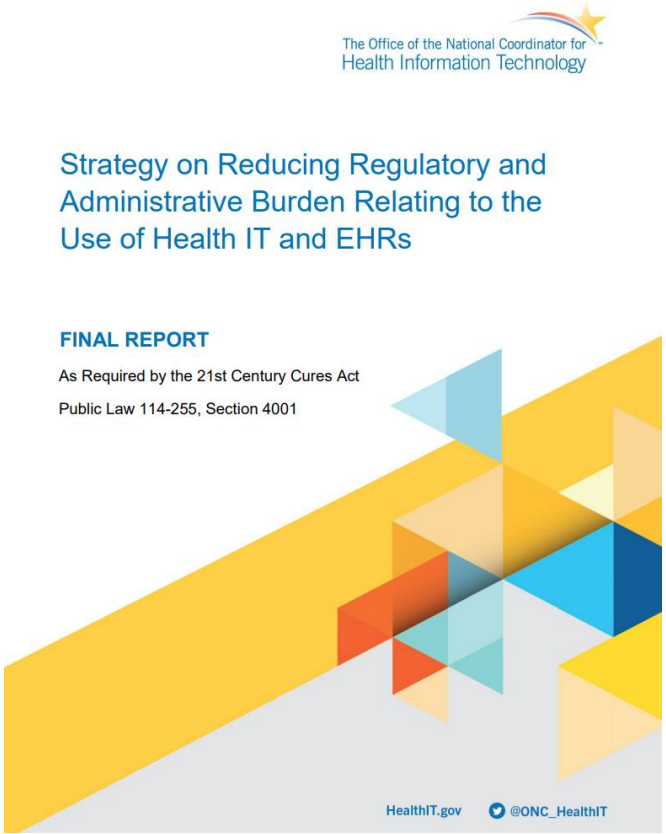
Code & Slide

Overview

- Why Language Model?
- What is a language model?
  - Word2Vec
  - GloVe
  - Transformer
  - BERT

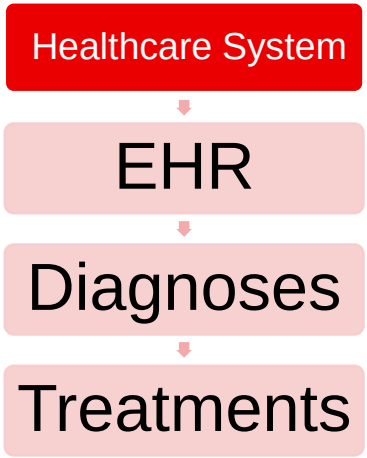
Why language models?

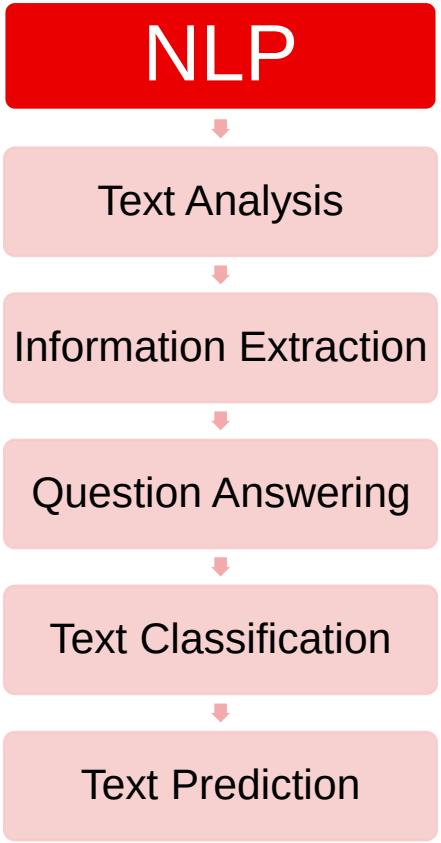
- Health Record System (EHR)
  - The Office of the National Coordinator for Health Information Technology
- Electronic Medical Record (EMR)
  - Electronic record of health-related information on an individual within one health care organization
- HHS Strategy on Reducing Regulatory and Administrative Burden Relating to the Use of Health IT and EHRs was released on February 21, 2020.



https://www.healthit.gov/playbook/

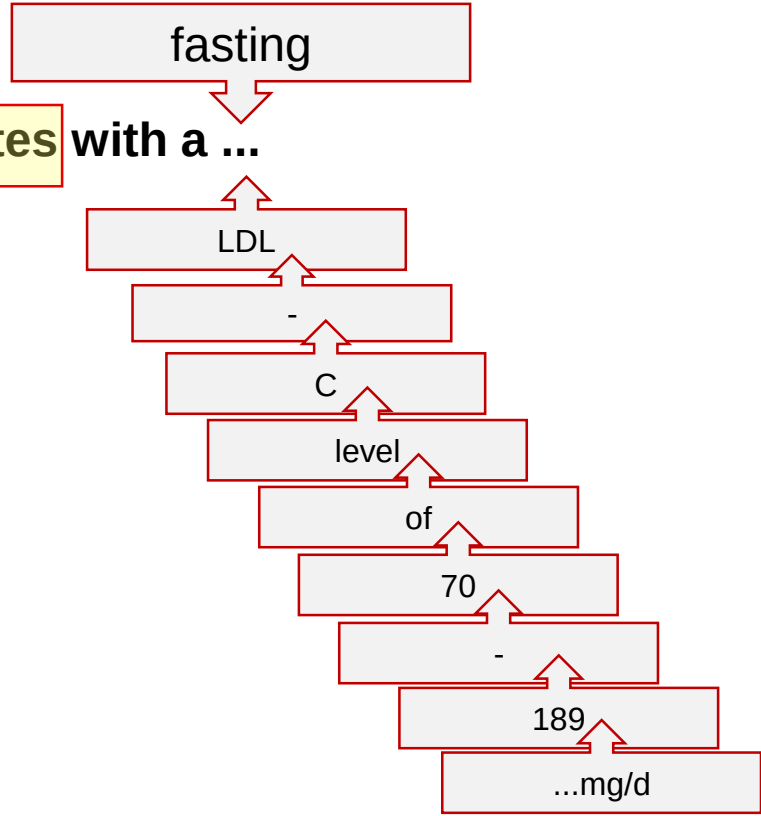
Utilizing NLP for Healthcare System





Language Modeling

Adults aged 40-75 years with a diagnosis of diabetes with a ...



Language Modeling

$k = 4$

Type 1 diabetes can develop at any age

However, some people with type 1 diabetes can develop insulin resistance.

type 1 diabetes can develop in people who have a particular HLA complex.

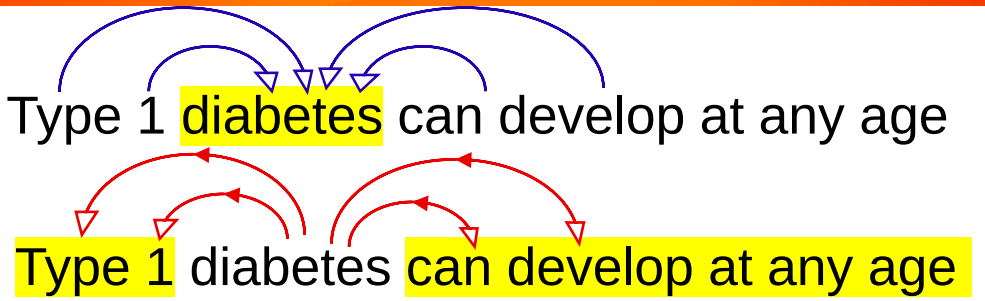
Detail about "diabetes" can be learned from the context (e.g., semantic similarity)

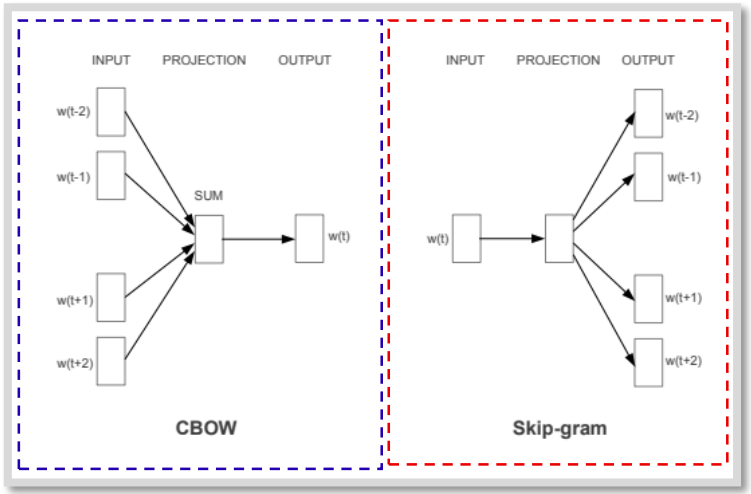
$$P(w_t|context)\ \forall t \in V$$

$$P(w_t|w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})$$

Language Modeling

- **Word2Vec**
- GloVe
- Transformer
- BERT





Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

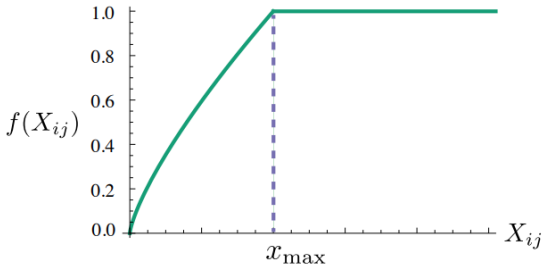
Language Modeling

- Word2Vec
- **GloVe**
- Transformer
- BERT

Frequent appearances of **ice** and **solid**

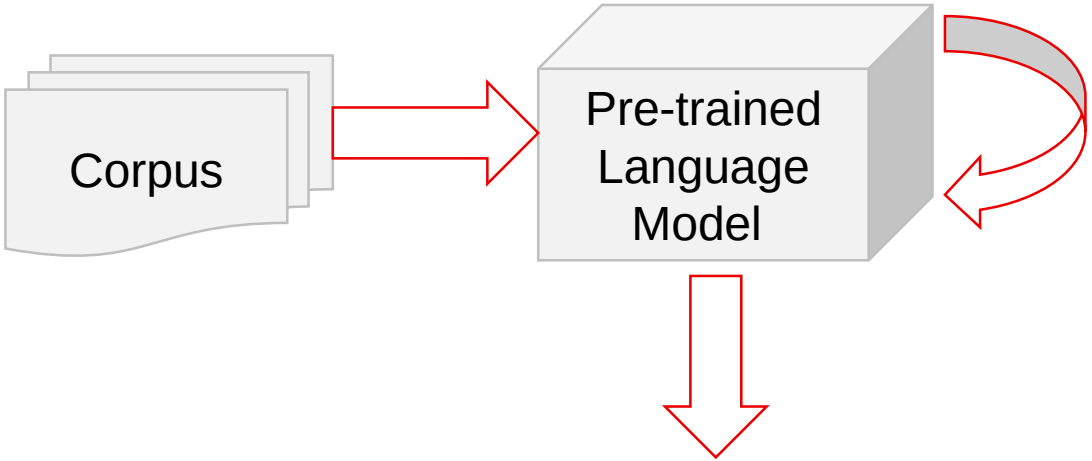
Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$



Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Text Analysis through Pre-trained Language Models

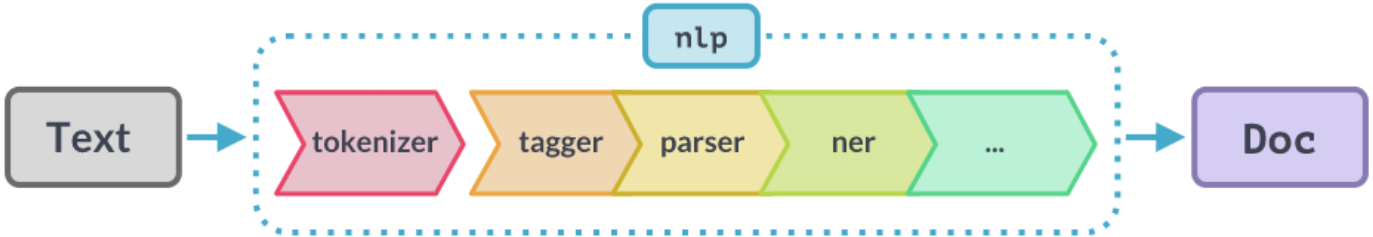


Tasks:

- Sentiment Analysis
- Similarity
- ...

Text Analysis

- Text Cleaning
- Lemmatization
- Tokenizer
- Semar



https://spacy.io/

Code #1

jupyter Tutorial_NLP_Summit_2022_c1 Last Checkpoint: 8 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Code

In [ ]:
```
import transformers
import tensorflow as tf
import keras as k
import sys
import seaborn as sns
import numpy as np
from spacy import displacy
import matplotlib.pyplot as plt
import logging
logging.basicConfig(level=logging.ERROR)
logger = logging.getLogger()
logger.setLevel(logging.ERROR)
```

In [ ]:
```
print(f"transormers:{transformers.__version__}")
print(f"tensorflow:{tf.__version__}")
print(f"keras:{k.__version__}")
```

Sample EHR

In [ ]:
```
#Ref: https://medicalcodify.com/eh/webchart.cgi?f=layoutnouser&func=&module=&tabmodule=&name=RXDBmain&searchterm
corpus= ["Percentage of the following patients - all considered at high risk of cardiovascular events - who were
```

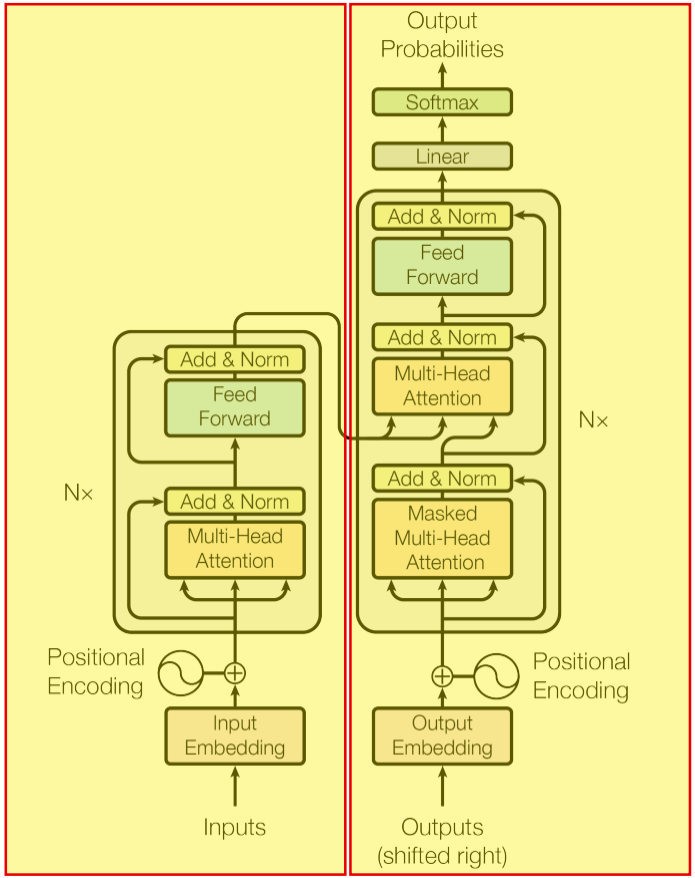
In [ ]:
```
corpus
```

Text Analysis

In [ ]:
```
import spacy
try:
    nlp = spacy.load("en_core_web_lg", )
except:
    !{sys.executable} -m spacy download en_core_web_lg
```

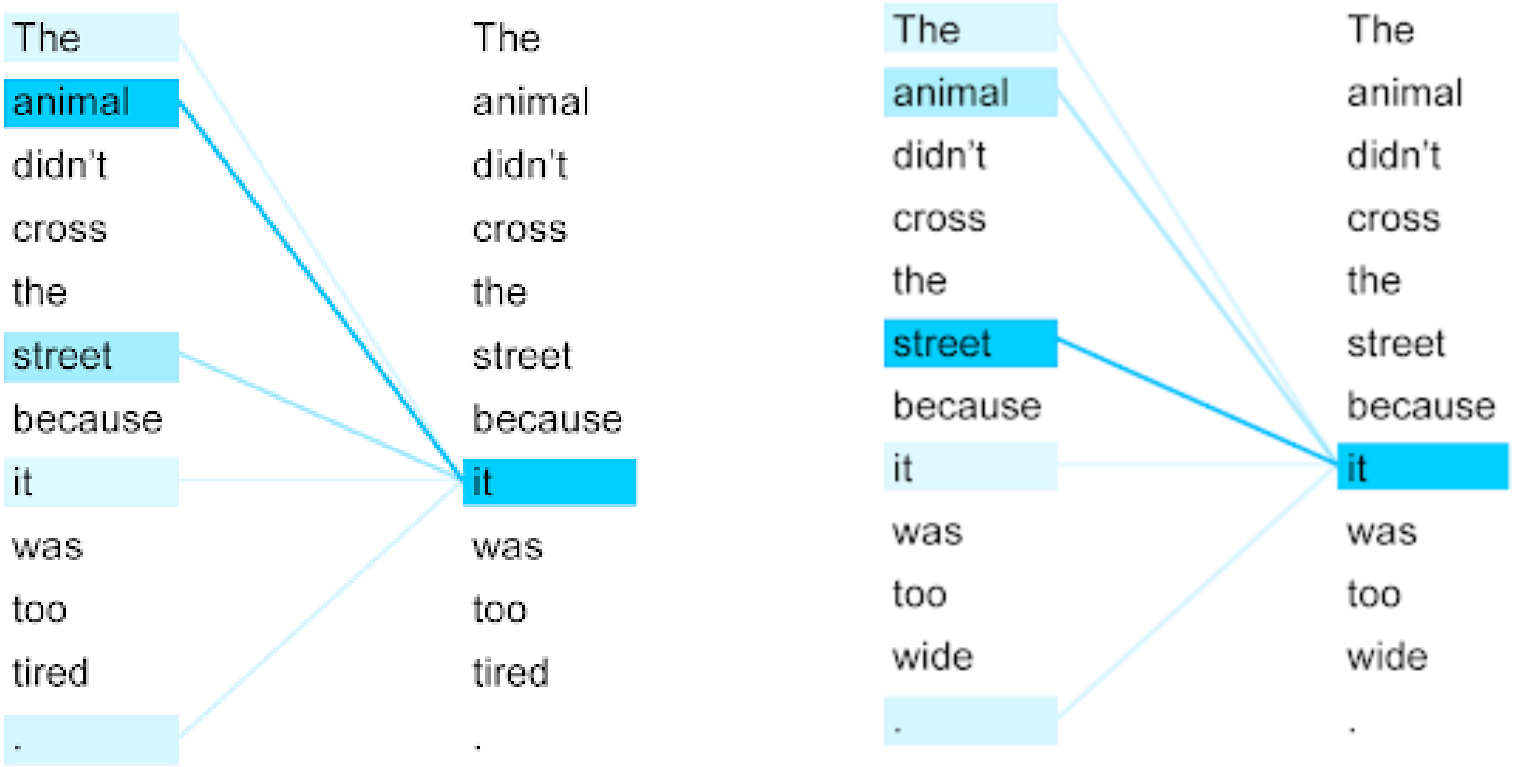
Language Modeling

- Word2Vec
- GloVe
- **Transformer**
- BERT



Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

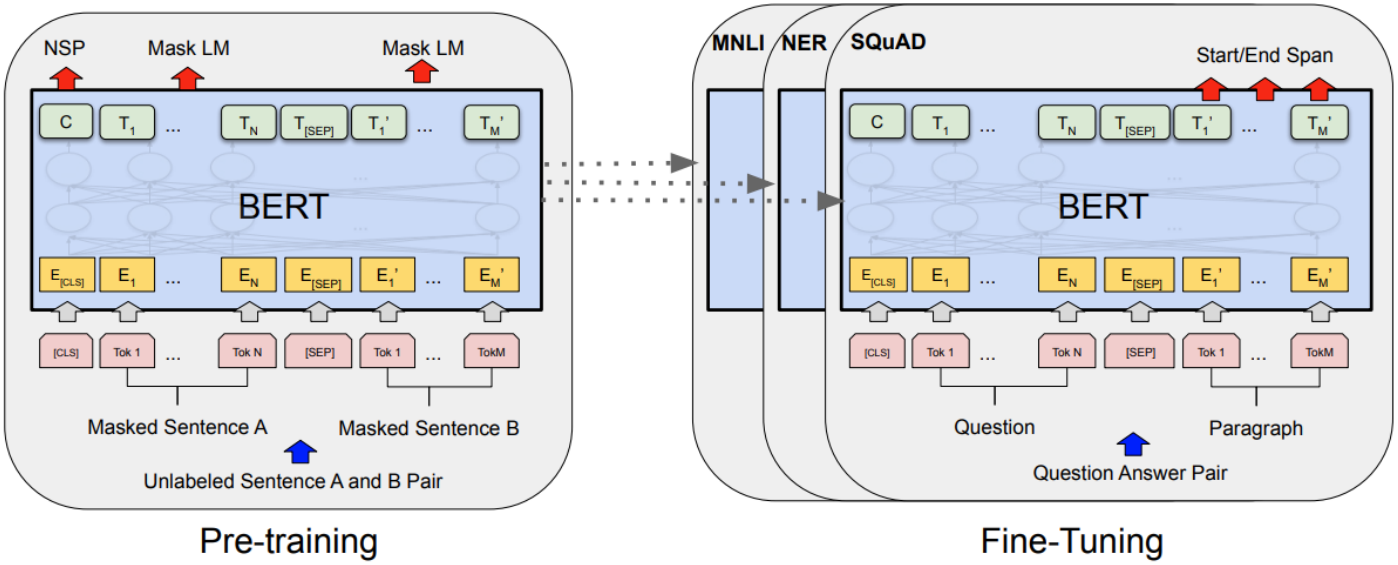
Transformer



https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

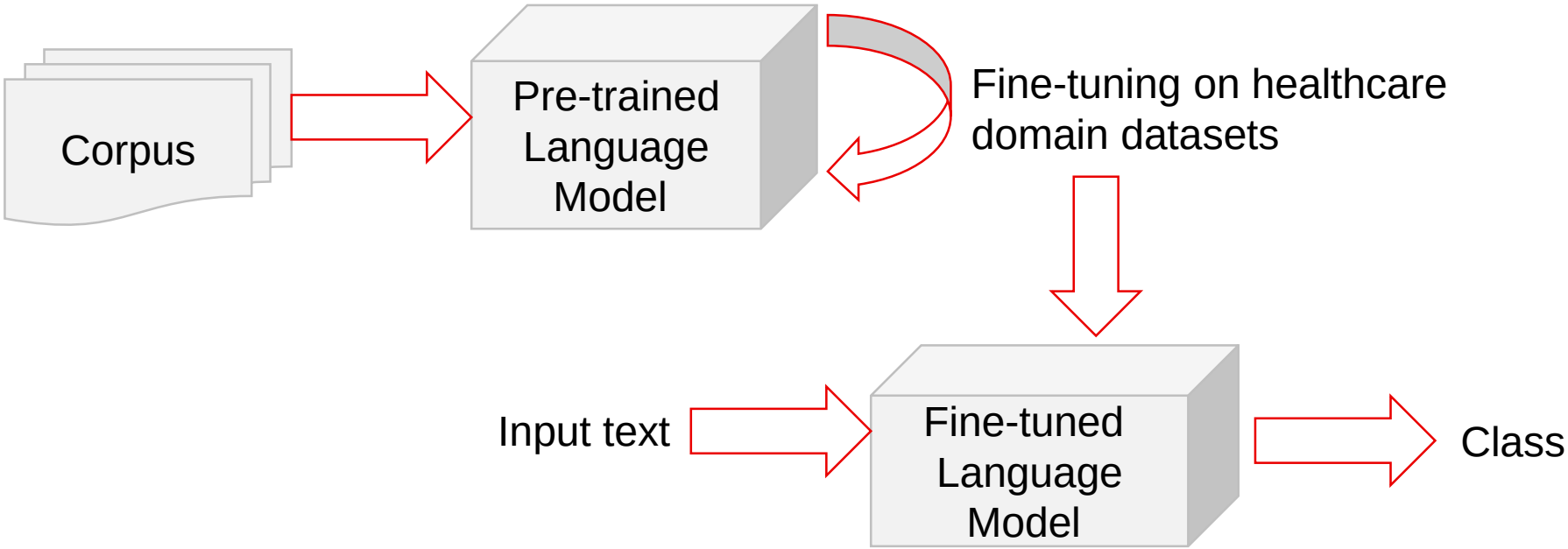
Language Modeling

- Word2Vec
- GloVe
- Transformer
- **BERT**



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Text Classification



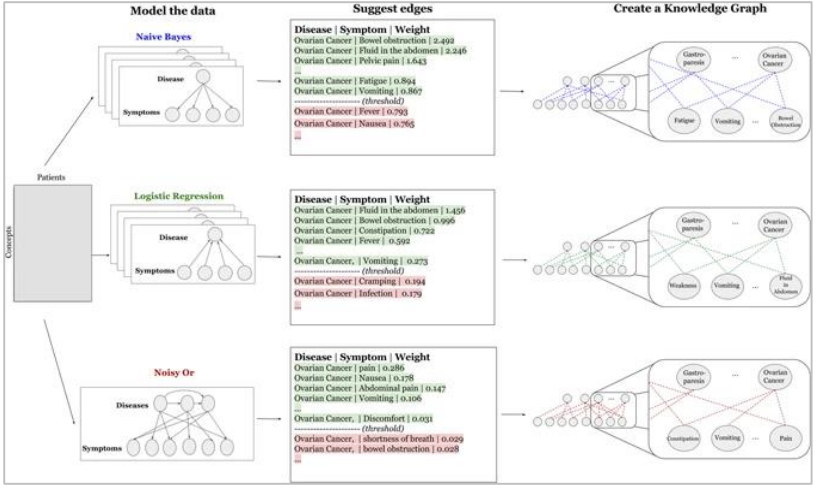
Question Answering

○ Structured Data

Y	filename	sex	age	age_years	corpus	group	child_TNW	child_TNS
1	922.cha	male	113	9.416666667	ENNI	SLI	1404	185
1	924.cha	male	112	9.333333333	ENNI	SLI	1162	125
1	926.cha	female	114	9.5	ENNI	SLI	729	80
1	928.cha	male	116	9.666666667	ENNI	SLI	592	63

Diagnose Specific Language Impairment in Children / Kaggle

Knowledge graph



Rotmensch, Maya, et al. "Learning a health knowledge graph from electronic medical records." *Scientific reports* 7.1 (2017): 1-11.

○ Unstructured Data

Description

Percentage of the following patients - all considered at high risk of cardiovascular events - who were prescribed or were on statin therapy during the measurement period: *Adults aged >= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR *Adults aged >= 21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level >= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR *Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL

https://medicalcodify.com/eh/webchart.cgi

Question Answering

○ Open Generative Question Answering

Description

Percentage of the following patients - all considered at high risk of cardiovascular events - who were prescribed or were on statin therapy during the measurement period: *Adults aged >= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR *Adults aged >= 21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level >= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR *Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL

Source: https://medicalcodify.com/eh/webchart.cgi

"What are previous diagnoses?"

Question Answering

- **Extractive Question Answering**

Description

Percentage of the following patients - all considered at high risk of cardiovascular events - who were prescribed or were on statin therapy during the measurement period: *Adults aged >= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR *Adults aged >= 21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level >= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR *Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL

Source: https://medicalcodify.com/eh/webchart.cgi

"What is the disease of the patients?"

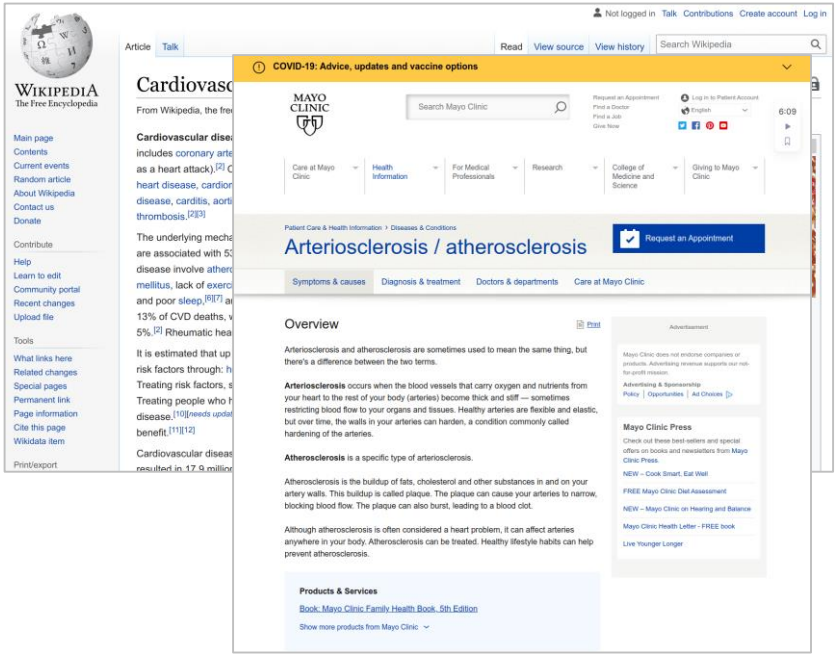
Question Answering

Closed Generative Question Answering

Description

Percentage of the following patients - all considered at high risk of cardiovascular events - who were prescribed or were on statin therapy during the measurement period: *Adults aged >= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR *Adults aged >= 21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level >= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR *Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL

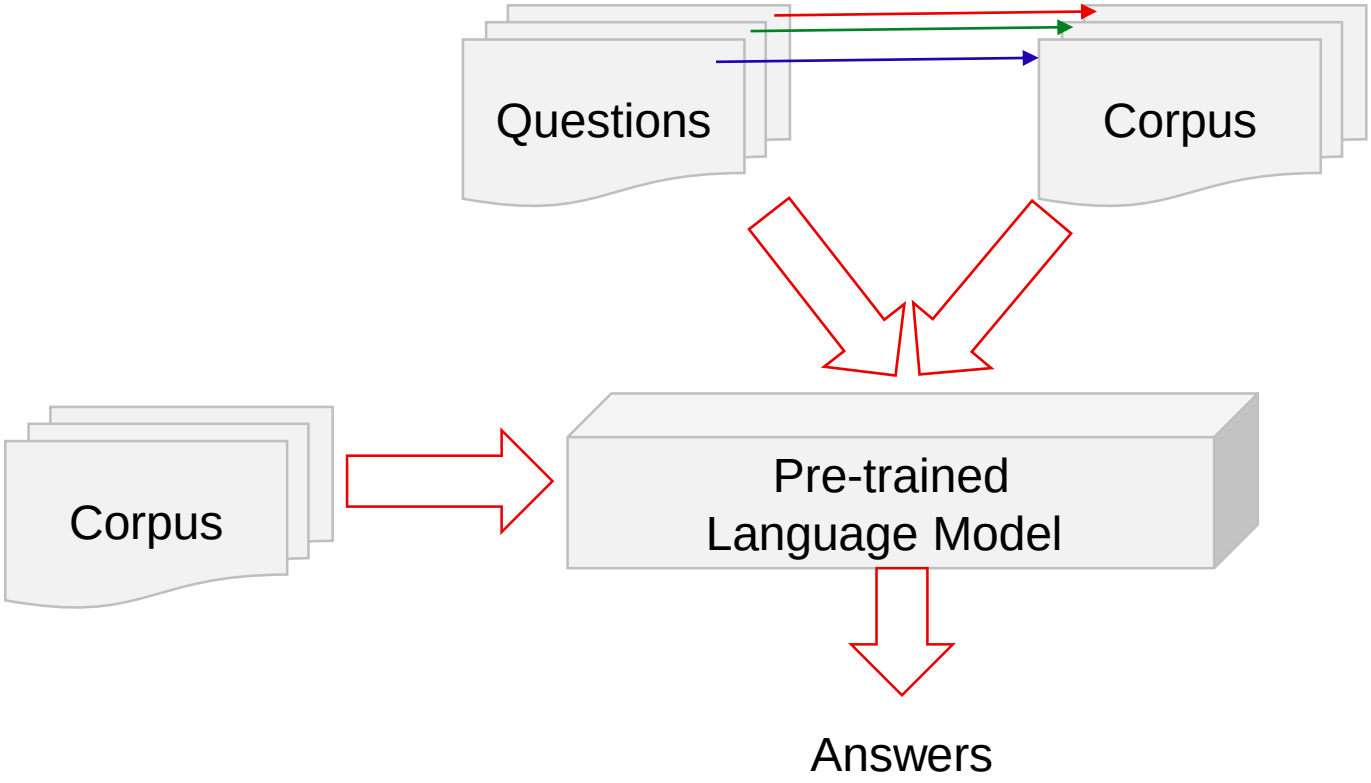
Source: https://medicalcodify.com/eh/webchart.cgi



"What are prevention methods for ASCVD?"

https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/symptoms-causes/syc-20350569
https://en.wikipedia.org/wiki/Cardiovascular_disease

Question Answering



Code #2

jupyter Tutorial_NLP_Summit_2022_c1 Last Checkpoint: 10 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Markdown

0.40659 , -0.73905 , 0.44078 , -0.012808], dtype=float32)

Question Answering

In [ ]:
```
from transformers import pipeline
```

In [ ]:
```
qa_model = pipeline("question-answering",
                    model="distilbert-base-cased-distilled-squad")
#for more detail refer to: https://github.com/huggingface/notebooks/blob/master/examples/question_answering.ipyn
```

In [ ]:
```
corpus[0]
```

In [ ]:
```
question_1="What are previous diagnosis?"
answer_1 = qa_model(question = question_1, context = corpus[0])
answer_1
```

In [ ]:
```
question_2="What is the disease of the patients?"
answer_2 = qa_model(question = question_2, context = corpus[0])
answer_2
```

In [ ]:
```
from sentence_transformers import SentenceTransformer, util
import numpy as np
```

In [ ]:
```
#ref: https://huggingface.co/allenai/biomed_roberta_base
model = SentenceTransformer("allenai/biomed_roberta_base")
```

In [ ]:
```
classifier = pipeline("text-classification", model = "roberta-large-mnli")
```

In [ ]:
```
mnli_context="Patients were previously diagnosed with atherosclerotic cardiovascular disease or were previously
```

In [ ]:
```
classifier(f"{mnli_context} Does patient previously diagnosed with diabetes?")
```

Text Classification

- **Entailment**

Description

Percentage of the following patients - all considered at high risk of cardiovascular events - who were prescribed or were on statin therapy during the measurement period: *Adults aged >= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR *Adults aged >= 21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level >= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR *Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL

Diagnoses Dataset
Heart Disease
Cancer
Diabetes
...

Code #3

jupyter Tutorial_NLP_Summit_2022_c1 Last Checkpoint: 11 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Notebook saved Trusted Python 3

Run Code

QA & Text Classification

In [ ]:
```
question_2, answer_2["answer"]
```

In [ ]:
```
corpus[0]
```

In [ ]:
```
diseases = ["Heart Disease", "Cancer", "Diabetes"]
```

In [ ]:
```
embeddings=[]
cosine_scores=[]
answer_embeding = model.encode(answer_2["answer"], convert_to_tensor=True)
for disease in diseases:
    target=model.encode(disease, convert_to_tensor=True)
    embeddings.append(target)
    cosine_scores.append(util.pytorch_cos_sim(target, answer_embeding))
```

In [ ]:
```
len(embeddings), embeddings[0].shape
```

In [ ]:
```
embeddings[0]
```

In [ ]:
```
cosine_scores
```

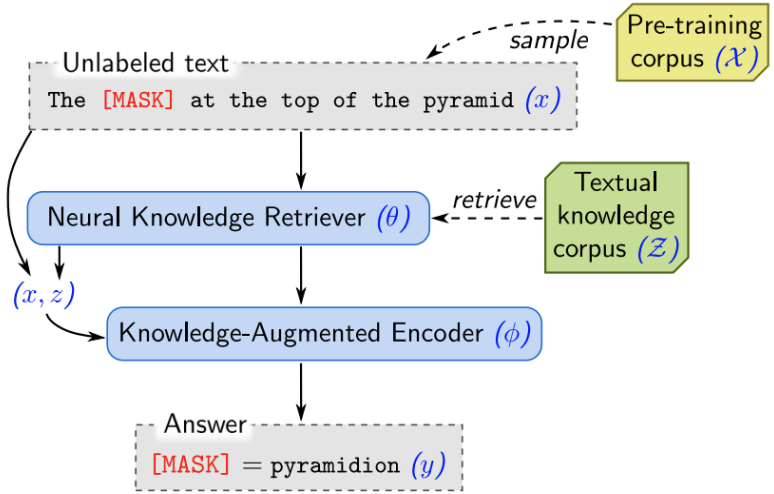
In [ ]:
```
for index in range(len(diseases)):
    print(f"{diseases[index]}:{cosine_scores[index]}")
```

In [ ]:
```
top_related_answer= np.argmax(cosine_scores)
diseases[top_related_answer]
```

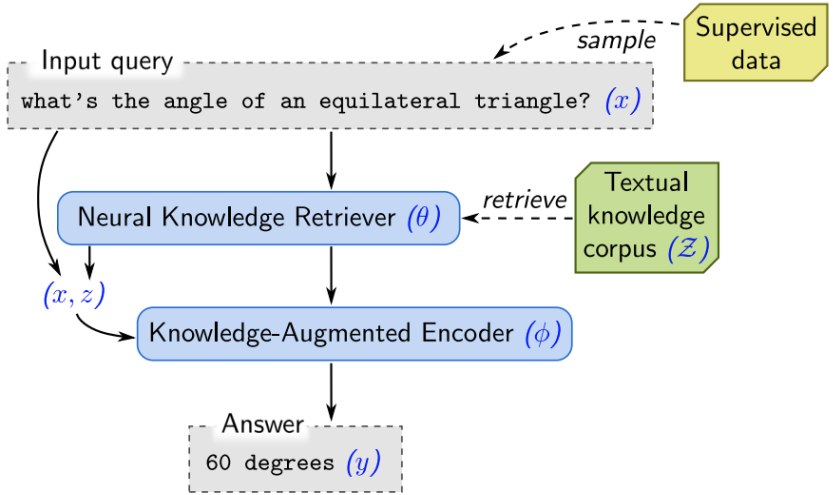
In [ ]:
```

```

QA & Text Classification / Training



Unsupervised pre-training



Supervised fine-tuning

Guu, Kelvin, et al. "Retrieval augmented language model pre-training." *International Conference on Machine Learning*. PMLR, 2020.

Code #4

jupyter Tutorial_NLP_Summit_2022_c1 Last Checkpoint: 12 hours ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [54]:
```
for index in range(len(diseases)):
    print(f"{diseases[index]}:{cosine_scores[index]}")
```

```
Heart Disease:tensor([[0.9217]])
Cancer:tensor([[0.9137]])
Diabetes:tensor([[0.9208]])
```

In [55]:
```
top_related_answer= np.argmax(cosine_scores)
diseases[top_related_answer]
```

Out[55]: 'Heart Disease'

Masked Language Model

In [ ]:
```
from transformers import pipeline
mlm = pipeline("fill-mask",
                        model="microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext")
```

In [ ]:
```
#original: all considered at high risk of cardiovascular
predicted_tokens = mlm("all considered at high [MASK]")
predicted_tokens = [_ for _ in predicted_tokens if _['token_str'] not in ['.', ',', ';']]
predicted_tokens
```

In [90]:
```

```

In [ ]:
```

```

Conclusion

- Based on use case, different language models can be used such as Word2Vec, GloVe, BERT, GPT
- Pre-trained language models can be used easily and works well across different domains
- The latest Pre-trained language models are expensive
- The pre-trained models can be fine-tune on specific down-stream tasks

Thank you

?

?Question?

?