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
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

Healthcare System
- EHR
  - Diagnoses
  - Treatments

NLP
- Text Analysis
- Information Extraction
- Question Answering
- Text Classification
- Text Prediction
Adults aged 40-75 years with a diagnosis of diabetes with a ...
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)

\[ P(w_t | \text{context}) \forall t \in V \]

\[ P(w_t | w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}) \]
Language Modeling

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

Type 1 diabetes can develop at any age

### Language Modeling

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

<table>
<thead>
<tr>
<th>Probability and Ratio</th>
<th>$k = solid$</th>
<th>$k = gas$</th>
<th>$k = water$</th>
<th>$k = fashion$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(k</td>
<td>ice)$</td>
<td>$1.9 \times 10^{-4}$</td>
<td>$6.6 \times 10^{-5}$</td>
<td>$3.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>steam)$</td>
<td>$2.2 \times 10^{-5}$</td>
<td>$7.8 \times 10^{-4}$</td>
<td>$2.2 \times 10^{-3}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>ice)/P(k</td>
<td>steam)$</td>
<td>8.9</td>
<td>$8.5 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

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

### Equation

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

---

Text Analysis through Pre-trained Language Models

Corpus

Pre-trained Language Model

Input text

Tasks:
• Sentiment Analysis
• Similarity
• ...

Results
Text Analysis

- Text Cleaning
- Lemmatization
- Tokenizer
- Semai

https://spacy.io/
Code #1

```python
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)
```

```python
print(f"transformers:{transformers.__version__}")
print(f"tensorflow:{tf.__version__}")
print(f"keras:{k.__version__}")
```

Sample EHR

```python
# Ref: https://medicalcodyf.com/eh/webchart.cgi?f=layoutuser&func=&module=&stabmodule=&name=RXDBmain&searchterm=

corpus= "Percentage of the following patients - all considered at high risk of cardiovascular events - who were 

```

Text Analysis

```python
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

The animal didn't cross the street because it was too tired.

The animal didn't cross the street because it was too wide.

Language Modeling

- Word2Vec
- GloVe
- Transformer
- BERT

Text Classification

- Corpus
- Pre-trained Language Model
- Fine-tuning on healthcare domain datasets
- Input text
- Fine-tuned Language Model
- Class
Question Answering

Structured Data

<table>
<thead>
<tr>
<th>filename</th>
<th>sex</th>
<th>age</th>
<th>corpus</th>
<th>child_TNW</th>
<th>child_TNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1922.cha</td>
<td>male</td>
<td>113</td>
<td>9.416666667 ENNI</td>
<td>SLI</td>
<td>1404</td>
</tr>
<tr>
<td>1924.cha</td>
<td>male</td>
<td>112</td>
<td>9.333333333 ENNI</td>
<td>SLI</td>
<td>1162</td>
</tr>
<tr>
<td>1926.cha</td>
<td>female</td>
<td>114</td>
<td>9.5 ENNI</td>
<td>SLI</td>
<td>729</td>
</tr>
<tr>
<td>1928.cha</td>
<td>male</td>
<td>116</td>
<td>9.666666667 ENNI</td>
<td>SLI</td>
<td>592</td>
</tr>
</tbody>
</table>

Diagnose Specific Language Impairment in Children / Kaggle

Knowledge graph

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
"What are previous diagnoses?"
"What is the disease of the patients?"

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

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
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://en.wikipedia.org/wiki/Cardiovascular_disease
Question Answering

Corpus → Pre-trained Language Model → Answers

Pre-trained Language Model

Corpus

Questions

Answers
Question Answering

```python
In [1]: from transformers import pipeline

In [2]: 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.ipynb

In [3]: corpus[0]

In [4]: question_1="What are previous diagnosis?"
   ...: answer_1 = qa_model(question = question_1, context = corpus[0])
   ...: answer_1

In [5]: question_2="What is the disease of the patients?"
   ...: answer_2 = qa_model(question = question_2, context = corpus[0])
   ...: answer_2

In [6]: from sentence_transformers import SentenceTransformer, util
   ...: import numpy as np

In [7]: #ref: https://huggingface.co/allenai/biomed_roberta_base
   ...: model = SentenceTransformer("allenai/biomed_roberta_base")

In [8]: classifier = pipeline("text-classification", model = "roberta-large-mnli")

In [9]: mnnli_context="Patients were previously diagnosed with atherosclerotic cardiovascular disease or were previously 
   ...: classifier(f"(mnnli_context) Does patient previously diagnosed with diabetes")
```
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
- ...

22
Code #3

```
# QA & Text Classification

In [ ]: 1 question_2, answer_2 = "answer"

In [ ]: 1 corpus[0]

In [ ]: 1 diseases = ['Heart Disease', 'Cancer', 'Diabetes']

In [ ]: 1 embeddings=[],
  2 cosine_scores=[],
  3 answer_embedding = model.encode(answer_2 = "answer", convert_to_tensor=True,
  4 for disease in diseases:
  5     target = model.encode(disease, convert_to_tensor=True)
  6     embeddings.append(target)
  7     cosine_scores.append(np.dot(embedding, answer_embedding))

In [ ]: 1 len(embeddings), embeddings[0].shape

In [ ]: 1 embeddings[0]

In [ ]: 1 cosine_scores

In [ ]: 1 for index in range(len(diseases)):
  2     print("diseases[index]: cosine_scores[index]")

In [ ]: 1 top_related_answer = np.argmax(cosine_scores)
  2     diseases[top_related_answer]
```
Unsupervised pre-training

The [MASK] at the top of the pyramid (x)


Supervised fine-tuning

Input query: "what's the angle of an equilateral triangle? (x)"

Answer: 60 degrees (y)
Code #4

```
In [54]:
  1) for index in range(len(diseases)):
  2)    print(f"{diseases[index]}:[cosine_scores[index]]")
  
Heart Disease:tensor([[0.9217]])
Cancer:tensor([[0.9137]])
Diabetes:tensor([[0.9208]])

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

Out[55]: 'Heart Disease'
```

**Masked Language Model**

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

In [ ]:
  1) #original: all considered at high risk of cardiovascular
  2) predicted_tokens = mlm("all considered at high [MASK]
  3) predicted_tokens = [token for token in predicted_tokens if token not in [',', '.', '']]

In [90]:

In [ ]:
```
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?