LEEDS Institute for Data Analytics

Research project: The use of natural language processing using CogStack to extract structured data from cancer pathology reports.

Introductory sentence:

80% of patient information is typically found within the free text documents of Electronic Health Records (EHR). Traditionally, this information is extracted manually which is time-consuming, inefficient and impractical at scale. This project aims to assess how Natural Language Processing (NLP) can be used to automate the extraction of relevant clinical information from EHR and the explainability of these NLP models.

Project overview

The introduction of the novel neural network architecture for sequence modeling i.e *Transformer* (Vaswani *et al.*, 2017), and transfer learning in NLP i.e *ULMFiT* (Howard and Ruder, 2018), has given birth to powerful language models such as *BERT* (Devlin *et al.*, 2019) which have revolutionized NLP from text generation to text classification, Named Entity Recognition, chatbots, and many more language problems. However, the widespread adoption of these methodologies in medical practice has been slow due to lack of expertise, intricacies of medical language, and skepticism due to lack of model explainability. This in turn has resulted in manual and rule-based information extraction techniques which are time-consuming and impractical at scale. This project aimed at:

- Assessing CogStack's MedCAT (Medical Concept Annotation Toolkit) and MedCATtrainer to annotate unstructured clinical texts such as pathology reports and clinical reports.
- Using the annotated texts to fine-tune a pre-trained clinical BERT model in order to perform tasks such as Text Classification and Named Entity Recognition.
- Assessing the model's predictions and explainability using Integrated Gradients.

Data and methods

Data

Due to unforeseen hurdles in accessing real-world pathology reports, this study resulted in using some open-source datasets from the CogStack repository (unannotated) and the BioNLP Shared Task 2013 dataset (annotated). The annotated dataset contained the following entities: *Cancer, Cell, Organ, Gene*.

Using MedCAT and MedCATtrainer to annotate unstructured texts

For the unannotated dataset, MedCAT and MedCATtrainer were used to annotate the individual texts for **Disease** concepts. MedCAT is an open-source Named Entity Recognition + Linking (NER+L) annotation tool (within the CogStack ecosystem) that can learn to extract concepts (e.g. disease, symptoms, medications) from free-text and link them to any biomedical ontology such as SNOMED-CT and UMLS. For example, the following hypothetical unstructured text:

```
"HISTORY OF PRESENT ILLNESS:, The patient is a 71-year-old Caucasian female with a history
of diabetes, osteoarthritis, atrial fibrillation, hypertension, asthma, obstructive sleep
apnea on CPAP, diabetic foot ulcer, anemia and left lower extremity cellulitis."
```

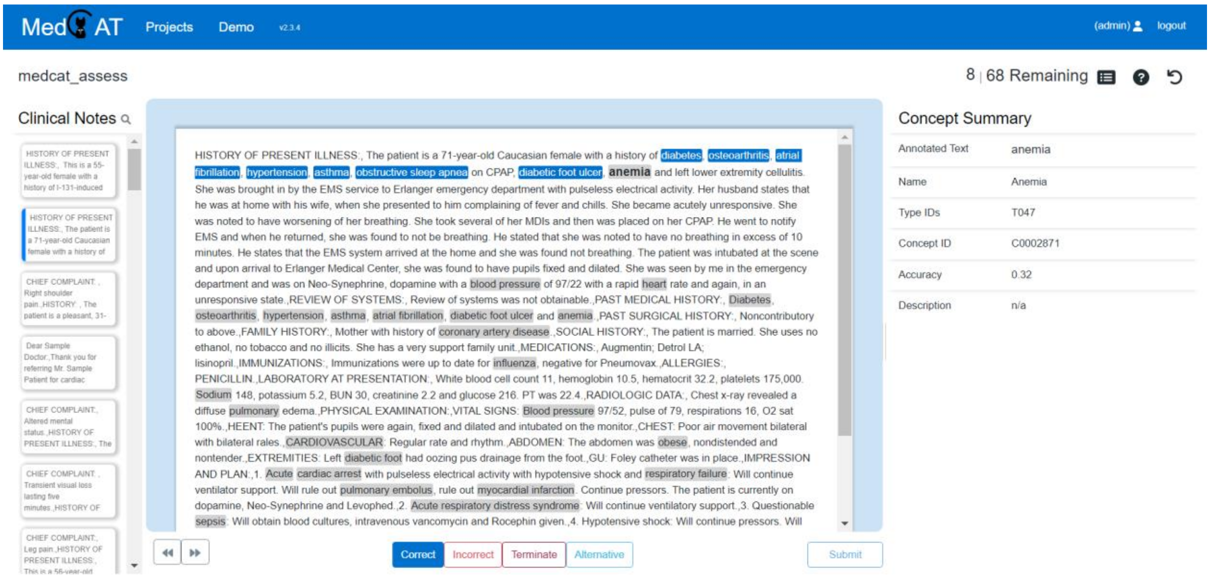
would be annotated as:

```
{'entities': {13: {'pretty_name': 'Diabetes',
   'cui': 'C0011847',
   'type_ids': ['T047'],
   'types': ['Disease or Syndrome'],
   'source_value': 'diabetes',
   'detected_name': 'diabetes',
   'acc': 0.5381721703086336,
   'context_similarity': 0.5381721703086336,
   'start': 93,
   'end': 101,
   'icd10': [],
   'ontologies': [],
   'snomed': [],
   'id': 13,
   'meta_anns': {'Status': {'value': 'Affirmed',
     'confidence': 0.9999875426292419,
     'name': 'Status'}}},
  14: {'pretty_name': 'Degenerative polyarthritis',
   'cui': 'C0029408',
   'type_ids': ['T047'],
   'types': ['Disease or Syndrome'],
   'source_value': 'osteoarthritis',
   'detected_name': 'osteoarthritis',
   'acc': 0.3615787619023934,
   'context_similarity': 0.3615787619023934,
   'start': 103,
   'end': 117,
   'icd10': [],
   'ontologies': [],
   'snomed': [],
   'id': 14,
   'meta_anns': {'Status': {'value': 'Affirmed',
     'confidence': 0.999998152256012,

     'name': 'Status'}}},

 ...........
```

As with all models, MedCAT too makes mistakes while annotating texts. MedCATtrainer allows domain experts to inspect, modify and improve a trained MedCAT model by either actively training an underlying MedCAT model or simply collecting and validating annotations extracted by a static MedCAT model. This is shown in Figure 1 below:



HISTORY OF PRESENT ILLNESS:, This is a 55-year-old female with a history of I-131-induced

HISTORY OF PRESENT ILLNESS:, The patient is a 71-year-old Caucasian female with a history of

CHIEF COMPLAINT: , Right shoulder pain.,HISTORY: , The patient is a pleasant, 31-

Dear Sample Doctor:,Thank you for referring Mr. Sample Patient for cardiac

CHIEF COMPLAINT:, Altered mental status.,HISTORY OF PRESENT ILLNESS:, The

CHIEF COMPLAINT: , Transient visual loss lasting five minutes.,HISTORY OF

CHIEF COMPLAINT:, Leg pain.,HISTORY OF PRESENT ILLNESS:, This is a 56-year-old

HISTORY OF PRESENT ILLNESS:, The patient is a 71-year-old Caucasian female with a history of diabetes, osteoarthritis, atrial fibrillation, hypertension, asthma, obstructive sleep apnea on CPAP, diabetic foot ulcer, **anemia** and left lower extremity cellulitis. She was brought in by the EMS service to Erlanger emergency department with pulseless electrical activity. Her husband states that he was at home with his wife, when she presented to him complaining of fever and chills. She became acutely unresponsive. She was noted to have worsening of her breathing. She took several of her MDIs and then was placed on her CPAP. He went to notify EMS and when he returned, she was found to not be breathing. He stated that she was noted to have no breathing in excess of 10 minutes. He states that the EMS system arrived at the home and she was found not breathing. The patient was intubated at the scene and upon arrival to Erlanger Medical Center, she was found to have pupils fixed and dilated. She was seen by me in the emergency department and was on Neo-Synephrine, dopamine with a blood pressure of 97/22 with a rapid heart rate and again, in an unresponsive state.,REVIEW OF SYSTEMS:, Review of systems was not obtainable.,PAST MEDICAL HISTORY:, Diabetes, osteoarthritis, hypertension, asthma, atrial fibrillation, diabetic foot ulcer and anemia.,PAST SURGICAL HISTORY:, Noncontributory to above.,FAMILY HISTORY:, Mother with history of coronary artery disease.,SOCIAL HISTORY:, The patient is married. She uses no ethanol, no tobacco and no illicits. She has a very support family unit.,MEDICATIONS:, Augmentin; Detrol LA; lisinopril.,IMMUNIZATIONS:, Immunizations were up to date for influenza, negative for Pneumovax.,ALLERGIES:, PENICILLIN.,LABORATORY AT PRESENTATION:, White blood cell count 11, hemoglobin 10.5, hematocrit 32.2, platelets 175,000. Sodium 148, potassium 5.2, BUN 30, creatinine 2.2 and glucose 216. PT was 22.4.,RADIOLOGIC DATA:, Chest x-ray revealed a diffuse pulmonary edema.,PHYSICAL EXAMINATION:,VITAL SIGNS: Blood pressure 97/52, pulse of 79, respirations 16, O2 sat 100%.,HEENT: The patient's pupils were again, fixed and dilated and intubated on the monitor.,CHEST: Poor air movement bilateral with bilateral rales.,CARDIOVASCULAR: Regular rate and rhythm.,ABDOMEN: The abdomen was obese, nondistended and nontender.,EXTREMITIES: Left diabetic foot had oozing pus drainage from the foot.,GU: Foley catheter was in place.,IMPRESSION AND PLAN:,1. Acute cardiac arrest with pulseless electrical activity with hypotensive shock and respiratory failure: Will continue ventilator support. Will rule out pulmonary embolus, rule out myocardial infarction. Continue pressors. The patient is currently on dopamine, Neo-Synephrine and Levophed.,2. Acute respiratory distress syndrome: Will continue ventilatory support.,3. Questionable sepsis: Will obtain blood cultures, intravenous vancomycin and Rocephin given.,4. Hypotensive shock: Will continue pressors. Will

Annotated Text	anemia
Name	Anemia
Type IDs	T047
Concept ID	C0002871
Accuracy	0.32
Description	n/a

Figure 1: MedCATtrainer validation interface

In this study, we validated annotations and downloaded them, ready to be used to *Fine-Tune/Supervise Train* a clinical BERT model. It is noteworthy that the same fine-tuning procedure can be applied to a MedCAT model. Additional details about the CogStack ecosystem can be found at (Kraljevic *et al.*, 2021).

Due to the limited number of unannotated documents, this study used the already annotated texts to illustrate the downstream processes of model fine-tuning, error analysis, and model interpretation.

Fine-Tuning a clinical BERT

In this study, the Bio+Clinical BERT model was used for Named Entity Recognition since it was shown to outperform BERT and Bio-BERT models in a variety of tasks (Alsentzer *et al.*, 2019). Bio+Clinical BERT (Alsentzer *et al.*, 2019) is a BERT model that was initialized from BioBERT and trained on all MIMIC notes. The model was trained for 12 epochs with a batch size of 32 and a learning rate of 0.001 using the train set as the training data. The model was then validated on the validation set and tested on the test set. The following results were obtained:

Metric	Value
F1-Score	0.898

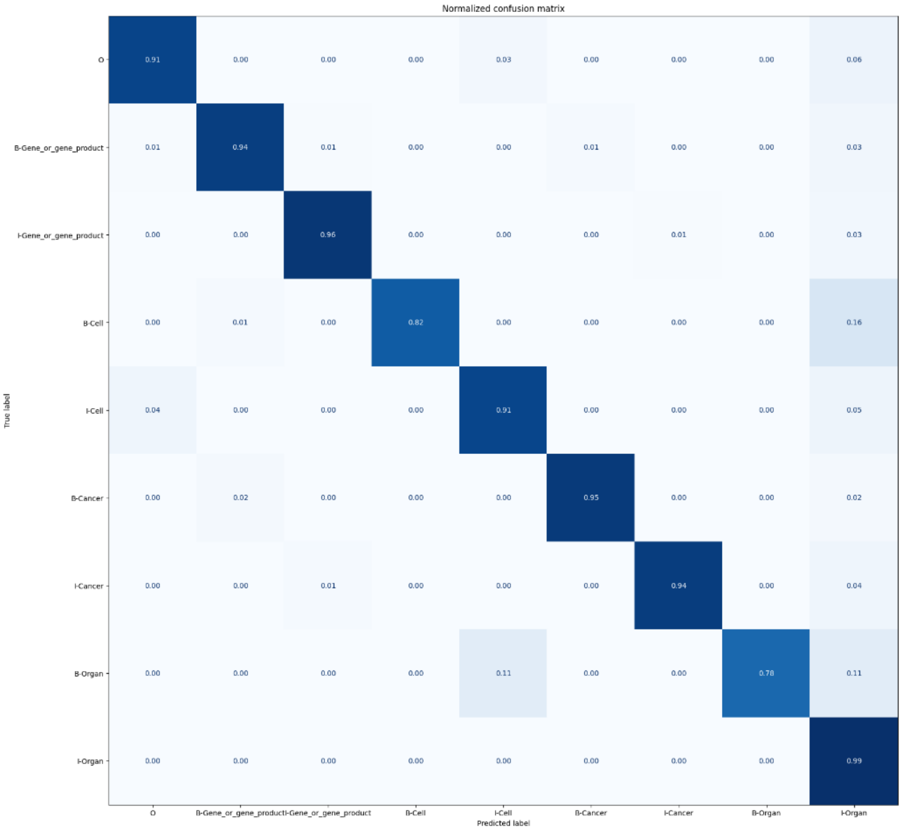


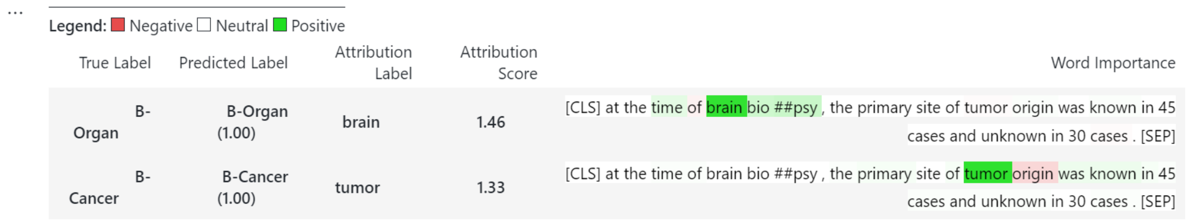
Figure 2: Confusion Matrix of the text entities

From the confusion matrix, we can see that the model tends to confuse B-Organ with I-Organ or I-Cell. Nonetheless, it does quite a good job at classifying the rest of the entities.

Model Interpretation

This study also explored why the model makes the predictions it does using a technique called integrated gradients. Integrated Gradients is a model interpretability algorithm that assigns an importance score to each input feature by approximating the integral of gradients of the model's output with respect to the inputs along the path from given references to inputs. Model interpretation was explored both at the individual sentence level and at the entire class entity.

At the sentence level, for each token of the sentence, the predicted entity is considered as the target and the attributions of the rest of the tokens are calculated. The attribution scores indicate which tokens contributed positively or negatively to the predicted class of the token. An example is shown below:



...

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-Organ	B-Organ (1.00)	brain	1.46	[CLS] at the time of brain bio ##psy , the primary site of tumor origin was known in 45 cases and unknown in 30 cases . [SEP]
B-Cancer	B-Cancer (1.00)	tumor	1.33	[CLS] at the time of brain bio ##psy , the primary site of tumor origin was known in 45 cases and unknown in 30 cases . [SEP]

At the class entity level, we aggregate the word attribution scores for a given entity over multiple random data splits and instances of the model, in order to extract the overall most predictive word features of a particular entity. As an example, the table below shows the top 10 keywords for the *Cancer* entity:

Keyword	Score
tumor	0.6512
resistant	0.6452
tumors	0.6061
antitumour	0.5747
oscc	0.5604
solid	0.3955
neoplasm	0.3848

This is still very much a work in progress since there is generally little work done to explore explanations of Clinical BERT models. Detailed theoretical and mathematical underpinnings of our quest to find stable word attributions for a given entity are discussed in an upcoming paper.

The code and documentation for the methodologies described above can be found at: https://github.com/R-icntay/cogstack_project

Key findings

- Preliminary results indicate that out of the box, MedCAT can perform Named Entity Recognition + Linking for most medical concepts. However, some subtle concepts such as *HER2* status in breast cancer could not be annotated out of the box. This can however be easily resolved by updating MedCAT's vocabulary and concept data base. Such flexibility allows MedCAT and MedCATtrainer to be adapted in the extraction of even domain-specific concepts such as breast cancer phenotypes.

- Once relevant clinical information has been extracted from EHR, various downstream clinical research can be done through the newly structured data and even other tools outside the CogStack ecosystem. For instance, using tools such as HuggingFace transformers and Captum, this study trains and assesses why a Clinical BERT model makes the predictions it does, by identifying the overall most predictive features of an entity.

Research project: The use of natural language processing using CogStack to extract structured data from cancer pathology reports.

Value of the research

The aim of this project was generally to assess whether and how CogStack could have transformative effects on clinical care delivery and clinical research. On real-world data and adapting some of the methodologies outlined in this project, this research could support clinical care delivery and research in the following ways:

a. Extracted data using Cogstack's MedCAT and MedCATtrainer could be used as the basis of summary information provided to clinicians as part of the delivery of routine care.
b. Structured data and linguistic concept mappings could be used to facilitate enhanced document search functionality within health care records to ensure key documents are easier to find for clinicians.
c. Clinically relevant extracted data could provide an excellent base for downstream clinical research tasks such as creating clinical models for classification, named entity recognition, text summarization, etc.

Insights

a. The work done in this project provided a reference point on how future studies could use CogStack to automate the extraction of relevant clinical concepts from unstructured datasets, which can improve clinical care delivery and clinical research.
b. An upcoming paper based on this study investigates global model explanations of how a clinical BERT model perceives an entire entity in a Named Entity Recognition task which can be helpful for clinicians to probe what are the most informative keywords for a given entity according to the model and whether this is clinically sensible.

Research theme

Health

People and Partners

Eric Wanjau, Data Scientist, Leeds Institute for Data Analytics, University of Leeds, UK

Geoff Hall, Professor of Digital Health and Cancer Medicine

Dr Serge Sharoff, Faculty of Arts, School of Languages

Dr Kieran Zucker, Clinical Research Fellow, LIMR, SoM Milton Hoz de Vila Eduardo, School of Computing

Funders

This work received funding from the Health Data Research UK (HDR-UK).

References

Alsentzer, E. *et al.* (2019) ‘Publicly Available Clinical’, in. doi: 10.18653/v1/w19-1909.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Howard, J. and Ruder, S. (2018) ‘Universal language model fine-tuning for text classification’, in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. doi: 10.18653/v1/p18-1031.

Kraljevic, Z. *et al.* (2021) ‘Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit’, *Artificial Intelligence in Medicine*. doi: 10.1016/j.artmed.2021.102083.

Vaswani, A. *et al.* (2017) ‘Attention is all you need’, in *Advances in Neural Information Processing Systems*.