Using Bidirectional Encoder Representations from Transformers (BERT) to remove personal data from healthcare records.

Ewart J Sheldon1,2, Anthony Shek3, Mohammed Al-Agil1, Vlad Dinu2, Sophie E Maxey1,2, Clodagh H McGuire1,2, Phil Davidson1,2, James TH Teo1

1: King’s College Hospital, Denmark Hill, Brixton, London SE5 9RS, 2: King’s College London, Strand, London WC2R 2LS, 3: The University of Manchester, Oxford Rd, Manchester M13 9PL. *ewart.sheldon@nhs.net

Introduction

- Patient data is stored in electronic healthcare records (EPR) which contain large amount of data that is valuable for research.
- Before this personal data can be used for secondary purposes it must be de-identified.
- This is a legal requirement and a patient safety concern.
- De-identification is currently done using rule-based approaches but this will miss novel situations like shortened nicknames, misspelt names, fragmented addresses.
- A natural language processing method such as BERT could help this as it looks at the context of words to identify personal data terms.

De-identification concept database

- UK government guidelines were used to specify key types of data that needed to be annotated (these comply with the NHS England confidentiality policy, and are similar to US Safe Harbour legislation):
  - Dates of Birth
  - Contact details
  - Names
  - Identifiers
  - Healthcare identifiers
  - Data was associated with either:
    - Patients
    - Relatives
    - Healthcare professionals (anyone involved in the care pathway)

Methods

- A concept database of personal data terms was developed (Fig 1)
- MedCAT and MedCATtrainer was used to assist in annotating identifiable data in unstructured clinical text documents (inpatient clinical notes & outpatient letters).
- This project is Governed by the patient-led KERRI committee at Kings College Hospital which audits adequate de-identification of clinical text.

Dataset description

- 2667 Documents extracted from Cogstack
- GP A&E Letters/Clinical note sections from 01/2016 – 12/2020 were uploaded into MedCATtrainer and annotated
- 56,128 annotations were generated
- 11,551 unique terms
- These terms related mainly to the healthcare professionals in the notes (53.95%), patients (45.88%) and the rest were relatives or careers (0.17%)
- All terms >1000 will be used to train BERT

Fig 1. Diagram of the de-identification hierarchical ontology structure. Five broad terms serve as the primary child nodes to which more specific terms child nodes, e.g. postcode, were connected. The terminal nodes were used to annotate documents with patient data terms.

Current/Future work

- The annotation output has been converted into BERT NER input and BERT will be trained on the data
- When BERT has annotated documents the patient data terms will be replaced by database concept name in patient documents
- Will be used with existing rules based systems to create a multi-layer anonymization system

Conclusion

- We’ve been able to build a patient data concept database
- We are now training BERT to replace these terms, intext, with these concepts. To retain the contextual information of the document for future downstream information extraction tasks.

References