A short introduction to the NHS Language Corpus project

Dr. Dan Schofield
Data Scientist
Analytics Unit - Innovation | NHSX

March 2021

An introduction to NHSX

NHSX is a joint unit bringing together teams from NHS England and NHS Improvement, and the Department of Health and Social Care to **drive the digital transformation of health and care**

We are a diverse team with a range of skills and expertise, including clinicians, technologists, policy experts, developers, data scientists and project managers

Motivation

Different language task settings require different training data or fine-tuned solutions: **Mental Health** notes are often long form, the **NHS.UK website** uses accessible language, **EHR entries** may be of variable length and contain technical abbreviations - ***NHS specific datasets***

Many modern large-scale Language Models gain success from exposure to **large amounts of unlabelled data** with various styles and sources and can then be fine-tuned to specific tasks

Successful Natural Language Processing (NLP) underpins many other important tasks:

- Automatic speech recognition and generated response
- Multi-modal approaches in model explainability in medical imaging
- Automated clinical coding
- And many more...

And some more...

Lots of medical text is **very sensitive** and takes significant effort to make safe to share (i.e. robust and safe de-identification is hard)

Some text data will be semi-structured, others fully unstructured, and could have been collected for different downstream tasks, so bringing them together would require the **appropriate metadata**

It is not always easy to quantify the **exact benefit gained** from task-specific training data to other tasks

Curated datasets used for research purposes **cannot always be shared further**

Some datasets will contain various **biases** that would need to be addressed

Testing the water… starting small

NHS Language Corpus Discovery - Four weeks with a small multidisciplinary team (luckily two pizzas is fine) working with open tools

Ingest examples of **open sources** from the internet - e.g. from NHS.UK, NHS Data Dictionary, etc.

Understand what **metadata or enrichment** is useful - is there a SNOMED code for that?

How best to **share** the outputs - can we, should we, how best to share?

What can we **learn** from the exercise?

Keep **building out** in the future - other internal projects may produce useful data sets

Vision

The **NHS Language Corpus** would look to be:

Open

Important to make this resource easily available to innovators and researchers in NLP healthcare space

Representative

Contains a range of sources and variability in language used in a given setting

Extensible

Collect a dataset that has a wide coverage as well as a large number of examples over time

Useful

Adds to the currently available resources constructively

What next?

Can add to this **accessible and shareable** datasets containing text from more clinical settings?

e.g. MIMIC, MedMentions

Collect examples of **synthetic text or successful de-identification** that have already been through Information Governance sign-off to learn from (both techniques and challenges)

Examples of datasets which sit closer to **patient-centred** interactions with the NHS

Ideas for data sources to add? Want to feed into enriching our **user stories**?

Work in the area of **speech-to-text** - who, how, where?

Questions, feedback, thoughts, or suggestions?

Connect with us

Web: **www.nhsx.nhs.uk**

Email: **daniel.schofield@nhsx.nhs.uk**

github.com/nhsx

@NHSX