Breaking barriers in applying NLP on Clinical Text: Practical experience in benchmarking NLP tools for radiology

ARLENE CASEY, RESEARCH FELLOW CLINICAL NLP, UNIVERSITY OF EDINBURGH

ARLENE.CASEY@ED.AC.UK

Clinical Natural Language Processing Research Group


Honghan Wu


Bea Alex


William Whiteley

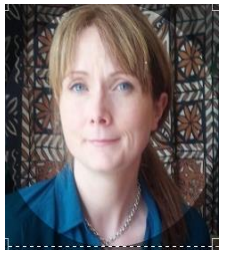

Grant Mair


Heather Walley


Michael Poon


Víctor Suárez-Paniagua


Hang Dong


Emma Davidson


Archie Campbell


Huayu Zhang


Richard Tobin


Andreas Grivas


Claire Grover


Matúš Falis

https://www.ed.ac.uk/usher/clinical-natural-language-processing @EdiClinicalNLP

Edinburgh CLINICAL CLINICAL CLINICAL NLP

Overview

- Breaking barriers in applying NLP on Clinical Text
- Practical experience in benchmarking NLP tools for radiology

Edinburgh CLINICAL CLINICAL CLINICAL NLP

Overview

- Breaking barriers in applying NLP on Clinical Text
- Practical experience in benchmarking NLP tools for radiology

A Systematic Review of Natural Language Processing Applied to Radiology Reports[1]

- Time Period: Jan 2015 Oct 2019
- Automated screening (4,799 ->397)
- Manual screening(274 -> **164**)
- Extracted 21 variables
  - anatomical region, NLP methods, data sizes, results, data and code availability

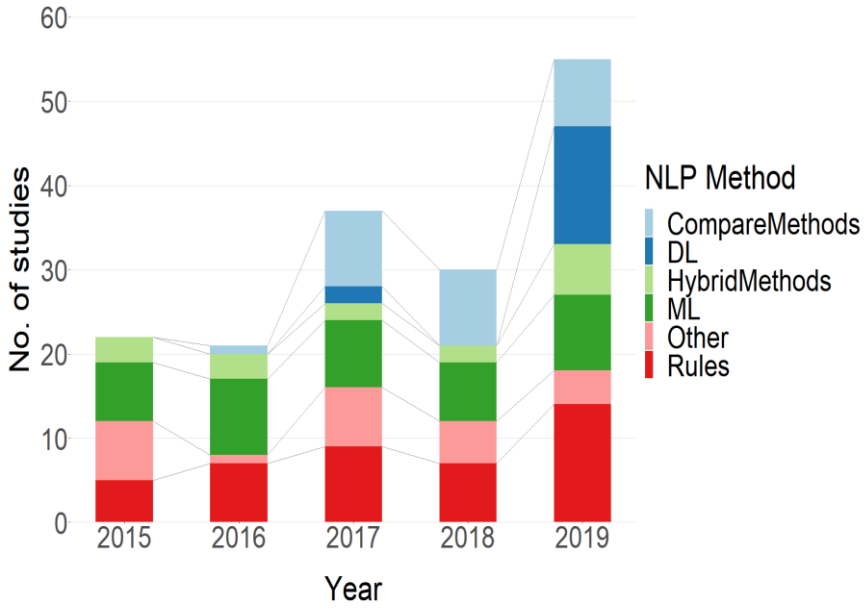
Observations on:

NLP Methods

Metrics & Reporting

[1] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu and Beatrice Alex (2021). **A Systematic Review of Natural Language Processing Applied to Radiology Reports**. 2102.09553.pdf (arxiv.org)

A Systematic Review of Natural Language Processing Applied to Radiology Reports[1]



DL – Use deep learning only
ML – Use machine learning (no deep learning),
Hybrid – combined different methods in overall solution
Compare – comparison of using methods in paper
Rules – rule-based only

Observations on:

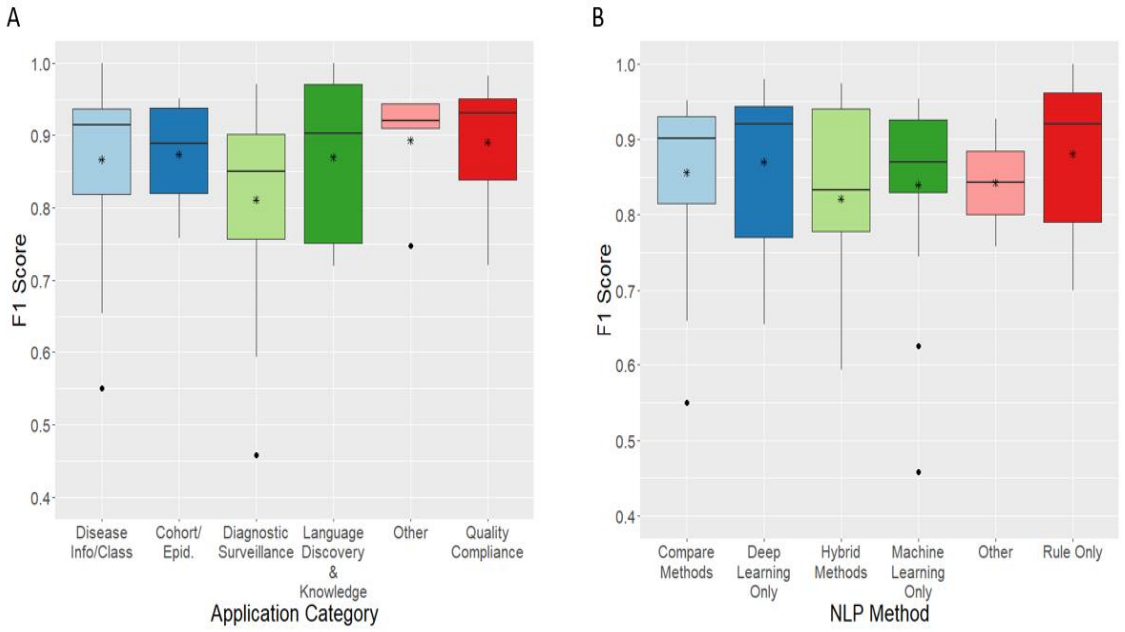
NLP Methods

Metrics & Reporting

- Deep learning increases but traditional methods still remain popular
  - Interpretability is more challenging with deep learning
  - Data scarcity plays a role
- Reproducibility for NLP is important
  - Data available (14) with most using single institution data
  - Externally Validated (10)
  - Code Available (14)

[1] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu and Beatrice Alex (2021). **A Systematic Review of Natural Language Processing Applied to Radiology Reports**. 2102.09553.pdf (arxiv.org)

A Systematic Review of Natural Language Processing Applied to Radiology Reports[1]



Mean and Median Values for F1 scores across Category and NLP Method
Vertical bar is mean value, * is the median value

Observations on:

NLP Methods

Metrics & Reporting

Comparison needs assessment on the same metrics, annotated data and comparable outcome

- Extremely difficult due to the heterogeneity of reporting
  - Metrics reported varied widely
  - Annotation varies widely
  - Outcomes vary in granularity and convention

[1] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu and Beatrice Alex (2021). **A Systematic Review of Natural Language Processing Applied to Radiology Reports**. 2102.09553.pdf (arxiv.org)

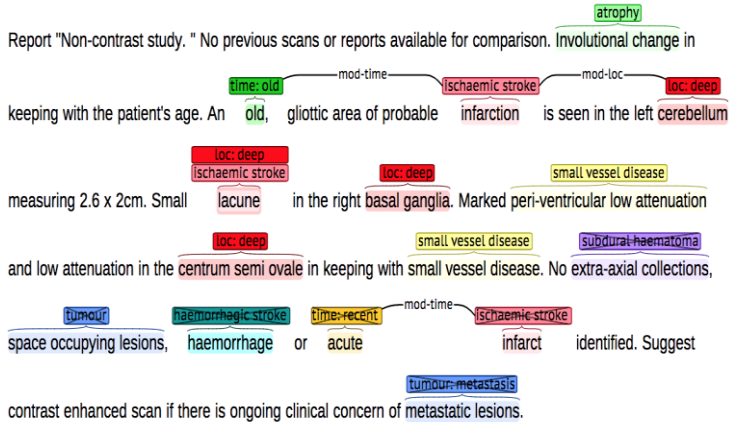
Breaking Barriers - Improving reproducibility and ultimately clinical application

Our role as researchers / practitioners in supporting

- Improving reproducibility
  - sharing of code and data
  - common conventions
  - reporting
- Shared tasks

Practical Experience in Benchmarking NLP Tools for Radiology

Comparison of NLP tools used for identifying cerebrovascular phenotypes on radiology reports



Using both brain image labelling and radiology report labelling as ground truth.

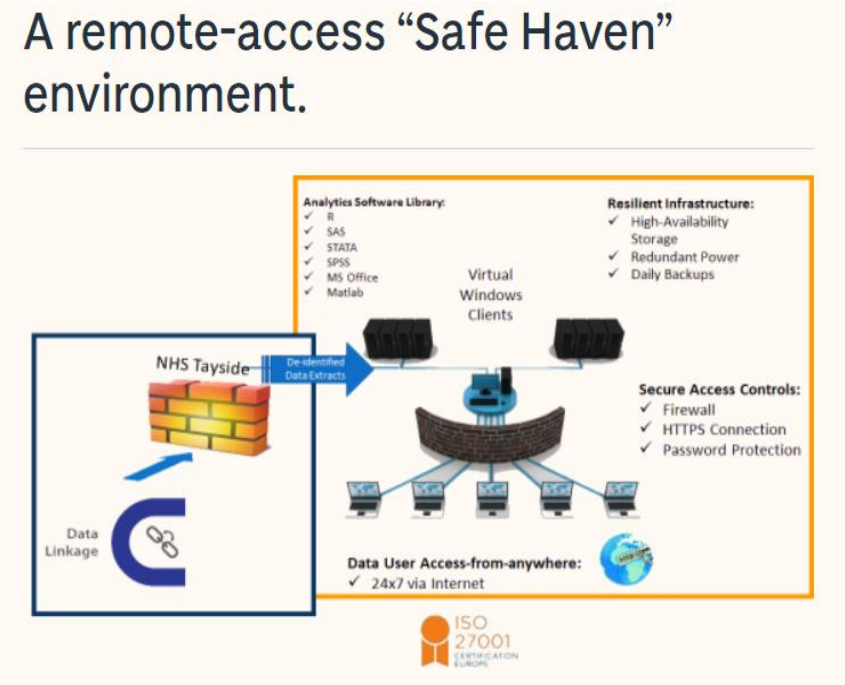
Project, Data and Access

The brain scans and reports have been obtained as part of a project to investigate patients presenting with delirium to NHS Fife. (CT scans)

In collaboration with:
Vera Cvoro, Delirium project lead Stroke Medicine, NHS Fife
Karen Ferguson, Neuroradiology, Edinburgh – Image reads

After data pre-processing:
2,345 records, just over 87% are over 70

Data located in a SafeHaven, Health Informatics Centre, University of Dundee



https://www.dundee.ac.uk/hic/hicsafehaven/

NLP Tools – Edinburgh Clinical NLP Tools

EdIE-R

MRI/CT scans
rule-based
24 phenotypes

EdIE-N

MRI/CT scans
neural model (bi-LSTM+CRF)
24 phenotypes

SemEHR

UMLS concepts mapped to entity types,
NLP2Phenome maps to phenotype labels

Named Entities	EdIE-R			EdIE-N			SemEHR			Inter-Annotator Agreement		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ESS test (266 reports)	**0.94**	**0.96**	**0.95**	0.82	0.92	0.86	0.91	0.94	0.92	0.96	0.98	0.97
Tayside test (700)	**0.99**	**0.95**	**0.97**	0.80	0.91	0.85	0.94	0.87	0.91	*0.95	*0.96	*0.96
Tayside+ test (300)	**0.94**	**0.91**	**0.93**	0.76	0.85	0.80	0.89	0.88	0.89	-	-	-

Precision (P),Recall (R), F1 (F1-score),*IAA on a sub-part of Tayside reports

NLP Tools from Others

ESPRESSO[1]

CT/MRI scans, U.S. Data
rule-based system using MedTagger (Mayo Clinic)
Silent brain infarct, White Matter Disease +Severity

Phenotype	Sensitivity	Specificity
Silent Brain Infarct	0.925	1.000
White Matter Disease	0.942	0.909

Ong et al.[2]

MRI, MRA,CT, CTA scans, U.S. data
GloVe embedding/RNN
Ischaemic stroke, MCA territory involvement, stroke acuity

Phenotype	Sensitivity	Specificity
Ischaemic Stroke	0.902	0.872
MCA	0.902	0.911
Acuity	0.911	0.689

[1]Fu, Sunyang & Leung, Lester & Wang, Yanshan & Raulli, Anne-Olivia & Kallmes, David & Kinsman, Kristin & Nelson, Kristoff & Clark, Michael & Luetmer, Patrick & Kingsbury, Paul & Kent, David & Liu, Hongfang. (2018). Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. 7. 10.2196/12109.

[2] Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports
Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, et al. (2020) Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLOS ONE 15(6): e0234908. https://doi.org/10.1371/journal.pone.0234908

NLP Tools from Others ?

ALARM[3]

MRI scans
Neural(bioBERT +custom attention)
abnormal/normal or 1 of 5 categories

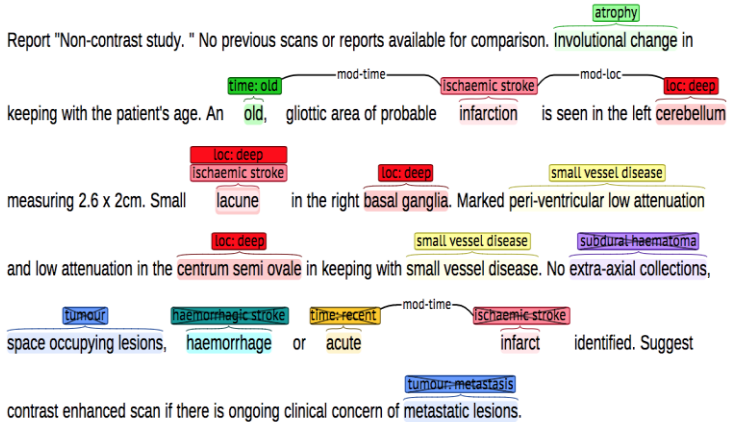
Named Entities	Normal /abnormal		Damage		Vascular		Mass		Acute Stroke		Fazekas	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
ALARM	**99.1**	**99.6**	92.6	94.3	**96.1**	95.7	**92.6**	96.4	94.5	100	**100**	99.3
Expert	77.2	98.9	96.2	97.1	84.6	99.3	77	100	97.2	100	96.1	100

[3] David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, Thomas C. Booth **Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM)** arXiv:2002.06588

Challenges:
Data, NLP Tools, Outcomes & Environment

Brain image observations need to be mapped to NLP tool outcomes – rule-based

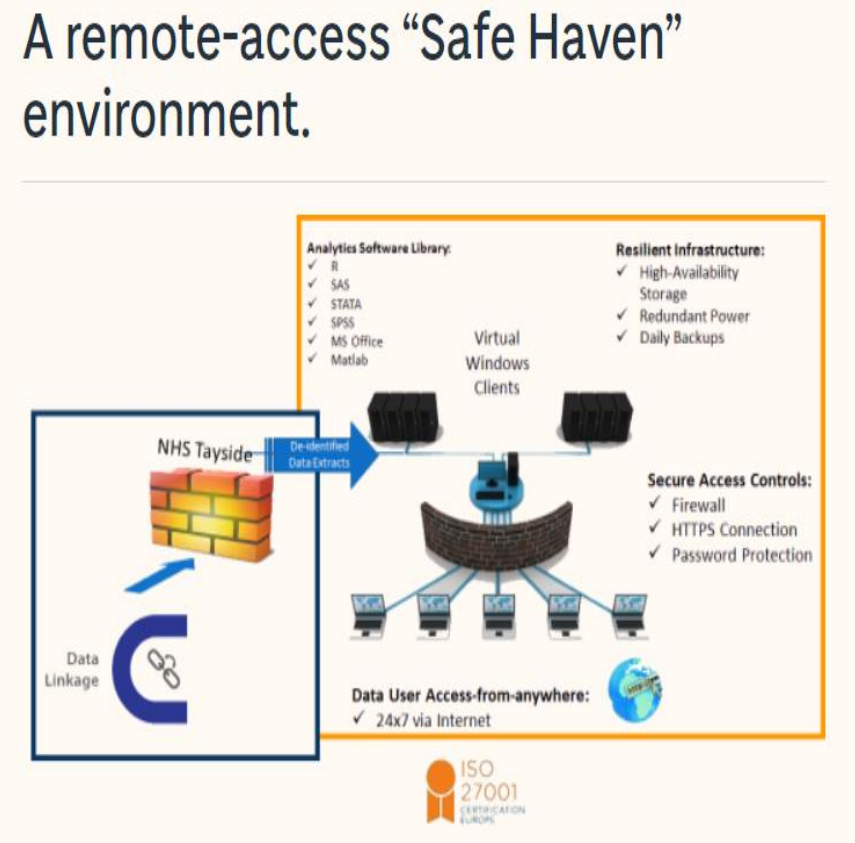
			Ischaemic stroke, deep, old	Ischaemic stroke, cortical, recent	Ischaemic stroke, cortical, old	Ischaemic stroke, unspecified	Haemorrhagic stroke, deep, recent	Haemorrhagic stroke, lobar, recent
Recent side	Recent lesion side	recentside						
Recent code (infarcts)	Recent lesion code	recentinfarctcode						
Recent type	Type of recent lesion[11]	recenttype		I		I	BGH or CbH	PH
Grouped types	type of recent lesions grouped for analysis[12]	condensed		LVDI		SVDI or LVDI	iCH	iCH



- Human annotation of radiology reports needs to be mapped to NLP tool outcomes
- Annotations are different for each NLP tool

How do you compare the outcomes of tools to each other?
Not necessarily about the best performing

Challenges:
Data, NLP Tools, Outcomes & Environment



https://www.dundee.ac.uk/hic/hicsafehaven/

- NLP Software is not standard
- Data & linkage
- Pre-processing and NLP free-text analysis experience

Results (preliminary)
Using Brain Image labelling (Gold Standard)

	EDIE-R		SemEHR	
Phenotype	PPV	Sensitivity	PPV	Sensitivity
Ischaemic - Any[1]	**88.69%**	**52.91%**	**79.13%**	**50.16%**
Ischaemic – deep recent	8.33%	4.48%		
Ischaemic –deep old	78.07%	38.56%		
Ischaemic –cortical recent	63.27%	21.83%		
Ischaemic –cortical old	77.50%	42.47%		
Haemorrhagic – Any	**32.93%**	**79.41%**	**55.56%**	**50.16%**
Subdural haematoma	28.36%	86.36%		
Subarachnoid haemorrhage[2]	**20.00%**	**71.43%**	**33.34%**	**42.86%**
Small Vessel Disease	90.06%	78.45%		
Atrophy	87.44%	75.98%		
Any Tumour[3]	34.21%	66.67%		

Phenotype	Image Readings	Edie-R
Ischaemic - Any	**1527**	**918**
Haemorrhagic – Any	**34**	**82**

[1] **Any label** is sub-labels plus underspecified for EDIE-R
[2] **Subarachnoid haemorrhage** – EDIE-R number is Subarachnoid haemorrhage other + Subarachnoid haemorrhage aneurysmal
[3] **Any tumour** EDIE-R number is tumour meningioma + tumour metastasis + tumour glioma + tumour other

Results
Using Brain Image labelling (Gold Standard)

Phenotype	ESPRESSO PPV	ESPRESSO Sensitivity
Silent Brain Infarct	90.10%	34.72%
White Matter Disease	74.05%	58.50%
White Matter Disease - Severity		
Mild	19.75%	4.16%
Moderate	28.81%	2.17%
Severe	87.23%	3.00%

Summary/Future

- Highlighted some challenges in NLP from observations on the literature
- Practical challenges in implementing and benchmarking NLP tools

- Ongoing work in benchmarking on the delirium study data
- Benchmarking for Generation Scotland data

Thank-You & Questions

Arlene.Casey@ed.ac.uk
@ArleneCasey

Edinburgh CLINICAL CLINICAL CLINICAL NLP
@EdiClinicalNLP

HDRUK Health Data Research UK

THE ALAN TURING INSTITUTE

MHDSS Mental Health Data Science Scotland

CHIEF SCIENTIST OFFICE

MRC Medical Research Council

generation SCOTLAND

[1] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu and Beatrice Alex (2021). **A Systematic Review of Natural Language Processing Applied to Radiology Reports**. 2102.09553.pdf (arxiv.org)