Annotation guidelines / Define concept

- Concept definition
  - Diagnosis, lab test, action, event…
- Variable definition
  - Diagnosis: explicitly mentioned or inferred
  - Lab test: exact numeric value or range or direction
  - Action: planned or occurred
  - Event: explicitly mentioned or inferred

Acknowledgements: Uni. of Utah DeCART summer school: https://github.com/jianlins/AnnotationNLP/blob/master/01_Introduction_to_Annotation.ipynb

Annotation guidelines / Define concept

- Annotation level/unit
  - Patient, document, sentence, phrase…
- Attribute definition
  - Polarity
  - Severity
  - Frequency
  - …

Acknowledgements: Uni. of Utah DeCART summer school: https://github.com/jianlins/AnnotationNLP/blob/master/01_Introduction_to_Annotation.ipynb

Annotation guidelines / Define concept

- Examples:
  - i2b2 2006 Smoking challenge:
    - Past Smoker
    - Current Smoker
    - Smoker
    - Non-Smoker
    - Unknown

    - What level/unit would be appropriate to annotate?

Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records.
J Am Med Inform Assoc. 2008;15(1):14–24. doi:10.1197/jamia.M2408

Annotation guidelines / Define concept

- Examples:
  - i2b2 2010 challenge:
    - Concepts: medical problem, treatment, test
    - Assertion: present, absent, possible in the patient; associated with someone else
    - Relation: improves, causes, worsens…

    - What level/unit would be appropriate to annotate?

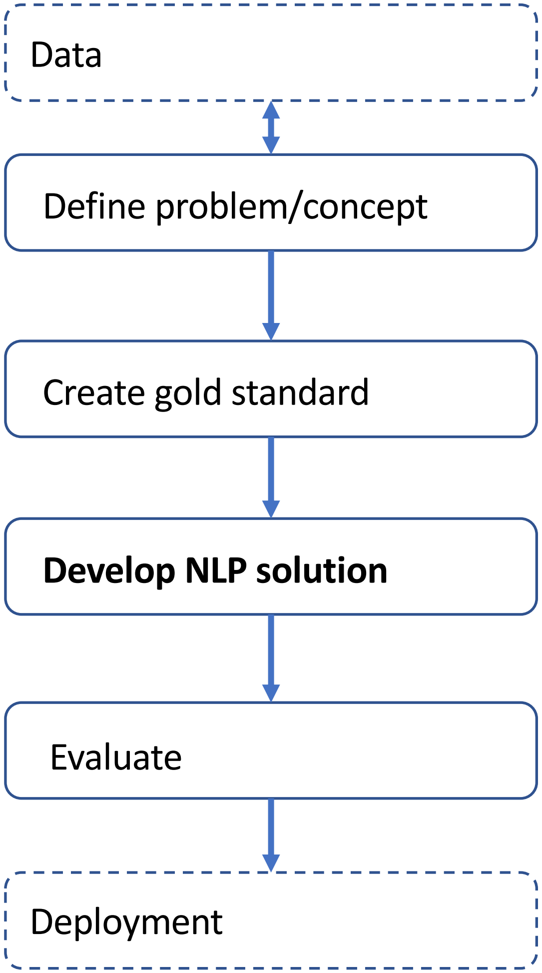
Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18(5):552–556. doi:10.1136/amiajnl-2011-000203

Annotation guidelines / Define concept

- Examples:
  - ShARe CLEF eHealth 2013 challenge:
    - Disorder mentions mapped to SNOMED-CT
    - What level/unit would be appropriate to annotate?

Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18(5):552–556. doi:10.1136/amiajnl-2011-000203

Natural Language Processing - workflow



- Develop NLP solution
  - Representation
    - What unit was annotated?
    - What would be an appropriate way to represent the data?
  - Entities
    - What is the distribution? Is the there a lot of variation, or are there clear patterns?
  - Approach: Patterns? Machine learning? Adaptation?
    - Off-the-shelf-tools often have default baseline representations, standard parameter settings etc – can this be a problem, or something you can re-use?

i2b2 2006 smoking challenge - distributions

Table 3

Table 3 Smoking Status Training and Test Data Distribution

Smoking Status	Training Data Frequency (%)	Test Data Frequency (%)
Past Smoker	36 (9)	11 (11)
Current Smoker	35 (9)	11 (11)
Smoker	9 (2)	3 (3)
Non-Smoker	66 (17)	16 (15)
Unknown	252 (63)	63 (61)
Total	398	104

Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records.
J Am Med Inform Assoc. 2008;15(1):14–24. doi:10.1197/jamia.M2408

i2b2 2010 concept challenge - approaches

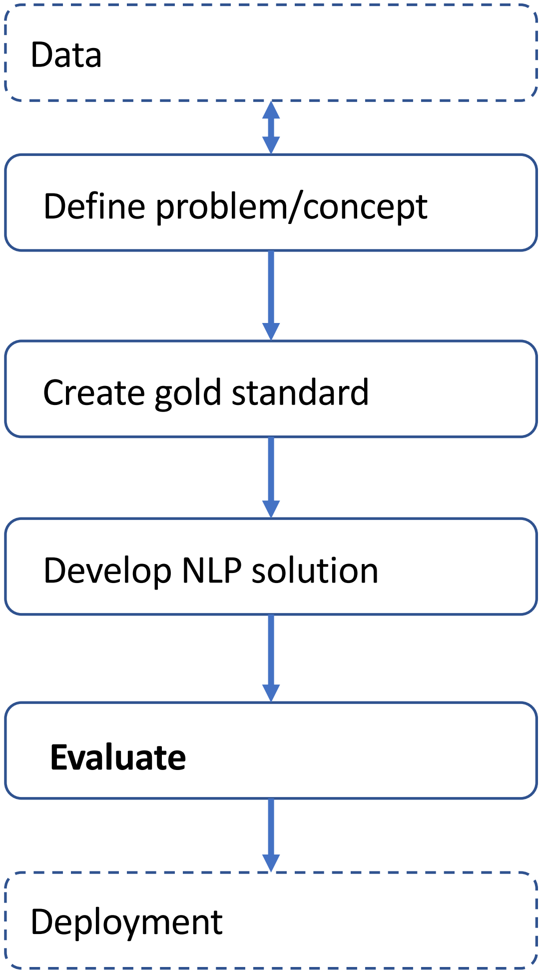
Table 2

Exact and inexact evaluation on the concept extraction task

Concept extraction System by	Medical experts	Method	External?	Exact F measure	Inexact F measure
deBruijn *et al*[25]	N	Semi-supervised	N	0.852	0.924
Jiang *et al*[16]	Y	Hybrid	Y	0.839	0.913
Kang *et al*[17]	N	Hybrid	Y	0.821	0.904
Gurulingappa *et al*[18]	N	Supervised	Y	0.818	0.905
Patrick *et al*[19]	N	Supervised	Y	0.818	0.898
Torii and Liu[20]	N	Supervised	N	0.813	0.898
Jonnalagadda and Gonzalez[21]	N	Semi-supervised	N	0.809	0.901
Sasaki *et al*[22]	N	Supervised	N	0.802	0.887
Roberts *et al*[23]	N	Supervised	N	0.796	0.893
Pai *et al*[24]	Y	Hybrid	N	0.788	0.884

Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18(5):552–556. doi:10.1136/amiajnl-2011-000203

Natural Language Processing - workflow



- Evaluation
  - What type of performance is acceptable? What type of evaluation is appropriate?
  - Intrinsic? Extrinsic?

  - Is the solution specific to a particular clinical population, or generic to the entire population?

i2b2 2010 concept challenge - evaluation

Table 2

Exact and inexact evaluation on the concept extraction task

Concept extraction					
System by	Medical experts	Method	External?	Exact F measure	Inexact F measure
deBruijn *et al*[25]	N	Semi-supervised	N	0.852	0.924
Jiang *et al*[16]	Y	Hybrid	Y	0.839	0.913
Kang *et al*[17]	N	Hybrid	Y	0.821	0.904
Gurulingappa *et al*[18]	N	Supervised	Y	0.818	0.905
Patrick *et al*[19]	N	Supervised	Y	0.818	0.898
Torii and Liu[20]	N	Supervised	N	0.813	0.898
Jonnalagadda and Gonzalez[21]	N	Semi-supervised	N	0.809	0.901
Sasaki *et al*[22]	N	Supervised	N	0.802	0.887
Roberts *et al*[23]	N	Supervised	N	0.796	0.893
Pai *et al*[24]	Y	Hybrid	N	0.788	0.884

Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18(5):552–556. doi:10.1136/amiajnl-2011-000203

i2b2 2006 smoking challenge - evaluation

Table 4

Table 4 Microaverages and Macroaverages for Precision, Recall, and F-Measure, Sorted by Microaveraged F-Measure

	Macroaveraged			Microaveraged		
Group Run	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Clark_3	0.81	0.73	0.76	0.90	0.90	0.90
Cohen_2	0.64	0.67	0.65	0.88	0.89	0.89
Aramaki_1	0.64	0.67	0.65	0.88	0.89	0.88
Cohen_1	0.64	0.65	0.64	0.88	0.88	0.88
Clark_2	0.76	0.69	0.72	0.87	0.88	0.88
Cohen_3	0.62	0.62	0.62	0.87	0.88	0.87
Wicentowski_1	0.58	0.61	0.59	0.85	0.87	0.86
Szarvas_2	0.59	0.60	0.59	0.85	0.87	0.85
Clark_1	0.69	0.65	0.66	0.86	0.87	0.85
Szarvas_3	0.56	0.58	0.57	0.84	0.86	0.84
Savova_1	0.62	0.60	0.60	0.84	0.86	0.84
Szarvas_1	0.56	0.58	0.57	0.83	0.86	0.84
Sheffer_1	0.59	0.59	0.58	0.83	0.86	0.84
Savova_2	0.56	0.57	0.56	0.81	0.84	0.82
Savova_3	0.55	0.55	0.55	0.80	0.83	0.81
Pedersen_1	0.55	0.56	0.54	0.82	0.82	0.81
Guillen_1	0.45	0.51	0.44	0.77	0.79	0.76
Carrero_1	0.52	0.47	0.48	0.74	0.77	0.75
Carrero_2	0.44	0.43	0.41	0.71	0.71	0.70

Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*. 2008;15(1):14–24. doi:10.1197/jamia.M2408

Data	
Source	South London and Maudsley hospital - CRIS
Governance/access procedure	Affiliation, approved research project, contact: xx.yy@zz.ac.uk
Content	Clinical notes: events, attachments
Size	500 annotated documents
Sampling procedure	All patients with diagnosis code xx (total yy), random sample of 500 (one document per patient)

NLP approach/model	
Objective/task	Smoking status (current, past, non-smoker)
Text/linguistic unit	Document
Gold standard	
Manual annotations? IAA?	Manual annotations, 2 clinicians, independent IAA: 77% F-score
Guidelines/definitions	URL

Model development	
Approach	Supervised machine learning SVM, liblinear, as implemented in scikit-learn, v. 3.2 Training/test split: 10-fold cross-validation
Parameters	Kernel: c: etc.
Prerequisites (preprocessing, etc)	nltk (v. 2.0) sentence splitting: punct_tokenizer, token

Data
Source
Governance/access procedure
Content
Size
Sampling procedure

NLP approach/model
Objective/task
Text/linguistic unit
Gold standard
Manual annotations? IAA?
Guidelines/definitions

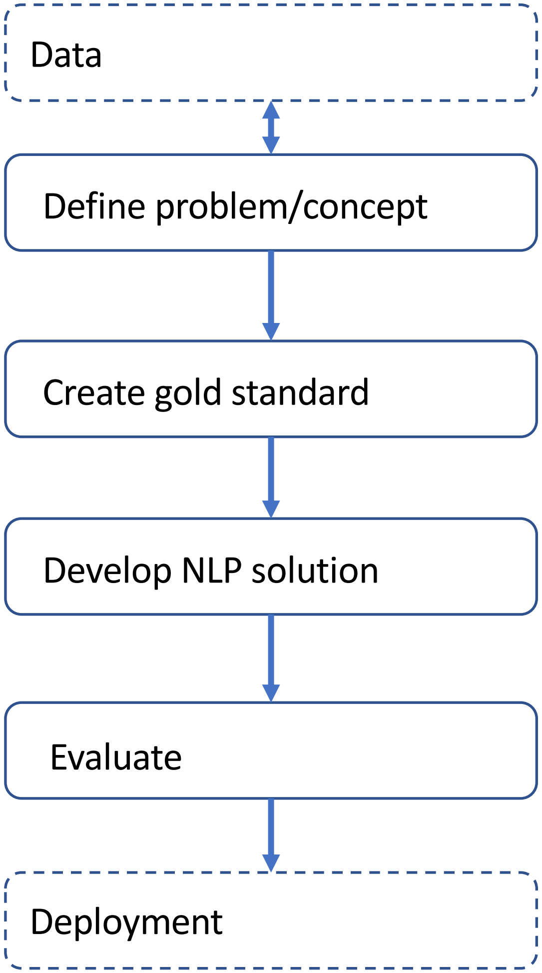
Model development
Approach (rule-based, machine learning – supervised, unsupervised)
Parameters
Prerequisites (preprocessing, etc)

NLP model development

Evaluation

Evaluation		
Intrinsic		
Criterion a		NLP task, e.g. De-identification
	Description	Quality of predictions/classification
	Metric	Precision, recall, F-score, accuracy, AUC
	Results	xx%
	Error analysis	Types of false positives, false negatives
Extrinsic		
Criterion b		Economic
	Description	Time to complete a task
	Method	Comparative – with/without NLP approach
	Metric	Time
	Results	X minutes faster with approach y
	Error analysis/comments	Advantages and disadvantages
Criterion c		Decision support
	Description	Alert to put patient on alternative treatment based on retrospective model using NLP to detect treatment response
	Method	Case-control
	Metric	Exposure measurement
	Results	No. of patients with improved health outcome
	Error analysis/comments	Advantages and disadvantages

Natural Language Processing - workflow



Example case:

A clinical NLP algorithm has developed to extract smoking status from EHRs and the algorithm has been made available. You want extract similar information from your EHR database for a particular clinical use-case.

What do you need to know about the algorithm and its development? How do you decide if it works well enough on your data?

NIHR | Maudsley Biomedical Research Centre

Thank you!

sumithra.velupillai@kcl.ac.uk