Semantic Annotation of Clinical Text: The CLEF Corpus

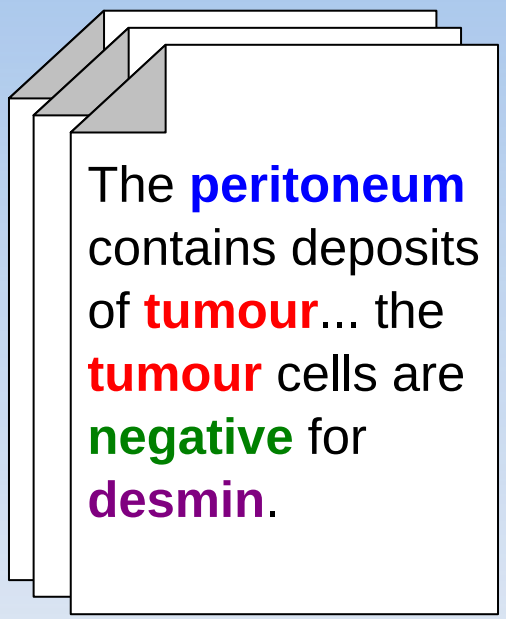
Angus Roberts, Robert Gaizauskas, Mark Hepple,
George Demetriou, Yikun Guo, Andrea Setzer and Ian Roberts

Natural Language Processing Group, University of Sheffield, UK

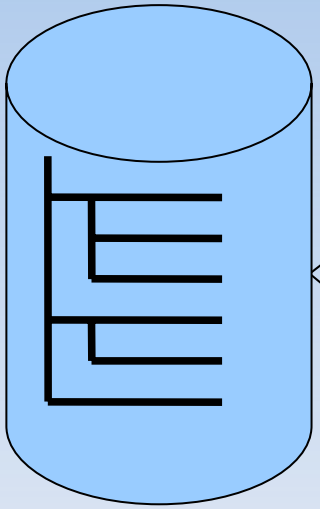
Introduction

- Background:
  - Information extraction and our application
- The CLEF (Clinical E-Science Framework) annotated corpus and gold standard
- Development methodology
- Some observations on annotators: results
- Annotation of temporal information
- Availability and conclusions

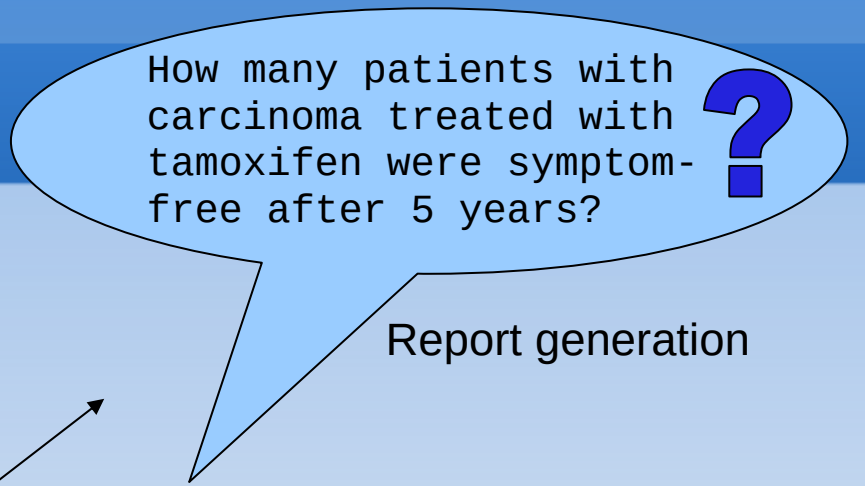
Application



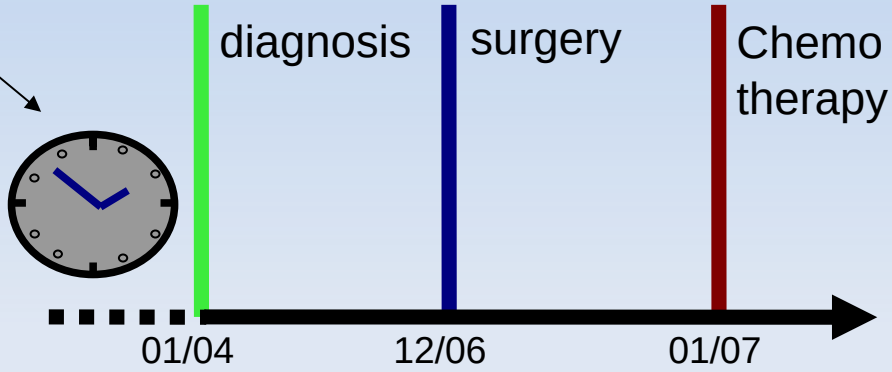
Information Extraction



CLEF EHR



Report generation



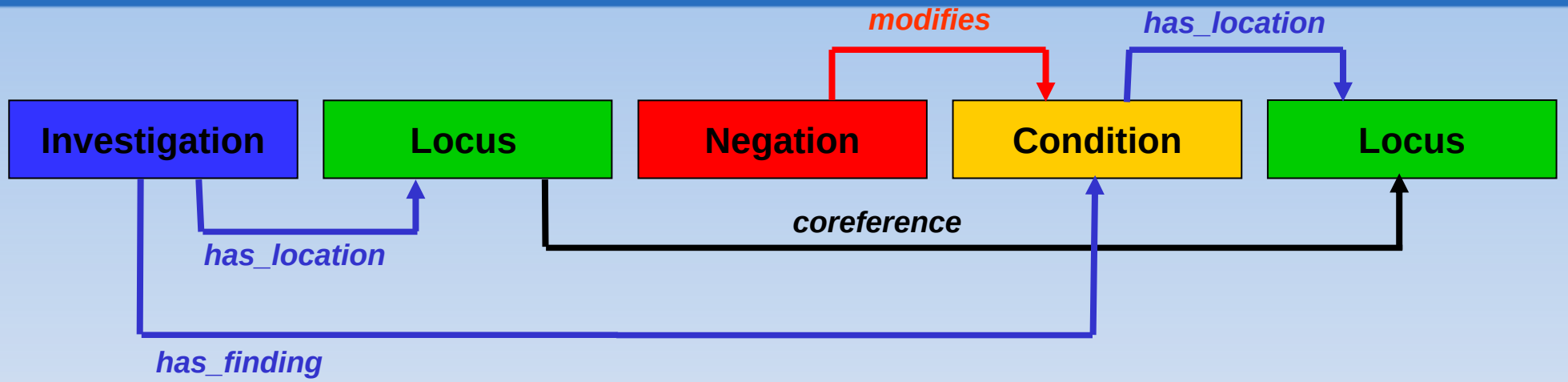
Chronicalisation

CLEF

MRC Medical Research Council

Entities, modifiers, relations, coreference

Punch biopsy of **skin**. **No** **lesion** on the **skin** surface following fixation.



- Cofererence, modifiers and relations allow for more sophisticated indexing and querying of reports

CLEF

MRC Medical Research Council

The CLEF Corpus

- Clinical text is hard to come by
- CLEF has a large corpus of clinical text

Document type	# of documents	tokens
Narratives	363K	63M
Imaging	187K	12M
Histopathology	15K	1.7M
Total	**566K**	**77M**

- Clearly, we can't manually annotate it all

CLEF

MRC Medical Research Council

The CLEF gold standard

- Principled selection of documents
- Mutiple text genres
- Multiple semantic types, relations, coreference
- Methodological approach to annotation
- Rigorous development of guidelines

MRC Medical Research Council

Document sampling

- Randomised and stratified selection of the whole corpus
- Minimum required to train statistical models
- Annotation is expensive!

Document type	# of documents
Narratives	50
Imaging	50
Histopathology	50
Total	150

CLEF

MRC Medical Research Council

Whole patients

- Some CLEF applications aggregate data across multiple documents on the same patient
- We have also annotated two whole patient records:

Document type	# of documents
Narratives	22
Imaging	14
Histopathology	2
Total	38

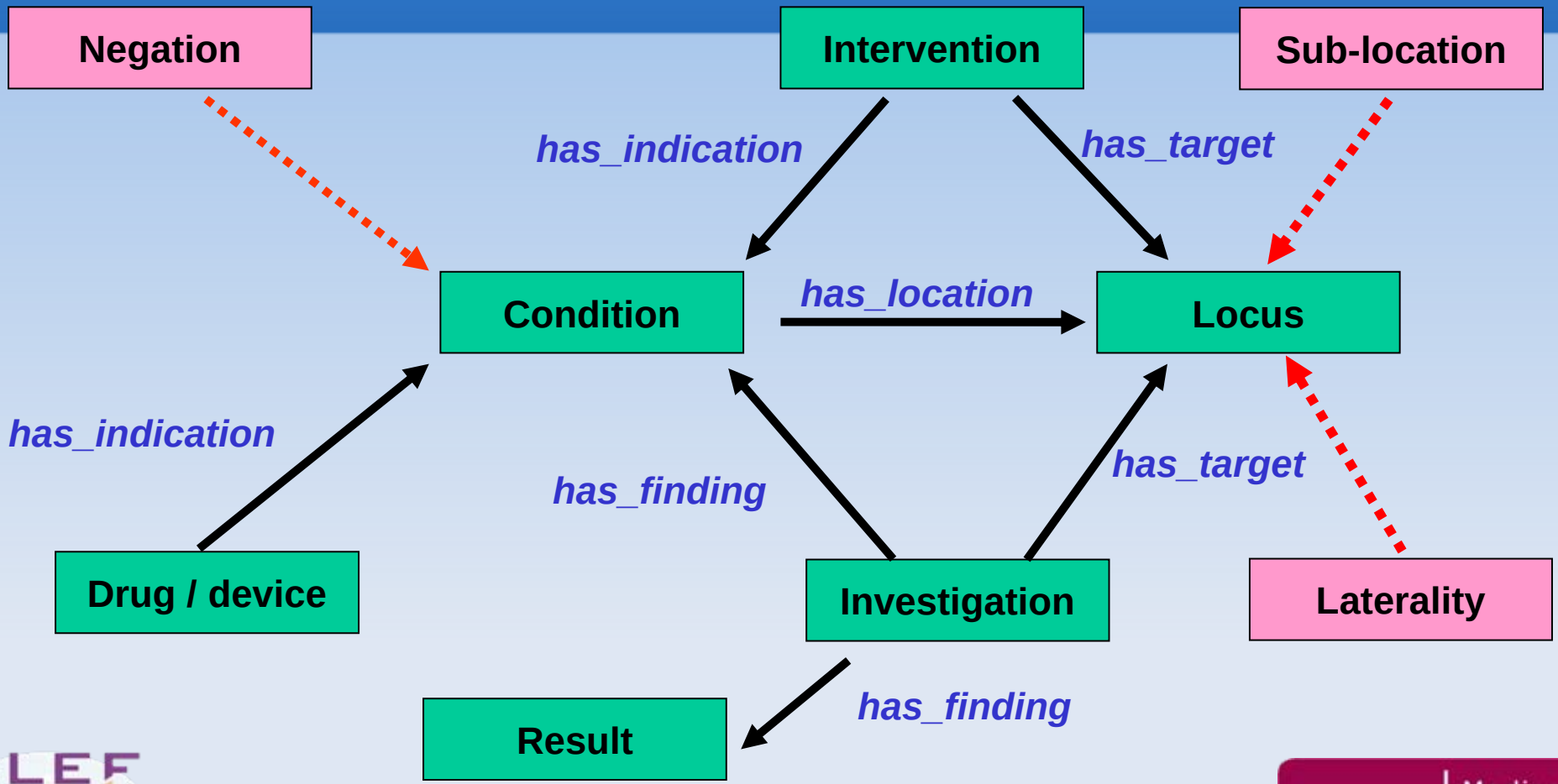
CLEF

MRC Medical Research Council

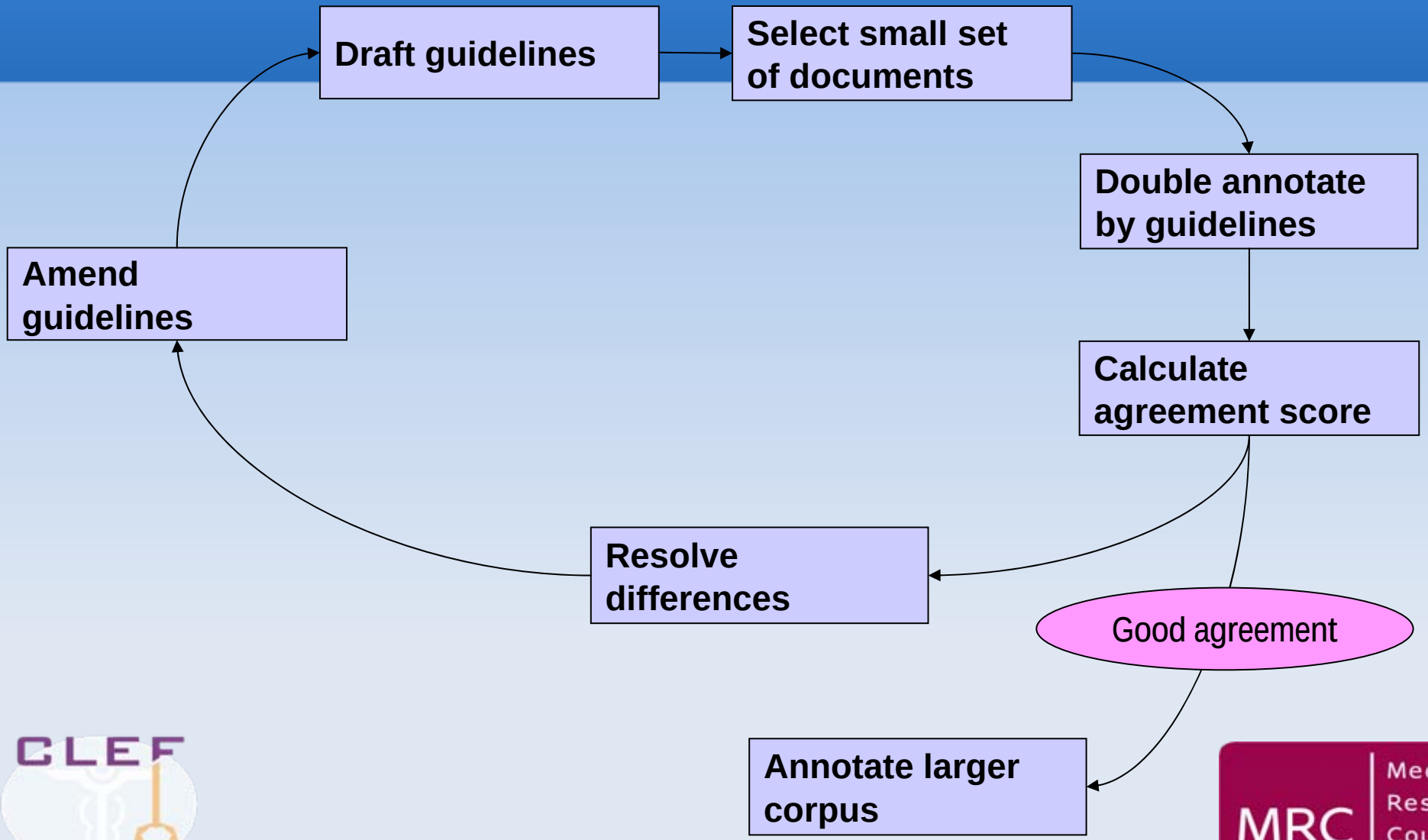
Annotation schema

- Developed through a requirements process
  - with end users of information extraction
- Schema is mapped to UMLS TUIs
- CUIs are added in a post-processing step

Annotation schema



Developing guidelines iteratively



Developing guidelines iteratively

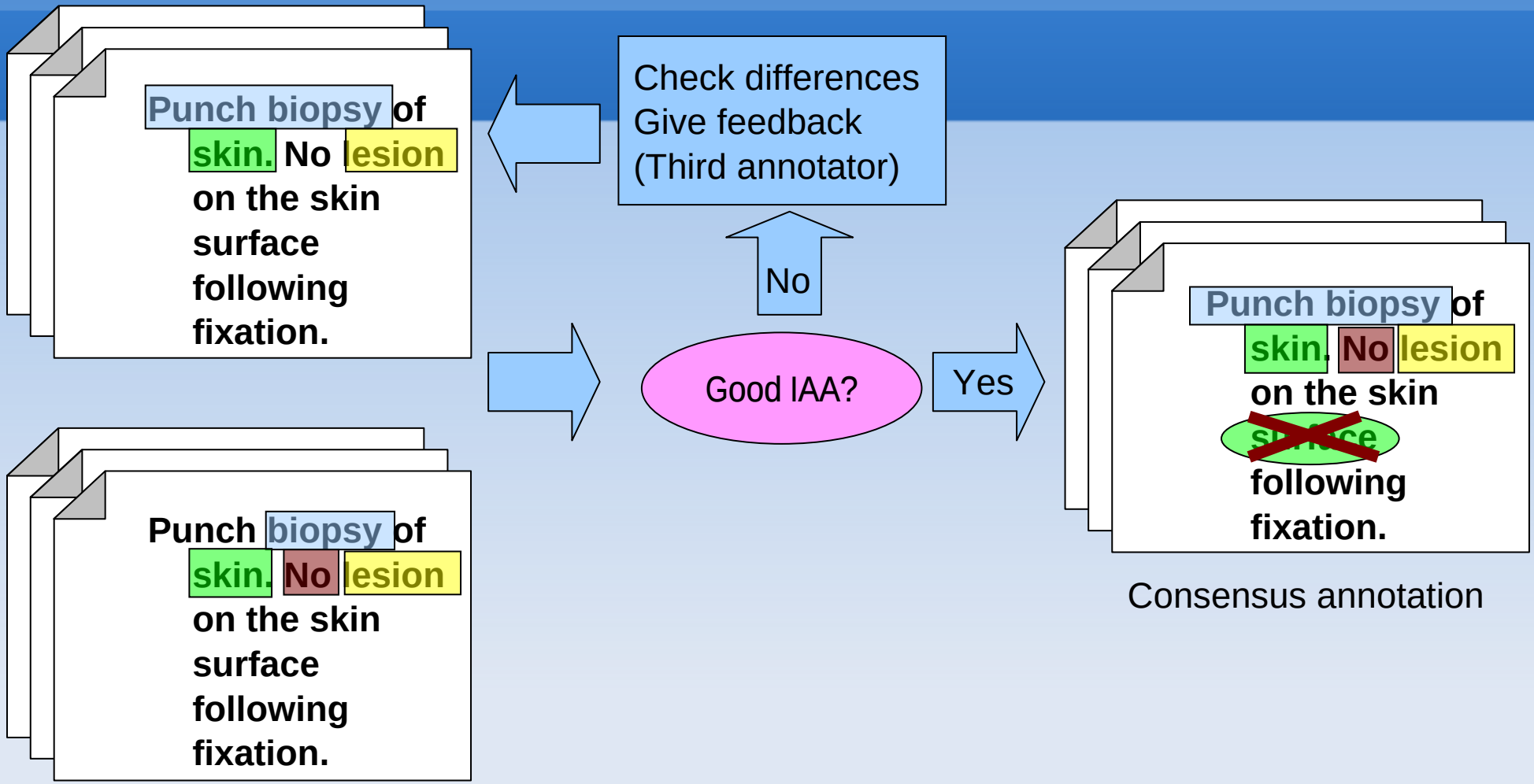
- Iterative development
  - Two senior annotators
  - 5 sets of documents (31 in total)
  - Amended guidelines at the end of each iteration
- Agreement score: % IAA

	Iteration				
	1	2	3	4	5
Entities	84	87	74	89	92
Relations	84	56	56	75	62 (73)

CLEF

MRC Medical Research Council

Consensus annotation



Consensus annotation

Tools

- Annotation: Knowtator text annotation tool
  - All annotation and consensus set creation
- Inter annotator agreement scoring
  - In-house scoring software
- Guidelines and feedback
  - Web site presenting cross-linked guidelines (wiki)
  - Feedback pages

CLEF

MRC Medical Research Council

Results: annotator expertise

- How does expertise affect agreement?
  - Senior development annotators
  - 3 annotators with minimal training

Sen2 (Senior 2)	77				
Clin (Clinician)	67	68			
BL (Biologist with linguistics)	76	80	69		
Ling (Linguist)	67	73	60	69	
Sen1 + Sen2 (Consensus)	85	89	68	78	73
	Sen1	Sen2	Clin	BL	Ling

Annotation of Temporal Information

- Guidelines were developed independently
- Automatic step:
  - Temporally Located CLEF entities (TLCs) (conditions, investigations and interventions) were imported from the annotated corpus
  - Time expressions were annotated by the GUTime tagger in accordance with the TimeML specification
- Manual step:
  - Annotators identified the temporal relations holding:
    - Between TLCs and the date of the letter (task A), and
    - Between TLCs and time expressions appearing in the same sentence (task B).
- To date 10 documents only have been annotated.

Distribution of Semantic Annotations

CLEF Gold Standard				
Entity	Narratives	Histopatho-logy	Radio-logy	Total
Condition	429	357	270	1056
Drug	172	12	13	197
Intervention	191	53	10	254
Investigation	220	145	66	431
Laterality	76	14	85	175
Locus	284	357	373	1014
Negation	55	50	53	158
Result	125	96	71	292
Sub-location	49	77	125	251
Relation				
has_finding	233	263	156	652
has_indication	168	47	12	227
has_location	205	270	268	743
has_target	95	86	51	232
laterality_mod	73	14	82	169
negation_mod	67	54	59	180
sub_loc_mod	43	79	125	247

Distribution of Temporal Annotations (1)

CTLink	Task A	Task B
After	5	18
Ended_by	3	0
Begun_by	4	0
Overlap	7	26
Before	5	135
None	4	8
Is_included	31	67
Unknown	6	14
Includes	13	137
Total	78	405

Distribution of CTLinks by type for tasks A & B.

CLEF

MRC Medical Research Council

Distribution of Temporal Annotations (2)

TLCs	Not hypothetical	243
	hypothetical	16
	Total	259
Time Expression	Duration	3
	DATE	52
	Total	55

Distribution of TLCs and temporal expressions.

Using the Corpus

- The gold standard corpus is used to train an IE system:
  - A ML layer that converts document annotations to SVM feature vectors and feeds classification results back into annotations.
  - A training subsystem that learns SVM models for tags.
  - A classification subsystem which takes features from pre-processed documents and trained SVM models to classify mentions/relations in text.

Preliminary F-measure results (with models trained/tested on incomplete gold standard):

- .71 over 5 clinical entity types
- .70 over 7 clinical relation types.

(see Roberts et al – LREC 2008, ACL-BioNLP 2008 for details)

Availability

- Gold standards of clinical text are not common
- Where they exist, use is normally restricted
- The CLEF gold standard:
  - Currently restricted
  - CLEF plans to develop a governance framework
  - This will take time!
- Annotation guidelines are available from the authors

Conclusions

- The annotated CLEF corpus is the richest resource of semantically marked up clinical text yet created:
  - Clinical entities and relations
  - Temporal entities and relations
- A rigorous and consistent methodology for gold standard development
- Challenges
  - Technical: consistency in relation annotation
  - Organisational: coordination of many annotators

Questions?

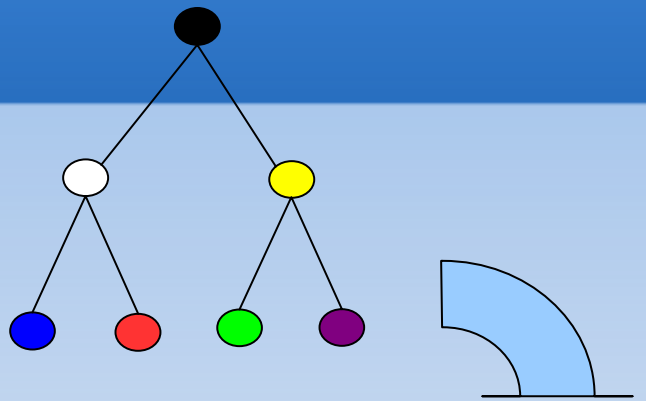
http://www.clinical-escience.org

http://www.clef-user.com

MRC | Medical Research Council

Clinical information extraction

The **peritoneum** contains deposits of **tumour**... the **tumour** cells are **negative** for **desmin**.



Test		**Result**
desmin	has finding	negative

Condition		**Locus**
tumour	has location	peritoneum
...	...	...

Randomised strata

- Not every random selection will do...
- The selection must reflect the whole corpus
- Randomised strata across two axes

Narrative subtype	% documents
To primary care	49
Discharge	17
Case note	15
Other letter	7
To consultant	6
To referrer	4
To patient	3

Neoplasm	% documents
Digestive	26
Breast	23
Haematopoetic	18
Respiratory etc	12
Female genital	12
Male genital	8

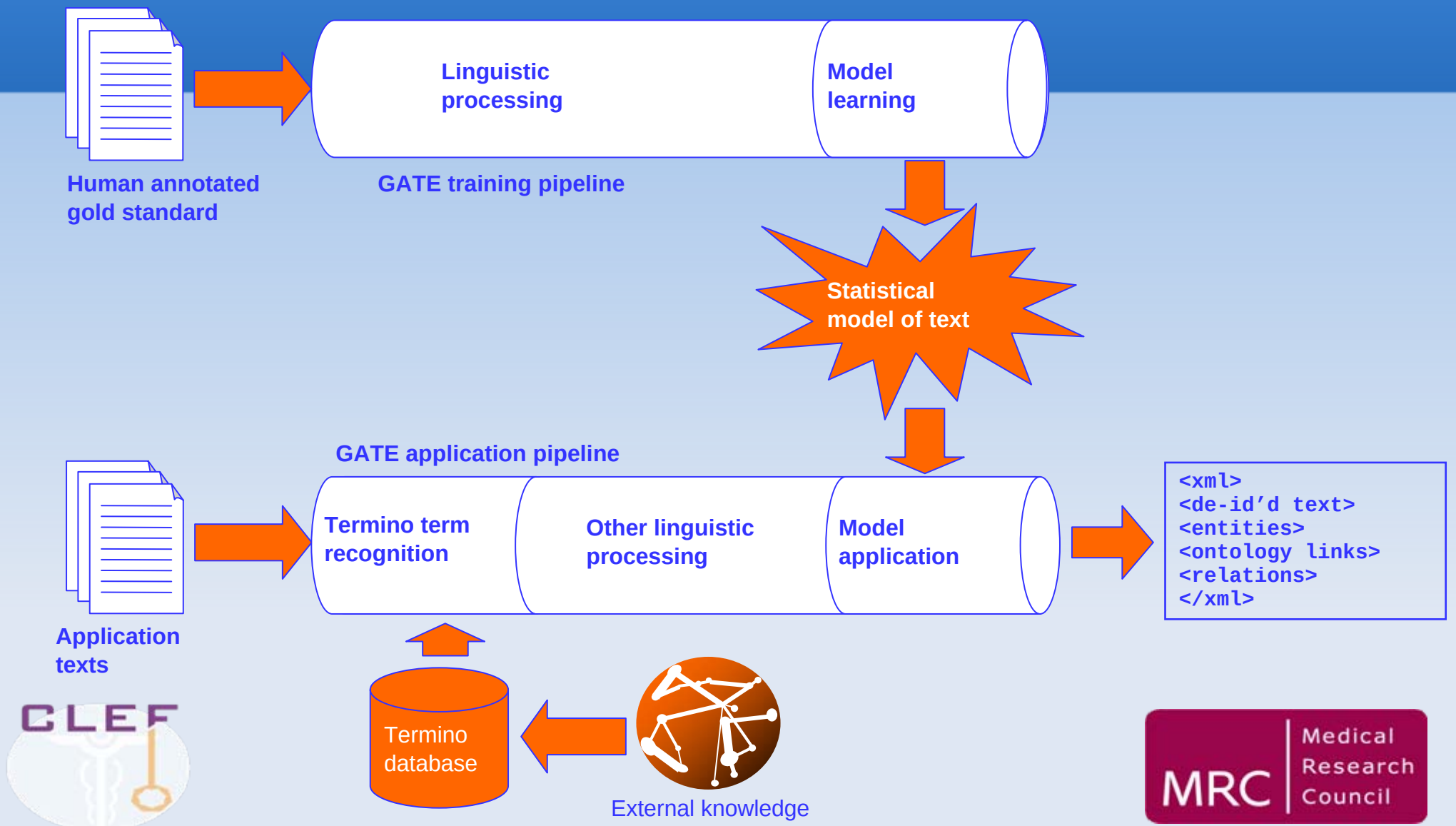
CLEF

MRC Medical Research Council

Annotation guidelines

- Consistency is critical to quality
- Documents need to be annotated in the same way
- Questions arise when annotating
  - e.g. when should a multi word expression be split?
- *Guidelines* detail *how* things should be annotated
  - and give a recipe to minimise errors
- Annotators are given structured training in annotation and the guidelines

System architecture



Annotating CUIs

- Separate post-processing task
- Automatic assignment of possible CUIs based on string match
- Manual: single annotation
  - confirmation
  - disambiguation
  - assignment where none found automatically

Text sub-genres

- Can guidelines developed on one genre be applied to another?
  - Developed guidelines over 5 iterations of narratives
  - Applied to imaging and histopathology reports

		IAA	
	Iterations	**Entities**	**Relationships**
Narratives	5	92	62
Imaging	2	90	84
Histopathology	2	88	70

CLEF

MRC Medical Research Council

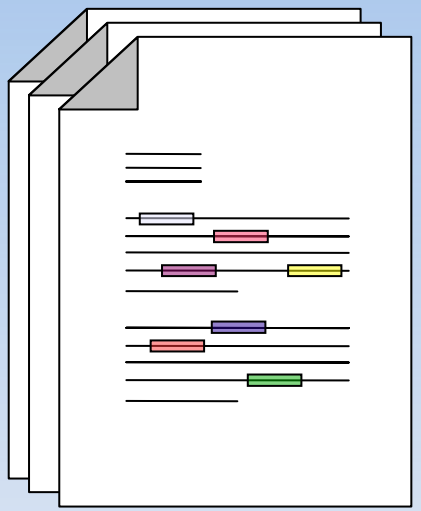
Results: annotator consistency

- How well do annotators agree?
  - Senior annotators vs 7 others, after training
  - Measured agreement with consensus

	Entities	Relationships
Senior 1	85	87
Senior 2	89	74
1	84	52
2	84	52
3	88	61
4	85	68
5	83	57
6	91	61
7	87	71

IE needs manually annotated text

- (usually...)
  - Learn models and patterns
  - Apply to unseen texts
    - `"X on the [locus]"`
      `=> X is a Condition`
    - Statistical models of context
  - Evaluation standard:
    - e.g. train on 90%, test on 10%
    - ten-fold cross validation



Human annotated gold standard

CLEF

MRC Medical Research Council