Scaling the Data Wall

Lynette Hirschman

The MITRE Corporation

April 23-24, 2012

NLM/NIH Workshop on
Natural Language Processing: State of the Art,
Future Directions and Applications for
Enhancing Clinical Decision-Making

MITRE

Outline

- Need for data
- Clinical data resources
- The clinical data wall
- Scaling to real applications

MITRE

Why Data?

- **Linguistic analysis**
  - Examples of (sub)language
- **Natural language processing system**
  - Examples of inputs and desired outputs
- **Evaluation**
  - Gold standard data of “correct” input/output pairs for comparison to system output

Evaluation is a routine part of NLP system development – like debugging

Challenge Evaluations Play a Different Role

- **Challenge evaluations**
  - Drive research progress of a subfield
  - Create communities and a market
  - Train the next generation of researchers
  - Create infrastructure

What infrastructure is needed to "unlock the patient record"?

MITRE

The Clinical Data Challenge

- **Medical NLP has been an active field since the '60s**
- **Large-scale terminological and ontological resources are available:**
  - MeSH, UMLS, SnoMED-CT
- **Medical journal articles are available:**
  - MEDLINE/PubMed, PubMedCentral
- **But until recently, there were *no sharable corpora* of clinical data (medical records)**
  - Making it impossible to share or compare results

Therefore, there were no shared evaluations

-- which limited progress

MITRE

Clinical Data: "Poster Child" For Challenge Evaluations?

- **Automated de-identification software facilitated removal of Protected Health Information (PHI)**
- **This made it possible to share corpora (under limited data use agreements)**
  - U Pitt corpus, Pediatric radiology, MIMIC II, i2b2
- **Corpora enabled Challenge Evaluations**
  - i2b2: Uzuner, Duvall, South;
  - Pediatric radiology: Pestian
- **Open source modules are becoming available**
  - NegEx, ConText, cTAKES[1], MASTIF[2]

[1]clinical Text Analysis and Knowledge Extraction System

[2]MITRE Assertion Status Tool for Interpreting Facts

MITRE

i2b2 Challenge Evaluations: Where Next?

Eval	Year	Task	Data sets	# Teams
1st	2006	De-identification	889 discharge summaries (de-identified with synthetic identifiers)	7 teams
		Smoking history	398/104 discharge summaries	11 teams
2nd	2008	Obesity and co morbidities	1237 discharge summaries	30 teams
3rd	2009	Medication extraction	Community annotation of 251 discharge summaries	20 teams
4th	2010	Concept extraction Assertion status Relation extraction	349 Training and 447 test reports:	22 teams 21 teams 16 teams
5th	2011	Coreference Emotion Extraction	500 pt notes annotated for coref 600 training; 300 test notes	20 teams 26 teams
6th	2012	Temporal relations		
		Clinical Trial Eligibility	TREC: Problem with data!	

MITRE

Automated De-identification Evaluation

Approach: Find and transform identifying information using natural language processing techniques (NLP)

HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old-woman with long standing hypertension who presented as a walk-in to me at the **Oak Valley Health Center** on **July 9th**. Recently had been started q.o.d. on Clonidine since **May 5th** to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the **Smith Cardiac Unit** for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. **Pearson** to follow.

- **i2b2 task: identifying PHI in narrative**
- **Practical applications: redacting or transforming PHI**
- **Automated de-identification tools becoming available**
  - **E.g., MIT de-id, U Pittsburgh DE-ID, MITRE's MIST, Emory University's HIDE**

Automated De-identification: What Did We Learn

- **This was a very successful evaluation**
  - Good performance using standard NLP measures: accuracy, precision/recall, f-measure
- **Automated de-id was usable for data sharing**
  - For MITRE, our open source de-identification module MIST[1] serves as the basis for many collaborations
    - Move the software to the data!
- **However – we still have no "extrinsic" evaluation**
  - Are these systems "good enough"?
  - And "good enough" for what?

Results are couched in "NLP" metrics – but IRBs need PHI exposure risk

[1]MITRE Identification Scrubber Toolkit

Evaluating the Evaluation: Coreference

i2b2 Coreference Tasks

1) **Identification of markables (mentions of person, problem, test, treatment, and pronouns)**

2) **Linkage of coreferring mentions of markables**

- **Person**
  - *****Dr. [Name XXX]***** reviewed the case. **He** recommends **the patient** should remain on the ventilator until **his** condition stabilizes.
- **Problem**
  - The **patient** had **a kidney stone** in ‘00, **he** presents today with **a left kidney stone**
- **Test**
  - **A CT** was done, **it** showed bilateral ground glass appearance
- **Treatment/Medication**
  - The **patient** takes **advil** at home so we gave **him** **ibuprofen 800 mg**

i2b2 Coreference: What Did We Learn?

The corpus

- "Person" category had most markables (and coref chains)
- But "person" can be treated as a 3-way classification task: patient, friend/family, provider
- Treatment of persons <u>may not require coreference module</u>

Results

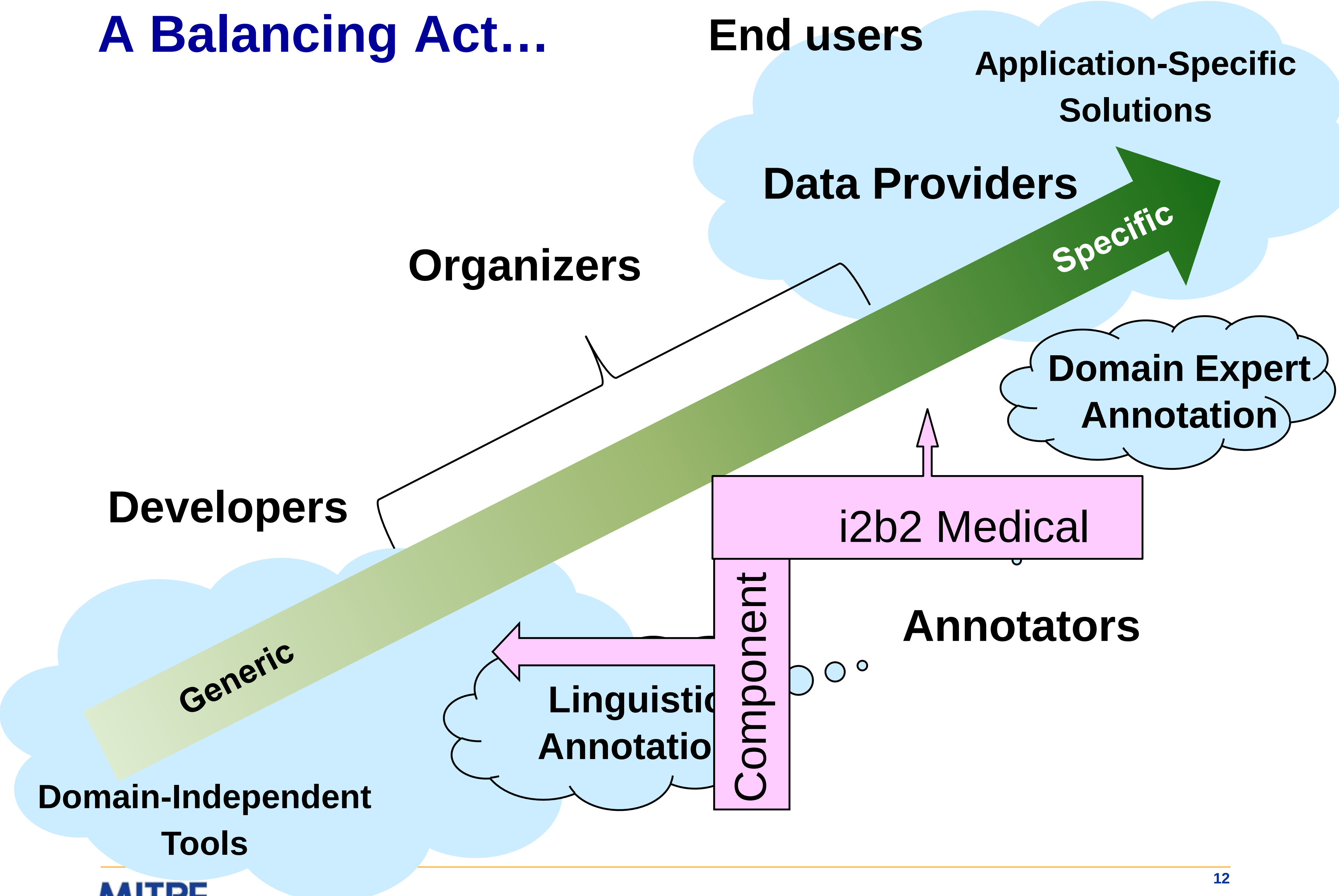
- Coref "score" was the arithmetic average of 3 metrics
- Scores on "end-to-end" i2b2 task were ~ 60% f-measure
- Comparable to corerefence evaluation on Newswire in 2001

What does this mean?

- That we made no progress since 2001?
- That this application didn't really need coreference?

We need an extrinsic (clinical task-based) evaluation!

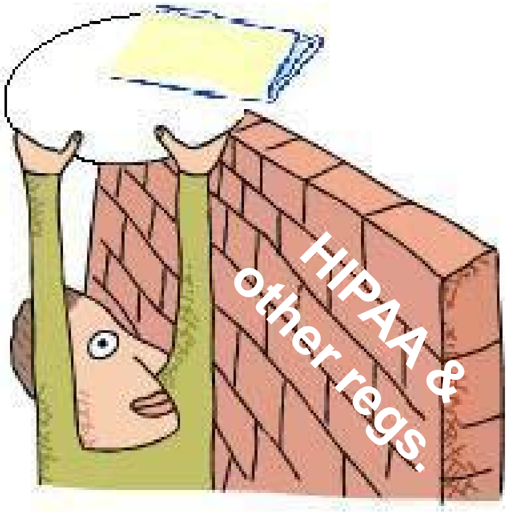
A Balancing Act…



Scaling the Clinical Data Wall

Records with protected health information (PHI) cannot be shared due to privacy constraints

Unstructured Medical Records containing PHI



Medical record de-identification is the rate-limiting step in many secondary use

- Clinical Applications
- Secondary Use
- NLP Research

Some Ways to Scale the Data Wall

- **Lower the barriers to IRB approval**
  - Release via limited data use (as done for i2b2, Pittsburgh data)
  - Develop metrics relevant to IRB concerns
- **Reduce the re-identification risk**
  - Interactive human review using automated de-identification tools
  - Selective extraction:
    - Extract clinically relevant information, leaving behind the PHI*
- **Move the software to the data**

* Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc. 2009 Jan-Feb;16(1):37-9.

Scaling the Data

- **Most real applications are "one-off"**
  - Data (patient records) can't be shared
  - Application is institution-specific
- **However, new multi-site projects are springing up**
- **SHARPn for secondary data use**
- **eMERGE: EMR and Genomics**
  - 7 groups pooling EMRs and biobank data to identify patient phenotypes* and associated genetic variations

*** Phenotype = patient characteristics (appearance, state of health/disease)**

MITRE

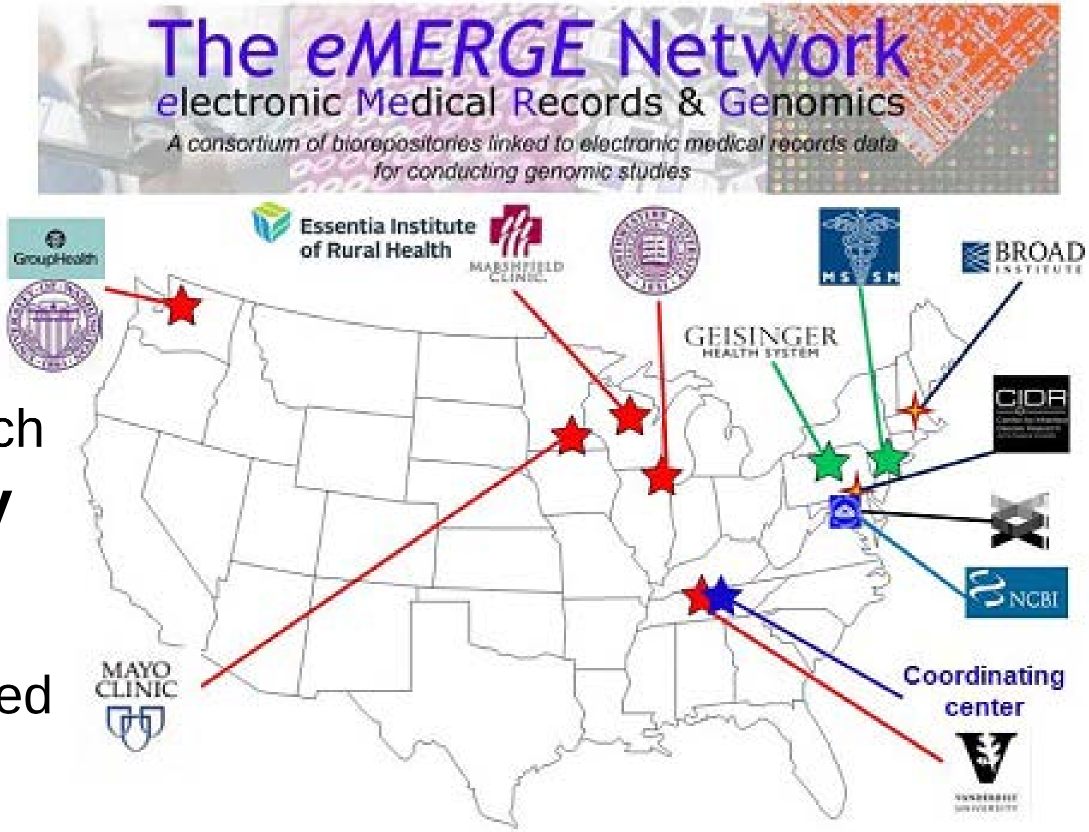
SHARPn*

- **SHARP Area 4: Secondary Use of EMRs**
  - Funded by Office of the National Coordinator as one of 5 Strategic Health Advanced IT Research Projects (SHARP)
  - PI: Prof Chris Chute, Mayo Clinic
  - Use case: phenotype extraction from EMRs
- **Data**
  - 360K notes from 10K patients (from 2 providers)
  - Pilot annotated corpus: 700 docs; 200K words
  - Stratified annotated corpus 1000 docs; 300k words
- **Annotation layers**
  - Linguistic: syntactic trees (treebanking), predicate-argument structure (propbanking), coreference
  - Medical: UMLS entities with mapping to SNOMED CT and RxNORM; UMLS relations; Clinical Element Model (CEM)

*http://informatics.mayo.edu/sharp/index.php/Main_Page

eMERGE Consortium

- **Combining**
  - DNA biorepositories
  - Electronic medical record (EMR) systems
  - For large-scale, high-throughput genetic research
- **Has a published library**
  - 13 phenotype extraction algorithms
  - From EMRs (both structured and unstructured data)



https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page
Funding from NHGRI, NIGMS

Scaling the Data for a “Real” Application

- **eMERGE applications**
  - Pool data across institutions to get sufficient statistical power
  - Extract patients and controls based on phenotype
  - Typical numbers: 3000 cases + 3000 controls
- **SHARPn has pool of 10K patients**
- **Could these provide challenge evaluation data sets?**
  - Data sharing issues – partially solved already (special eMERGE data use agreement)
  - Coarse-grained clinical annotations available at patient level for “displays phenotype” and “control”

MITRE

Scaling the Annotation: the Challenges

- **Annotation Cost**
  - Even with tools, annotation is still very expensive
    - Optimistic estimate: $1/patient note/layer
    - 30 notes/patient annotated w 3 layers: $90 per patient
- **How can we scale the annotations?**
  - Use machine-assisted human review
    - Don’t annotate de novo
  - Do less – fewer layers, coarser granularity
  - Leverage naturally occurring annotations
    - E.g., records binned for cohort selection
  - Develop better algorithms to learn from noisy annotation

Recommendations

- **Scale the data wall**
  - Acquire medically relevant collections of clinical data
  - Using automated methods to minimize re-identification risk
- **Scale for real applications**
  - Evaluate NLP systems for utility, cost-effectiveness on “extrinsic” clinical applications, e.g.,
    - Phenotype identification (SHARP, eMERGE)
    - Clinical decision making (if data are available)
- **Scale the annotations**
  - Develop cost-effective “minimalist” annotation strategies
    - Take advantage of “naturally occurring” (partially) annotated corpora
    - With new algorithms for learning from noisy, coarse-grained annotations

Acknowledgements

- **Many thanks to my MITRE colleagues working on:**
  - **Automated Fact Extraction from Medical Records** (responsible for MASTIF for assertion status): Cheryl Clark, John Aberdeen, Matt Coarr, Robyn Kozierok, David Tresner-Kirsch, Ben Wellner, Alex Yeh
  - **De-identification** (responsible for MIST for de-identification): John Aberdeen, Sam Bayer, Cheryl Clark, Ben Wellner
  - **Unlocking the Patient Record for Translational Medicine**: John Aberdeen, Sam Bayer, John Burger, Cheryl Clark, David Tresner-Kirsch, Ben Wellner
- **Also thanks to our external collaborators, including David Carrell, Wendy Chapman, Len D'Avolio, Brad Malin, Guergana Savova, Ozlem Uzuner**

Back Up