Comparing Concept Normalization Accuracy and Speed for Medical Problems and Medication Allergies

Stéphane M. Meystre, MD, PhD[1,2], Andrew Trice, BS[1], Youngjun Kim, PhD[1], Paul Heider, PhD[1]

[1] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC [2] Clinacuity Inc., Charleston, SC

AMIA Virtual Informatics Summit

March 24, 2020 Session VS05

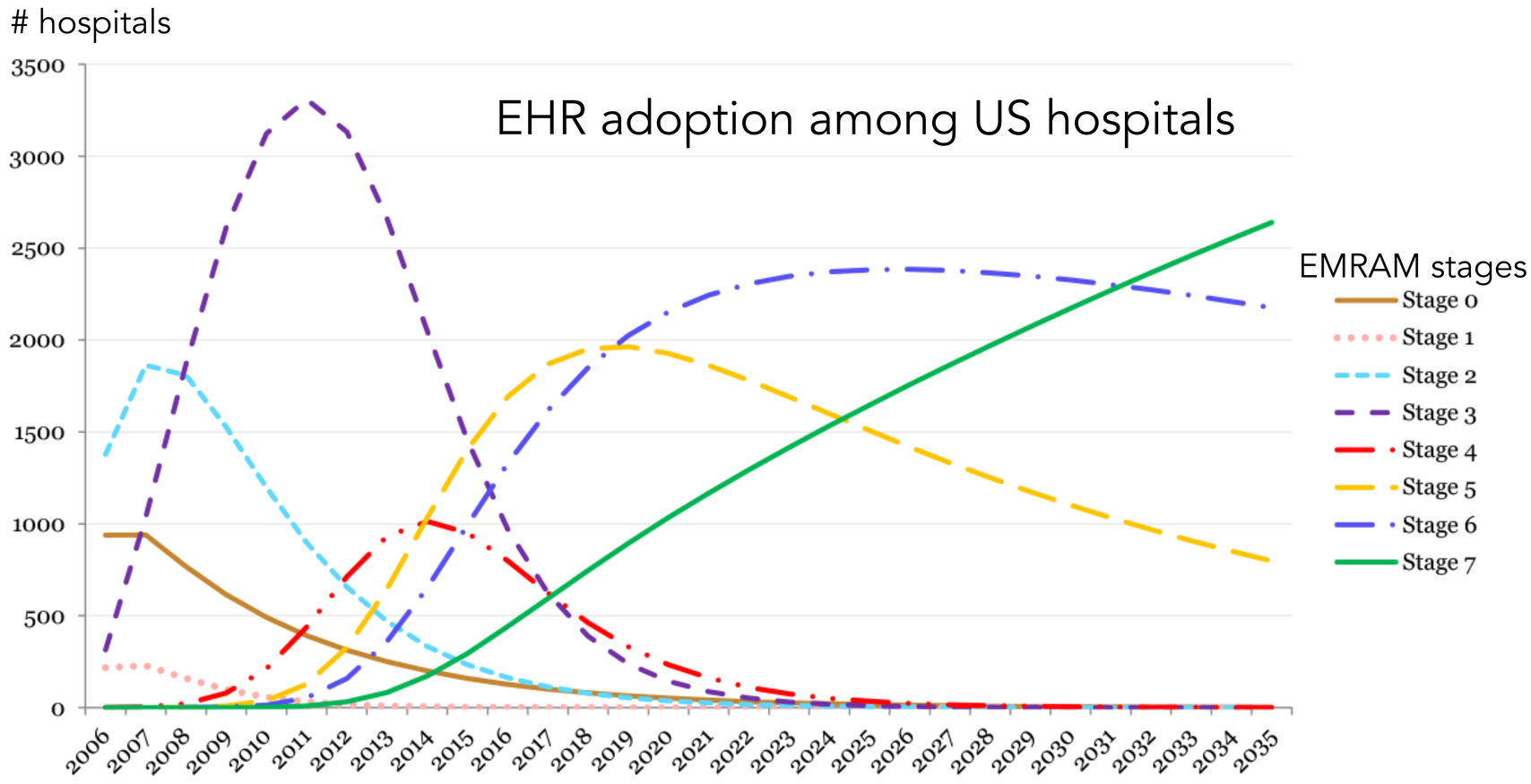
Disclosures

Stephane Meystre is a shareholder of Clinacuity, Inc.

No other relevant relationships with commercial interests to disclose .

Introduction

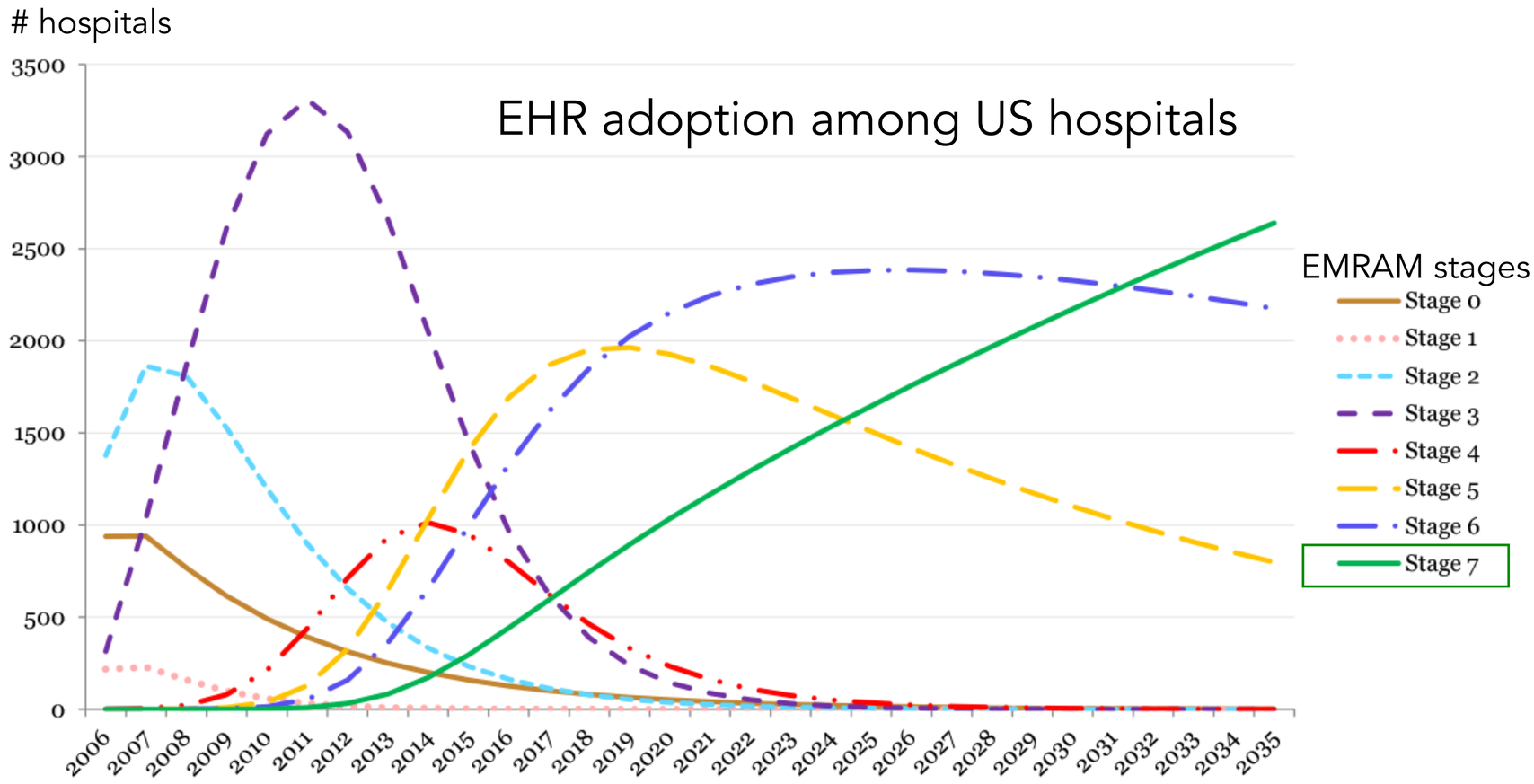
Large quantities of clinical information available and strong incentives for reuse



Kharrazi H, Gonzalez CP, Lowe KB, et al. Forecasting the Maturation of Electronic Health Record Functions Among US Hospitals: Retrospective Analysis and Predictive Model. J Med Internet Res 2018;20(8):e10458

Introduction

Large quantities of clinical information available and strong incentives for reuse



Fully electronic EHR environment with data sharing

Kharrazi H, Gonzalez CP, Lowe KB, et al. Forecasting the Maturation of Electronic Health Record Functions Among US Hospitals: Retrospective Analysis and Predictive Model. J Med Internet Res 2018;20(8):e10458

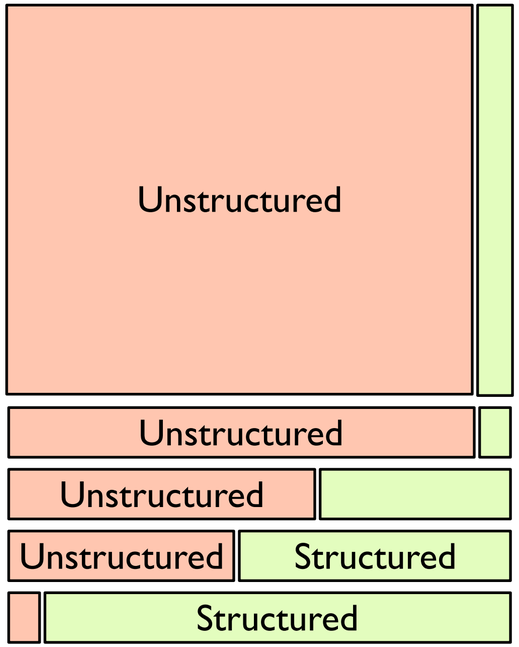
Introduction

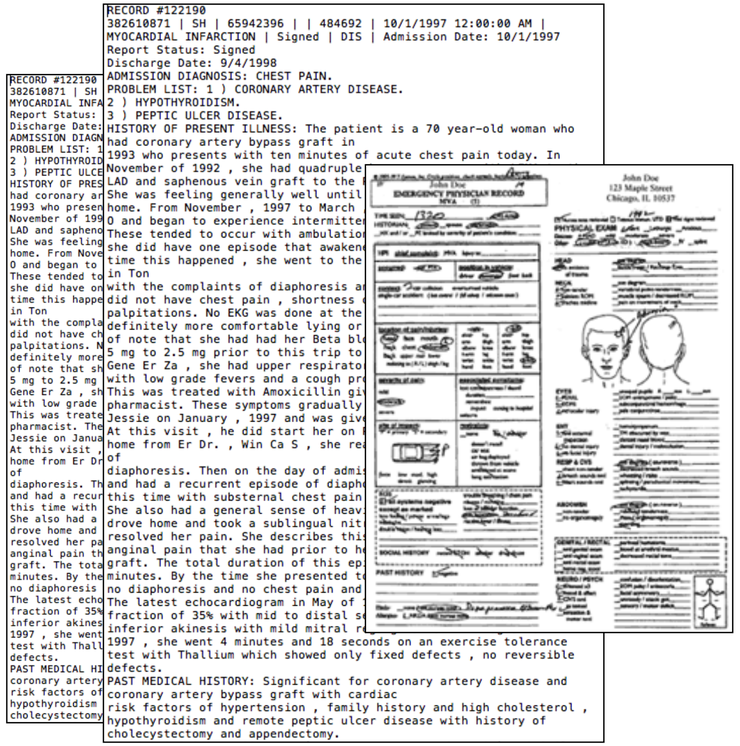
Large quantities of clinical information available (cont.)

But most clinical information is unstructured, requiring **information extraction**

Clinical text is unstructured EHR data

- Documents
  - History and Physicals
  - Clinical notes, Consult notes
  - Operative reports
  - Surgical pathology reports
  - Progress notes, Letters
  - Orders
  - Discharge summaries
- Imaging / Radiology
- Prescriptions (pharmacy; CPOE)
- Laboratory results
- Administrative information




RECORD #122190
382610871 | SH | 65942396 | | 484692 | 10/1/1997 12:00:00 AM |
MYOCARDIAL INFARCTION | Signed | DIS | Admission Date: 10/1/1997
Report Status: Signed
Discharge Date: 9/4/1998
ADMISSION DIAGNOSIS: CHEST PAIN.
PROBLEM LIST: 1) CORONARY ARTERY DISEASE.
2) HYPOTHYROIDISM.
3) PEPTIC ULCER DISEASE.
HISTORY OF PRESENT ILLNESS: The patient is a 70 year-old woman who had coronary artery bypass graft in
1993 who presents with ten minutes of acute chest pain today. In
November of 1992 , she had quadruple
LAD and saphenous vein graft to the
She was feeling generally well until
home. From November , 1997 to March
0 and began to experience intermitte
These tended to occur with ambulatio
she did have one episode that awaken
time this happened , she went to the
in Ton
with the complaints of diaphoresis a
did not have chest pain , shortness
palpitations. No EKG was done at the
definitely more comfortable lying or
of note that she had had her Beta bl
5 mg to 2.5 mg prior to this trip to
Gene Er Za , she had upper respirato
with low grade fevers and a cough pr
This was treated with Amoxicillin gi
pharmacist. These symptoms gradually
Jessie on January , 1997 and was giv
At this visit , he did start her on
home from Er Dr. , Win Ca S , she re
of
diaphoresis. Then on the day of admi
and had a recurrent episode of diaph
this time with substernal chest pain
She also had a general sense of heav
drove home and took a sublingual nit
resolved her pain. She describes thi
anginal pain that she had prior to h
graft. The total duration of this ep
minutes. By the time she presented t
no diaphoresis and no chest pain and
The latest echocardiogram in May of
fraction of 35% with mid to distal s
inferior akinesis with mild mitral r
1997 , she went 4 minutes and 18 seconds on an exercise tolerance
test with Thallium which showed only fixed defects , no reversible
defects.
PAST MEDICAL HISTORY: Significant for coronary artery disease and
coronary artery bypass graft with cardiac
risk factors of hypertension , family history and high cholesterol ,
hypothyroidism and remote peptic ulcer disease with history of
cholecystectomy and appendectomy.

Introduction

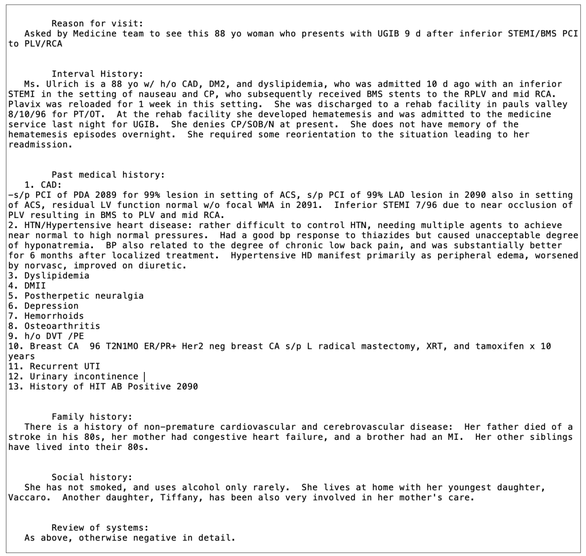
Clinical information extraction

Typically requires a **dictionary lookup** linking text to standard terminologies, searching the text for mentions of concept terms from the *dictionary* (i.e., standard terminology). It is sometimes also called "concept extraction," "concept normalization" or even "concept recognition" even if the latter would be closer to "named-entity recognition" or "entity recognition."

Introduction

Clinical information extraction

Typically requires a **dictionary lookup** linking text to standard terminologies, searching the text for mentions of concept terms from the *dictionary* (i.e., standard terminology). It is sometimes also called "concept extraction," "concept normalization" or even "concept recognition" even if the latter would be closer to "named-entity recognition" or "entity recognition."



Reason for visit:
Asked by Medicine team to see this 88 yo woman who presents with UGIB 9 d after inferior STEMI/BMS PCI to PLV/RCA

Interval History:
Ms. Ulrich is a 88 yo w/ h/o CAD, DM2, and dyslipidemia, who was admitted 10 d ago with an inferior STEMI in the setting of nauseau and CP, who subsequently received BMS stents to the RPLV and mid RCA. Plavix was reloaded for 1 week in this setting. She was discharged to a rehab facility in pauls valley 8/10/96 for PT/OT. At the rehab facility she developed hematemesis and was admitted to the medicine service last night for UGIB. She denies CP/SOB/N at present. She does not have memory of the hematemesis episodes overnight. She required some reorientation to the situation leading to her readmission.

Past medical history:
1. CAD:
-s/p PCI of PDA 2089 for 99% lesion in setting of ACS, s/p PCI of 99% LAD lesion in 2090 also in setting of ACS, residual LV function normal w/o focal WMA in 2091. Inferior STEMI 7/96 due to near occlusion of PLV resulting in BMS to PLV and mid RCA.
2. HTN/Hypertensive heart disease: rather difficult to control HTN, needing multiple agents to achieve near normal to high normal pressures. Had a good bp response to thiazides but caused unacceptable degree of hyponatremia. BP also related to the degree of chronic low back pain, and was substantially better for 6 months after localized treatment. Hypertensive HD manifest primarily as peripheral edema, worsened by norvasc, improved on diuretic.
3. Dyslipidemia
4. DMII
5. Postherpetic neuralgia
6. Depression
7. Hemorrhoids
8. Osteoarthritis
9. h/o DVT /PE
10. Breast CA 96 T2N1M0 ER/PR+ Her2 neg breast CA s/p L radical mastectomy, XRT, and tamoxifen x 10 years
11. Recurrent UTI
12. Urinary incontinence
13. History of HIT AB Positive 2090

Family history:
There is a history of non-premature cardiovascular and cerebrovascular disease: Her father died of a stroke in his 80s, her mother had congestive heart failure, and a brother had an MI. Her other siblings have lived into their 80s.

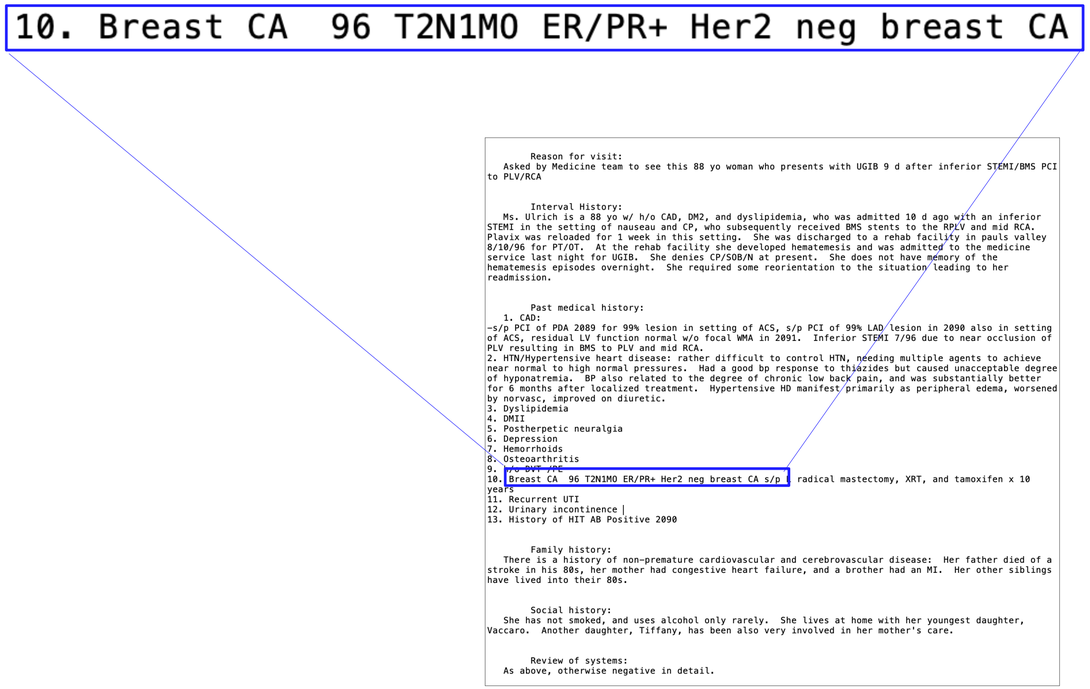
Social history:
She has not smoked, and uses alcohol only rarely. She lives at home with her youngest daughter, Vaccaro. Another daughter, Tiffany, has been also very involved in her mother's care.

Review of systems:
As above, otherwise negative in detail.

Introduction

Clinical information extraction

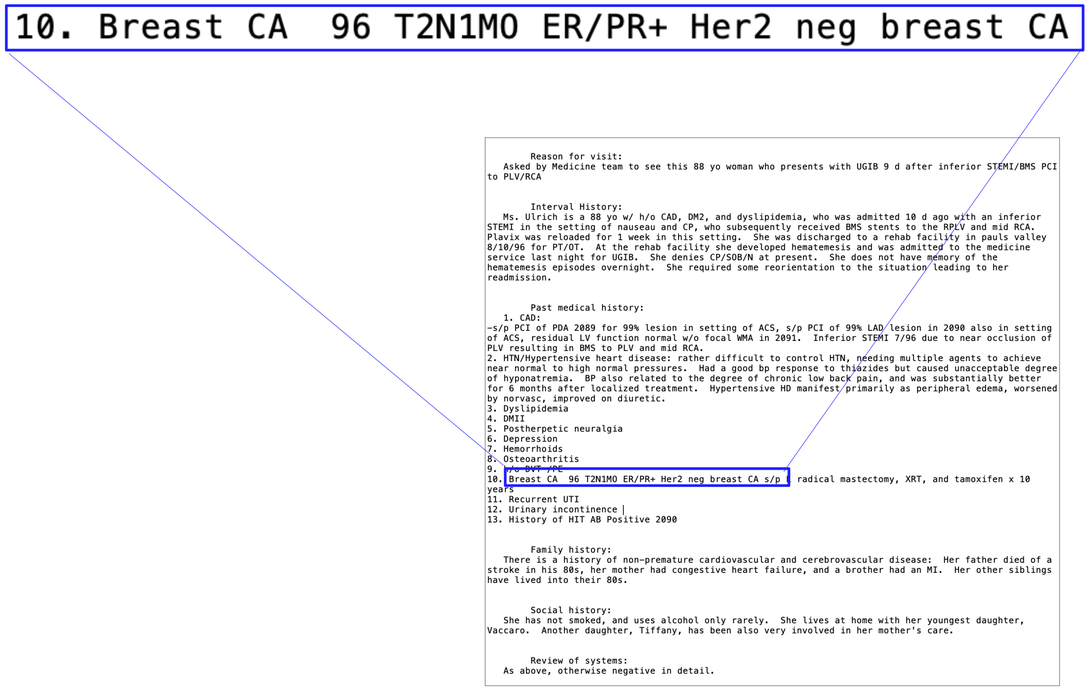
Typically requires a **dictionary lookup** linking text to standard terminologies, searching the text for mentions of concept terms from the *dictionary* (i.e., standard terminology). It is sometimes also called "concept extraction," "concept normalization" or even "concept recognition" even if the latter would be closer to "named-entity recognition" or "entity recognition."



Introduction

Clinical information extraction

Typically requires a **dictionary lookup** linking text to standard terminologies, searching the text for mentions of concept terms from the *dictionary* (i.e., standard terminology). It is sometimes also called "concept extraction," "concept normalization" or even "concept recognition" even if the latter would be closer to "named-entity recognition" or "entity recognition."



- Breast cancer (254837009 SNOMED-CT "Malignant neoplasm of breast (disorder))
- T2 category (67673008)
- N1 category (53623008)
- M0 category (30893008)

etc.

Introduction

Clinical information extraction (cont.)

Most natural language processing (NLP) software applications used with EHR text include some dictionary lookup. These applications include prominent examples such as:

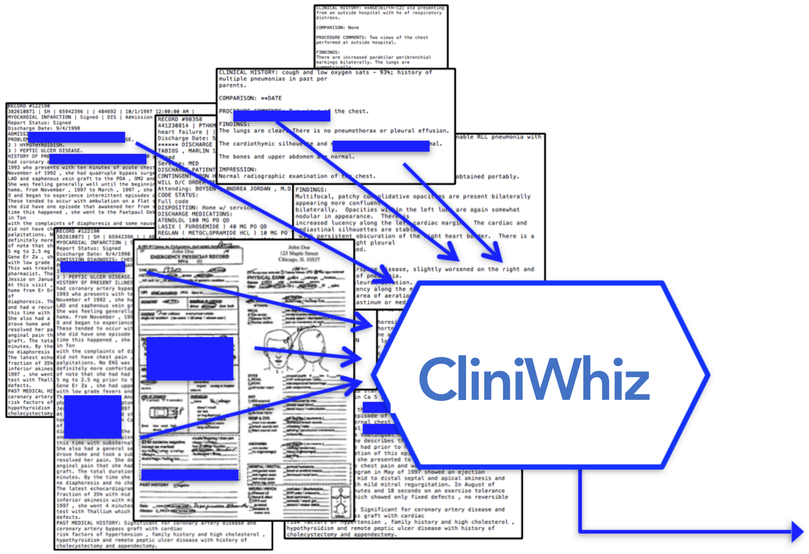
- MetaMap,
- MedLEE (now commercially available as REVEAL, from Health Fidelity, Palo Alto, CA),
- NOBLE Coder,
- NCBO Annotator,
- Textractor,
- Apache cTAKES.

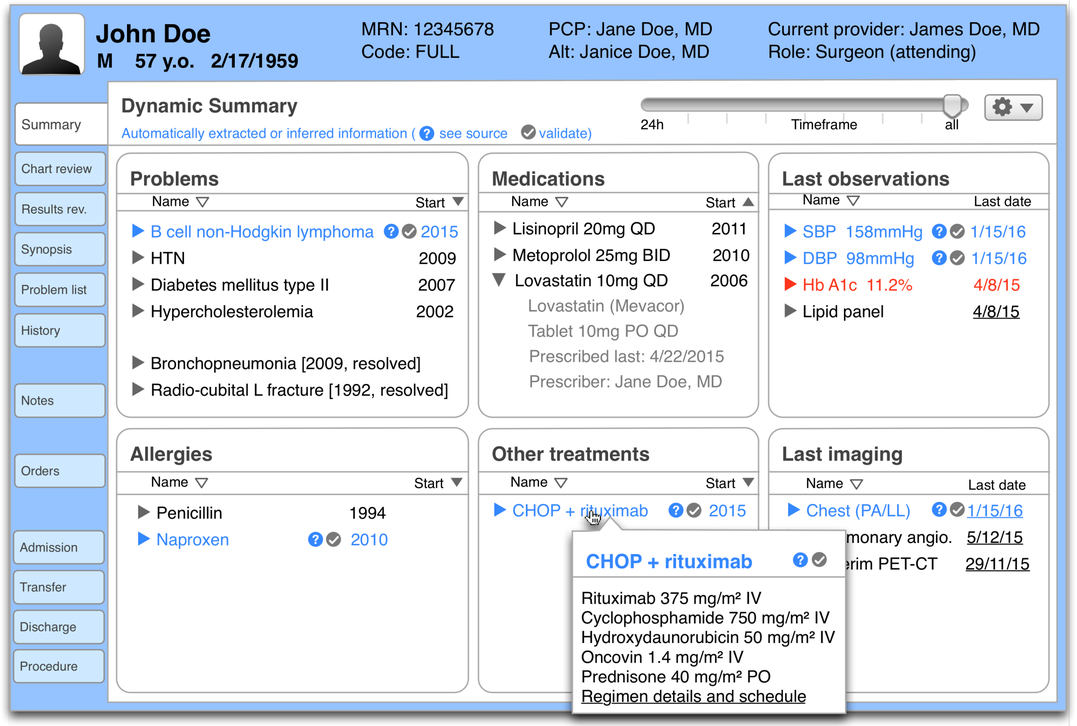
Some use their own dictionary lookup algorithm while others use existing algorithms like Apache Lucene or Apache UIMA ConceptMapper.

Introduction

Clinical information extraction for structured summarization of key patient information

Work realized in the context of a larger project to automatically extract information from the EHR with high accuracy to then improve the completeness and timeliness of lists of medical problems and allergies.





Methods

Variety of dictionary lookup tools, text corpora, and dictionaries used in this study

Dictionary lookup tools:

- Apache Lucene,
- Apache UIMA ConceptMapper
- Apache cTAKES (fast lookup)

Dictionaries:

- SNOMED-CT CORE subset
- Custom focused dictionary

Clinical text corpora:

- University of Utah corpus
- Medical University of South Carolina corpus

Methods

Dictionary lookup tools

Apache Lucene: popular and powerful text search engine library used by numerous websites and applications (e.g., LinkedIn and Twitter). Used in Textractor and in our prototype application for extracting medical problems and allergens from clinical notes.

In CliniWhiz, we combine it with a normalization process that includes abbreviation expansion, stemming, removal of punctuation, lowercasing, reordering of tokens and removal of stopwords. Also uses noun phrase chunks and named entities detected by a machine learning classifier.

Apache UIMA ConceptMapper is a dictionary lookup tool (part of Apache UIMA) that is also powerful and highly configurable, capable of non-contiguous terms mapping and fast performance.

Apache cTAKES is a popular open source clinical NLP application (built on Apache UIMA) offering a fast dictionary lookup module in its latest version (4.0).

Methods

Clinical text corpora

- Utah corpus (770 clinical notes)
- MUSC corpus (522 clinical notes)

Both de-identified and annotated for

a selection of medical problems and allergens.

	MUSC corpus	Utah corpus
Training set		
Notes count	247	495
Avg. words count	746	905
Problem annotations	4777	9344
Allergen annotations	101	70
Test set		
Notes count	275	275
Avg. words count	717	904
Problem annotations	5793	5361
Allergen annotations	126	63

Dictionaries

- SNOMED-CT CORE (6,117 concepts with 106,616 terms)
- Focused dictionary was semi-automatically built using a set of 168 problems and 138 allergens expanded using the UMLS Metathesaurus (24,833 concepts with 134,408 terms)

Results

Concept normalization speed:

Measured in seconds per note and seconds per 5,000 characters to account for note size differences between corpora

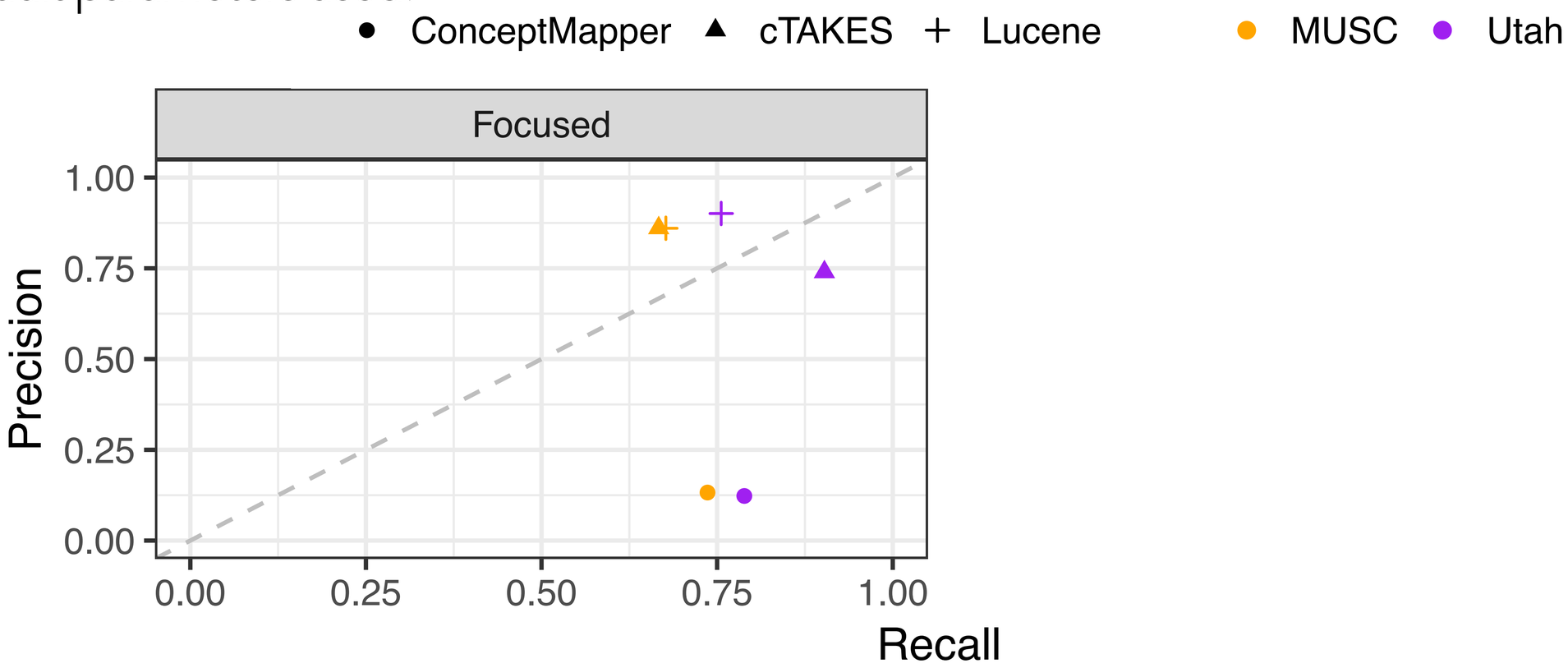
	SNOMED CORE				Custom dictionary				Average
	Utah corpus		MUSC corpus		Utah corpus		MUSC corpus		
	s/note	s/5K char	s/note	s/5K char	s/note	s/5K char	s/note	s/5K char	s/5K char
Lucene v7.7 normalized	0.924	0.797	2.264	2.356	1.111	0.959	2.379	2.476	1.647
Lucene v7.7 no normalization	0.008	0.007	0.009	0.009	0.007	0.006	0.009	0.009	0.008
ConceptMapper	0.003	0.003	0.003	0.003	0.007	0.006	0.007	0.008	0.004
cTAKES fast lookup	0.022	0.019	0.023	0.023	0.010	0.008	0.010	0.011	0.015

Results

Concept extraction accuracy:

When only considering the identification of mentions of problems or allergens with overlapping text spans, recall ranged from 66.7% to 90.3% with the focused dictionary.

Default parameters used.

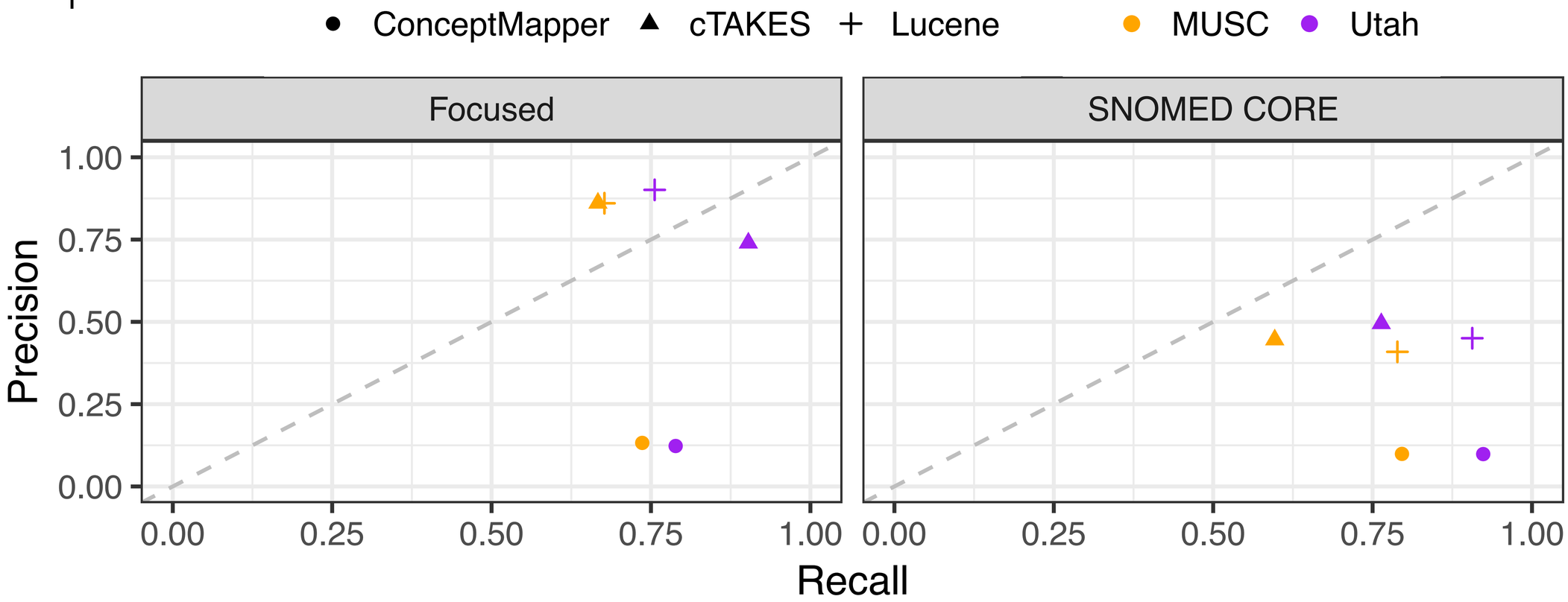


Results

Concept extraction accuracy:

When only considering the identification of mentions of problems or allergens with overlapping text spans, recall ranged from 59.65% to 91.52% with the SNOMED CORE dictionary

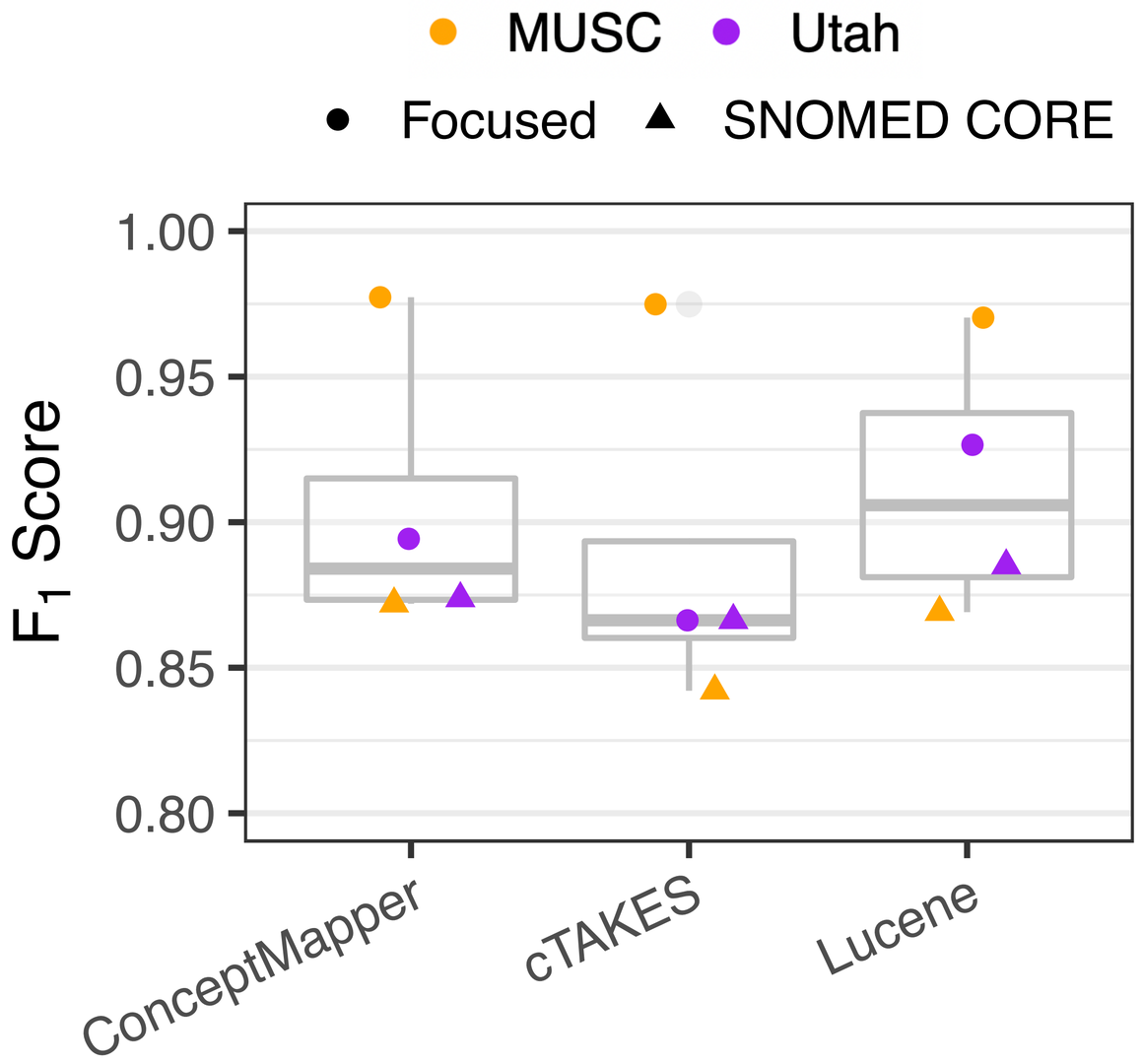
Default parameters used.



Results

Concept normalization accuracy:

When assessing the mapping of mentions of problems or allergens identified in the previous step with UMLS Metathesaurus concepts, the F_1-measure ranged from 84.2% to 97.49%.



Discussion and Conclusion

Very large variation in processing speed was measured, mostly caused by normalization processes.

When comparing similar dictionary lookup processes without normalization, differences in speed shrank, but accuracy was also negatively affected.

Using Lucene without normalization caused a drop in mention identification recall (3.7-9.2% less), a slight increase in mention identification precision (0.3-4.7% more), and an increase in concept normalization F_1-measure of 5.2%.

Limitations: No medication causing allergy filtering used, causing low precision. SNOMED CORE had only partly overlapping coverage causing lower recall and precision.

These very large variations in accuracy and processing speed motivated an extensive study of the impact of dictionary lookup algorithms and parameters, dictionaries used and corpora characteristics.

Acknowledgments

Paul Heider and Youngjun Kim (NLP experts)

Andrew Trice and Gary Underwood (Software developers and machine learning experts)

Work supported in by the National Cancer Institute (R41CA180190)

Thank you!

Questions and comments: meystre@musc.edu
Profile: https://profiles.healthsciencessc.org/Stephane.Meystre
Lab website: http://meystrelab.org