Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES)

Guergana K. Savova, PhD
Pei Chen
Boston Children's Hospital
Harvard Medical School

Guergana.Savova@childrens.harvard.edu
chenpei@apache.org

Acknowledgments

- NIH
  - Multi-source integrated platform for answering clinical questions (MiPACQ) (NLM RC1LM010608)
  - Temporal Histories of Your Medical Event (THYME) (NLM 10090)
  - Shared Annotated Resources (ShARe) (NIGMS R01GM090187)
  - Informatics for Integrating Biology and the Bedside (i2b2) (NLM U54LM008748)
  - Electronic Medical Records and Genomics (eMERGE) (NIH 1U01HG006828)
  - Pharmacogenomics Research (PGRN) (NIH 1U01GM092691-01)
- Office of the National Coordinator of Healthcare Technologies (ONC)
  - Strategic Healthcare Advanced Research Project: Area 4, Secondary Use of the EMR data (SHARPn) (ONC 90TR0002)
- Industry
  - IBM UIMA grant
- Institutions contributing de-identified clinical notes
  - Mayo Clinic, Seattle Group Health Cooperative, MIMIC project (Beth Israel)

Outline

- Current Healthcare Challenges
- Apache cTAKES
- Technical details
- Demo

Patient January 16, 2006

Total weight of printed pages presented for review:
5 lbs.

Image courtesy of Piet C. de Groen

Patient January 16, 2006

Total number of X-rays presented for review:
16,902

Image courtesy of Piet C. de Groen

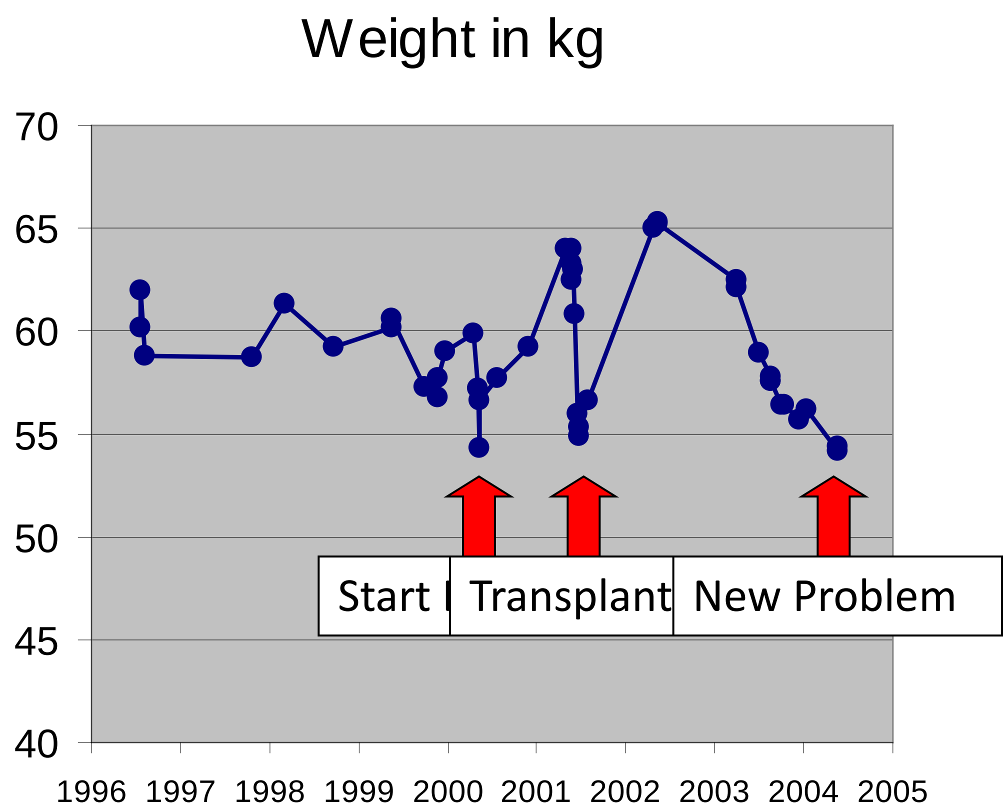
Questions

- What is exactly the patient's problem?
  - Are liver tests and weight loss due to Lipitor?
  - When did she use Lipitor?
  - What was the weight on what date?

- Impossible to review all notes!
  - Which notes are relevant to current symptoms?
  - Which have notes have weights and drug information?

EHR/Data Warehouse to the rescue!

- Structured Data
- Demographics
- ICD9 Codes
- Patient Vitals
  - weight



Slide courtesy of Piet C. de Groen

What happened to Cholesterol?

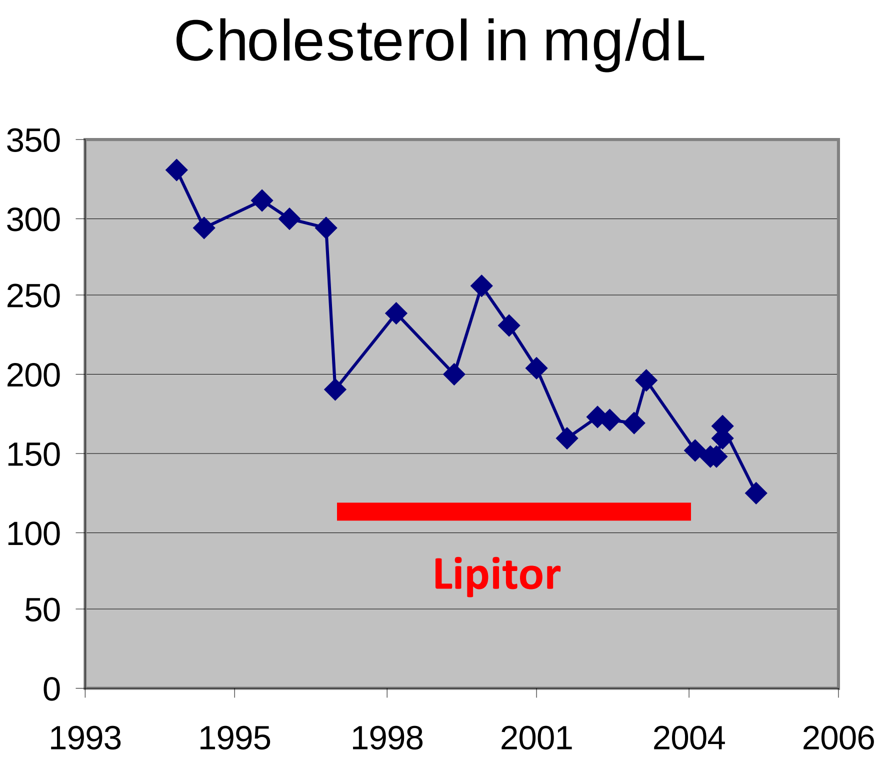
- She was on Lipitor, but:
  - When was it discontinued?
  - Did it do anything to her lipid levels?

NLP to the rescue!

- Sort 33 identified Clinical Notes on date
- First note is from 1997
  - Lipitor is highlighted in the note
  - …Dr. X recommended discontinuation of Pravachol and initiation of Lipitor … have written a prescription for Lipitor …
- Last note is from 2005
  - … Lipitor was discontinued in 2004 …
  - March 2004 note confirms discontinuation

Complete Picture

- Demographics
  - Paitent ID #
- Tests
  - Cholesterol exists
- Clinical Notes
  - "Lipitor"
- Result
  - 22 cholesterol levels
  - 243 notes: 33 mentioned "**Lipitor**"



Slide courtesy of Piet C. de Groen

NLP Areas of Research

- Part of speech tagging
- Parsing – constituency and dependency
- Predicate-argument structure (semantic role labeling)
- Named entity recognition
- Word sense disambiguation
- Relation discovery and classification
- Discourse parsing (text cohesiveness)
- Language generation
- Machine translation
- Summarization
- Creating datasets to be used for learning
  - a.k.a. computable gold annotations
  - Active learning

NLP: Example 1

I	saw	the	man	with	the	telescope.
w1	w2	w3	w4	w5	w6	w7
pronoun	verb	article	noun	prep	article	noun

NP VP NP

NP VP NP PP

NLP: Example 2

I	saw	the	man	with	the	stethoscope.
w1	w2	w3	w4	w5	w6	w7
pronoun	verb	article	noun	prep	article	noun

$\underbrace{\text{I}}_{\text{NP}}\ \underbrace{\text{saw}}_{\text{VP}}\ \underbrace{\text{the man with the stethoscope.}}_{\text{NP}}$

How do we get the semantics?

Clinical Text Analysis and Knowledge Extraction System (cTAKES)

JAMIA, 2010

Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications

Guergana K Savova,[1] James J Masanz,[1] Philip V Ogren,[2] Jiaping Zheng,[1] Sunghwan Sohn,[1] Karin C Kipper-Schuler,[1] Christopher G Chute[1]

► Additional tables and appendices are published online only. To view these files please visit the journal online (http://jamia.bmj.com).

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA
[2]Computer Science Department, University of Colorado, Denver, Colorado, USA

Correspondence to
Guergana Savova, Children's Hospital Informatics Program, Children's Hospital Boston and Harvard Medical School, 300 Longwood Avenue, Enders 138, Boston, MA 02115, USA; guergana.savova@childrens.harvard.edu

The annotation guidelines will be made available at http://www.ohnlp.org after manuscript publication. The clinical corpus created from Mayo Clinic notes is not released with cTAKES. For model-building purposes, that corpus was anonymized per Safe Harbor Health Insurance Portability and Accountability Act[76] guidelines. Technical details and discussions on technical topics related to cTAKES are posted on the Forums at http://www.ohnlp.org.

Received 30 October 2009
Accepted 29 June 2010

ABSTRACT

We aim to build and evaluate an open-source natural language processing system for information extraction from electronic medical record clinical free-text. We describe and evaluate our system, the clinical Text Analysis and Knowledge Extraction System (cTAKES), released open-source at http://www.ohnlp.org. The cTAKES builds on existing open-source technologies—the Unstructured Information Management Architecture framework and OpenNLP natural language processing toolkit. Its components, specifically trained for the clinical domain, create rich linguistic and semantic annotations. Performance of individual components: sentence boundary detector accuracy=0.949; tokenizer accuracy=0.949; part-of-speech tagger accuracy=0.936; shallow parser F-score=0.924; named entity recognizer and system-level evaluation F-score=0.715 for exact and 0.824 for overlapping spans, and accuracy for concept mapping, negation, and status attributes for exact and overlapping spans of 0.957, 0.943, 0.859, and 0.580, 0.939, and 0.839, respectively. Overall performance is discussed against five applications. The cTAKES annotations are the foundation for methods and modules for higher-level semantic processing of clinical free-text.

INTRODUCTION

The electronic medical record (EMR) is a rich source of clinical information. It has been advocated that EMR adoption is a key to solving problems related to quality of care, clinical decision support, and reliable information flow among individuals and departments participating in patient care.[1] The abundance of unstructured textual data in the EMR NLP system designed to process and extract semantically viable information to support the heterogeneous clinical research domain and to be sufficiently scalable and robust to meet the rigors of a clinical research production environment. This paper describes and evaluates our system—the clinical Text Analysis and Knowledge Extraction System (cTAKES).

BACKGROUND

The clinical narrative has unique characteristics that differentiate it from scientific biomedical literature and the general domain, requiring a focused effort around methodologies within the clinical NLP field.[2] Columbia University's proprietary Medical Language Extraction and Encoding System (MedLEE)[3] was designed to process radiology reports, later extended to other domains,[4] and tested for transferability to another institution.[5] MedLEE discovers clinical concepts along with a set of modifiers. Health Information Text Extraction (HITEx)[6 7] is an open-source clinical NLP system from Brigham and Women's Hospital and Harvard Medical School incorporated within the Informatics for Integrating Biology and the Bedside (i2b2) toolset.[8] IBM's BioTeKS[9] and MedKAT[10] were developed as biomedical-domain NLP systems. SymText and MPLUS[11 12] have been applied to extract the interpretations of lung scans[13] to detect pneumonia[14] and central venous catheters mentions.[15] Other tools developed primarily for processing biomedical scholarly articles include the National Library of Medicine MetaMap,[16] providing mappings to the Unified Medical Language System (UMLS) Metathesaurus concepts,[17 18] those from the National Center for Text Mining (NaCTeM),[19] JULIE lab,[20] and

OPEN ACCESS

JAMIA, 2013

Towards comprehensive syntactic and semantic annotations of the clinical narrative

Daniel Albright,[1] Arrick Lanfranchi,[1] Anwen Fredriksen,[1] William F Styler IV,[1] Colin Warner,[2] Jena D Hwang,[1] Jinho D Choi,[3] Dmitriy Dligach,[4] Rodney D Nielsen,[1,5] James Martin,[3] Wayne Ward,[3] Martha Palmer,[1] Guergana K Savova[4]

[1]Department of Linguistics, University of Colorado, Boulder, Colorado, USA
[2]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Department of Computer Science University of Colorado, Boulder, Colorado, USA
[4]Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[5]Department of Computer Science and Engineering, University of North Texas, Texas, USA

Correspondence to Dr Guergana K Savova, Boston Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02114, USA; Guergana.Savova@childrens.harvard.edu

Received 3 September 2012
Revised 27 December 2012
Accepted 28 December 2012

ABSTRACT

Objective To create annotated clinical narratives with layers of syntactic and semantic labels to facilitate advances in clinical natural language processing (NLP). To develop NLP algorithms and open source components.

Methods Manual annotation of a clinical narrative corpus of 127 606 tokens following the Treebank schema for syntactic information, PropBank schema for predicate-argument structures, and the Unified Medical Language System (UMLS) schema for semantic information. NLP components were developed.

Results The final corpus consists of 13 091 sentences containing 1772 distinct predicate lemmas. Of the 766 newly created PropBank frames, 74 are verbs. There are 28 539 named entity (NE) annotations spread over 15 UMLS semantic groups, one UMLS semantic type, and the Person semantic category. The most frequent annotations belong to the UMLS semantic groups of Procedures (15.71%), Disorders (14.74%), Concepts and Ideas (15.10%), Anatomy (12.80%), Chemicals and Drugs (7.49%), and the UMLS semantic type of Sign or Symptom (12.46%). Inter-annotator agreement results: Treebank (0.926), PropBank (0.891–0.931), NE (0.697–0.750). The part-of-speech tagger, constituency parser, dependency parser, and semantic role labeler are built from the corpus and released open source. A significant limitation uncovered by this project is the need for the NLP community to develop a widely agreed-upon schema for the annotation of clinical concepts and their relations.

Conclusions This project takes a foundational step towards bringing the field of clinical NLP up to par with NLP in the general domain. The corpus creation and NLP components provide a resource for research and application development that would have been previously impossible.

other), the level of certainty associated with an event (confirmed, possible, negated) as well as textual mentions that point to the same event. We describe our efforts to combine annotation types developed for general domain syntactic and semantic parsing with medical-domain-specific annotations to create annotated documents accessible to a variety of methods of analysis including algorithm and component development. We evaluate the quality of our annotations by training supervised systems to perform the same annotations automatically. Our effort focuses on developing principled and generalizable enabling computational technologies and addresses the urgent need for annotated clinical narratives necessary to improve the accuracy of tools for extracting comprehensive clinical information.[1] These tools can in turn be used in clinical decision support systems, clinical research combining phenotype and genotype data, quality control, comparative effectiveness, and medication reconciliation to name a few biomedical applications.

In the past decade, the general natural language processing (NLP) community has made enormous strides in solving difficult tasks, such as identifying the predicate-argument structure of a sentence and associated semantic roles, temporal relations, and coreference which enable the abstraction of the meaning from its surface textual form. These developments have been spurred by the targeted enrichment of general annotated resources (such as the Penn Treebank (PTB)[2]) with increasingly complex layers of annotations, each building upon the previous one, the most recent layer being the discourse level.[3] The emergence of other annotation standards (such as PropBank[4] for the annotation of the sentence predicate-argument structure) has brought new progress in the annotation of semantic informa-

About Apache cTAKES

ctakes.apache.org

cTAKES™

The Apache Software Foundation
http://www.apache.org/

General
- About
- Getting Started
- Downloads
- Glossary
- Archives

Community
- Get Involved
- Bug Tracker
- Mailing Lists
- People
- License
- History
- Community FAQs

Users
- User Guide
- User FAQs

Developers
- Developer Guide
- Developer FAQs

PMC
- PMC FAQs
- Release Guide

ASF
- Apache Software Foundation
- Thanks
- Become a Sponsor

Welcome to Apache cTAKES

Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) is an open-source natural language processing system for information extraction from electronic medical record clinical free-text. It processes clinical notes, identifying types of clinical named entities from various dictionaries including the Unified Medical Language System (UMLS) - medications, diseases/disorders, signs/symptoms, anatomical sites and procedures. Each named entity has attributes for the text span, the ontology mapping code, subject (patient, family member, etc.) and context (negated/not negated, conditional, generic, degree of certainty). Some of the attributes are expressed as relations, for example the location of a clinical condition (locationOf relation) or the severity of a clinical condition (degreeOf relation).

Apache cTAKES was built using the Apache UIMA Unstructured Information Management Architecture engineering framework and Apache OpenNLP natural language processing toolkit. Its components are specifically trained for the clinical domain out of diverse manually annotated datasets, and create rich linguistic and semantic annotations that can be utilized by clinical decision support systems and clinical research. cTAKES has been used in a variety of use cases in the domain of biomedicine such as phenotype discovery, translational science, pharmacogenomics and pharmacogenetics.

Apache cTAKES employs a number of rule-based and machine learning methods. Apache cTAKES components include:

1. Sentence boundary detection
2. Tokenization (rule-based)
3. Morphologic normalization
4. POS tagging
5. Shallow parsing
6. Named Entity Recognition
   - Dictionary mapping
   - Semantic typing is based on these UMLS semantic types: diseases/disorders, signs/symptoms, anatomical sites, procedures, medications
7. Assertion module
8. Dependency parser
9. Constituency parser
10. Semantic Role Labeler
11. Coreference resolver
12. Relation extractor
13. Drug Profile module
14. Smoking status classifier

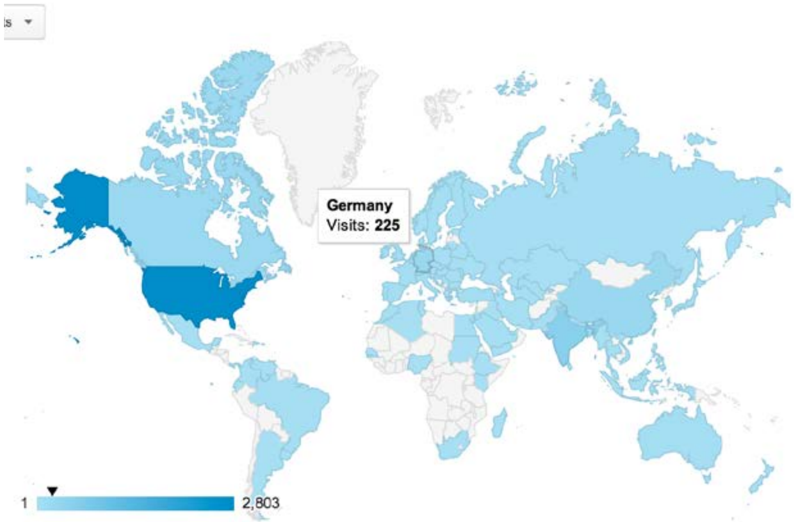
The goal of cTAKES is to be a world-class natural language processing system in the healthcare domain. cTAKES can be used in a great variety of retrievals and use cases. It is intended to be modular and expandable at the information model and method level. The cTAKES community is committed to best practices and R&D (research and development) by using cutting edge technologies and novel research. The idea is to quickly translate the best performing methods into cTAKES code.

ctakes.apache.org

Recent Developments

- cTAKES
  - Top-level Apache Software Foundation project (as of March 22, 2013)
  - many new components for semantic processing
  - multi-institutional contributions (not an exhaustive list and in no particular order)
    - Boston Childrens Hospital
    - Mayo Clinic
    - University of Colorado
    - MITRE
    - MIT
    - Seattle Group Health Cooperative
    - University of California, San Diego
    - …

Apache cTAKES Usage



1.	United States	2,803	2.85	00:03:12	43.92%
2.	India	499	2.32	00:02:58	54.11%
3.	China	242	1.74	00:01:07	88.43%
4.	Germany	225	2.04	00:01:19	48.44%
5.	Canada	222	2.46	00:02:47	29.73%
6.	United Kingdom	111	2.10	00:01:39	78.38%
7.	South Korea	67	2.28	00:01:21	59.70%
8.	Japan	66	2.58	00:01:57	53.03%
9.	Taiwan	58	2.16	00:01:51	53.45%
10.	(not set)	58	1.74	00:00:42	96.55%
11.	France	48	1.81	00:01:03	70.83%
12.	Australia	45	2.40	00:04:24	73.33%
13.	Turkey	45	3.60	00:01:14	42.22%
14.	Italy	42	2.36	00:00:58	92.86%
15.	Spain	41	3.10	00:04:39	70.73%
16.	Brazil	39	2.26	00:02:17	94.87%
17.	Sweden	30	3.53	00:03:33	66.67%
18.	Russia	28	1.86	00:01:19	96.43%
19.	Switzerland	27	2.93	00:01:27	74.07%
20.	Greece	26	3.88	00:04:04	46.15%

Why ASF?

ASF provides necessary parts for a community driven project to succeed:

- Infrastructure
  - Compile Servers
  - Jira Issues Tracking
  - Mail Servers/Mailing Lists
  - SVN/MVN Repositories
  - Wiki
- Governance Framework
  - Meritocracy
  - Voting process
  - Organization Structure
    (user | developer | committer | PMC member | PMC chair | ASF member)

http://www.apache.org/foundation/how-it-works.html

The Apache Way

- collaborative software development
- commercial-friendly standard license
- consistently high quality software
- respectful, honest, technical-based interaction
- faithful implementation of standards
- security as a mandatory feature
- keep things as public as possible

apache.org/foundation/how-it-works.html#management

Get Involved!

- You don't need to be a software developer to contribute to Apache cTAKES
  - provide feedback
  - write or update documentation
  - help new users
  - recommend the project to others
  - test the code and report bugs
  - fix bugs
  - give us feedback on required features
  - write and update the software
  - create artwork
  - anything you can see that needs doing
- All of these contributions help to keep a project active and strengthen the community.

Mailing Lists

Subscribe:

- Development List: dev-subscribe@ctakes.apache.org
- Commits List: commits-subscribe@ctakes.apache.org
- Users List: user-subscribe@ctakes.apache.org

cTAKES: Components

- **Sentence boundary detection (OpenNLP technology)**
- **Tokenization (rule-based)**
- **Morphologic normalization (NLM's LVG)**
- **POS tagging (OpenNLP technology)**
- **Shallow parsing (OpenNLP technology)**
- **Named Entity Recognition**
    - Dictionary mapping (lookup algorithm)
    - Machine learning (MAWUI)
    - types: diseases/disorders, signs/symptoms, anatomical sites, procedures, medications
- **Negation and context identification (NegEx)**
- **Dependency parser**
- **Constituency parser**
- **Dependency based Semantic Role Labeling**
- **Relation Extraction**
- **Coreference module**
- **Drug Profile module**
- **Smoking status classifier**
- **Clinical Element Model (CEM) normalization module**

cTAKES Technical Details

- Open source
  - Apache Software Foundation project
  - Java 1.6 or higher
  - Dependency on UMLS which requires a UMLS license (free)
- Framework
  - Apache Unstructured Information Management Architecture (UIMA) engineering framework
- Methods
  - Natural Language Processing methods (NLP)
  - Based on standards and conventions to foster interoperability
- Application
  - High-throughput system

Toolkits used

- Don't reinvent the Wheel!
  - UIMA
  - UIMA-AS
  - OpenNLP
  - clearTK
  - uimaFIT

    Component implementation, instantiation, definition, execution via Java code w/o xml descriptors.

    Utils

Medication CEM template

- associatedCode
- Change_status
- Conditional
- Dosage
- Duration
- End_date
- Form
- Frequency
- Generic
- Negation_indicator
- Route
- Start_date
- Strength
- Subject
- Uncertainty_indicator

Procedure CEM template

- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Device
- End_date
- Generic
- Method
- Negation_indicator
- Relative_temporal_context
- Start_date
- Subject
- Uncertainty_indicator

Sign/Symptom CEM template

- Alleviating_factor
- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Course
- Duration
- End_time
- Exacerbating_factor
- Generic
- Negation_indicator
- Relative_temporal_context
- Severity
- Start_time
- Subject
- Uncertainty_indicator

Lab CEM template

- Abnormal_interpretation
- associatedCode
- Conditional
- Delta_flag
- Estimated_flag
- Generic
- Lab_value
- Negation_indicator
- Ordinal_interpretation
- Reference_range_narrative
- Subject
- Uncertainty_indicator

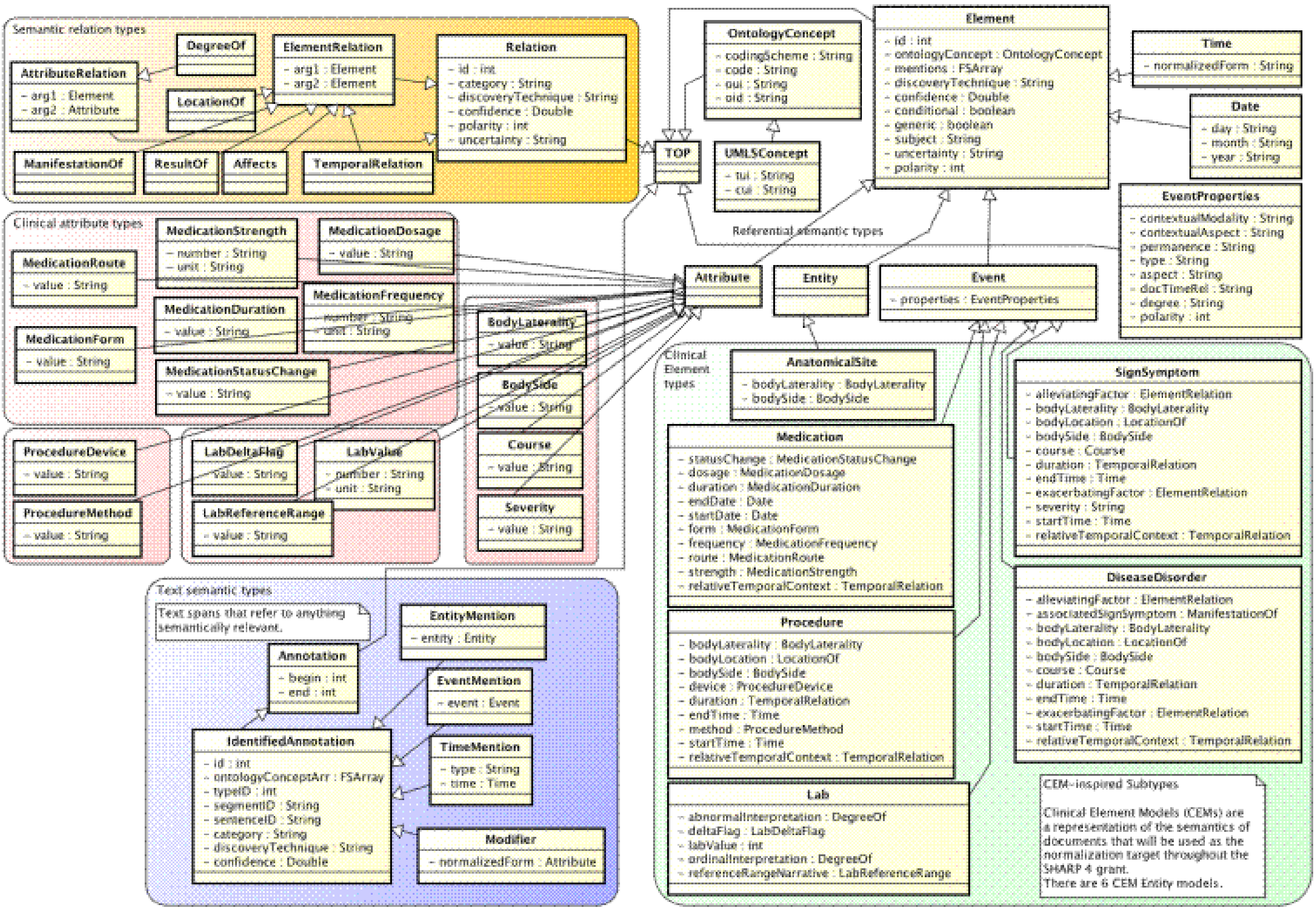
Disease/Disorder CEM template

- Alleviating_factor
- Associated_sign_or_symptom
- associatedCode
- Body_laterality
- Body_location
- Body_side
- Conditional
- Course
- Duration
- End_time
- Exacerbating_factor
- Generic
- Negation_indicator
- Relative_temporal_context
- Severity
- Start_time
- Subject
- Uncertainty_indicator

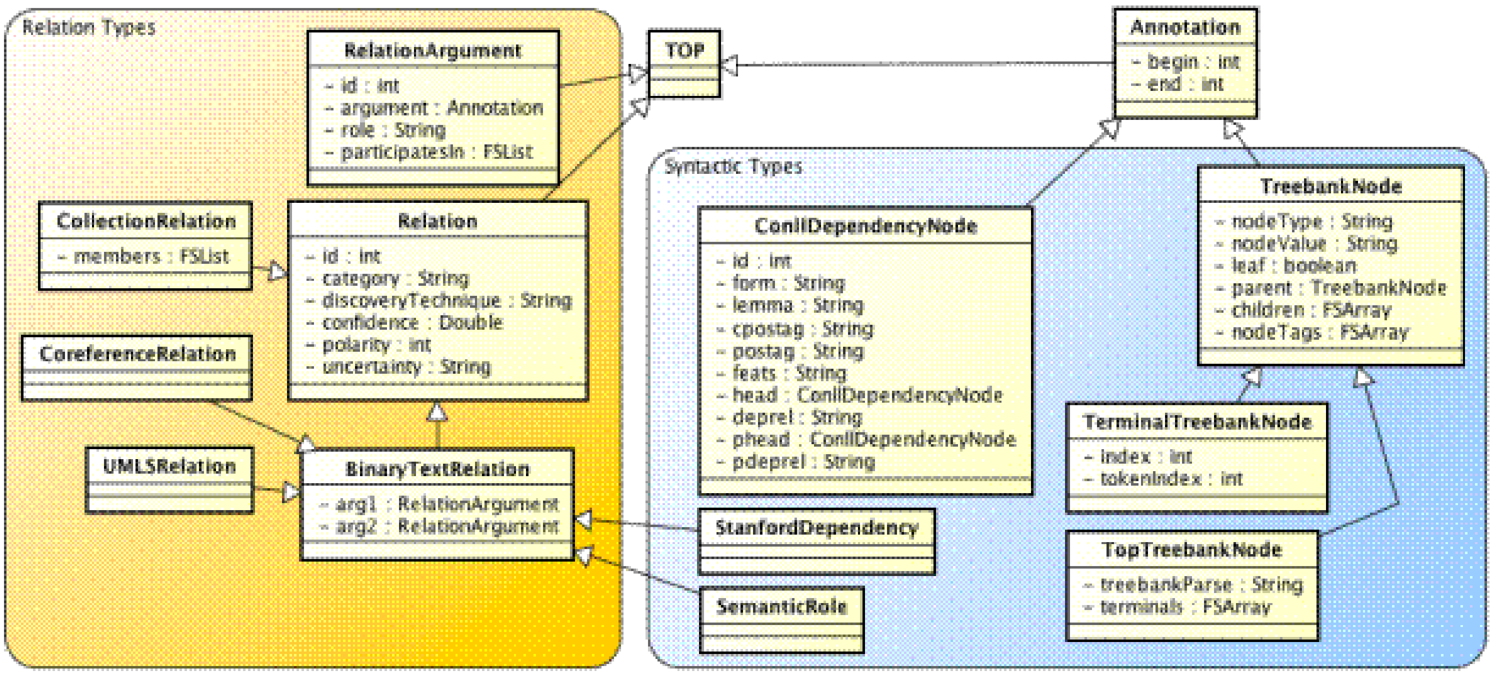
Anatomical Site CEM template

- associatedCode
- Body_laterality
- Body_site
- Conditional
- Generic
- Negation_indicator
- Subject
- Uncertainty_indicator

cTAKES Type System



Additional Spanned Types



UMLS, Named Entity Recognition

UMLS Semantic Types, Groups and Relations

- UMLS (Unified Medical Language System) was developed to help with cross-linguistic translation of medical concepts
- Bodenreider and McCray (see Table 1 and Figure 3) http://semanticnetwork.nlm.nih.gov/SemGroups/Papers/2003-medinfo-atm.pdf
- http://clear.colorado.edu/compsem/documents/umls_guidelines.pdf

UMLS Example

- The patient underwent a radical tonsillectomy (with additional right neck dissection) for metastatic squamous cell carcinoma. He returns with a recent history of active bleeding from his oropharynx.

Example UMLS annotations:

<u>Entities</u>
[patient]: Person
[radical tonsillectomy (with additional right neck dissection)]: Procedure
[radical tonsillectomy]: Procedure
[additional right neck dissection]: Procedure
[right neck]: Anatomy
[metastatic squamous cell carcinoma]: Disorder
[active bleeding from his oropharynx]: Disorder
[active bleeding]: Disorder
[oropharynx]: Anatomy

UMLS Terminology Services

- https://uts.nlm.nih.gov/home.html
  - Colorectal cancer
  - Ascending colon
  - MS
- Named entities
  - Mentions that belong to a particular semantic type (Ms. Smith – Person; colorectal cancer – Disease/Disorder; ascending colon – anatomical site; joint pain – sign/symptom)
  - Anything that can be referred to with a proper name

Named Entity Recognition

- Methods for discovering mentions of particular semantic types
  - Finding the spans of text that constitute the entity mention
  - Classifying the entities according to their semantic type
- Ambiguity in NER
  - MS
    - Patient diagnosed with MS
    - Ms Smith was diagnosed with RA

Normalization of Named Entities

- Assigning an ontology code to varied surface forms
  - Patient diagnosed with *RA (C0003873)*
  - Patient diagnosed with *Rheumatoid Arthritis (C0003873)*
  - Patient diagnosed with *atrophic arthritis (C0003873)*

Attributes: Negation and Uncertainty

- Negation – entity mention is negated
  - Patient denies *foot joint pain*.
    - foot joint pain, negated
    - C0458239, negated
- Uncertainty – degree of uncertainty is associated with the entity mention
  - Results suggestive of *colorectal cancer*.
    - colorectal cancer, probable
    - C1527249, probable

Relation Extraction (UMLS)

- Upcoming JAMIA manuscript

Dligach, Dmitriy; Bethard, Steven; Becker, Lee; Miller, Timothy; Savova, Guergana. (in press). Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association.

DISORDER

PHENOMENON

PERSON

“The patient has strep throat which is hindering her eating. We are treating it with Azithromycin.”

CHEMICAL/DRUG

DISRUPTS

“The patient has strep throat which is hindering her eating.
We are treating it with Azithromycin.”

MANAGES/TREATS

UMLS Relations

- UMLS relations of interest:
  - *LocationOf(anatomical site, disease/disorder)*
  - *LocationOf(anatomical site, sign/symptom)*
  - *DegreeOf(modifier, disease/disorder)*
- Examples:
  - *LUNGS: Equal AE bilaterally, no rales, no rhonchi.*
  - *LocationOf(lungs, rales)*
  - *LocationOf(lungs, rhonchi)*
- DegreeOf relation
  - *Severe headache*
  - *DegreeOf(severe, headache)*

Modifiers

- DegreeOf
  - Modifiers
  - Entities
- Modifier discovery module
  - Implemented in cTAKES
  - BIO (Begin, Inside, Outside) representation
  - Word features
  - Algorithm: SVM
- Informal evaluation results

Relation Learning

- Statistical classifier
  - Input: a pair of entities
  - Output: relation / no relation label
- Training
  - Pair up all entity pairs
  - Assign a gold relation label (including NONE)
  - Downsample
  - Train an SVM model
- Testing
  - Pair up all entities in test set
  - Pass to the model
  - Assign label

Features

- Word features
  - Words of mentions
  - Context words
  - Distance
- Named entity features
  - Entity types
  - Entity context
- POS features
  - POS tags of entities
  - POS tags between entities
- Dependency features
  - Distance to common ancestor
  - Dependency path features
  - Governing/depedent word
- Chunking features
  - Head word of phrases between entities
  - Phrase head context
- Wikipedia features
  - Entity similarity
  - Article titles

Annotated Data

- SHARP

Total notes	Instances of LocationOf	Instances of DegreeOf
80	1852	308

- ShARe
  - Anatomical Sites and Disease/Disorders

Total notes	Instances of LocationOf	Instances of DegreeOf
130	2190	702

Evaluation

- Two-fold cross validation
- LibSVM
- Parameter search
  - Kernel (Linear/RBF)
  - SVM Cost parameter
  - RBF gamma parameter
  - Probability of keeping a negative example
- Evaluation on gold entities

Results

	F1 Score	
	SHARP	ShARe
LocationOf relation	0.71	0.88
DegreeOf relation	0.93	0.94

- Best parameters
  - Linear kernel
  - Downsampling rate: 0.5
- Best features
  - Entity features
  - Word features

Upcoming

- Events
- Temporal Expression and their normalization
- Viz tool
- Question-answering (way in the future)

Applications in Biomedicine

- Translational science and clinical investigation
  - Patient cohort identification
  - Phenotype extraction
  - Linking patient's phenotype and genotype
  - eMERGE, PGRN, i2b2, SHARP
- Meaningful use of the EMR
- Comparative effectiveness
- Epidemiology
- Clinical practice
- …..

Processing Clinical Notes

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed. Since then, self-monitoring of blood glucose (SMBG) showed blood glucose levels of 250-270 mg/dL. She was referred to an endocrinologist for further evaluation.

On examination, she was normotensive and not acutely ill. Her
body mass index (BMI) was 18.7 kg/m2 follo
weight loss. Her thyroid was symmetrically
reflexes absent. Her blood glucose was 272 m
hemoglobin A1c (HbA1c) was 10.3%. A lipid p
cholesterol of 261 mg/dL, triglyceride level of 321 mg/
level of 48 mg/dL, and an LDL of 150 mg/dL. Thyroid
was normal. Urinanalysis showed trace ketones.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years, and limited her alcohol intake to 1 drink daily. Her mother's brother was diabetic.

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed. Since then, self-monitoring of blood glucose (SMBG) showed blood glucose levels of 250-270 mg/dL. She was referred to an endocrinologist for further evaluation.

On examination, she was normotensive and not acutely ill. Her body mass index (BMI) was 18.7 kg/m2 following a recent 10 lb weight loss. Her thyroid was symmetrically enlarged and ankle reflexes absent. Her blood glucose was 272 mg/dL, and her hemoglobin A1c (HbA1c) was 10.3%. A lipid profile showed a total cholesterol of 261 mg/dL, triglyceride level of 321 mg/dL, HDL level of 48 mg/dL, and an LDL of 150 mg/dL. Thyroid function was normal. Urinanalysis showed trace ketones.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years, and limited her alcohol intake to 1 drink daily. Her mother's brother was diabetic.

Clinical Element Model

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	patient
relative temporal context:	3 months ago
negation indicator:	not negated

Medication CEM	
text:	Glyburide
code:	315989
subject:	patient
frequency:	once daily
negation indicator:	not negated
strength:	2.5 mg

Tobacco Use CEM	
text:	smoking
code:	365981007
subject:	patient
relative temporal context:	25 years
negation indicator:	not negated

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	family member
relative temporal context:	
negation indicator:	not negated

A 43 year old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dl. Glyburide 2.5 mg once daily was prescribed.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years

Her mother's brother was diabetic.

Comparative Effectiveness

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	patient
relative temporal context:	3 months ago
negation indicator:	not negated

Medication CEM	
text:	Glyburide
code:	315989
subject:	patient
frequency:	once daily
negation indicator:	not negated
strength:	2.5 mg

Tobacco Use CEM	
text:	smoking
code:	365981007
subject:	patient
relative temporal context:	25 years
negation indicator:	not negated

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	family member
relative temporal context:	
negation indicator:	not negated

Compare the effectiveness of different treatment strategies (e.g., modifying target levels for glucose, lipid, or blood pressure) in reducing cardiovascular complications in newly diagnosed adolescents and adults with type 2 diabetes.

Compare the effectiveness of traditional behavioral interventions versus economic incentives in motivating behavior changes (e.g., weight loss, smoking cessation, avoiding alcohol and substance abuse) in children and adults.

Meaningful Use

Disorder CEM

text:	diabetes mellitus
code:	73211009
subject:	patient
relative temporal context:	3 months ago
negation indicator:	not negated

Medication CEM

text:	Glyburide
code:	315989
subject:	patient
frequency:	once daily
negation indicator:	not negated
strength:	2.5 mg

Tobacco Use CEM

text:	smoking
code:	365981007
subject:	patient
relative temporal context:	25 years
negation indicator:	not negated

Disorder CEM

text:	diabetes mellitus
code:	73211009
subject:	family member
relative temporal context:	
negation indicator:	not negated

- Maintain problem list
- Maintain active med list
- Record smoking status
- Provide clinical summaries for each office visit
- Generate patient lists for specific conditions
- Submit syndromic surveillance data

Clinical Practice

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	patient
relative temporal context:	3 months ago
negation indicator:	not negated

Medication CEM	
text:	Glyburide
code:	315989
subject:	patient
frequency:	once daily
negation indicator:	not negated
strength:	2.5 mg

- Provide problem list and meds from the visit

Example: Cohort Identification

- > 30MM records
- UIMA-AS
  - Scale out entire pipeline
  - Large Batch Processing
  - Dedicated Cluster(s) running LSF
    - > 96 concurrent pipelines
  - Custom start/stop scripts
- Future: UIMA-DUCC

Apache cTAKES Parallel Processing

- Background:
  - UIMA (2006)
  - UIMA-AS (2008)
  - Dedicated Cluster vs Grid Computing
- Future:
  - UIMA-DUCC (2013)
    (Distributed UIMA Cluster Computing)

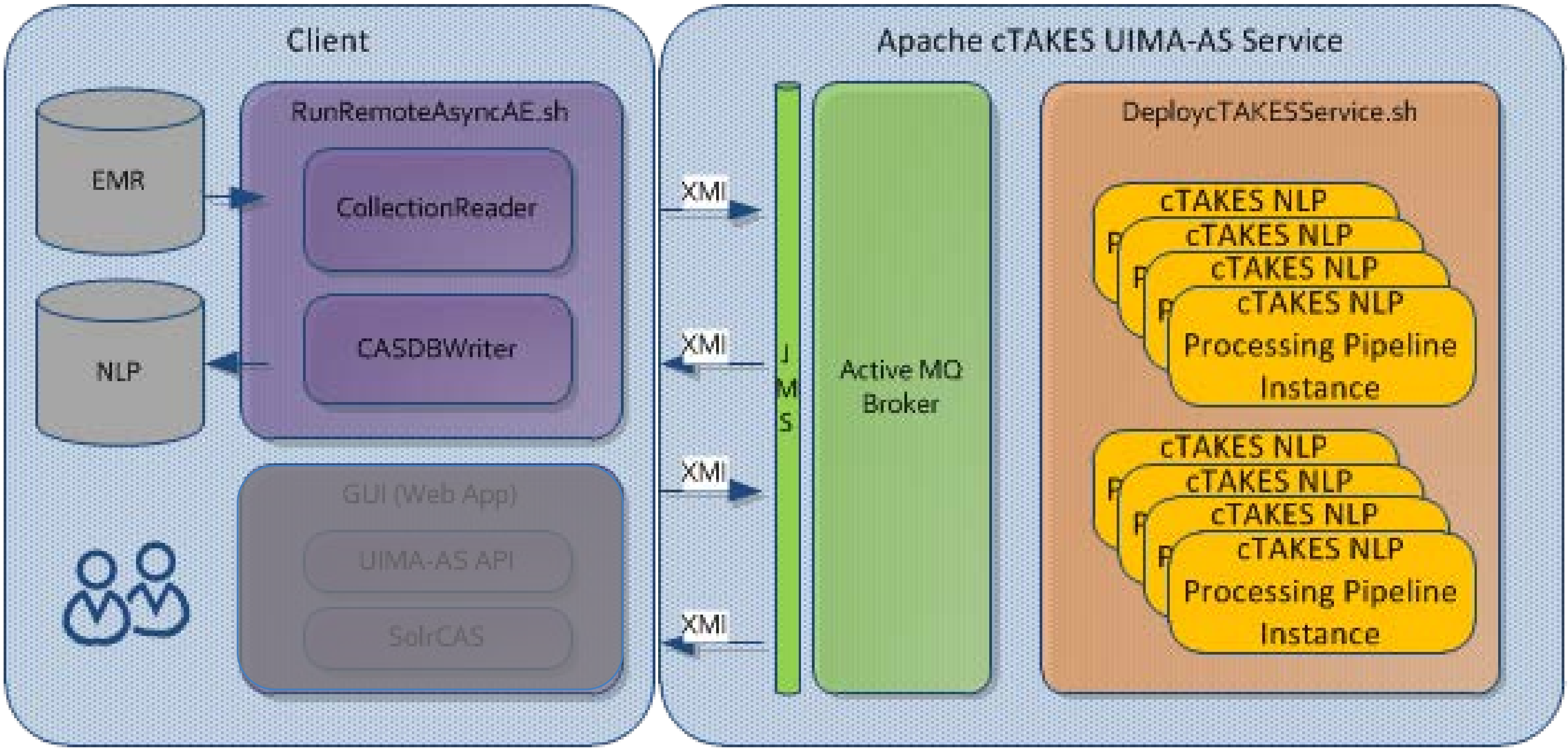
What is UIMA (you – eee –muh)?

- Unstructured Information Management Architecture
- Open source scaleable and extensible platform
- Create, integrate and deploy unstructured information management solutions
- Many Open Source projects based on UIMA

Why UIMA?

- Interoperability – Many developers adopting UIMA
  - Easy to share and re-use resources
- Precisely controlled work flow
- Good scalability abilities
- Easy to utilize modules created by 3rd party developers
- Ongoing active development on new resources

Apache cTAKES UIMA-AS



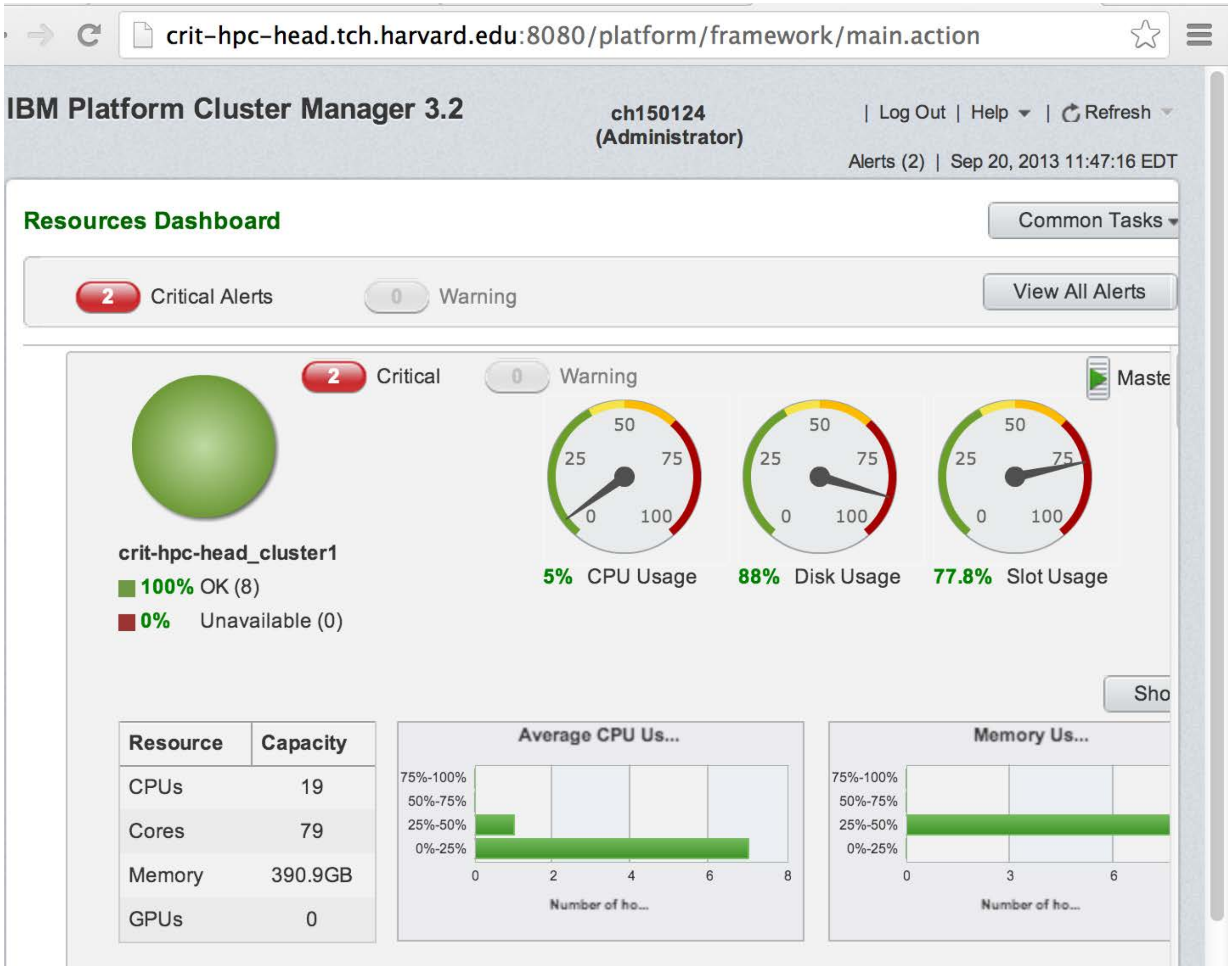
Apache cTAKES Pipeline Deploy

- Define Pipeline (AggregatePlaintextUMLSProcessor.xml)
  - Collection Reader (CR)
  - Analysis Engine(s) (AE)
  - Cas Consumer (CC)
- Define Deploy Descriptor (DeployAggregatePlaintextUMLStoDb.xml)
  - BrokerURL
  - Input/Output Queue
- Start MQ Broker
- Deploy!

UIMA-AS Cluster Helper Scripts

```
# set directories
export ROOT_DIR=/shared/chip_nlp
export UIMA_HOME=$ROOT_DIR/app/uima-as/v2.4.0
export ACTIVEMQ_HOME=$ROOT_DIR/app/activemq/apache-activemq-5.8.0
export CTAKES_AS_HOME=$ROOT_DIR/app/ctakes-as
export CTAKES_HOME=$ROOT_DIR/app/ctakes/apache-ctakes-3.1.0-incubating-SNAPSHOT

export ACTIVEMQ_OPTS_MEMORY="-Xms1G -Xmx2G"

# set ctakes-as defaults to be overridden per user
export CTAKES_AS_QUEUE=org.apache.ctakes.service.pipeline.input.queue
export CTAKES_AS_PORT=51515
export CTAKES_AS_WORK_DIR=~

# Memory allocation for each Pipeline
export UIMA_JVM_OPTS="-Xms256M -Xmx2G"
# set up ctakes-as cluster configuration
export CTAKES_BROKER_NODE=compute001
export CTAKES_WORKER_NODES="compute001 compute002 compute003 compute004"
export CTAKES_WORKER_PIPES="1 2 3 4 5 6 7 8 9 10 11 12"
export CTAKES_READER_NODE=compute001
# Use a pool with twice as many CAS Objects as there are Pipelines
export CAS_COUNT=96

# attempt setup of user-specific environment
. $CTAKES_AS_HOME/setUserEnv.sh

# final ctakes-as setup is based upon values either default or user
export CTAKES_BROKER_URL=tcp://$CTAKES_BROKER_NODE:$CTAKES_AS_PORT
export CTAKES_AS_LOG_DIR=$CTAKES_AS_WORK_DIR/log
export ACTIVEMQ_BASE=$CTAKES_AS_WORK_DIR/activemq

# Each user should have a ctakes-as/ directory and setProjectEnv.sh
. $CTAKES_AS_WORK_DIR/setProjectEnv.sh
[ch150124@crit-hpc-head ctakes-as]$
```

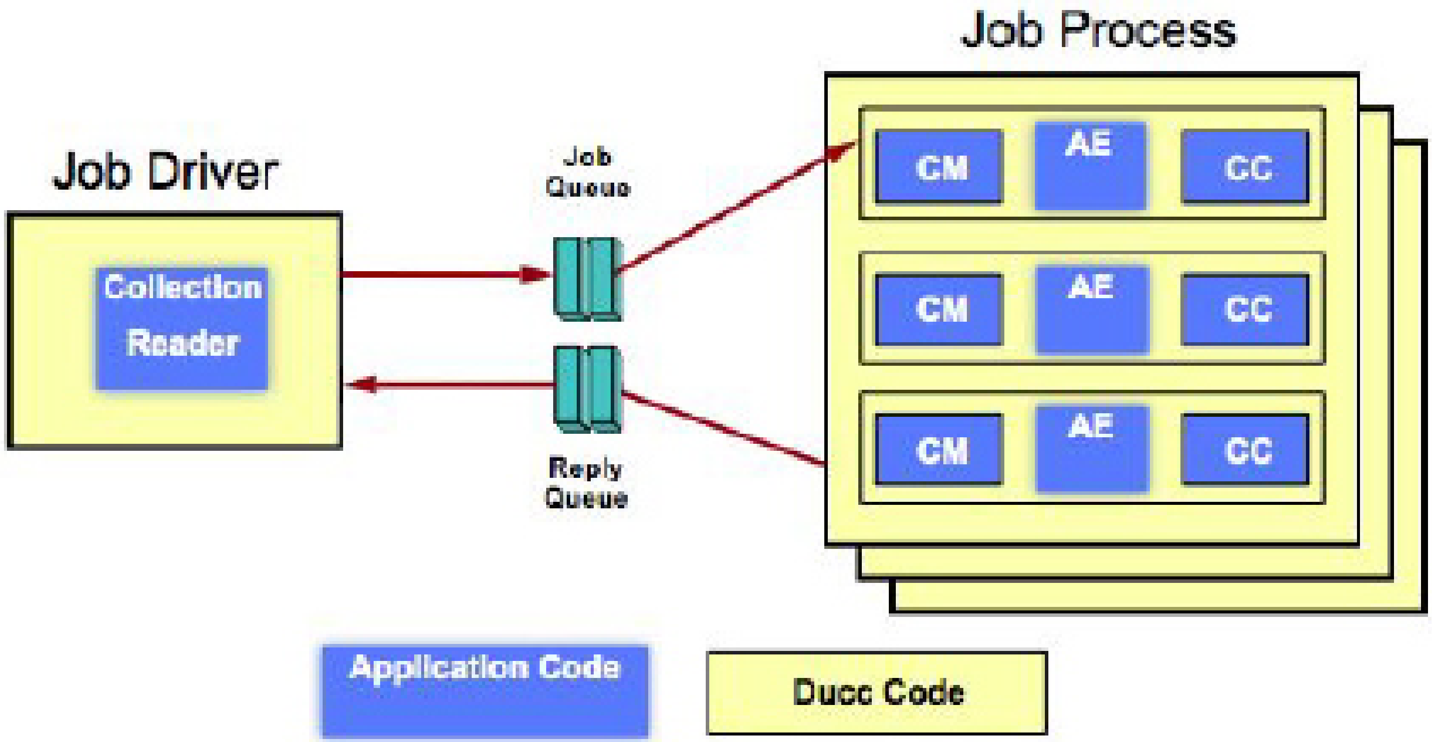
Dedicated Cluster(s) running LSF



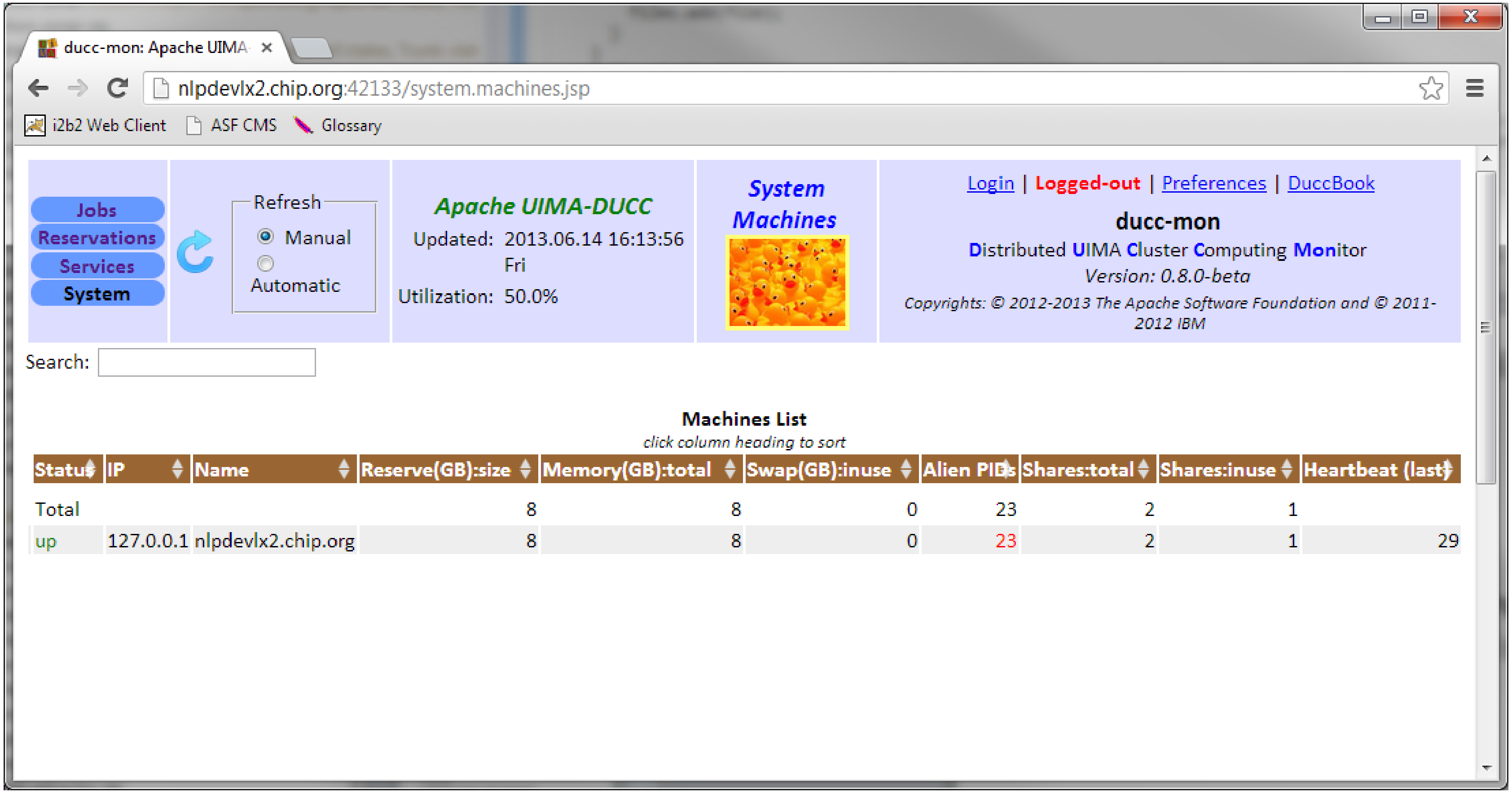
Error Handling

- & Recovery

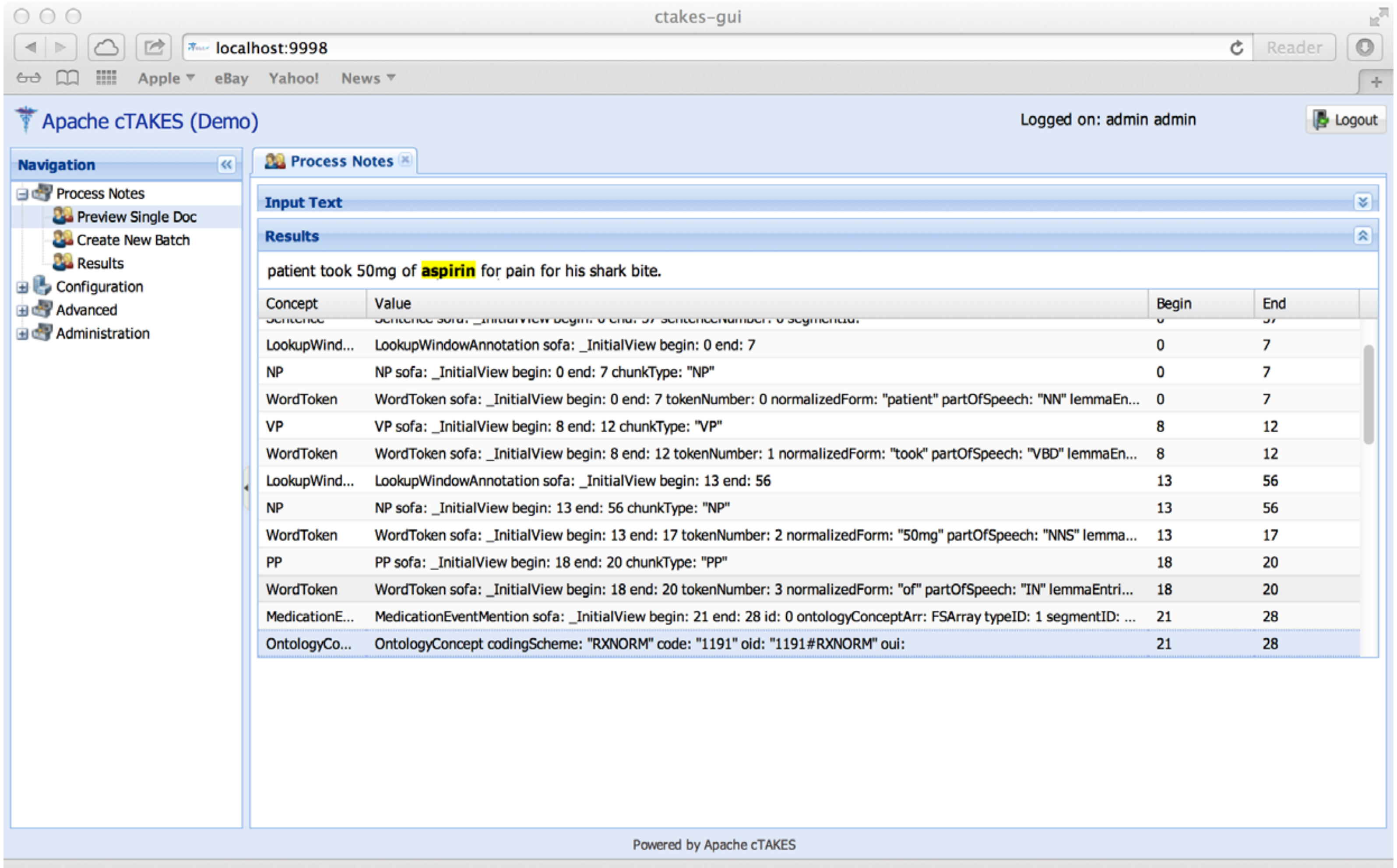
Future: UIMA-DUCC



Future: UIMA-DUCC

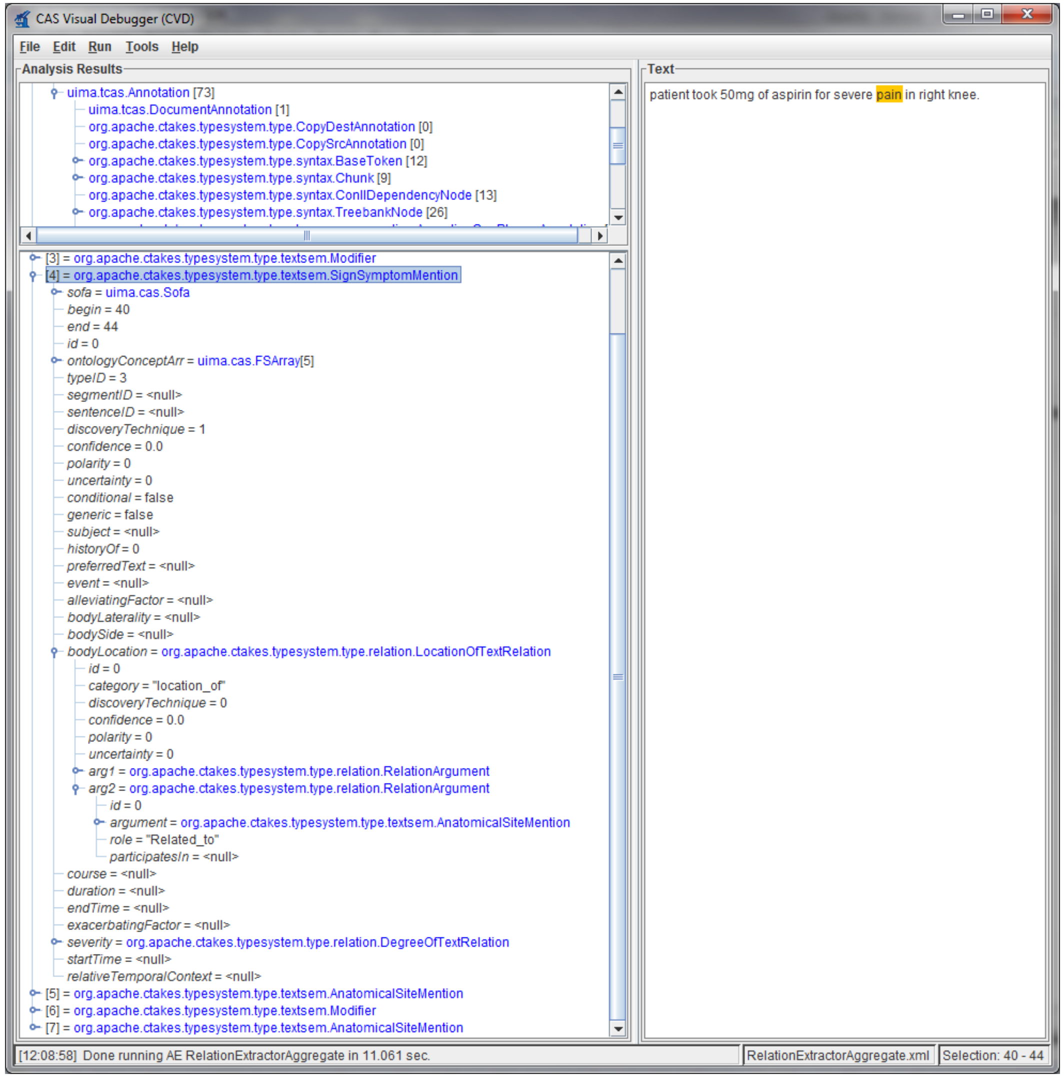


Demo



ctakes-gui
localhost:9998
Apple eBay Yahoo! News
Apache cTAKES (Demo)
Logged on: admin admin
Logout
Navigation
Process Notes
Preview Single Doc
Create New Batch
Results
Configuration
Advanced
Administration
Process Notes
Input Text
Results
patient took 50mg of aspirin for pain for his shark bite.
Concept | Value | Begin | End
LookupWind... | LookupWindowAnnotation sofa: _InitialView begin: 0 end: 7 | 0 | 7
NP | NP sofa: _InitialView begin: 0 end: 7 chunkType: "NP" | 0 | 7
WordToken | WordToken sofa: _InitialView begin: 0 end: 7 tokenNumber: 0 normalizedForm: "patient" partOfSpeech: "NN" lemmaEn... | 0 | 7
VP | VP sofa: _InitialView begin: 8 end: 12 chunkType: "VP" | 8 | 12
WordToken | WordToken sofa: _InitialView begin: 8 end: 12 tokenNumber: 1 normalizedForm: "took" partOfSpeech: "VBD" lemmaEn... | 8 | 12
LookupWind... | LookupWindowAnnotation sofa: _InitialView begin: 13 end: 56 | 13 | 56
NP | NP sofa: _InitialView begin: 13 end: 56 chunkType: "NP" | 13 | 56
WordToken | WordToken sofa: _InitialView begin: 13 end: 17 tokenNumber: 2 normalizedForm: "50mg" partOfSpeech: "NNS" lemma... | 13 | 17
PP | PP sofa: _InitialView begin: 18 end: 20 chunkType: "PP" | 18 | 20
WordToken | WordToken sofa: _InitialView begin: 18 end: 20 tokenNumber: 3 normalizedForm: "of" partOfSpeech: "IN" lemmaEntri... | 18 | 20
MedicationE... | MedicationEventMention sofa: _InitialView begin: 21 end: 28 id: 0 ontologyConceptArr: FSArray typeID: 1 segmentID: ... | 21 | 28
OntologyCo... | OntologyConcept codingScheme: "RXNORM" code: "1191" oid: "1191#RXNORM" oui: | 21 | 28
Powered by Apache cTAKES

Demo


CAS Visual Debugger (CVD)
File Edit Run Tools Help
Analysis Results
uima.tcas.Annotation [73]
uima.tcas.DocumentAnnotation [1]
org.apache.ctakes.typesystem.type.CopyDestAnnotation [0]
org.apache.ctakes.typesystem.type.CopySrcAnnotation [0]
org.apache.ctakes.typesystem.type.syntax.BaseToken [12]
org.apache.ctakes.typesystem.type.syntax.Chunk [9]
org.apache.ctakes.typesystem.type.syntax.ConllDependencyNode [13]
org.apache.ctakes.typesystem.type.syntax.TreebankNode [26]
[3] = org.apache.ctakes.typesystem.type.textsem.Modifier
[4] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
sofa = uima.cas.Sofa
begin = 40
end = 44
id = 0
ontologyConceptArr = uima.cas.FSArray[5]
typeID = 3
segmentID = <null>
sentenceID = <null>
discoveryTechnique = 1
confidence = 0.0
polarity = 0
uncertainty = 0
conditional = false
generic = false
subject = <null>
historyOf = 0
preferredText = <null>
event = <null>
alleviatingFactor = <null>
bodyLaterality = <null>
bodySide = <null>
bodyLocation = org.apache.ctakes.typesystem.type.relation.LocationOfTextRelation
id = 0
category = "location_of"
discoveryTechnique = 0
confidence = 0.0
polarity = 0
uncertainty = 0
arg1 = org.apache.ctakes.typesystem.type.relation.RelationArgument
arg2 = org.apache.ctakes.typesystem.type.relation.RelationArgument
id = 0
argument = org.apache.ctakes.typesystem.type.textsem.AnatomicalSiteMention
role = "Related_to"
participatesIn = <null>
course = <null>
duration = <null>
endTime = <null>
exacerbatingFactor = <null>
severity = org.apache.ctakes.typesystem.type.relation.DegreeOfTextRelation
startTime = <null>
relativeTemporalContext = <null>
[5] = org.apache.ctakes.typesystem.type.textsem.AnatomicalSiteMention
[6] = org.apache.ctakes.typesystem.type.textsem.Modifier
[7] = org.apache.ctakes.typesystem.type.textsem.AnatomicalSiteMention
Text
patient took 50mg of aspirin for severe pain in right knee.
[12:08:58] Done running AE RelationExtractorAggregate in 11.061 sec.
RelationExtractorAggregate.xml
Selection: 40 - 44

END

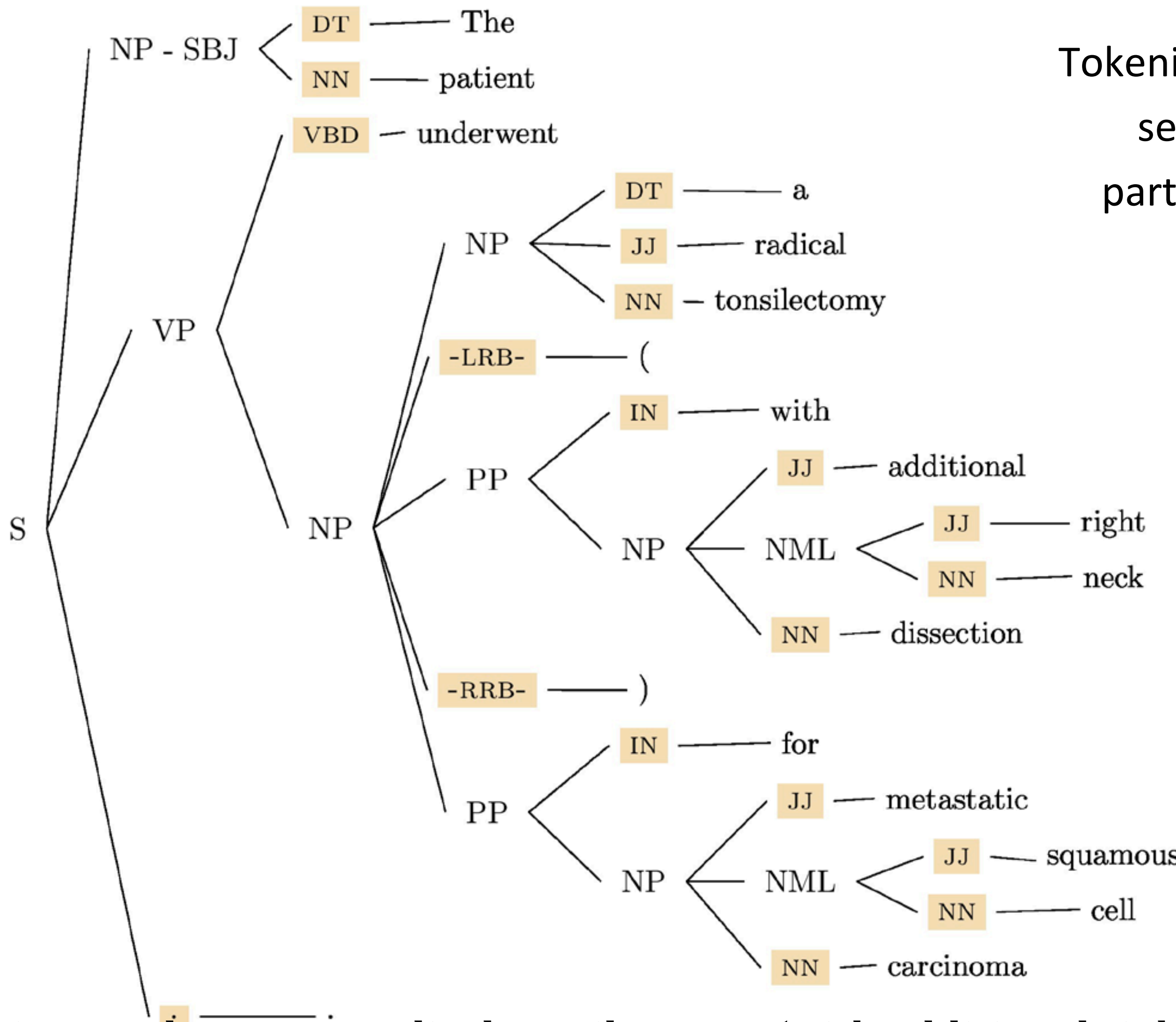
Treebank Annotations

Treebank Annotations

- Consist of part-of-speech tags, phrasal and function tags, and empty categories organized in a tree-like structure
- Adapted Penn's POS tagging guidelines, bracketing guidelines, and associated addenda
- Extended the guidelines to account for domain-specific characteristics

http://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf

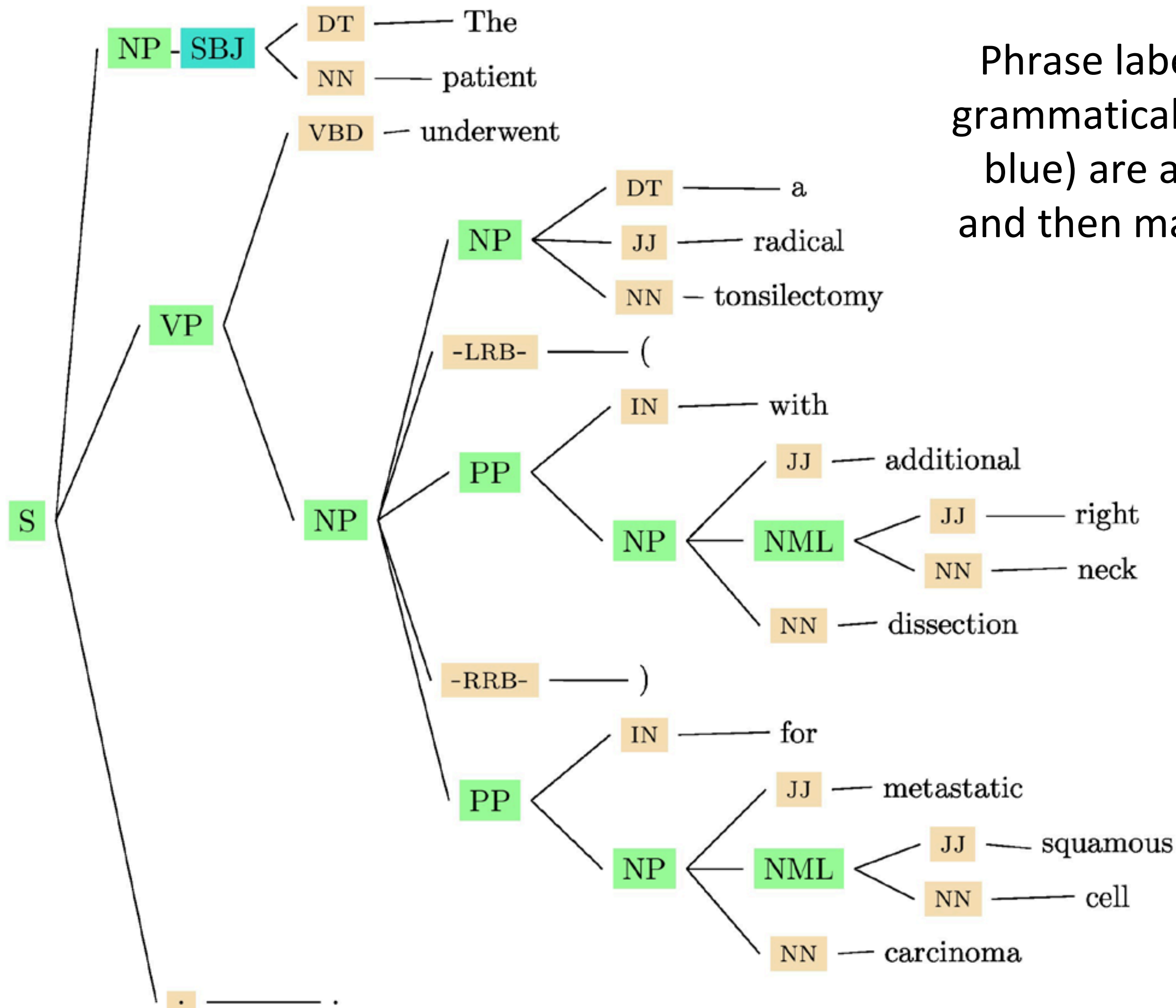
Treebank Review



Tokenization, sentence segmentation, and part of speech labels (in brown) are all done in an initial pass.

The patient underwent a radical tonsilectomy (with additional right neck dissection) for metastatic squamous cell carcinoma .

Treebank Review



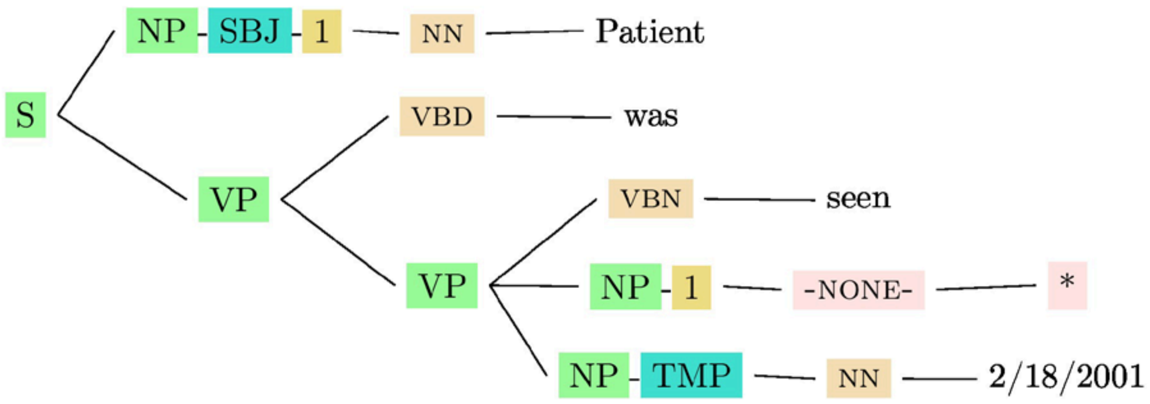
Phrase labels (in green) and grammatical function tags (in blue) are added by a parser and then manually corrected

The patient underwent a radical tonsilectomy (with additional right neck dissection) for metastatic squamous cell carcinoma .

Treebank Review

In that second pass, new tokens are added for implicit and empty arguments (in red), and grammatically linked elements are indexed (in yellow)

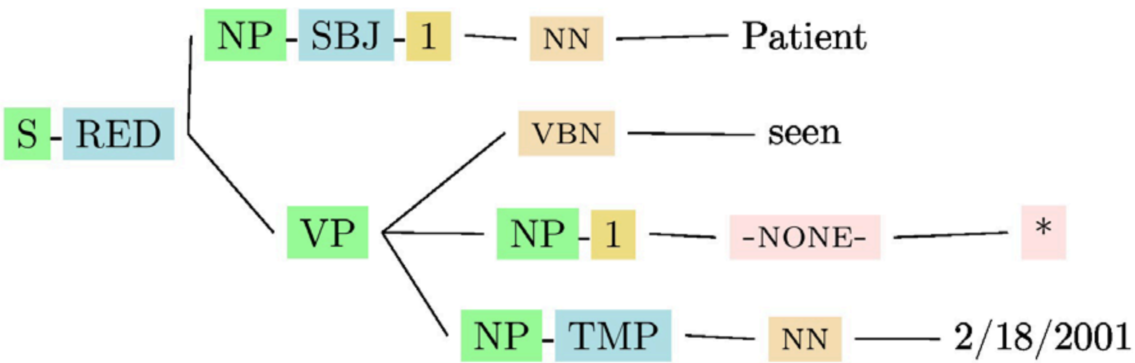
Patient was seen 2/18/2001



Clinical Additions – S-RED

Clinical language is highly reduced, and often elides copula ("to be").
-RED tag was introduced to mark clauses with elided copulas.

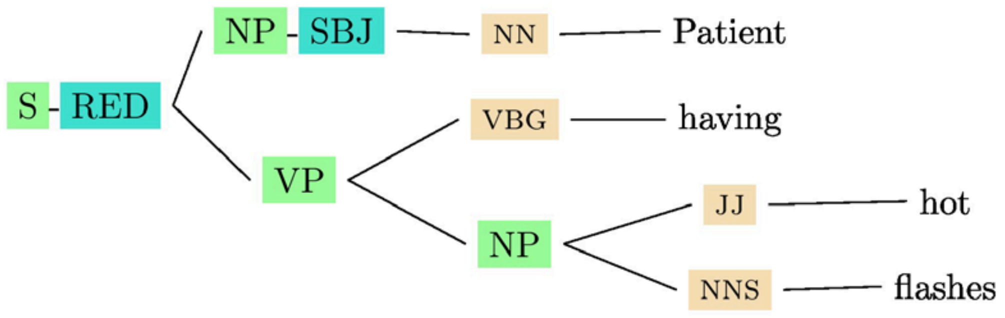
Patient (was) seen 2/18/2001



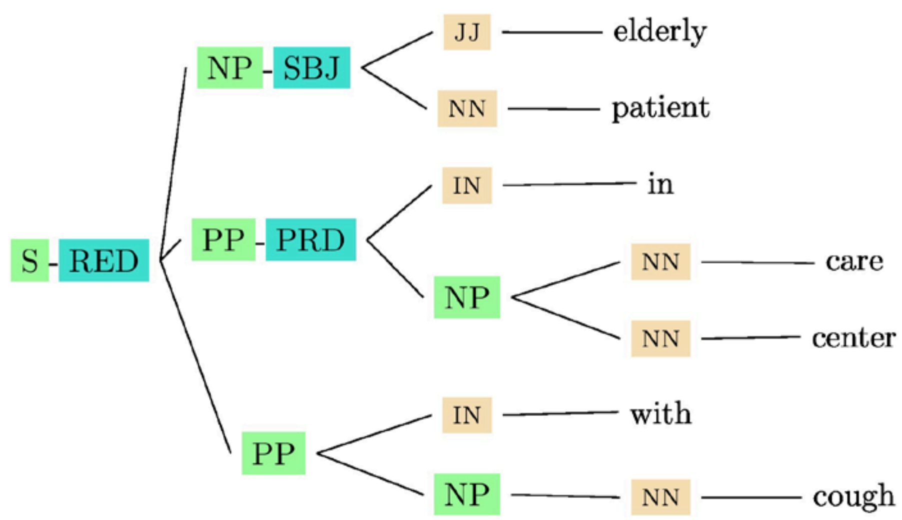
Clinical Additions – S-RED

-RED tags are used for all elisions of the copula, including passive voice, progressive (top example) and equational clauses (bottom example).

Patient (is) having hot flashes

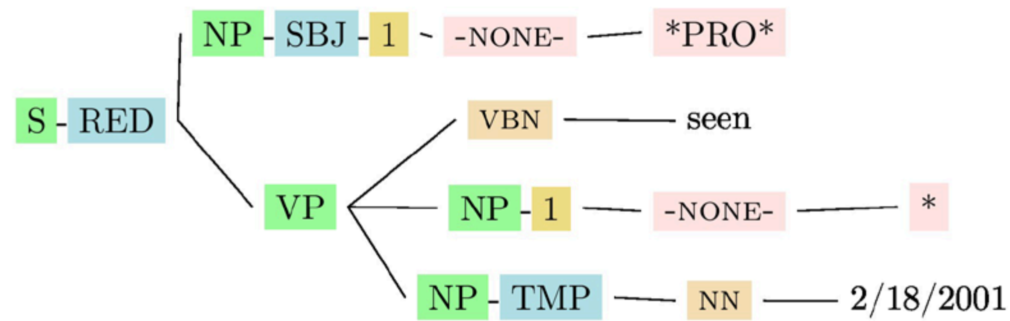


Elderly patient (is) in care center with cough

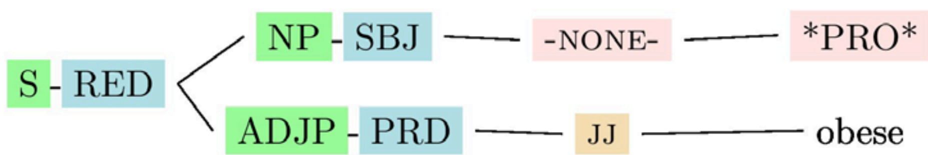


Clinical Additions – Null Arguments

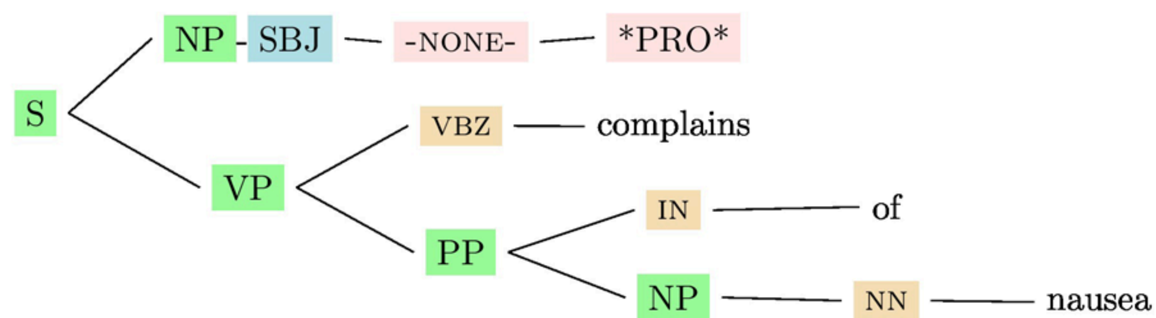
Dropped subjects are very common in this data, and *PRO* tags are added to represent them.



(*PRO*) (was) Seen 2/18/2001

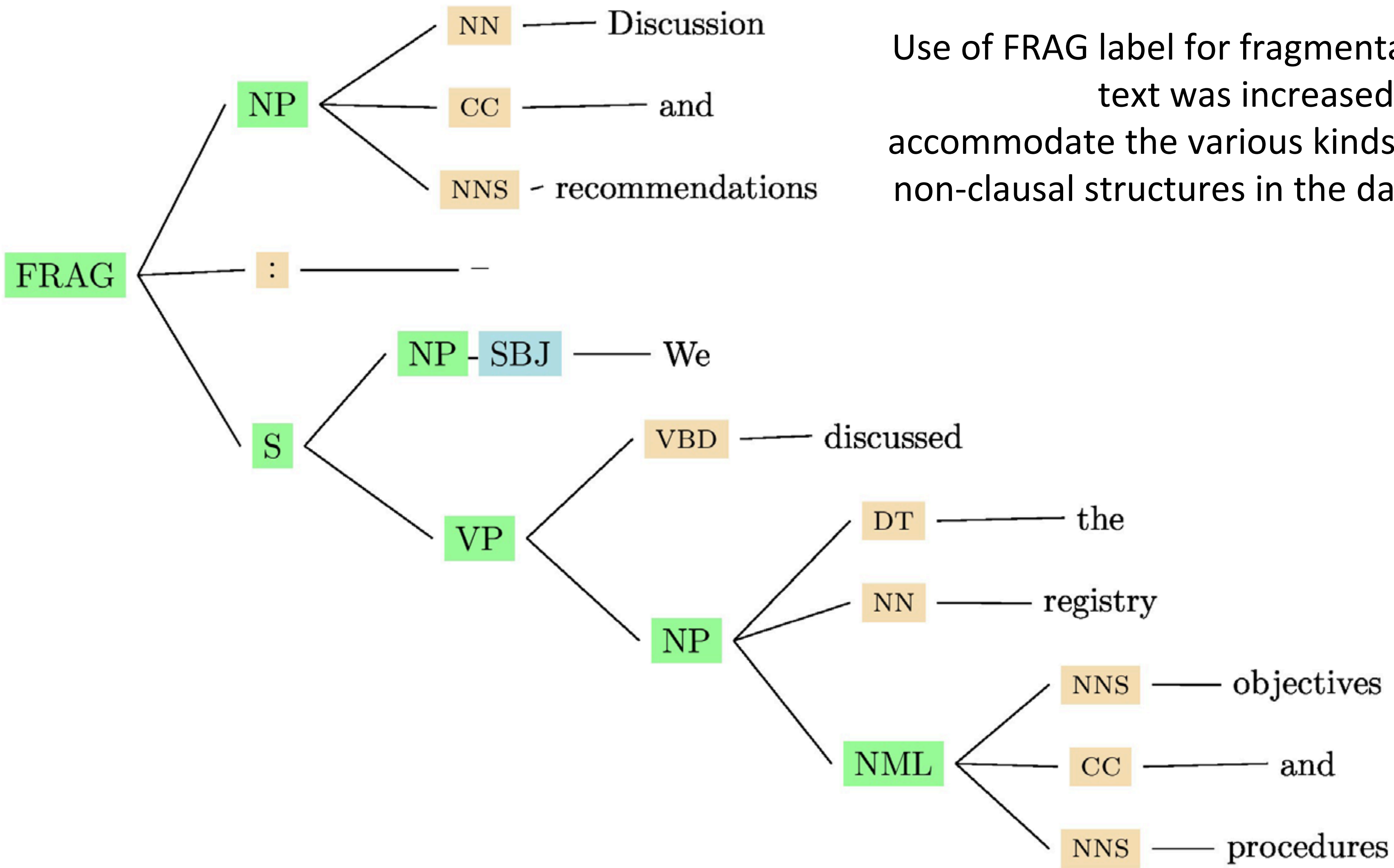


(*PRO*) (is) Obese



(*PRO*) Complains of nausea

Clinical Additions – FRAG



Use of FRAG label for fragmentary text was increased to accommodate the various kinds of non-clausal structures in the data.

Discussion and recommendations: We discussed the registry objectives and procedures.

Propbank Annotations

What is Propbank?

- *who did what to whom when where and how*
- A database of syntactically parsed trees annotated with semantic role labels
- All arguments are annotated with semantic roles in relation to their predicate structure
- This provides training data that can identify predicate-argument structures for individual verbs.

Propbank Labels

- Labels do not change with predicate
- Meanings of core arguments 2-5 change with predicate
- Arg0 proto-agent for transitive verbs
- Arg1 proto-patient for transitive verbs
- Meanings of Adjunctive args do not change
- http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf

Propbank Labels

- **Arg0 = agent**
- **Arg1 = theme / patient**
- **Arg2 = benefactive / instrument/ attribute / end state**
- **Arg3 = start point / benefactive / attribute**
- **Arg4 = end point**
- **ArgM = modifier**

Propbank Labels

ARG0(agent)	Adverbial	Manner
ARG1(patient)	Cause	Modal
ARG2	Direction	Negation
ARG3	Discourse	Purpose
ARG4	Extent	Temporal
	Location	Predication

Why Propbank?

- Identifying a commonalities in predicate-argument structures:

Agent diagnosing

[Dr.Z] diagnosed [Jack's bronchitis]

Person diagnosed

Disease

[Jack] was diagnosed [with bronchitis] [by Dr.Z]

[Dr. Z's] diagnosis [of Jack's bronchitis] allowed her to treat him with the proper antibiotics.

Stages of the Propank process

- Frame Creation

Predicate: ***undergo***

undergo: Frames file for 'undergo' based on sentences in financial subcorpus.

Roleset id: undergo.01 , *experience, undergo*, vncls: , framnet: Undergoing

undergo.01: No Vncls. Comparison with 'go through'.

Roles:

Arg1: *experiencer*
Arg2: *experienced*

Example: pretty typical

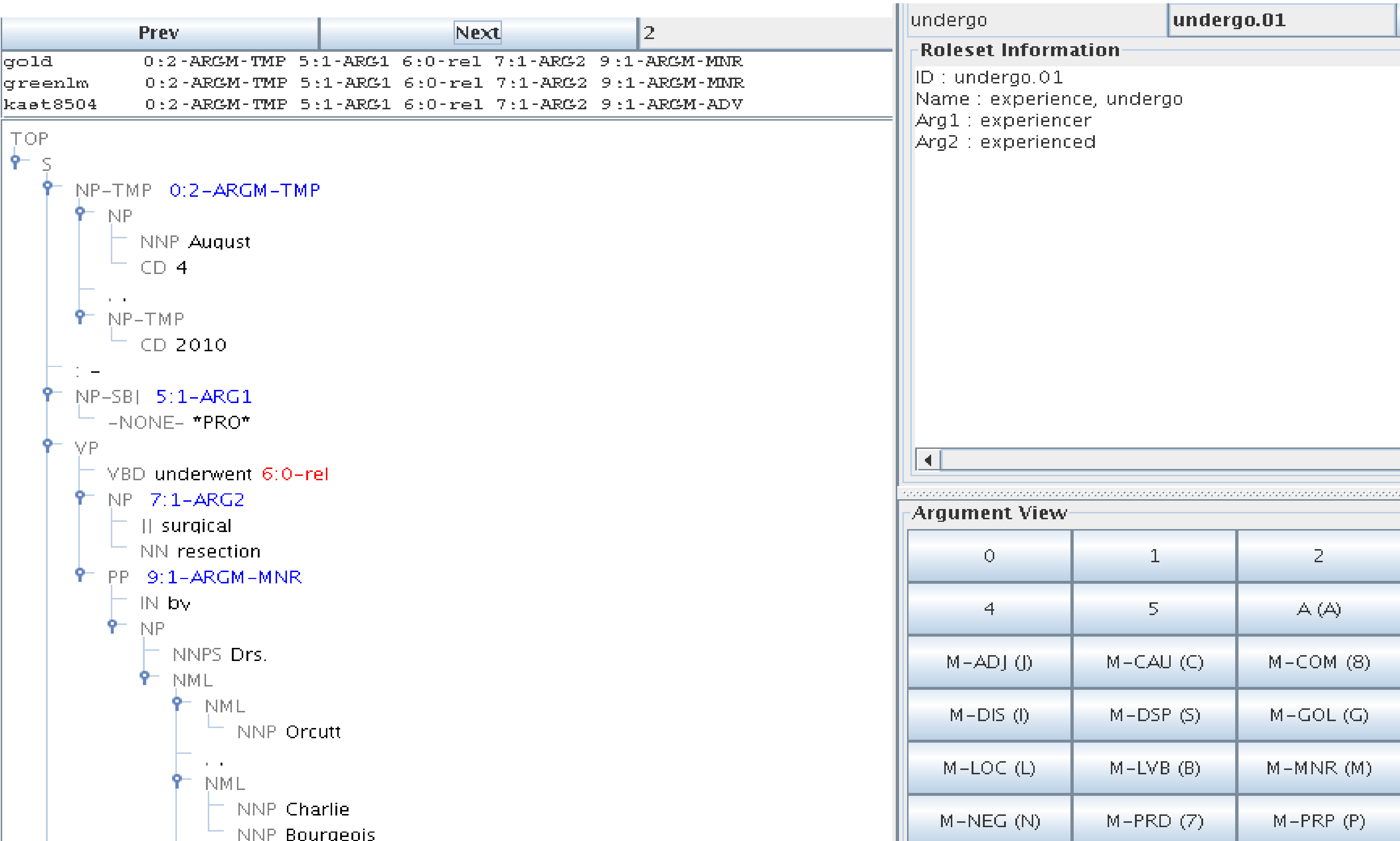
Periods before the advent of futures or program trading were often more volatile, usually when fundamental market conditions were undergoing change (1973-75, 1937-40, and 1928-33 for example).

Arg1: fundamental market conditions
Rel: undergoing
Arg2: change

Stages of Propbank

- Annotation
  - Data is double annotated
  - Annotators
    1. Determine and select the sense of the predicate
    2. Annotate the arguments for the selected predicate sense
- Adjudication
  - After data is annotated, it is passed to an adjudicator who resolves differences between the two annotators
  - This creates the *gold standard* – corrected, finished training data

Annotation Example



OPEN ACCESS

JAMIA, 2013

Towards comprehensive syntactic and semantic annotations of the clinical narrative

Daniel Albright,[1] Arrick Lanfranchi,[1] Anwen Fredriksen,[1] William F Styler IV,[1] Colin Warner,[2] Jena D Hwang,[1] Jinho D Choi,[3] Dmitriy Dligach,[4] Rodney D Nielsen,[1,5] James Martin,[3] Wayne Ward,[3] Martha Palmer,[1] Guergana K Savova[4]

[1]Department of Linguistics, University of Colorado, Boulder, Colorado, USA
[2]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Department of Computer Science University of Colorado, Boulder, Colorado, USA
[4]Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[5]Department of Computer Science and Engineering, University of North Texas, Texas, USA

Correspondence to
Dr Guergana K Savova, Boston Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02114, USA; Guergana.Savova@childrens.harvard.edu

Received 3 September 2012
Revised 27 December 2012
Accepted 28 December 2012

ABSTRACT

Objective To create annotated clinical narratives with layers of syntactic and semantic labels to facilitate advances in clinical natural language processing (NLP). To develop NLP algorithms and open source components.

Methods Manual annotation of a clinical narrative corpus of 127 606 tokens following the Treebank schema for syntactic information, PropBank schema for predicate-argument structures, and the Unified Medical Language System (UMLS) schema for semantic information. NLP components were developed.

Results The final corpus consists of 13 091 sentences containing 1772 distinct predicate lemmas. Of the 766 newly created PropBank frames, 74 are verbs. There are 28 539 named entity (NE) annotations spread over 15 UMLS semantic groups, one UMLS semantic type, and the Person semantic category. The most frequent annotations belong to the UMLS semantic groups of Procedures (15.71%), Disorders (14.74%), Concepts and Ideas (15.10%), Anatomy (12.80%), Chemicals and Drugs (7.49%), and the UMLS semantic type of Sign or Symptom (12.46%). Inter-annotator agreement results: Treebank (0.926), PropBank (0.891–0.931), NE (0.697–0.750). The part-of-speech tagger, constituency parser, dependency parser, and semantic role labeler are built from the corpus and released open source. A significant limitation uncovered by this project is the need for the NLP community to develop a widely agreed-upon schema for the annotation of clinical concepts and their relations.

Conclusions This project takes a foundational step towards bringing the field of clinical NLP up to par with NLP in the general domain. The corpus creation and NLP components provide a resource for research and application development that would have been previously impossible.

other), the level of certainty associated with an event (confirmed, possible, negated) as well as textual mentions that point to the same event. We describe our efforts to combine annotation types developed for general domain syntactic and semantic parsing with medical-domain-specific annotations to create annotated documents accessible to a variety of methods of analysis including algorithm and component development. We evaluate the quality of our annotations by training supervised systems to perform the same annotations automatically. Our effort focuses on developing principled and generalizable enabling computational technologies and addresses the urgent need for annotated clinical narratives necessary to improve the accuracy of tools for extracting comprehensive clinical information.[1] These tools can in turn be used in clinical decision support systems, clinical research combining phenotype and genotype data, quality control, comparative effectiveness, and medication reconciliation to name a few biomedical applications.

In the past decade, the general natural language processing (NLP) community has made enormous strides in solving difficult tasks, such as identifying the predicate-argument structure of a sentence and associated semantic roles, temporal relations, and coreference which enable the abstraction of the meaning from its surface textual form. These developments have been spurred by the targeted enrichment of general annotated resources (such as the Penn Treebank (PTB)[2]) with increasingly complex layers of annotations, each building upon the previous one, the most recent layer being the discourse level.[3] The emergence of other annotation standards (such as PropBank[4] for the annotation of the sentence predicate-argument structure) has brought new progress in the annotation of semantic informa-

Boston Children's Hospital

Harvard-MIT Division of Health Sciences and Technology

Select Publications on cTAKES Methods

- Dligach, Dmitriy; Bethard, Steven; Becker, Lee; Miller, Timothy; Savova, Guergana. (in press). Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association.
- Miller, Timothy; Bethard, Steven; Dligach, Dmitriy; Pradhan, Sameer; Lin, Chen; and Savova, Guergana. 2013. Discovering narrative containers in clinical text. BioNLP workshop at the Association for Computational Linguistics conference, August 3-9, Sofia, Bulgaria. http://aclweb.org/anthology/W/W13/W13-1903.pdf
- Albright, Daniel; Lanfranchi, Arrick; Fredriksen, Anwen; Styler, William; Warner, Collin; Hwang, Jena; Choi, Jinho; Dligach, Dmitriy; Nielsen, Rodney; Martin, James; Ward, Wayne; Palmer, Martha; Savova, Guergana. 2013. Towards syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association. 2013;0:1–9. doi:10.1136/amiajnl-2012-001317 http://jamia.bmj.com/cgi/rapidpdf/amiajnl-2012-001317?ijkey=z3pXhpyBzC7S1wC&keytype=ref
- Stephen T Wu, Vinod C Kaggal, Dmitriy Dligach, James J Masanz, Pei Chen, Lee Becker, Wendy W Chapman, Guergana K Savova, Hongfang Liu and Christopher G Chute. 2012. A common type system for clinical Natural Language Processing. Journal of Biomedical Semantics. MS ID: 1651620874755068
- Miller, Timothy; Dligach, Dmitriy; Savova, Guergana. 2012. Active learning for Coreference Resolution in the Biomedical Domain. BioNLP workshop at the Conference of the North American Association of Computational Linguistics (NAACL 2012). Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pp. 73-81.
- Zheng, Jiaping; Chapman, Wendy; Miller, Timothy; Lin, Chen; Crowley, Rebecca; Savova, Guergana. 2012. A system for coreference resolution for the clinical narrative. Journal of the American Medical Informatics Association. doi:10.1136/amiajnl-2011-000599
- Jinho D. Choi, Martha Palmer, “Getting the Most out of Transition-based Dependency Parsing”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 687-692, Portland, Oregon, 2011.
- Jinho D. Choi, Martha Palmer, “Transition-based Semantic Role Labeling Using Predicate Argument Clustering”, Proceedings of ACL workshop on Relational Models of Semantics (RELMS'11), 37-45, Portland, Oregon, 2011.
- Savova, Guergana; Masanz, James; Ogren, Philip; Zheng, Jiaping; Sohn, Sunghwan; Kipper-Schuler, Karin and Chute, Christopher. 2010. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications Journal of the American Medical Informatics Association 2010;17:507-513 doi:10.1136/jamia.2009.001560

Select Publications on cTAKES Applications

- Carrell, David; Halgrim, Scott; Tran, Diem-Thy; Buist, Diana SM; Chubak, Jessica; Chapman, Wendy; Savova, Guergana. In Press. Using Natural Language Processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. American Journal of Epidemiology.
- Chen, Lin; Karlson, Elizabeth; Canhao, Helena; Miller, Timothy; Dligach, Dmitriy; Chen, Pei; Guzman Perez, Raul; Cai, Tianxi; Weinblatt, Michael; Shadick, Nancy; Plenge, Robert; Savova, Guergana. 2013. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PlosOne. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0069932
- Ananthakrishnan, Ashwin; Cai, Tianxi; Cheng, Su-Chun; Chen, Pei; Savova, Guergana; Guzman Perez, Raul; Gainer, Vivian; Murphy, Shawn; Szolovits, Peter; Xia, Zongqi; Shaw, Stanley; Churchill, Susanne; Karlson, Elizabeth; Kohane, Isaak; Plenge, Robert; Liao, Katherine. 2012. Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: a Novel Informatics Approach. Journal of Inflammatory Bowel Diseases.
- Savova, Guergana; Olson, Janet; Murphy, Sean; Cafourek, Victoria; Couch, Fergus; Goetz, Matthew; Ingle, James; Suman, Vera; Chute, Christopher and Weinshilboum, Richard. 2011. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer. Journal of American Medical Informatics Association. 19:e83-e89 doi:10.1136/amiajnl-2011-000295
- Sohn, Sunghwan; Kocher, Jean-Pierre; Chute, Christopher; Savova, Guergana. 2011. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. Journal of American Medical Informatics Association. 2011 Dec;18 Suppl 1:i144-9. doi: 10.1136/amiajnl-2011-000351. Epub 2011 Sep 2.
- Cheng, Lionel; Zheng, Jiaping; Savova, Guergana and Erickson, Bradley. 2010. Discerning tumor status from unstructured MRI reports – completeness of information in existing reports and utility of natural language processing. Journal of Digital Imaging of the Society of Imaging Informatics in Medicine, ISSN: 0897-1889: 23(2), 119-133. PMID: 19484309. (Best paper 2010 award of the Journal of Digital Imaging). http://www.ncbi.nlm.nih.gov/pubmed/19484309