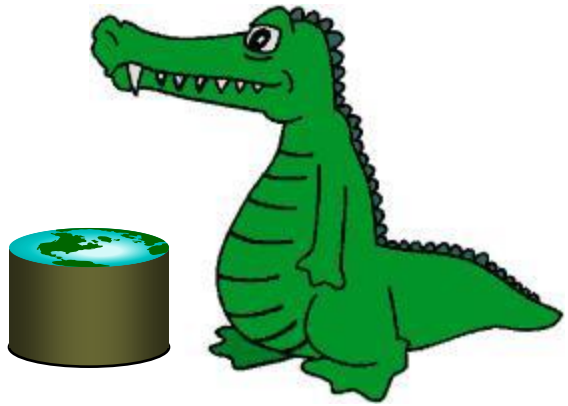
CAP4770/5771
Introduction to Data Science
Fall 2015

University of Florida, CISE Department
Prof. Daisy Zhe Wang



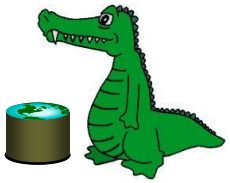
NLP and cTAKES for Biomedical Text Analysis

Biomedical Challenges

NLP techniques

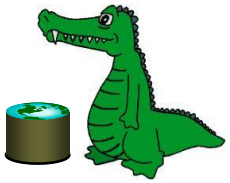
cTakes and applications

Based on slides from Dr. Guergana K. Savova from Boston Children's Hospital/Harvard Medical School and Pei Chen from Apache cTAKES



Outline

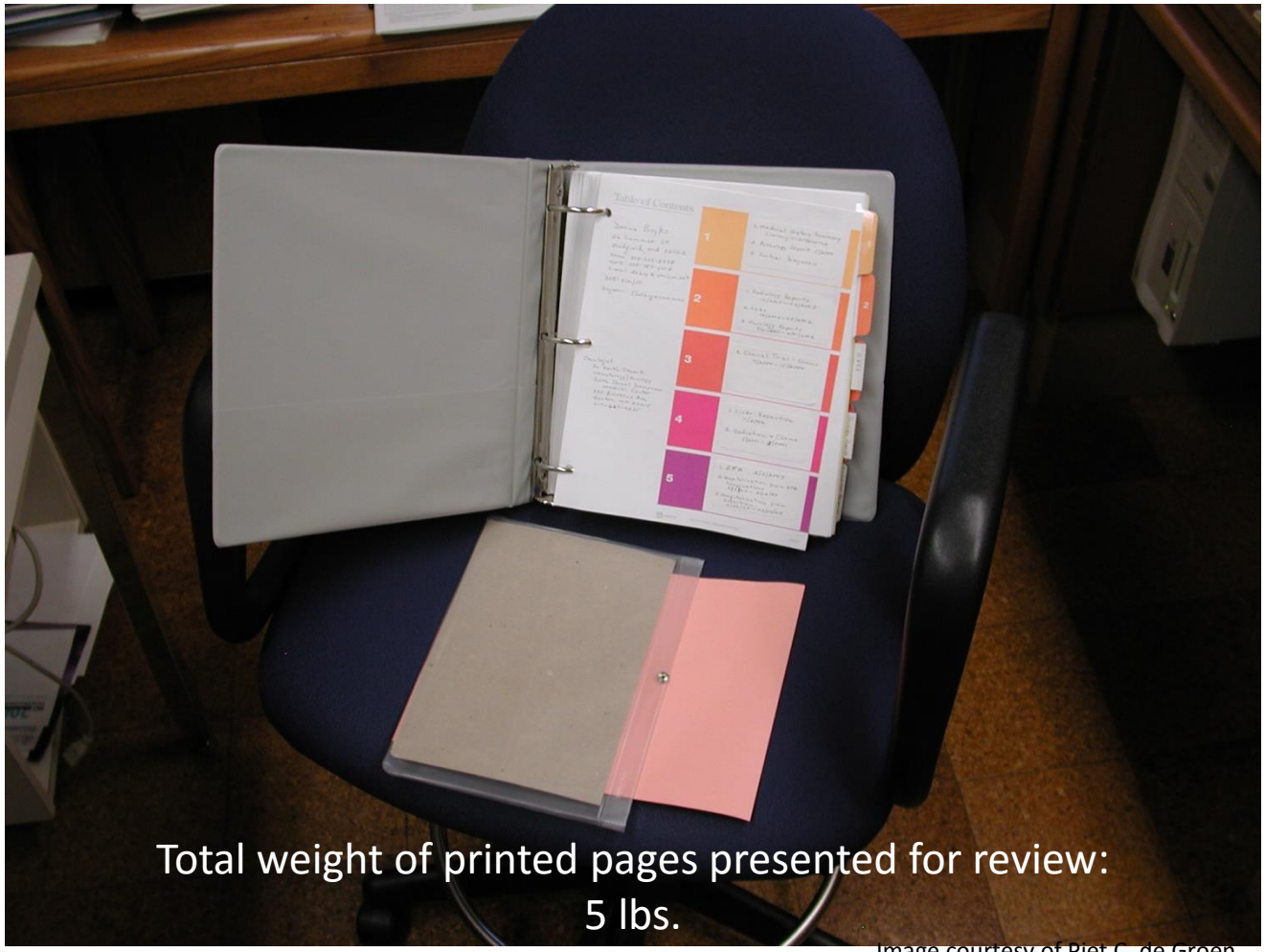
- Current Healthcare Challenges
- Computer Science concepts and techniques (e.g., IR, NLP, IE, DB)
- Apache cTAKES Overview and Applications
- Technical details



Outline

- Current Healthcare Challenges
- Computer Science concepts and techniques (e.g., IR, NLP, IE, DB)
- Apache cTAKES Overview and Applications
- Technical details

Patient January 16, 2006

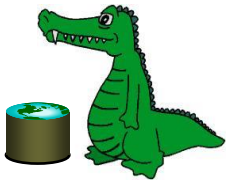


Total weight of printed pages presented for review:
5 lbs.

Image courtesy of Piet C. de Groen

Patient January 16, 2006

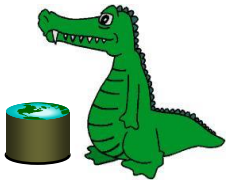
Image courtesy of Piet C. de Groen



Questions

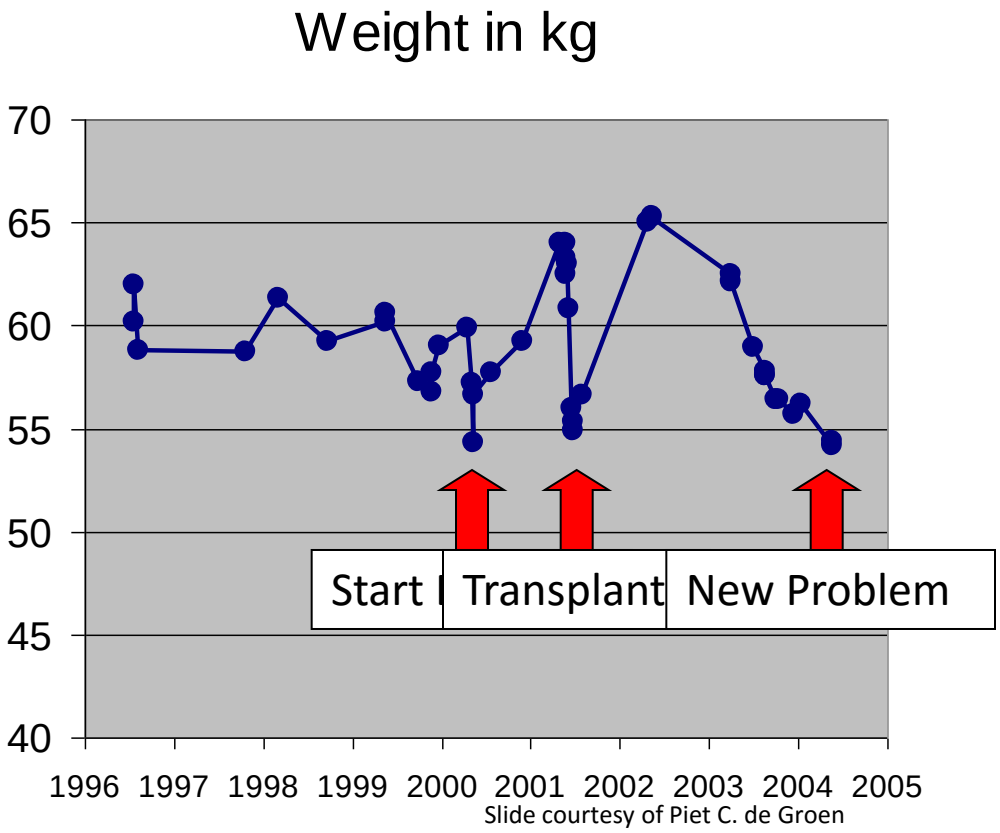
- What is exactly the patient's problem?
  - Are liver tests and weight loss due to Lipitor?
  - When did she use Lipitor?
  - What was the weight on what date?

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

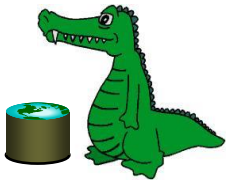


EHR/Data Warehouse to the rescue!

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

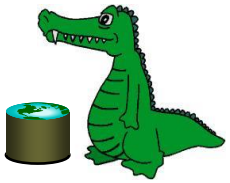


Slide courtesy of Piet C. de Groen



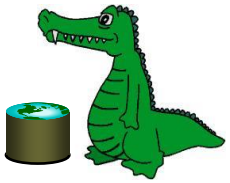
What happened to Cholesterol?

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



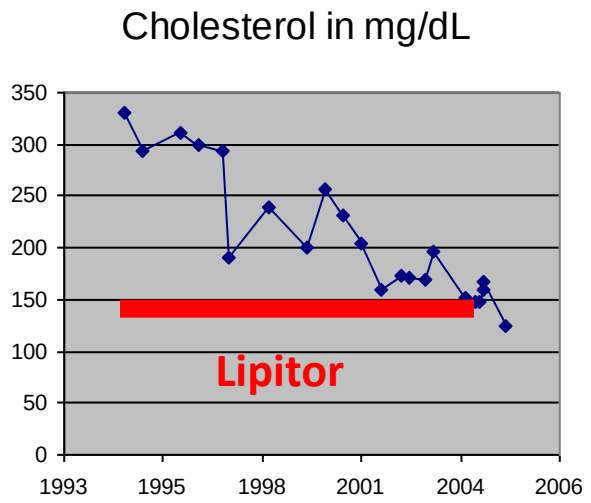
NLP to the rescue!

- 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

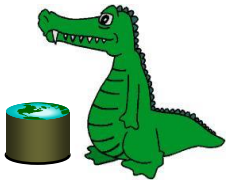


Complete Picture

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

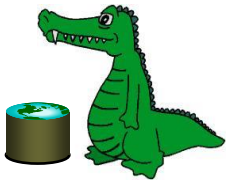


Slide courtesy of Piet C. de Groen



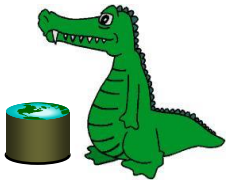
Outline

- Current Healthcare Challenges
- Computer Science concepts and techniques (e.g., IR, NLP, IE, DB)
- Apache cTAKES Overview and Applications
- Technical details



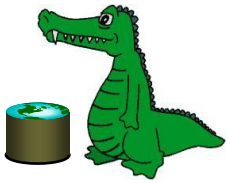
Definitions

- Information Extraction (IE)
  - Extracting existing facts from unstructured or loosely structured text into a structured form
- Information Retrieval (IR)
  - Finding documents relevant to a user query
- Named Entity Recognition (NER)
  - Discovery of groups of textual mentions that belong to certain semantic class
- Natural Language Processing (NLP)
  - Computational methods for text processing based on linguistically sound principles
  - Clinical NLP – NLP for the clinical narrative
  - Biomedical NLP – NLP for the clinical narrative and biomedical literature



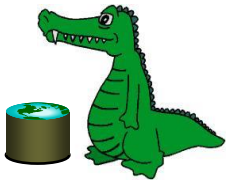
Problem Space

- Structured information
  - Relational databases
  - Easy to extract information from them
- Semi-structured information
  - Loosely formatted XML, CSV tables
  - Not challenging to extract information
- Unstructured information
  - Scholarly literature, clinical notes, research reports, webpages
  - Majority of information is unstructured!!
  - Real challenge to extract the information



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

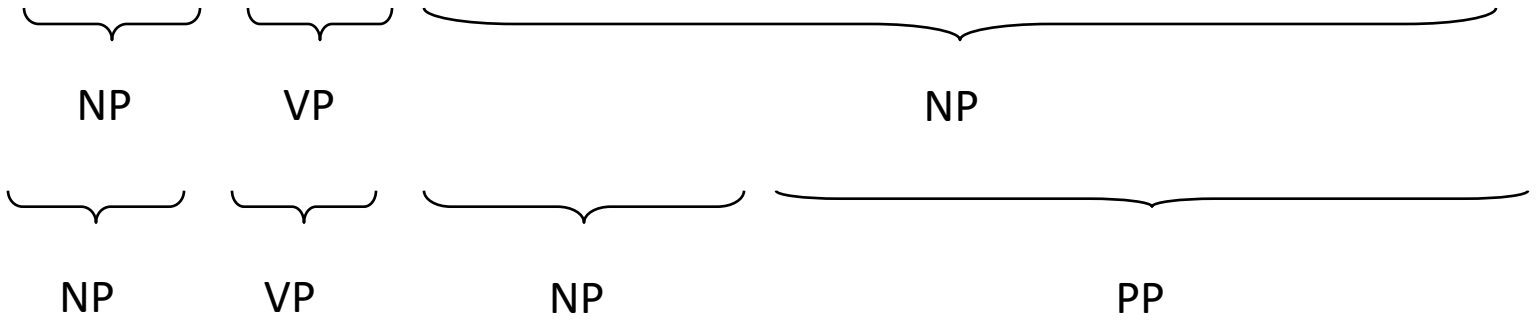


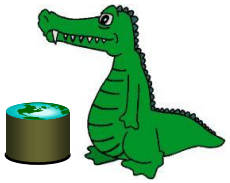
NLP Example

I saw the man with the telescope.

w1 w2 w3 w4 w5 w6 w7

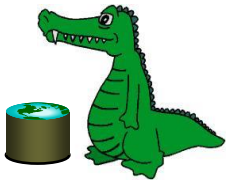
pronoun verb article noun prep article noun





NLP methods

- Rule-based
  - Regular Expression Pattern matching (e.g., "\b(lipitor|Lipitor)\b")
  - Dictionaries (e.g., drug names, ICD10 codes)
- Statistical Machine Learning
  - HMM: Hidden Markov Models
  - Linear-CRF: Linear-Chain Conditional Random Fields
  - Viterbi algorithms
- Hybrid



Why NLP? Understand semantics

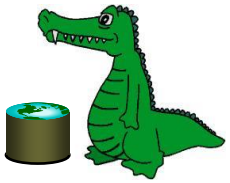
- From keyword search to language understanding
  - Negation (and any other similar context)

    The patient denies headache, earache, sore throat, fever, rash, hallucinations, stomachache, cough and any pneumonia-related symptoms
  - Inverted syntax

    Colon, ascending and descending, biopsy
  - Relation discovery

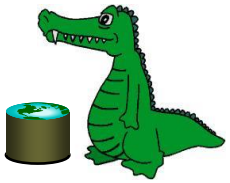
    Tamoxifen is used in the treatment of breast cancer.
  - Morphologic variations

    runs, running, ran, run -> mapped to the same base form
  - Higher level discourse phenomena: synonyms, anaphora relations, temporal relations, document summarization



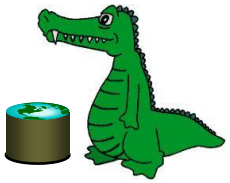
Outline

- Current Healthcare Challenges
- Computer Science concepts and techniques (e.g., IR, NLP, IE, DB)
- Apache cTAKES Overview and Applications
- Technical details



cTAKES Overview

- Release 1.0 developed at Mayo (~2010)
- Goal:
  - Phenotype extraction
  - Generic – to be used for a variety of retrievals and use cases
  - Expandable – at the information model level and methods
  - Modular
  - Cutting edge technologies – best methods combining existing practices and novel research with rapid technology transfer
  - Best software practices (80M+ notes)



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
    - …

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[26] 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

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.

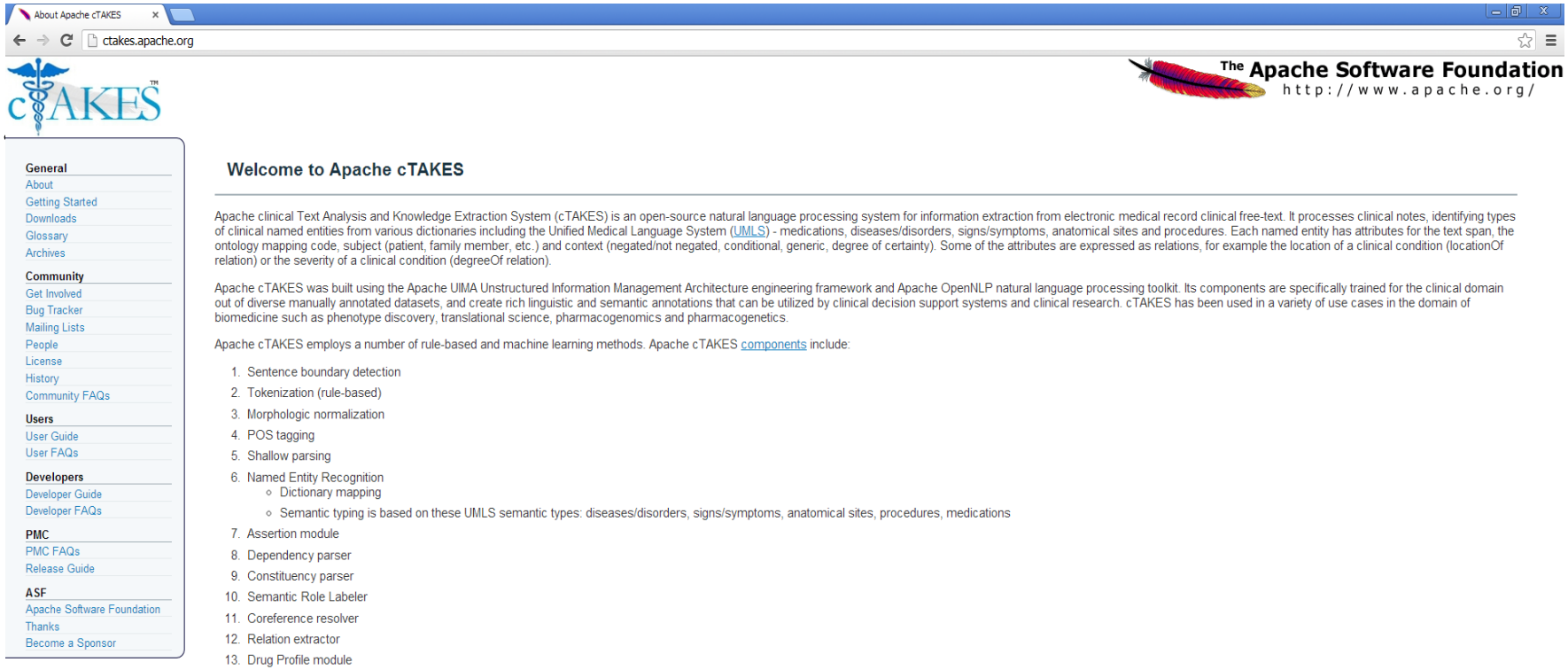
Methods Manual annotation of a clinical narrative corpus of 127 606 tokens following the Treebank schema for syntactic information, PropBank schema for predicate-argument structures, and the Unified Medical Language System (UMLS) schema for semantic information. NLP components were developed.

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.

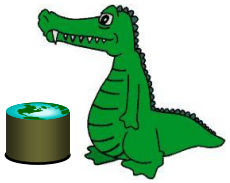
other), the level of certainty associated with an event (confirmed, possible, negated) as well as textual mentions that point to the same event. We describe our efforts to combine annotation types developed for general domain syntactic and semantic parsing with medical-domain-specific annotations to 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-



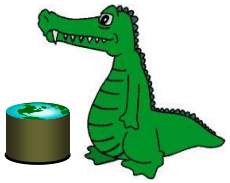
Start

ctakes.apache.org



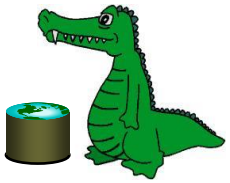
cTAKES: Components

- **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**



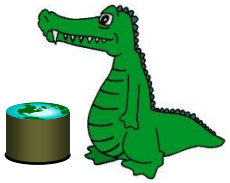
cTAKES Technical Details

- 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



What is UIMA (you – eee –muh)?

- Unstructured Information Management Architecture from IBM Research
- Open source scalable 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

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