Apache Clinical Text Analysis and Knowledge Extraction System

takes.apache.org
The patient underwent a CT scan in April which did not reveal lesions in his liver.

<table>
<thead>
<tr>
<th>Entity Recognition</th>
<th>Procedure</th>
<th>Disease / Disorder</th>
<th>Anatomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMLS ID: C0040405</td>
<td>Negated: no</td>
<td>Negated: yes</td>
<td>Negated: no</td>
</tr>
<tr>
<td>Subject: patient</td>
<td>Subject: patient</td>
<td></td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UMLS Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesions LOCATED AT liver</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temporal Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT scan WITHIN April</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coreferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient SAME AS his</td>
</tr>
</tbody>
</table>
The patient is a 53-year-old man referred by Dr. Dood for recently diagnosed colorectal cancer. The patient was well till 6 months ago, when he started having a little blood with stool. He initially thought it was hemorrhoids related, and saw his primary physician after a few weeks of symptoms. As physical examination by the primary care provider did show small hemorrhoids, but as stool was strongly positive for blood, he was referred for a repeat colonoscopy. Due to family circumstances, this was postponed for a few months. He had a prior colonoscopy at age 50, which revealed 5 or 6 polyps, all adenomatous, and the maximum size of the largest polyp was about 1 cm. This polyp was located in the cecum. The repeat colonoscopy on March 26, 2014, showed a 4 cm likely colorectal cancer in the cecum, just behind the ileocecal valve. Biopsies were positive for grade 3 out of 4 adenocarcinoma. Stains for HRPC genes were negative. Patient did bring the slides of the colonoscopy biopsies but left these in his hotel room.

The patient underwent a staging CT scan early April, which did not reveal obvious intra-abdominal spread and large lymph nodes, or metastatic lesions in his liver. We do not have the images available today. The patient is now here at our institution for evaluation of possible surgery.

Service Date/Time: 02-May-2014 11:33
Provider: Johnny Cash, MD
Page: 4
Service: GI Type/Desc: MIS Status: Trg Revision #: 2
[System id: "20114"]

[Start section id="20112"]
CHIEF COMPLAINT/PURPOSE OF VISIT
#1 Colorectal Cancer
[System id: "20112"]

[Start section id="20106"]
HISTORY OF PRESENT ILLNESS
The patient is a 53-year-old man referred by Dr. Dood for recently diagnosed colorectal cancer. The patient was well till 6 months ago, when he started having a little blood with stool. He initially thought it was hemorrhoids related, and saw his primary physician after a few weeks of symptoms. As physical examination by the primary care provider did show small hemorrhoids, but as stool was strongly positive for blood, he was referred for a repeat colonoscopy. Due to family circumstances, this was postponed for a few months. He had a prior colonoscopy at age 50, which revealed 5 or 6 polyps, all adenomatous, and the maximum size of the largest polyp was about 1 cm. This polyp was located in the cecum. The repeat colonoscopy on March 26, 2014, showed a 4 cm likely colorectal cancer in the cecum, just behind the ileocecal valve. Biopsies were positive for grade 3 out of 4 adenocarcinoma. Stains for HRPC genes were negative. Patient did bring the slides of the colonoscopy biopsies but left these in his hotel room.

The patient underwent a staging CT scan early April, which did not reveal obvious intra-abdominal spread and large lymph nodes, or metastatic lesions in his liver. We do not have the images available today. The patient is now here at our institution for evaluation of possible surgery.

[System id: "20103"]
cTAKES Global Users

Users have many use cases ...

Point of Care is not (yet) one of them
cTAKES Growth

- cTAKES created at the Mayo Clinic in 2006
- Dr. Savova moves to Boston Children’s Hospital in 2010
- Article on cTAKES (Savova et al, 2010) highest cited in the Journal of American Medical Informatics Association in 2013
- cTAKES becomes an official Apache project in 2013

Participating in Core Projects
- PGRN
- eMerge Network
- i2b2

Participating in NLP challenges
- Shared Annotated Resources
- SemEval 2015
Why Apache

Free Resources and Support

– Version Control
– Web Server
– Wiki / Documentation
– Email List Servers
– Access to Software
– Legal Support

Global Presence
Name Recognition

Core, Dedicated Group of NLP Researchers and Software Developers

So we are golden, right?
Maintaining Quality

Open Source Community
- Test cTAKES
- Report and Fix Bugs
- Add Enhancements
- Provide email Help
- Write Documentation
- Contribute to Official Site
- Contribute Entire Modules

Core Group handles most of the Effort
Funding

Apache Provides Support

Open Source Community Contributes

The Core Group is not free ...

Funding through

- Grants, with Specific Aims on improving Clinical NLP
- Running cTAKES for large and small Funded Projects
- *No* Funding from our Institutions
  
  But we are working on it ...
- Creating a Data Distribution Center
- Creating a Processing Center
Challenges

• Answering Questions
  – Wait for Community

• Fixing
  – Bait the Community

• Enhancing
  – Ask Community for similarities
  – Create seed, ask Community for assistance
  – Offer to absorb 3rd party tools as parts of cTAKES
    • yTEX NLP Tools
    • Scrubber De-Identification

• Funding
  – See previous slide
Questions?
User Community

• Types of users
  – Developers
  – End users -- biomedical investigators, data warehouse managers, point-of-care clinicians
  – Active user and developer mailing lists
Applications

• Patient cohort identification from the EMR – eMERGE, PGRN, i2b2
• Analysis of rare diseases – Phelan McDermut Syndrome
• Pharmacovigilance – adverse events from twitter, MTX liver toxicity
• Patient-facing applications – patient-interpretable clinical notes
• Point-of-care – summarization
• Question-answering
• Quality metrics
Apache cTAKES

- Natural language processing system (NLP)
- Extracts information from electronic medical record clinical text
- Top level open source project in the Apache Software Foundation
- Diverse and Global User Base
Key Features

- **Powerful**: discover codable entities, events, attributes and relations
- **Fast**: process batches at 50,000 clinical notes per hour
- **Scalable**: run on clusters, queue systems and cloud computing services
- **Modular**: use only the components needed
- **Portable**: run on any major computer platform
- All machine learning models are trained on expert-annotated gold standard data
The patient underwent a CT scan in April which did not reveal lesions in his liver.
The patient underwent a CT scan in April which did not reveal lesions in his liver.
The patient underwent a **CT scan** in April which did not reveal **lesions** in his **liver**.

<table>
<thead>
<tr>
<th>Boundary Detection</th>
<th>The patient underwent a CT scan in April which did not reveal lesions in his liver.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization</td>
<td>DT patient undergo CT scan in April which did not reveal lesions in his liver.</td>
</tr>
<tr>
<td>Normalization</td>
<td>DT NN undergo VBD DT NN CT scan IN April WDT did VBD not reveal VBD lesion NNS IN PRP$ NN Liver .</td>
</tr>
<tr>
<td>Part-of-speech Tagger</td>
<td>CT scan Procedure UMLS ID: C0040405</td>
</tr>
</tbody>
</table>

**cTAKES** can normalize to domain ontologies such as SNOMED-CT and RxNORM.
The patient underwent a CT scan in April which did not reveal lesions in his liver.

Chunking:
- NP
- VP
- NP
- PP
- NP
- VP
- NP

Constituency Parsing:
- S
- NP
- DT
- NN
- VP
- ..........
- ..........
The patient underwent a CT scan in April which did not reveal lesions in his liver.
The patient underwent a CT scan in April which did not reveal lesions in his liver.
Clinical Element Model Template

**Sign/Symptom**
- Alleviating Factor
- Associated Code
- 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**
- Abnormal Interpretation
- Associated Code
- Conditional
- Delta Flag
- Estimated flag
- Generic

**Procedure**
- Associated Code
- Body Laterality
- Body Location
- Body Side
- Conditional
- Device
- End Date
- Generic
- Negation Indicator
- Relative Temporal Context
- Start Date
- Subject
- Uncertainty Indicator

**Disease/Disorder**
- Alleviating Factor
- Associated Sign or Symptom
- Associated Code
- 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**
- Associated Code
- Body Laterality
- Body Site
- Conditional
- Generic
- Negation Indicator
- Subject
- Uncertainty Indicator

**Medication**
- Associated Code
- Change Status
- Conditional
- Dosage
- Duration
- End Date
- Form
- Frequency
- Generic
- Negation Indicator
- Route
- Start Date
- Strength
- Subject
- Uncertainty Indicator
FHIR®

FHIR Composition representing the document
System Integration

cTAKES has been integrated with

• i2b2 platform

• TranSMART

Partial TranSMART environment
Measures of Success

- Goal – enable groundbreaking medical research
- Publications using the tool
- Mailing list activity around the tool
Sustainability Plans

- The Apache Community
- In an active state of fund-raising
Lower Entry Barrier

• Based on community feedback
  – Improve ease of use
  – API
  – GUI
  – Documentation and how to videos
Community Contributions

- Apache Software Foundation principles
  - Do-acracy
  - Community testing