TT07

# A Semantic NLP Approach for Structuring and Analysis of FDA Meeting Minutes Documents

Presenters:
Michelle Shen, FDA/CDER/OND
Suresh Subramani, PhD, FDA/NCTR/DBB

# Disclaimer

The information in these materials is not a formal dissemination of the U.S. Food and Drug Administration.

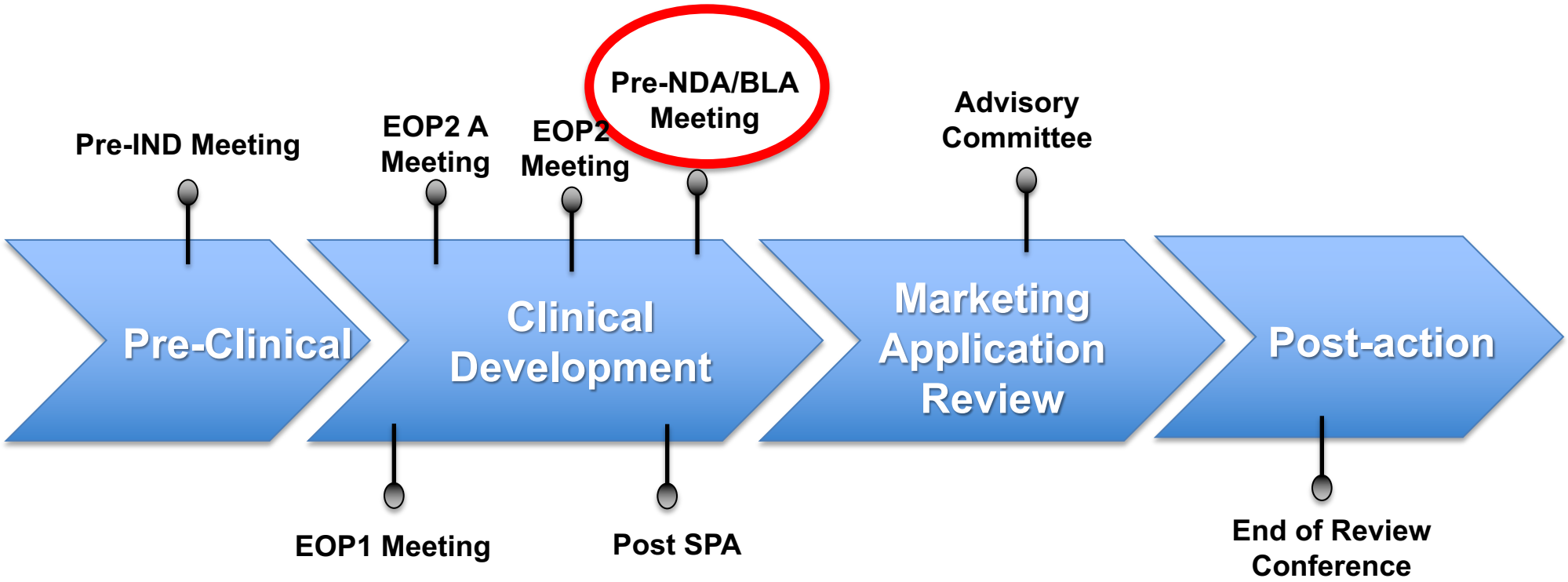
The views expressed in this presentation are the authors' and do not necessarily represent the official views or policies of the U.S. Food and Drug Administration.

The presenters are ORISE Fellows and have no conflicts of interest, financial or otherwise, that may be presented in this activity.

# Background

- The mission of the Center for Drug Evaluation and Research (CDER) is to ensure that drugs marketed in this country are safe and effective
- Prior to approval, drugs are submitted and reviewed through a new drug application (NDA) or biologics license agreement (BLA)
- There are many scheduled interactions between Sponsors and FDA during the drug development process
- One way for Sponsors to interact with the FDA during the development process is through formal meetings (~2500-3000 per year)

# Drug Development Meeting Timeline



# Meeting Minutes

- Meetings with Sponsors during the Investigational New Drug (IND) phase are documented as Meeting Minutes and archived in FDA's Document Archiving, Reporting and Regulatory Tracking System (DARRTS)

- When preparing for reviews and industry meetings, reviewers often search minutes for past regulatory decisions for reference

# Challenge: Meeting Minutes Accessibility

Challenge:

Meeting Minutes documents contain complex information that is not easily retrievable with free-text searches

Solution:

Use natural language processing (NLP) to extract document information, which combined with established ontologies, will allow for document retrieval through improved search capabilities including hierarchical search

# Focus Areas for NLP Extraction

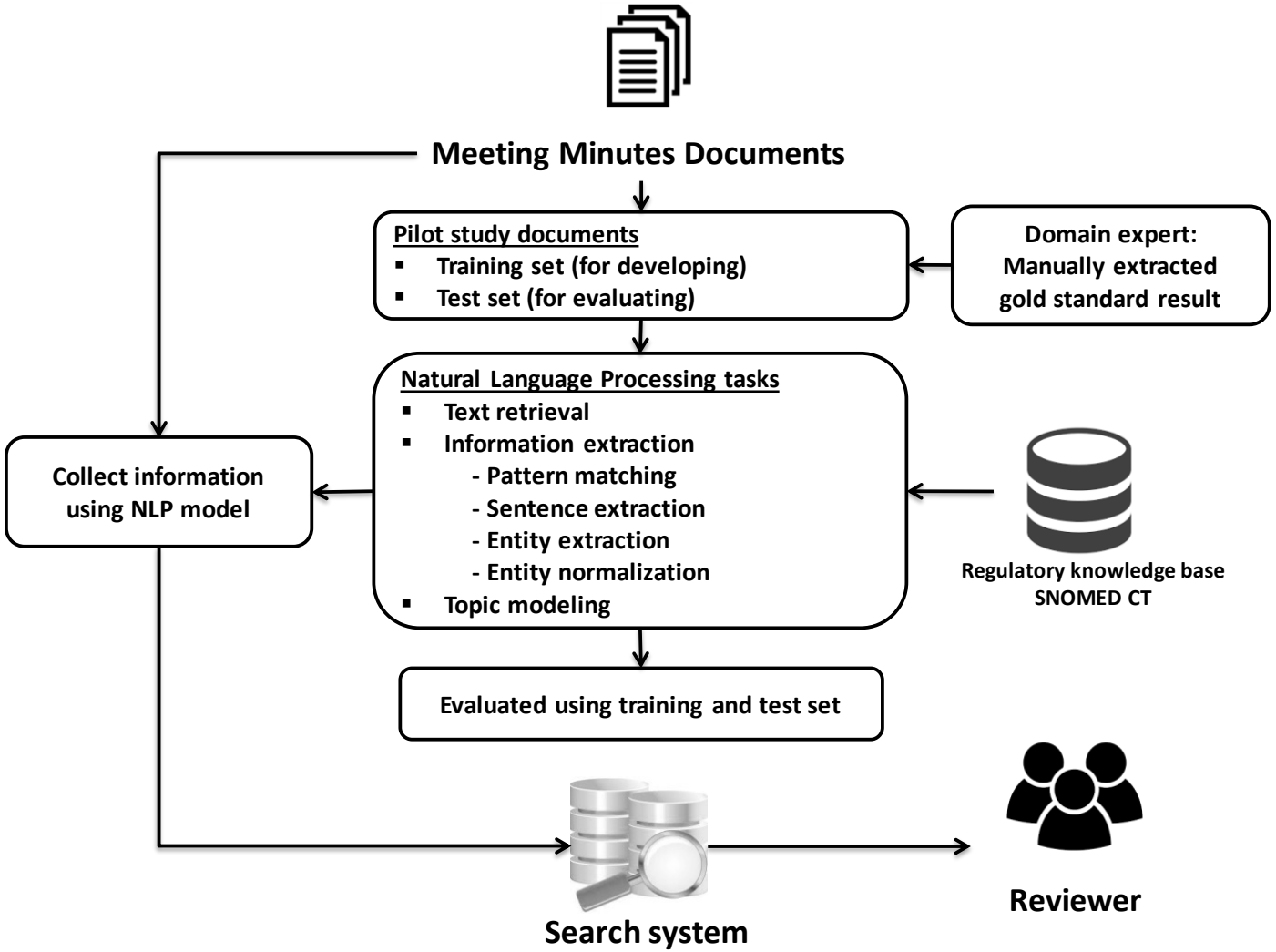
1. Document metadata (semi-structured)
   - To extract metadata from the Meeting Minutes documents for indexing in a database
   - To map critical fields to an established ontology (e.g., Indication to SNOMED CT Fully Specified Names)

2. Q&A section (unstructured)
   - Synthesize and incorporate knowledge contained within the PDF documents to inform review work

# Pilot Scope Criteria

- 230 Pre-NDA and Pre-BLA Meeting Minutes from 2015 and 2016
- 46 documents (20%) were randomly selected for a training and test set
  - 23 training for developing the algorithm
  - 23 test for evaluating the trained algorithm
- Gold standards for training and test sets were generated by domain experts

# Natural Language Processing (NLP) Information Extraction Model



# TEXT RETRIEVAL

# Download and Conversion of Documents

- PDF documents were individually downloaded from DARRTS

- PDF documents were converted into plain text using Apache® PDFBox [1], a Java PDF library for convenient text extraction

# SEMI-STRUCTURED TEXT EXTRACTION: DOCUMENT METADATA

Objective: To extract the fields and field-values of interest

# Semi-structured Text Extraction Steps

1. Creation of template-based pattern
2. Pattern matching
3. Sentence extraction for Indication
4. Entity extraction and normalization for Indication

Semi-structured Text Extraction: Metadata

# 1. CREATION OF TEMPLATE-BASED PATTERN

# DOCUMENT INFORMATION PAGE

This page is for FDA internal use only. **Do NOT send this page with the letter.**

| | |
|---|---|
| **Application #(s):** | PIND/IND######<br>NDA######/S-###<br>BLA######/S-### |
| **Communication Type:** | XXX |
| **Communication Group:** | XXX |
| **Communication Name:** | XXX |
| **Communication ID:** | XXX |
| **Drafted by:** | XXX |
| **Clearance History:** | XXX |
| **Finalized:** | XXX |
| **Filename:** | XXX |
| **PDUFA Goal Impact** | XXX |
| **Signatory Authority:** | XXX |
| **Use Statement:** | XXX |
| **Notes:** | • **XXX** |

Version: 04/18/2016

# END OF DOCUMENT INFORMATION PAGE

**The letter begins on the next page**

**Document Information Page:**

Contains administrative information about the Meeting Minutes document for internal tracking

# DOCUMENT INFORMATION PAGE

This page is for FDA internal use only. **Do NOT send this page with the letter.**

| | |
|---|---|
| **Application #(s):** | PIND/IND######<br>NDA######/S-###<br>BLA######/S-### |
| **Communication Type:** | XXX |
| **Communication Group:** | XXX |
| **Communication Name:** | XXX |
| **Communication ID:** | XXX |
| **Drafted by:** | XXX |
| **Clearance History:** | XXX |
| **Finalized:** | XXX |
| **Filename:** | XXX |
| **PDUFA Goal Impact** | XXX |
| **Signatory Authority:** | XXX |
| **Use Statement:** | XXX |
| **Notes:** | • **XXX** |

Version: 04/18/2016

# END OF DOCUMENT INFORMATION PAGE

**The letter begins on the next page**

Reference ID: XXXXXX

**Document Information Page:**

Contains administrative information about the Meeting Minutes document for internal tracking

PIND/IND######
NDA######/S-###
BLA######/S-###

**MEETING MINUTES**

SPONSOR NAME
Attention: CONTACT NAME
TITLE
SPONSOR ADDRESS

Dear CONTACT:

[Text]

We also refer to the meeting/telecon between representatives of your firm and the FDA on MEETING DATE. The purpose of the meeting was to INSERT PURPOSE.

A copy of the official minutes of the meeting/telecon is enclosed for your information. Please notify us of any significant differences in understanding regarding the meeting outcomes. If you have any questions, call me/NAME, Regulatory Project Manager at NUMBER.

Sincerely,

CPMS OR RPM OR MEETING CHAIR SIGNATURE BLOCK
TITLE
Division of DIVISION NAME
Office of Drug Evaluation XX
Center for Drug Evaluation and Research

Enclosure:
Meeting Minutes

**Cover Letter Page:**

Cover letter for the document from the Agency to the sponsor that provides a summary of content, general discussion, or action items documented in the Meeting Minutes document

PIND/IND######
NDA######/S-###
BLA######/S-###

**MEETING MINUTES**

SPONSOR NAME
Attention: CONTACT NAME
TITLE
SPONSOR ADDRESS

Dear CONTACT:

[Text]

We also refer to the meeting/telecon between representatives of your firm and the FDA on MEETING DATE. The purpose of the meeting was to INSERT PURPOSE.

A copy of the official minutes of the meeting/telecon is enclosed for your information. Please notify us of any significant differences in understanding regarding the meeting outcomes. If you have any questions, call me/NAME, Regulatory Project Manager at NUMBER.

Sincerely,

CPMS OR RPM OR MEETING CHAIR SIGNATURE BLOCK
TITLE
Division of DIVISION NAME
Office of Drug Evaluation XX
Center for Drug Evaluation and Research

Enclosure:
Meeting Minutes

**Cover Letter Page:**

Cover letter for the document from the Agency to the sponsor that provides a summary of content, general discussion, or action items documented in the Meeting Minutes document

# MEMORANDUM OF MEETING MINUTES

**Meeting Type:** [Identify type of meeting, such as A, B, C]
**Meeting Category:** [Identify category of meeting, i.e., Pre-IND, End of Phase 2, Pre-NDA, etc.]

**Meeting Date and Time:** [Insert meeting date and time]
**Meeting Location:** [Insert meeting location]

**Application Number:** [Insert application number]
**Product Name:** [Insert product name]
**Indication:** [Insert indication]
**Sponsor/Applicant Name:** [Insert sponsor name]

**Meeting Chair:** [Insert the meeting leader's name.]
**Meeting Recorder:** [Insert the meeting recorder's name.]

**FDA ATTENDEES**

[redacted]

**SPONSOR ATTENDEES**

[redacted]

**Memo Page:**

Meta-information about the meeting and drug product for which the Meeting Minutes document was recorded

# MEMORANDUM OF MEETING MINUTES

| | |
|---|---|
| **Meeting Type:** | [Identify type of meeting, such as A, B, C] |
| **Meeting Category:** | [Identify category of meeting, i.e., Pre-IND, End of Phase 2, Pre-NDA, etc.] |
| **Meeting Date and Time:** | [Insert meeting date and time] |
| **Meeting Location:** | [Insert meeting location] |
| **Application Number:** | [Insert application number] |
| **Product Name:** | [Insert product name] |
| **Indication:** | [Insert indication] |
| **Sponsor/Applicant Name:** | [Insert sponsor name] |
| **Meeting Chair:** | [Insert the meeting leader's name.] |
| **Meeting Recorder:** | [Insert the meeting recorder's name.] |

**FDA ATTENDEES**

[redacted]

**SPONSOR ATTENDEES**

[redacted]

**Memo Page:**

Meta-information about the meeting and drug product for which the Meeting Minutes document was recorded

Semi-structured Text Extraction: Metadata

# 2. PATTERN MATCHING

# Field Name Possibility Pattern Matching

First iteration patterns:

- Attempted to extract information based on field names identified from the Meeting Minutes standard template [2]
- Variation in formatting and versioning led to a failure to extract complete information

Second iteration patterns:

- Used manually added patterns found from first iteration incomplete fields in the training set

# Example: Memo Page Possibility Patterns List

| Field Name | Original Identified Patterns in Iteration 1 (Template) | Possibility Patterns Added to Iteration 2 (Training set) |
|---|---|---|
| Memorandum of Meeting Minutes | MEMORANDUM OF MEETING MINUTES | • AMENDED MEMORANDUM OF MEETING MINUTES<br>• MEMORANDUM OF TELECONFERENCE MINUTES<br>• MEMORANDUM OF TELECONFERENCE<br>• PRELIMINARY MEETING COMMENTS<br>• TO PRELIMINARY MEETING COMMENTS<br>• MEETING COMMENTS |
| Product Name | Product Name: | • Product:<br>• Products:<br>• Drug: |
| Sponsor/Applicant Name | Sponsor/Applicant Name: | • Sponsor Name:<br>• Sponsor/Applicant:<br>• Sponsor: |

Semi-structured Text Extraction: Metadata

# 3. SENTENCE EXTRACTION FOR INDICATION

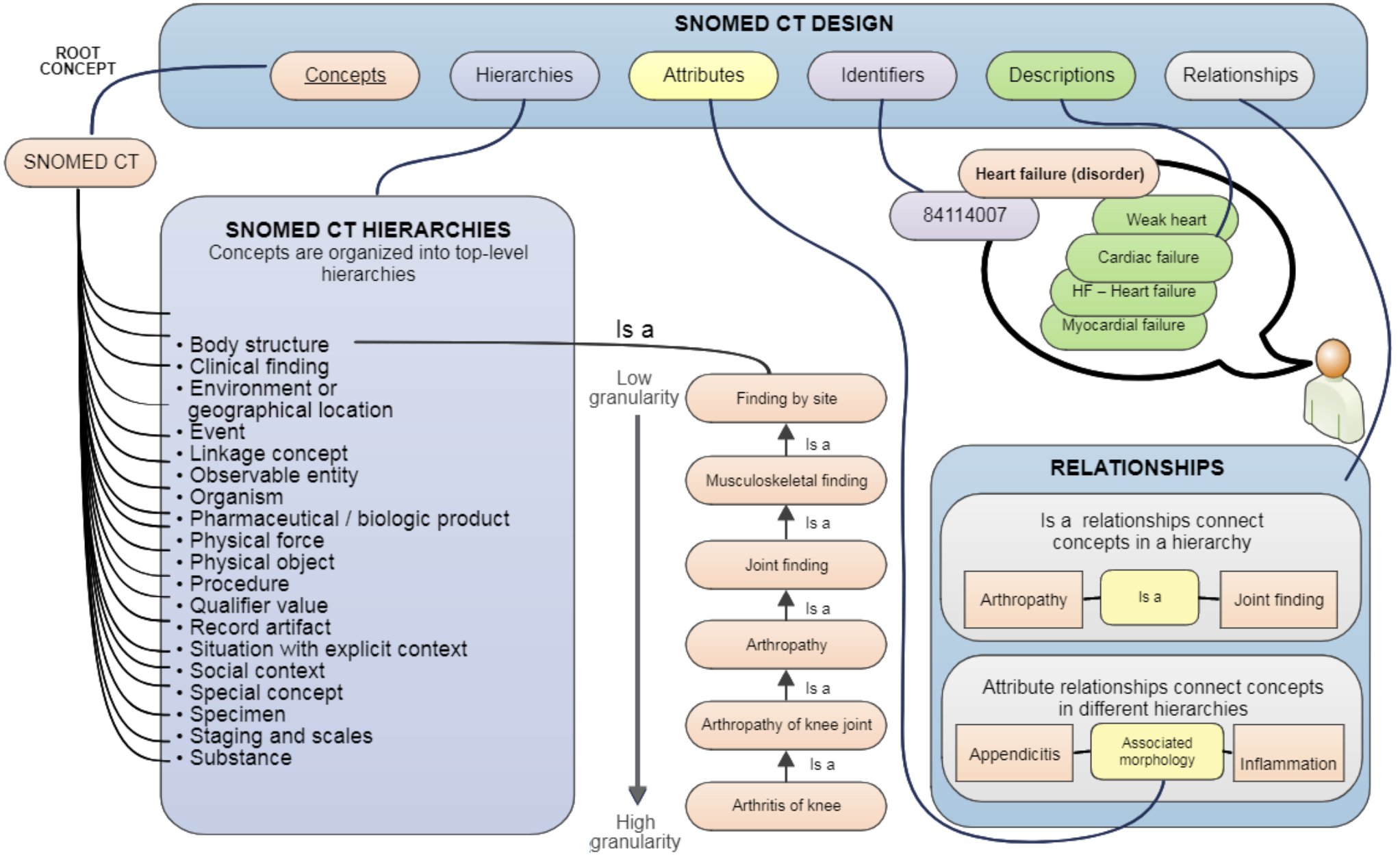
# Significance of Mapping Indication to SNOMED CT

- SNOMED CT, or the Systematized Nomenclature of Medicine -- Clinical Terms [3], is a standardized, multilingual vocabulary of clinical terminology used by physicians and other health care providers for electronic exchange of clinical health information

- Concepts are organized in hierarchies from more general concepts to the more detailed concepts

- SNOMED CT relationships link concepts to other concepts with related meanings

- Mapping Indication to SNOMED CT concepts allows reviewers to search Meeting Minutes across related Indications




https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics

# Sentence Extraction

- Indication is a free text field with one or more sentences

- Using pattern matching, extracted Indication field text was split into single sentences using Perl module Lingua::EN::Sentence [4]

Semi-structured Text Extraction: Metadata

# 4. ENTITY EXTRACTION AND NORMALIZATION

# Entity Tagging and Identification

- Perl Lingua::StanfordCoreNLP [5] module was modified for sentence parsing and used to generate a constituent tree of tagged subjects, objects, and parts of speech from extracted Indication sentences

- Tagged noun phrases (NP) and verb phrases (VP) were isolated

# Entity Tagging

Original text:

```
Treatment of non-small cell lung cancer.
```

Stanford NLP Algorithm constituent tree output:

```
(ROOT
 (NP
   (NP (NNP Treatment))
   (PP (IN of)
     (NP (JJ non-small) (NN cell) (NN lung) (NN cancer)))
   (. .)))
```

# Retaining Noun and Verb Phrases

Original output (line numbers added):

```
1 (ROOT
2  (NP
3    (NP (NNP Treatment))
4    (PP (IN of)
5      (NP (JJ non-small) (NN cell) (NN lung) (NN cancer)))
6    (. .)))
```

Retained output after filtering results (line numbers added):

```
3    (NP (NNP Treatment))
5      (NP (JJ non-small) (NN cell) (NN lung) (NN cancer)))
```

# SNOMED CT Disorder Dictionary and String Matching

- A SNOMED CT Disorder Normalized Dictionary was generated from SNOMED CT US Edition
  - Downloaded 196,533 Synonyms for 72,551 Active Disorder Concepts
  - Contained list of Fully Specified Names for Disorder (Disease) Concepts, averaging 2-3 Synonyms per Concept
- Perl Regex::PreSuf [6] was used to generate regular expressions from the Disorder Dictionary and to match to retained NP and VP
- Additional string matching rules were applied for complex matching

# String Matching Dictionary Rules

| Rule | Preprocessed Text Example | Postprocessed Text Example |
|---|---|---|
| 1. Normalization of case | DIABETES MELLITUS | diabetes mellitus |
| 2. Replacement of hyphens with spaces | Drug-induced dyskinesia | Drug induced dyskinesia |
| 3. Removal of stop words | Skin disorder of umbilicus | Skin disorder umbilicus |
| 4. Elimination of word delimiters (semicolons, colons, and commas) | Myeloid sarcoma, disease | Myeloid sarcoma disease |

# Added Dictionary Rules

| Rule | Preprocessed Text Example | Postprocessed Text Example |
|---|---|---|
| 1. Replacement of synonym words: “disorder” to “disease” and vice versa | Bipolar disease | Bipolar disorder |
| 2. Accounting for both singular and plural forms | Acute coronary syndrome | Acute coronary syndromes |
| 3. Allowing for word permutations of up to 5 words | Ricin poisoning | Ricin poisoning; poisoning Ricin |

# FREE TEXT EXTRACTION: Q&A SECTION

Objective: To correctly identify Q&A sections in the document body and accurately extract questions and answers

# Overview of Free Text Extraction

1. Creation of template-based pattern

2. Pattern matching and entity extraction

Free Text Extraction: Q&A Section

# 1. CREATION OF TEMPLATE-BASED PATTERN

# Meeting Minute Body General Structure

**1.0 BACKGROUND**
[Text]

**2.0 DISCUSSION**
**2.1 Category/Discipline A**
***Question 1:*** [Text]
***FDA Response to Question 1:*** [Text]
***Discussion:*** [Text]

***Question 2:*** [Text]
***FDA Response to Question 2:*** [Text]
***Discussion:*** [Text]

**2.2 Category/Discipline B**
***Question 1:*** [Text]
***FDA Response to Question 1:*** [Text]
***Discussion:*** [Text]

**3.0 OTHER IMPORTANT MEETING LANGUAGE SECTIONS**
[Text]

**4.0 ISSUES REQUIRING FURTHER DISCUSSION**
[Text]

**5.0 ACTION ITEMS**
[Text]

| Action Item/Description | Owner | Due Date |
|---|---|---|
| [Text] | [Text] | [Date] |
| [Text] | [Text] | [Date] |

**6.0 ATTACHMENTS AND HANDOUTS**
[Text]

# Body Content Q&A Section Locations

| Meeting Minutes Template Section | Description |
|---|---|
| 1.0 BACKGROUND | Details meeting purpose, meeting context, history of events leading to the meeting, and context for product development. |
| 2.0 DISCUSSION<br>Category/Discipline A<br>Question 1:<br>FDA Response to Question 1:<br>Discussion: | This is the section where sponsor questions and FDA responses are recorded. |
| 4.0 ISSUES REQUIRING FURTHER DISCUSSION | If there are additional discussions on the date of the teleconference or face-to-face meeting, the comments are usually recorded in this area. Additional Comments is often in this section. |

Free Text Extraction: Q&A Section

# 2. PATTERN MATCHING AND ENTITY EXTRACTION

# Pattern List 1

The first pattern list, Pattern List 1, was generated based on document structure

- Assumed that documents adhered strictly to template structuring and numbering

Pattern List 1 failed to capture complete information

- Document formatting deviations from the standard template pattern identified as cause of incomplete information capture

# Pattern List 2

- Improved Pattern List 1 by identifying where training set did not capture complete information using the template pattern
- Pattern List 2 incorporated additional patterns from training set to account for variable nature of body content and structure
- Created Heading and Middle Heading possibility patterns lists

## 1.0 BACKGROUND

[Text]

## 2.0 DISCUSSION

### 2.1 Category/Discipline A

***Question 1:*** [Text]

***FDA Response to Question 1:*** [Text]

***Discussion:*** [Text]

***Question 2:*** [Text]

***FDA Response to Question 2:*** [Text]

***Discussion:*** [Text]

### 2.2 Category/Discipline B

***Question 1:*** [Text]

***FDA Response to Question 1:*** [Text]

***Discussion:*** [Text]

Key:

**Heading**

**Middle Heading**

# Headings Possibility Patterns List

| Heading Type | Description | Example (Heading Patterns List) |
|---|---|---|
| Document Section Headers | Divides the Meeting Minutes into several sections; based on standard FDA templated structured sections. | • "Background"<br>• "Discussion"<br>• "Questions and Responses" (Q&A Section)<br>• "Additional Comments"<br>• "Meeting Discussion"<br>• "General Discussion"<br>• "Additional Comments" |
| Category Headers | Used to classify Q&A section questions relating to a specific topic. | • "Safety"<br>• "Regulatory"<br>• "Administrative"<br>• "Regulatory History"<br>• "Datasets" |
| Discipline Headers | Used to classify Q&A section questions relating to a specific discipline. | • "Nonclinical"<br>• "Clinical"<br>• "Statistics"<br>• "Clinical Pharmacology"<br>• "Chemistry, Manufacturing, and Controls"<br>• "Pharmacokinetics"<br>• "Biometrics"<br>• "Toxicology" |

# Middle Headings Possibility Patterns Examples

| Questions Possibility Patterns | Responses Possibility Patterns |
|---|---|
| • Question 3:<br>• Question 3.0:<br>• Question 4a:<br>• Sponsor Question 1a:<br>• SPONSOR QUESTION 1:<br>• 2. Background:*<br>• 2.0 Background:*<br>• 2. Question<br>• [Sponsor Name] Question 3:<br>• [Reviewer Name] Question 13a:<br>• Q2:<br>• 1. Does the Agency agree th…?<br>• 3) Does the Agency agree th…?<br>• b. Does the Agency agree th…? | • Response to Question:<br>• FDA Response:<br>• FDA Response to Question 14:<br>• FDA Response to Question 14b:<br>• FDA's Preliminary Response Sent on [date]:<br>• Sponsor's Clarifying Request for Question 1:<br>• [Sponsor Name]'s emailed response of [date]:<br>• Sponsor Response (via email [date]): |

# TOPIC MODELING

Objective: Evaluate common elements among content in Meeting Minute documents

# Data Preparation

- To lower case, tokening, stemming and removing general stop-words
- Filtering with specific start-words and stop-words
- Low (5) and high (1000) thresholds
- Final vocabulary: 3707 words

# Modeling and Visualization

- Modeling
  - LDA [7] in Mallet [8]
  - Gibbs sampling [9]
  - Asymmetric alpha = 0.1
  - Beta = 0.01
  - Topic number = 20
  - Number of iteration = 3,000
- Visualization
  - Top-5 most probable words
  - Word clouds

Semi-structured Text Extraction: Metadata

# RESULTS & DISCUSSION

# NLP Performance Metrics

- Domain experts manually created a gold standard for the metadata sections of document
- Algorithm extraction values were manually compared with the gold standard to determine True Positives (TPs), False Positives (FPs), True Negatives (TNs), and False Negatives (FNs) for each extracted field
- 1518 total fields (33 per document) were evaluated

- Accuracy, $A = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$
- Precision, $P = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- Recall, $R = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
- F−score, $F = \frac{2PR}{P+R}$

# NLP performance evaluation of semi-structured field-value extraction in document metadata pages

| | DOCUMENT INFORMATION PAGE | | COVER LETTER PAGE | | MEMO INFORMATION PAGE | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | Training | Test | Training | Test | Training | Test | Training | Test |
| True Positives (TPs) | 309 | 298 | 179 | 177 | 246 | 250 | 734 | 725 |
| False Positives (FPs) | 0 | 1 | 1 | 3 | 0 | 0 | 1 | 4 |
| False Negatives (FNs) | 0 | 1 | 4 | 4 | 0 | 1 | 4 | 6 |
| True Negatives (TNs) | 13 | 22 | 0 | 0 | 7 | 2 | 20 | 24 |
| Total # Fields | 322 | 322 | 184 | 184 | 253 | 253 | 759 | 759 |
| | | | | | | | | |
| **Accuracy %** | **100** | **99.4** | **97.3** | **96.2** | **100** | **99.6** | **99.3** | **98.7** |
| **Precision %** | **100** | **99.7** | **99.4** | **98.3** | **100** | **100** | **99.9** | **99.5** |
| **Recall %** | **100** | **99.7** | **97.8** | **97.8** | **100** | **99.6** | **99.5** | **99.2** |
| **F-score %** | **100** | **99.7** | **98.6** | **98.1** | **100** | **99.8** | **99.7** | **99.3** |

# Discussion

- Reported accuracy, precision, recall, and F-scores for extraction of test set semi-structured field-values in metadata pages were >90%

- Fields of particular interest (Application Number, Meeting Type, Meeting Category, and Indication) did not have FPs or FNs in all training and test documents

- FPs and FNs occurred most frequently for other fields less critical to understanding decision making and recommendations

# Challenges

1. Document archival in PDF format
   - Unstructured data
   - Tables and figures
2. Versioning inconsistencies in templates
3. Fields allow entry of free text
   - Breaking of grammatical rules
4. Strict conformance to standard template guidelines (numbering and structuring) was variable
   - Template allowed for free-text editing of field-values
5. Automated extraction is more difficult due to multiple methods
6. Validation required manual creation of gold standard

# Conclusion: Accomplishments

Designed, developed, and evaluated an automated text mining tool that uses NLP to help with CDER's knowledge management efforts

- Model extracted FDA Meeting Minutes metadata with high precision and recall
- Proof of concept was developed for extracting Q&A sections of Meeting Minutes using rule-based pattern matching

# Conclusion: Future Plans

1. Extract metadata field information from additional documents
   - Store extracted metadata field information in a database
   - Grouping of documents by Indication, meeting type, pharmacologic class

2. Evaluate accuracy and comprehensiveness of extracted Q&A sections to further refine the model
   - Information can be leveraged from this knowledge base and synthesized to inform review work and development of standard meeting communications

# References


[1] Apache Software Foundation. (2016). Apache PDFBox 2.0.6. Retrieved from https://pdfbox.apache.org/

[2] Food and Drug Administration. (2018, March 14). Meeting Minutes (COR-MEET-03).

[3] International Health Terminology Standard Development Organization. (2017). Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT) US Edition: March 2017 Release. Retrieved from https://www.nlm.nih.gov/healthit/snomedct/us_edition.html

[4] Yona, S. (2016). Lingua::EN::Sentence - split text into sentences. Retrieved from http://search.cpan.org/~kimryan/Lingua-EN-Sentence-0.30/lib/Lingua/EN/Sentence.pm

[5] Raisanen, K. (2011-2013). Lingua::StanfordCoreNLP - A Perl interface to Stanford's CoreNLP tool set. Retrieved from http://search.cpan.org/~kal/Lingua-StanfordCoreNLP-0.10/lib/Lingua/StanfordCoreNLP.pm

[6] Hietaniemi, J. Regex::PreSuf - create regular expressions from word lists. Retrieved from http://search.cpan.org/~jhi/Regex-PreSuf-1.17/PreSuf.pm

[7] Blei, D.M., Ng, A.Y. and Jordan M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

[8] McCallun, A.K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved from http://mallet.cs.umass.edu

[9] Griffiths, T.L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101(suppl. 1)*, 5228-5235.

# Acknowledgments

**CDER, Office of New Drugs**

**Biomedical Informatics and Regulatory Review Science Team**

Vaishali Popat, MD, MPH

Jeffry Florian, PhD

Jessica Voqui, PharmD, MS

**National Center for Toxicology Research**

**Division of Biostatistics and Bioinformatics**

Joe Meehan

Weida Tong, PhD

Weizhong Zhao, PhD

**Other Acknowledgments**

G-SRS Team

HIVE Team