PhUSE US Connect 2018

Paper TT07

A semantic NLP approach for structuring and analysis of FDA Meeting Minutes documents

Michelle Shen, FDA/CDER/OND, Silver Spring, USA
Suresh Subramani, FDA/NCTR/DBB, Jefferson, USA
Weizhong Zhao, FDA/NCTR/DBB, Jefferson, USA
Jeffry Florian, FDA/CDER/OND, Silver Spring, USA
Jessica Voqui, FDA/CDER/OND, Silver Spring, USA
Joe Meehan, FDA/NCTR/DBB, Jefferson, USA
Weida Tong, FDA/NCTR/DBB, Jefferson, USA
Vaishali Popat, FDA/CDER/OND, Silver Spring, USA

ABSTRACT

The Center for Drug Evaluation and Research (CDER) in the Food and Drug Administration (FDA) conducts approximately 2,500 to 3,000 formal industry meetings each year. In these meetings, reviewers are asked to provide general guidance and to comment on a variety of impactful issues related to the ongoing clinical development program, such as new clinical trial protocol design, adequacy of evidence for safety and efficacy, and Agency policies and procedures. To ensure consistency in responses across regulatory submissions, the reviewer must review prior communications for the same or similar products, drug classes, and indications. Current search tools produce inconsistent results and return a list of documents that must then be downloaded individually and reviewed. This is a laborious and time-intensive process. The Meeting Minutes Project evaluates the use of natural language processing (NLP) to organize previous Agency responses in a way that facilitates the identification and analysis of relevant precedent cases.

1. INTRODUCTION

The mission of the Center for Drug Evaluation and Research (CDER) in the Food and Drug Administration (FDA) is to protect and promote public health. One way to accomplish the mission is by ensuring that human drugs are safe and effective for their intended use. A critical part of this process is providing advice to drug companies that sponsor drug development programs through formal meetings with industry. The number of formal meetings between CDER and sponsors has increased over the past decade from approximately 1,800 to almost 3,000 formal industry meetings per year. Upon completion, minutes from these meetings are entered and archived into the FDA's Document Archiving, Reporting and Regulatory Tracking System (DARRTS). When preparing for industry meetings, reviewers often review minutes from past regulatory decisions to ensure consistent advice is provided to sponsors. Available tools for assisting reviewers in locating these documents have various limitations, including cumbersome user interfaces, an inability to link key fields of interest (e.g. drug class and indication) across applications, search restrictions in metadata fields containing incomplete or unstandardized free text, and returning of many false positive hits. The goal of this project is to develop a contextual, ontology-connected, reviewer-friendly, graphical user interface (GUI) search system intended for use by CDER reviewers to facilitate convenient identifying, extracting, refining, normalizing, viewing, and searching of Meeting Minutes documents. Meeting Minutes documents are archived in Portable Document Format (PDF), a format that is not conducive to extracting text via standard text mining. We use a pattern matching natural language processing (NLP) approach to extract useful metadata and to gain an understanding of the way in which existing search and archival systems at FDA capture information.

We present here two NLP approaches for capturing semi-structured and unstructured text in PDF documents with large variations in document structure. The first approach uses pattern matching and string matching methods to extract existing metadata within semi-structured fields of documents containing structural, wording, and versioning differences and to map extracted metadata to a relevant ontology. The second approach is a proof of concept to locate and extract the free-text Question and Answer (Q&A) section of the document, to isolate individual questions and corresponding answers, and to assign the questions to categories and disciplines similarly located in the free text. The proof of concept aims to overcome the lack of clear delineation of sections in the Q&A section.

2. METHODS

In this pilot, NLP capabilities were leveraged to discover and capture text information inaccessible to existing search engines at FDA, to assist in the development of a regulatory ontology for enhanced reviewer accessibility, and to isolate and categorize the Q&A sections of Meeting Minutes documents so that questions can be sorted by discipline and by therapeutic area (i.e., review Division).

NLP comprises but is not limited to several related methods, such as text retrieval, text preprocessing, sentence extraction, entity extraction, information normalization, and topic modeling [1]. The methods mentioned were employed during the text mining process for this project.

2.1 PILOT STUDY DOCUMENT SELECTION

When the sponsor of a new drug believes that enough evidence on the drug's safety and effectiveness has been obtained to meet FDA's requirements for marketing approval, the sponsor submits to FDA a new drug application (NDA) or biologics license agreement (BLA). Pre-NDA or pre-BLA meetings typically serve as the final meeting prior to submission of the application.

The pilot consisted of 230 Pre-NDA and Pre-BLA meeting minutes from the years 2015 and 2016. Of the 230 Meeting Minutes, 46 (roughly 20%) were randomly selected for a training and test set. Of these 46 documents, 23 were used to train the algorithm, and 23 were used to test the trained algorithm. We consulted four domain experts, subject matter experts who were knowledgeable and experienced in identifying the salient information from meeting minutes, to manually create gold standards for the training and test sets.

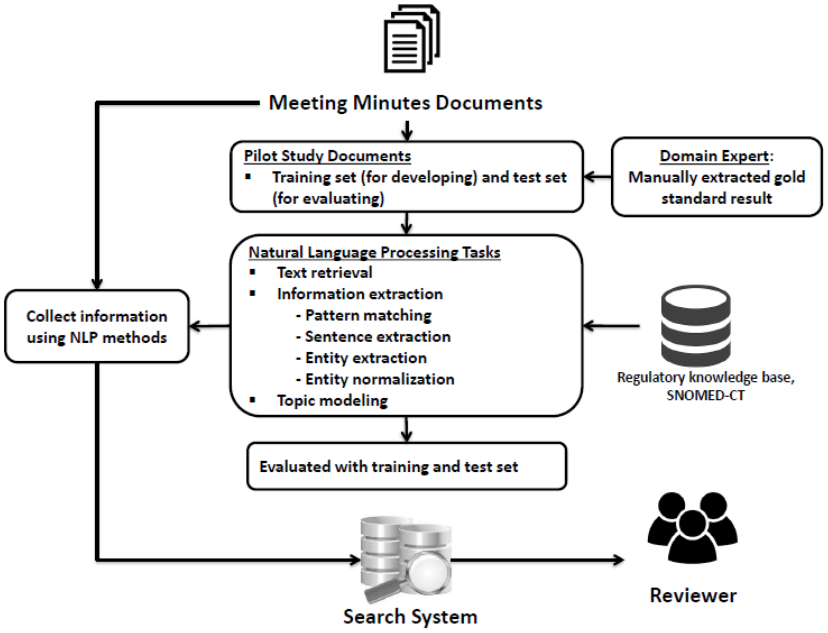

Figure 1: System Architecture Design

2.2 TEXT RETRIEVAL

PDF documents were individually downloaded from the DARRTS. Downloaded documents were manually perused and noted to have slight template versioning differences. All PDF documents were converted into plain text using Apache® PDFBox [2], a Java PDF library for convenient text extraction.

2.3 INFORMATION EXTRACTION

A general overview of the structure and content of Meeting Minutes documents, as defined by the Meeting Minutes template [3] can be found below in Table 1.

Table 1: A general overview of the Meeting Minutes template.

Meeting Minutes Template Section	Description	Subfields
DOCUMENT INFORMATION PAGE	Administrative information about the Meeting Minutes document for internal tracking.	Fourteen fields of interest were identified: Application #(s) Communication Type Communication Group Communication Name Communication ID Drafted by Clearance History Finalized Filename PDUFA Goal Impact Signatory Authority Use Statement Version Reference ID
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.	Eight fields of interest were identified: Letter Sponsor Full Address Letter Sponsor Name Letter Sponsor Attention Contact Name Letter Content Meeting Chair Signature Name Meeting Chair Signature Title Letter Signature Division Letter Signature Office
MEMO PAGE (MEMORANDUM OF MEETING MINUTES)	Meta-information about the meeting and drug product for which the Meeting Minutes document was recorded.	Eleven fields of interest were identified: Memo Meeting Type Meeting Type Meeting Category Meeting Date and Time Meeting Location Application Number Product Name Indication Sponsor/Applicant Name Meeting Chair Meeting Recorder
1.0 BACKGROUND	Details meeting purpose, meeting context, history of events leading to the meeting, and context for product development.	No fields of interest are extracted from this section.
2.0 DISCUSSION Category/Discipline A Question 1: FDA Response to Question 1: Discussion:	This is the section where sponsor questions and FDA responses are recorded.	Fields of interest include discipline, question from the sponsor, and FDA responses.
3.0 OTHER IMPORTANT MEETING LANGUAGE SECTIONS	Guidelines for documenting other important meeting language sections, such as specific requirements, regulatory pathways, or protocols, among others.	No fields of interest are extracted from this section.
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.	Fields of interest include discipline, question from the sponsor, and FDA responses.

5.0 ACTION ITEMS	Describes action items stated during the meeting.	No fields of interest are extracted from this section.
6.0 IMPORTANT ATTACHMENTS AND HANDOUTS	Attachments or handouts are described and attached.	No fields of interest are extracted from this section.

2.3.1 EXTRACTION OF SEMI-STRUCTURED FIELDS: METADATA

The Meeting Minutes template generally consists of three structured pages at the beginning of the document (Table 1):

1. A Document Information Page, comprising semi-structured fields with administrative information;
2. A Cover Letter from the Agency to the sponsor summarizing content, general discussion, or action items documented in the Meeting Minutes;
3. A Memo Page, detailing meta-information about the application, meeting details, and drug.

These pages were identified as pages of interest from which to extract metadata. Specific recurring fields from these pages were selected for automated text extraction.

2.3.1A PATTERN MATCHING METHODS: SEMI-STRUCTURED FIELDS

Meta-information from the documents was extracted using pattern matching methods in order to gather accurate metadata to attach to and augment searchability of Meeting Minutes documents in the database. The first iteration of pattern matching, Iteration 1, attempted to extract information based on field names identified from one sample document in the training set (n=23). This method assumed Meeting Minutes documents closely abided by the standard internal template (i.e. patterns generated based on one document template were expected to apply to others). However, this method was unsuccessful in capturing complete field-values, as conformance to the internal template formatting was manually evaluated and determined to vary widely.

It was noted that several versions of the standard Meeting Minutes template had been used among training set documents. A total of five different template versions were identified, even though the training set constituted only Meeting Minutes from 2015 to 2016. Version differences included changes in wording or phrasing, differences in naming conventions for fields of interest, and differences in the ordering of structured fields. These changes resulted in an inability for the initial approach in Iteration 1 to identify page information and fields of interest.

A second pattern list was generated from the training set based on mismatches discovered in Iteration 1. Fields that were not extracted or that were extracted incompletely in Iteration 1 were manually evaluated, collected, and added to a possibility patterns list so that synonymous or similar ways of describing a field of interest would be identified. In Iteration 2, a pattern matching algorithm written in Perl programming language was trained by manually inspecting cases where fields were not successfully extracted and adding those patterns to the possibility patterns lists. The trained algorithm recognized and identified possibility patterns for each field of interest. This algorithm was then tested on the test set of Meeting Minutes documents (n=23). Examples of anticipated pattern differences in fields of interest based on initial template analysis from Iteration 1 and a subset of possibility patterns by section for critical elements added in Iteration 2 are listed in Table 2. The Iteration 2 pattern matching algorithm was used to distinguish document section headers, allowing separation of the Document Information, Cover Letter, and Memo Pages from the remaining document.

Table 2: Iteration 2 pattern list for a subset of critical elements with additional possibility patterns generated from analysis of missing information from Iteration 1.

Template Field Name	**Field Type**	**Parent**	**Original Identified Patterns in Iteration 1**	**Possibility Patterns Added to Iteration 2**
DOCUMENT INFORMATION PAGE	Section header	---	DOCUMENT INFORMATION PAGE	---
Application Number	Subfield	Document Information Page	Application #(s): [*PIND*|*IND*|*NDA*|*BLA*####]	• Application: • Application #: • Application Number: • Application Numbers: • Application Number(s):
MEMORANDUM OF MEETING MINUTES	Section header	---	MEMORANDUM OF MEETING MINUTES	• AMENDED MEMORANDUM OF MEETING MINUTES • MEMORANDUM OF TELECONFERENCE MINUTES • MEMORANDUM OF TELECONFERENCE • PRELIMINARY MEETING COMMENTS

				• TO PRELIMINARY MEETING COMMENTS • MEETING COMMENTS
Product Name	Subfield	Memo Page	Product Name:	• Product: • Products: • Drug:
Sponsor/Applicant Name	Subfield	Memo Page	Sponsor/Applicant Name:	• Sponsor Name: • Sponsor/Applicant: • Sponsor:
Indication	Subfield	Memo Page	Indication:	• Proposed Indications: • Proposed Indication: • Indications:
Meeting Type	Subfield	Memo Page	Meeting Type:	• Meeting Type/Category:
Meeting Category	Subfield	Memo Page	Meeting Category:	• Meeting Type/Category:
Meeting Date and Time	Subfield	Memo Page	Meeting Date and Time:	• Meeting Date: • Teleconference Date and Time: • Teleconference Date: • T-Con Date:
Meeting Location	Subfield	Memo Page	Meeting Location:	• Location: • Meeting Format:

2.3.1B SENTENCE EXTRACTION: INDICATION

A specific field of interest in the document was Indication. This field conveys information regarding the intended use of the drug; it is important in searching across meeting minutes and interactions involving products for or related to the same indication. To facilitate grouping of Indications across documents, it is necessary both to extract Indication text from the document and to normalize Indications to a common ontology.

Because Indication is a free text field that may contain one or more sentences, sentences were extracted from the plain text files for next level entity extraction and normalization using the Perl module, Lingua::EN::Sentence [4]. In cases where the Indication contained more than one sentence or line, Lingua::EN::Sentence was used to split the text into individual sentences.

2.3.1C ENTITY EXTRACTION AND NORMALIZATION: INDICATION

After splitting into sentences, Indications were normalized to Disorder Concepts in the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) US Edition [5]. Because Indication is a free text field in the Meeting Minutes template, there was a need to normalize language for the same Indication into one standardized Concept.

A Stanford NLP Parser [6] module in Perl, Lingua::StanfordCoreNLP [7], was used to tag subjects, objects, and parts of speech from previously extracted Indication sentences. The parser was modified for sentence parsing and used to generate a constituent tree of tagged entities from extracted Indication sentences.

Example 1

Original text:

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

Stanford NLP Algorithm 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    (. .)))
```

Once parts of speech were tagged using the Stanford Parser, noun and verb phrases were isolated and retained, while other identified parts of speech were discarded.

Example 2

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)))
```

A SNOMED CT Disorder Normalized Dictionary, consisting of a list of Fully Specified Names (FSNs) for Disorder (Disease) Concepts and all respective Synonyms, was generated to map extracted Indication text to SNOMED CT FSNs. Perl Regex::PreSuf [8] expressions were used to create regular expressions from the Disorder Dictionary for pattern matching and for dictionary term matching.

To normalize Indication terms with SNOMED CT, each group of noun phrases (NP) and verb phrases (VP) identified in the constituent tree using the Stanford NLP Parser was searched against the Disorder Dictionary. A set of four string matching dictionary rules [9] was applied using Perl Regex to match the noun and verb phrases to Disorder Dictionary Concepts (Table 3).

Table 3: String matching dictionary rules initially applied.

Rule	**Example (preprocessing : postprocessing)**
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

Matching using the above rules was unsuccessful in mapping Indication entities to SNOMED CT Concepts. To further improve upon performance, additional string rules were implemented to account for the variable ways in which indications may be described (Table 4). Indications successfully mapped to SNOMED CT Concepts were also associated with their unique SNOMED Concept IDs [5].

Table 4: Additional dictionary rules implemented in addition to Table 3 string matching rules.

Rule	**Example (preprocessing : postprocessing)**
1. Replacement of synonym words: “disorder” to “disease” and vice versa	Sickle cell disease : Sickle cell disorder Bipolar disorder : Bipolar disease
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

2.3.2 EXTRACTION OF UNSTRUCTURED FIELDS: QUESTION AND ANSWER SECTION

The body of a Meeting Minutes document consists of a Q&A section that documents sponsor questions to FDA regarding a multitude of topics. The questions are generally directed toward a specific discipline and refer to current findings from the application or future plans from sponsors. Following a sponsor question, there is typically a documented response from the FDA. Meetings characteristically have multiple questions, and as such, these sections are usually recorded in question-and-answer format, with the FDA response to a sponsor question directly succeeding

each question. Topics discussed in the Q&A section are not fixed, but instead can range from sponsors seeking FDA guidance regarding submission guidelines to trial-specific questions regarding individual clinical trial development.

2.3.2A PATTERN MATCHING, SENTENCE EXTRACTION, AND ENTITY EXTRACTION: QUESTION AND ANSWER SECTION

In order to isolate the Q&A section, pattern matching was initiated using the Meeting Minutes template as a guide. A Perl algorithm was written to identify document content based solely on section number and title in the standard template, targeting sections 1.0, 2.0, and 4.0 (Table 1).

This pattern (Pattern List 1) was tested on training set documents but was unsuccessful in identifying questions and responses in many of the documents. Similar to the attempt to extract from structured fields, this failure was due to document formatting deviations from the standard template pattern. A second set of pattern lists (Pattern List 2) was generated by manually identifying information not captured using the Pattern List 1 algorithm after identifying causes of extraction failure. The lists in Pattern List 2 incorporated additional patterns to account for the variable nature of the body content and structure. These lists were primarily divided between Heading patterns and Middle Heading patterns. Heading patterns were generated from manual analyses of documents. Pattern List 2 constituted a separate possibility list of patterns for each categorization of interest for both Headings and Middle Headings.

Table 5: Pattern List 2 Heading Tag conditions. Heading types in the Q&A section and examples of possibility patterns for each type of Heading.

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

The Perl module, Lingua::EN::Sentence [4], was used to split the extracted plain text from its native paragraph form into individual sentences. Possibility pattern cues for Middle Headings were generated (Table 6).

Table 6: Pattern List 2 Middle Heading Tag conditions. Based on possibility pattern cues used to identify Q&A-relevant sections in Meeting Minutes documents that are not Headings (Table 5). Patterns matching those listed in the Possibility Pattern Cues column, as well as text immediately following these cues, were flagged as Q&A segments. Numbers represented in this table are for demonstrative purposes only, but may be any number in the actual documents.

Section (Middle Heading Tag)	**Possibility Pattern Cues (Middle Heading Patterns List)**
Question	Question 3: Question 3.0: Question 4a: Sponsor Question 1a: SPONSOR QUESTION 1: 2. Background:* 2.0 Background:* 2. Question [Sponsor Name] Question 3: [Reviewer Name] Question 13a: Q2: 1. Does the Agency agree th…?

	3) Does the Agency agree th…? b. Does the Agency agree th…?
Response	Response to Question: FDA Response: FDA Response to Question 14: FDA Response to Question 14b: FDA's Preliminary Response Sent on [date]: Sponsor's Clarifying Request for Question 1: [Sponsor Name]'s emailed response of [date]: Sponsor Response (via email [date]):
Discussion	Discussion: Meeting Discussion: Meeting Discussion for Question 6: Meeting Discussion on [date]: Discussion regarding Question 4:
Comments (Additional Comments)	Comments: Additional comments: Additional FDA comments: We have the following additional comments: Post-Meeting Comments for Question 1: General [Discipline] Comment:

*Sometimes, questions were not explicitly preceded by any information designating them as questions ("Question 3:"), so the section header, "Background," is the only text preceding these types of questions and therefore used to flag anteceding information as a question.

Using Pattern List 2 (Tables 5 and 6), the Perl pattern matching module first checked individual sentences within each document to determine whether the sentence contained patterns indicative of the presence of a Q&A section. Once affirmed, each sentence was tagged as a Heading, Middle Heading, Sentence Content, or Negative Heading based on tagging conditions (Table 7). In addition, Block of New Space Events were tagged.

Table 7: Tagging conditions for Q&A sentences.

Tag	**Tagging Conditions**	**Example**
Heading	First line after a Block of New Space event (double new line) that is 64 characters or less, and 1. The line is in all capital letters and cannot be found as a Heading in the Heading Pattern List (Table 5). -OR- 2. The text following the line can be identified in the pattern list as a Middle Heading (Table 6).	Any headers in the Heading Pattern List (Table 5). May be Section Headers, Category Headers, or Discipline Headers (discipline names).
Middle Heading	Fulfills conditions for Heading but can be found in the Middle Heading Pattern List (Table 6).	Any Middle Heading Pattern List possibility pattern cues (Table 6).
Sentence Content	1. Text following a Heading that is not a Middle Heading until the next Block of New Space event. -OR- 2. Text following a Middle Heading until the next Block of New Space event.	Content occurring after any Middle Heading Pattern List possibility pattern cues (Table 6), not including the cues themselves.
Block of New Space Event	Double new line event. Used to denote a line of blank space in the text, after which usually follows a Heading or Middle Heading.	---
Negative Heading	False positive preventative measure for Heading that triggers when something fulfills Heading tag requirements but contains words from the Negative Heading List.	Negative Heading List: • Table • Figure • [Month] (any month) • Slide • http://

Sentences tagged as Headings were further categorized as Section Headers, Category Headers, or Discipline Headers (Table 5). Tagged sentences were sorted into a spreadsheet with sentence content, tags, and document metadata in preparation for further analysis.

2.4 TOPIC MODELING

Topic modeling was explored to generate groups of topics with common regulatory themes. These themes were evaluated for usefulness for guidance development and in generating a regulatory lexicon.

2.4.1 DATA PREPARATION

A preprocessing procedure of transforming words into lower case, tokening, stemming and removing stop words was performed on the PDF documents that had been converted to plain text files. Further filtering of topic words was performed by a two-filter implementation of one start-word list and one stop-word list provided by domain experts. The first list, called *start-words*, contained specific words of interest around which domain experts were interested in building topics, such as safety and efficacy, labeling, guidance, datasets and data analysis, and adverse events. The other, called *stop-words*, specified frequently occurring words to be omitted. This word list included sponsor company names and words frequently used in template language or template field names. The first filter removed words found to occur in the *stop-words* list. The second filter removed words occurring less than 5 times (low threshold) or greater than 1,000 times (high threshold). After the preprocessing procedure, a vocabulary with 3,707 words was obtained from 230 documents.

2.4.2 TOPIC MODELING

For the processed document set, topic modeling with Latent Dirichlet Allocation (LDA) [10] was utilized to model the relationships between documents and words. In this study, the LDA program implemented in Mallet [11] was applied for topic modeling. In Mallet, Gibbs sampling [12], a special case of the MCMC approach, was utilized to calculate the two matrices. The number of iterations was set to 3,000 in Gibbs sampling and other parameters were set to default values in Mallet in all calculations. Two matrices were obtained using LDA. The first matrix, the document-topic matrix, described topic distribution in each document. The second matrix, a topic-word matrix, described word distribution for each topic. To test the best fitting model, the measurement perplexity [10] was used to determine the appropriate topic number for the document set.

2.4.3 TOPIC ANALYSIS

Two approaches were used to represent obtained topics: 1) creation of a list of top-5 most probable words in each topic and 2) generation of word clouds.

3. RESULTS AND DISCUSSION

3.1 PERFORMANCE EVALUATION OF SEMI-STRUCTURED FIELD-VALUE EXTRACTION

Applying the trained algorithm to the test set, field-values extracted from semi-structured fields were output in a CSV table. Due to the confidential nature of Meeting Minutes document content, this table of results cannot be publicly disseminated.

Semi-structured field-value extraction was evaluated for performance in both training and test sets. The algorithm extraction value was manually compared with the gold standard to determine True Positives (TPs), False Positives (FPs), True Negatives (TNs), and False Negatives (FNs).

$$\text{Accuracy,} \quad A = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision,} \quad P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall,} \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-score,} \quad F = \frac{2PR}{P + R}$$

Table 8 shows accuracy, precision, recall, and F-scores calculated based on the trained algorithm field-value extraction for the 23-test document set. Thirty-three field-values were extracted from each document, totaling 759 fields from each set of 23 test documents. Performance metrics from the 23-training document set after training the algorithm using domain expert feedback are also shown.

Table 8: 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	99.38	97.28	96.20	100	99.60	99.34	98.68
Precision %	100	99.66	99.44	98.33	100	100	99.86	99.45
Recall %	100	99.66	97.81	97.79	100	99.60	99.45	99.17
F-score %	100	99.66	98.62	98.06	100	99.80	99.66	99.31

The reported accuracy, precision, recall, and F-scores for the extraction of test set semi-structured field-values in the metadata pages of Meeting Minutes documents were above 90%. The 99.45% precision score on test documents suggests that a high percentage of returned field-values are correct. The 99.17% recall score on test documents suggests that the model returns a very high proportion of positives. Thus, the trained algorithm appears able to detect positives consistently and appears to have a high chance of a correctly retrieving the field-value, given its detection of a field-value.

An important consideration is that this extraction serves as the sole source of metadata for certain fields of interest crucial to linking of the Meeting Minute with other databases and documents. These fields include: Application Number, Meeting Type, Meeting Category, and Indication. Other fields, such as Letter Sponsor Attention Contact Name, Letter Full Address, and Letter Sponsor Name, may be of lesser importance. The fields of interest, which included Application Number, Meeting Type, Meeting Category, and Indication, had an accuracy of 100% in all training and test documents. FPs and FNs occurred most frequently for Letter Sponsor Attention Contact Name, Letter Full Address, and Letter Sponsor Name, and other fields less critical to understanding decision making and recommendations across indications, drug classes, meeting types, and products.

While the algorithm successfully extracted indications from the documents, evaluation of the text normalization is still ongoing at this time. Domain experts are in the process of manually reviewing indication mappings from the extracted indication fields to the SNOMED CT Disorder term FSN mappings to ensure that the matching of indication to normalized term is accurate.

3.2 PERFORMANCE EVALUATION OF QUESTION AND ANSWER SECTION EXTRACTION

The Q&A section extraction was originally intended as a proof of concept, due to the highly variable nature of document formatting around the Q&A sections. While the ability to easily sort and peruse Q&A content by discipline is invaluable, there was also a need to determine if extraction of the Q&A section would be a feasible endeavor with existing time and resource limitations.

While this possibility was being evaluated, resources dedicated to the domain expert were limited; thus, one document gold standard was generated and used to train the algorithm. With the learning that has occurred in generating Pattern List 2 and reevaluation of the potential success rate of this endeavor, domain experts are in the process of generating a gold standard using a different set of training documents, after which the model will be retrained using the new training set documents and evaluated for performance.

3.3 CHALLENGES

3.3.1 INFORMATION EXTRACTION CHALLENGES

Document archival in PDF format and versioning inconsistencies in templates used for each Meeting Minutes document led to difficulties extracting text based on template cues alone (Table 2). The FDA Meeting Minutes templates, in all iterations, contain structured fields that allow for free text editing of field-values, and in some documents, the field name itself was modified. For instance, in field-value extraction, “Indication:” would be rewritten as “Proposed Indication:” instead of following the template specification.

Similarly, in Q&A extraction, the template structure implied that the Q&A segment was present only under the “2.0 DISCUSSION” section and, when applicable the “4.0 ISSUES REQUIRING FURTHER DISCUSSION”; however, this was not consistent throughout analysis of the training set. Though some documents adhered to the standard template numbering and structuring (Table 1), other documents replaced section headers with free text, changed the numbering scheme, or even deleted the numbering scheme or section headers entirely.

Strict conformance to template guidelines was variable, causing extraction to identify fields incorrectly or miss them altogether when using the templated format as a guide. This was circumvented through the creation of possibility patterns lists using the standard template as a base and adding rule-based patterns during training (Table 2, Table 5, Table 6). We anticipate that training on a larger quantity of documents will continue to help refine the algorithm, as more possibilities patterns are collected and added to the patterns list.

3.3.2 INFORMATION NORMALIZATION CHALLENGES

Regex dictionary matching was initially unsuccessful, returning few matches and missing matches to many SNOMED Concepts. Further inspection of missing matches determined that syntax greatly impacted matching. Colons, hyphenations, and trademark symbols (which would convert from "™" in the original PDF to "TM" in plain text), especially, led to many false negative results during testing; therefore, string matching rules were applied (Table 3). Free text fields allowed for misspellings of Indications or other words. Lack of name standardization in clinical naming for some disease names also resulted in false negatives. There were many indications and diseases with several names, some of which were permutations of others. Additional rules were needed (Table 4) to account for nuanced differences in clinical terminologies to facilitate mapping of extracted indications to SNOMED CT Concepts.

4. CONCLUSION

We presented an NLP approach for extracting fields and field-values from irregular, semi-structured PDF documents and mapping indications to an existing ontology. The pipeline we have developed to extract fields and field-values from Meeting Minutes documents allows for extraction of metadata from PDF documents with high precision and recall. We plan to apply the trained algorithm to the remaining 184 documents in the pilot set to enhance our ruleset for atypically formatted Meeting Minutes documents and to refine our algorithm iteratively. The metadata collected from the algorithm output will be quality checked by domain experts and entered into a secure database that will allow for searching and filtering of text in any extracted field. After domain experts evaluate the SNOMED CT ontology mappings, the database will be linked to a local copy of SNOMED CT so that users can perform Indication searches augmented by SNOMED CT's synonym and relationship framework.

We have also provided a proof-of-concept approach that uses pattern matching and sentence tagging to consistently isolate a loosely-defined free-text Q&A section. We were able to develop our proof of concept for extracting the Q&A sections of Meeting Minutes using rule-based pattern matching. We plan to evaluate the Q&A sections from all 230 of our pilot pre-NDA/pre-BLA meeting documents to synthesize and incorporate the knowledge into standard pre-NDA and pre-BLA meeting communications. For the first time, reviewers and regulatory authorities will be able to sort Agency-sponsor questions and answers across Meeting Minutes by discipline and by relevant topic categories for downstream analyses. This will help regulators and sponsors to address frequently encountered issues in a proactive manner. Furthermore, this set of information, formerly only accessible through opening individual PDF documents, may ultimately be made more accessible after postprocessing and sorting into our Meeting Minutes search engine.

Topic modeling was used to explore topic subgroups of recurring regulatory terms around which a regulatory ontology may be constructed. Currently, there exists no ontology consisting of regulatory language, or language used to discuss the drug or biologic application process. This ontology will comprise language describing commonly searched topics in meeting minutes, such as safety, labeling, discipline, and application timeline information. Domain experts attempted to summarize common themes from frequent retrieval topics in the meeting minutes produced from topic modeling. Results using current parameters were inconclusive. However, topic modeling may be a viable discovery tool if applied to the Q&A section of the meeting minutes filtered by discipline. We believe that this approach of focusing on related topics within each review domain will yield topic groups that are more relevant to the domains.

REFERENCES

[1] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Tripple, E.D., Gutierrez, J.B., Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Computation and Language (cs.CL*). arXiv:1707.02919.

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

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

[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] 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

[6] Kumar V. (2011, March 6). NLP::StanfordParser. Retrieved from http://search.cpan.org/~vikas/NLP-Service-0.02/lib/NLP/StanfordParser.pm

[7] 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

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

[9] Subramani, S., Kalpana, R. and Natarajan, J. (2014). ProNormz--an integrated approach for human proteins and protein kinases normalization. *Journal of Biomedical Informatics*, *47*, 131-138.

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

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

[12] 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

Vahan Simonyan
HIVE Team
G-SRS Team

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Michelle Shen
FDA/CDER/OND
10903 New Hampshire Avenue
Silver Spring, MD 20993
(301) 796-5094
Michelle.Shen@fda.hhs.gov