Extracting Body Function from Clinical Text

Guy Divita[1], Jessica Lo[1], Chunxiao Zhou[1], Kathleen Coale[1] and Elizabeth Rasch[1]
[1] *Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, Maryland, USA*

Abstract
This paper describes finding Body Function (BF) mentions within clinical text. Body Function is noted in clinical documents to provide information on potential pathologies within underlying body systems or structures. BF mentions are embedded in highly formatted structures where the formats include implied scoping boundaries that confound existing NLP segmentation and document decomposition techniques. We have created two extraction systems: a dictionary lookup rule-based version, and a conditional random field (CRF) approach based on training from manual annotations. Training and test data utilized the NIH Clinical Center Rehabilitation Medicine Department records. Results of these systems provide a baseline for future work to improve document decomposition techniques.

Keywords
Natural Language Processing, Body Function, ICF

1. Introduction

Body functions are the physiological or psychological functions of body systems[1]. Body functions are mentioned in clinical text when there is concern for or documentation of pathologies around body function or body function assessment. Body Function information is commonly collected during physical exams to provide information on potential pathologies within underlying body systems or structures.

Our motivation came from a request from the Social Security Administration to retrieve BF mentions within their documents to support existing efforts to enhance their disability claims adjudication process. While there is a question around the utility of body function information as it relates to disability adjudications, we are motivated to work on this task as a mechanism to improve the algorithms that support BF extraction, namely sectionizing, sentence chunking, and context scoping annotators using BF mentions as the use case. BF mentions are often embedded in complex formatted text in the form of lists, slot-values, and oddly punctuated sentences in clinical notes. This paper reports on the systems developed to capture this information before making improvements to the document decomposition tasks.

Our conceptual framework for BF comes from the International Classification of Functioning, Disability and Health(ICF) [2]. While there are many specific kinds of body function, we set out to find mentions of *strength*, *range of motion (ROM)*, and *reflexes* because of their relevance to the current disability adjudication business process. Within these mentions, we label the body function **type** (strength, range of motion, reflex), the body **location**, and any associated **qualifiers**.

2. Prior Work

There is little prior work specifically extracting body function from clinical notes. Some work has been done extracting other ICF defined areas using traditional rule-based techniques as well as deep learning methods. Kukafka, Bales, Burkhardt and Friedman report on modifying MedLEE to automatically identify five ICF codes from Rehab Discharge summaries[3]. Newman-Griffis and Fosler-Lussier describe linking physical activity reports to ICF codes using more recent language models and embeddings[4].

The NLP platform employed for this work was adapted from the V3NLP Framework[5] and Sophia[6] which were used for symptom extraction and finding mentions of sexual trauma in veteran

clinical notes. The framework employed is built upon UIMA[7], so resembles the cTAKES[8] system closely, but has a pedigree from UMLS concept extraction in biomedical literature (MetaMap)[9].

3. Corpus and Manual Annotations

The NIH Clinical Center Rehabilitation Medicine notes, from which our corpus was drawn, are indicative of any hospital's rehab notes, in that, the services provided are in support of patients in need of rehab. While the document formatting is idiosyncratic, the terminology is in line with what is being seen in SSA claimant data, which is composed from a national pool of clinical records from an extremely heterogeneous set of providers.

3.1. Manual Annotations

Table 1
Distribution of Manual Annotations in NIH Records

Annotation Type	Training	Testing
BF Mention	1014	403
Strength	917	417
ROM	482	221
Reflex	32	8
Body Location	1001	361
Qualifiers	1014	660

3.2. Annotation Guidelines

A body function mention is identified when there is mention of body function type, and a qualifier, and optionally, one or more body locations within the scope of a phrase or sentence. These mentions are only annotated from objective (clinician-observed) information.

Laterality and similar modifiers are to be lumped with body location as body function locations are typically modified with descriptors such as left, right, both, proximal, and distal. There were exceptions, where a mention is made without one of the necessary components or where those components are inferred, when it is thought that some body function type information is indicated. For example, the mention *Neurological: Negative* is marked. Neurological here infers a location and a body function type.

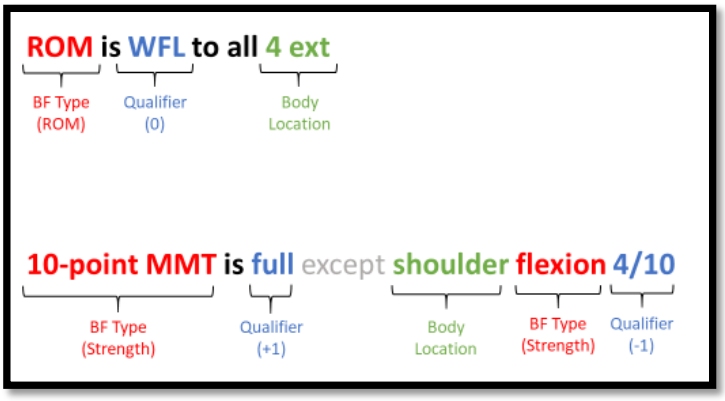

Figure 1: Body Function Annotation Examples

3.2.1. Qualifiers and Polarity

Body Function qualifiers include terms like *positive*, *improved*, as well as values from test scores such as degrees of range or values from test scores such as the (clinically applied) manual muscle tests which look like 5/5 or 9/10. Each qualification is classified into below-level function (-1), ambiguous (0), or at- or above- level function (+1). As a side note, strength values between 0/5 and 4+/5 are given a -1 qualifier as they are *less than normal* as quantified on the manual muscle test scale. Only a 5/5 would be marked as +1.

3.2.2. Underspecified and Ambiguous Mentions

There are body function mentions found in clinical text that say something about body function incompletely. This comes in at least three varieties. A mention like *raise arm overhead* is either strength or range of motion but is not more specific. A mention that is qualified as *assessed* is not sufficient to assign a polarity of 1 or -1.

3.2.3. Implicit Semantics

There are mentions that contain implicit body function type and implicit body location. For example, *grip strength* implicitly indicates the hand as a body location. A mention like *a fist can be made with both hands* implicitly indicates strength and/or range of motion.

3.2.4. Interrater Reliability

This initial set of records was annotated by a fellow with some medical training guided by domain experts (ER, KC). While the body function annotation task included two annotators, only one annotator worked on this NIH corpus. Late in process, a second annotator was trained and interrater reliability endeavors were done using NIH data. The interrater reliability (F1) scores across all the labels between the two annotators macro and micro scores on the NIH corpus were .70. For this work, the only useful take-away is that two annotators can comparably continue to annotate NIH data, and that the task is not too complex.

4. Methods

We have created two systems, a dictionary lookup rule-based version and a conditional random field (CRF) approach based on training from manual annotations. Neither of these approaches use an underlying, pre-built language model. The assumption is that the verbiage and context around body function mentions are in very different contexts than how they would appear in any of the pre-built language models available to us.

4.1. Rule Based, Dictionary Lookup System

This system is constructed from a V3NLP Framework, a UIMA NLP suite of annotators, pipeline connection utilities and readers and writers. The evolution of V3NLP Framework is now branded Framework Legacy. The pipeline created stitched together (mechanical annotators) to decompose the clinical text into its constituent parts, including sections, sentences, phrases, tokens and dictionary looked up terms. The intelligence of the system uses a dictionary lookup annotator which relies upon lexicons. It is worth sharing the pedigree of each of those lexica.

4.1.1. Lexica

A separate lexicon (one for each) was created for Strength, ROM, reflex, body location, qualifier terms, as well as pain, balance, and coordination.

For most of these, a top level (seed) term was identified in the UMLS. All semantically decedent UMLS terms were extracted from the UMLS and added to each respective lexicon. The resulting terms were fed through a lexical variant generation tool (LVG)[10] to create fruitful variants[11]. Each of these extracted and generated terms were labeled with Strength, ROM, Body Location and such. Metadata including the UMLS identifiers and UMLS semantic types were retained for pedigree sake. When a term is found in the text, a *lexicalElement* annotation is created and tagged with one or more categories (Strength, ROM, Body Location, etc.) and the metadata garnered from the UMLS.

4.1.1.1. Body Location Lexicon

The SNOMED terminology has a modifier (Body Structure) attached to a number of their terms. All of these body structure tagged terms were extracted, then manually culled to remove those terms which would not be relevant. These culled terms involved terms with *cell* and *cell structure*, *cardiac*, *vein*. This yielded 53,641 terms with UMLS concept identifiers. Thirty-six body laterality terms including *left-sided*, *proximal* and *distal* were manually added to cover parts of body location expressions in the text. An additional 24 terms were added to cover body location expressions found in the training set. These were mostly abbreviations like *r le* and a few more colloquial terms like *core* and *quad*. There are 53,704 terms total in the body location lexicon.

4.1.1.2. Body Strength Lexicon

The bulk of Body Strength terms gathered from the UMLS came from terms with the token *strength* in them. There are 4739 such terms. Some of these, admittedly are overly broad, for example having to do with the strength of contractions, and the strength of medication. A number of these were manually filtered out. Sixty-two terms were added from expressions seen in the text, not otherwise found. These were mostly in the form *[body location] [extensor|extensors|extension|extensions|ext]*. Note that 34 terms having to do with *muscle weakness* were included as part of the body strength lexicon. There were a total of 4802 body strength terms.

4.1.1.3. Range of Motion Lexicon

Descendent terms of Range of Motion were gathered from SNOMED-CT[12] (screen scraping from the SNOMED CT Browser[13][14]). These were augmented from terms in the UMLS with range of motion, extension, flexion as part of the term. It should be noted that a number of these terms came from MEDCIN[15] in particular. While most of the terms came from SNOMED, LOINC[16], the National Cancer Institute Thesaurus[17], Ontology of Consumer Health Vocabulary (OCHY)[18], along with MeSH[19], and ICD-10CM[20] had some coverage. Sixteen additional terms were added to cover range of motion expressions found in the training text. There are 793 range of motion terms in the range of motion lexicon.

4.1.1.4. Body Function Qualifiers Lexicon

Many body function qualifiers are numeric and are covered by regular expression mechanisms to identify units of measure. To this end, a lexicon of units of measure is being used to identify the units. That lexicon is derived from the Unified Code for Units of Measure (UCUM) provided by the National Library of Medicine. This resource was altered for the body function task. All single letter units were commented out, because they were causing too many false positives. In addition, the terms *feet* and *foot* and *field* were likewise commented out. The UCUM lexicon includes 946 entries.

A lexicon needed to be gathered to cover the non-numeric qualifiers. 9755 terms which had a semantic type of *Qualifier* were taken from the UMLS. This was augmented by terms that descended from concepts *weakness*, *observation of reflex*, and *hyperflexia*. In all, 2807 concepts from the UMLS were gathered. 104 terms were added to this resource to cover terms in the training text that were not already known as a body qualifier. The body qualifier lexicon had 2910 terms.

4.1.1.5. Non Body Function Lexicon

It was useful to gather terms that when identified, would rule out a body function expression. These were labeled as *Confounding Terms*. Top among these terms was *NIHFA score* and *MRI*. There were 17 such terms added. An additional 34 terms were added, without any semantic label, to chunk up the text to prevent fallacious qualifiers, mostly around temporal entities. Among these were terms like *alert and oriented x 3* and *age-matched norms*. There were 32 such terms added. A small lexicon of 62 pain terms that indicate pain of some sort was gathered. One term in this lexicon turns out to be ambiguous with a body strength term *pinching*.

A number of qualifiers were erroneously identified because they were within expressions that also included body location and sometimes strength but were referring to balance and coordination. A small lexicon of 13 balance and coordination terms were gathered to identify balance and coordination terms rather than strength or range of motion terms to combat these errant qualifiers to block them from becoming part of a strength, rom, or reflex mention.

4.1.2. Preprocessing Annotators

The redaction done on this data was overly aggressive. Templated redacted forms were found within section names and body function mentions as well as other locations. An annotator was created to label redacted forms, to allow those spans of text to become invisible and are temporarily removed for downstream annotators.

A particularity of this data set is that redactions are in Section Names such as *History of [First Name id = XXXXX] Illness:* . Taking out the redaction enabled many sections to be correctly identified.

The dataset we work with has a particular idiosyncrasy: it contains no newlines. While not usually worth noting this level of data quality control and normalization, it is noted here because our other datasets that come from a variety of providers from around the world also occasionally include documents that have no newlines.

An annotator was created to infer when there were no newlines in a document and inject newlines around section names from a rough lookup of section names and simplistic regular expressions. This aided in subsequent section boundary and section name identification because the existing sectionizing mechanism requires newlines to be present.

4.1.3. Body Function Pipeline

The body function pipeline's purpose is to identify BF mentions. That is, an utterance that includes a body location, a body function type (such as strength, range of motion or reflex) along with some kind of qualifier related to the body function type.

The body function pipeline has been appended to the pipeline that decomposes the text into sections, sentences, slot:values, lists, phrases, terms, and tokens (see appendix for details). The body function pipeline relies upon having terms in the document already looked up and classified prior to the next set of annotators and knowing what sections those terms occurred in.

The guidelines and subsequent manual annotations created a Body Function Type label, with a type attribute which has an enumerated value as one of *Body Strength, Range of Motion* or *Reflex*. For the convenience of building the tool from existing components, those attributes were turned into labels.

The guidelines indicate some sections to ignore. These include Goals, Plan, Education, Family History, Medications, Referrals, Interventions, Gait, Balance, Coordination, Mobility, Motor learning, Motor Function, Follow-up and Recommendation sections. While Balance Coordination are body

function, they are not included as the initial ones (strength, rom, reflex) we are addressing. One oddity: there were several mentions in the training set that came from a common section labeled *Impressions and Plan*. Impressions and plan sections were not filtered out.

The Body Function Location, Strength, Range of Motion and Reflex Annotators each create their respective annotations from terms noted to have those categories as attributes from the upstream term lookup step. Annotations were not made from sections that were specifically noted to be ignored and annotations were not made from mentions that were not about the patient. As noted above, all terms have as an attribute the section they are within and if the term refers to the patient or not.

4.1.4. Body Function Qualifiers

Finding BF qualifiers is more complicated. Sometimes there is both a strength and pain mention in the same sentence where the qualifier is really about the pain, not the strength. Although less frequent, mentions about coordination and balance were found with strength and range of motion mentions in the same statement and the qualifiers were not about the body function type we were looking for.

To thwart these confusions, mentions that were categorized as *pain*, *coordination*, or *sensation* when found within a window of six tokens of the other body function type kind of mentions would inhibit the creation of a qualifier. To this end, a small lexicon of pain and coordination terms was created to support this. While this works well, it is noted that some terms such as *pinch* were found to be about pain or strength depending upon context beyond the scope of this task.

There were a number of qualifier candidates that occurred in statements that had mentions of a body function type and a body location, but the qualifiers were not about the body function type. Common among these were mentions of patient ages and dates. There were a number of confounding terms also found to be in the vicinity of BF type mentions that when seen, would indicate the qualifier would not be attributed to the BF type. A lexicon of such confounding terms was created and used. Such terms include *fine motor activities* and *MRI*.

Scoping rules are a common theme to NLP, making it important to accurately attribute the scope of the section where a mention is found as well as the associated sentence or slot:value. There are a number of cases where the text is not pristine sentences, lists, or slot:values. Occasionally there were texts where there were no sentence breaks or multiple colons causing the sentence scoping to go awry. There were a number of these cases where no qualifier was found for a body function type within scope. In these cases, a modification was added to the scope of where to look for a qualifier candidate when no qualifier could be found within a sentence. Looking to the right by 266 characters (empirically set) to find a qualifier for a body function type improved performance.

4.2. CRF Model

Stanford's Named Entity Recognizer[21], which relies upon an underlying Conditional Random Field (CRF)[22] statistical machine learning modeling algorithm has been chosen as the machine learning approach for us to start with.

A UIMA based NLP pipeline was created to chunk the text into tokens. Those tokens that came from manually annotated body function mentions were marked. All tokens were used as the fodder for the CRF model. The tokens were classified in the BIO fashion. Those tokens that were part of potential body function mentions were marked where those tokens that began a mention were marked with Begin, the middle tokens were marked with Inside, and all the rest of the tokens were marked with an Outside classification. In addition to the BIO, all permutations of the body function classes and BIO were used. For example, Begin-Body-Function-Type-Strength, Inside-Body-Function-Type-Strength, Begin-Body-Function-Location, Inside-Body-Function-Location.

5. Results

5.1. Rule-based System: Token Based Matching Criteria

Table 3

Rule-based System: Token-based Body Function Evaluation

Label	F-1 Score	Recall	Precision
BF Mention	.6125	.9452	.4532
Qualifiers	.5699	.8593	.4263
Type	.6378	.8888	.4974
Body Location	.4696	.8287	.3276

5.2. CRF-based System: Token Based Matching Criteria

Table 4

Label	F-1 Score	Recall	Precision
BF Mention	.6578	.8376	.5415
Qualifiers	.6586	.8442	.5399
Type	.7287	.8657	.6291
Body Location	.5776	.7548	.4677

6. Rule-based System: Failure Analysis

Particular attention is being paid here to the qualifiers because it is the lynch pin to creating mentions for the most part. The rest of this section has to do with failures with qualifiers.

6.1.1. False Negatives

The most prevalent term missed was *weakness*. There were 34 cases where *weakness* was correctly identified, but 93 cases where identifying *weakness* was a false positive. In the cases where weakness was missed, five of the seven cases also involved confounding mentions related to balance, coordination, and pain. One was a scoping case where a list of test results followed *no weakness identified on R side of body*. As it turns out *no weakness* is in the Body Qualifier lexicon as a qualifier, but not also tagged as *strength* as it should have been. One case of weakness incorrectly attributed to a patient mention was triggered by the word *note* in the sentence.

6.1.2. False Positives

As mentioned above, there were 93 cases involving *weakness*. The majority of these turned out to be within statements the patient made, either within chief complaints, or *patient reports weakness*, or within quoted expressions. While an annotator was created to assign attribution of who authored the statement and mentions were marked, the rule to filter these patient attributed mentions was not working. Scoping issues arose, particularly with scoping in or out section names. Thirty-two cases were caused by scoping, where a series of values either delimited by colons, semi-colons or periods limited the scope of what those numbers were referring to. For example, for *... 3 trials : Right : 60 , 60 , 62 Left: 60, 43, 50 Gauge was measured in ....* where it was missed that the scope of these were grip strength measurements.

7. Discussion

It is noted that the most challenging part of this task is identifying the qualifiers. This is where most of the formatting and therefore scoping challenges arose.

This initial work has led to the guidelines being altered for body function type categorization going forward based on discussions of how to handle ambiguous statements that, for example can refer to both strength and range of motion.

The initial guidelines included marking section names as part of a mention, as when the section name was *Strength:*. After some discussion, we have come up with a new annotation type: *Relevant Context* to cover elements like section names which set the semantic context of mentions to come, but, themselves are not mentions.

The timing of this task was such that this set of 500 was annotated by one annotator before a second annotator came aboard. The inter-rater reliability was done to insure, going forward, that the annotators are consistent.

The CRF results are acknowledged to be better than the rule based results but provide little insight into the task. The CRF modeling also has provided challenges where there are limitations to how many labels can be modeled given our current computing resources. The rule-based version will continue to provide a baseline to benchmark gains to the system due to alterations in the document decomposition tasks for scoping contexts to sentences, slot:values, tables, and sections.

8. Future Work

There are a number of next steps for this work. The first is to model the qualifier attributes (-1,0,1). We will loop back and alter annotations within this set to accommodate changes to the guidelines and re-run.

Strength, ROM, and reflexes are not the only body function types that can be retrieved. This work can be expanded to include balance and coordination information.

We intend to release the body function software after work to improve document decomposition techniques has improved the performance of this baseline.

9. Conclusions

We describe in this paper a rule based extraction tool developed to find body function mentions that include strength, range of motion, and reflex. We developed this work learning from NIH Clinical Center Rehabilitation Medicine notes and are adapting it to find body function mentions in SSA claimant records.

10. Acknowledgements

We would like to thank Rafael Jimenez Silva for staging the annotation tasks, ensuring an equitable distribution of records across training and testing, doing the IIR work and providing corpus and quality assurance statistics.

Supported by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration.

Appendix: Syntatic Pipeline Defined

This body of work relies on an NLP UIMA based pipeline that decomposes text into its constituent parts. This pipeline is substantially similar to what was used in the Sophia pipeline, but is outlined here because some of the components have been added.

The Syntactic Pipeline

For the most part, the annotators listed here do obvious tasks that need no further explanation. There are exceptions and white lies of course, which will be noted for the seemingly mundane tasks for Tokenization, Sentence Chunking, and Date and Time identification. As it turns out, within the richly heterogenous data we are processing, those tasks are not as straightforward and error-free as is ultimately needed.

a. Line Annotator with Blank Lines

This annotator creates annotations for each line in a document. It does not strip empty lines out. Having line annotations enables an algorithm to walk through lines of text. Multiple blank lines indicate a topic shift. Thus, one needs to keep track of those kinds of lines, rather than filter them out, when looking for paragraph breaks. This annotator does not work well when there are no newlines in text, as is the case for the BTRIS data we are using. Special ameliorations are needed for such data.

b. RegEx Shape Annotator

The regular expression shape annotator creates annotations for emails, phone, URLs, zip codes and common redaction artifacts found in clinical text. Shapes are pseudo lexical entities, have meaning, but would not normally be looked up in a dictionary, which would distinguish them from lexical entities that come from a dictionary lookup. This annotator Identifies the easy things you don't want which makes the task of identifying things you do want easier. Identifying these entities makes sure that downstream annotators do not erroneously pick up entities that are these.

c. Date and Time Annotator

This annotator identifies dates and times via regular expressions.

d. Token Annotator

This annotator chunks the text into space delimited units. It creates word tokens and white space tokens. The tokenizer used here also creates attributes describing if the token has punctuation, is only punctuation, has numbers, is only numbers, starts with upper case, is camel case, and ends with sentence ending punctuation.

A technical note: this tokenizer is the V3NLP Framework Tokenizer, a tokenizer tuned for clinical text that has a legacy from MMTx and MetaMap.

Tokenizers play an unspoken, but big role in errors downstream, and no tokenizer does a perfect job with clinical text. This tokenizer informally compared to the python based *scispacy* language model driven tokenizer. Both tokenizers had failures with different difficult to parse texts, with neither exhibiting brilliance, one way or another. As a consequence, this legacy version of the tokenizer continues to be used, in great part because it is much faster and has a much smaller memory footprint than the wrapped *scispacy* tokenizer.

e. Date by Lookup Annotator

This annotator identifies parts of temporal expressions by items listed in a date lexicon as being a date. These include the obvious – names of the months and days and holiday names.

f. Date and Time by Token Annotator

There are oddball dates that get missed by the regular expression annotator before tokenization. This annotator identifies dates that bounded by each token.

g. Checkbox Annotator

This annotator identifies and analyzes mentions like *Smoking: yes [ ] no [x]*. The annotator identifies the heading, each of the options, and which option was marked. It identifies whether the options have a positive or negative polarity to them. If so, it takes the polarity of the marked option and applies that polarity to the heading. In this example, *smoking* gets negated because the no box was marked, noting that *no* has a negative polarity.
[Note: This annotator was turned off for this work partly because the BTRIS data did not have checkbox mentions that were relevant to the task.]

h. Slot:Value Annotator

The slot:value annotator identifies and analyzes slot and value entities into a content heading entity and an answer entity. Example: *Denies Alcohol*: *yes.*

Slot:value entities are telegraphic sentences which lack an explicit verb. They are quick methods of data capture and easy comprehension but do not syntactically parse in the same way sentences in prose do.

There are a lot of variations to slot: value formats within clinical text in general, and within the BTRIS dataset. Getting this structure correct is paramount. However, there are many ambiguous examples which flummox the current iteration of this annotator.

i. Sentence Chunker

This annotator identifies sentences within the text. Embedded within this task, are also the identification of lists and list elements. Like the slot:value annotator, correctly identifying the bounds of when a sentence begins and ends is paramount. The variation of text found in clinical text have flummoxed all the sentence chunkers tried thus far. None have worked 100% of the time. Many of the downstream errors are attributed to sentence chunking failures.

j. Term Annotator

This annotator chunks together tokens into terms based on dictionary lookup. Categorization and syntactic information from the dictionary are tagged onto the terms created.

The UMLS SPECIALIST Lexicon, by default, is employed to chunk general English into terms. There are annotator specific lexica also employed, including a date lexicon, a lexicon of section names, a lexicon of assertion terms. Most of the pipelines employ 20 lexica of one kind or another.

k. Assertion Evidence Annotator

This is one of two annotators that work in conjunction with each other. This annotator identifies evidence for negation, conditional statements, hypothetical statements, whether the mention is about the patient (subject), whether the mention is historical, and who is saying the mention.

The algorithm employed is a re-write of Wendy Chapman's ConTEXT algorithm in java. The Lexica came from her rules, and greatly augmented from work done by three groups at the University of Utah combining each group's rules.

Who is saying the mention (attribution) is the newest extension to this algorithm and was done for this project. The annotation guidelines stipulated to ignore patient authored statements, thus, the need to identify who is saying what. While it is not completely straight forward to identify patient reported mentions, there are clues or evidence, including trigger statements such as "patient reports", and patient notes". Also, any mentions that come from the subjective portion of SOAP notes are a-priori ruled as patient reported. The rules used for this work were adopted from work done to determine the difference between a sign vs a symptom and work done to determine if the statement is about the patient vs someone else.

Spoiler alert: the second annotator, the assertion annotator, is much further downstream in the pipeline.

l. Unit-of-Measure Annotator

This annotator identifies things that are measured, are like terms, but not something to be looked up. These include numeric test results, pulse rates, ejection fractions, or degrees of range of motion.

This annotator employs, for the most part, a combination of dictionary lookup for the units part, and regular expression for the numeric parts. The dictionary used for this is a snapshot of NLM's UCUM resource. Not perfect, but useful.

m. Term Shapes Annotator

This annotator identifies spelled out numbers and units of measure ranges.

n. Punctuation Terms Annotator

This is a corrective annotator: it creates terms that are only punctuation like +++. The current lexical lookup ignores runs of only punctuation, thus making it impossible to create terms that are only punctuation. There are many test results that are only punctuation. This annotator was created specifically for this task to pick up such entities.

o. Person Tokens Annotator

This annotator identifies persons in the text.

Note: The BTRIS data has persons already redacted, so this annotator is not useful currently and was turned off for this work.

p. Slot:Value Repairs

There are various failings of the current slot:value annotator that these corrective annotators fix, using downstream annotations not available to the slot:value annotator when it runs in the sequence in the pipeline. This annotator fixes some of the failures that are fixable.

q. CCDA Section Header Annotator

This annotator creates section headers based, for the most part, on dictionary lookup. The annotator uses an augmented version of HL7's list of approved section headings. The list was augmented a lot for this task because OT/PT specific sections do not appear within the CCDA domain (yet).

r. CCDA Panel Section Header Annotator

Panels are sections within clinical documents that list test results for blood tests, primarily.
This annotator creates headers for panel sections.
Note: Panels are ignored for this work, and this annotator is turned off.

s. CCDA Section Annotator

This annotator creates section zones from the end of the section name down to just before the beginning of the next section name.

t. Sentence Section Repair

This is a corrective annotator. Once section headings are determined, there is need to adjust (erroneous) sentence boundaries to exclude section names.

u. Quoted Utterance Annotator

This annotator creates quoted text. Symptoms are typically found in "quoted text", so it's useful to find them.
Note: Quoted text does not play a role in the Body Function task and though it is on, this feature is not used downstream.

v. Sentence Repairs

This is a corrective annotator. This annotator removes lists that only have one element to them and turns those back into sentences. Sentences that end with a number also caused issues because the numbers look like list delimiters. So lists that have list delimiters like "1. 2." that have the list delimiter ordering out of order are likely not lists, but sentences that end with numbers. Sentences that have tabs in them are likely to be from multi-column formats, where, within the process of OCRing them, the OCR software injected tabs to indicate a new column.

w. Assertion Annotator

This annotator, part two of the two assertion annotators, creates assertion attributes to all annotations based on the assertion evidence noted from the assertion evidence annotator.

x. Section Name in Terms Attribute Annotator

It is useful to know what section a term is mentioned in. This is useful to filter out mentions found that come from sections you do not care about. This annotator adds the section name to each term in the document. This is done outside the term annotator, which happens before the section zones are computed

References

[1] NIH SEER Training Modules, Anatomy & Physiology, Intro to the Human Body, Body Functions & Life Processes: training.seer.cancer.gov/anatomy/body/functions.html (last accessed 2021/05/06)

[2] Üstün, T. Bedirhan, et al. "The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health." *Disability and rehabilitation* 25.11-12 (2003): 565-571.

[3] Kukafka, Rita, et al. "Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health." *Journal of the American Medical Informatics Association* 13.5 (2006): 508-515.

[4] Newman-Griffis, Denis, and Eric Fosler-Lussier. "Automated Coding of Under-Studied Medical Concept Domains: Linking Physical Activity Reports to the International Classification of Functioning, Disability, and Health." *Frontiers in digital health* 3 (2021): 24.

[5] Divita, Guy, et al. "v3NLP Framework: tools to build applications for extracting concepts from clinical text." *eGEMs* 4.3 (2016).

[6] Divita, Guy, et al. "Sophia: A expedient UMLS concept extraction annotator." *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, 2014.

[7] Ferrucci, David, and Adam Lally. "UIMA: an architectural approach to unstructured information processing in the corporate research environment." *Natural Language Engineering* (2004): 1-26.

[8] Savova, Guergana K., et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *Journal of the American Medical Informatics Association* 17.5 (2010): 507-513.

[9] Aronson, Alan R. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.

[10] McCray, Alexa T., Suresh Srinivasan, and Allen C. Browne. "Lexical methods for managing variation in biomedical terminologies." *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1994.

[11] Lexical Variant Generation Documentation: Fruitful Variants. lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/UDF/flow/fG.html. (last accessed 2021/05/06)

[12] Donnelly, Kevin. "SNOMED-CT: The advanced terminology and coding system for eHealth." *Studies in health technology and informatics* 121 (2006): 279.

[13] Richesson, Rachel, et al. "A web-based SNOMED CT browser: distributed and real-time use of SNOMED CT during the clinical research process." *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, 2007.

[14] SNOMED-CT Online Browser, browser.ihtsdotools.org/ (last accessed 2021/05/06)

[15] Goltra, Peter S. *MEDCIN: a new nomenclature for clinical medicine*. Springer Science & Business Media, 2012.

[16] McDonald, Clement J., et al. "LOINC, a universal standard for identifying laboratory observations: a 5-year update." *Clinical chemistry* 49.4 (2003): 624-633.

[17] Golbeck, Jennifer, et al. "The National Cancer Institute's thesaurus and ontology." *Journal of Web Semantics First Look 1_1_4* (2003).

[18] Amith, Muhammad, et al. "Ontology of Consumer Health Vocabulary: providing a formal and interoperable semantic resource for linking lay language and medical terminology." *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019.

[19] Lipscomb, Carolyn E. "Medical subject headings (MeSH)." *Bulletin of the Medical Library Association* 88.3 (2000): 265.

[20] Cartwright, Donna J. "ICD-9-CM to ICD-10-CM codes: what? why? how?." (2013): 588-592.

[21]Finkel, Jenny Rose, Trond Grenager, and Christopher D. Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 2005.
[22]Wallach, Hanna M. "Conditional random fields: An introduction." *Technical Reports (CIS)* (2004): 22.