NLP OF RADIOLOGY REPORTS (PAD term spotter)

Summary:

Pipeline assesses presence of phrases indicative of peripheral arterial disease (PAD) in one or more sentences contained in radiology related documents.

Overview:

The 'PAD term spotter' pipeline processes radiology note textual extractions specifically pertaining to the diagnosis, treatment, etc. of lower limb artery stenosis/occlusion. The main feature is classifying each document for the presence of PAD. Descriptive text of diagnosis's and illness terms are paired with the site designated terms to build a relational tie, indicating a 'hit'.

Background:

The system was trained with 455 documents that were manually abstracted to establish a gold standard and achieved about a 90% accuracy using four different classifications of PAD (unknown, probable, PAD yes, PAD no).

Following is represents the algorithm used by Mayo implementation of the pipeline post processing (see the customization example in section below – additionally, a more detailed explanation of the post processing algorithm is provided below):

If there is no related exam type
OR no positive or negative evidence,
classify as UNKNOWN;
If there is positive evidence,
classify as POS;
If there is explicit evidence of negation of positive evidence
OR if a lower solo extremity exam with the discovery of no stenosis, probable or negated related PAD term exists, classify as NEG;
If there is no positive or negative evidence
OR if there is negation of positive evidence in an Ultrasound or Vascular Interventional Radiology report,
classify as UNKNOWN;
Otherwise,
classify as PROB based on severity modifier;

The best performance has been achieved by combining the 'unknown' and 'PAD no' into one 'unknown' category at the consumption time post analysis to provide per patient classification (more details under customized example section below).

cTAKES requires Java version 1.5 (aka Java 5) or later and Apache UIMA 2.2.x (https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP_Documentation_and_Downloads).

PAD Term dictionaries:

The terms being looked up have been added to dictionary e.g:
"PAD term spotter\resources\lookup\radiology\pad_anatomical_sites.csv"
Using format as follows:
"first-word|first-word plus terms|value"
'value' indicates how the term should be used when paired with the paired relational term, a context value, or some other special cases which will be discussed in more detail below. For example, the entry "common|common iliac artery|6" would be a valid entry in the 'pad_anatomical_sites.csv' file. This entry would be used to pair up with terms contained in the sister term dictionary, 'pad_disorders.csv' file.

Keep in mind that the 'first-word' term can be hyphenated, but that term must exist in the "core\resources\tokenizer\hyphenated.txt" file contained in the 'core' project (Note if the hyphenated version didn't exist then the above example should be expressed as "first|first-word plus terms|value").

Dictionary values:

The primary use of the value contained in each record within the term dictionaries is to discern whether the entity belongs to the anatomical site (6) or disorder (2) collection. The related annotated named entity discovered within the analyzed text will be tagged accordingly with the 'typeID' field. It can be used to filter terms by utilizing the 'ANN_PART_ONE_TYPES_TO_IGNORE' and 'ANN_PART_TWO_TYPES_TO_IGNORE' specified in the 'PAD_Hit.xml' annotator or other special handling.

The values coded for specific function include:

For the 'pad_anatomical_sites.csv':

- '6' - this is the default value for the anatomical site terms.
- 'STAND_ALONE' - this value indicates that the term should be able to be used as a 'hit' even if there is no corresponding disorder term found. For example, the term 'femoral-femoral artery' is strong enough evidence that a PAD related diagnosis is being discussed in the text. Therefore, the dictionary will contain the field: 'femoral-femoral|femoral-femoral artery|STAND_ALONE' to handle this case.
- '9' - this value acts as a negation case to eliminate those mentions that contain this term in the phrase. For instance, the dictionary contains the field: 'vein|vein|9', so that a phrase such as "iliac vein" would not be considered for a relational pairing, since PAD only involves artery disorders.

For the 'pad_disorder.csv':

- '2' - this is the default value for the disorder terms.
- "STAND_ALONE' - this indicates that the term should be able to be used as a 'hit' even if there is no corresponding site term found. For example, the term ' patent graft' is strong enough evidence that a PAD related diagnosis is being discussed in the text. Therefore, the dictionary will contain the field: 'patent| patent graft|STAND_ALONE' to handle this case.
- '7' - this value is for terms that should not indicate a 'no PAD' situation when certain context terms are applied (more discussion concerning context terms below). For instance, the term 'patent' can be an indication of a relatively health artery, since the term describes a clear or non-blocked situation. However, having a graft that is patent would not necessary indicate a case where PAD doesn't exist, especially, since the presence of a graft, in itself, may be an indication that there was a problem with the artery that needed to be circumvented.
- '8' - this value is related to terms which contain the term 'stenosis' or 'stenoses' with a descriptive term that indicates more than a moderate case. So pairing with terms such as 'moderate' or 'severe' are a strong indication a possible case. This is especially helpful in cases where a tie breaker is needed due to contradictory evidence in the record.

Dictionary customization:

As mentioned above the dictionary terms for this project are provided in flat text files in 'PAD term spotter\resources\lookup\radiology\pad_*.csv'. If a new dictionary is created, the descriptors related to the dictionary lookup must be updated accordingly.

Limiting lookup to Sections:

If you need the results to be limited to specific sections, you need to specify that in the file "PAD term spotter\resources\lookup\radiology\LookupDesc_csv.xml". This would typically pertain to clinical notes which have designated section header identification. The radiology notes used in the running of this pipeline have specific text based rules as defined in the 'SubSectionPadIdFSM' (see additional information below).

How does the annotator derive a "HIT"?
PAD_Hit.xml the descriptor for 'edu.mayo.bmi.uima.termspotter.ae.PADHit' Annotator takes a handful of parameters.
Params: WINDOW_SIZE, ANNOTATION_TYPE, ANNOTATION_PART_ONE_OF_PAIR, ANNOTATION_PART_TWO_OF_PAIR, may be others
a) If the annotations of type part_one and part_two fall in the window_size of type annotation_type, it is considered a hit.
b) If the annotation is defined as stand alone, then it does not require to be part of a pair to be considered a hit.

Customized Examples:
Use of customized terms, values, properties specifically for targeting the Mayo Radiology notes:

Overall the pipeline has been developed and trained with the Mayo patient radiology EMR, therefore, customization for other patient records in other hospital settings may necessary.

The following classes and files have Mayo specific site and terminology terms that are being leveraged, especially as it pertains to the
subsection handling:

1) '/core/src/edu/mayo/bmi/fsm/machine/SubSectionPadIdFSM.java'
   -Terms; "smh","rmh","gonda","romayo" are indicative of names of buildings on the Mayo campus which are used to mark subsection begin/end
   - terms; "indications","bleindications","exam","showing" are special terms which often contain the terms being screened for relating to PAD, but since they are titles of examinations, revision sections, and generalized screenings they are to be ignored in the Mayo cohort.

2) '/PAD term spotter/src/edu/mayo/bmi/uima/pad/impl/PADConsumerImpl.java'
   - Terms; "indications:" and "showing" are special terms which often contain the terms being screened for relating to PAD, but since they are titles of examinations, revision sections, and generalized screenings they are to be ignored in the Mayo cohort.

3) '/PAD term spotter/resources/lookup/radiology/ExamTitleWords.txt'
   - Comma delimited terms which represent key values to distinguish the type of radiology examination being utilized:
   US_EXAM (ultrasound), LOWER_EXT (lower extremity), US_LOWER_EXT (ultrasound lower extremity), US_LOWER_SOLO (ultrasound lower extremity one side only), CT_EXAM (CAT scan), CT_EXAM_SOLO (CAT Scan one side only)

4) '/PAD term spotter/resources/lookup/radiology/ExamsForPAD.csv'
   - Provides a list of valid examination codes to act as a filter to eliminate the need to parse records not related to PAD. This improves performance and minimizes the probability of false positive mentions.

5) '/PAD term spotter/src/edu/mayo/bmi/uima/termspotter/cc/PADOffSetsRecord.java'
   - Provides complex example of balancing context values, examination types, and various other variables relating to PAD to provide a classification of 'unknown' (-1), 'probable' (prob), 'PAD yes' (positive numeric value), 'PAD no' (0).

- Provides offset information for the 'hits' found related to PAD which is helpful to implement evaluation tools to highlight the actual text 'hits'.
    - If no matching term is found then it is represented as a '-1'
        - For example the following indicates a site term was found with no corresponding disorder:
            - (-1:325-344)|SIMPLE_SEGMENT|**NO TERM**|right thigh and leg

- Since each patient will typically have several radiology notes processed by the system, the precedence order in the event of conflicts between two or more classifications needs to be handled. A post processing step is also provided that shows how the classifier combines several record level classifications into one patient level one Thus the set precedence is as PAD-Yes > PAD-prob > PAD-no/unknown (combined category).

Sample Collection Processing Engine (CPE) "Radiology_sample_generic.xml" :

A sample CPE has been provided along with de-identified patient information to provide a means to test the pipeline after you have set up the environment. In order to run this sample you will need to specify the path where you installed the cTAKES projects in place of the '<path to your eclipse environment>' within the existing paths provided in the 'Input File Name:" and "Filter Exam Types:" fields. Additionally, specify the path and file name to output the record level radiology on your system.

Following is a description of the fields available in the 'Collection Reader' panel:

'Descriptor:' : The name of the collection reader that provides the additional fields and ties to the java class responsible for running the code which loads the records into cas memory. In this example the file '../collection_reader/RadiologyRecordsCollectionReader.xml' is specified.

'Input File Name:' (Required): The name and path of the file which contains the records which make up the radiology notes being processed. Each line of the record is considered a separate examination and will be classified as such if the record is not being filtered (see filtering options below). 'PAD term spotter/data/SampleInputRadiologyNotes.txt' is a de-identified sample of three records.

'Language:' (Optional): This will explicitly set a language if used, but do not use since only the English language has been incorporated.

'Comment String:' (Optional): The reader will filter/skip lines that begin with this case sensitive literal string. This prevents header columns from being processed as well as the records.

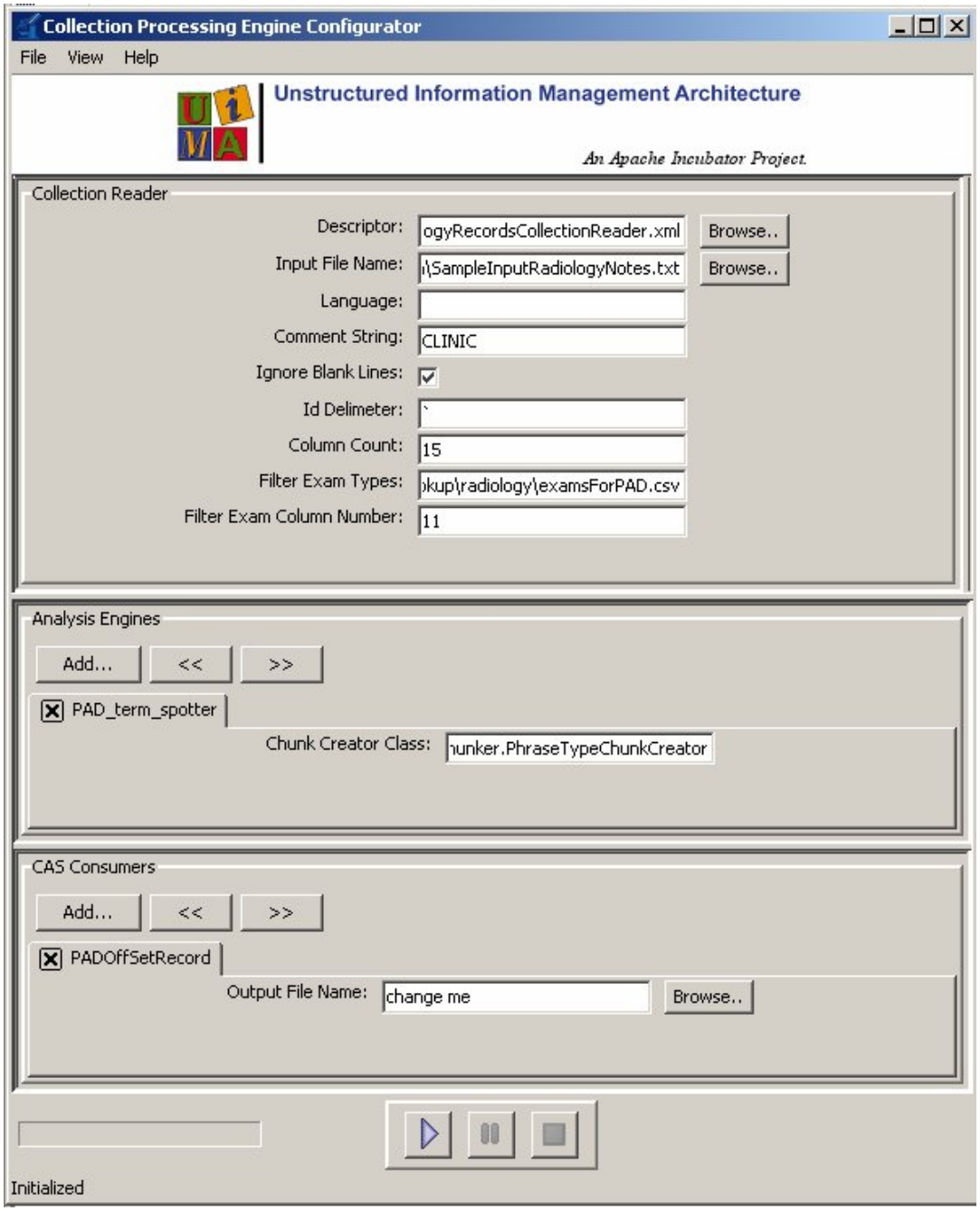
'Ignore Blank Lines:' (Optional): Will prevent blank rows from being processed which may cause interruptions to the subsequent processes.

'Id Delimiter:' (Optional): Specifies what character will be used to delimit the identification column (first column) or all fields, depending upon if values are specified for the remaining fields in this panel. For this sample the '`' is specified.

'Column Count:' (Optional): Indicates the number of columns, delimited using the value in 'Id Delimiter', should be skipped to locate the actual contents of the radiology examination. If this value is null or blank then only the first column will be skipped. In the provided sample the 15 field will designate the contents of the radiology record.

'Filter Exam Types:' (Optional): Provides a list of valid examination codes to act as a filter to eliminate the need to parse records not related to PAD. This improves performance and minimizes the probability of false positive mentions.

'Filter Exam Column Number:' (Optional): The column count of the radiology record, delimited using the value in 'Id Delimiter', used as input to compare the 'Filter Exam Types' provided above. The 11 column provides the 'testcode' information for mapping in the sample.



The CAS Consumer, "PADOffSetRecord" is provided as a means to post process the results and provides the following features:

1 - Record by record level classification for PAD
2 - Site and disorder terms along with offset information (useful for debugging)
3 - Overall patient level classification based on record classification

Record by Record level classification for PAD :

The private method 'calculateRecordLevelClassification' is responsible for providing the PAD classification given the factors;

* Type of examination US_EXAM (ultrasound), LOWER_EXT (lower extremity), US_LOWER_EXT (ultrasound lower extremity), US_LOWER_SOLO (ultrasound lower extremity one side only), CT_EXAM (CAT scan), CT_EXAM_SOLO (CAT Scan one side only)

* Number of hits, and, if applicable, the difference if both exist ("numberOfHits=total number of confirmed positive PAD hits" and "differentialHitCount=total positive PAD minus total negative PAD hits or if , no PAD hits, the number of terms only plus number of sites only found")

* Number of term only mentions, site only mentions, probable evidence, mentions of vein related terms, and mention of stent related terms ("numberOfTermOnly=terms found outside of hit", "numberOfLocationOnly=sites found outside of hit", "numberOfProbable=count probable elements", "numberVeinMentions=count of vein type terms", and "noMentionOfStents=count of stent type terms")

* Whether there is a mention of veins or stenosis related terms ("noMentionOfVeins=void of mention of vein type terms" “locationTermsOnly=location terms only”, and "noMentionOfStenosis=void of stenosis type terms")

* Whether there is evidence of probable, negation and possible negation ("foundEvidenceOfProbable=at least one probably element found", "haveNegativeCases=at least one negative term or site found", "possibleNegativeCases=negation is found, but not adjacent to stenosis, patent, or a site excluded term");

* noPadMentionThreshold is a somewhat arbitrary value used to measure a relative weight of evidence against other counts.

Algorithm for 'calculateRecordLevelClassification'’:

Site and disorder terms along with offset information (useful for debugging)

Provides offset information for the ‘hits’ found related to PAD which is helpful to implement evaluation tools to highlight the actual text ‘hits’.

- If no matching term is found then it is represented as a ‘-1’
  For example the following indicates a site term was found with no corresponding disorder:
  (-1:325-344)|SIMPLE_SEGMENT|**NO TERM**|right thigh and leg

Overall patient level classification based on record classification

Since each patient will typically have several radiology notes processed by the system, the precedence order in the event of conflicts between two or more classifications needs to be handled. A post processing step is also provided that shows how the classifier combines several record level classifications into one patient level one.

Thus the set precedence is as PAD-Yes > PAD-prob > PAD-no/unknown (combined category).