Natural Language Processing of Clinical Data

Giheon Koh[1], Ahmad Baghal[2]

[1]Department of Mathematics, University of Central Arkansas, Conway, AR, 72035

[2]Department of Biomedical Informatics, Univeristy of Arkansas for Medical Sciences, Little Rock, AR, 72211

Abstract

Problem: For the fast adoption of unstructured Electronic Medical Records (EMRs), it is necessary to extract accurate information from EMRs to support automated systems at the point of care and to enable secondary use of EMRs for clinical and translational research.

Objectives: For this research, I designed the Natural Language Processing protocol model with applying machine learning approaches and challenged to figure out as a minimum test error rate as possible while testing EMR samples in the protocol.

Methods: For the platform to build up, I used the Python programming language with the Natural Language Tool Kit (NLTK) package. For the experiment, I used 20 de-identified pathology report samples. For the workflow, I searched and extracted cardinal quantity values from the samples, then extracted the strings for the final diagnosis. When extracting the diagnosis information, I used Support Vector Machine (SVM) algorithm. While working in the SVM model, I calculated the test error to verify how significant the model is. For the entire protocol, I designed it based on the decision tree algorithm to extract necessary information from the unstructured EMRs.

Results: For the result, I extracted the patient's name, medical record number, gender, age, body part, procedure, diagnosis description, grade, date and note ID information from each sample, and stored them in database.

Introduction

For the last centuries, the healthcare system has been rapidly grown. For its growth, the relative healthcare data were desirably and necessarily stored to adapt throughout the evolution of healthcare technology. As a tool being adapted to the technology, Electronic Medical Record (EMR) has been providing clinicians a lot of benefits to understand patients by tracking over the histories from many healthcare providers. Through the EMRs, clinicians were efficiently able to identify general information such as treatments and medical histories of patients. However, with the rapid adoptions of unstructured or free text formatted EMRs which contain more detailed information about patients, there were some concerns raised. First, for the numbers of medical records, clinicians had to spend lots of time to read, classify and summarize information for each patient. Second, it was inevitable to extract valid and accurate information and knowledge from EMRs to support clinicians' decisions at the point of care. Third, it was desirable to enable secondary use of EMRs for clinical and translational research. To solve these concerns, Natural Language Processing tool has been developing to facilitate information extraction from such unstructured free texts.

Laurie Miles, the head of analytics for big data specialist at SAS [1], mentioned, "About 75% of data is unstructured, coming from sources such as text, voice, and video." Due to the massive data are hidden in unstructured texts, it became essential to develop tools to extract information, and Natural Language Processing (NLP) presents as the solution. According to the definition of Natural Language Processing by SAS institute Inc. [2], NLP is "a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.” Thus, NLP helps clinicians to be able to interpret free text or human language and make it analyzable.

Saying for a short brief of Natural Language Processing history with referring the work of Nadkarni et al. [3], it started in the 1950s as an intersectional tool of computer and linguistics. In 1956, Chomsky published a book, *Syntactic Structures*, and suggested the revolutionized linguistic concepts which include that a computer understands a human language. Since then, many programming languages were suggested and developed. After fourteen years, Natural Language Processing was researched linked with statistics for its popularity. In the 21st century, there were many proposals for the neural language model and Apple’s Siri, the first successful NLP and artificial intelligence assistant, came out to the world. The research of NLP is still an ongoing project to minimize the test error and extract as accurate information as possible.

According to Kreimeyer et al. [4], utilizing Natural Language Processing for text mining have many advantages. First, it will help to reduce time for manual expert review. For clinicians, processing numbers of EMRs for patients is a highly time-consuming task. However, when doing this with the automated processing system which converts the unstructured data to the structured, it will result in a lot of reduction of time for expert review and also more flexibility of secondary use of such data for large scale automated processing. Utilizing NLP is also advantageous for gaining more knowledge about patients. Sometimes, it is possible that a clinician misses important information from a free-text medical report. To prevent this, NLP will organize all necessary information and store them into the database.

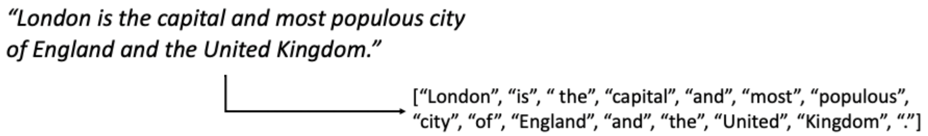
Despite these advantages, NLP still has some challenges. Because of a free text’s poor structures, abundant shorthand, and domain-specific vocabularies, it is quite a bit hard to figure out zero test error rate, meaning that, some cases miss capturing important terms or certain captures of not important terms. For this research, I designed the NLP protocol model with applying machine learning approaches and challenged to figure out as a minimum test error rate as possible while testing EMR samples in the protocol.

Important Concepts

To understand the Natural Language Processing protocol, the basic NLP concepts should be priory studied. There are many concepts in NLP but studying a few important concepts may work to understand the entire NLP protocol. The listed concepts are arranged based on the procedure of building up the protocol: Corpus, Tokenization, Stop words, Normalization, Stemming, Lemmatization, and Part-of-Speech tagging.

The first concept is the corpus. A corpus is a body written or spoken material upon which a linguistic analysis is based, meaning that statistics is accumulated on natural language text. In practice, the medical record sample is referred to the corpus.

The second concept is tokenization. The idea of tokenization is to process word recognition by splitting strings into smaller pieces called tokens. As shown in *Figure 1*, a string is split into tokens.

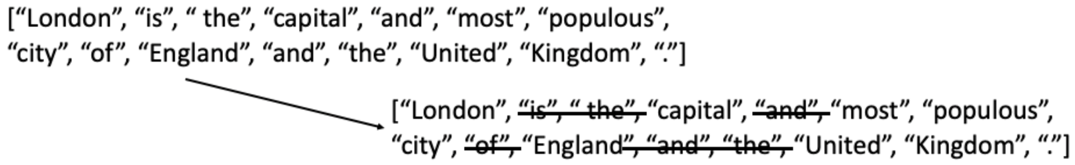


“London is the capital and most populous city of England and the United Kingdom.”

[“London”, “is”, “ the”, “capital”, “and”, “most”, “populous”, “city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”, “.”]

Figure 1 Tokenization

The third concept is the stop words. The stop words are generally the most common in a language which should be filtered out before further processing of text because they contribute little to the overall meaning. For the examples of stop words, there are ‘the’, ‘a/ an’, ‘of’, ‘and’, ‘or’, ‘am/ is/ are’, and other things. So, for the general process is shown in *Figure 2*, the stop words are deleted from the tokenized sentence.

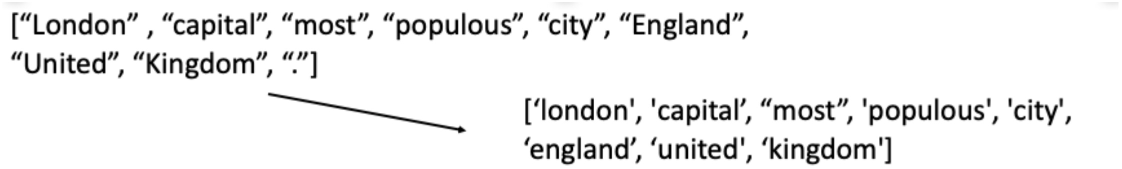


[“London”, “is”, “ the”, “capital”, “and”, “most”, “populous”,
“city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”, “.”]

[“London”, ~~“is”, “ the”,~~ “capital”, ~~“and”,~~ “most”, “populous”,
“city”, ~~“of”,~~ “England~~”, “and”, “the”,~~ “United”, “Kingdom”, “.”]

Figure 2 Filtering against Stop words

The fourth concept is normalization. Before further processing, texts need to be normalized, meaning that the testing samples should be laid on equal footing. In the process of normalization, it mainly refers to the tasks such as converting all texts to the same case either uppercase or lowercase, removing all punctuations, expanding contractions, and converting numbers to their word equivalents. For example, *Figure 3* shows that all uppercases are converted to the lowercases, and punctuations are removed.

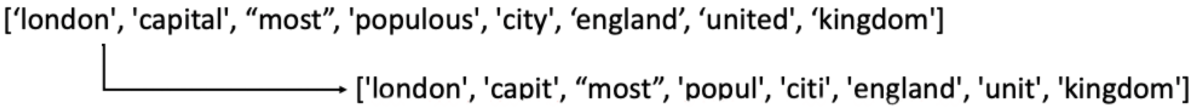


[“London” , “capital”, “most”, “populous”, “city”, “England”,
“United”, “Kingdom”, “.”]

[‘london', 'capital’, “most”, 'populous', 'city',
‘england’, ‘united', ‘kingdom']

Figure 3 Normalization

The stemming is the process of eliminating affixes such as suffixes, prefixes, infixes, and circumfixes from a word to obtain a word stem. The stemming is necessary to recognize two other words are meaning the same. For example, there are two words: ran and running. For a computer, they are different because of the different spellings and the lengths of words. However, when looking over their words’ stems, they are fundamentally same. Through the stemming process, a computer can recognize that those words mean the same thing. In practice, *Figure 4* shows an example. The words, ‘capital’, ‘populous’, ‘city’ and ‘united’, are converted to their stem words.

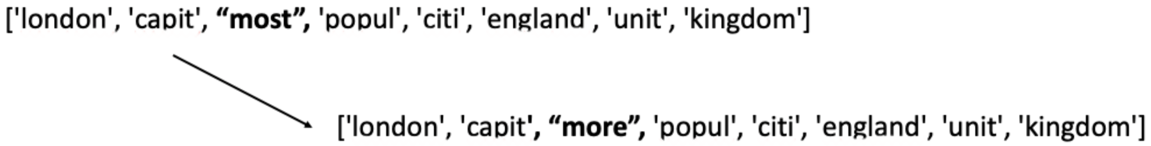


[‘london', 'capital', “most”, 'populous', 'city', ‘england’, ‘united', ‘kingdom']

['london', 'capit', “most”, 'popul', 'citi', 'england', 'unit', 'kingdom']

Figure 4 Stemming

Lemmatization is the similar process with the stemming process, meaning that the purposes of stemming and lemmatization are the same, but the difference is that the lemmatization captures the canonical forms based on a word’s lemma. In linguistic definition, a lemma is *“the base form under which the word is entered in a dictionary.”* For an example of lemmatization, a word, ‘better’, will be processed to ‘good’ after lemmatizing. In the practice shown in *Figure 5*, the word, ‘most’, is converted to the ‘more’ based on its word lemma.

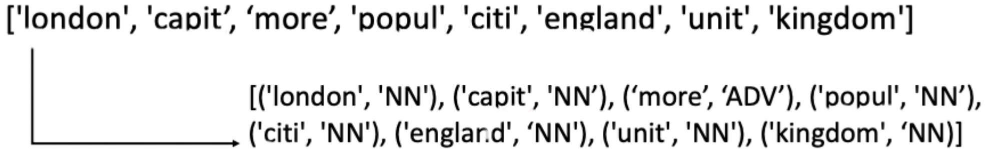


```
['london', 'capit', "most", 'popul', 'citi', 'england', 'unit', 'kingdom']
```

```
['london', 'capit', "more", 'popul', 'citi', 'england', 'unit', 'kingdom']
```

Figure 5 Lemmatization

For the last concept, the Part-Of-Speech (POS) tagging is the process to assign category tags to the tokens of a sentence based on the general rule of Part of Speech. By conducting the POS tagging to each word, it helps a computer to recognize what to extract. In this process, what a computer mainly takes care are the nouns, verbs, and adjectives. In *Figure 6*, 'NN' means the common noun, and 'ADV' means adverb. For additional information, aside from the 'NN' and 'ADV', there are many types of nouns, verbs, and adjectives based on the detail branches of the general rule of Part of Speech.



```
['london', 'capit', 'more', 'popul', 'citi', 'england', 'unit', 'kingdom']
```

```
[('london', 'NN'), ('capit', 'NN'), ('more', 'ADV'), ('popul', 'NN'),
('citi', 'NN'), ('england', 'NN'), ('unit', 'NN'), ('kingdom', 'NN)]
```

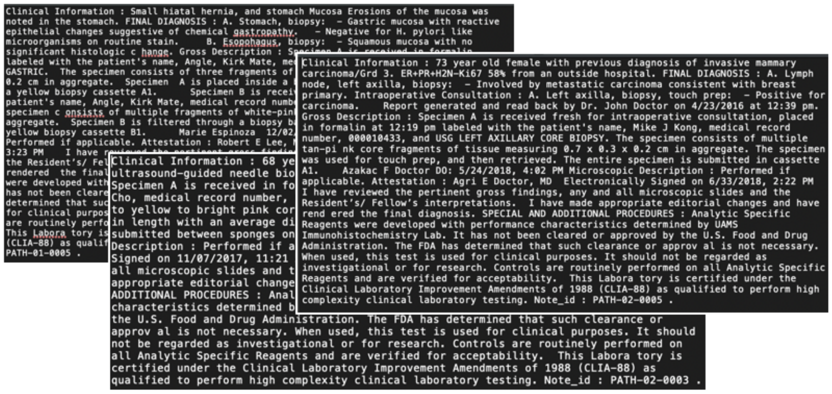
Figure 6 Part-Of-Speech (POS) Tagging

Technology Applied

For this research project, I used the Python programming language because the Natural Language Processing (NLP) requires Machine-Learning (ML) algorithms for automated processing of large-scale data, and Python provides both NLP and ML tools in one platform. For the NLP tool, I used the Natural Language Tool Kit (NLTK) package because of its utility and popularity.

Samples Description

For this experiment, I used 20 de-identified pathology reports. I could not use the actual pathology reports because the actual pathology reports refer to the real patients' information which is confidential. As shown in *Figure 7*, the samples are quite a bit unstructured and have a different arrangement of information. It is because the structures of such pathology reports depend on patients' diagnosis or the provider that entered the text.



```
Clinical Information : Small hiatal hernia, and stomach Mucosa Erosions of the mucosa was
noted in the stomach. FINAL DIAGNOSIS : A. Stomach, biopsy:  – Gastric mucosa with reactive
epithelial changes suggestive of chemical gastropathy.   – Negative for H. pylori like
microorganisms on routine stain.     B. Esophagus, biopsy:  – Squamous mucosa with no
significant histologic c hange. Gross Description : Specimen A is received in formalin
labeled with the patient's name, Angle, Kirk Mate, me
GASTRIC.  The specimen consists of three fragments of
0.2 cm in aggregate.  Specimen  A is placed inside a
a yellow biopsy cassette A1.      Specimen B is recei
patient's name, Angle, Kirk Mate, medical record numbe
specimen c onsists of multiple fragments of white-pin
aggregate.  Specimen B is filtered through a biopsy b
yellow biopsy cassette B1.     Marie Espinoza  12/02
Performed if applicable. Attestation : Robert E Lee,
3:23 PM    I have
the Resident's/ Fel
rendered  the final
were developed with
has not been cleare
determined that suc
for clinical purpos
are routinely perfo
This Labora tory is
(CLIA-88) as qualif
PATH-01-0005 .
```

```
Clinical Information : 68 ye
ultrasound-guided needle bio
Specimen A is received in fo
Cho, medical record number,
to yellow to bright pink cor
in length with an average di
submitted between sponges on
Description : Performed if a
Signed on 11/07/2017, 11:21
all microscopic slides and t
appropriate editorial change
ADDITIONAL PROCEDURES : Anal
characteristics determined b
the U.S. Food and Drug Administration. The FDA has determined that such clearance or
approv al is not necessary. When used, this test is used for clinical purposes. It should
not be regarded as investigational or for research. Controls are routinely performed on
all Analytic Specific Reagents and are verified for acceptability.  This Labora tory is
certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA-88) as
qualified to perform high complexity clinical laboratory testing. Note_id : PATH-02-0003 .
```

```
Clinical Information : 73 year old female with previous diagnosis of invasive mammary
carcinoma/Grd 3. ER+PR+H2N-Ki67 58% from an outside hospital. FINAL DIAGNOSIS : A. Lymph
node, left axilla, biopsy:  – Involved by metastatic carcinoma consistent with breast
primary. Intraoperative Consultation : A. Left axilla, biopsy, touch prep:  – Positive for
carcinoma.    Report generated and read back by Dr. John Doctor on 4/23/2016 at 12:39 pm.
Gross Description : Specimen A is received fresh for intraoperative consultation, placed
in formalin at 12:19 pm labeled with the patient's name, Mike J Kong, medical record
number, 000010433, and USG LEFT AXILLARY CORE BIOPSY. The specimen consists of multiple
tan-pi nk core fragments of tissue measuring 0.7 x 0.3 x 0.2 cm in aggregate. The specimen
was used for touch prep, and then retrieved. The entire specimen is submitted in cassette
A1.    Azakac F Doctor DO: 5/24/2018, 4:02 PM Microscopic Description : Performed if
applicable. Attestation : Agri E Doctor, MD  Electronically Signed on 6/33/2018, 2:22 PM
I have reviewed the pertinent gross findings, any and all microscopic slides and the
Resident's/ Fellow's interpretations.  I have made appropriate editorial changes and have
rend ered the final diagnosis. SPECIAL AND ADDITIONAL PROCEDURES : Analytic Specific
Reagents were developed with performance characteristics determined by UAMS
Immunohistochemistry Lab. It has not been cleared or approved by the U.S. Food and Drug
Administration. The FDA has determined that such clearance or approv al is not necessary.
When used, this test is used for clinical purposes. It should not be regarded as
investigational or for research. Controls are routinely performed on all Analytic Specific
Reagents and are verified for acceptability.  This Labora tory is certified under the
Clinical Laboratory Improvement Amendments of 1988 (CLIA-88) as qualified to perform high
complexity clinical laboratory testing. Note_id : PATH-02-0005 .
```

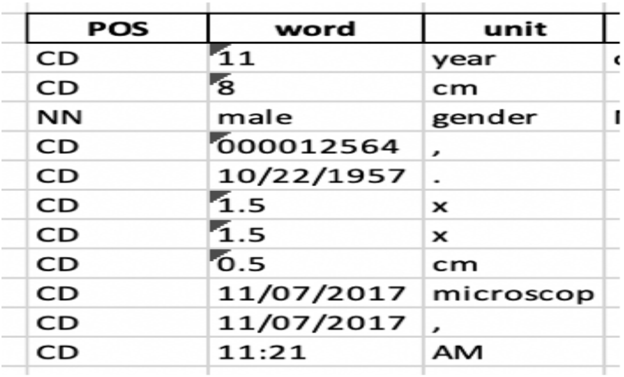
Figure 7 Examples of deidentified pathology reports

Workflow

The workflow is categorized into four stages: Preprocess, Information extraction based on Part-Of-Speech (POS) tags, Information extraction from diagnosis description, and Exporting the mined data.

According to Assale et al. [5], the preprocessing stage includes data cleaning, data integration, data reduction, and data transformation. The stage aims two major things: to make data cleaner in terms of noise, inconsistency and incompleteness, and to improve the speed and accuracy of data mining, dealing with heterogeneous data and its redundancy. In practice, setting the stemming, lemmatization standards and the list of stop words is the first step. These standards and list are provided by the Natural Language Tool Kit (NLTK) package in Python. Then, importing the Unified Medical Language System (UMLS) is the second step. The UMLS is a compendium of many controlled vocabularies in the biomedical sciences and used to train testing samples to extract only important terminologies in a diagnosis description. After setting and importing these, building up the algorithm to test sentence by sentence in a testing sample and word by word in a processing sentence is the next step. The mechanism of this algorithm is that while looping for every sentence in a sample, there is another looping for every single word inside a sentence. For a pointer in the looping for sentences iterating, the pointed sentence is tokenized, and pointer inside the sentence is generated. For iterating words in the tokenized sentence, the pointed word is tested to be filtered out against stop words. Then, for the filtered words, they are lemmatized. After the process, the Part-Of-Speech tags are assigned to each word.

In the information extraction based on POS tags stage, the algorithm is designed to search for cardinal quantity values which are the numbers and categorized as nouns. The reason why it searches for the cardinal quantity values is that the numbers in medical reports are typically more meaningful about patients than any words. When extracting the cardinal quantity or numeric values in a sample, the age, medical record number, and procedure date and time are extracted on average. After that, the algorithm searches gender information among the nouns. For this, I set a list of words indicating gender such as ‘male’, ‘female’, ‘man’, ‘woman’, and such things. Thus, based on those words, the algorithm searches for the gender information. In *Figure 8*, it shows the table for extracted information based on the POS tags.



POS	word	unit
CD	11	year
CD	8	cm
NN	male	gender
CD	000012564	,
CD	10/22/1957	.
CD	1.5	x
CD	1.5	x
CD	0.5	cm
CD	11/07/2017	microscop
CD	11/07/2017	,
CD	11:21	AM

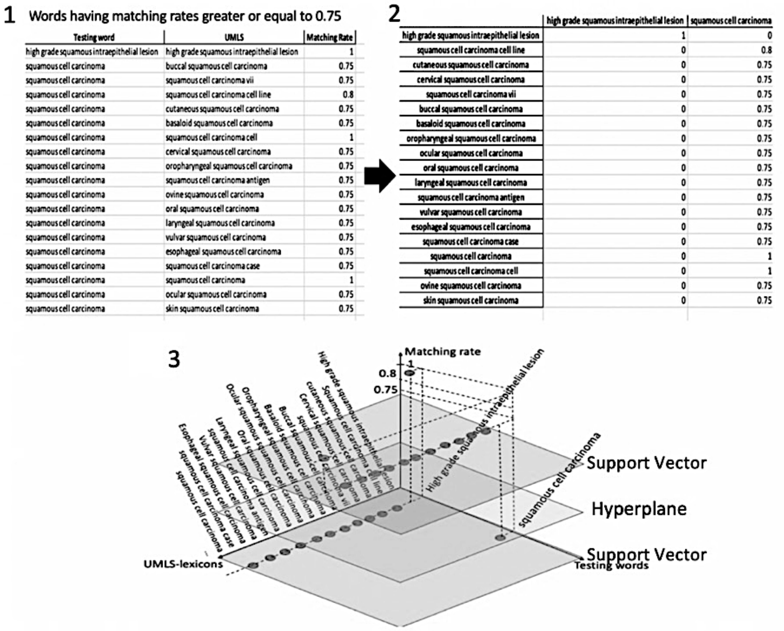
Figure 8 Table for Extracted Information based on the POS tags

In the information extraction from the diagnosis description stage, the first step is to extract the strings for the final diagnosis from a sample. Throughout all samples, there is a common indicator for the final diagnosis section. For example, most of the final diagnosis sections are formed in such:

> … FINAL DIAGNOSIS : A. Cervix, biopsy: - At least high grade squamous intraepithelial lesion (CIN III) in a background of extensive necrosis, see comment. Comment : The specimen is predominately necrotic; however, there are small superficial fragments of severely dysplastic epithelium present. These findings are consistent with squamous cell carcinoma (unsampled) ...

From the sample, there is the section indicator at the head which is 'FINAL DIAGNOSIS'. Based on the indicator, the algorithm extracts the strings for the final diagnosis section. From the extracted strings, the next step is to compart body part, clinical procedure information and diagnosis descriptions for each alphabetical index whose numbers depend on the numbers of diagnoses. With finding the minimal common pattern in the final diagnosis for each index, the next string after a string indicating alphabetical index is the body part, the following string to the body part is the clinical procedure and the strings behind the hyphen are the diagnosis description for the body part in the same index.

While the body part and clinical procedure information are extracted, the strings for diagnosis description should be filtered for the important terms, and to do this, the Support-Vector Machine (SVM) algorithm is applied. SVM is one of the machine learning approaches that analyze data used for classification or grouping. The way of classifying is that the SVM algorithm creates a regression model-based optimal compartment called a hyperplane between the data points and generates the support vectors based on the closest points on each comparted side. The test error is calculated from the margin of support vectors. Thus, for this research, SVM is used to calculate the test error of classification. Beforehand conducting SVM, since SVM is a supervised learning algorithm, a training dataset should be defined. So, as the training dataset, UMLS which is imported in the preprocessing stage is set. When the training dataset is defined, each word is tested for the word's matching rate with the words in UMLS.



1 Words having matching rates greater or equal to 0.75

Testing word	UMLS	Matching Rate
high grade squamous intraepithelial lesion	high grade squamous intraepithelial lesion	1
squamous cell carcinoma	buccal squamous cell carcinoma	0.75
squamous cell carcinoma	squamous cell carcinoma vii	0.75
squamous cell carcinoma	squamous cell carcinoma cell line	0.8
squamous cell carcinoma	cutaneous squamous cell carcinoma	0.75
squamous cell carcinoma	basaloid squamous cell carcinoma	0.75
squamous cell carcinoma	squamous cell carcinoma cell	1
squamous cell carcinoma	cervical squamous cell carcinoma	0.75
squamous cell carcinoma	oropharyngeal squamous cell carcinoma	0.75
squamous cell carcinoma	squamous cell carcinoma antigen	0.75
squamous cell carcinoma	ovine squamous cell carcinoma	0.75
squamous cell carcinoma	oral squamous cell carcinoma	0.75
squamous cell carcinoma	laryngeal squamous cell carcinoma	0.75
squamous cell carcinoma	vulvar squamous cell carcinoma	0.75
squamous cell carcinoma	esophageal squamous cell carcinoma	0.75
squamous cell carcinoma	squamous cell carcinoma case	0.75
squamous cell carcinoma	squamous cell carcinoma	1
squamous cell carcinoma	ocular squamous cell carcinoma	0.75
squamous cell carcinoma	skin squamous cell carcinoma	0.75

2

	high grade squamous intraepithelial lesion	squamous cell carcinoma
high grade squamous intraepithelial lesion	1	0
squamous cell carcinoma cell line	0	0.8
cutaneous squamous cell carcinoma	0	0.75
cervical squamous cell carcinoma	0	0.75
squamous cell carcinoma vii	0	0.75
buccal squamous cell carcinoma	0	0.75
basaloid squamous cell carcinoma	0	0.75
oropharyngeal squamous cell carcinoma	0	0.75
ocular squamous cell carcinoma	0	0.75
oral squamous cell carcinoma	0	0.75
laryngeal squamous cell carcinoma	0	0.75
squamous cell carcinoma antigen	0	0.75
vulvar squamous cell carcinoma	0	0.75
esophageal squamous cell carcinoma	0	0.75
squamous cell carcinoma case	0	0.75
squamous cell carcinoma	0	1
squamous cell carcinoma cell	0	1
ovine squamous cell carcinoma	0	0.75
skin squamous cell carcinoma	0	0.75

3
Matching rate
1
0.8
0.75
High grade squamous intraepithelial lesion
Squamous cell carcinoma cell line
cutaneous squamous cell carcinoma
Cervical squamous cell carcinoma
squamous cell carcinoma vii
Buccal squamous cell carcinoma
Basaloid squamous cell carcinoma
Oropharyngeal squamous cell carcinoma
Ocular squamous cell carcinoma
Oral squamous cell carcinoma
Laryngeal squamous cell carcinoma
squamous cell carcinoma antigen
Vulvar squamous cell carcinoma
Esophageal squamous cell carcinoma
squamous cell carcinoma case
squamous cell carcinoma
High grade squamous intraepithelial lesion
squamous cell carcinoma
Support Vector
Hyperplane
Support Vector
UMLS-lexicons
Testing words

Figure 9

1. Matching rate table for testing words and UMLS-lexicons, 2. Transformed table for the input of SVM model, 3. Visualization of the SVM model

Moving back to the example, when filtering the result table for the matching rate greater or equal to 0.75, *Figure 9-1* is generated. To fit this table as the input of the SVM model, the table should be transformed into a contingency table *(Figure9-2)*. When the table is visualized, it presents in *Figure 9-3*. In *Figure 10,* the hyperplane is created based on the linear regression model and the support vectors are generated on the closest data points on each side. In the final output, the terms upon the upper support vectors are chosen to be tested for their significances.

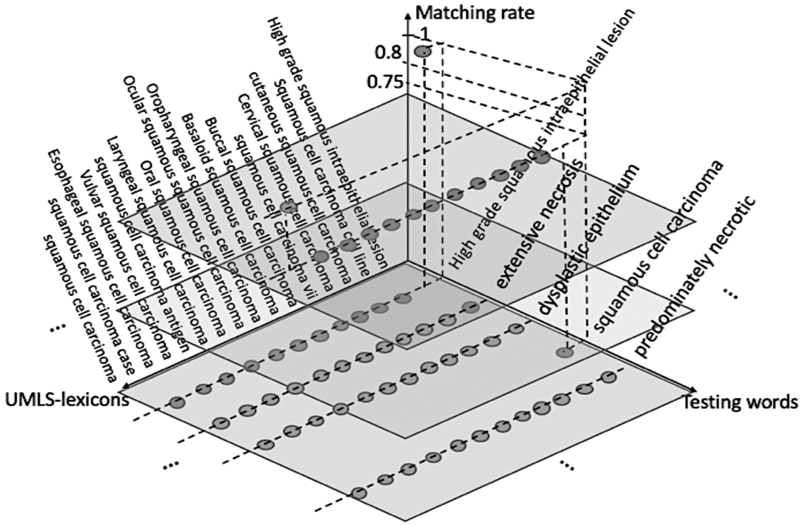

Matching rate
1
0.8
0.75
High grade squamous intraepithelial lesion
Squamous cell carcinoma cell line
cutaneous squamous cell carcinoma
Cervical squamous cell carcinoma vii
squamous cell carcinoma
Buccal squamous cell carcinoma
Basaloid squamous cell carcinoma
Oropharyngeal squamous cell carcinoma
Ocular squamous cell carcinoma
Oral squamous cell carcinoma
Laryngeal squamous cell carcinoma antigen
squamous cell carcinoma
Vulvar squamous cell carcinoma
Esophageal squamous cell carcinoma case
squamous cell carcinoma
...
UMLS-lexicons
High grade squamous intraepithelial lesion
extensive necrosis
dysplastic epithelium
squamous cell carcinoma
predominately necrotic
...
Testing words
...
...

Figure 10

When the SVM model is set, the algorithm calculates a test error of the SVM model. The *Figure 11* is the recommended formula from the research of Gaonkar and Davatzikos (2013) [6] for the approximation of permutation testing for SVMs. In the formula, *E*[*P* (*Error*)] is a measure of the generalization/test error of the SVM, E[Number of support vectors] is the mean number of support vectors as permutating, and the number of training samples is the number of permutating samples. Therefore, for the example in *Figure 10*, there are only two linear support vectors when permutating because those support vectors are based on the linear regression model. The number of training samples is the factorial of the testing samples which is 20, thus, 20! = 2.432902008 E+18, therefore, *E*[*P* (*Error*)] ≤ 1/20! ≈ 0. For the result of the example, the test error of the SVM model is less or equal to 0. For all tests, the terms of diagnosis description in final output is chosen based on the models which have the test error less than 0.1.

$$E\left[P\left(Error\right)\right] \leq \frac{E[Number\ of\ support\ vectors]}{Number\ of\ training\ samples}$$

Figure 11 Formula for test error of SVM with permutation

For the last stage, the algorithm reanalyzes and extract the information for medical note ID, a medical record number of a patient, and age information from the table for extracted information based on the POS tags. Then, it finally merges all harvested information in one data frame and exports it to Microsoft Excel file.

For the entire protocol, it is designed following the decision tree in *Figure 12.* Every word after filtered out against stop words and normalized, the words are tested for their types of classification based on the unit defined, indicators set, and general definitions in medical records.

When a word reaches to a node, if the node is one of the underlined, the word is stored as information, or skipped.

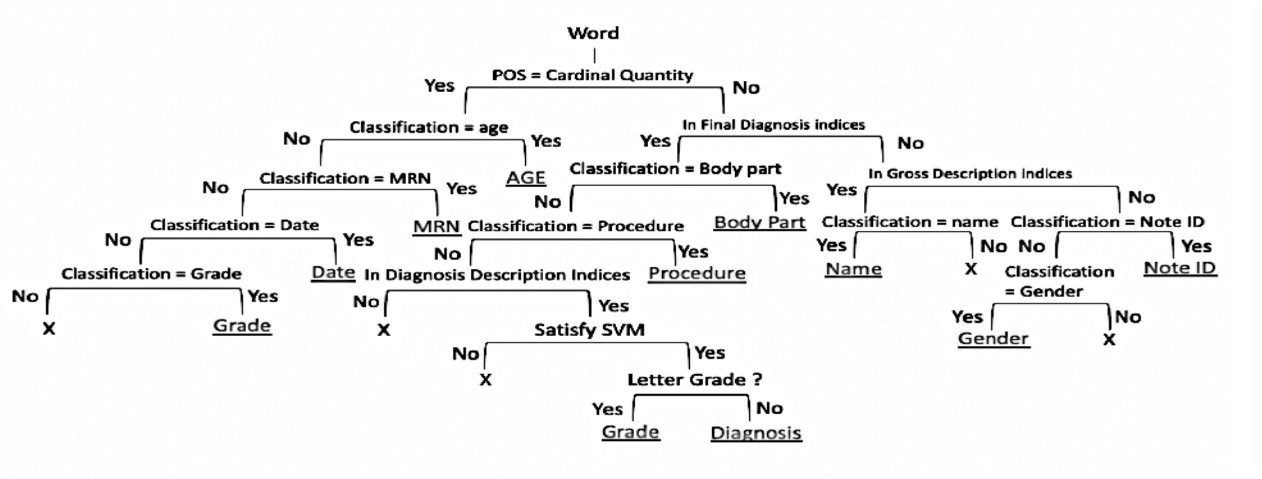
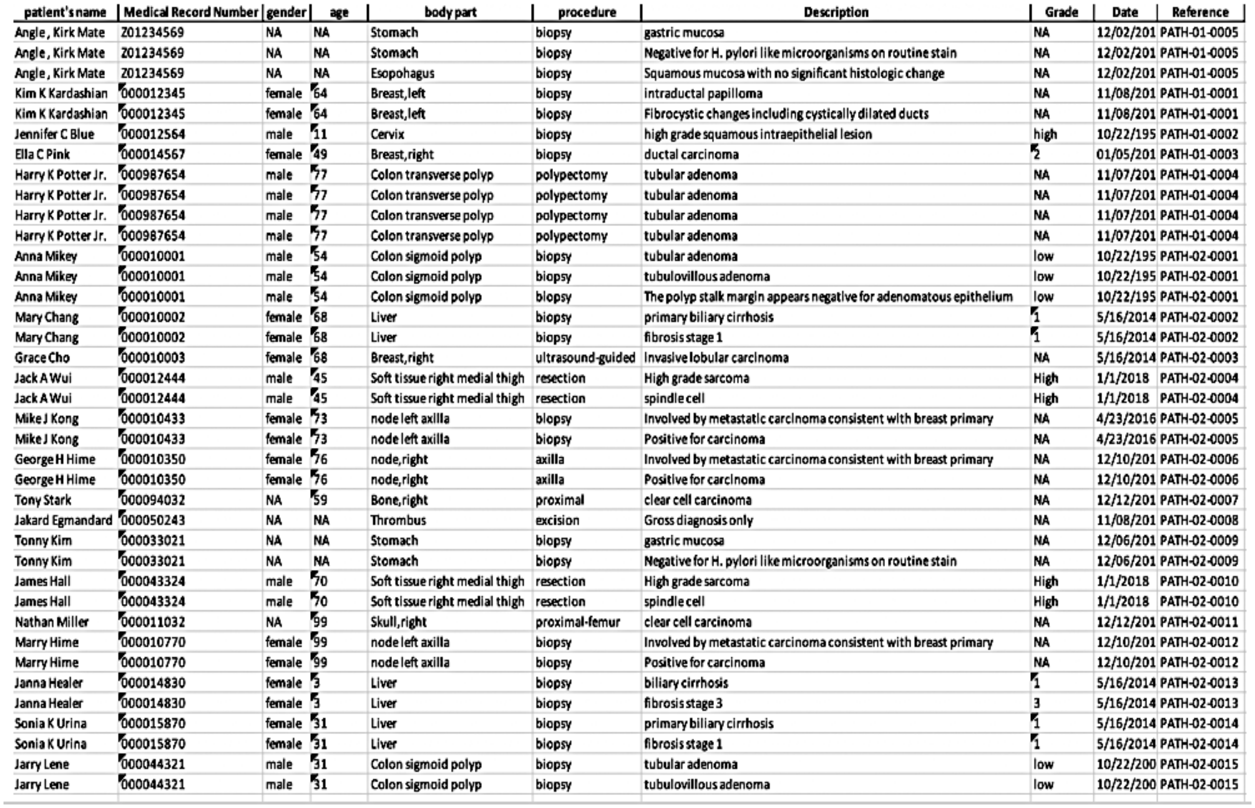


Figure 12 Decision Tree Model

Final Result

As going through the workflow, the outcome of processing 20 de-identified pathology reports in the protocol is generated. For each column, it refers to patient's name, medical record number, age, body part, procedure, diagnosis description, the grade of diagnosis, procedure date, and medical record ID.



patient's name	Medical Record Number	gender	age	body part	procedure	Description	Grade	Date	Reference
Angle, Kirk Mate	Z01234569	NA	NA	Stomach	biopsy	gastric mucosa	NA	12/02/201	PATH-01-0005
Angle, Kirk Mate	Z01234569	NA	NA	Stomach	biopsy	Negative for H. pylori like microorganisms on routine stain	NA	12/02/201	PATH-01-0005
Angle, Kirk Mate	Z01234569	NA	NA	Esopohagus	biopsy	Squamous mucosa with no significant histologic change	NA	12/02/201	PATH-01-0005
Kim K Kardashian	000012345	female	64	Breast,left	biopsy	intraductal papilloma	NA	11/08/201	PATH-01-0001
Kim K Kardashian	000012345	female	64	Breast,left	biopsy	Fibrocystic changes including cystically dilated ducts	NA	11/08/201	PATH-01-0001
Jennifer C Blue	000012564	male	11	Cervix	biopsy	high grade squamous intraepithelial lesion	high	10/22/195	PATH-01-0002
Ella C Pink	000014567	female	49	Breast,right	biopsy	ductal carcinoma	2	01/05/201	PATH-01-0003
Harry K Potter Jr.	000987654	male	77	Colon transverse polyp	polypectomy	tubular adenoma	NA	11/07/201	PATH-01-0004
Harry K Potter Jr.	000987654	male	77	Colon transverse polyp	polypectomy	tubular adenoma	NA	11/07/201	PATH-01-0004
Harry K Potter Jr.	000987654	male	77	Colon transverse polyp	polypectomy	tubular adenoma	NA	11/07/201	PATH-01-0004
Harry K Potter Jr.	000987654	male	77	Colon transverse polyp	polypectomy	tubular adenoma	NA	11/07/201	PATH-01-0004
Anna Mikey	000010001	male	54	Colon sigmoid polyp	biopsy	tubular adenoma	low	10/22/195	PATH-02-0001
Anna Mikey	000010001	male	54	Colon sigmoid polyp	biopsy	tubulovillous adenoma	low	10/22/195	PATH-02-0001
Anna Mikey	000010001	male	54	Colon sigmoid polyp	biopsy	The polyp stalk margin appears negative for adenomatous epithelium	low	10/22/195	PATH-02-0001
Mary Chang	000010002	female	68	Liver	biopsy	primary biliary cirrhosis	1	5/16/2014	PATH-02-0002
Mary Chang	000010002	female	68	Liver	biopsy	fibrosis stage 1	1	5/16/2014	PATH-02-0002
Grace Cho	000010003	female	68	Breast,right	ultrasound-guided	Invasive lobular carcinoma	NA	5/16/2014	PATH-02-0003
Jack A Wui	000012444	male	45	Soft tissue right medial thigh	resection	High grade sarcoma	High	1/1/2018	PATH-02-0004
Jack A Wui	000012444	male	45	Soft tissue right medial thigh	resection	spindle cell	High	1/1/2018	PATH-02-0004
Mike J Kong	000010433	female	73	node left axilla	biopsy	Involved by metastatic carcinoma consistent with breast primary	NA	4/23/2016	PATH-02-0005
Mike J Kong	000010433	female	73	node left axilla	biopsy	Positive for carcinoma	NA	4/23/2016	PATH-02-0005
George H Hime	000010350	female	76	node,right	axilla	Involved by metastatic carcinoma consistent with breast primary	NA	12/10/201	PATH-02-0006
George H Hime	000010350	female	76	node,right	axilla	Positive for carcinoma	NA	12/10/201	PATH-02-0006
Tony Stark	000094032	NA	59	Bone,right	proximal	clear cell carcinoma	NA	12/12/201	PATH-02-0007
Jakard Egmandard	000050243	NA	NA	Thrombus	excision	Gross diagnosis only	NA	11/08/201	PATH-02-0008
Tonny Kim	000033021	NA	NA	Stomach	biopsy	gastric mucosa	NA	12/06/201	PATH-02-0009
Tonny Kim	000033021	NA	NA	Stomach	biopsy	Negative for H. pylori like microorganisms on routine stain	NA	12/06/201	PATH-02-0009
James Hall	000043324	male	70	Soft tissue right medial thigh	resection	High grade sarcoma	High	1/1/2018	PATH-02-0010
James Hall	000043324	male	70	Soft tissue right medial thigh	resection	spindle cell	High	1/1/2018	PATH-02-0010
Nathan Miller	000011032	NA	99	Skull,right	proximal-femur	clear cell carcinoma	NA	12/12/201	PATH-02-0011
Marry Hime	000010770	female	99	node left axilla	biopsy	Involved by metastatic carcinoma consistent with breast primary	NA	12/10/201	PATH-02-0012
Marry Hime	000010770	female	99	node left axilla	biopsy	Positive for carcinoma	NA	12/10/201	PATH-02-0012
Janna Healer	000014830	female	3	Liver	biopsy	biliary cirrhosis	1	5/16/2014	PATH-02-0013
Janna Healer	000014830	female	3	Liver	biopsy	fibrosis stage 3	3	5/16/2014	PATH-02-0013
Sonia K Urina	000015870	female	31	Liver	biopsy	primary biliary cirrhosis	1	5/16/2014	PATH-02-0014
Sonia K Urina	000015870	female	31	Liver	biopsy	fibrosis stage 1	1	5/16/2014	PATH-02-0014
Jarry Lene	000044321	male	31	Colon sigmoid polyp	biopsy	tubular adenoma	low	10/22/200	PATH-02-0015
Jarry Lene	000044321	male	31	Colon sigmoid polyp	biopsy	tubulovillous adenoma	low	10/22/200	PATH-02-0015

Figure 13 Final Output

Challenges

While building up the Natural Language Processing protocol, it was quite a bit challenged to deal with its unstructured format, abundant of shorthand, wrong grammars and spellings, misuse of hyphen, and others. Because the contents of the same type of reports are differently organized depending on patients' diagnoses or the provider that entered the text, it causes the high-test error rate. For future research, building on the advanced protocol dealing with these variances will be studied.

Conclusion

Because of the rapid adoption of the unstructured Electronic Medical Record, Natural Language Processing has been developed to extract important and dependable information from such free texts. For this project, I applied Support Vector Machine to extract the accurate and important terms from the texts and Decision Tree algorithms to automate the processing of large-scale data. As a result, I extracted the patient's name, MRN, gender, age, body part, procedure, diagnosis description, grade, date, and note ID. While doing the research, there were many challenges to deal with high variances in the free text. With researching more natural language processing cases, approaches and protocols, building up the advanced NLP protocol will be continued in future research.

Acknowledgment

This project was supported by the Arkansas INBRE program, with the National Institute of General Medical Sciences -NIGMS (P20 GM103429) from the National Institutes of Health. I also would like to acknowledge Dr. Ahmad Baghal, Dr. Shaymaa Al-shukri, and Annu Kumari in the Department of Biomedical Informatics at UAMS for supporting the project.

References

[1] Wall, M. (2014, March 04). Big Data: Are you ready for blast-off? Retrieved June 18, 2019, from https://www.bbc.com/news/business-26383058

[2] What is Natural Language Processing? Retrieved July 9, 2019, from https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

[3] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544–551. doi:10.1136/amiajnl-2011-000464

[4] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., . . . Botsis, T. (17 july 2017). Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information. *Journal of Biomedical Informatics,*1-16. doi:10.1016

[5] Assale, M., Dui, L. G., Cina, A., Seveso, A., & Cabitza, F. (17 april 2019). The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Frontiers in Medicine,*1-23. doi:10.3389

[6] Gaonkar B, Davatzikos C. Deriving statistical significance maps for SVM based image classification and group comparisons. *Med Image Comput Comput Assist Interv*. 2012;15(Pt 1):723–730.