SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Data Analytics

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

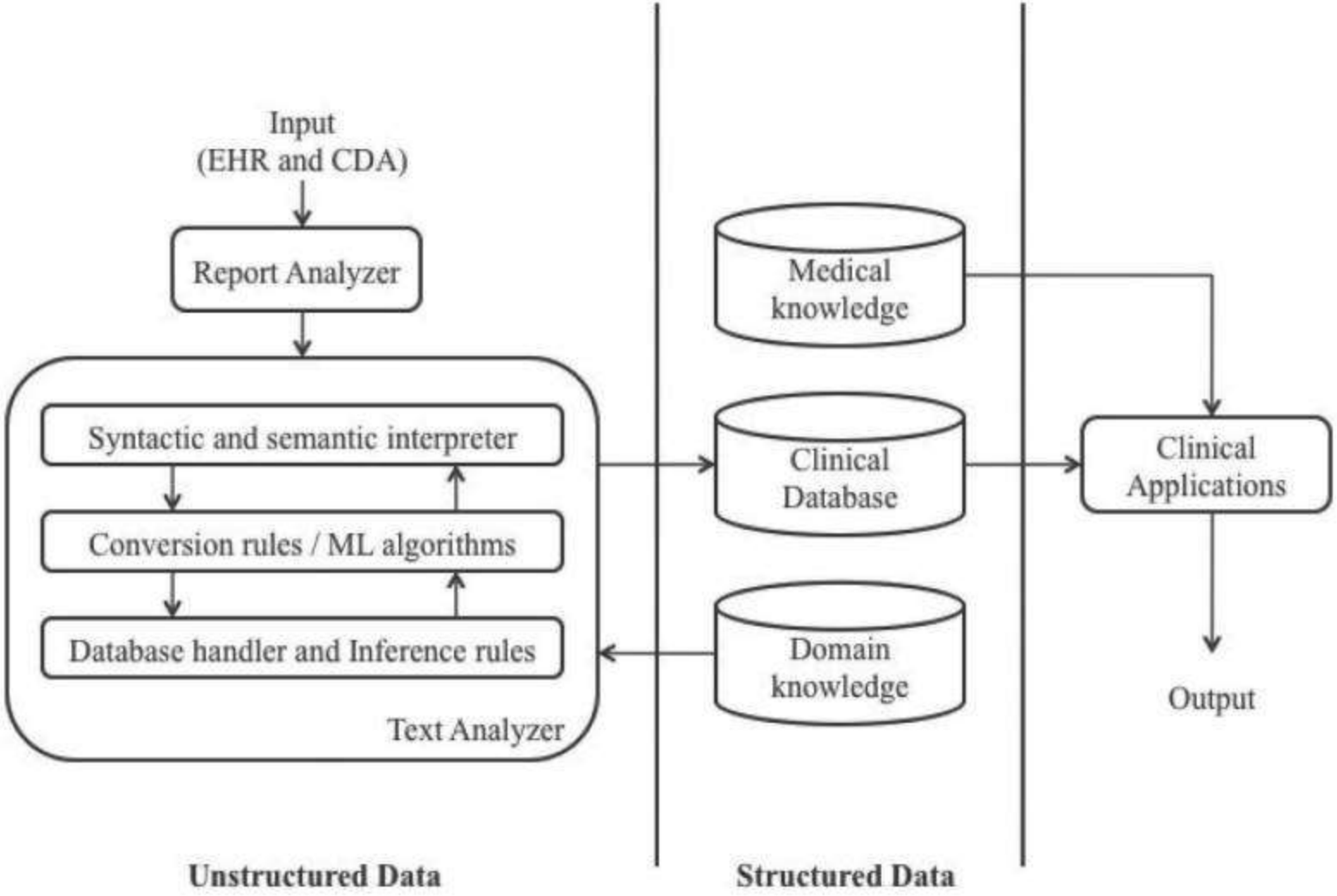
Somaiya TRUST

Natural Language Processing and Data Mining for Clinical Text

General Discussion over NLP

- Student Should Add some points to it

- The input for a NLP system is the unstructured natural text from a patient's medical record that is given to a report analyzer to identify segments and to handle textual irregularities such as tables, domain specific abbreviations, and missing punctuation.
- The core of the NLP engine is the text analyzer that uses the syntactic and semantic knowledge associated with the domain knowledge to extract information.
- In text analyzer a syntactic and semantic interpreter captures the respective details and generates a deeper structure such as a constituent tree or dependency tree.
- The conversion rules or ML algorithms accept this deep structure and encode the clinical information to make it compatible for the database storage.
- The database handler and inference rules work to generate a processed form from the storage point of view.


Input
(EHR and CDA)
Report Analyzer
Syntactic and semantic interpreter
Conversion rules / ML algorithms
Database handler and Inference rules
Text Analyzer
Medical knowledge
Clinical Database
Domain knowledge
Clinical Applications
Output
Unstructured Data
Structured Data

Report Analyzer

- The clinical text differs from the biomedical text with the possible use of pseudo tables, i.e., natural text formatted to appear as tables, medical abbreviations, and punctuation in addition to the natural language.
- The text is normally dictated and transcribed to a person or speech recognition software and is usually available in free-text format.
- Some clinical texts are even available in the image or graph format.
- For example, the radiology report is a primary means of communication between a radiologist and the referring physician.

- However, the data is required to be in a structured format for the purposes of research, quality assessment, interoperability, and decision support systems.
- As a result, NLP processing techniques are applied to convert the unstructured free-text into a structured format. Numerous NLP applications require preprocessing of the input text prior to the analysis of information available.
- The first and foremost task of report analyzer is to preprocess the clinical input text by applying NLP methodologies. The major preprocessing tasks in a clinical NLP include text segmentation, text irregularities handling, domain specific abbreviation, and missing punctuations

Text Analyzer

- Text analyzer is the most important module in clinical text processing that extracts the clinical information from free-text and makes it compatible for database storage.
- The analyzer is meant to perform an automatic structuring of clinical data into predefined sections in addition to text processing and extraction of clinical information.
- An initial preprocessing of the text is necessary to map the medical concepts in narrative text to an unstructured information management architecture (UMLS) Metathesaurus.
- The UMLS (http://uima.apache.org) is the most commonly employed application for a concept extraction pipeline.
- The components in the pipeline are tokenization, lexical normalization, UMLS Metathesaurus look-up, and concept screening and medical subject heading (MeSH) conversion as described below:
  - Tokenization component splits query into multiple tokens.
  - Lexical normalization component converts words to their canonical form.
  - UMLS Metathesaurus look-up for concept screening.
  - Semantic grouping of UMLS concepts.
  - Mapping of the screened concepts to MeSH headings

Core NLP components

Input Text → Pt took aspirin 325 mg for knee pain

Tokenization →

Pt	*took*	aspirin	325	mg	for	knee	pain

Morphological analysis →

Pt	*take*	aspirin	325	mg	for	knee	pain

Lexical analysis, Syntactic analysis and Semantic analysis →

```
(ROOT
  (S
    (NP (NNP Pt))
    (VP (VBD took)
      (NP
        (NP
          (QP (CD aspirin) (CD 325))
          (NN mg))
        (PP (IN for)
          (NP (NN knee) (NN pain)))))))
```

- Morphological Analysis
  - This is the process of converting a sentence into a sequence of tokens that are mapped to their canonical base form (e.g., cures = cure + s).
  - This is the first stage of processing the input text in most of the NLP systems to reduce the number of words/tokens needed for the lexical analysis.
  - An alternate approach is to implement a POS tagger to identify syntactic POS of the words and possibly their canonical forms

- Lexical Analysis
- The words or phrases in the text are mapped to the relevant linguistic information such as syntactic information, i.e., noun, verb, adverb, etc., and semantic information i.e., disease, procedure, body part, etc.
- Lexical analysis is achieved with a special dictionary called a lexicon, which provides the necessary rules and data for carrying out the linguistic mapping.
- The development of maintenance of a lexicon requires extensive knowledge engineering and effort to develop and maintain.
- The National Library of Medicine (NLM) maintains the Specialist Lexicon [13] with comprehensive syntactic information associated with both medical and English terms.

- Syntactic Analysis
  - The word “syntax” refers to the study of formal relationships between words in the text.
  - The grammatical knowledge and parsing techniques are the major key elements to perform syntactic analysis.
  - The context free grammar (CFG) is the most common grammar used for syntactic analysis. CFG is also known by various other terms including phrase structure grammar (PSG) and definite clause grammar (DCG).
  - The syntactic analysis is done by using two basic parsing techniques called top-down parsing and bottom-up parsing to assign POS tags (e.g., noun, verb, adjective, etc.) to the sequence of tokens that form a sentence and to determine the structure of the sentence through parsing tools.

- Semantic Analysis
- It determines the words or phrases in the text that are clinically relevant, and extracts their semantic relations.
- The natural language semantics consists of two major features: (1) the representation of the meanings of a sentence, which can allow the possible manipulations (particularly inference) and (2) relating these representations to the part of the linguistic model that deals with the structure (grammar or syntax).
- The semantic analysis uses the semantic model of the domain or ontology to structure and encodes the information from the clinical text.
- The semantic model is either frame oriented or conceptual graphs. The generated structured output of the semantic analysis is subsequently used by other automated processes.

- Data Encoding
- The process of mining information from EHR requires coding of data that is achieved either manually or by using NLP techniques to map free-text entries with an appropriate code.
- The coded data is classified and standardized for storage and retrieval purposes in clinical research.
- Manual coding is normally facilitated with search engines or pick-up list
- NLP techniques make use of a wide range of available medical vocabularies such as ICD-10-CM, SNOMED, UMLS, and even locally developed vocabularies.
- The automatic ending of clinical text concepts to a standardized vocabulary is an area of interest for many NLP research teams.

Clinical NLP system	Purpose
LSP-MLP	NLP system for extraction and summarization of signs/symptoms and drug information, and identification of possible medication and side effects
MedLEE	A semantically driven system used for (1) extracting information from clinical narrative reports, (2) participating in an automated decision-support system, and (3) allowing NLP queries
cTAKES	Mayo clinical Text Analysis and Knowledge Extraction System
SPRUS	A semantically driven IE system
SymText	NLP system with syntactic and probabilistic semantic analysis driven by Bayesian Networks
SPECIALIST	A part of UMLS project with SPECIALIST lexicon, semantic network, and UMLS Metathesaurus
IndexFinder	A method for extracting key concepts from clinical text for indexing
KnowledgeMap	A full-featured content management system to enhance the delivery of medical education contents
Lexical Tools	A set of fundamental core NLP tools for retrieving inflectional variants, uninflectional forms, spelling variants, derivational variants, synonyms, fruitful variants, normalization, UTF-8 to ASCII conversion, and many more
MetaMap	A highly configurable program to map biomedical text to UMLS Metathesaurus concepts

Mining Information from Clinical Text

- Clinical text mining is an interdisciplinary area of research requiring knowledge and skills in computer science, engineering, computational linguistics, and health science.
- It is a subfield of biomedical NLP to determine classes of information found in clinical text that are useful for basic biological scientists and clinicians for providing better health care

- Text mining and data mining techniques to uncover the information on health, disease, and treatment response support the electronically stored details of patients' health records.
- A significant chunk of information in HER and CDA are text and extraction of such information by conventional data mining methods is not possible.
- The semi-structured and unstructured data in the clinical text and even certain categories of test results such as echocardiograms and radiology reports can be mined for information by utilizing both data mining and text mining techniques.

- Information Extraction
  - Information extraction (IE) is a specialized field of NLP for extracting predefined types of information from the natural text.
  - It is defined as the process of discovering and extracting knowledge from the unstructured text
  - IE differs from information retrieval (IR) that is meant to be for identifying and retrieving relevant documents.
  - In general, IR returns documents and IE returns information or facts.
  - A typical IE system for the clinical domain is a combination of components such as tokenizer, sentence boundary detector, POS tagger, morphological analyzer, shallow parser, deep parser (optional), gazetteer, named entity recognizer, discourse module, template extractor, and template combiner.

- Preprocessing
  - The primary source of information in the clinical domain is the clinical text written in natural language.
  - However, the rich contents of the clinical text are not immediately accessible by the clinical application systems that require input in a more structured form. An initial module adopted by various clinical NLP systems to extract information is the preliminary preprocessing of the unstructured text to make it available for further processing.
  - The most commonly used preprocessing techniques in clinical NLP are spell checking, word sense disambiguation, POS tagging, and shallow and deep parsing

- Spell Checking
- The misspelling in clinical text is reported to be much higher than any other types of texts. In addition to the traditional spell checker, various research groups have come out with a variety of methods for spell checking in the clinical domain: UMLS-based spell-checking error correction tool and morpho-syntactic disambiguation tools

- Word Sense Disambiguation
  - The process of understanding the sense of the word in a specific context is termed as word sense
  - disambiguation. The supervised ML classifiers and the unsupervised approaches automatically perform the word sense disambiguation for biomedical terms.

- POS Tagging
- An important preprocessing step adapted by most of the NLP systems is POS tagging that reads the text and assigns the parts of speech tag to each word or token of the text.
- POS tagging is the annotation of words in the text to their appropriate POS tags by considering the related and adjacent words in a phrase, sentence, and paragraph.

Note

- A much more elaborated set of tags is provided by complete Brown Corpus tag-set (www.hit.uib.no/icame/brown/bcm.html) with eighty seven basic tags and Penn Treebank tag-set (www.cis.upenn.edu/treebank) with forty five tags.

- Shallow and Deep Parsing
  - Parsing is the process of determining the complete syntactic structure of a sentence or a string of symbols in a language.
  - Parser is a tool that converts an input sentence into an abstract syntax tree such as the constituent tree and dependency tree, whose leafs correspond to the words of the given sentence and the internal nodes represent the grammatical tags such as noun, verb, noun phrase, verb phrase, etc.
  - Most of the parsers apply ML approaches such as PCFGs (probabilistic context-free grammars) as in the Stanford lexical parser nd even maximum entropy and neural network.
  - Few parsers even use lexical statistics by considering the words and their POS tags.
  - Such taggers are well known for overfitting problems that require additional smoothing.

- An alternative to the overfitting problem is to apply shallow parsing, which splits the text into nonoverlapping word sequences or phrases, such that syntactically related words are grouped together.
- The word phrase represents the predefined grammatical tags such as noun phrase, verb phrase, prepositional phrase, adverb phrase, subordinated clause, adjective phrase, conjunction phrase, and list marker
- The benefits of shallow parsing are the speed and robustness of processing. Parsing is generally useful as a preprocessing step in extracting information from the natural text

List of POS Tags

TABLE 7.2: List of POS Tags

Tag	Description
$	Dollar
‘	open quotation mark
”	closing quotation mark
(	open parenthesis
)	closed parenthesis
,	comma
–	dash
.	sentence terminator
:	colon or ellipsis
CC	conjunction, coordinating
CD	numeral, cardinal
DT	determiner
EX	existential there
FW	foreign word
IN	preposition or conjunction, subordinating
JJ	adjective or numeral, ordinal

JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	model auxiliary
NN	noun, common, singular or mass
NNP	noun, proper, singular
NNPS	noun, proper, plural
NNS	noun, common, plural
PDT	predeterminer
POS	genitive marker
PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
SYM	symbol
TO	to as preposition or infinitive marker
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle or gerund
VBN	verb, past participle
VBP	verb, present tense, not 3rd person singular
VBZ	verb, present tense, 3rd person singular
WDT	WH-determiner
WP	WH-pronoun
WP$	WH-pronoun, possessive
WRB	WH-adverb

Tag	Description
$	Dollar
´ ’	open quotation mark
”	closing quotation mark
(	open parenthesis
)	closed parenthesis
,	comma
–	dash
.	sentence terminator
:	colon or ellipsis
CC	conjunction, coordinating
CD	numeral, cardinal
DT	determiner
EX	existential there
FW	foreign word
IN	preposition or conjunction, subordinating
JJ	adjective or numeral, ordinal
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	model auxiliary
NN	noun, common, singular or mass
NNP	noun, proper, singular
NNPS	noun, proper, plural
NNS	noun, common, plural
PDT	predeterminer
POS	genitive marker
PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
SYM	symbol
TO	to as preposition or infinitive marker
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle or gerund
VBN	verb, past participle
VBP	verb, present tense, not 3rd person singular

Context-Based Extraction

- Concept Extraction
- Extracting concepts (such as drugs, symptoms, and diagnoses) from clinical narratives constitutes a basic enabling technology to unlock the knowledge within and support more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment. The first and foremost module in clinical NLP following the initial text preprocessing phase is the identification of the boundaries of the medical terms/phrases and understanding the meaning by mapping the identified term/phrase to a unique concept identifier in an appropriate ontology

Association Extraction

- Clinical text is the rich source of information on patients' conditions and their treatments with additional information on potential medication allergies, side effects, and even adverse effects.

Coreference Resolution

- Coreferential expressions are common in clinical narratives and therefore understanding coreference relations plays a critical role in the discourse-level analysis of clinical documents, such as compiling a patient profile.
- Since the language and description style in clinical documents differ from common English, it is necessary to understand the characteristics of clinical text to properly perform coreference resolution.
- A comprehensive methodological review of coreference resolution developed for general English can be applied for coreference resolution in the clinical domain.
- The existing methodologies for coreference resolution are:
  1. Heuristics-based approaches based on linguistic theories and rules
  2. Supervised machine learning approaches with binary classification of markable mention/entity pairs or classification by ranking markables
  3. Unsupervised machine learning approaches, such as nonparametric Bayesian models or expectation maximization clustering.

Negation

- “Negation” is an important context that plays a critical role in extracting information from the clinical text.
- Many NLP systems incorporate a separate module for negation analysis in text preprocessing
-

Temporality Analysis

- Temporal resolution for events and time expressions in clinical notes is crucial for an accurate summary of patient history, better medical treatment, and further clinical study.
- Discovery of a temporal relation starts with extracting medical events and time information and aims at building a temporal ink (TLINK) between events or between events and time expressions.
- Clinical practice and research would benefit greatly from temporal expression and relation detection.

Extracting Codes

- Extracting codes is a popular approach that uses NLP techniques to extract the codes mapped to controlled sources from clinical text.
- The most common codes dealing with diagnoses are the International Classification of Diseases (ICD) versions 9 and 10 codes.
- The ICD is designed to promote international comparability in the collection, processing, classification and presentation of mortality statistics.
- ICD-10 is the latest revised codes available with coding for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury (http://apps.who.int/classifications/icd10/browse/2010/en).

Current Methodologies

- Rule-Based Approaches
  - Rule-based approaches rely on a set of rules for possible textual relationships, called patterns, which encode similar structures in expressing relationships.
  - The set of rules are expressed in the form of regular expressions over words or POS tags. In such systems, the rules extend as patterns by adding more constraints to resolve few issues including checking negation of relations and determining direction of relations.
  - The rules are generated in two ways: manually constructed and automatically generated from the training dataset

- Pattern-Based Algorithms
  - The second popular approach for extracting information from the clinical text is the patternbased algorithm.
  - A set of word patterns are coded based on the biomedical entities and their relation keywords to extract special kinds of interactions.
  - These approaches can vary from simple sentencebased extraction to more advanced extraction methods using POS tagging with additional linguistic information.

- Machine Learning Algorithms

Input variables: $\boldsymbol{x} = (x_1, x_2, \ldots x_m)$
Hidden variables: $\boldsymbol{h} = (h_1, h_2, \ldots h_k)$
Output variables: $\boldsymbol{y} = (y_1, y_2, \ldots y_n)$

Challenges of Processing Clinical Reports

- Domain Knowledge
- Confidentiality of Clinical Text
- Abbreviations
- Diverse Formats
- Expressiveness

TABLE 7.4: PHI Identifiers Related to Confidentiality of Clinical Text

PHI identifiers
1. Name of the patient
2. All geographical identifiers smaller than a state except for the initial three digits of a zip code
3. Dates (other than year)
4. Phone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers (SSN)
8. Medical record numbers
9. Health insurance beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web URLs
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger, retinal, and voice prints
17. Full face photographic images/any comparable images
18. Any other unique identifying numbers, characteristics, or codes except for the unique codes assigned by the investigator to code the data

- Intra- and Interoperability
  Interpreting Information

Question ?