Non-small cell lung cancer and small cell lung cancer. Each has a separate staging system that doctors use to classify how advanced the cancer is.

Techniques:

1. Named Entity Recognition
2. rule-based methods
3. shallow classifiers
4. dictionary-based method
5. IE extraction methods: cTAKES, GATE framework, MedEx, TEMPTING(extracting temporal relations useful in tracking the progression of the disease from patient discharge summaries),

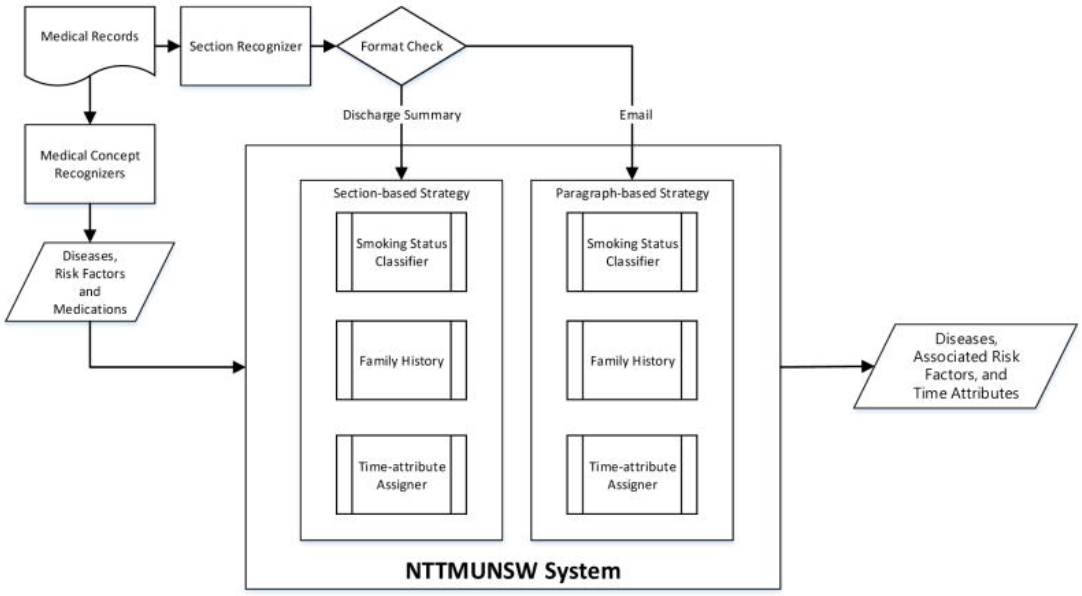
Article Notes:
real-world progression (rwP) dates

Reference on breast cancer:

1. https://ascopubs.org/doi/pdf/10.1200/CCI.20.00139
2. https://www.researchgate.net/publication/351596637_Automated_NLP_Extraction_of_Clinical_Rationale_for_Treatment_Discontinuation_in_Breast_Cancer
3. https://github.com/clinicalml/oncology_rationale_extraction

Links for various NLP techniques:
ncbi.nlm.nih.gov/pmc/articles/PMC6528438/#ref77

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4977838/
One general approach for progression tracking is to first recognize all temporal expressions, and then assign each to the nearest target concept.



1) mention concept recognizer (eg. diabetes, CAD) Tried approaches:

1. dictionary based(220 terms + old research terms)
2. Machine learning manually annotated data used, using IOB2 tagging scheme

MI model - part-of-speech tags, shallow parser tags, dictionary matching and bag of words to build the model

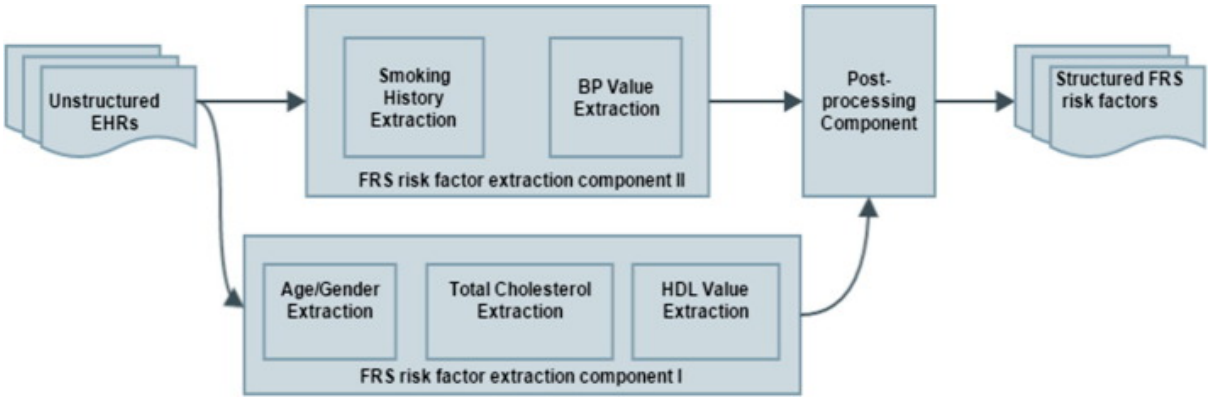
2) Risk factors recognition: (eg. numeric=BP)
for non-numeric type risk factors - Dictionary based approach used
Summary of the targeted diseases and their corresponding risk factor definitions

Category	Risk Factor	Numeric Value
Diabetes	High **A1C**	>=6.5
Diabetes	High **glucose**	>126
Hyperlipidemia	High **cholesterol**	>=240
Hyperlipidemia	High **LDL**	>=100 mg/dL
Hypertension	High **blood pressure**	>=140/90 mm/hg
Obesity	**BMI**	> 30
Obesity	Waist circumference	Men:>= 40 inches; Women: >= 35 inches

3) Context-aware time attribute assignment:
For instance, our risk factor recognizer will recognize the entire statement “HBA1C 05/25/2092 7.30” as a High A1C risk factor (7.30 > 6.5), which indicates that the risk factor was observed on 05/25/2092.

2. Relation extraction from discharge summary:
https://pubmed.ncbi.nlm.nih.gov/24060600/

3. Coronary artery disease risk assessment from unstructured electronic health records using text mining



4. NLP with lung cancer
https://ascopubs.org/doi/full/10.1200/CCI.20.00020
For each note, curators recorded whether the assessment/plan indicated any cancer, progression/worsening of disease, and/or response to therapy or improving disease.

5. https://www.sciencedirect.com/science/article/pii/S1532046416301381?via%3Dihub

6. Context-based date extraction: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153063/
7. https://ascopubs.org/doi/full/10.1200/CCI.19.00147

Questions / To Do:

1. There are three main things to extract:
    a. **Progression has occurred**. If the doctor states that progression has not occurred, it is not of particular interest to us, aside from the fact that we should not extract that progression has occurred.
    Expected Output - interested in only yes(based on rule-based) or no.
    Steps:
    - Rule based method used to define target variable (assign labels to the data)
    - CNN for classification of defined labels
        1. **identity assessment plan**
            - rule based method(dictionary of words)
            - RNN trained after removing key phrases - to predict whether each remaining word in each note was part of the assessment/plan
            - model can detect plan/asses without having key phrases
        2. **predict presence and status of cancer and progression/worsening**
            - CNN to predict cancer is there or not
            - whether cancer was progressing or worsening
            - whether cancer was responding or improving
            - lasso linear regression is used to predict outcome of CNN

    Rule based methods:
        - Dictionary based approach
        - Annotation
        - NLP methods - Regex

    b. **Date of progression**. The date of progression is sometimes mentioned explicitly in the note, but it is often not. When it is not mentioned, we can infer the date of progression from other data we have. We can use a hierarchical approach to assign the date of progression:
        i. If the date of progression is explicitly mentioned (e.g., “progression as of 21/9/21”), extract it, else,
        ii. If the date of the scan is explicitly mentioned (e.g., “CT scan dated 15/9/21 compared to prior scan shows progressive disease”), extract it, else,

iii. If the progression is relevant to the current visit but date is not explicitly mentioned (e.g., “patient shows signs of progressing”), extract the note date, else,

iv. If the progression is mentioned for some time in the past but no date is given, extract "No date provided"

c. **Medication regimen**. We need to be able to tie the progression to the cancer drug regimen the patient was on at the time of progression, e.g., to be able to say that progression occurred while the patient was receiving Keytruda. The regimen may consist of one or more drugs, and cancer treatments are often given as a combination of 2+ drugs. To extract the regimen, we can use the hierarchical approach of:

Note: only medication

i. If the medication regimen is provided in the note mentioning progression, extract it, else

ii. If the patient’s medication records contain information about the cancer drugs they are receiving, extract it - not using NLP, but applying rules to data

2. How do we identify progression? - See 1a above

a. Doctors determine progression based on an increase in tumor size or the spread of disease to other organs. We will not use this approach because we are using real-world data, where tumor sizes are not consistently reported. Instead, we will rely on a doctor’s explicit statement that progression has occurred, based on the understanding that the physician has reviewed all relevant reports and has synthesized this information to make the determination.

3. Context-based associate date - see 1b above

Associated date can be extracted after identifying the progression using the following methods.

reference:

I2b2 nlp challenge, extract three types (before, after and overlap)

Clinical tempeval challenge, temporal information extraction and temporal relation tasks.

Temporal based relation extraction:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3756273/

Steps:

1. Rule based information extraction(open source medtagger)

- identify events from EHR
- Remove Negated events(needs to get more idea)
- extract dates associated with events
    - if we already have a predefined sections
    - extract dates within and around the these sentences
    - also extract linked dates with extracted events

2. Then we need to normalize the date into proper format.(MedTime was used to normalize the dates)

4. Multiple progression notes
    a. Problem 1 - in their notes, doctors may mention progression that has happened a long time ago or recently. We therefore need to be able to distinguish when it occurred and which cancer drugs they were on at the time of progression.
    b. Problem 2 - doctors will repeat the same information across different notes. They tend to copy and paste prior summaries into the next summary, which could make it seem like there were multiple progression events, even though all notes are referring to the same event.
    c. Problem 3 - multiple progressions are possible. A patient can be on Treatment X and have the disease progress. Next, they are given Treatment Y, and again they have progression. We will want to differentiate and capture both progression events. Doctors tend to report the events in chronological order (oldest to newest) but that may not always be the case.
5. Progression "sentiment"
    a. Progression is a worsening of disease, so there may be sentiment-relevant terms that help us identify progression where the term is not explicitly mentioned, e.g., patient worsening, deterioration, or decline.
    b. We should, however, focus our efforts on explicit mentions of progression.
6. Rules-based approach to identifying progression
    a. A list of inclusion and exclusion criteria are provided in the Excel file located in the same folder as this file (see file named "Rules-Based Progression Identification 20210920.xlsx")
7. Do we also want to check the presence of cancer?

    We already predefined set of data

lung cancer diagnosis can be extracted using:
- identify sentences with mention of lung cancer with a custom dictionary of terms that describe the histological cell types of lung cancer, examinations of lung cancer, symptoms of lung cancer and positive malignancy of lung tumor.
For clinical notes, the whole dictionary was used associated with the mention of "diagnosis" to identify the sentence with lung cancer diagnosis. After extracting lung cancer diagnosis, we then extracted dates associated with this event from both pathology reports and clinical notes.

8. Are we interested in response to treatment?

Inputs from Kiran:

CT scan (date): progression ... date
- extract

progression - in front progression word

where ct scan date is discharge summary

ct scan (any scan)
date
progression

- if scan word is not there
- pet ct
- ct
- mri

progression keyword:
Increase FDG uptake or Increase uptake
previous word should - pet-ct
not ct or mri

Alternative Models

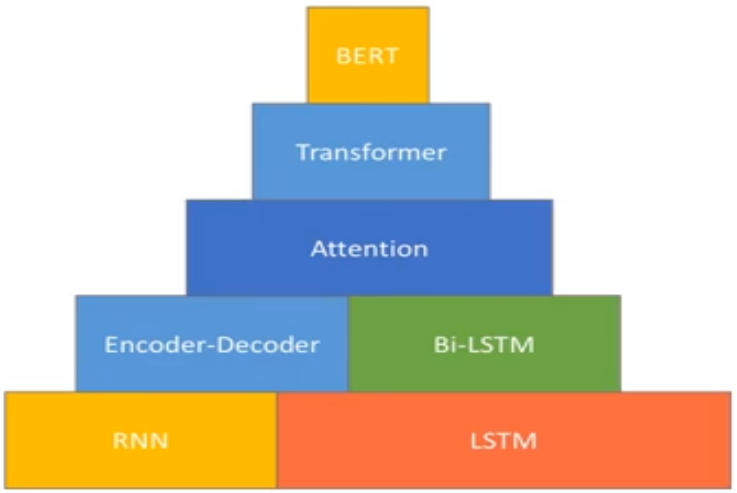


Fig1: Bert Mountain

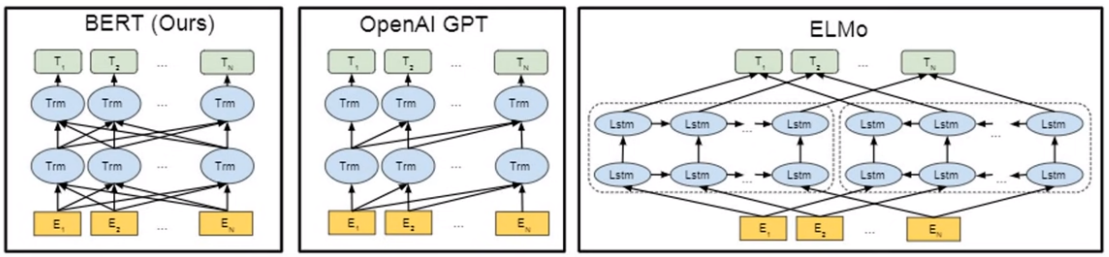


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

Fig2: latest models

1. BERT: Bidirectional Encoder Representations from Transformers

[CLS] I'm going to [MASK] play cricket. [SEP] Would you like [MASK] to play? [SEP]

- Predict [MASK] using information from left and right.

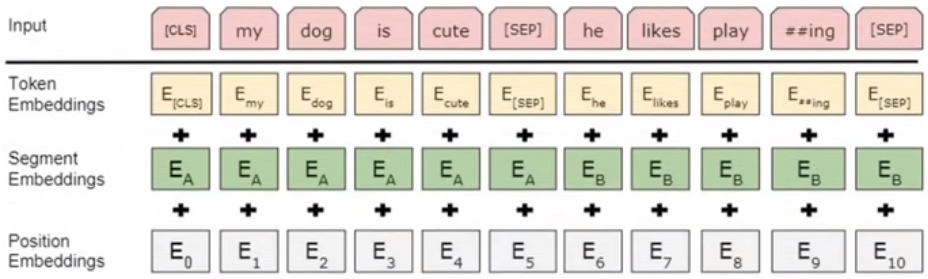


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Input: input sentence
Token embeddings : word wise embeddings
Segment embedding: label of setence - A and Sentence- B

1.1 Bert Sequence Classification
Reference 1
Reference 2
Workpeace embedding: `['Who', 'is', 'S', '##hak', '##a', 'Khan', '?']`

1.2 Entity Extraction

2. BioBERT

Paper: https://arxiv.org/abs/1901.08746

BioBERT is a domain-specific language representation model pre-trained on large-scale biomedical corpora.

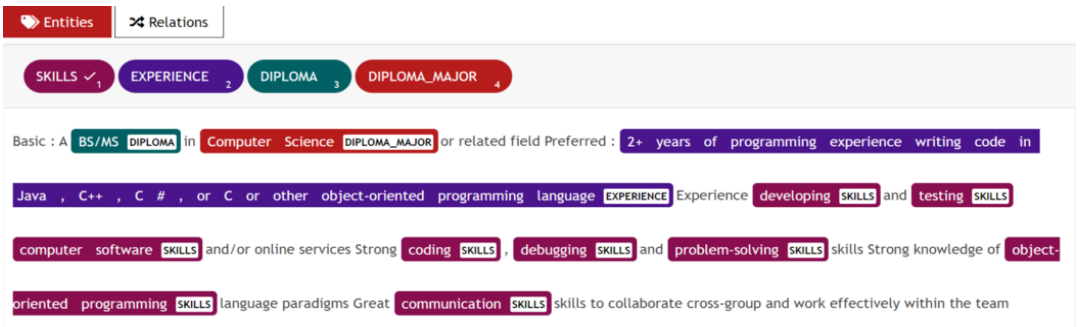
BioBERT embeddings:

- Token and word or sentence level embeddings from BioBERT model (Biomedical Domain).

2.1 Biomedical text classification
2.2 Entity extraction

3. fine-tune BERT using spaCy 3

- Requires IOB format json data
- https://ubiai.tools/ tool is mostly used to create annotated data



- Convert spacy annotation in [iob format]

4. CNN classification

Paper: https://ascopubs.org/doi/full/10.1200/CCI.20.00020
Problem statement:
curate/extract: presence of any cancer, cancer progression/worsening and cancer response/improvement from medical oncologist notes.

Steps:

1. RNN was trained using unlabeled notes to extract the assessment/plan from each note.
2. CNN were trained on labeled assessments/plans to predict the probability that each curated outcome was present.
3. overall survival were measured using Cox models among patients receiving palliative-intent systemic therapy

Paper approach:

Expected Output - interested in only yes(based on rule-based) or no.

Steps:

- Rule based method used to define target variable (assign labels to the data)
  - Dictionary based approach
  - Annotation
  - NLP methods - Regex
- CNN for classification of defined labels
  1. **identity assessment plan**
     - rule based method(dictionary of words)
     - RNN trained after removing key phrases - to predict whether each remaining word in each note was part of the assessment/plan
     - model can detect plan/asses without having key phrases
  2. **predict presence and status of cancer and progression/worsening**
     - CNN to predict cancer is there or not
     - whether cancer was progressing or worsening
     - whether cancer was responding or improving
     - lasso linear regression is used to predict outcome of CNN

Medical record curation:-

For each medical oncology note starting from diagnosis of lung cancer, curators reviewed each note to determine whether cancer was present and, if so, whether cancer was “improving/responding,” “stable/no change,” “mixed,” “progressing/worsening/enlarging,” or “not stated/indeterminate” in comparison with the most recent previous assessment. Curators were instructed to evaluate the “assessment/plan” section of the note only.

Identification of Assessment/Plan: RNN

- rules-based classifier was applied, in which the first occurrence of any of the following key phrases were used to define the beginning of the assessment/plan for a given note: “a/p,” “assessment/plan,” “assessment,” “assessment and plan,” “impression and plan,” “in summary,” and “plan.”
- words in each note were then divided into those occurring before the key phrase (not part of the assessment/plan) and those occurring after the key phrase (part of the assessment/plan)
- The key phrase was then removed from each note, and a recurrent neural network was trained to predict whether each remaining word in each note was part of the assessment/plan,

Performance of pregressing/worsening :

- The performance of this model for evaluating cancer progression and response was superior compared with results obtained when training the CNN against the entire medical oncology note, without limiting the input to the assessment/plan;
- performance for evaluating the presence of any cancer was similar for both methods

What we have:

1. Dictionary/Regex based curated notes where we’re assigning the label
    a. Progression - yes
    b. Progression - no
    c. other