NLP SUMMIT HEALTHCARE

Patient Similarity through Representation Learning from Medical Records

Nasser Ghadiri

Associate Professor | Isfahan University of Technology

Research Fellow | Western Sydney University

Talk Outline

1. Patient Similarity through Representation Learning from Medical Records
    - ➔ Focusing on the temporal aspect at different levels of detail
    - ➔ Two downstream tasks

2. Applying unsupervised key-phrase methods on concepts extracted from discharge sheets
    - ➔ Using pre-trained language models

Patient Similarity through Representation Learning from Medical Records

Team members

Hoda Memarzadeh
PhD student

Nasser Ghadiri
Isfahan University of Technology | IUT · Department of Electrical and Computer Engineering
Associate Professor
Machine learning, Biomedical text mining

Maryam Lotfi
15.44 · Assistant Professor
University of Isfahan, Shahreza campus, com| engineering group

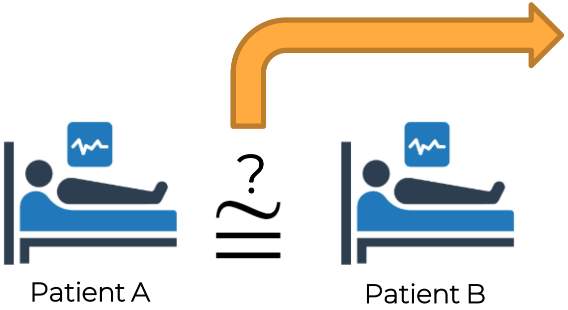
Assoc.-Prof. Dr. Matthias Samwald
Medical University of Vienna
Verified email at meduniwien.ac.at - Homepage
biomedical informatics   artificial intelligence

+ Medical experts who helped us in annotation of data in the Phenotyping Project

NLP SUMMIT

The Problem: Patient Similarity

- How similar patients are to each other based on their Electronic Health Records (EHR)
- A key mechanism with diverse applications



EHR A vs. EHR B

Precision medicine
Tailor treatments and interventions to individual patients

Healthcare resource allocation
By clustering patients ➔ improve efficiency, reduce costs, and enhance patient outcomes

Disease surveillance
Identify clusters of cases, monitor disease spread, and develop targeted interventions

Patient stratification
Identify patient subgroups ➔ different treatment approaches or targeted interventions

Challenge 1: Unstructured data in EHR

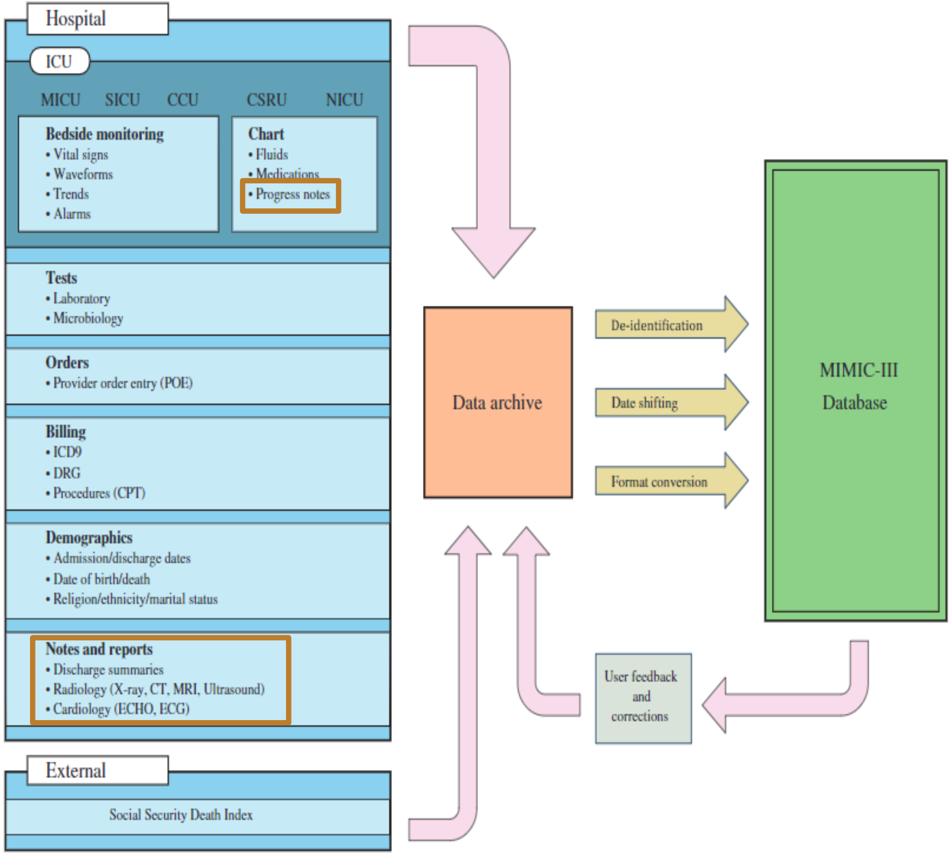
- Structured data, including:
- ICD codes
- Laboratory results
- Medications

➔ Simpler processing

- Unstructured data, including:
- Clinician progress notes
- Discharge summaries

➔ Needs complex processing - NLP

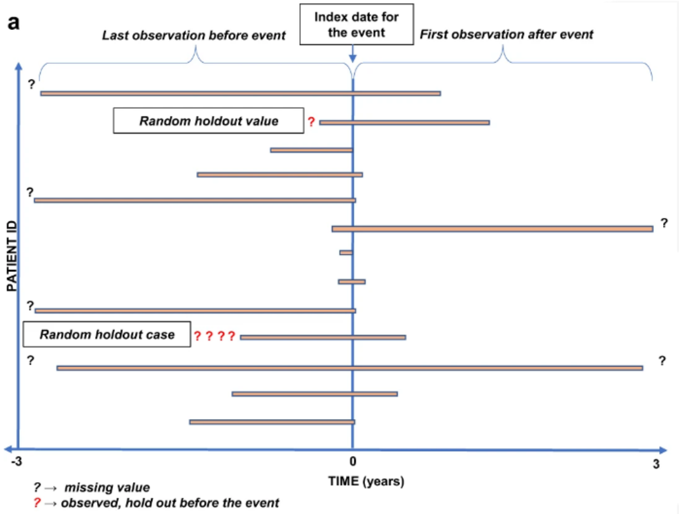
The **temporal** aspect of the recorded events



https://www.nature.com/articles/sdata201635

The Temporal aspect of EHR

- Enables longitudinal patient care: Track patients' health status and treatment progress over time
- Facilitates clinical decision-making: A comprehensive record of a patient's health history
  ➔ Clinicians make more informed decisions
- Supports research and quality improvement: identifying trends in disease prevalence or treatment outcomes
- Supports regulatory and legal requirements: Ensures that a complete record of a patient's health history is available to support these requirements



https://www.nature.com/articles/s41746-021-00518-0/figures/1

NLP SUMMIT

Challenge 2: Temporal levels of detail

- Long-term trends: Analyzing EHR data over longer time periods, such as years or decades ➔ identify long-term trends in disease prevalence, treatment outcomes, and resource utilization
- Medium-term outcomes: Analyzing EHR data over months or years ➔ provide insights into patient outcomes, including treatment effectiveness and disease progression
- Short-term changes: Analyzing EHR data over days or weeks ➔ provide insights into acute changes in patient health status and treatment responses.
- Real-time monitoring: Identify and respond to acute changes in patient health status, such as in the case of intensive care units or emergency departments

The challenge: What if multiple levels of time need to be integrated for some tasks?

NLP SUMMIT

Some existing methods for the two challenges

- Research

    Temporal Tree (Pokharel et. al 2020) / Univ of Queensland : only structured EHR data

    TAPER (Darabi et. Al 2020) / UCLA : structured and unstructured parts of EHR, but isolated representations

- Industry

    Partial support of modeling temporal data by some tools like:

    Amazon Comprehend Medical

    https://aws.amazon.com/about-aws/whats-new/2020/03/announcing-time-expression-for-amazon-comprehend-medical/

    SparkNLP and Google Cloud

    https://medium.com/spark-nlp/comparison-of-key-medical-nlp-benchmarks-spark-nlp-vsaws-google-cloud-and-azure-cab5619d2bf6

NLP SUMMIT

Filling the gap

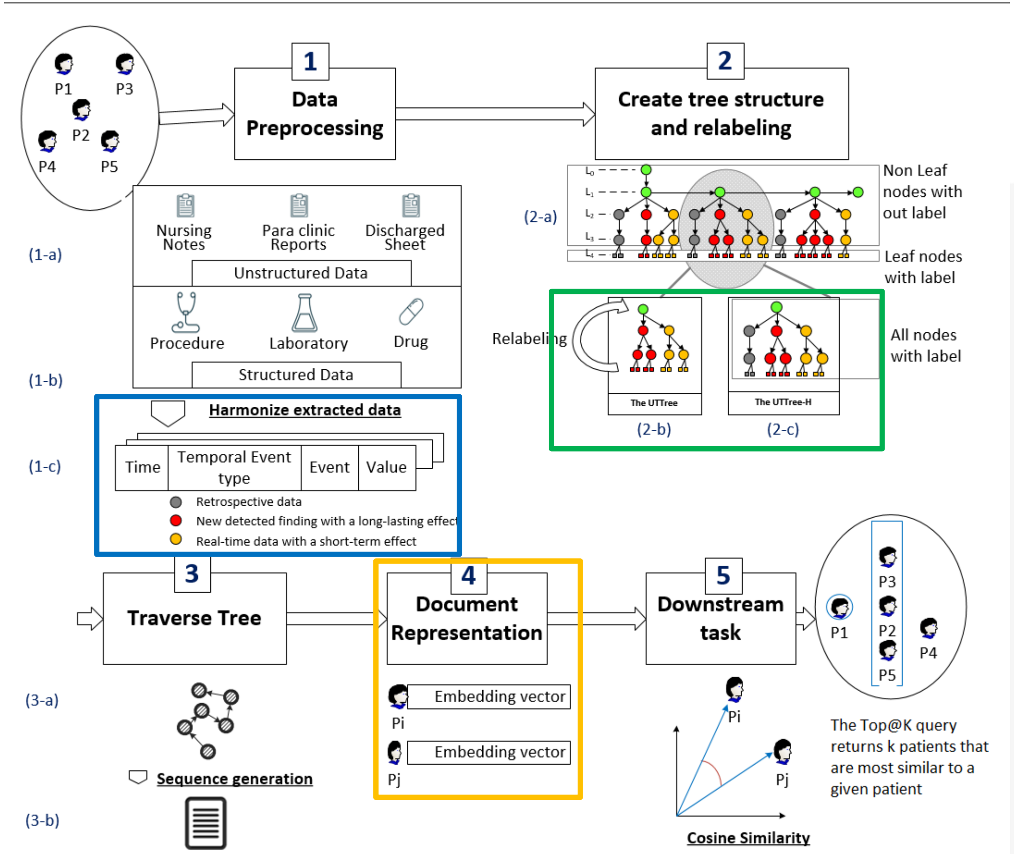
- Our model integrates the structured and unstructured data through the tree structure for the first time to produce a unified representation vector for both types of data, based on the temporal aspect
- Fewer studies used external knowledge sources, such as UMLS, in the clinical processing of notes. Our model is based on this enriched processing
- Previous studies: same weight to all parts of clinical notes

  Family history, illness history, and current referral are not equally important.

  ➔ We improve this by computing and assigning different weights to the EHR sections

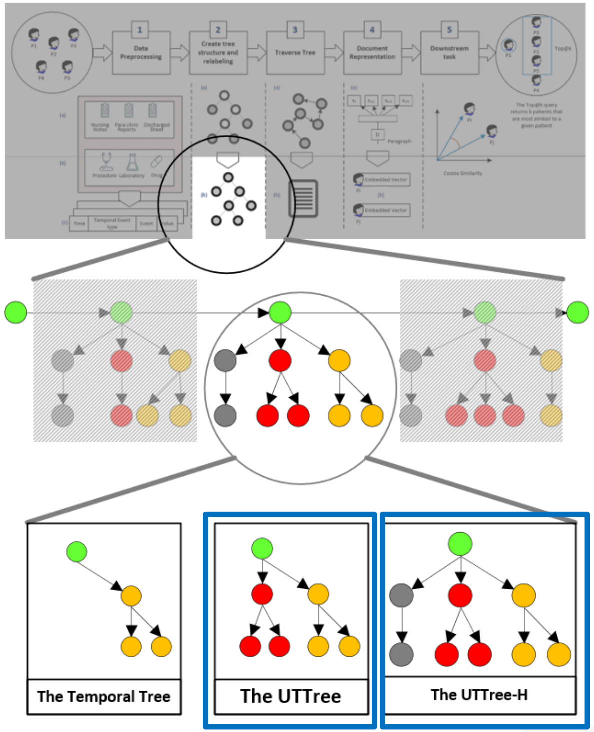
The pipeline

1. Harmonizing EMR data for each patient as quadruples
2. Tree construction and the relabeling processes
3. Tree traversal to build rich sequences
4. Generate embedding vectors (*currently*: doc2vec, see future work)
5. Computing patient similarity



The new temporal trees

- UTTree: a novel representation for EMR that integrates unstructured and structured data in tree-based structured / Create sequences using a new relabeling approach
- UTTree-H: Enriching the UTTree model with the historical data in EMR
- Evaluated the produced embedding on downstream tasks including patient similarity and mortality prediction



Relabeling illustrated

- The non-leaf nodes of the **higher levels** are labeled in order (based on *Weisfeiler-Lehman graph kernels*)
- Relabeling in UTTree uses the current visit information
- Relabeling UTTree-H adds the past **history** of the patient

Quadruples that participate in the construction of the subtree(t_{i+1})

Time	Temporal Event Type	Event	Value
2018-7-30	Retrospective data with long-lasting effect	DiseaseDisorderMention	Diabetes mellitus type 2
2018-7-30	New finding with long-lasting effect	Demographic Category	F_4
2018-8-1	New finding with long-lasting effect	DiseaseDisorderMention	Diabetic neuropathy
2018-8-2	Real-time data with a short-term effect	MainDrug	Insulin
2018-8-2	Real-time data with a short-term effect	Glucose	High

Resulting temporal sequences

Seq1: Disease, Diabetesmellitustype2, Female,4, Glucose, High, MainDrug, Insulin, Disease, Diabetesmellitustype2, Female,4, Disease, Diabetneuropathy, Gloucose, High, MainDrug, Insulin, Disease, Diabetesmellitustype2, Female,4, Disease, Diabetneuropathy, Symptom, Fatique, Glocose, Normal

Seq2: DiseaseDiabetesmellitustype2, Female4, GlucoseHigh, MainDrugInsulin, DiseaseDiabetesmellitustype2, Female4,DiseaseDiabetneuropathy, GloucoseHigh, MainDrugInsulin, DiseaseDiabetesmellitustype2, Female4, DiseaseDiabetneuropathy, SymptomFatique, GlocoseNormal,

Seq3: DiseaseDiabetesmellitustype2, Female4, GlucoseHighMainDrugInsulin, DiseaseDiabetesmellitustype2, Female4DiseaseDiabetneuropathy, GloucoseHighMainDrugInsulin, DiseaseDiabetesmellitustype2, Female4DiseaseDiabetneuropathySymptomFatique, GlocoseNormal,

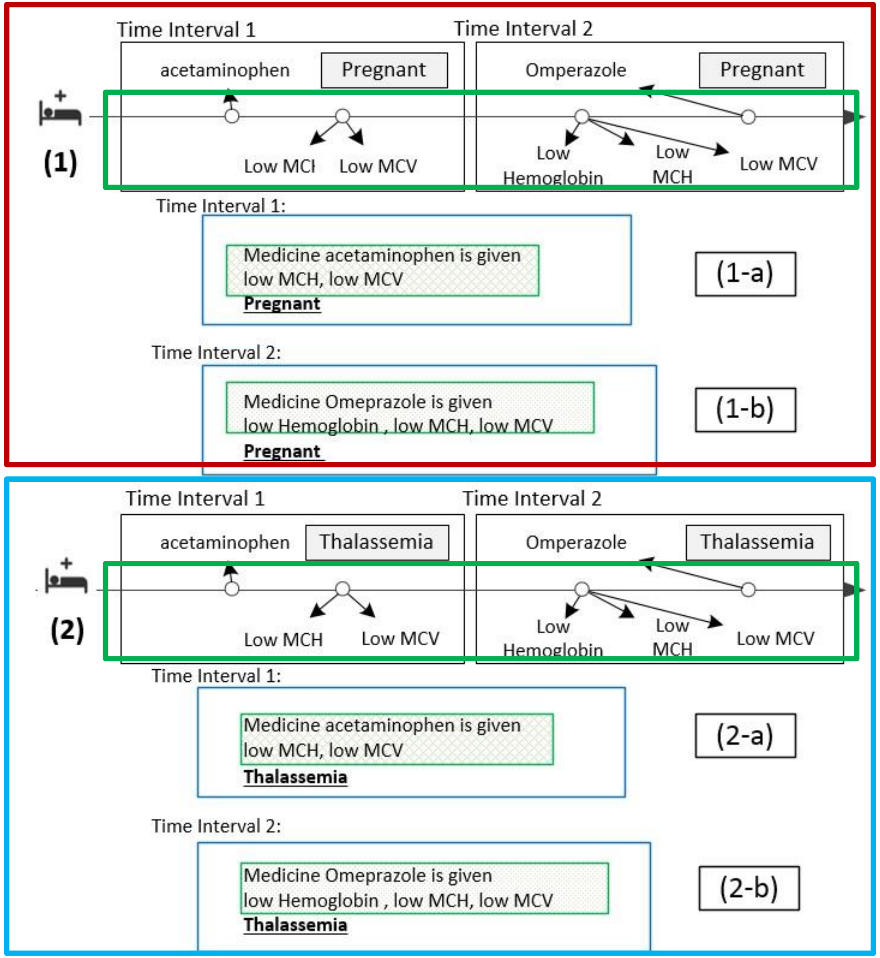
Seq4: DiseaseDiabetesmellitustype2Female4GlucoseHighMainDrugInsulin, DiseaseDiabetesmellitustype2Female4DiseaseDiabetneuropathyGloucoseHighMainDrugInsulin, DiseaseDiabetesmellitustype2Female4DiseaseDiabetneuropathySymptomFatiqueGlocoseNormal,

- Temporal sequences in the BFS order generated for the tree with four levels
- All leaf and non-leaf nodes of the tree are labeled in step 3 of the workflow
- A sequence is generated for each tree level when the tree is traversed in the BFS order
- The sequences combine to form the final document for each patient

UTTree-H: adding patient history

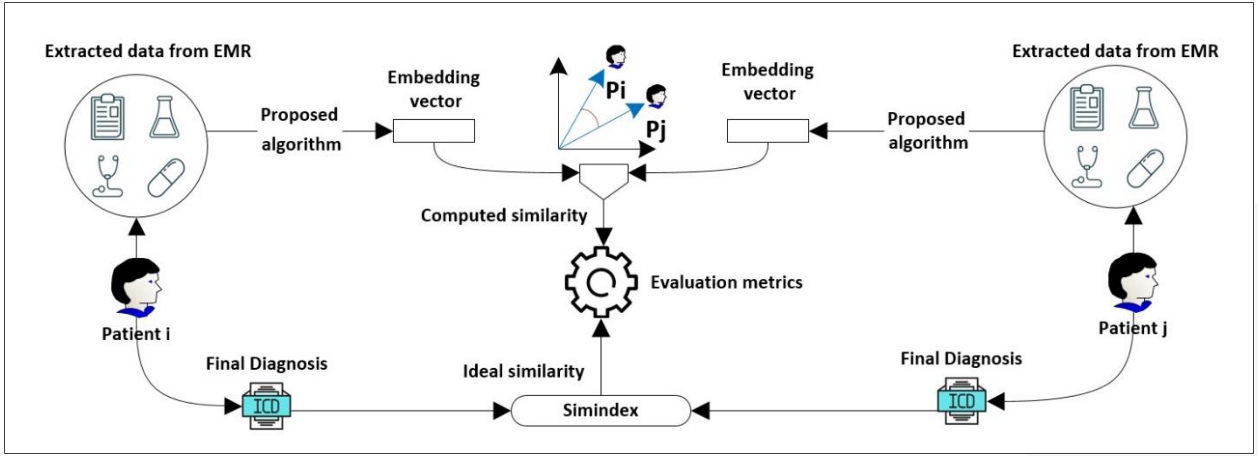
- Patient 1 is pregnant
- Patient 2 has thalassemia
- Have similar clinical manifestations at the time of admission
- But they require different treatment methods due to their different medical histories

- UTTree-H incorporates the patient history into the creation of labels, resulting in a unique final sequence for each patient.



NLP SUMMIT

Evaluation of patient similarity



- Using all available **final diagnosis codes** and assigning a weight to each code
- based on diagnosis code priority in each patient's EMR
- Cosine angle between two **embedding vectors** is calculated for each patient
- Compared to the gold standard's ideal similarity

NLP SUMMIT

Evaluation – patient similarity

- Resulted in:

Lower mean-squared error (**MSE**)

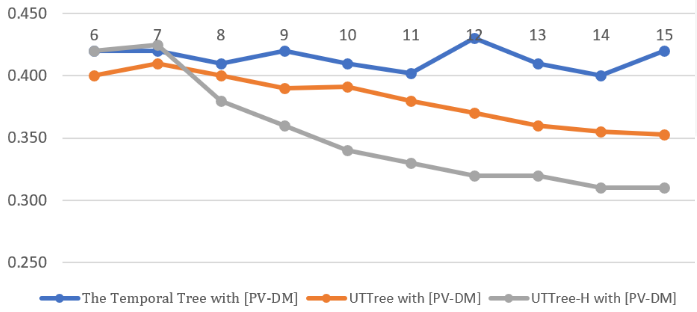
Higher precision, and normalized discounted cumulative gain (**NDCG**) relative to baseline

Two downstream tasks: **patient similarity** and **mortality prediction** (next slide)

The effect of the **number of extracted words** on the error rate

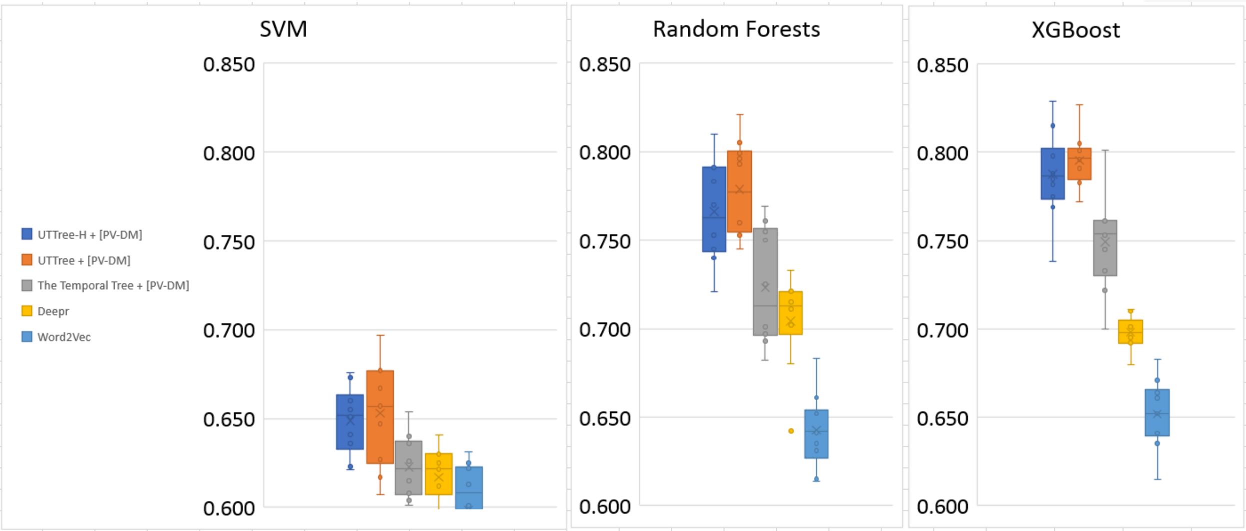
UTTree-H

MSE	MSE@1	MSE@5	MSE@10	MSE@20
TFIDF_Structured and Unstructured data	0.310	0.128	0.078	0.050
UTTree _ [PV-DBOW]	0.239	0.096	0.069	0.045
UTTree-H _ [PV-DBOW]	0.234	0.094	0.061	0.039
UTTree _ [PV-DM]	0.235	0.091	0.064	0.040
UTTree-H _ [PV-DM]	0.232	0.093	0.061	0.038
nDCG	**nDCG@1**	**nDCG@5**	**nDCG@10**	**nDCG@20**
TFIDF_Structured and Unstructured data	0.421	0.419	0.412	0.406
UTTree _ [PV-DBOW]	0.481	0.471	0.449	0.431
UTTree-H _ [PV-DBOW]	0.491	0.481	0.449	0.437
UTTree _ [PV-DM]	0.495	0.483	0.455	0.438
UTTree-H _ [PV-DM]	0.492	0.482	0.454	0.437



NLP SUMMIT

Evaluation - mortality prediction



- Predicting mortality in MIMIC-III patients
- UTTree-H and UTTree outperform other methods
- The XGboost classifier performed better

NLP SUMMIT

Limitations and future directions

- This work focuses on the use and processing of clinical text data ➔ Potential for enhanced representation methods (autoregressive, GPT, etc.) *
- At a higher level, integrating multimodal data (such as clinical, imaging, and molecular profiling) is crucial to understand complicated diseases and providing accurate diagnoses

Knowledge and Information Systems
https://doi.org/10.1007/s10115-022-01740-2

REGULAR PAPER

A study into patient similarity through representation learning from medical records

Hoda Memarzadeh[1] · Nasser Ghadiri[1] · Matthias Samwald[2] · Maryam Lotfi Shahreza[3]

Received: 8 June 2022 / Revised: 24 July 2022 / Accepted: 7 August 2022

https://link.springer.com/article/10.1007/s10115-022-01740-2

Source code:
https://github.com/HodaMemar/Patient-Similarity-through-Representation

- For the **biomedical** domain, GPT-3 underperformed in-domain pretraining such as BioBERT - See: Moradi, Milad, et al. "Gpt-3 models are poor few-shot learners in the biomedical domain." *arXiv preprint arXiv:2109.02555* (2021).

NLP SUMMIT

Applying unsupervised key-phrase methods on concepts extracted from discharge sheets

The problem

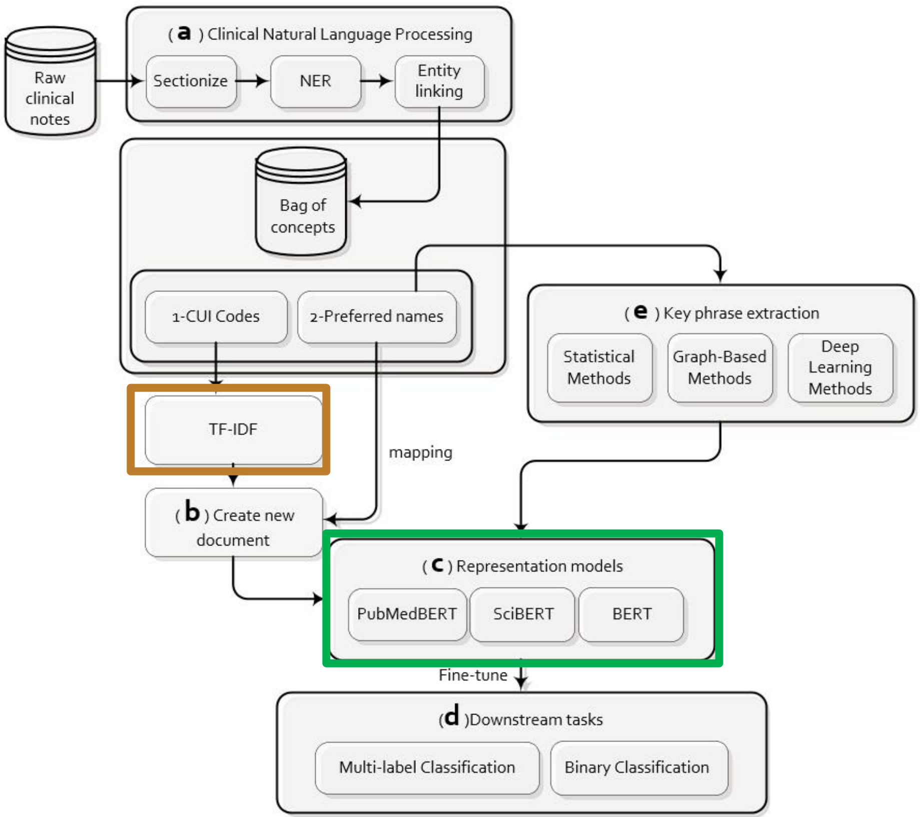
- Clinical notes: various scientific levels and writing styles
- Named Entity Recognition and entity linking are critical steps BUT they can produce repetitive and low-value concepts
- The need to identify the section in which each content is recorded and critical concepts to extract meaning from clinical texts

Image

The solution

Most clinical concepts are in the form of multi-word expressions ➔ accurate identification requires the user to specify an n-gram range

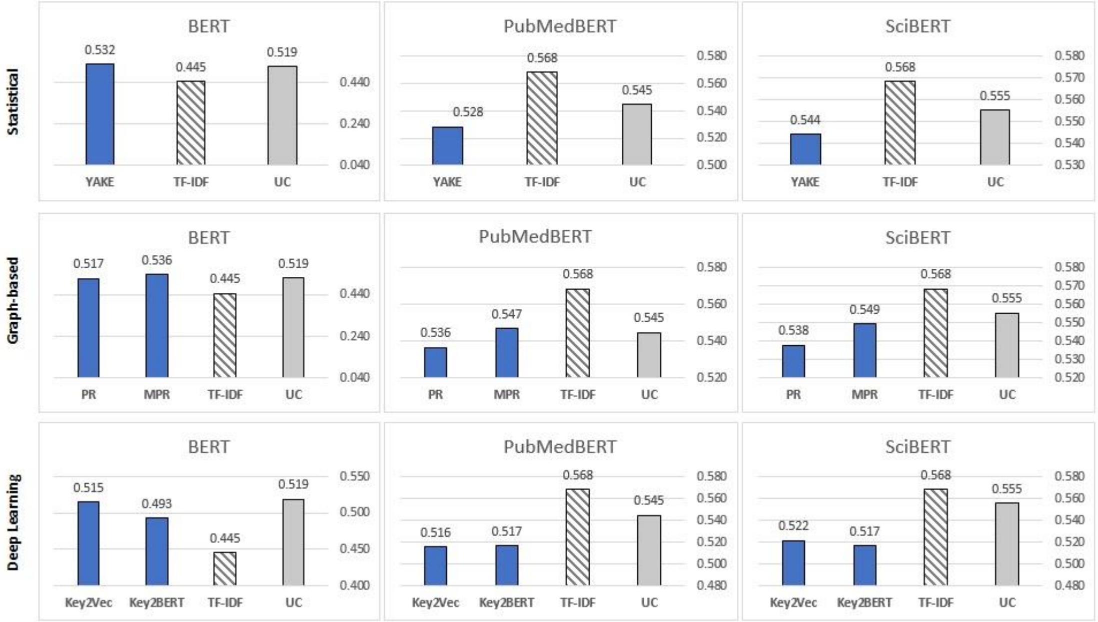
a) Discharge sheet ➔ converted to a bag of concepts in the NLP pipeline
b) The dataset of CUI concepts codes is processed by the TF-IDF algorithm to detect ones with scores above the threshold (higher scores for key concepts)
c) Three clinical transformer-based representation models fine-tune the generated dataset
d) Two types of downstream tasks (multiple and binary classifications) using the capabilities of transformer-based models
e) Compare with keyphrase extraction methods that directly run



NLP SUMMIT

Results – multi-label classification

- The proposed method's superiority was shown in combination with the SciBERT model
- The results offer an insight into the efficacy of general extracting essential phrase methods for clinical notes

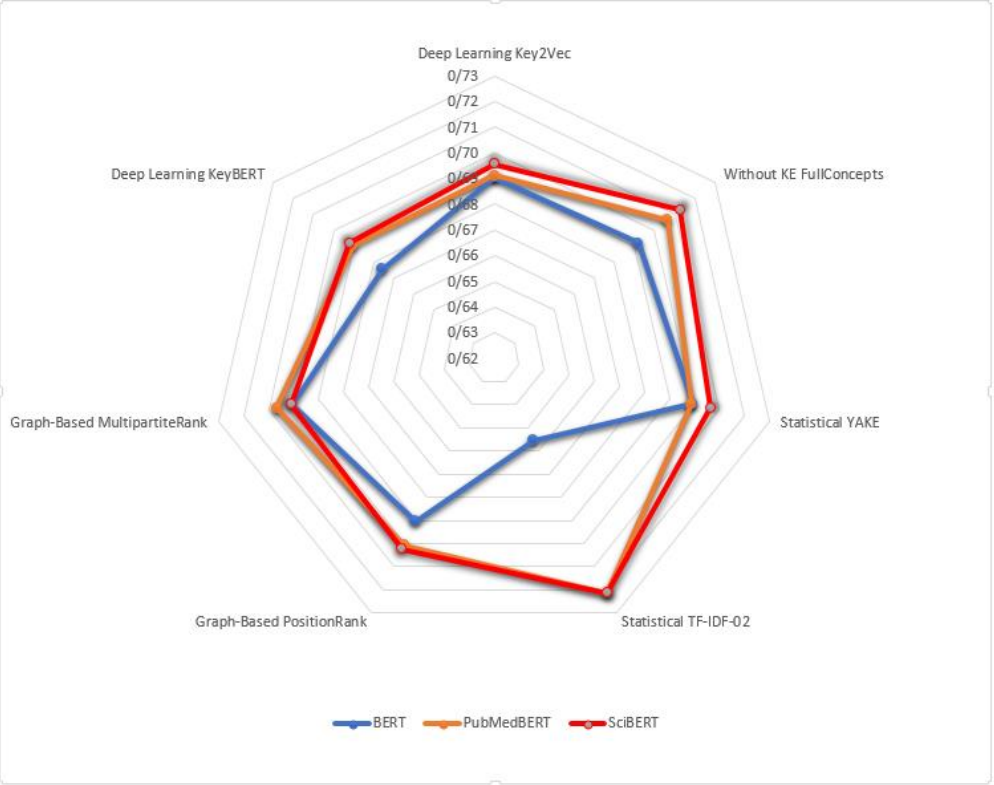


F1 measure in multi-label classification using unsupervised keyphrase extraction methods and fine-tuned clinical transformer-based models

Results – combinations with transformer models

- The performance comparison of all input datasets in combination with transformer-based models

- Best F1 results were obtained by combining **statistical** keyphrase extraction with the **SciBERT** representation model



NLP SUMMIT

Output of other methods

Method	Output
CUIs	C0002893,C0007012,C0027270,C0032326,C0013604,C0037197,C0011777,C1264639,C0017887,C1761613,C0700124,C0201950,C0018824,C1272695,C0010068,C0336779,C0023031,C0026266,C0032227,C3275121,C0264956,C0032285,C0439688,C0301362,C0005367,C0010054,C0031039,C0024129
Preferred Names of CUIs	Refractory anemias, carbon dioxide, nicotinamide adenine inucleotide (nad), pneumothorax, edema, structure of sinus of valsalva, dexamethasone, date/time, nitroglycerin, conjunctival hyperemia, dilated, cholesterol measurement test, heart valve disease, done (qualifier value), coronary heart disease, machine, lanthanum, mitral valve insufficiency, pleural effusion disorder, one vessel coronary disease, atheroma, pneumonia, atelectatic, bromdimethoxyamphetamine, bicarbonates, coronary arteriosclerosis, pericardial effusion, lung volume measurements,
Keybert n-gram=4	atelectatic bromdimethoxyamphetamine bicarbonates coronary, bromdimethoxyamphetamine bicarbonates coronary, bromdimethoxyamphetamine bicarbonates coronary arteriosclerosis, bicarbonates coronary arteriosclerosis pericardial, bicarbonates coronary arteriosclerosis, coronary heart disease, cholesterol measurement test heart, bicarbonates coronary, coronary disease, qualifier value coronary heart, pneumonia atelectatic bromdimethoxyamphetamine bicarbonates, measurement test heart, heart disease machine lanthanum, cholesterol measurement test, arteriosclerosis pericardial effusion lung, nad pneumothorax, valve disease done qualifier, nad pneumothorax edema, cholesterol measurement, heart disease,
Keybert n-gram=3	bromdimethoxyamphetamine bicarbonates coronary, bicarbonates coronary arteriosclerosis, coronary heart disease, bicarbonates coronary, coronary disease, measurement test heart, cholesterol measurement test, nad pneumothorax, nad pneumothorax edema, cholesterol measurement, heart disease, coronary arteriosclerosis pericardial, pericardial effusion lung, qualifier value coronary, coronary heart, coronary arteriosclerosis, arteriosclerosis pericardial, disease done qualifier, arteriosclerosis pericardial effusion, pericardial effusion,
Keybert n-gram=2	bicarbonates coronary, coronary disease, nad pneumothorax, cholesterol measurement, heart disease, coronary heart, coronary arteriosclerosis, arteriosclerosis pericardial, pericardial effusion, test heart, bromdimethoxyamphetamine bicarbonates, pericardial, coronary, insufficiency pleural, cholesterol, pneumonia atelectatic, heart valve, effusion lung, pneumothorax, arteriosclerosis,
Keybert n-gram=1	pericardial, coronary, cholesterol, pneumothorax, arteriosclerosis, nicotinamide, bicarbonates, anemias, bromdimethoxyamphetamine, nitroglycerin, lung, mitral, pleural, hyperemia, lanthanum, edema, insufficiency, pneumonia, atelectatic, disease,

- The KeyBERT method with **different n-grams**.
- Includes vocabulary that is trained on many public documents
- Most of the generated phrases are meaningless.

NLP SUMMIT

Output of proposed method

Tokenizer	Vocabulary	Pre-train	Token for “coronary arteriosclerosis”
BERT	BERT-BASE		[‘corona’, ‘##ry’, ‘arte’, ‘##rio’, ‘##sc’, ‘##ler’, ‘##osis’]
SciBERT	SciVocab	Papers from the biomedical domain and computer science	[‘coronary’, ‘arterios’, ‘##cle’, ‘##rosis’]
Bio+Clinical	BioBERT BERT-BASE	All MIMIC III, Only the discharge summaries in MIMIC III	[‘co’, ‘##rona’, ‘##ry’, ‘art’, ‘##eri’, ‘##os’, ‘##cle’, ‘##rosis’]
BLUEBERT	BERT-BASE	PubMed abstracts, clinical notes from the MIMIC III dataset	[‘corona’, ‘##ry’, ‘arte’, ‘##rio’, ‘##sc’, ‘##ler’, '##osis']
PubMed-BERT	BERT-BASE	PubMed (abstracts and full biomedical articles) (3.1B words)	['coronary', 'arteri', '##osclerosis']
UMLS-BERT	Bio+Clinical-BERT	Patient notes and diagnostic test reports from the MIMIC III	['co', '##rona', '##ry', 'art', '##eri', '##os', '##cle', '##rosis']

Comparison of Token for “*coronary arteriosclerosis*” in vocabularies used by the standard BERT, SciBERT, and PubMedBERT.

Paper: http://arxiv.org/abs/2303.08928
Source code: https://github.com/HodaMemar/A3

NLP SUMMIT

Takeaways

- Patient Similarity: A key mechanism for many tasks in the healthcare system

➔ The challenges of unstructured textual data and multi-level temporality addressed by UTTree, UTTree-H

- Applying unsupervised key-phrase methods on concepts extracted from discharge sheets

➔ Assigning weights to extract more important concepts

NLP SUMMIT

Thank you!

Nasser Ghadiri

nghadiri@gmail.com

in nasserghadiri

NLP SUMMIT HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org