Introduction

My background

- Clinical
  - Community health volunteer: Worked with various patient populations, e.g., prison population, LGBTQ+ patients, and also geriatric patients, providing HIV/HCV testing and counseling for over 5 years
- AI/NLP
  - Artificial Intelligence Scientist in Biotech/Pharma for 2 years (the vants) <- MLE at ESI <- speaker at Grace Hopper Celebration in 2018
  - Graduated from NYU with thesis MS in Biomedical Informatics in 2018 (*Zero Start: Deep Learning Models to Predict End of Life from Clinical Text)* and was contracted to deploy this work at various NYU Langone hospitals

Some topics you can expect from today:

Clinical NLP - unsupervised learning, text classification, named entity recognition, machine translation, various deep learning architectures (e.g., 1D CNNs, Seq2Seq, BERT-based and more!), data augmentation and generative type models, model deployment, clinical significance & collaboration (if time permits)

YEAR 2016

NIH Endowment Scholarship

The purpose of the scholarship is to provide financial support to high achieving, qualified students from socially or economically disadvantaged groups as defined by the National Institutions of Health (NIH).

I joined a top ten pharmacy program and was involved in many extracurriculars as well as wet lab research.


THE DEFINING
MOMENT

Deep Learning Models to Predict End of Life from Clinical Text

Isabel Kayu Metzger, MS , Seda Bilaloglu, MS , Vincent J Major, MS, Himanshu Grover, PhD, Yindalon Aphinyanaphongs, MD/PhD
Department of Population Health, NYU Langone Health, New York, NY

Abstract

Accurate prognosis upon admission can help patients with serious diseases and their families to receive the care they need as they approach end of life. Admission notes hold key descriptions related to illness or problems and can be used to predict the course of the disease using machine learning techniques. In this paper, we explored deep

AMIA (2018)

SMM4H Shared Task 2020 - A Hybrid Pipeline for Identifying Prescription Drug Abuse from Twitter: Machine Learning, Deep Learning, and Post-Processing

Isabel Metzger[1,5], Emir Y. Haskovic[2], Allison Black[3], Whitley M. Yi[4], Rajat S. Chandra[1], Mark T. Rutledge[1], William McMahon[1], and Yindalon Aphinyanaphongs[5]

[1]Sumitovant Biopharma, Inc., New York, NY
[2]Lokavant, New York, NY
[3]Roivant Sciences, Inc., New York, NY
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, CO
[5]Department of Population Health, NYU Langone Health, New York, NY
{*isabel.metzger, rajat.chandra, mark.rutledge, bill.mcmahon*}*@sumitovant.com,*
emir.haskovic@lokavant.com,
allison.black@roivant.com,
whitley.yi@ucdenver.edu,
yin.a@nyulangone.org

Association for Computational Linguistics, ACL (2020)

Expanding the Reach of Structured EHR Data with Clinical Notes: Improving End-of-Life Prediction

Seda Bilaloglu, Vincent Major, Himanshu Grover, Isabel Kayu Metzger and Yindalon Aphinyanaphongs
Department of Population Health, NYU Langone, New York, NY

Abstract

Appropriate treatment decisions and end-of-life planning for patients with serious, life-limiting disease rely on accurate prognosis estimates. Many existing methods rely on structured electronic health record data which may limit generalizability across sites and restrict performance for patients with less documented history. Clinical notes may help to 'level the playing field'. We use History and Physical (H&P) notes written within 16 hours of hospitalization to predict 60-day, all-cause mortality. We test several neural network approaches and observe little improvement over a CNN by adding bi-directional recurrence or convolutional attention. The CNN was prospectively validated against an existing system using structured data. The CNN reports

care planning, code status or advance directives, these patients may receive unwanted aggressive treatment. Precise identification of high-risk patients can break this cycle by encouraging appropriate end-of-life care.

Clinical risk tools often provide a score (Charlson et al. 1987; Knaus et al. 1985; Morita et al. 1999) to stratify patients into risk groups. Numerous machine learning methods also exist but many are limited to specific populations by disease or acuity (Ghassemi et al. 2014; Makar et al. 2015; Elfiky et al. 2017; Parikh et al. 2019). Several general approaches have been proposed for use to prompt clinicians to consider end-of-life planning (Avati et al. 2018; Wegier et al. 2019; Courtright et al. 2019; Major and Aphinyanaphongs 2020). Each of these works rely on structured electronic

Flairs Proceedings, AAAI 2021

Question to SQL Query:
A Clinical Natural Language Processing Interface over Electronic Health Records Database

Isabel Metzger[1], Whitley Yi, PharmD[2]
New York University[1], NYU Langone Health[2], UNC Health Care[2]

ABSTRACT

Patient and clinician-generated narratives are considered "noisy text", filled with domain-specific abbreviations and misspellings. Regardless, the presence of noise allows for knowledge discovery. When incorporated into machine learning based systems, along with human clinical expertise, it has the potential to aid in clinical decision making and directly influence patient care. **An interface across the electronic health record (EHR) that would translate questions from "natural language" to SQL queries, while accounting for domain-specific abbreviations, arises from realizing the need for a tool to aid clinicians in obtaining information for critical patient care decisions faster.** For this reason, we built a prototype interface on the Medical Information Mart for Intensive Care III (MIMIC III) database, a publicly-available database of EHRs of critical care patients. We created a corpus via crowdsourcing of what is now approximately 4,000 questions (from clinician, research, and financial investigator perspectives) for data and statistics on interactions between patient demographics, treatments, comorbidities, and more. This dataset was manually extended by annotating the natural language questions with SQL which is used to query the MIMIC database. Medical concepts are recognized and normalized using a named entity recognition (NER) model pre-trained on requested annotated

METHODS

Named Entity Recognition → Seq2Seq with Attention → Connect to DB and Eval

How many patients died with <PROBLEM_1> after receiving <PROCEDURE_1>?

```
select count(distinct subject_id)
from PATIENTS
where expire_flag=1 and subject_id IN
  (select distinct subject_id
  from DIAGNOSES_ICD
  where icd9_code IN
    (select distinct icd9_code
    from D_ICD_DIAGNOSES
    where long_title like '%PROBLEM_1%')
and subject_id IN
  (select distinct subject_id
  from PROCEDURES_ICD
  where icd9_code IN
    (select distinct icd9_code
```

PROBLEM_1: Sepsis
PROCEDURE_1: Catheter

DB → 43

RESULTS & CO

- Bi-LSTM CRF on
  - Accuracy = 0.9
  - F-1 Score = 0.8
- Bi-LSTM CRF on
  - Accuracy = 0.7
  - F-1 Score = 0.7
- Seq2Seq with att
  - Accuracy of ob
  - Accuracy of SC

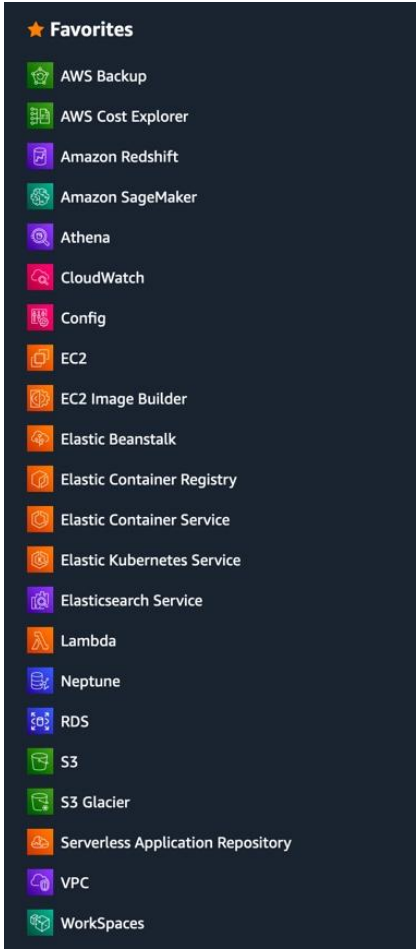
To understand the simulated experime batch is then seque pairs of natural lan

Northeast Health Summit hosted by IBM & Brown University (2019)

ML Ops Experience

AWS Tech Stack

- Training either on Virtual Machines (EC2) instance and Sagemaker
- Dockerized models (ECR)
- Storage in S3 - Model, Data, Predictions
- Amazon Workspace
- Create API for my models so that the digital innovators could integrate them into their dashboards as well

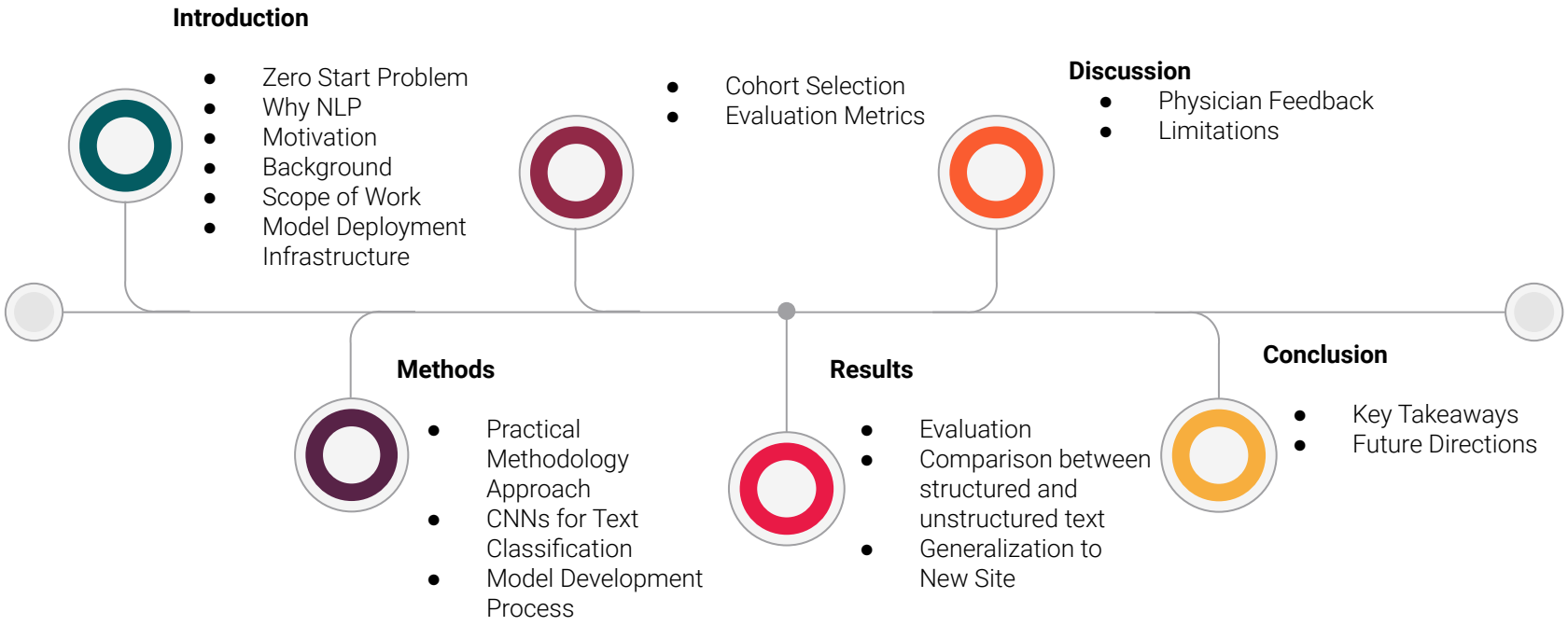


2017 - 2018

ZERO START:

Deep Learning Models to Predict Mortality from Clinical Text

Overview


Introduction
Zero Start Problem
Why NLP
Motivation
Background
Scope of Work
Model Deployment Infrastructure
Cohort Selection
Evaluation Metrics
Discussion
Physician Feedback
Limitations
Methods
Practical Methodology Approach
CNNs for Text Classification
Model Development Process
Results
Evaluation
Comparison between structured and unstructured text
Generalization to New Site
Conclusion
Key Takeaways
Future Directions

ZERO START PROBLEM refers to when there are no previous records about a patient who visits us and thus we cannot use any data besides the note made upon admission to the facility. EPIC (healthcare software) alert system will not be triggered in these situations.

- **the goal of this work: Can we predict 2-month mortality from unstructured data (clinical text), more specifically the History & Present illness (H&P) note?**

Motivation

Why is this work important to do?

Why is this novel?

Most mortality predictive models fall into two extreme categories:

- Short term: (1-6 days) (e.g., rapid deterioration)
- Long term (12 months) (palliative care)

This work predicts for 2-6 months.

Why is this useful?

- Physician agreement with a recommendation for palliative care is highest at a prediction window of 2 months.
- Research shows that interventional palliative care is optimal when administered at least 6 months prior to mortality.

Categories requested by physicians

Term	Definition
High Risk	Predicted to die within 2 months.
Appropriate	Expected to die within 6 monhts. GOC/ ACP warranted.
Inappropriate	Neither of the above.
Helpful	High Risk or Appropriate

Why NLP?

“Unstructured EHR” such as medical notes may provide unique insight and possibly more information than “structured”

A publication using the MIMIC III medical notes to **predict sepsis** within 24 hours found that the unstructured text data **performed better** than the structured tables. (Culliton, 2017)

We need to talk about death

Prognosis in Practice

- Clinicians are often inaccurate when predicting end-of-life (only right ~50% of the time). (Nicola White et. Al, 2016).

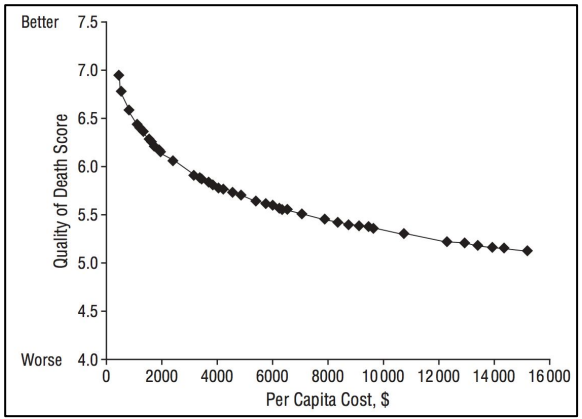
What is important to terminally ill patients?

- **“Not to be kept alive on life support when there is little hope for a meaningful recovery”** (55.7%)
- **“That information about your disease be communicated to you by your doctor in an honest manner”** (44.1%)
- **“To complete things and prepare for life's end — life review, resolving conflicts, saying goodbye”** (43.9%)

Costs

- A disproportionate amount of healthcare cost occurs in the last month of life (Zhang et al. 2009).
  - One study investigating the total cost of treatment in final week of life with, or without, an EOL conversation described:

    higher costs were negatively associated with caregiver reported quality of death, and EOL conversations reduced total cost by 36% ($2780 vs. $1925).



(Zhang et al. 2009)

Score Based Metrics

The problem:

- Points-based metrics discretize real world physiology into coarse, weighted bins.
  - Makes them easy to use and interpretable
  - Restricted to integer weights, a small number of variables, a small number of bins

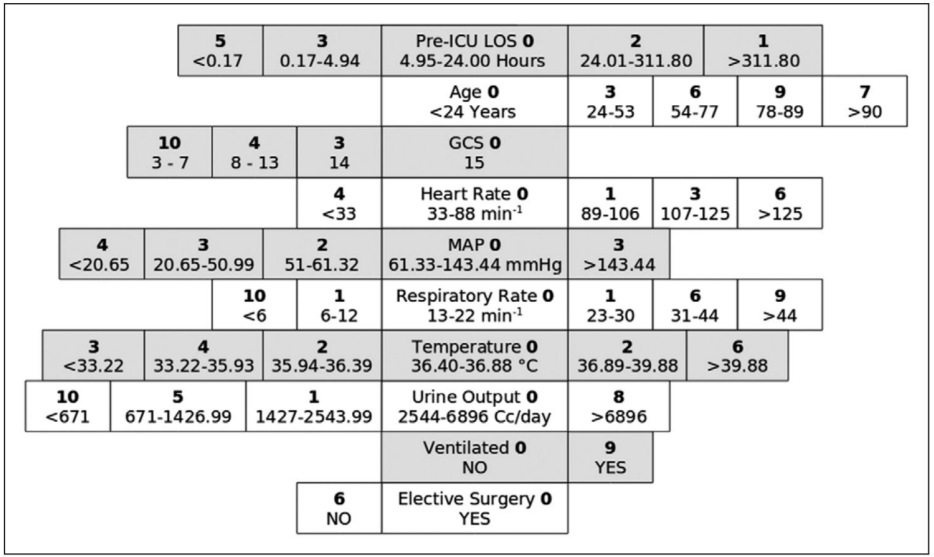


Figure 1. Component weights and bins for the Oxford Acute Severity of Illness Score (OASIS). The *bold values* are the individual scores assigned to an associated range of measured values. For each variable, the worst score across the first day should be used to tabulate OASIS. The final OASIS score is the sum of all the component weights. LOS = length of stay, GCS = Glasgow Coma Score, MAP = mean arterial pressure.

(Johnson et al. 2013)

Scope of Work

SCOPE OF WORK

CENTER FOR HEALTHCARE INNOVATION AND DELIVERY SCIENCE (CHIDS) of NYU LANGONE HEALTH is contracting with Isabel Metzger to build 2 month mortality machine learning based models from initial history and physical notes of admitted patients. Accurate mortality prediction with initial notes would allow models to be built for patients that do not have prior data. These models help providers and patients coordinate in delivering supportive care that align with patient wishes. Specifically we will

(1) integrate a model into our scalable text classification infrastructure,
(2) emit model classifications at a performance threshold daily to an inpatient team for feedback,
(3) deliver a publication ready writeup of the model and results.

Model Deployment Infrastructure

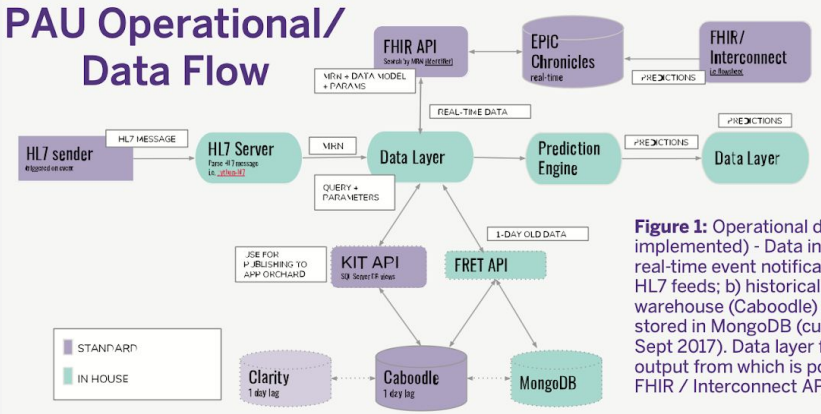


Figure 1: Operational data workflow (partially implemented) - Data infrastructure consumes a) real-time event notifications directly from Epic via HL7 feeds; b) historical data from Epic data warehouse (Caboodle) or Epic Clarity snapshot stored in MongoDB (currently has data from 2014-Sept 2017). Data layer feeds the prediction engine, output from which is populated back into Epic via FHIR / Interconnect API.

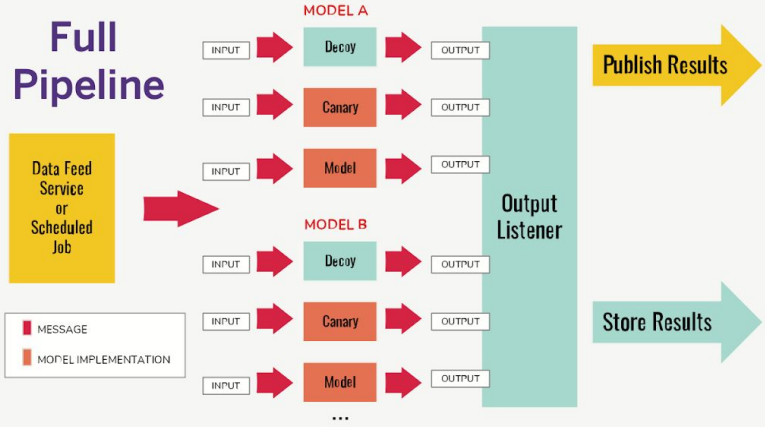


Figure 2: Production Model Deployment Architecture. Models implement a standardized interface and are deployed as containerized services that: a) consume input data stream; b) publish results for downstream services to listen and act on. Production models can be supplemented with Decoys (to capture raw input for later re-use) and Canaries (to monitor drift in model or data). Streaming also facilitates comparing new test models that consume same input stream as production model.

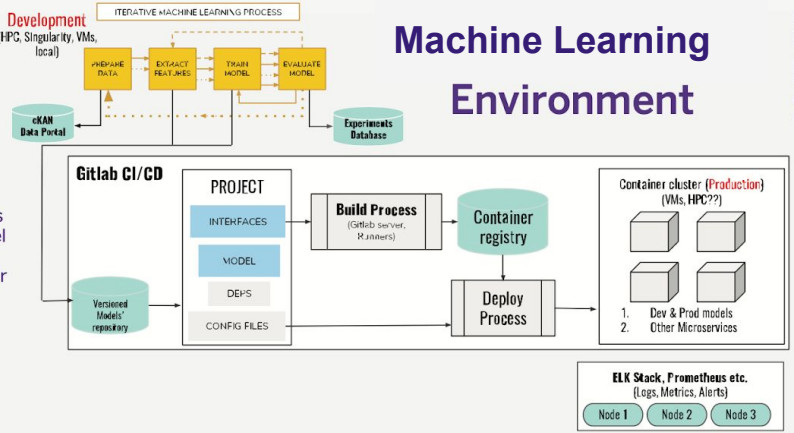


Figure 2: The Vision - A Dev cycle using High Performance Computing (HPC), Singularity, Virtual Machines (VMs) and local resources to (1) build datasets and store them (cKAN data portal), (2) build machine learning models, and (3) store the experiments (experiments database). Once a model is built, the Gitlab CI/ CD allows versioning models, containerizing their deployment and finally pushing them into a Prod Container cluster for scalable deployment through microservices. Finally logs and monitoring are captured using ELK stack and Prometheus.

Practical Methodology

Practical Methodology

“Don’t Be a Hero - Best practices and literature review”

• Successfully applying deep learning requires more than just a good knowledge of what algorithms exist and the principles that explain how they work

• We also need to know how

- to choose an algorithm for a particular application
- to monitor and respond to feedback obtained from experiments in order to improve a machine learning system

• During development of deep learning systems, we need to decide:

- whether to gather more data
- increase or decrease model capacity
- add or remove regularizing features
- debug the software implementation of the model

• Understand what task you are solving what model architecture you should use and best practices

Model Development Process

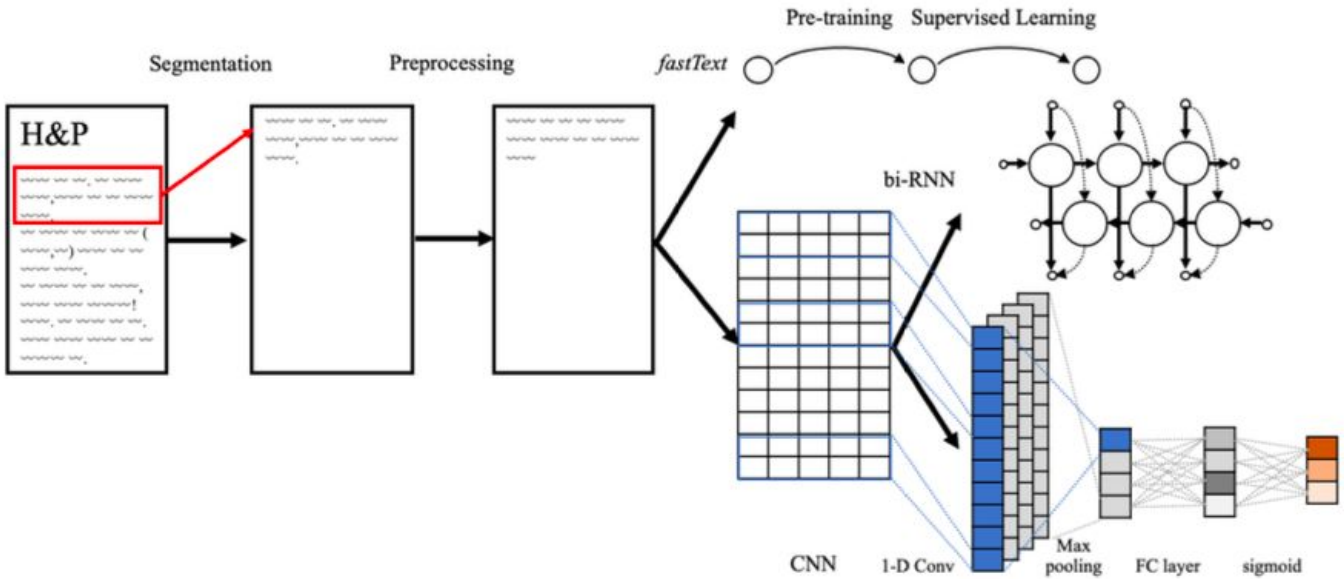


Figure 1: Preprocessing and model development workflow

CNNs in Natural Language Processing

- Character/Byte level, e.g., SMILES CC1=C(C=C(C=C1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]
- To reduce the vocabulary size
- Orthology
- **When you want things to run faster *and significantly less computationally expensive***

Convolutional Neural Networks for Sentence Classification (Kim, 2014)

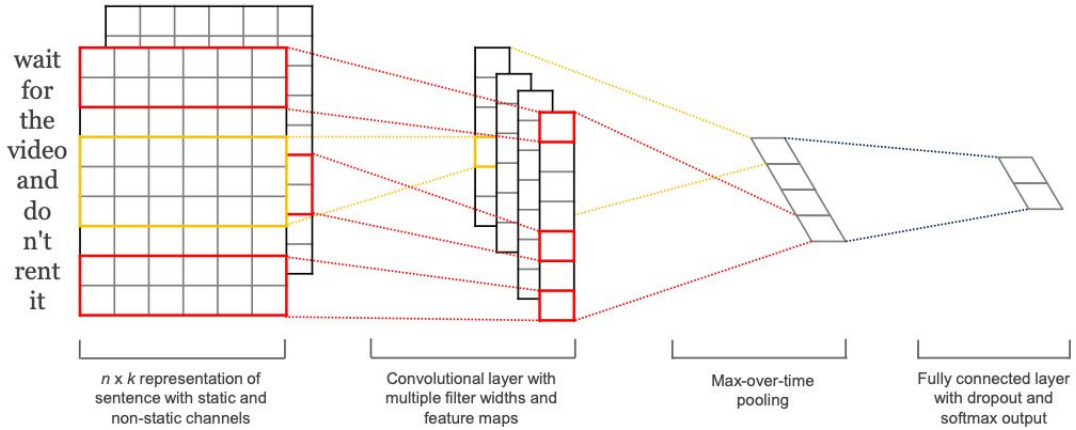
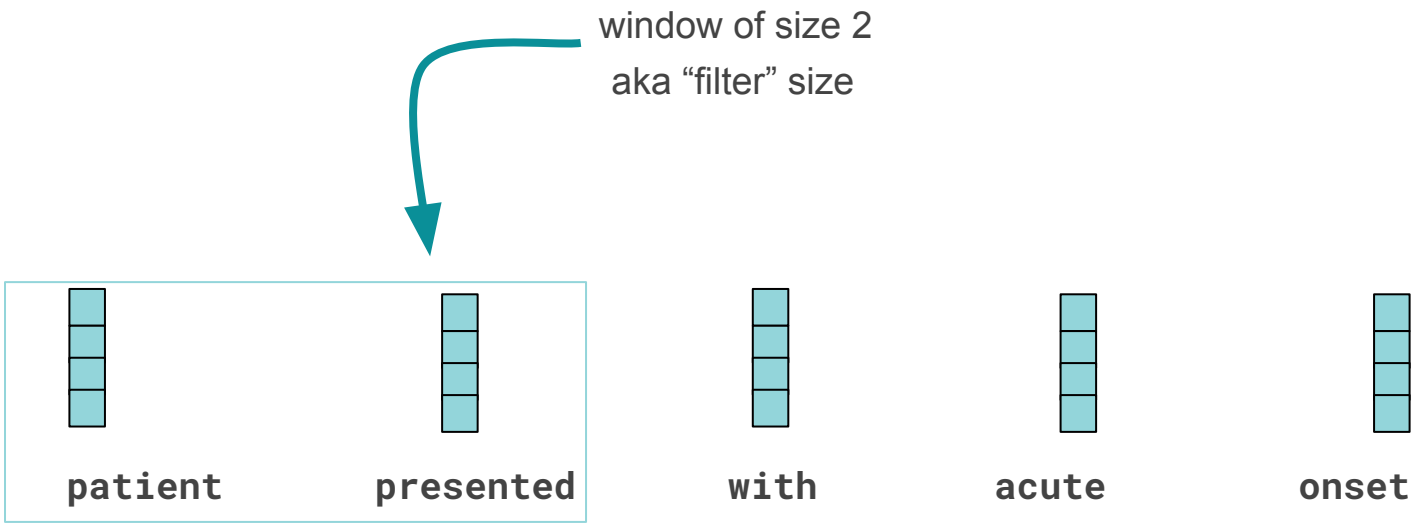


Figure 1: Model architecture with two channels for an example sentence.

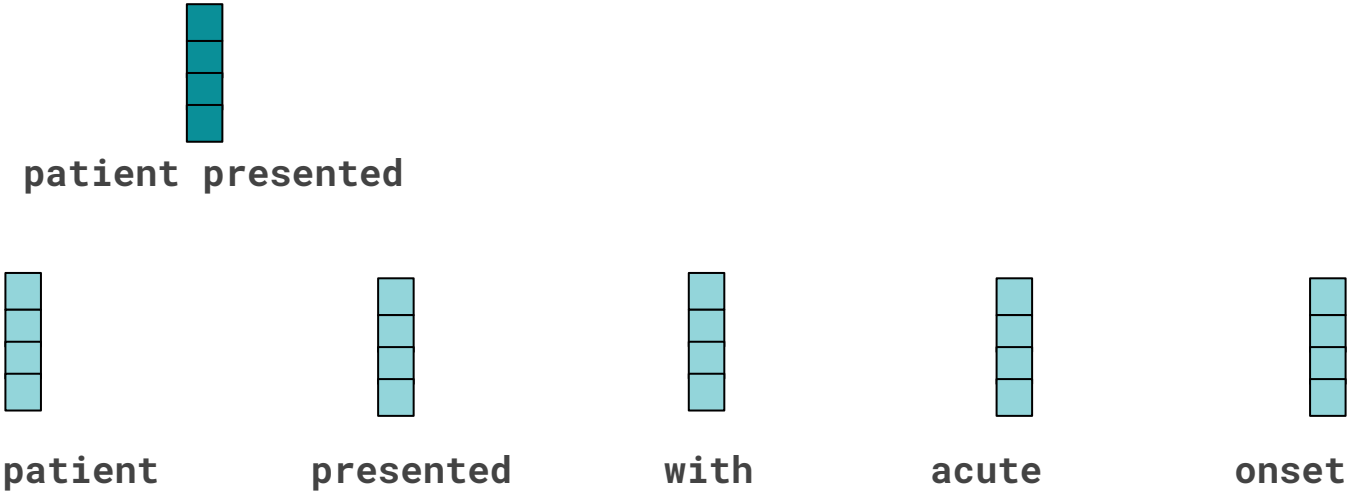
How to convolve on text

“channel” size is 4 in this example

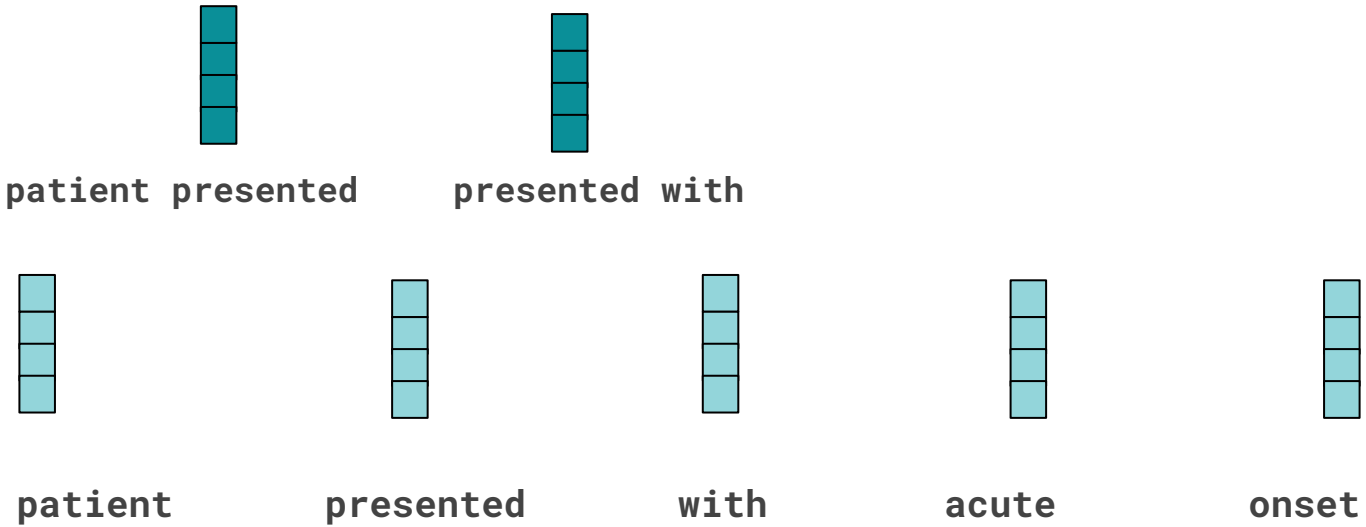
How to convolve on text



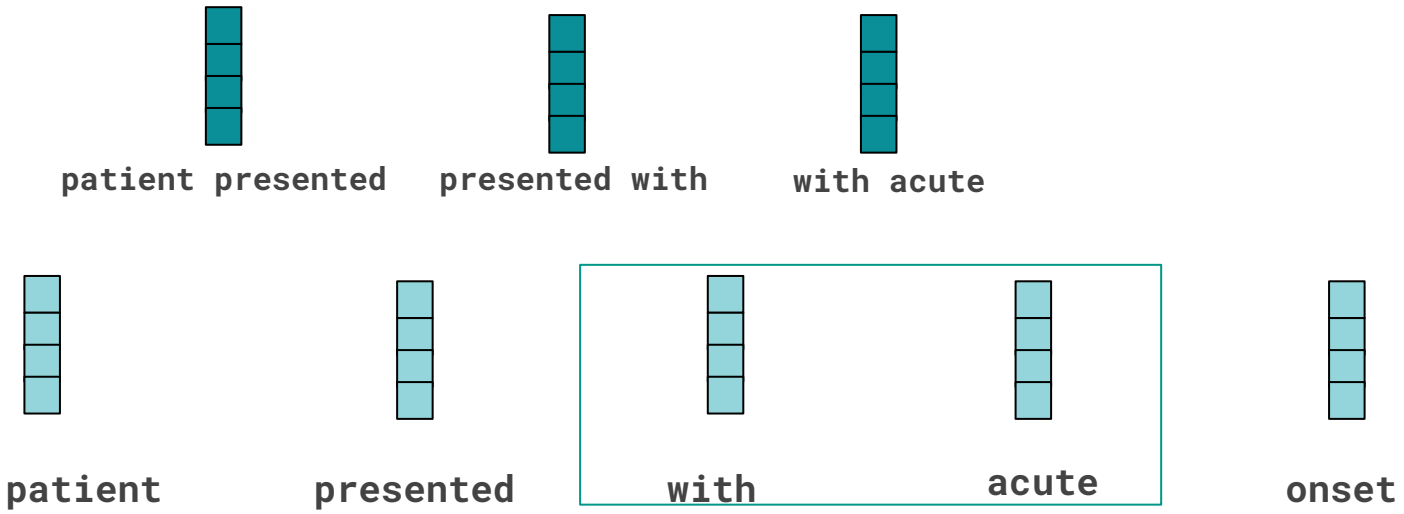
How to convolve on text



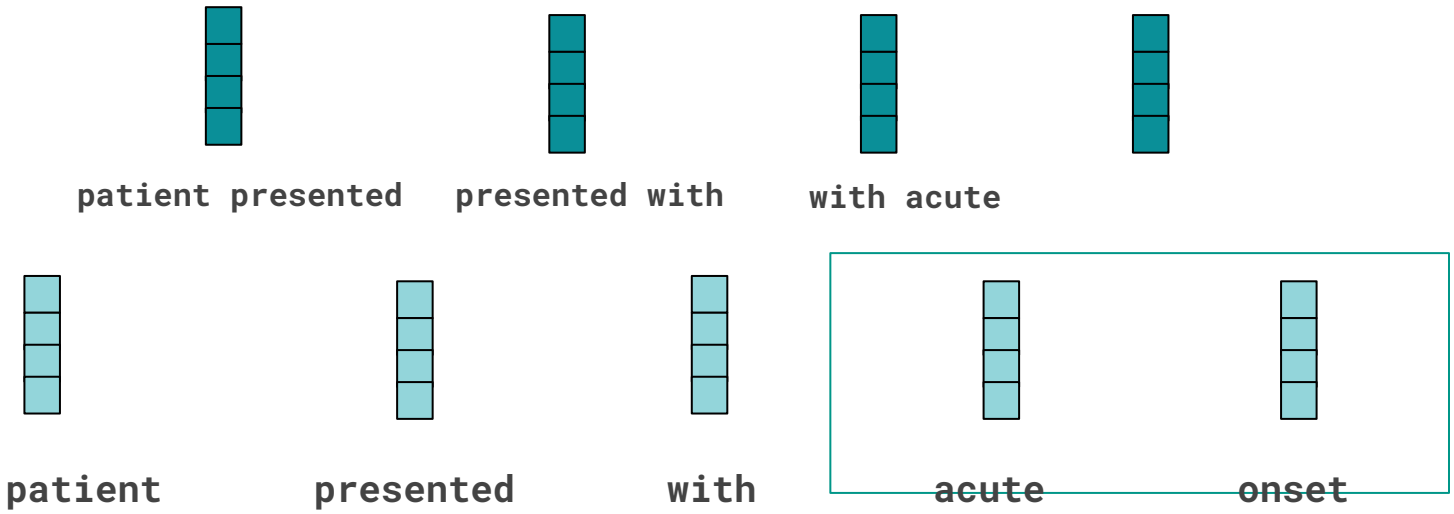
How to convolve on text



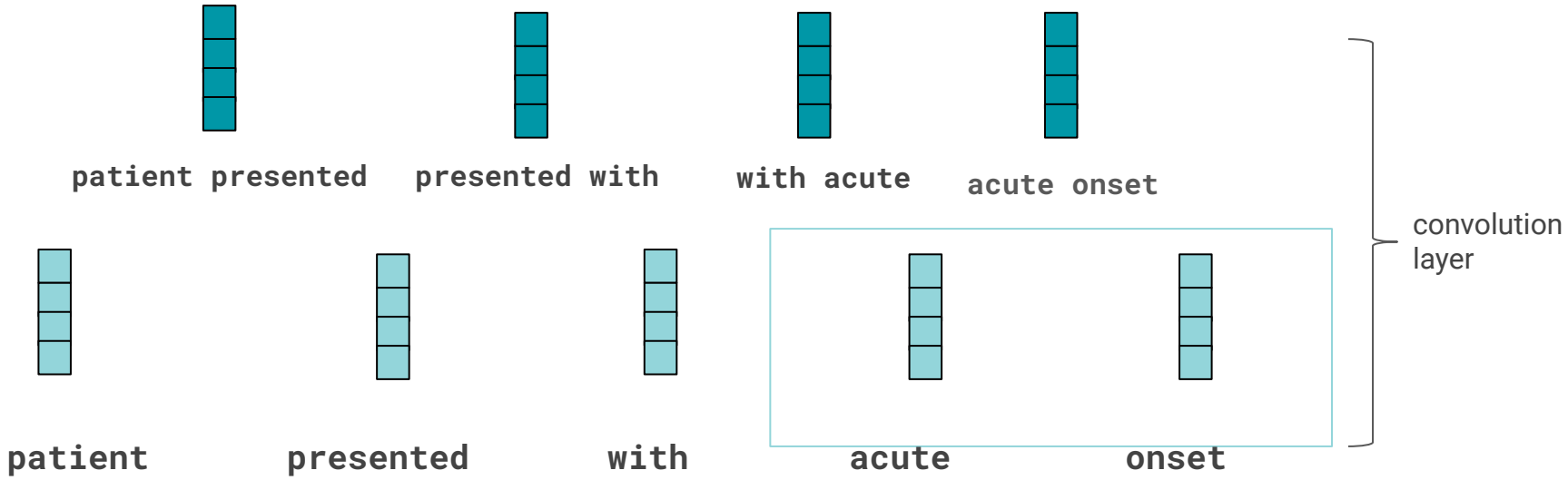
How to convolve on text



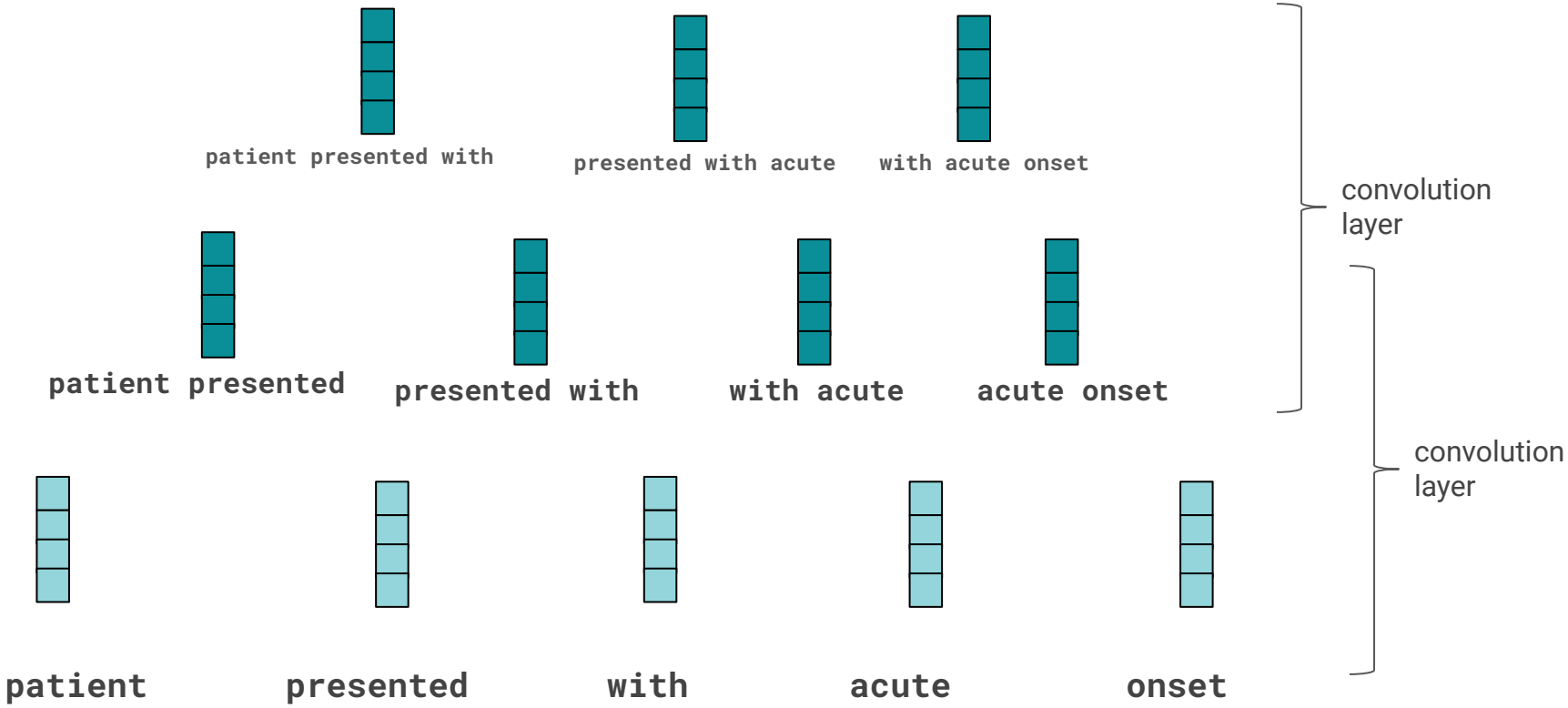
How to convolve on text



How to convolve on text



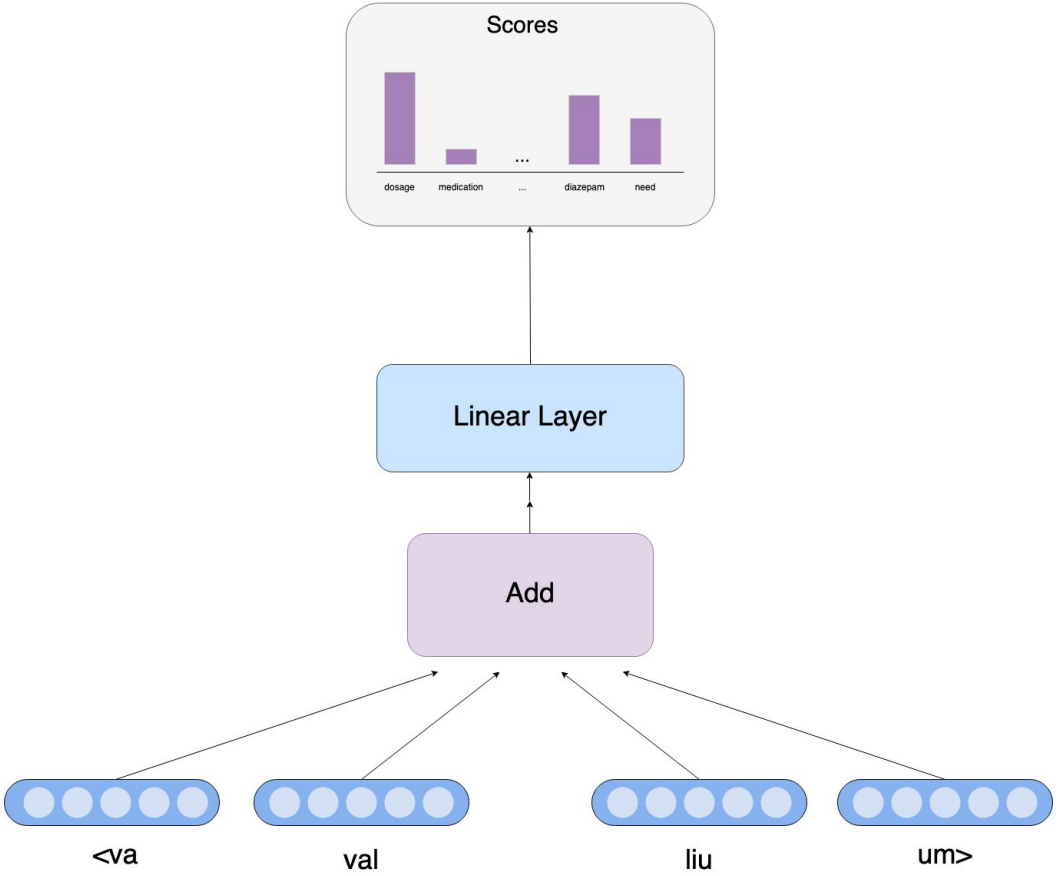
How to convolve on text



Unsupervised Learning

Word Embeddings with fastText

- Word2Vec < Glove < fastText
- 100 and 300 dim skip-gram model with negative sampling and subword enrichment
- training set H&P notes (62k documents and 5M words) + all MIMIC-III Critical Care Database (Johnson et. al. 2016)(2M documents and 5M Words)



Qualitative Evaluation of Word Embeddings

Testing clinical domain-specific acronym against facebook's fastText pretrained embeddings

(300 dim) MIMIC III
+ NYU embeddings

Query word? dnr
dni 0.942678
hcp 0.749637
resuscitate 0.711382
code 0.65287
intubate 0.638306
cmo 0.637459
dnri 0.612248
reintubate 0.606106
hcps 0.603236
hospice 0.58464

fastText's English
Wikipedia 300 dim

Query word? dnr
dnl 0.954677
dnssec 0.940462
cwp 0.936214
dpb 0.934438
hvdc 0.929971
bvu 0.927638
dnq 0.927279
tpb 0.925606
pkc 0.924558
hvd 0.923596

DNR = do not resuscitate
DNI = do not intubate
HCP = health care provide
CMO = comfort measures only

Cohort Selection

- Restricted to notes between 0-16 hours upon admission
- Notes with more than 50 words (after removing addendums and attestations)
- Restricted to most common author types: Physician, Fellow, Resident, Physician Assistant, and Nurse Practitioner

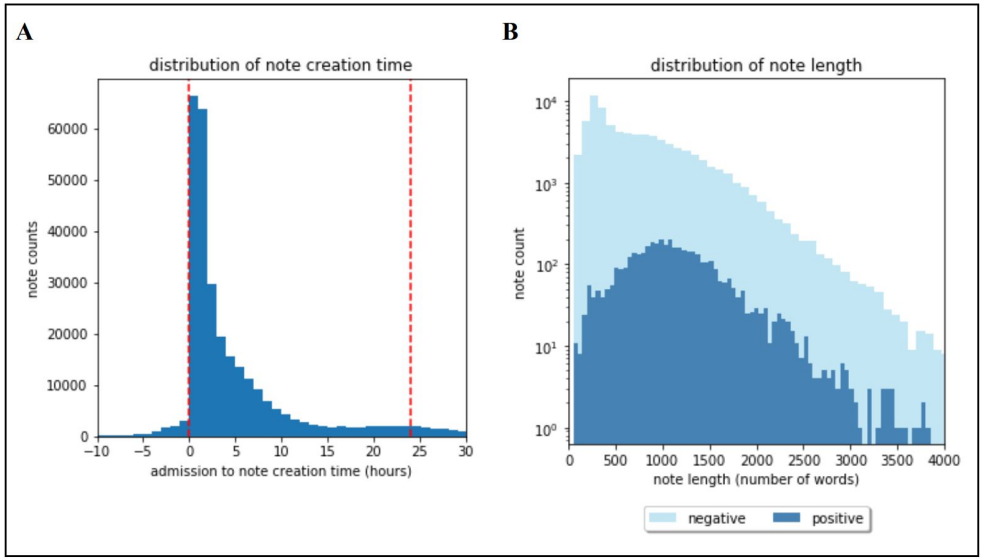


Figure 1. *A) Time taken between admission and note creation, B) distribution of note length*

Cohort Selection

- Hospital admissions from January 1st, 2013 to December 31st, 2017 (5 years)
- Death outcomes from social security data and institutional data (positive labels)
- Separation performed temporarily and at patient level to prevent data leakage between sets (Neto et. al. 2019)

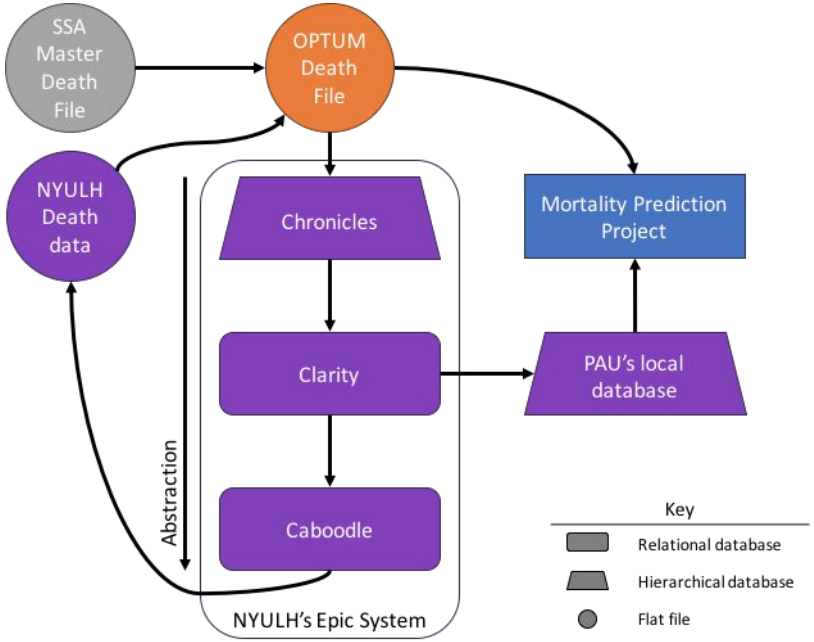


Table 1: Datasets of H&P notes and 60-day patient outcomes used for model development.

	Training	Validation	Test	Total	
Positive	2,679	468	1,055	4,202	(5.1%)
Negative	50,140	8,973	19,473	78,586	
Total	52,819	9,441	20,528	82,788	

Results

Evaluation Metrics

- End-of-life is a rare outcome which can skew evaluation metrics such as accuracy
- Visualized with receiver operating characteristic (ROC) and measured by the AUROC
- As this model is potentially helpful to recommend an intervention to predicted positives, precision-recall curves (PRC) and AUPRC and the max-F1 score are employed

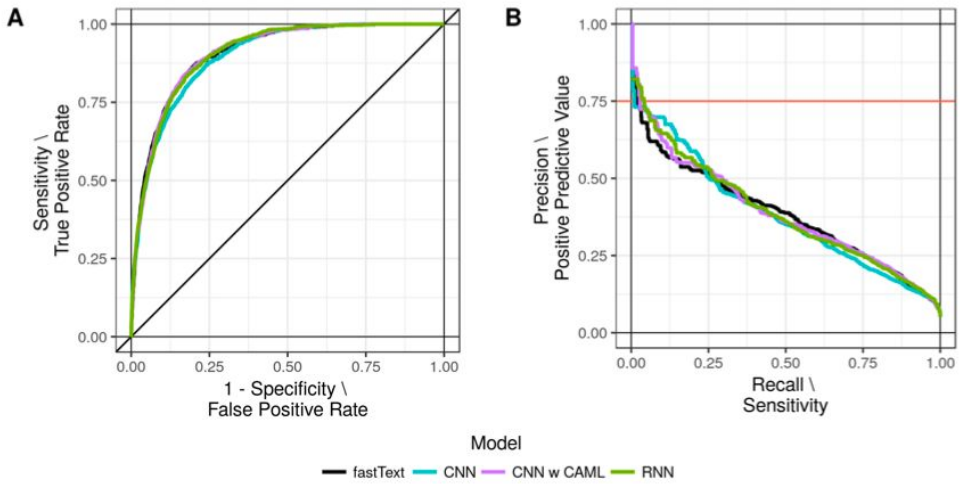


Figure 2: Test set evaluation metrics. A) ROC and B) PRC.

Table 2: Test set evaluation metrics.

Model	**AUROC** [95% CI]	**AUPRC** [95% CI]	**max F-1** [95% CI]
CNN	0.899 [0.890, 0.908]	0.381 [0.348, 0.421]	0.418 [0.394, 0.450]
RNN (bi-GRU)	0.907 [0.899, 0.915]	0.388 [0.357, 0.427]	0.421 [0.399, 0.452]
CNN with CAML	0.908 [0.900, 0.917]	0.388 [0.354, 0.424]	0.425 [0.401, 0.454]

CNNs trained & predicted 75% faster than RNN with comparable results, thus this model was selected for prospective validation

Performance on New Site (New Paper)

Evaluation of the model on an entirely different hospital

Table 3: Prospective validation results.

Metric	H&P CNN	Structured Data
Total admissions	65,727	
Predictions	57,997	53,446
Admissions predicted	37,720 (57.4%)	53,446 (81.3%)
Timing (hrs) median [IQR]	29.6 [19.0, 36.8]	0.03 [0.02, 0.85]
High-risk admissions	80 (0.21%)	131 (0.25%)
AUROC [95% CI]	0.860 [0.847, 0.873]	0.806 [0.793, 0.820]
AUPRC [95% CI]	0.314 [0.282, 0.353]	0.179 [0.157, 0.204]
Max F-1 [95% CI]	0.377 [0.352, 0.409]	0.228 [0.210, 0.254]

- Identified 26 high-risk patients using H&P approach vs 1 from the structured data model
- The text-based approach outperforms an structured data system and generalizes better to a new hospital location

Comparison of the models that performed on a new unseen 4th hospital (8 months)

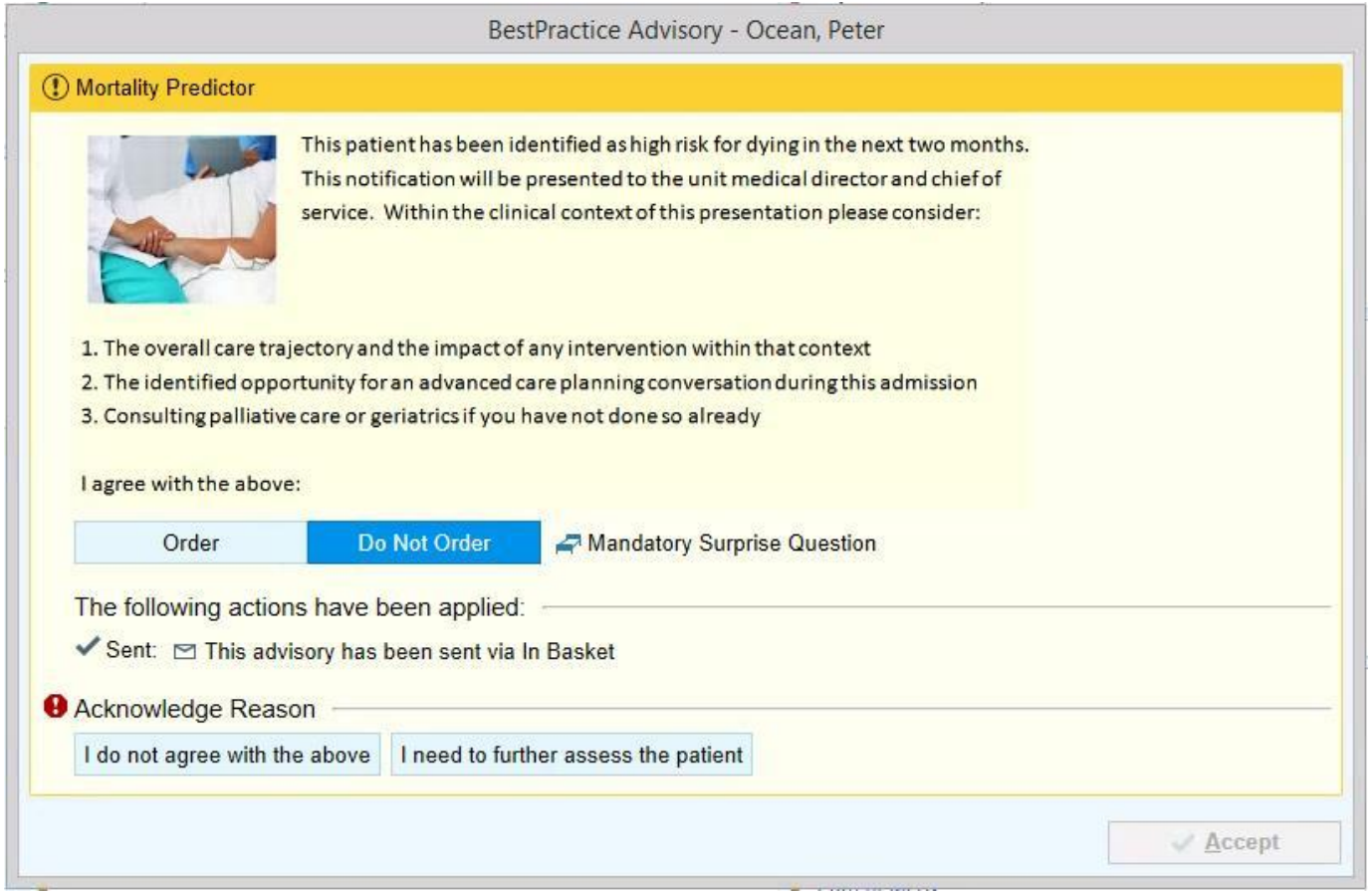
Why this is relevant to Covera Health?

- Although notes can be written differently the model still generalizes quite well
- Note the same for the model built on structured data
    - Not necessarily the same for hospital - billing behavior are often dictated by the contract with payers

Clinical Significance and Feedback

Feedback from Clinicians

“One case (I can provide the details if that would be helpful), at the beginning of the admission to me would not have triggered that pt was an end of life patient, but as the admission unfolded this became clear as pt became more acutely ill. Since you are running this tool retrospectively I am curious if this patient would have flagged at the start of pt admission.”



- Presented to the attending of record
- Interruptive (cannot away)
  - ***One chance t further asses***

Prompts a MSQ (mandatory surprise question) order and suggests an ACP conversation and consulting Palliative Care/Geriatrics

References

Avati A, Jung K, Harman S, et al. (2017) Improving Palliative Care with Deep Learning. arXiv [cs.CY]. Available from: http://arxiv.org/abs/1711.06402.

Choi E, Bahadori MT, Schuetz A, et al. (2016) Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR workshop and conference proceedings 56: 301–318.

Detsky ME, Harhay MO, Bayard DF, et al. (2017) Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an I Admission. JAMA: the journal of the American Medical Association 317(21): 2187–2195.

Detsky ME, Harhay MO, Bayard DF, et al. (2017) Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an I Admission. JAMA: the journal of the American Medical Association 317(21): 2187–2195.

Ghassemi MM, Richter SE, Eche IM, et al. (2014) A data-driven approach to optimized medication dosing: a focus on heparin. Intensive care medicine 40(9): 1332–133

Huang J, Osorio C, Sy LW. (2018) An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes. arXiv [cs.CL]. Available from: https://arxiv.org/abs/1802.02311.

Johnson AEW, Pollard TJ, Shen L, et al. (2016) MIMIC-III, a freely accessible critical care database. Scientific data 3: 160035.

Li J, Chen X, Hovy E, et al. (2016) Visualizing and Understanding Neural Models in NLP. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 681–691.

Longo D, Fauci A, Kasper D, et al. (2011) Harrison's Principles of Internal Medicine, 18th Edition. McGraw Hill Professional.
Konkle B. Chapter 115. Disorders of Platelets and Vessel Wall. In: Longo DL, Kasper DL, Jameson JL, Fauci AS, Hauser SL, eds. Harrison's Principles of Internal Medicine. 18th ed. New York: McGraw-Hill; 2012.http://www.accesspharmacy.com/content.aspx?aID=9100733. Accessed January 30, 2013.

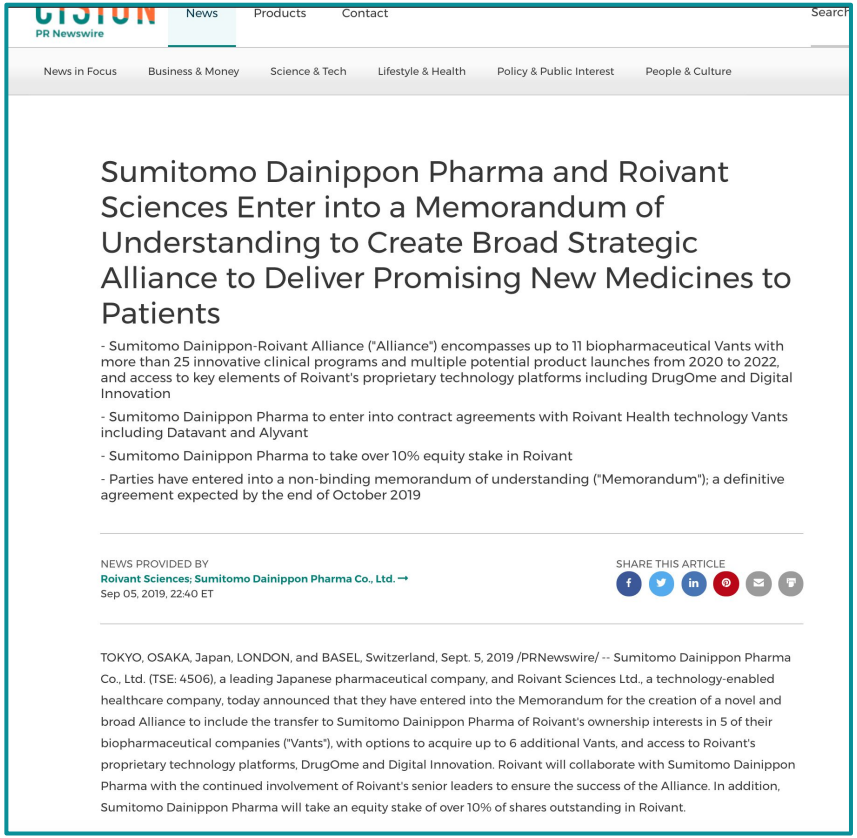
Mullenbach J, Wiegreffe S, Duke J, et al. (2018) Explainable Prediction of Medical Codes from Clinical Text. arXiv [cs.CL]. Available from: https://arxiv.org/abs/1802.056

Purushotham S, Meng C, Che Z, et al. (2017) Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets. arXiv [cs.LG]. Available from: http://arxiv.org/abs/1710.08531.

Smythe MA, Priziola J, Dobesh PP, et al. (2016) Guidance for the practical management of the heparin anticoagulants in the treatment of venous thromboembolism. Journal of thrombosis and thrombolysis 41(1): 165–186.

AI Scientist - Journey into Vant Alliance

Roivant Sciences



PR Newswire

News | Products | Contact | Search

News in Focus | Business & Money | Science & Tech | Lifestyle & Health | Policy & Public Interest | People & Culture

Sumitomo Dainippon Pharma and Roivant Sciences Enter into a Memorandum of Understanding to Create Broad Strategic Alliance to Deliver Promising New Medicines to Patients

- Sumitomo Dainippon-Roivant Alliance ("Alliance") encompasses up to 11 biopharmaceutical Vants with more than 25 innovative clinical programs and multiple potential product launches from 2020 to 2022, and access to key elements of Roivant's proprietary technology platforms including DrugOme and Digital Innovation
- Sumitomo Dainippon Pharma to enter into contract agreements with Roivant Health technology Vants including Datavant and Alyvant
- Sumitomo Dainippon Pharma to take over 10% equity stake in Roivant
- Parties have entered into a non-binding memorandum of understanding ("Memorandum"); a definitive agreement expected by the end of October 2019

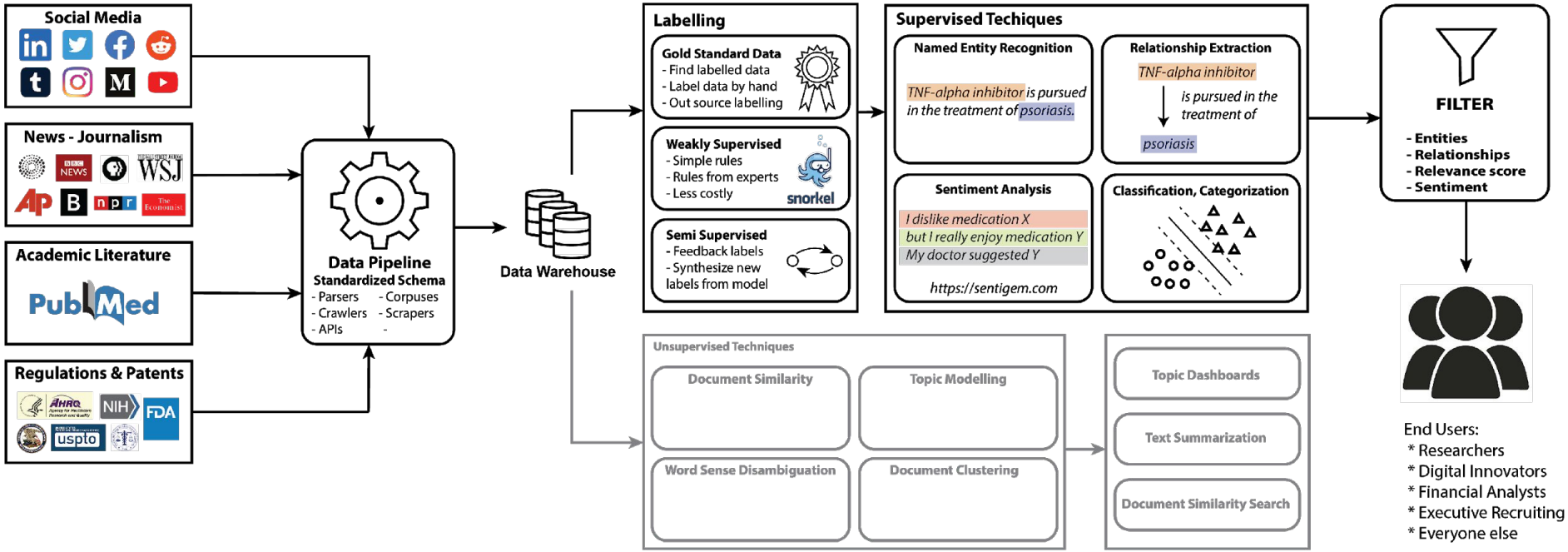
NEWS PROVIDED BY
Roivant Sciences; Sumitomo Dainippon Pharma Co., Ltd.
Sep 05, 2019, 22:40 ET

SHARE THIS ARTICLE

TOKYO, OSAKA, Japan, LONDON, and BASEL, Switzerland, Sept. 5, 2019 /PRNewswire/ -- Sumitomo Dainippon Pharma Co., Ltd. (TSE: 4506), a leading Japanese pharmaceutical company, and Roivant Sciences Ltd., a technology-enabled healthcare company, today announced that they have entered into the Memorandum for the creation of a novel and broad Alliance to include the transfer to Sumitomo Dainippon Pharma of Roivant's ownership interests in 5 of their biopharmaceutical companies ("Vants"), with options to acquire up to 6 additional Vants, and access to Roivant's proprietary technology platforms, DrugOme and Digital Innovation. Roivant will collaborate with Sumitomo Dainippon Pharma with the continued involvement of Roivant's senior leaders to ensure the success of the Alliance. In addition, Sumitomo Dainippon Pharma will take an equity stake of over 10% of shares outstanding in Roivant.

Sumitovant

"The Alliance"


Social Media
News - Journalism
Academic Literature
PubMed
Regulations & Patents
Data Pipeline
Standardized Schema
- Parsers
- Corpuses
- Crawlers
- Scrapers
- APIs
-
Data Warehouse
Labelling
Gold Standard Data
- Find labelled data
- Label data by hand
- Out source labelling
Weakly Supervised
- Simple rules
- Rules from experts
- Less costly
snorkel
Semi Supervised
- Feedback labels
- Synthesize new labels from model
Supervised Techiques
Named Entity Recognition
TNF-alpha inhibitor is pursued in the treatment of psoriasis.
Relationship Extraction
TNF-alpha inhibitor
is pursued in the treatment of
psoriasis
Sentiment Analysis
I dislike medication X
but I really enjoy medication Y
My doctor suggested Y
https://sentigem.com
Classification, Categorization
Unsupervised Techniques
Document Similarity
Topic Modelling
Word Sense Disambiguation
Document Clustering
Topic Dashboards
Text Summarization
Document Similarity Search
FILTER
- Entities
- Relationships
- Relevance score
- Sentiment
End Users:
* Researchers
* Digital Innovators
* Financial Analysts
* Executive Recruiting
* Everyone else

Generating Toxic Molecules

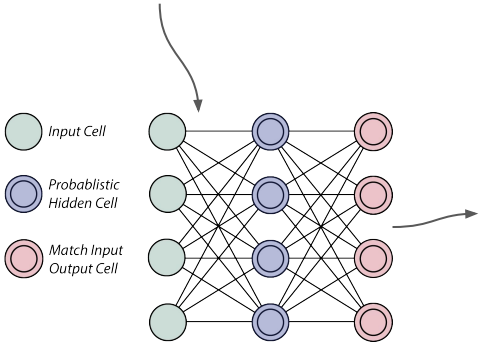
- Authors: Izzy Metzger *Computational Research*, Zach Carpenter *Roivant Health*
  - paper we follow is: https://arxiv.org/abs/1610.02415
  - other papers we utilize: https://link.springer.com/article/10.1007%2FBF00332918
    - https://github.com/HIPS/molecule-autoencoder
      - Note: deepchem has a vae model that follows the Aspuru-Guzuki Framework similarily to ours but we didn't implement deepchem in this (we could in the future if everyone prefers that framework)
    - we trained the model using the latest chembl database (chembl 25)
    - This notebook shows the output of thatmodel
    - we will use the same latent dimension as in the paper (292)
    - OVERALL DESIGN:
      - From each toxic smile in the dataset sample the latent space next to that toxic smile to get auto-generated toxish molecules
      - In particular, we create 1000 new toxish smiles using this method with our vae model (that we trained over the weekend on the latest chembl db)
      - Each of those generated smiles are then checked to see if they are "working" via rdkit/ and get rid of the broken smiles (for e.g., 1000 generated toxish smiles in this example (which uses the first toxic (label==1) smile in the latest master table)
      - We plot valid smiles (note some of these are real and already exist and some of them are not real)
      - we can also interpolate two smiles (like in the paper-- note our results are not as good as teh paper but we didn't train the model as long)
    - NEXT STEPS:
    - Determine measures of significance
      - Defining a toxic metric for our new smiles
        - Some we can validate, others we can look at tanimoto similarity
          - https://stackoverflow.com/questions/51681659/how-to-use-rdkit-to-calculte-molecular-fingerprint-and-similarity-of-a-list-of-s
          - Tanimoto method is then used to produce a dissimilarity matrix related to the Jaccard dissimilarity
          - Tanimoto dissimilarity is a modified Hamming dissimilarity
        - Perhaps using LIME in some way?
      - The paper postedd describes how we could optimize a molecule produced by our model for a particular molecular property (e.g., skin permeability or even solubility .. possibility toxicity?)

Zach Carpenter now the CEO of Vant.AI

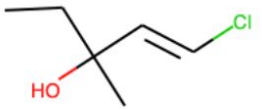
name	canonical	isomeric	toxicity	one_hot	designation	Mol_ID	ABC	...	SRW
S=C=Nc1c2c(ccc1)cccc2	C1=CC=C2C(=C1)C=CC=C2N=C=S	C1=CC=C2C(=C1)C=CC=C2N=C=S	Hepatotoxicity	1	experimental/Liu smiles	Mol0	9.818615	...	9.3825
c1(c(cc(cc1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[...	CC1=C(C=C(C=C1[N+](=O)[O-])[N+](=O)[O-])[N+](=...	CC1=C(C=C(C=C1[N+](=O)[O-])[N+](=O)[O-])[N+](=...	Hepatotoxicity	1	experimental/Liu smiles	Mol1	11.877237	...	9.6377
c1(c(cc(cc1)[N+](=O)[O-])[N+](=O)[O-])O	C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])O	C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])O	Hepatotoxicity	1	experimental/Liu smiles	Mol2	9.618017	...	9.3008
O(CCO)CC	CCOCCO	CCOCCO	Hepatotoxicity	1	experimental/Liu smiles	Mol3	3.535534	...	6.6080
Oc1cc2c(cc1)cccc2	C1=CC=C2C=C(C=CC2=C1)O	C1=CC=C2C=C(C=CC2=C1)O	Hepatotoxicity	1	experimental/Liu smiles	Mol4	8.554231	...	9.2257

Variational AutoEncoders

CCC(O)(\\C=C\\Cl)C



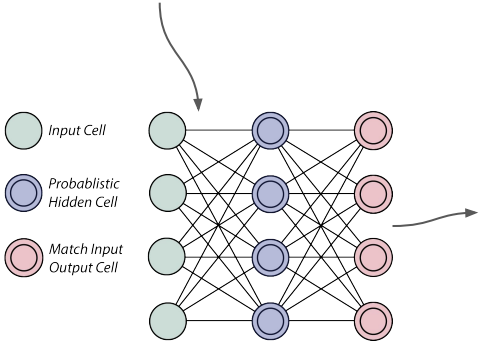

```
array([[ 0.05363178,  0.0078782 , -0.07700312,  0.05649856,  0.13152978,
         0.07801478, -0.07673378, -0.04488152,  0.04152625,  0.05349444,
        -0.04960863,  0.00735946, -0.03721184,  0.06771883, -0.00867662,
        -0.11634302, -0.07169832, -0.12267046,  0.06091165,  0.06741397,
         0.05132991, -0.11788134, -0.04044838, -0.06816725,  0.02068123,
         0.05099233,  0.18753207, -0.13674015, -0.12235098,  0.0800745 ,
         0.05587533,  0.10807797, -0.03181177, -0.1161322 , -0.138663  ,
        -0.00987642,  0.00706142,  0.19610442,  0.0629833 , -0.05595475,
         0.01440607,  0.01905868, -0.07988203,  0.05325389,  0.04673581,
        -0.13123605,  0.04979331,  0.10067537,  0.11553518,  0.03937402,
         0.01276392,  0.17929122,  0.08065402,  0.0651684 , -0.11405068,
         0.12338633, -0.09664553, -0.04511814, -0.05998807,  0.10256029,
         0.10905663, -0.06486235, -0.12057194, -0.08205813, -0.0080945 ,
        -0.00697902, -0.14520209,  0.00825424,  0.01416555, -0.10578428,
        -0.01481704, -0.11556359, -0.03562332, -0.04776421, -0.03348266,
         0.13707504,  0.16890147,  0.14823346,  0.12322275, -0.03067702,
         0.04986928, -0.00574416, -0.02069437, -0.01018216,  0.0542658 ,
         0.02804135, -0.05027196,  0.08862189,  0.0130431 , -0.03631146,
         0.01536888,  0.05524712,  0.08686657,  0.04180029,  0.0685646 ,
        -0.13962124,  0.00456896,  0.0190599 ,  0.20133126,  0.12001671,
         0.01564296, -0.02090459,  0.07305951,  0.00762762,  0.08478145,
        -0.01544177, -0.03805823,  0.05480809,  0.02945027,  0.04178546,
        -0.0258328 , -0.06435941,  0.04975348,  0.00826132, -0.0732389 ,
         0.12626693,  0.03737481, -0.00254592,  0.03650747,  0.01470203,
         0.01745417,  0.01659185, -0.01863235, -0.09864108, -0.02229385,
         0.02127865, -0.08553039, -0.01834461,  0.02821246,  0.17647676,
         0.05512521,  0.00616604, -0.04119888,  0.02849228, -0.04119967,
        -0.01906603, -0.02676078, -0.01619593, -0.04449657, -0.06778533,
         0.01248086,  0.13863096,  0.09380561,  0.01931311,  0.01201088,
         0.07519506, -0.04049241,  0.00817274, -0.12936787,  0.10033616,
```



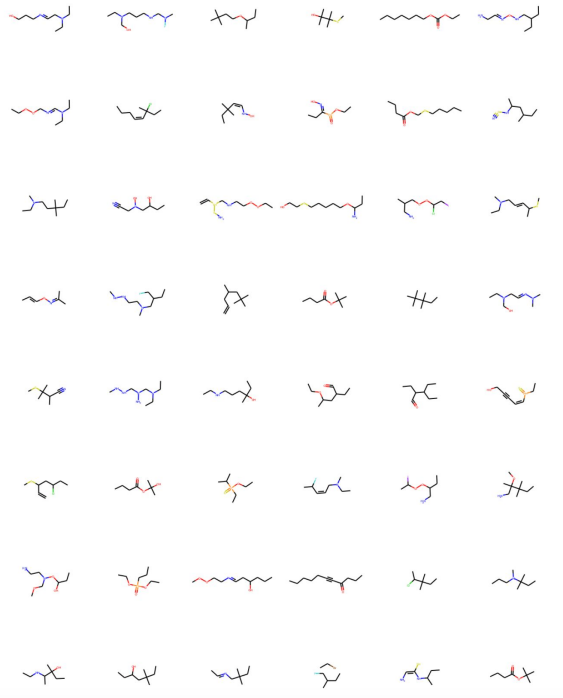
- Converting SMILE to latent representation to molecule and Vice Versa

Variational AutoEncoders

CCC(O)(\\C=C\\Cl)C




```
array([[ 0.05363178,  0.0078782 , -0.07700312,  0.05649856,  0.13152978,
         0.07801478, -0.07673378, -0.04488152,  0.04152625,  0.05349444,
        -0.04960863,  0.00735946, -0.03721184,  0.06771883, -0.00867662,
        -0.11634302, -0.07169832, -0.12267046,  0.06091165,  0.06741397,
         0.05132991, -0.11788134, -0.04044838, -0.06816725,  0.02068123,
         0.05099233,  0.18753207, -0.13674015, -0.12235098,  0.0800745 ,
         0.05587533,  0.10807797, -0.03181177, -0.1161322 , -0.138663  ,
        -0.00987642,  0.00706142,  0.19610442,  0.0629833 , -0.05595475,
         0.01440607,  0.01905868, -0.07988203,  0.05325389,  0.04673581,
        -0.13123605,  0.04979331,  0.10067537,  0.11553518,  0.03937402,
         0.01276392,  0.17929122,  0.08065402,  0.0651684 , -0.11405068,
         0.12338633, -0.09664553, -0.04511814, -0.05998807,  0.10256029,
         0.10905663, -0.06486235, -0.12057194, -0.08205813, -0.0080945 ,
        -0.00697902, -0.14520209,  0.00825424,  0.01416555, -0.10578428,
        -0.01481704, -0.11556359, -0.03562332, -0.04776421, -0.03348266,
         0.13707504,  0.16890147,  0.14823346,  0.12322275, -0.03067702,
         0.04986928, -0.00574416, -0.02069437, -0.01018216,  0.0542658 ,
         0.02804135, -0.05027196,  0.08862189,  0.0130431 , -0.03631146,
         0.01536888,  0.05524712,  0.08686657,  0.04180029,  0.0685646 ,
        -0.13962124,  0.00456896,  0.0190599 ,  0.20133126,  0.12001671,
         0.01564296, -0.02090459,  0.07305951,  0.00762762,  0.08478145,
        -0.01544177, -0.03805823,  0.05480809,  0.02945027,  0.04178546,
        -0.0258328 , -0.06435941,  0.04975348,  0.00826132, -0.0732389 ,
         0.12626693,  0.03737481, -0.00254592,  0.03650747,  0.01470203,
         0.01745417,  0.01659185, -0.01863235, -0.09864108, -0.02229385,
         0.02127865, -0.08553039, -0.01834461,  0.02821246,  0.17647676,
         0.05512521,  0.00616604, -0.04119888,  0.02849228, -0.04119967,
        -0.01906603, -0.02676078, -0.01619593, -0.04449657, -0.06778533,
         0.01248086,  0.13863096,  0.09380561,  0.01931311,  0.01201088,
         0.07519506, -0.04049241,  0.00817274, -0.12936787,  0.10033616,
```



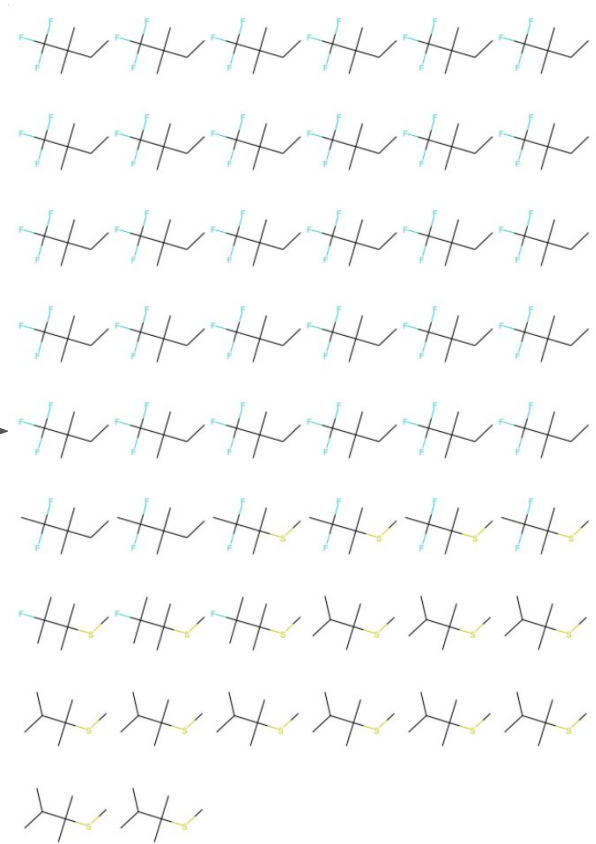
Looking for 1k toxish neighbours

- Generating 1k random 292 dimension continuous arrays with mean=tox_latent and stdev==0.1 to sample the latent space next to the tox example.

VAE

C1=CC=C2C=C(C=CC2=C1)O

S=C=Nc1c2c(ccc1)cccc2



- **Interpolation of two toxic molecules**
- **(combinations can be endless!)**

30+ models I developed during my two years in Computational Research Team (DrugOME) in areas (1 Natural Language Processing & 2 Drug Development)

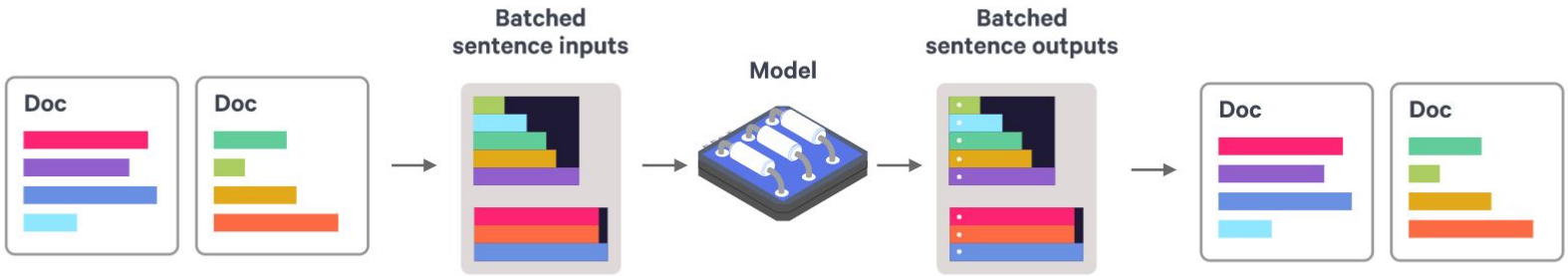
Language Modelling & Transformer-Based Projects

- Reading Comprehension (Question and Answering)
- Textual Entailment within Pharma Patents, Regulatory Documents and More
- PICO (Patient, Intervention, Cohort, and Outcome) NER and Relationship Extraction
- Medication, Dosage, etc NER & Clinical Relationship Extraction

Transformers Tokenization

```
# bert-base-uncased

['[CLS]', 'laced', 'with', 'dreams', '-', 'dripping', 'in', 'reality', ',', 'the', 'american',
'dream', 'reign', '##ites', 'after', '9', '.', '11', 'with', 'a', 'true', 'story', 'about',
'the', 'devil', 'ray', "'", 's', 'mid', '-', 'life', 'rookie', ',', 'jimmy', 'morris', '.',
'[SEP]']

# gpt2

['<|endoftext|>', 'L', 'aced', 'Ġwith', 'Ġdreams', 'Ġ-', 'Ġdripping', 'Ġin', 'Ġreality', ',',
'Ġthe', 'ĠAmerican', 'ĠDream', 'Ġreign', 'ites', 'Ġafter', 'Ġ9', '.', '11', 'Ġwith', 'Ġa',
'Ġtrue', 'Ġstory', 'Ġabout', 'Ġthe', 'ĠDevil', 'ĠRay', "'s", 'Ġmid', '-', 'life', 'Ġrookie', ',',
'ĠJimmy', 'ĠMorris', '.', '<|endoftext|>']

# xlnet-base-cased

['<cls>', '_Lac', 'ed', '_with', '_dreams', '_', '-', '_dripping', '_in', '_reality', ',', '_the',
'_American', '_Dream', '_reign', 'ites', '_after', '_9', '.', '11', '_with', '_a', '_true',
'_story', '_about', '_the', '_Devil', '_Ray', "'", 's', '_mid', '-', 'life', '_rookie', ',',
'_Jimmy', '_Morris', '.', '</s>']
```

Sorting & Batch Training



Q&A

Thank you Covera Health!

Natural Language Interface for Electronic Health Records

Overview - NLI for EHR

1

Introduction

(1) Motivation of Work
(2) Data Sources

2

Methods

(1) Corpus Creation
(2) Question-SQL Pair Creation
(3) Named Entity Recognition
(4) Semantic Parsing

3

Results

(1) NER on i2b2 Dataset
(2) NER on Question-SQL Corpus
(3) Seq2Seq on Question-SQL Corpus

4

Conclusion

(1) Final Takeaways

Introduction

Motivation of Work:
Why a Natural Language Interface for EHR?

- Build a prototype interface on the Medical Information Mart for Intensive Care III (MIMIC III) database, a publicly-available database of EHRs of critical care patients.
- “Natural Language Question” to SQL

Data Sources

- **MIMIC-III Critical Care Database**
  - ~60,000 ICU patients
  - 28 tables: includes admission dates and times, laboratory tests, medications, transfers, and more

- **2012 i2b2 Clinical NLP benchmark dataset**
  - Entity tags include:
    1. problems (“upper quadrant pain”, “diabetes”)
    2. treatments (“blood transfusion”, “aspirin”)
    3. tests (“EKG”, “INR”, “cardiac enzymes”)
    4. clinical departments (“surgery”, “ICU”)
    5. evidentials (“presented” in “Patient presented with…”)
    6. occurrences (events that happen to the patient: “admission”, “transfer”, “follow-up”)

Methods

Question-SQL Pair Creation

Question Types

1. Demographics, which include data such as insurance, race, religion, gender, age
2. Pharmacological Treatments
3. Procedures
4. Comorbidities (*other conditions/diseases patient may have*)
5. Patient Medical History (for example, surgical history)

Question-SQL Pair Dataset Creation

- Crowd-sourced questions
- Constructed “templates”
- Wrote SQL query pairs
- Utilized paraphrasing techniques to extend dataset

How many patients with Diabetes were given Insulin

```
NER:
How many patients with PROBLEM@1
were given TREATMENT@1
{PROBLEM@1 : Diabetes,
TREATMENT@1 : Insulin}
```

```
Seq2Seq:
SELECT count(DISTINCT hadm_id)
FROM DIAGNOSES_ICD
WHERE icd9_code IN
(SELECT DISTINCT icd9_code
FROM D_ICD_DIAGNOSES
WHERE long_title LIKE '%PROBLEM@1%')
AND hadm_id IN
(SELECT DISTINCT hadm_id
FROM PRESCRIPTIONS
WHERE drug LIKE '%TREATMENT@1%')
```

Question-SQL Pair Creation

How many patients suffering from PROBLEM_1 were administered TREATMENT_1

Paraphrases using PPDB (**Paraphrase Database**) (Ganitkevitch et al)
How many people suffering from PROBLEM_1 were administered TREATMENT_1
How many patients suffering from PROBLEM_1 were treated TREATMENT_1
Count of patients suffering from PROBLEM_1 were administered TREATMENT_1

```
select count(distinct hadm_id) from DIAGNOSES_ICD where icd9_code IN (select distinct icd9_code from D_ICD_DIAGNOSES where long_title like '%PROBLEM_1%') and hadm_id IN (select distinct hadm_id from PRESCRIPTIONS where drug like '%TREATMENT_1%')
```

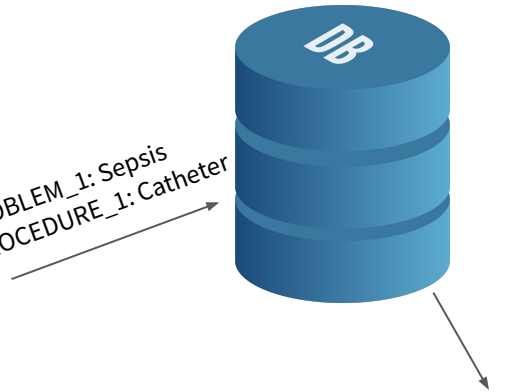
Overall Pipeline

Name Entity Recognition → Seq2Seq with attention → Connect to DB and evaluate

How many patients died with <PROBLEM_1> after receiving <PROCEDURE_1>

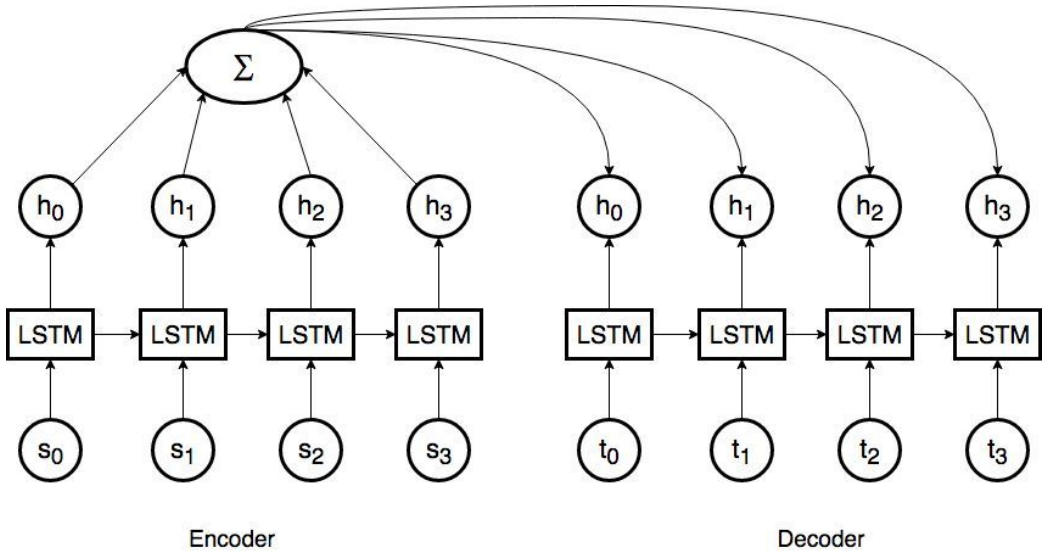
PROBLEM_1: Sepsis
PROCEDURE_1: Catheter

```
select count(distinct subject_id)
from PATIENTS
where expire_flag=1 and subject_id IN
   (select distinct subject_id
   from DIAGNOSES_ICD
   where icd9_code IN
      (select distinct icd9_code
      from D_ICD_DIAGNOSES
      where long_title like 'PROBLEM_1')
   and subject_id IN
      (select distinct subject_id
      from PROCEDURES_ICD
      where icd9_code IN
         (select distinct icd9_code
         from D_ICD_PROCEDURES
         where long_title like 'PROCEDURE_1')))
```



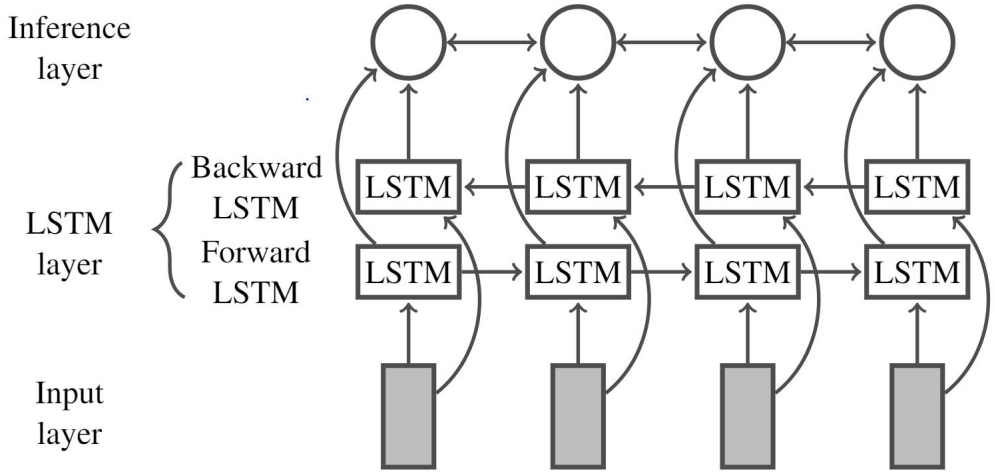
Seq2Seq with attention

- Dong and Lapata(2016) proposed Seq2Seq models with attention for semantic parsing
- Comparatively similar results on GEO(~500) and JOBS(~600)
- Question fed to Encoder. Hidden representation with attention to Decoder



Named Entity Recognition

- Trained on i2b2 2012 clinical NER task (Sun et. al, 2013)
- Dataset has 6 entity types:
  problems, tests, treatments, clinical departments, evidentials, occurrences
- Pretrained word embeddings on MEDLINE and Wikipedia
- Bi-LSTM with Conditional Random Field



Example of NER Output

```
text="medical center" 0:3 --> tag="clinical_dept"
text="physical trauma activation" 9:11 --> tag="problem"
text="w pmhx afib" 24:26 --> tag="problem"
text="coumadin" 28:28 --> tag="treatment"
~~text="dementia" 30:30 --> tag="treatment" ????~~
text="breast ca biba s / p fall" 32:38 --> tag="problem"
~~text="this visit" 79:80 --> tag="problem" ????~~
text="atrial fibrillation" 141:142 --> tag="problem"
text="guilty about drinking" 295:297 --> tag="problem"
text="bun" 544:544 --> tag="test"
text="creatinine" 546:546 --> tag="test"
text="wbc" 567:567 --> tag="test"
text="hct" 569:569 --> tag="test"
text="plt" 571:571 --> tag="test"
text="results" 579:579 --> tag="occurrence"
text="inr" 590:590 --> tag="test"
text="npo" 732:732 --> tag="treatment"
text="ivf" 734:734 --> tag="treatment"
```

Seq2Seq with Attention Architecture

- Dropout on non-recurrent connections for regularization, as suggested by Pham et al. (2014).
- Beam search is used for decoding the SQL queries after learning