



Spark NLP for Healthcare Data Scientists

April 17-18, 2024

Veysel Kocaman, PhD
Head of Data Science
veysel@johnsnowlabs.com

Hasham UI Haq
Sr. Data Scientist

Muhammet Santas
Data Scientist

Gursey Pirge, PhD
Sr. Data Scientist

Youssef Mellah, PhD
Data Scientist

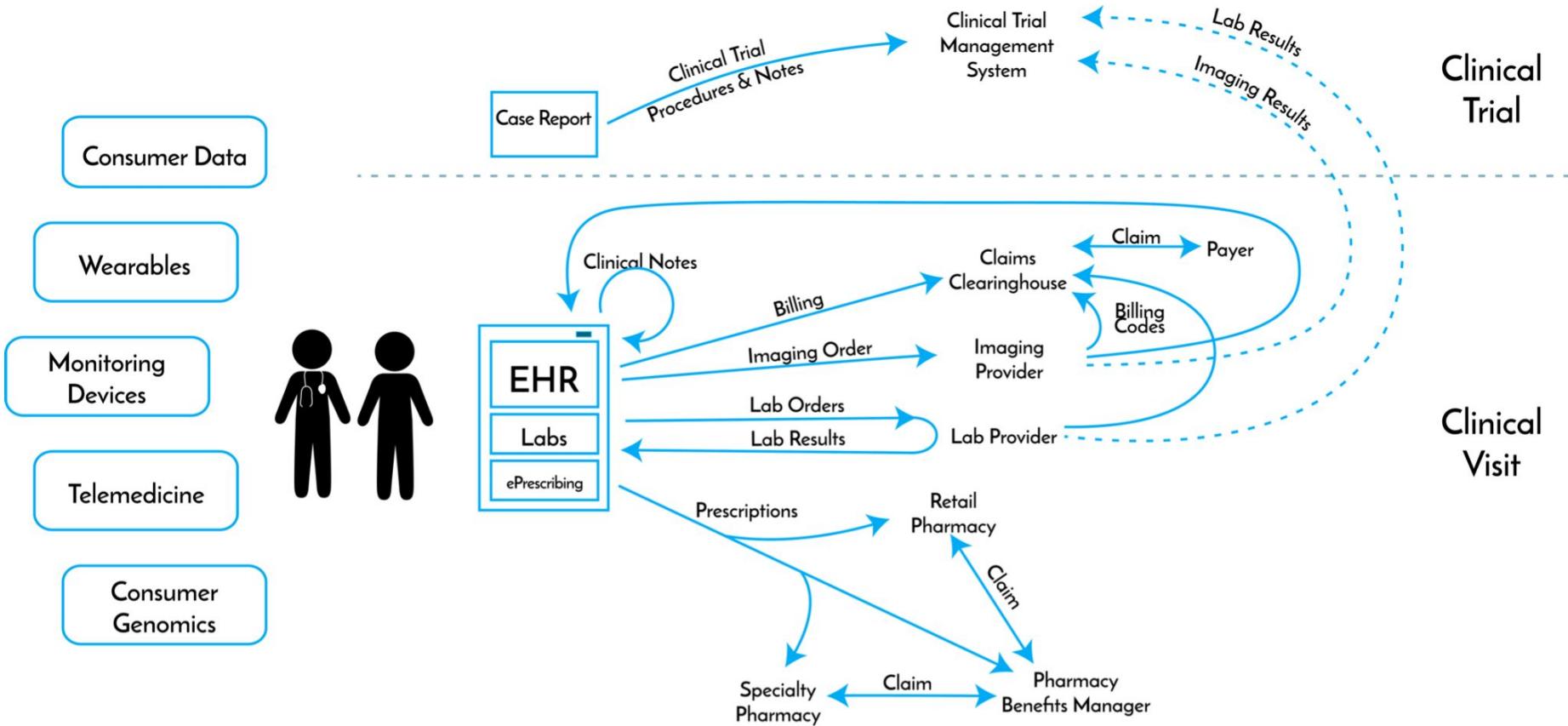
Agenda

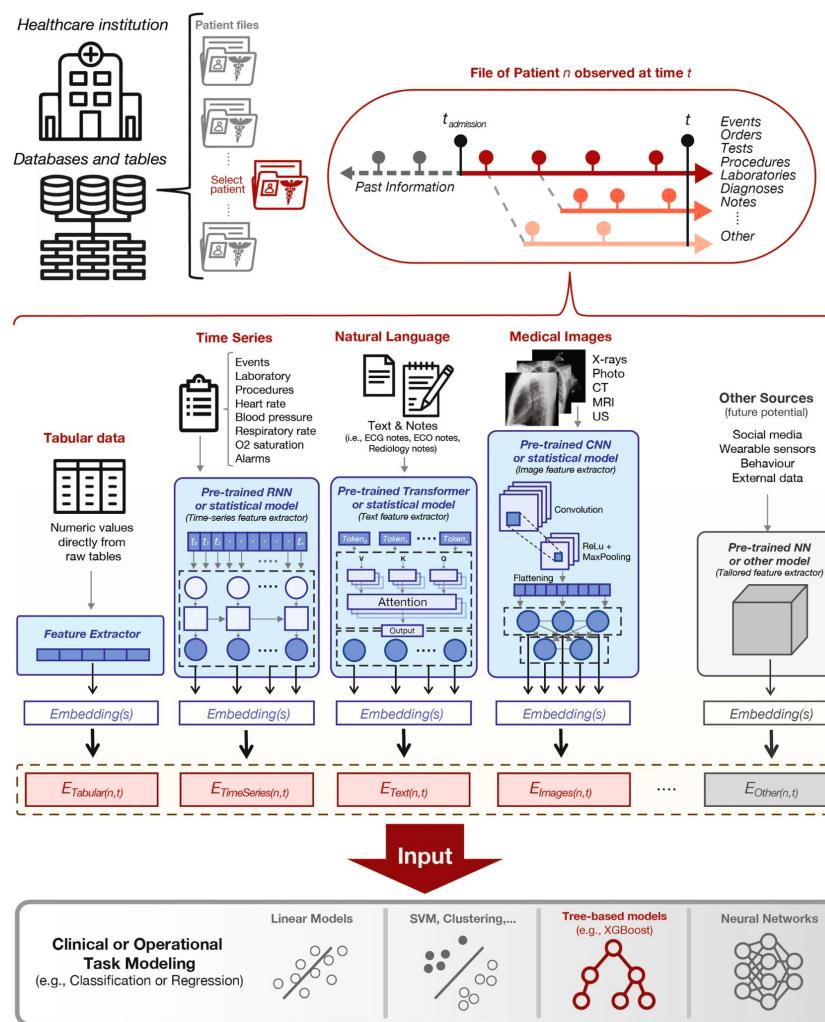
Day	Dur.	Topic	Notebooks	Instructor
April 17 th	50 min	- Intro to John Snow Labs and Spark NLP - Healthcare NLP in Spark NLP	-	Veysel
	50 min	- Clinical Named Entity Recognition	01.0, 10.0, 01.5	Gurseyv
	50 min	- Clinical Assertion Status Model - Clinical Relation Extraction Model	02.0, 03.0	Gurseyv
	50 min	- Clinical Entity Resolution - Medicare Risk Adjustment	05.0, 05.1, 05.2, 05.3	Muhammet
April 18 th	50 min	- Retrieval Augmented Generation with Spark NLP	30.0, spark-nlp-workshop/ generative-ai	Muhammet
	50 min	- LLMs (Text Generation, Text2SQL, Summarization, QA)	23.0, 24.0, 25.1, 29.0	Hasham
	50 min	- De-Identification and Obfuscation of PHI	04.0, 04.1	Youssef
	50 min	- Text Classification (Section Headers&Classifier) - Few Shot Classifier	08.0, 08.3, 08.6, 18.1	Gurseyv

Part - I

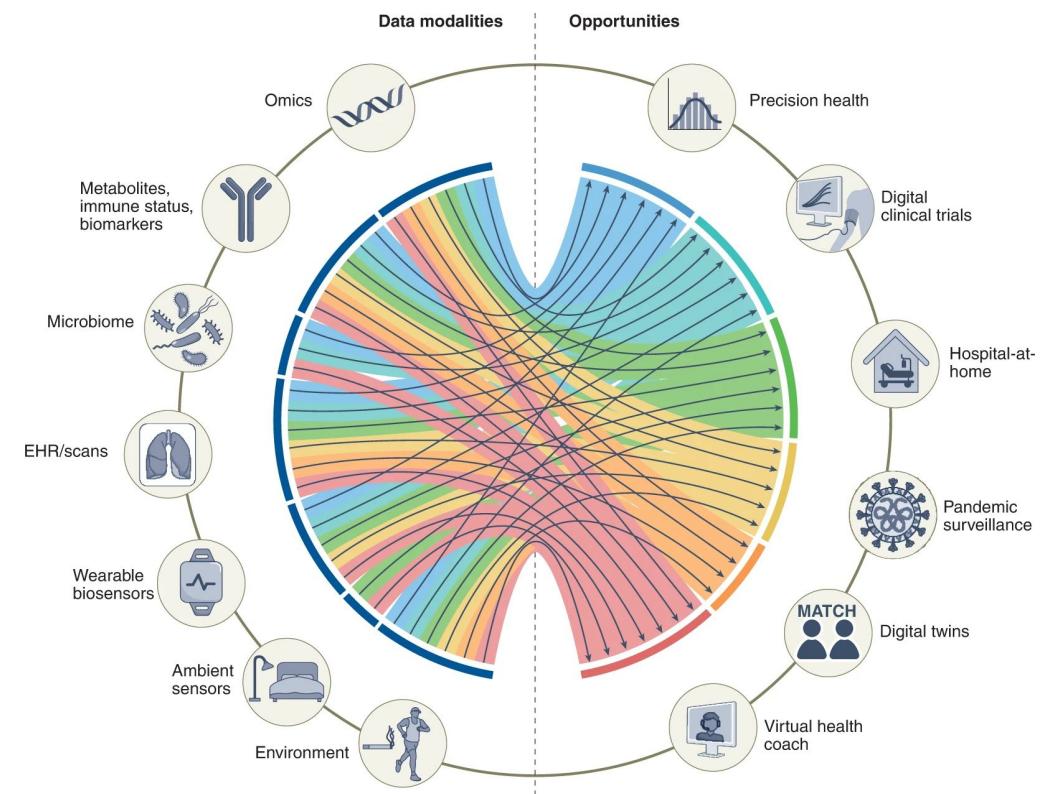
- ❖ Modular Approach to Solve Healthcare Problems in NLP
- ❖ Common medical NLP use cases
- ❖ Intro to John Snow Labs
- ❖ Overview and key concepts in Healthcare NLP

Data Origination and Exchange



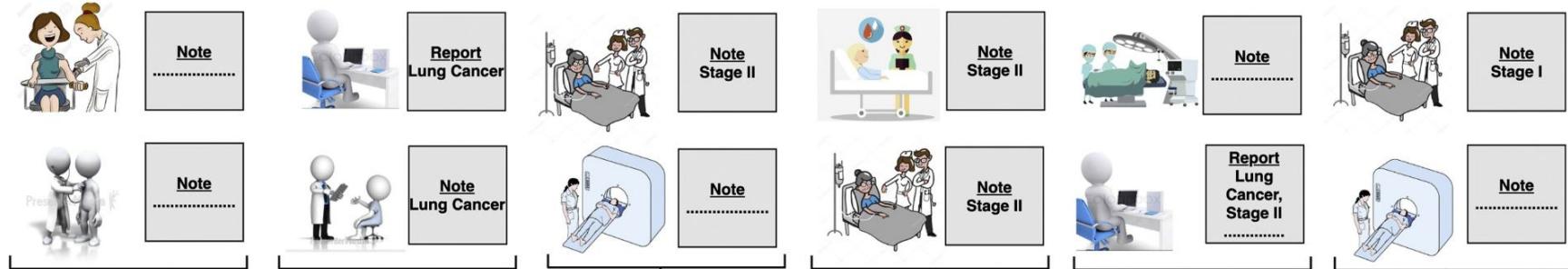


Multimodality in Healthcare



Putting the clinical facts on a timeline

Natural History

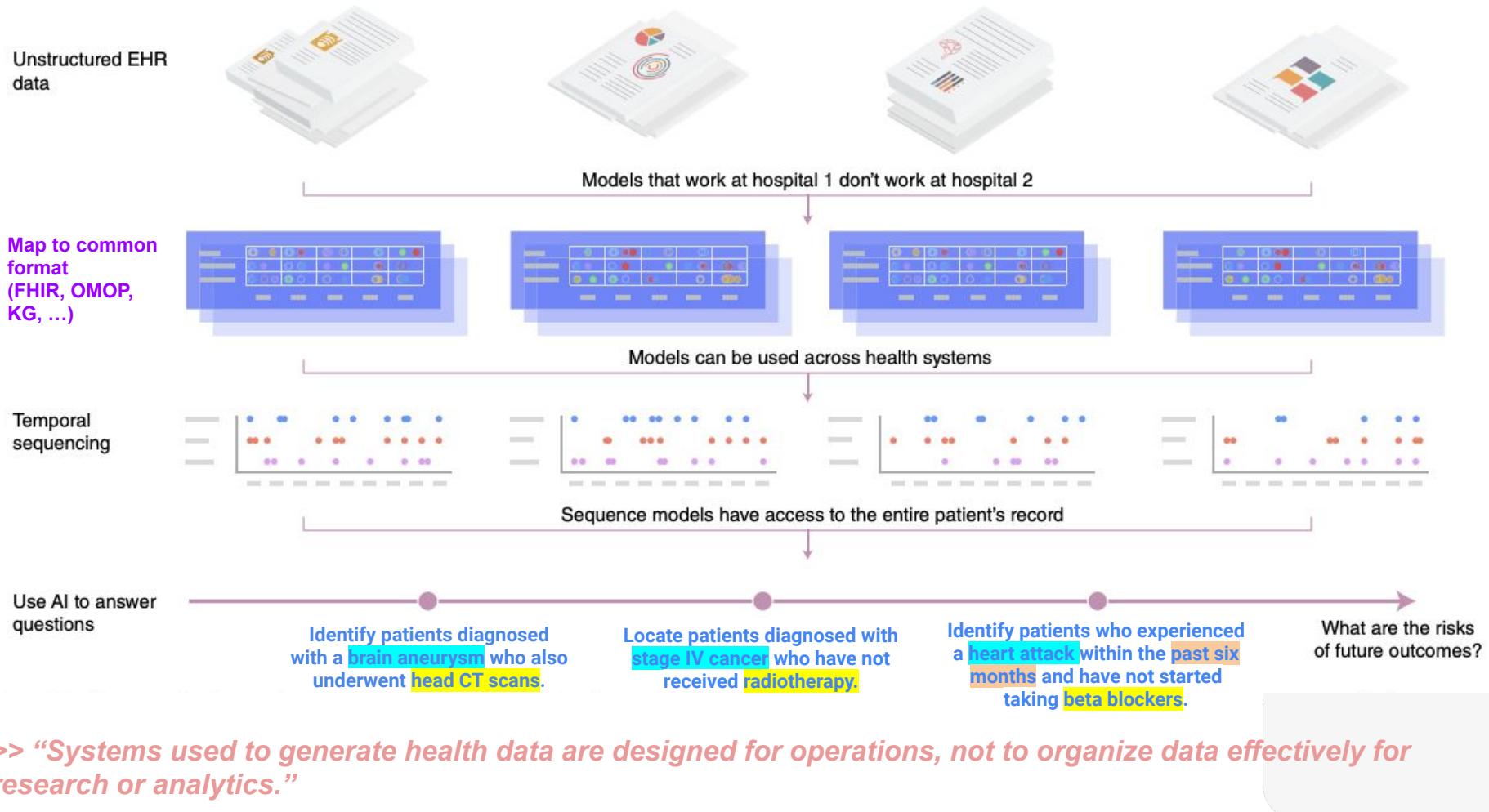


Medical Timeline

Lung Cancer
Diagnosis

Tumor Stage II

Tumor Stage I



Healthcare NLP

Clean & structured data



Raw & unstructured data



Healthcare data



- Less than **50% of the structured data** and less than **1% of the unstructured data** is being leveraged for decision making in companies (HBR). This is even worse in healthcare.
- NLP is ultra domain specific, so train your own models.

Why is language understanding hard?

Human Language is:

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific

Healthcare specific needs:

1. Core Annotators

Part of speech, spell checking, ...

2. Vocabulary

Ontologies, relationships, word embeddings, ...

3. ML & DL Models

Named entity recognition, entity resolution, ...

ED Triage Notes
states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting
Since yesterday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back
Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue. Visiting from Japan on business."



Features	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy

NLP in Healthcare

Case: Predicting if a patient would develop a metastasis on certain sites.



Annotate your own data and train a custom NER model

Recently diagnosed, stage 4 adenocarcinoma of both lungs with metastasis to bone.
CT scan shows no indication of mets on brain.

Extract named entities with Spark NLP *NERDL* and assign assertion statuses with *AssertionDL* model

Feature extraction & engineering

Prediction: Bone metastasis on June 2018



Text embeddings thru clinical word embeddings

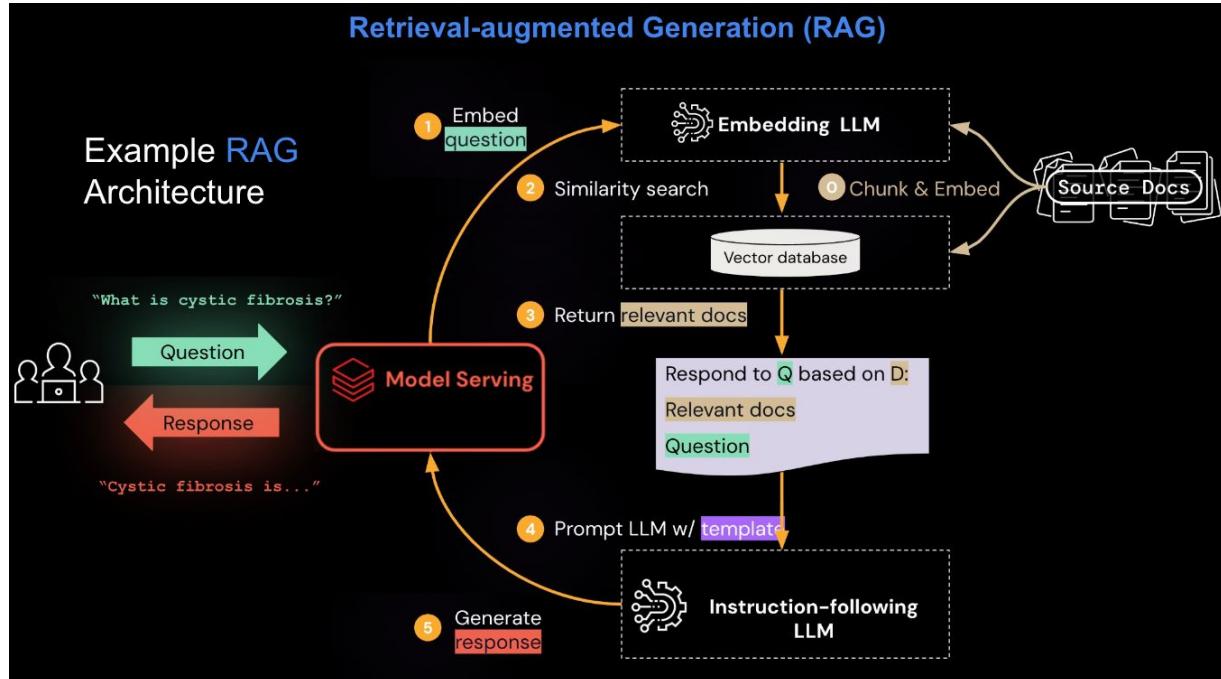
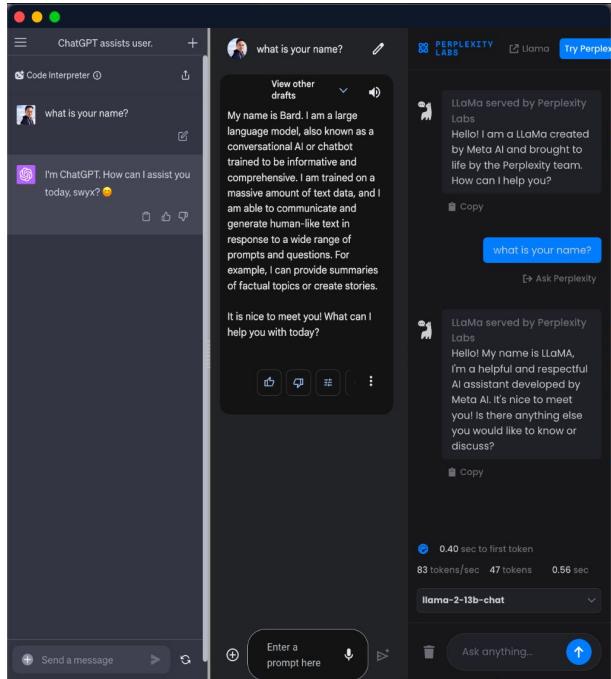
- The frequency of clinical visits using document dates
- The number of positive site (organ) entities (hits by window)
- The number of Radiology/Oncology/Pathology reports in the last x days
- The number of tests applied in the last x days
- The number of diseases detected in the last x days
- Family history and social health determinants, etc.
- Date extraction and normalization

Answering Patient Level Questions via Chatbot

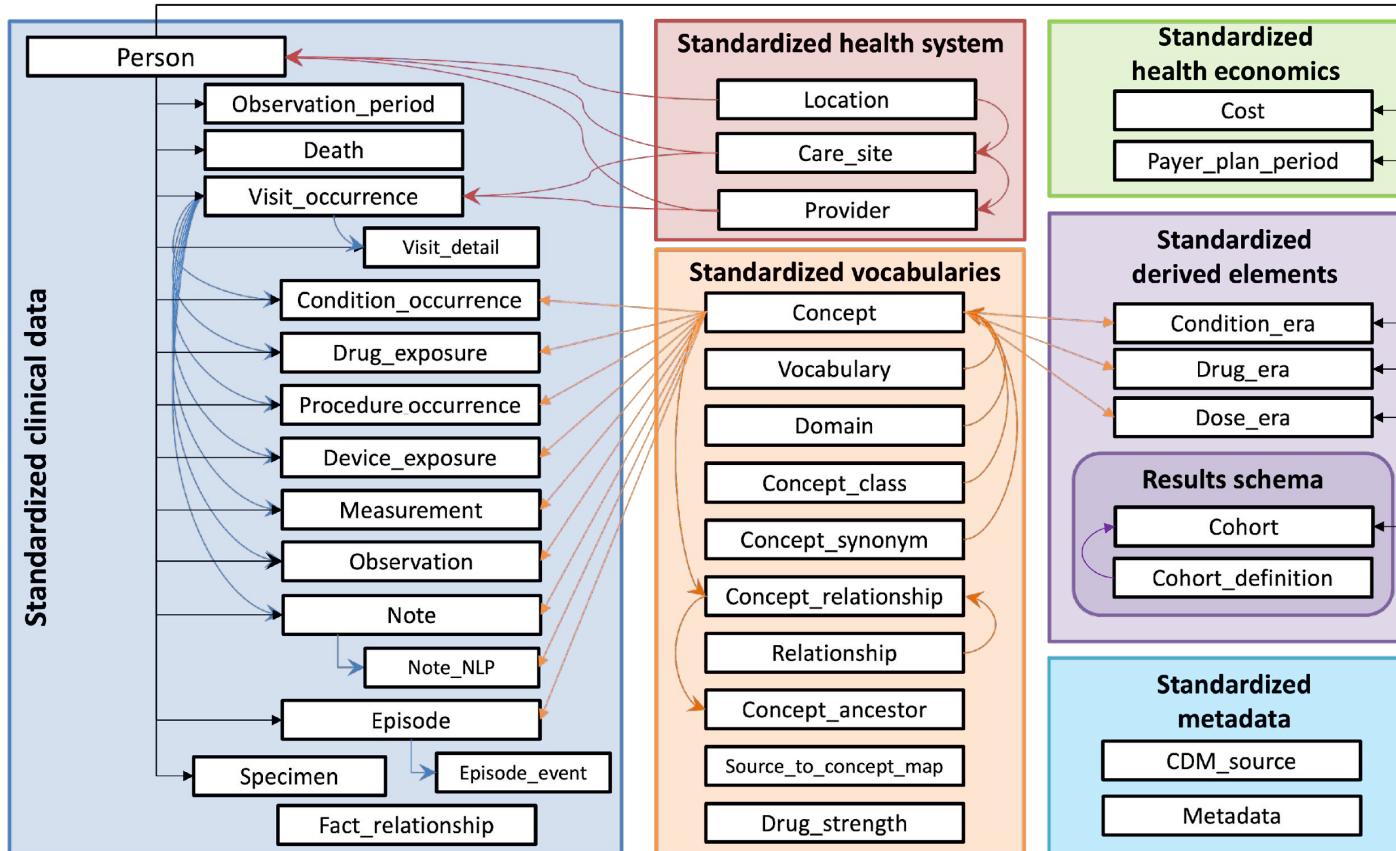
The screenshot shows a web browser window with the URL chat.johnsnowlabs.com. The main content area features a purple header bar with the text "Talk to your Medical Chatbot" and a subtext "Your personal medical assistant - available 24/7 to provide instant answers to patient's health-related questions". Below this is a large, light blue input field containing the placeholder text "Ask me anything about medical data ...". To the left of the input field are several small, semi-transparent icons: a blue gear, a blue plus sign, a blue speech bubble with three dots, a blue person icon, and a blue gear with a circular arrow. In the bottom right corner of the input field, there are two small green circular icons with white symbols. The browser's address bar and toolbar are visible at the top.

No LLM or RAG application can answer this question alone !

>> Give me all the patients who have *type 2 diabetes*, using *metformin* for the *last 3 years*, and also *recently diagnosed stage-IV lung cancer*?



Answering Patient Level Questions from Raw Clinical Data



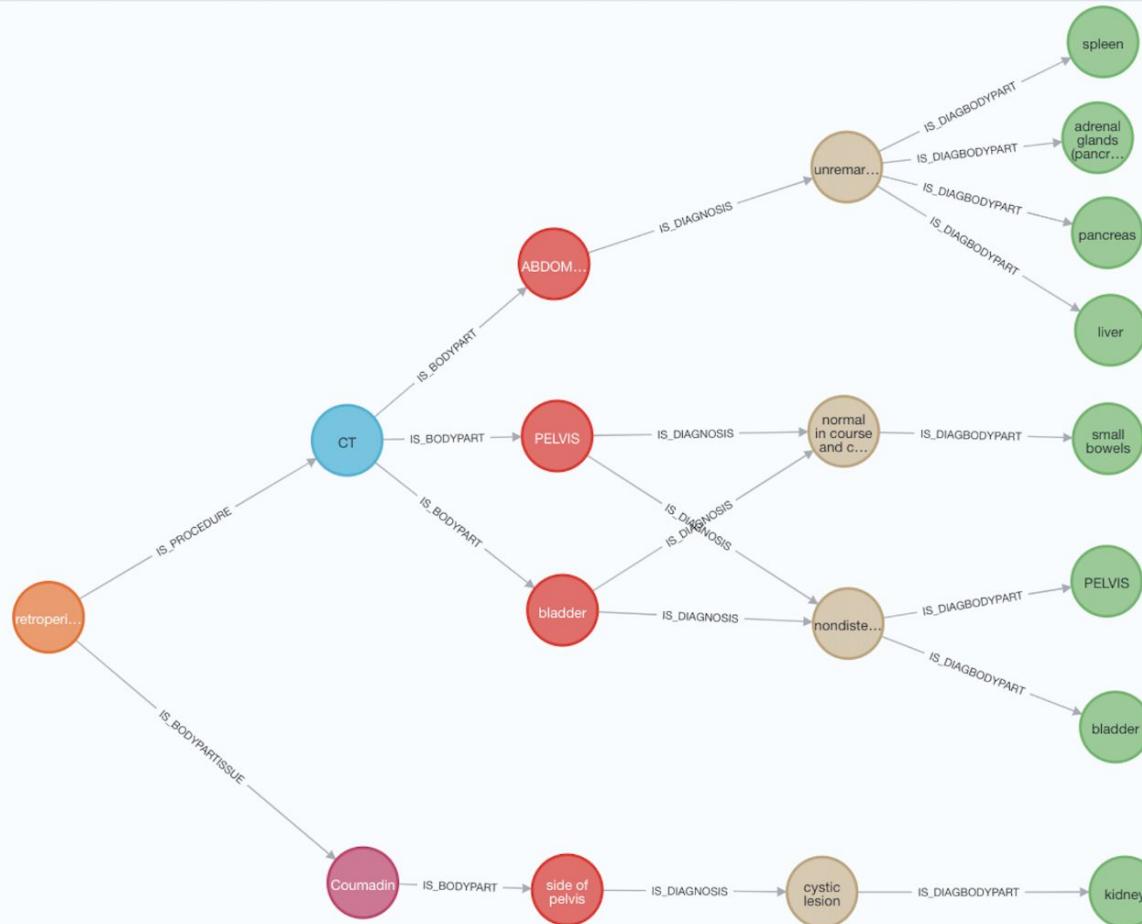
OMOP CDM
Observational
Medical
Outcomes
Partnership -
Common Data
Model

39 tables

433 fields

7 categories

CDM v5.4



REASON FOR EXAM: Evaluate for retroperitoneal hematoma on the right side of pelvis, the patient has been following, is currently on Coumadin.

CT ABDOMEN: There is no evidence for a retroperitoneal hematoma.

The liver, spleen, adrenal glands, and pancreas are unremarkable.

Within the superior pole of the left kidney, there is a 3.9 cm cystic lesion.

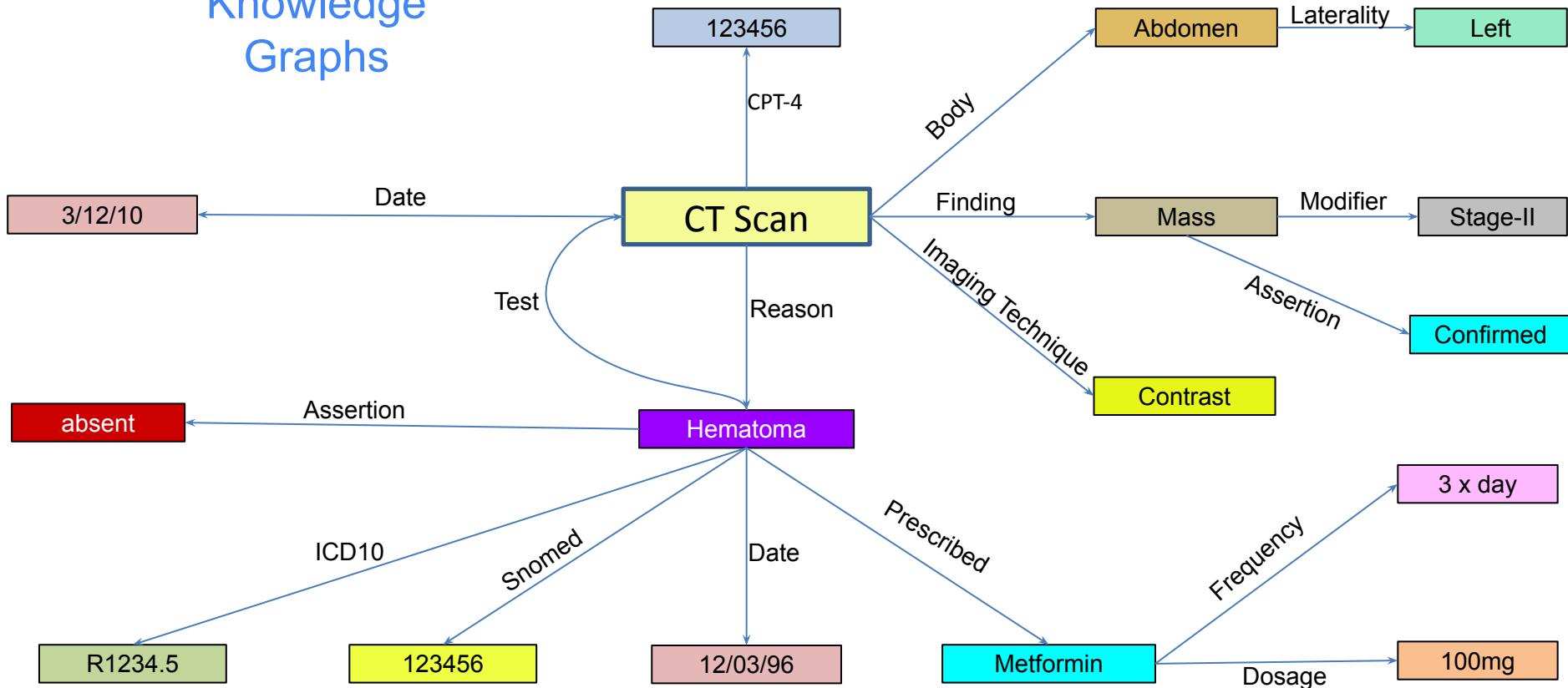
A 3.3 cm cystic lesion is also seen within the inferior pole of the left kidney.

No calcifications are noted. The kidneys are small bilaterally.

CT PELVIS: Evaluation of the bladder is limited due to the presence of a Foley catheter, the bladder is nondistended.

The large and small bowels are normal in course and caliber. There is no obstruction.

Knowledge Graphs



NLP in Healthcare

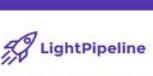
Entity Recognition	Entity Linking	Assertion Status	Relation Extraction
40 units DOSAGE of insulin glargine DRUG at night FREQUENCY	Suspect diabetes SNOMED-CT: 473127005 Lisinopril 10 MG RxNorm: 316151 Hyponatremia ICD-10: E87.1	Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY	
De-Identification	Question Answering	Summarization	Data Enrichment
Katia was born on April 29th PATIENT was born on DATE Olga was born on March 28th	Do preoperative stains reduce arterial fibrillation after CABG? YES		Amoxicillin → RxNorm: 722 → drug class: antibiotic → brand: Amoxil, Larotid

Spark NLP for Healthcare

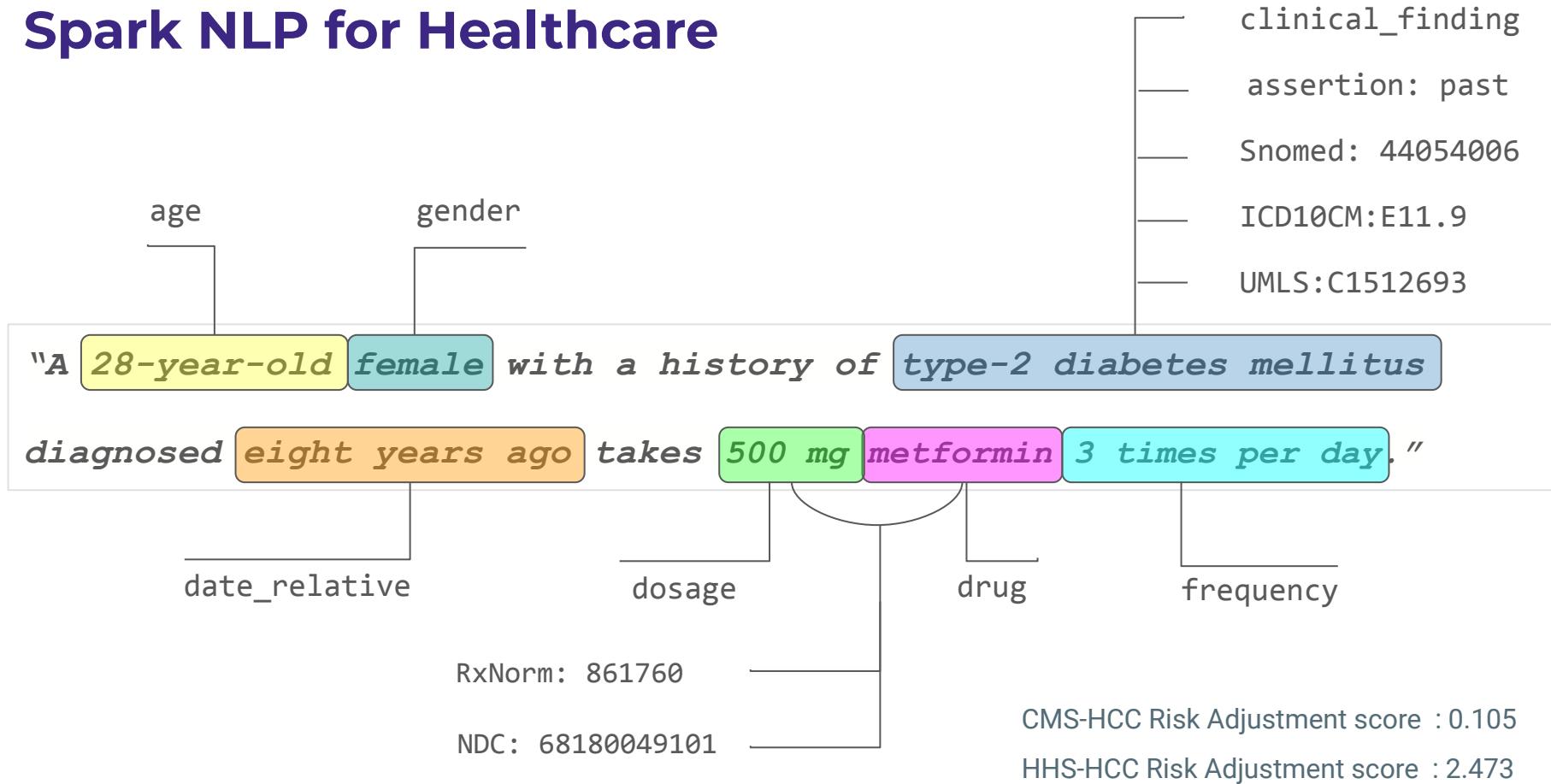
Spark NLP for Healthcare provides

- accurate,
- scalable,
- private,
- tunable,
- modular

software library that helps healthcare & pharma organizations build longitudinal patient records and knowledge graphs on real-world EHR data.

Clinical Entity Recognition	Clinical Entity Linking	Assertion Status	Relation Extraction						
<p>40 units DOSAGE of insulin glargine DRUG at night FREQUENCY</p>	<p>Suspect diabetes SNOMED-CT: 473127005 Lisinopril 10 MG RxNorm: 316151 Hyponatremia ICD-10: E87.1</p>	<p>Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY</p>							
Algorithms		Content							
<p>Extract Knowledge</p> <ul style="list-style-type: none"> • Entity Linker • Entity Disambiguator • Document Classifier • Contextual Parser 		<p>De-identify text</p> <ul style="list-style-type: none"> • Structured Data • Unstructured Text • Obfuscator • Generalizer 							
<p>Split Text</p> <ul style="list-style-type: none"> • Sentence Detector • Deep Sentence Detector • Tokenizer • nGram Generator 		<p>Clean Medical Text</p> <ul style="list-style-type: none"> • Spell Checking • Spell Correction • Normalizer • Stopword Cleaner 							
<p>Clinical Grammar</p> <ul style="list-style-type: none"> • Stemmer • Lemmatizer • Part of Speech Tagger • Dependency Parser 		<p>Find in Text</p> <ul style="list-style-type: none"> • Text Matcher • Regex Matcher • Date Matcher • Chunker 							
Trainable & Tunable	Scalable to a Cluster	Fast Inference	Hardware Optimized						
									
Community		NLP SUMMIT							
									
<p>2100+ Pretrained Models</p> <table border="1"> <tbody> <tr> <td>Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections</td><td>Anatomy: Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical</td></tr> <tr> <td>Drugs: Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects</td><td>Demographics: Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs</td></tr> <tr> <td>Risk Factors: Smoking, Obesity, Diabetes, Hypertension, Substance Abuse</td><td>Sensitive Data: Patient Name, Address, Phone, Email, Dates, Providers, Identifiers</td></tr> </tbody> </table>				Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	Anatomy: Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical	Drugs: Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects	Demographics: Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs	Risk Factors: Smoking, Obesity, Diabetes, Hypertension, Substance Abuse	Sensitive Data: Patient Name, Address, Phone, Email, Dates, Providers, Identifiers
Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	Anatomy: Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical								
Drugs: Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects	Demographics: Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs								
Risk Factors: Smoking, Obesity, Diabetes, Hypertension, Substance Abuse	Sensitive Data: Patient Name, Address, Phone, Email, Dates, Providers, Identifiers								

Spark NLP for Healthcare



Academic Activities & Benchmarks



Preparing for the Next Pandemic: Transfer Learning from Existing Diseases via Hierarchical Multi-Modal BERT Models to Predict COVID-19 Outcomes

Khushbu Agarwal¹, Sutanay Choudhury^{1*}, Sindhu Tipirneni², Pritam Mukherjee³, Colby Ham¹, Suzanne Tamang¹, Matthew Baker⁴, Siyi Tang⁵, Veysel Kocaman⁷, Olivier Gervais^{1,2}, Robert Rallo¹, and Chandan K Reddy¹

¹Pacific Northwest National Laboratory, Richland, 99354, USA

²Department of Computer Science, Virginia Tech, Arlington, 22203, USA

³Stanford Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, 94305, USA

⁴Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA

⁵Department of Electrical Engineering, Stanford University, Stanford, 94305, USA

⁶Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, 94305, USA

⁷John Snow Labs, Delaware City, 19968, USA

stanford.cs@stanford.edu



American Medical
Informatics
Association

Tracking the Evolution of COVID-19 via Temporal Comorbidity Analysis from Multi-Modal Data

Sutanay Choudhury¹, Khushbu Agarwal¹, Colby Ham¹, Pritam Mukherjee², Siyi Tang³, Sindhu Tipirneni³, Chandan Reddy¹, Suzanne Tamang², Robert Rallo¹, Veysel Kocaman⁴,
¹Pacific Northwest National Laboratory; ²Stanford University; ³Virginia Tech;

John Snow Labs

Introduction

We aim to characterize the evolution in the effectiveness of treatment for different patient groups over the course of the COVID-19 pandemic. In contrast to most existing studies¹, we study the evolution of patient trajectories based on unique sets of frequent comorbid conditions discovered from the data. Further, we study the association between frequent co-morbid conditions to the length of stay (LOS) as a measure of treatment efficacy, for poor COVID-19 related outcomes.

Journal of Biomedical Semantics

SOFTWARE

Accurate Clinical and Biomedical Named Entity Recognition at Scale

Veysel Kocaman* and David Talby

*Correspondence:
veysel@johnsnowlabs.com
John Snow Labs, Lewes, DE, USA
Full list of author information is available at the end of the article

Scientific Document Understanding (SDU) at AAAI

Deeper Clinical Document Understanding Using Relation Extraction

Hasham Ul Haq, Veysel Kocaman, David Talby

John Snow Labs Inc.
16192 Coastal Highway
Lewes, DE, USA 19958
{hasham, veysel, david}@johnsnowlabs.com

Abstract

The surging amount of biomedical literature & digital clinical records presents a growing need for text mining techniques that can not only identify but also semantically enrich

publications and literature are growing rapidly, there still lacks structured knowledge that can be easily processed by computer programs. Relation Extraction becomes even more pertinent in biomedical research as it can provide the criti-



New State-of-the-art (SOTA) Benchmarks



- ✓ 6 academic publications & events and 1 patent application, 20+ medium blogposts
- ✓ new SOTA benchmarks on Clinical NER challenges (i2b2 2010 Clinical, i2b2 2014 Deid, n2c2 2018 Medication)
- ✓ new SOTA benchmarks on Adverse Drug Reaction NER datasets (ADE, CADEC, SMM4H)
- ✓ new SOTA benchmarks on Adverse Drug Reaction classification datasets (ADR, CADEC)
- ✓ new SOTA benchmarks on Clinical Relation Extraction datasets (i2b2, temporal, ADE, Posology, PGR – 5 out of 7)



Health
Intelligence
(W3PHIAI-22)
at AAAI

Mining Adverse Drug Reactions from Unstructured Mediums at Scale

Hasham Ul Haq, Veysel Kocaman, David Talby

John Snow Labs Inc.
16192 Coastal Highway
Lewes, DE, USA 19958
{hasham, veysel, david}@johnsnowlabs.com

ADR's has been estimated to cost \$156 billion each year in the United States alone (van Der Hoof et al. 2006).

Finding all ADR's of a drug before it is marketed is not practical for several reasons. First, the number of human subjects going through clinical trials is often too small to detect rare ADR's. Second, many clinical trials are short-lasting while some ADR's take time to manifest. Third,

John Snow Labs

is the team behind Spark NLP

Most popular

NLP library in
the enterprise

O'Reilly Media

59% share

of Healthcare NLP
teams use Spark NLP

Gradient Flow

#1 Accuracy

on 20 benchmarks in
peer-reviewed papers

Papers with Code

TRUSTED BY



Imperial College
London



STANFORD
UNIVERSITY

Customer Case Studies - Life Science Industry

Safety

- Identify Opioid-related AEs from Unstructured Text;
- Social Media Listening;
- Semi-automate AE Monitoring



Genentech
A Member of the Roche Group



Digital Health

- Accelerate Biomedical Innovation by Combining NLP & Knowledge Graphs



Clinical Trial Operations

- Automate Classification and Entity Extraction from Clinical Trial Documents;
- Accelerate Clinical Trial Recruitment;
- Inclusive Clinical Research



Real World Evidence

- Regulatory Grade Oncology Data Curation
- De-Identify Doctor Notes in German Language



Medical Affairs

- Build an Integrated Data Approach to Medical Affairs



Financial Operations

- Automate Invoice Classification for Submission Compliance



Customer Case Studies - Healthcare Industry

De-Identification

- De-Identify 700M Patient Notes



Clinical Data Abstraction

- Extract what, when, why and how from Radiology Reports in Real World Data Acquisition Projects



Patient Safety

- Identify opioid-related adverse events from unstructured text in electronic health records



Clinical Decision Support

- A Real-time NLP-based Clinical Decision Support Platform for Psychiatry and Oncology
- Identify Housing Insecurity and other Social Determinates of Health from Free-Text notes



Medical Affairs

- Optimize hospital patient flow models



Financial Operations

- Accelerate Clinical Risk Adjustment Through NLP





Introducing Spark NLP

PyPI link

<https://pypi.org/project/spark-nlp>

Total downloads
97,813,020

downloads 98M

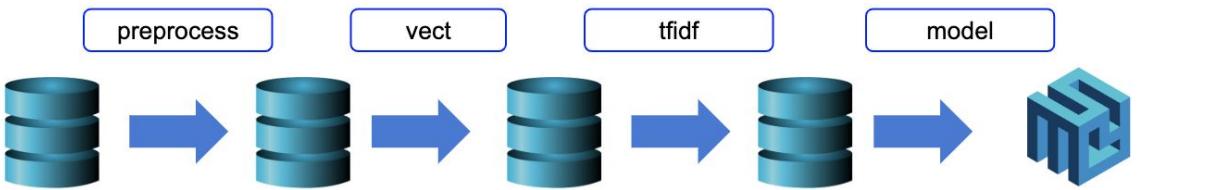
downloads/month 4M

downloads/week 805k

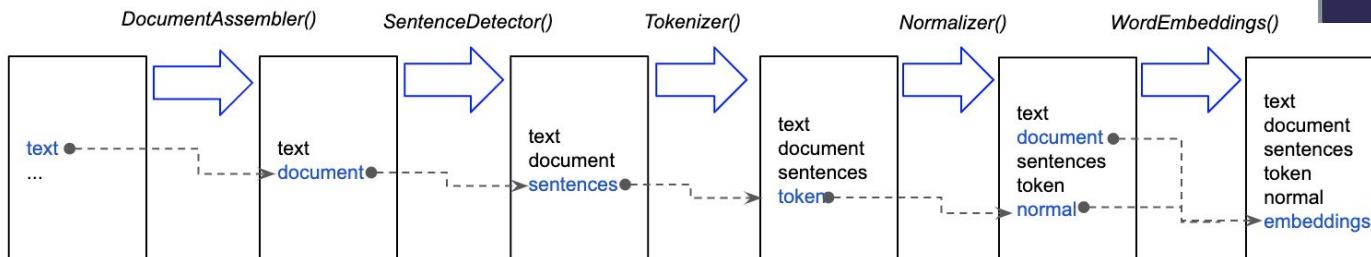
- Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. (initial release: Oct 2017)
 - A single unified solution for all your NLP needs
 - Take advantage of transfer learning and implementing the latest and greatest SOTA algorithms and models in NLP research
 - The most widely used NLP library in industry (3 yrs in a row) - downloaded 98+ million times !
 - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
 - Full-time development team (a new release every other week)

Introducing Spark NLP

Pipeline of annotators



```
from pyspark.ml import Pipeline
documentAssembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")
sentenceDetector = SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentences")
tokenizer = Tokenizer() \
    .setInputCols(["sentences"]) \
    .setOutputCol("token")
normalizer = Normalizer()\
    .setInputCols(["token"])\
    .setOutputCol("normal")
word_embeddings=WordEmbeddingsModel.pretrained()\
    .setInputCols(["document","normal"])\
    .setOutputCol("embeddings")
nlpPipeline = Pipeline(stages=[documentAssembler,
    sentenceDetector,
    tokenizer,
    normalizer,
    word_embeddings,
])
nlpPipeline.fit(df).transform(df)
```



Spark NLP for Healthcare

Named Entity Recognition

ICD10 Resolver

Snomed Resolver

UMLS Resolver

Assertion Status Detection

Risk Adj. Module

RxNorm Resolver

Relationship Extraction

clinical_finding

Snomed: 44054006

ICD10CM:E11.9

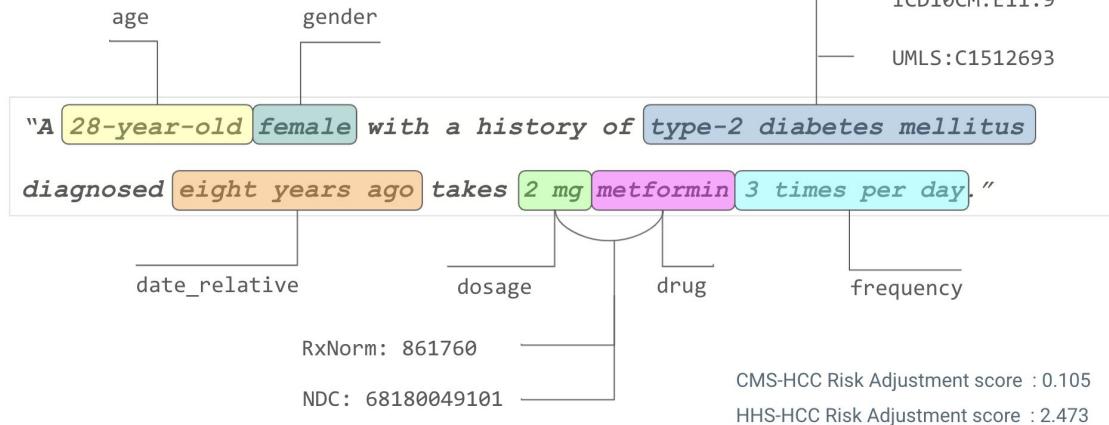
UMLS:C1512693

Sentence Splitter

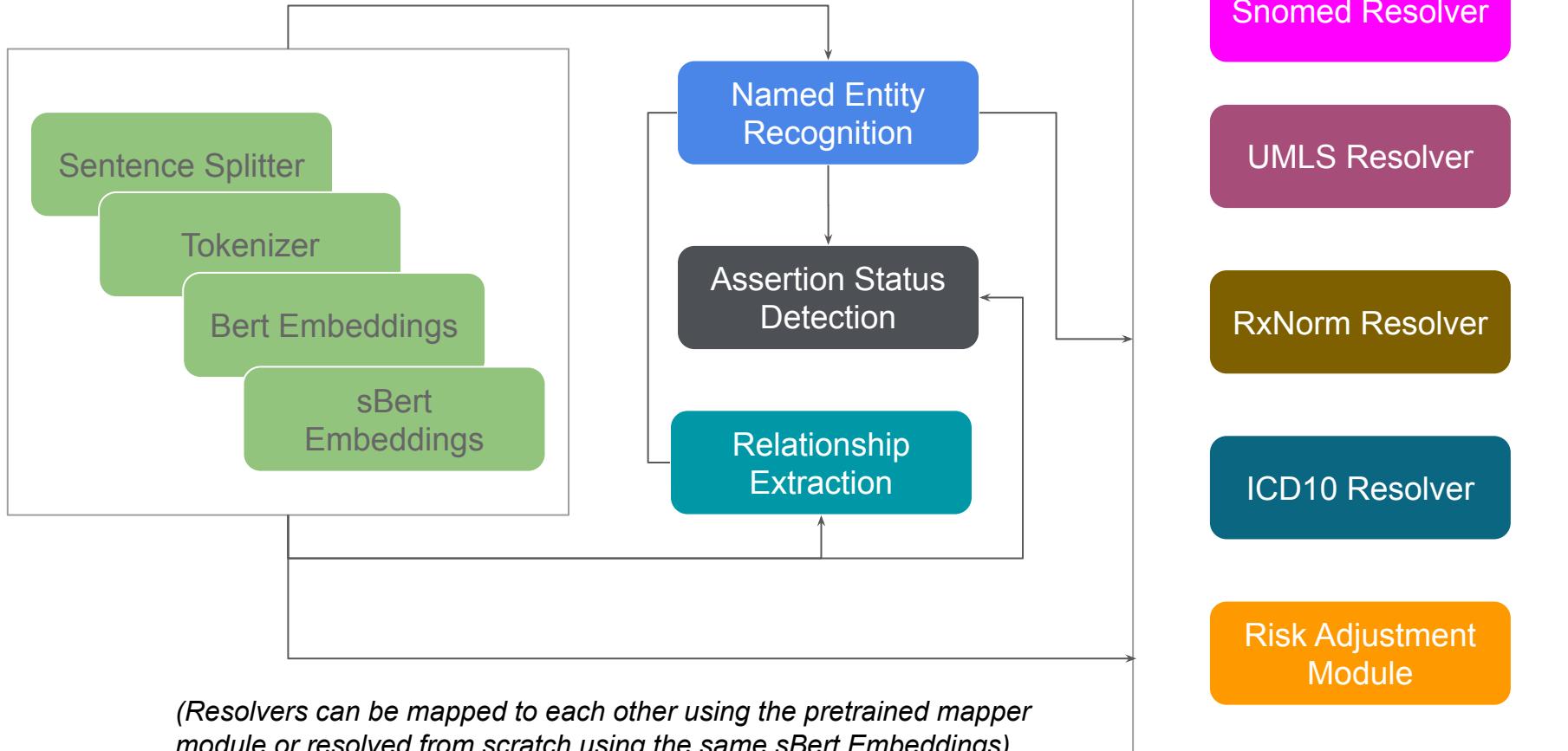
Tokenizer

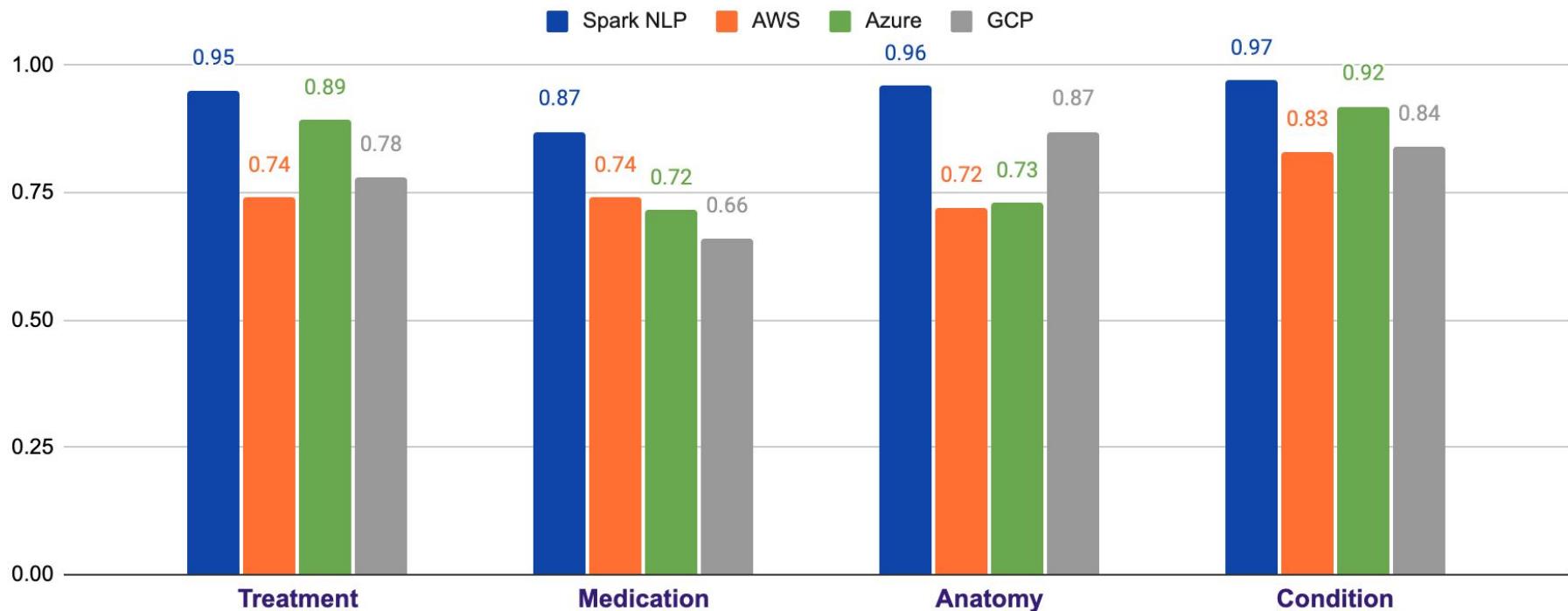
Bert Embeddings

sBert Embeddings



Spark NLP for Healthcare





Pretrained NER Models (300+)

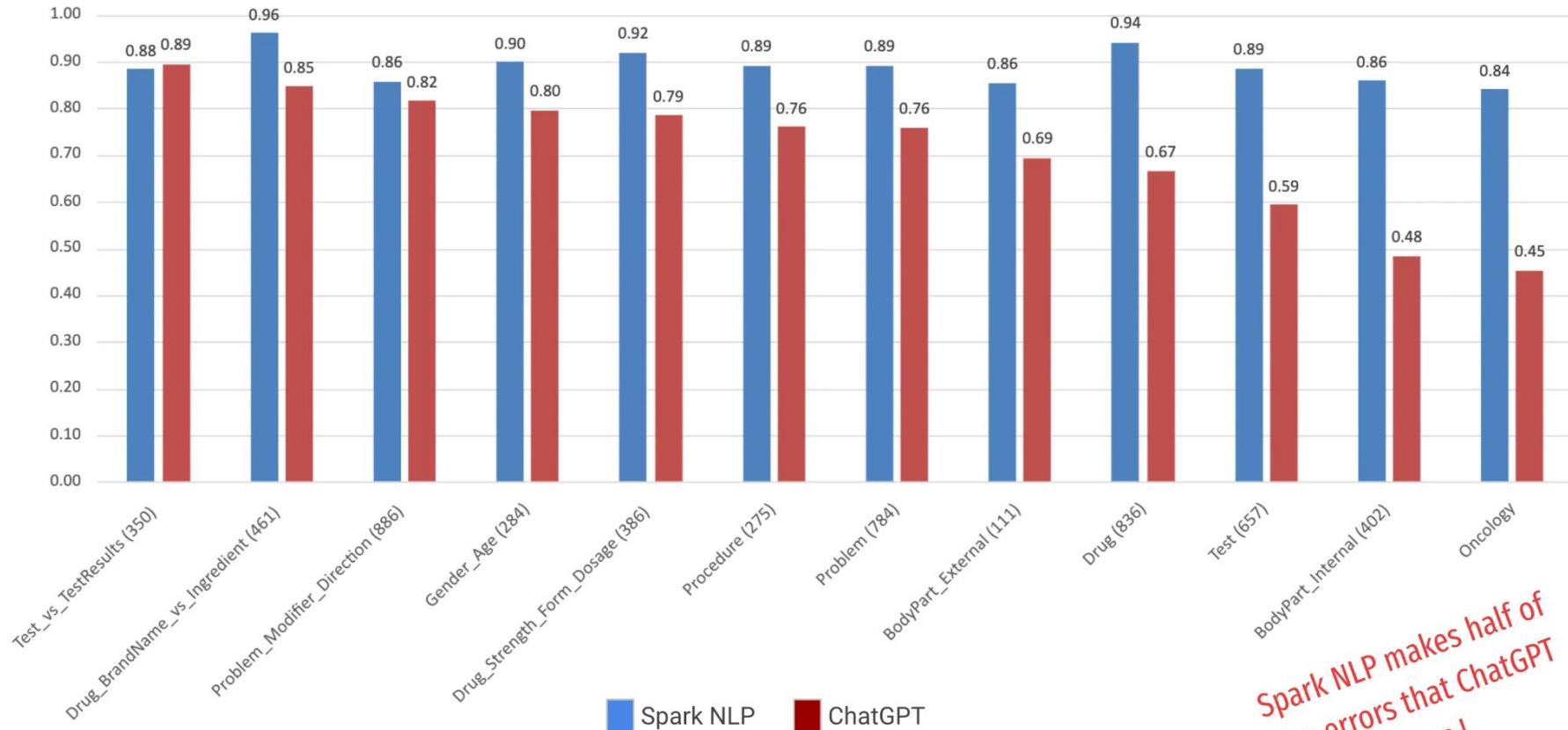
ner_ade_clinical
ner_posology_greedy
ner_risk_factors
jsl_ner_wip_clinical
ner_human_phenotype_gene_clinical
jsl_ner_wip_greedy_clinical
ner_cellular
ner_cancer_genetics
jsl_ner_wip_modifier_clinical
ner_drugs_greedy
ner_deid_sd_large
ner_diseases
nerdl_tumour_demo
ner_deid_subentity_augmented
ner_jsl_enriched
ner_genetic_variants
ner_bionlp
ner_measurements_clinical
ner_diseases_large
ner_radiology
ner_deid_augmented
ner_anatomy
ner_chemprot_clinical

ner_posology_experimental
ner_drugs
ner_deid_sd
ner_posology_large
ner_deid_large
ner_posology
ner_deidentify_dl
ner_deid_enriched
ner_bacterial_species
ner_drugs_large
ner_clinical_large
jsl_rd_ner_wip_greedy_clinical
ner_medmentions_coarse
ner_radiology_wip_clinical
ner_clinical
ner_chemicals
ner_deid_synthetic
ner_events_clinical
ner_posology_small
ner_anatomy_coarse
ner_human_phenotype_go_clinic
ner_jsl_slim
ner_jsl
ner_jsl_greedy
ner_events_admission_clinical

500+ entities from 300+ NER models

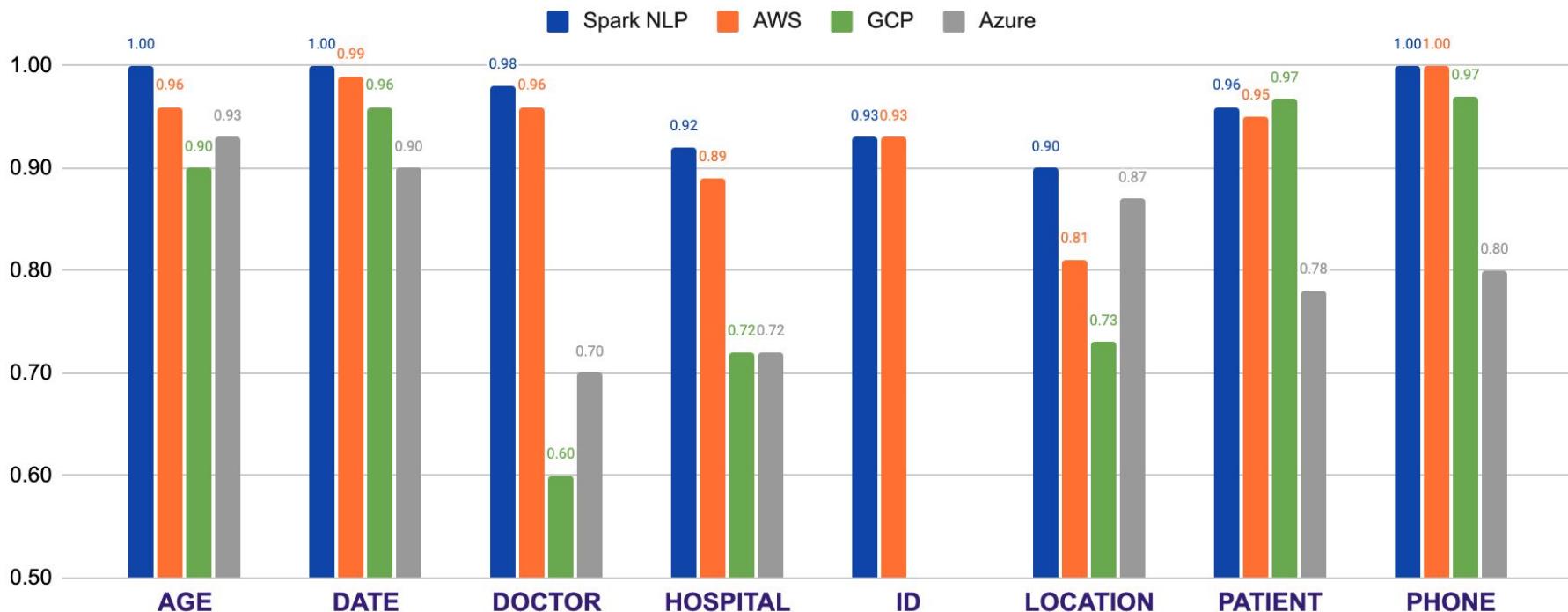


Spark NLP for Healthcare vs ChatGPT (GPT 3.5) on Clinical Entities



Spark NLP makes half of
the errors that ChatGPT
does!

De-Identification Benchmarks (en)



Spark NLP in Action

Spark NLP - English → Recognize Entities

Spark NLP, English

- Infer Meaning & Intent
- Classify Sentiments
- Recognize Entities
- Detect Sentiment & Emotion
- Analyze Spelling & Grammar

Spark NLP, World Languages

- Identify & Translate Languages
- European Languages
- East Asian Languages
- Language of India
- Middle Eastern Languages
- Language of Africa

Spark NLP for Healthcare

- Recognize Clinical Entities
- Recognize Biomedical Entities
- De-identification
- Resolve Entities to Codes
- Recognize Social Determinants
- Extract Entities
- Soft & Clean Medical Text
- Analyze non-English Medical Text

Spark OCR

- Extract Text from Documents
- Enhance Low-Quality Images
- Extract Tables & Structured Data
- Analyze non-English Medical Text

John Snow Labs NLP Documentation

John Snow Labs NLP Documentation

Spark NLP

Healthcare NLP
Legal NLP
Finance NLP

Visual NLP

Annotation Lab

NLP Server

John Snow Labs NLP

NLP Models Hub

A place for sharing and discovering Spark NLP models and pipelines

Show: All | models & pipelines in: All Languages | for: All Spark NLP versions

4,390 Models & Pipelines Results:

Model Type	Description	Date	Task	Language	Edition
SUPPORTED	Sentence Entity Resolver for UMLS CUI Codes (Clinical Drugs)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for UMLS CUI Codes (Disease or Syndrome)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for RxNorm (product_base_cased_mit embeddings)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Longformer Token Classification Base - NER CoNLL	10/2021	Named Entity Recognition	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Longformer Token Classification Base - NEB CoNLL	10/2021	Named Entity Recognition	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for RxNorm (NDC)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3

nlp.johnsnowlabs.com/licensed/api/python/reference/autosummary/sparknlp_jsl.annotator.html

JohnSnowLabs / spark-nlp-workshop Public

Code Issues Pull requests Discussions Actions Projects Wiki Security

master · 75 branches · 7 tags · Go to file · Add file · Code ·

galiph Merge pull request #400 from JohnSnowLabs/galiph · 1,601 commits

- data Add files via upload
- dbnabrics Update benchmark.md
- java add java examples
- jupyter Update docker-compose.yaml
- mlu mlu notebooks updated
- platforms added more info for troubleshooting
- scala update scala codes
- tutorials Notebooks updated with v3.3.0
- zeppelin clean notebook
- .gitattributes Ignore html from linguist-vendorized
- .gitignore removed outdated notebooks
- ISSUE_TEMPLATE.md Create ISSUE_TEMPLATE.md
- LICENSE Initial commit
- README.md Update README.md
- colab_setup.sh Update colab_setup.sh
- js_colab_setup.sh Update js_colab_setup.sh
- js_colab_setup_with_OCR.sh Update js_colab_setup_with_OCR.sh
- js_sagemaker_setup.sh Corrected paths for spark binaries
- js_sagemaker_setup_3.0.tsh added script
- js_sagemaker_setup_with_OCR.sh Update js_sagemaker_setup_with_OCR.sh

sparknlp_jsl.annotator

com.johnsnowlabs.nlp.annotators

C **MedicalNerApproach** Companion object **MedicalNerApproach**

```
class MedicalNerApproach extends AnnotatorApproach[MedicalNerModel] with NerApproach[MedicalNerAnnotator] with Logging with ParamAndFeatureWriterTrait with DefaultLicense
```

Linear Superatypes

Filter all members

anno

```
val inspectionEntityTypes: Array[String]
Input annotator types: DOCUMENT, TOKEN, WORD_EMBEDDINGS

val outputAnnotationsType: String
Input annotator types: NAMED_ENTITY
```

getParam

```
def getBatchSize: Int
Batch size

def getConfigProtoBytes: Option[Array[Byte]]
Config proto from tensorflow, serialized into byte array.

def getDropout: Float
Dropout coefficient

def getEnableNLPyOptimizer: Boolean
Memory Optimizer

def getEnableOutputLogs: Boolean
Whether to output to annotators log folder

def getIncludeAllConfidenceScores: Boolean
```

<https://nlp.johnsnowlabs.com/demos>

Spark NLP in Action

Run 300+ live demos and notebooks

Spark NLP: English

Spark NLP: World Languages

Clinical NLP

De-Identification

Diagnoses & Procedures

Drugs & Adverse Events

Labs, Tests, and Vitals

Analyze Clinical Notes

Radiology

Oncology

Resolve Entities to Terminology Codes

Databricks Solution Accelerators

Social Determinants of Health

Risk Factors

LangTest

Explore Healthcare NLP Models



Medical Large Language Models

Explore the use of Medical Large Language Models for tasks like Text Summarization, Generation, and Question Answering.

[Live Demo](#)

[Colab](#)



Detect Entities in Clinical Text

Identify 77 entity types including Symptom, Treatments, Test, Oncological, Procedure, Diabetes, Drug, Dosage, Date, Imaging Finding, and more.

[Live Demo](#)

[Colab](#)



Information Extraction in Oncology

Detect clinical entities and relationships related to cancer staging, grading, histology, tumor characteristics, biomarkers, treatments, and outcome measures.

[Live Demo](#)

[Colab](#)



De-identify Clinical Notes in Different Languages

De-identify and obfuscate protected health information (PHI) in English, Spanish, French, Italian, Portuguese, Romanian, and German texts.

[Live Demo](#)

[Colab](#)



Adverse Drug Event Detection

Detect adverse reactions from drugs described in the clinical text, online reviews, and social media posts.

[Live Demo](#)

[Colab](#)



Voice of the Patients

Extract and classify healthcare-related terms from documents written by patient such as questions, reviews, messages, and social media posts.

[Live Demo](#)

[Colab](#)



Social Determinants of Health

Extract Social Determinants of Healthcare such as employment, education, social support, housing, financial hardship, substance abuse, demographics, and more.

[Live Demo](#)

[Colab](#)



Calculate Medicare HCC Risk Score

Automatically calculate patient risk adjustment scores, using ICD codes of diseases that are extracted from clinical notes about a patient.

[Live Demo](#)

[Colab](#)



Recommend Available Models for Your Text

This pipeline is used to explore all the available pretrained entity recognition models at once. It recommends which models will provide results on a given document.

[Live Demo](#)

[Colab](#)

Colab Notebooks

RUNNING CODE:

<https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/healthcare-nlp>

 master / spark-nlp-workshop / healthcare-nlp /

BOOKMARK:

<https://nlp.johnsnowlabs.com/models>

<https://nlp.johnsnowlabs.com/docs/en/quickstart>

<spark-nlp.slack.com>

[Go to file](#) [Add file](#) [...](#)

 Open in Colab

► Clinical Named Entity Recognition

- [01.0.Clinical_Named_Entity_Recognition_Model.ipynb](#)
- [01.1.prepare_CoNLL_from_annotations_for_NER.ipynb](#)
- [01.2.Resume_MedicalNer_Model_Training.ipynb](#)
- [01.3.BertForTokenClassification_NER_SparkNLP_with_Transformers.ipynb](#)
- [01.4.ZeroShot_Clinical_NER.ipynb](#)

► Clinical Assertion

- [02.0.Clinical_Assertion_Model.ipynb](#)
- [02.1.Scope_window_tuning_assertion_status_detection.ipynb](#)

► Clinical Relation Extraction

- [03.0.Clinical_Relation_Extraction.ipynb](#)
- [03.1.Clinical_Relation_Extraction_BodyParts_Models.ipynb](#)
- [03.2.Clinical_RE_Knowledge_Graph_with_Neo4j.ipynb](#)
- [03.3.ZeroShot_Clinical_Relation_Extraction.ipynb](#)
- [03.4.Resume_RelationExtractionApproach_Training.ipynb](#)

► Clinical De-identification

- [04.0.Clinical_Deidentification.ipynb](#)
- [04.1.Clinical_Multi_Language_Deidentification.ipynb](#)
- [04.2.Clinical_Deidentification_SparkNLP_vs_SpaCy_vs_Scrubadub_vs_Presidio_Comparison.ipynb](#)
- [04.3.Clinical_Deidentification_SparkNLP_vs_Cloud_Providers_Comparison.ipynb](#)
- [04.4.Clinical_Deidentification_Improvement.ipynb](#)
- [04.5.Clinical_Deidentification.Utility_Module.ipynb](#)

► Clinical Entity Resolution

- [05.0.Clinical_Entity_Resolvers.ipynb](#)
- [05.1.Clinical_Entity_Resolver_Model_Training.ipynb](#)
- [05.2.Finetuning_Clinical_Entity_Resolver_Model.ipynb](#)
- [05.3.Calculate_Medicare_Risk_Adjustment_Score.ipynb](#)
- [05.4.Sentence_Entity_Resolvers_with_EntityChunkEmbeddings.ipynb](#)
- [05.5.Improved_Entity_Resolvers_in_SparkNLP_with_sBert.ipynb](#)
- [05.6.Improved_Entity_Resolution_with_SentenceChunkEmbeddings.ipynb](#)

► Entity Code Mapping

- [06.0.Chunk_Mapping.ipynb](#)
- [06.1.Code_Mapping_Pipelines.ipynb](#)

► Large Language Modeling

- [23.0.Medical_Question_Answering](#)
- [23.1.Porting_QA_Models_From_Text_Generator_Backbone](#)
- [24.0.Medical_Text_Summarization](#)
- [24.1.Medical_Text_Summarization_with_Abstractive_Approach](#)
- [24.2.Medical_Text_Summarization_with_Extractive_Approach](#)
- [24.3.Comparison_Medical_Text_Summarization](#)
- [25.0.Biogpt_Chat_JSL](#)
- [25.1.Medical_Text_Generation](#)

► Topic Based Notebooks

- [21.0.Oncology_Model.ipynb](#)
- [22.0.Adverse_Drug_Event_ADE_NER_and_Classifier.ipynb](#)
- [26.0.Voice_of_Patient_Models.ipynb](#)
- [27.0.Social_Determinant_of_Health_Models.ipynb](#)

► Clinical Pipelines

- [07.0.Pretrained_Clinical_Pipelines.ipynb](#)
- [07.1.Pretrained_NER_Profiling_Pipelines.ipynb](#)

► Classifiers

- [08.0.Clinical_Text_Classification_with_SparkNLP.ipynb](#)
- [08.1.Text_Classification_with_DocumentMLClassifier.ipynb](#)
- [08.2.Generic_Classifier.ipynb](#)
- [08.3.MedicalBertForSequenceClassification_in_SparkNLP.ipynb](#)
- [08.4.Gender_Classifier.ipynb](#)
- [08.5.Text_Classification_with_Contextual_Window_Splitting.ipynb](#)
- [08.6.Text_Classification_with_FewShotClassifier](#)

► Normalizers

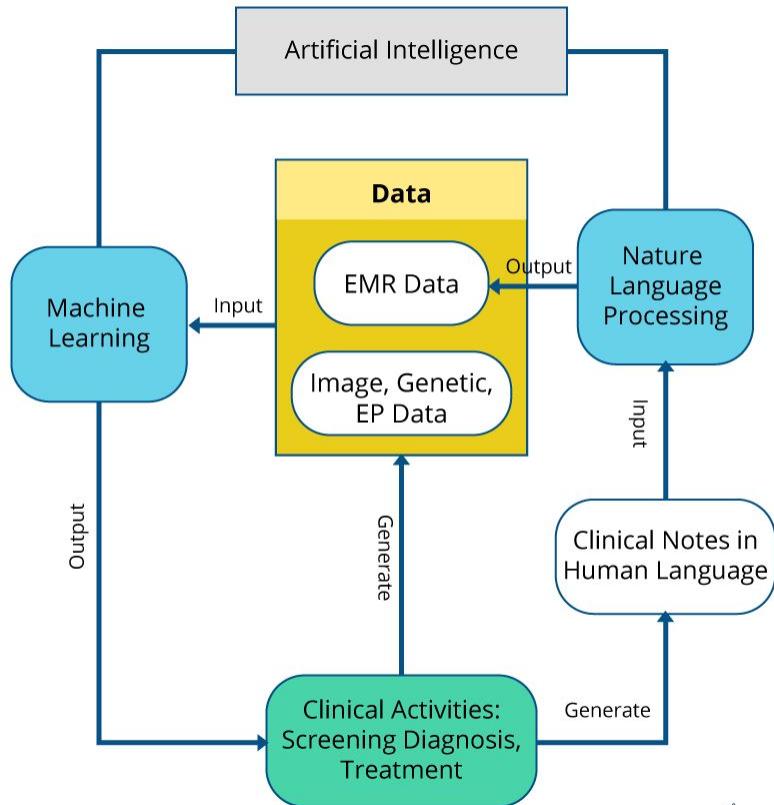
- [13.0.Date_Normalizer.ipynb](#)
- [14.0.Drug_Normalizer.ipynb](#)

► Auxillary Notebooks

- [15.0.EntityRuler_with_Clinical_NER_Models.ipynb](#)
- [16.0.Coreference_Resolution_with_Clinical_NER_Models.ipynb](#)
- [17.0.Graph_builder_for_DL_models.ipynb](#)
- [18.0.Chunk_Sentence_Splitter.ipynb](#)
- [19.0.Chunk_Key_Phrase_Extraction.ipynb](#)
- [20.0.Named_Entity_Disambiguation.ipynb](#)
- [28.0.Model_Download_Helpers.ipynb](#)

Clinical Named Entity Recognition (NER)

Clinical Named Entity Recognition (NER)



The patient was prescribed 1 capsule of Advil for 5 days . He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely fro 3 months .

Color codes:FREQUENCY, DOSAGE, DURATION, DRUG, FORM, STRENGTH, **Posology NER**

No findings in urinary system , skin color is normal , brain CT and cranial checks are clear . Swollen fingers and eyes . Extensive stage small cell lung cancer . Chemotherapy with carboplatin and etoposide . Left scapular pain status post CT scan of the thorax .

Color codes:Organ, Organism_subdivision, Organism_substance, PathologicalFormation, Anatomical_system, **Anatomy NER**

A . Record date : 2093-01-13 , David Hale , M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

Color codes:STREET, DOCTOR, AGE, HOSPITAL, PATIENT, DATE, MEDICALRECORD, **PHI NER**

NLP in Healthcare

"Mother with a lung cancer, a patient is diagnosed as breast cancer in 1991 and then admitted to Mayo Clinic in Oct 2000, went under chemo for 6 months, discharged in April 2001 with a prescription of 2 mg metformin 3 times per day."

Named Entities

Mother with a lung cancer **ONCOLOGICAL** , a pregnant **PREGNANCY** patient is diagnosed as breast cancer **ONCOLOGICAL** in **1991 DATE** and then admitted **ADMISSION_DISCHARGE** to Mayo Clinic **CLINICAL_DEPT** in Oct **2000 DATE** , went under chemo **TREATMENT** for 6 months **DURATION** , discharged **ADMISSION_DISCHARGE** in **April 2001 DATE** with a prescription of **2 mg STRENGTH** metformin **DRUG_INGREDIENT** **3 times per day FREQUENCY** .

Clinical Named Entity Recognition (NER)

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

D

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

ner_clinical

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

ner_jsl

Pretrained Clinical NER Models

1 jsl_ner_wip_clinical	2 jsl_ner_wip_greedy_clinical	3 jsl_ner_wip_modifier_clinical	4 jsl_rd_ner_wip_greedy_clinical	5 ner_abbreviation_clinical
6 ner_abbreviation_emb_clinical_large	7 ner_abbreviation_emb_clinical_medium	8 ner_ade_clinical	9 ner_ade_clinicalbert	10 ner_ade_emb_clinical_large
11 ner_ade_emb_clinical_medium	12 ner_ade_healthcare	13 ner_anatomy	14 ner_anatomy_coarse	15 ner_anatomy_coarse_en
16 ner_anatomy_emb_clinical_large	17 ner_anatomy_emb_clinical_medium	18 ner_anatomy_en	19 ner_aspect_based_sentiment	20 ner_bacterial_species
21 ner_bacterial_species_emb_clinical_large	22 ner_bacterial_species_emb_clinical_medium	23 ner biomarker	24 ner_biomedical_bc2gm	25 ner_bionlp
26 ner_bionlp_en	27 ner_cancer_genetics	28 ner_cellular	29 ner_cellular_emb_clinical_large	30 ner_cellular_emb_clinical_medium
31 ner_cellular_en	32 ner_chemd_clinical	33 ner_chemicals	34 ner_chemicals_emb_clinical_large	35 ner_chemicals_emb_clinical_medium
36 ner_chemprot_clinical	37 ner_chexpert	38 ner_clinical	39 ner_clinical_en	40 ner_clinical_large
41 ner_clinical_large_en	42 ner_clinical_trials_abstracts	43 ner_covid_trials	44 ner_crft	45 ner_deid_augmented
46 ner_deid_enriched	47 ner_deid_generic_glove	48 ner_deid_large	49 ner_deid_large_emb_clinical_large	50 ner_deid_large_emb_clinical_medium
51 ner_deid_sd	52 ner_deid_sd_large	53 ner_deid_subentity_augmented	54 ner_deid_subentity_augmented_i2b2	55 ner_deid_subentity_glove
56 ner_deid_synthetic	57 ner_deidentify_dl	58 ner Demographic_extended_healthcare	59 ner_diseases	60 ner_diseases_en
61 ner_diseases_large	62 ner_drugprot_clinical	63 ner_drugs	64 ner_drugs_en	65 ner_drugs_greedy
66 ner_drugs_large	67 ner_drugs_large_en	68 ner_eu_clinical_case	69 ner_eu_clinical_condition	70 ner_events_admission_clinical
71 ner_events_clinical	72 ner_events_healthcare	73 ner_financial_contract	74 ner_genetic_variants	75 ner_healthcare
76 ner_healthcare_en	77 ner_human_phenotype_gene_clinical	78 ner_human_phenotype_go_clinical	79 ner_jsl	80 ner_jsl_emb_clinical_large
81 ner_jsl_emb_clinical_medium	82 ner_jsl_en	83 ner_jsl_enriched	84 ner_jsl_enriched_en	85 ner_jsl_greedy
86 ner_jsl_limited_80p_for Benchmarks	87 ner_jsl_slim	88 ner_living_species	89 ner_measurements_clinical	90 ner_medmentions_coarse
91 ner_nature_nero_clinical	92 ner_nihss	93 ner_oncology	94 ner_oncology_anatomy_general	95 ner_oncology_anatomy_general_healthcare
96 ner_oncology_anatomy_granular	97 ner_oncology biomarker	98 ner_oncology biomarker_healthcare	99 ner_oncology_demographics	100 ner_oncology_diagnosis
01 ner_oncology_emb_clinical_large	102 ner_oncology_emb_clinical_medium	103 ner_oncology_limited_80p_for Benchmarks	104 ner_oncology_posology	105 ner_oncology_response_to_treatment
06 ner_oncology_test	107 ner_oncology_therapy	108 ner_oncology_tnm	109 ner_oncology_unspecific_posology	110 ner_oncology_unspecific_posology_healthcare
11 ner_pathogen	112 ner_posology	113 ner_posology_emb_clinical_large	114 ner_posology_emb_clinical_medium	115 ner_posology_en
16 ner_posology_experimental	117 ner_posology_greedy	118 ner_posology_healthcare	119 ner_posology_large	120 ner_posology_small
21 ner_radiology	122 ner_radiology_wip_clinical	123 ner_risk_factors	124 ner_risk_factors_en	125 ner_sdoh
26 ner_sdoh_mentions	127 ner_supplement_clinical	128 ner_vop	129 ner_vop_anatomy	130 ner_vop_anatomy_emb_clinical_large
31 ner_vop_anatomy_emb_clinical_medium	132 ner_vop_clinical_dept	133 ner_vop_clinical_dept_emb_clinical_large	134 ner_vop_clinical_dept_emb_clinical_medium	135 ner_vop_demographic
36 ner_vop_demographic_emb_clinical_large	137 ner_vop_demographic_emb_clinical_medium	138 ner_vop_emb_clinical_large	139 ner_vop_emb_clinical_medium	140 ner_vop_problem
41 ner_vop_problem_emb_clinical_large	142 ner_vop_problem_emb_clinical_medium	143 ner_vop_problem_reduced	144 ner_vop_problem_reduced_emb_clinical_large	145 ner_vop_problem_reduced_emb_clinical_medium
46 ner_vop_temporal	147 ner_vop_temporal_emb_clinical_large_final	148 ner_vop_temporal_emb_clinical_medium	149 ner_vop_test	150 ner_vop_test_emb_clinical_large
51 ner_vop_test_emb_clinical_medium	152 ner_vop_treatment	153 ner_vop_treatment_emb_clinical_large	154 ner_vop_treatment_emb_clinical_medium	155 ner_vop_wip_emb_clinical_large
56 ner_vop_wip_emb_clinical_medium	157 nerdl_tumour_demo			

Pretrained Clinical NER Models

- Multilingual Clinical NER Models

index	model	index	model	index	model	index	model
1	ner_clinical	2	ner_clinical_bert	3	ner_clinical_trials_abstracts	4	ner_deid_generic
5	ner_deid_generic_bert	6	ner_deid_subentity	7	ner_deid_subentity_bert	8	ner_diag_proc
9	ner_eu_clinical_case	10	ner_eu_clinical_condition	11	ner_healthcare	12	ner_healthcare_slim
13	ner_living_species	14	ner_living_species_300	15	ner_living_species_bert	16	ner_living_species_roberta
17	ner_negation_uncertainty	18	ner_neoplasms	19	ner_pharmacology	20	ner_traffic

- BioBert NER Models

index	model	index	model	index	model	index	model
1	jsl_ner_wip_greedy_biobert	2	jsl_rd_ner_wip_greedy_biobert	3	ner_ade_biobert	4	ner_anatomy_biobert
5	ner_anatomy_coarse_biobert	6	ner_anatomy_coarse_biobert_en	7	ner_bionlp_biobert	8	ner_cellular_biobert
9	ner_chemprot_biobert	10	ner_clinical_biobert	11	ner_deid_biobert	12	ner_deid_enriched_biobert
13	ner_diseases_biobert	14	ner_events_biobert	15	ner_human_phenotype_gene_biobert	16	ner_human_phenotype_go_biobert
17	ner_jsl_biobert	18	ner_jsl_enriched_biobert	19	ner_jsl_greedy_biobert	20	ner_living_species_biobert
21	ner_posology_biobert	22	ner_posology_large_biobert	23	ner_profiling_biobert	24	ner_risk_factors_biobert

- BertForTokenClassification Clinical NER models

index	model	index	model	index	model	index	model
1	bert_token_classifier_ade_tweet_binary	2	bert_token_classifier_drug_development_trials	3	bert_token_classifier_ner_ade	4	bert_token_classifier_ner_ade_binary
5	bert_token_classifier_ner_anatem	6	bert_token_classifier_ner_anatomy	7	bert_token_classifier_ner_bacteria	8	bert_token_classifier_ner_bc2gm_gene
9	bert_token_classifier_ner_bc4chemd_chemicals	10	bert_token_classifier_ner_bc5cdr_chemicals	11	bert_token_classifier_ner_bc5cdr_disease	12	bert_token_classifier_ner_bionlp
13	bert_token_classifier_ner_cellular	14	bert_token_classifier_ner_chemicals	15	bert_token_classifier_ner_chemprot	16	bert_token_classifier_ner_clinical
17	bert_token_classifier_ner_clinical_trials_abstracts	18	bert_token_classifier_ner_deid	19	bert_token_classifier_ner_drugs	20	bert_token_classifier_ner_jnlpba_cellular
21	bert_token_classifier_ner_jsl	22	bert_token_classifier_ner_jsl_slim	23	bert_token_classifier_ner_linnaeus_species	24	bert_token_classifier_ner_living_species
25	bert_token_classifier_ner_ncbi_disease	26	bert_token_classifier_ner_pathogen	27	bert_token_classifier_ner_species	28	bert_token_classifier_ner_supplement

Approach	embeddings	# of models
BiLSTM-CNN-Char	Clinical (glove)	200+
BiLSTM-CNN-Char	Biobert	20+
Bert for Token Cls.	Biobert	30+
Total		250+

NER JSL

Let's show an example of `ner_js1` model that has about 80 clinical entity labels by changing just only the model name.

Entities

Injury_or_Poisoning	Direction	Test	Admission_Discharge	Death_Entity
Relationship_Status	Duration	Respiration	Hyperlipidemia	Birth_Entity
Age	Labour_Delivery	Family_History_Header	BMI	Temperature
Alcohol	Kidney_Disease	Oncological	Medical_History_Header	Cerebrovascular_Disease
Oxygen_Therapy	O2_Saturation	Psychological_Condition	Heart_Disease	Employment
Obesity	Disease_Syndrome_Disorder	Pregnancy	ImagingFindings	Procedure
Medical_Device	Race_Ethnicity	Section_Header	Symptom	Treatment
Substance	Route	Drug_Ingredient	Blood_Pressure	Diet
External_body_part_or_region	LDL	VS_Finding	Allergen	EKG_Findings
Imaging_Technique	Triglycerides	RelativeTime	Gender	Pulse
Social_History_Header	Substance_Quantity	Diabetes	Modifier	Internal_organ_or_component
Clinical_Dept	Form	Drug_BrandName	Strength	Fetus_NewBorn
RelativeDate	Height	Test_Result	Sexually_Active_or_Sexual_Orientation	Frequency
Time	Weight	Vaccine	Vital_Signs_Header	Communicable_Disease
Dosage	Overweight	Hypertension	HDL	Total_Cholesterol
Smoking	Date			

Attention (aka Bert) is all you need ?

ner model	embeddings_clinical (BLSTM-CNN-Char)		biobert (BLSTM-CNN-Char)		BertForTokenClassification (SOTA)	
	micro	macro	micro	macro	micro	macro
ner_jsl	0.878	0.814	0.862	0.711	0.88	0.71
ner_jsl_slim	0.87	0.766	0.86	0.778	0.89	0.75
ner_deid	0.94	0.77	0.93	0.77	0.75	0.63
ner_drug	0.964	0.964	0.912	0.911	1	0.98
ner_ade	0.84	0.807	0.839	0.819	0.89	0.84

* On average, the GLoVe embeddings are 30% faster during training compared to BERT embeddings, and more than 5x faster during inference, while being on-par in terms of F1 score.

Clinical Named Entities – Spark NLP vs Others

Spark NLP

Google	Azure	AWS
PROBLEM	DIAGNOSIS SYMPTOM_OR_SIGN ALLERGEN	MEDICAL_CONDITION_DX_NA MEDICAL_CONDITION_SIGN MEDICAL_CONDITION_SYMPTOM
PROCEDURE	TREATMENT_NAME	PROCEDURE_NAME TREATMENT_NAME
MEDICINE	MEDICATION_CLASS MEDICATION_NAME	MEDICATION_BRAND_NAME MEDICATION_GENERIC_NAME
ANATOMICAL_STRUCTURE	BODY_STRUCTURE	SYSTEM_ORGAN_SITE
LABORATORY_DATA BODY_MEASUREMENT	EXAMINATION_NAME	TEST_NAME
SEVERITY	CONDITION_QUALIFIER CONDITION_SCALE	MEDICAL_CONDITION_ACUITY
MED_DOSE MED_TOTALDOSE MED_STRENGTH MED_UNIT	DOSAGE	MEDICATION_DOSAGE MEDICATION_STRENGTH MEDICATION_RATE
MED_FREQUENCY	FREQUENCY	MEDICATION_FREQUENCY
MED_FORM	MEDICATION_FORM	MEDICATION_FORM
MED_ROUTE	MEDICATION_ROUTE	MEDICATION_ROUTE_OR_MODE
MED_DURATION	TIME	MEDICATION_DURATION
LAB_VALUE MED_VALUE	MEASUREMENT_VALUE	TEST_VALUE
LAB_UNIT BM_UNIT	MEASUREMENT_UNIT	TEST_UNIT



Symptom, Disease_Syndrome_Disorder, Heart_Disease, VS_Finding, Communicable_Disease, Hypertension, Diabetes, Kidney_Disease, Cerebrovascular_Disease, Injury_or_Poisoning, Psychological_Condition, Total_Cholesterol, Hyperlipidemia, Obesity, Oncological, Pregnancy, EKG_Findings, Death_Entity, ImagingFindings, Female_Reproductive_Status, Fetus_NewBorn, Pregnancy_Delivery_Puerperium, Overweight, Puerperium



Test, Test_Result, Treatment, Pulse, Imaging_Technique, Labour_Delivery, Temperature, Blood_Pressure, Oxygen_Therapy, Weight, LDL, O2_Saturation, BMI, Vaccine, Respiration, Triglycerides

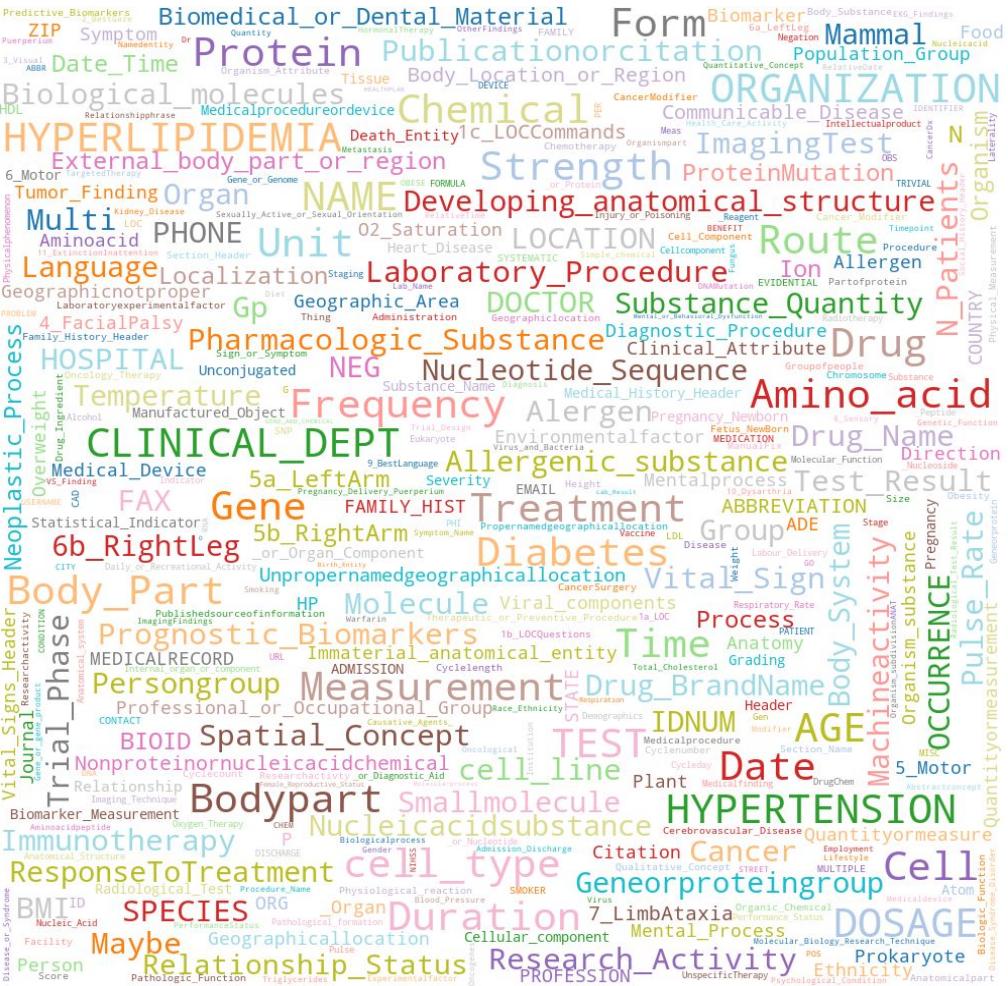
Mother with a lung cancer ONCOLOGICAL , a pregnant PREGNANCY patient is diagnosed as breast cancer ONCOLOGICAL in 1991 DATE and then admitted ADMISSION_DISCHARGE to Mayo Clinic CLINICAL_DEPT in Oct 2000 DATE , went under chemo TREATMENT for 6 months DURATION , discharged ADMISSION_DISCHARGE in April 2001 DATE with a prescription of 2 mg STRENGTH metformin DRUG_INGREDIENT 3 times per day FREQUENCY .

Clinical Named Entities

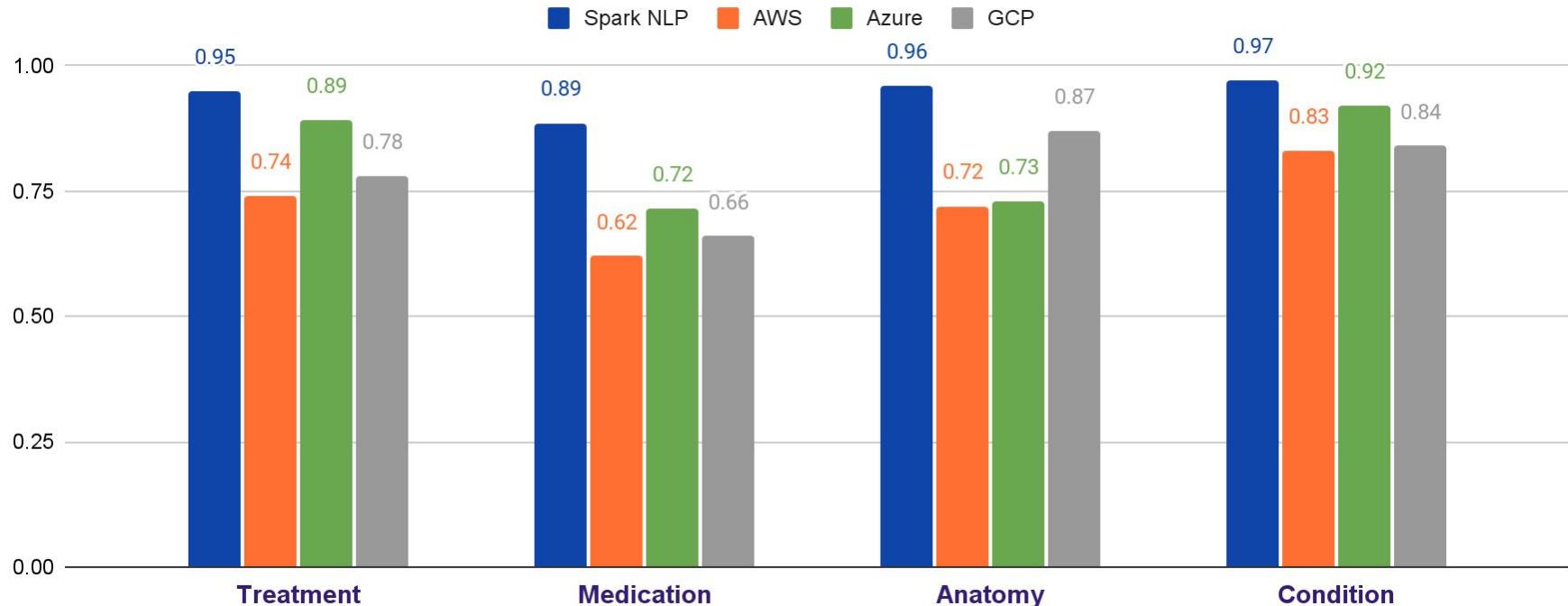
Spark NLP vs Others

Google	Azure	AWS
PROBLEM	DIAGNOSIS SYMPTOM_OR_SIGN ALLERGEN	MEDICAL_CONDITION_DX_NA MEDICAL_CONDITION_SIGN MEDICAL_CONDITION_SYMPTOM
PROCEDURE	TREATMENT_NAME	PROCEDURE_NAME TREATMENT_NAME
MEDICINE	MEDICATION_CLASS MEDICATION_NAME	MEDICATION_BRAND_NAME MEDICATION_GENERIC_NAME
ANATOMICAL_STRUCTURE	BODY_STRUCTURE	SYSTEM_ORGAN_SITE
LABORATORY_DATA BODY_MEASUREMENT	EXAMINATION_NAME	TEST_NAME
SEVERITY	CONDITION_QUALIFIER CONDITION_SCALE	MEDICAL_CONDITION_ACUITY
MED_DOSE MED_TOTALDOSE MED_STRENGTH MED_UNIT	DOSAGE	MEDICATION_DOSAGE MEDICATION_STRENGTH MEDICATION_RATE
MED_FREQUENCY	FREQUENCY	MEDICATION_FREQUENCY
MED_FORM	MEDICATION_FORM	MEDICATION_FORM
MED_ROUTE	MEDICATION_ROUTE	MEDICATION_ROUTE_OR_MODE
MED_DURATION	TIME	MEDICATION_DURATION
LAB_VALUE MED_VALUE	MEASUREMENT_VALUE	TEST_VALUE
LAB_UNIT BM_UNIT	MEASUREMENT_UNIT	TEST_UNIT

400+ entities from 100+ models



NER Benchmarks



Spark NLP vs AWS vs GCP vs Academic

		Spark NLP	Competition Best	Last Best
Clinical Concept Extraction	2010 i2b2/VA	0.876	0.852	0.862
De-Identification	2014 n2c2	0.961	0.936	0.955
Medication Extraction	2018 n2c2	0.899	0.896	0.896

Entity	Sample	Spark NLP Clinical Models			AWS Medical Comprehend			GCP Healthcare API		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Problem	4891	0.726	0.585	0.648	0.539	0.478	0.507	0.850	0.516	0.642
Test	5903	0.782	0.662	0.717	0.594	0.703	0.644	0.576	0.461	0.512
Drug	10284	0.946	0.882	0.913	0.815	0.910	0.860	0.962	0.885	0.922
Avg. F1				0.759			0.670			0.692

Biomedical Named Entity Recognition

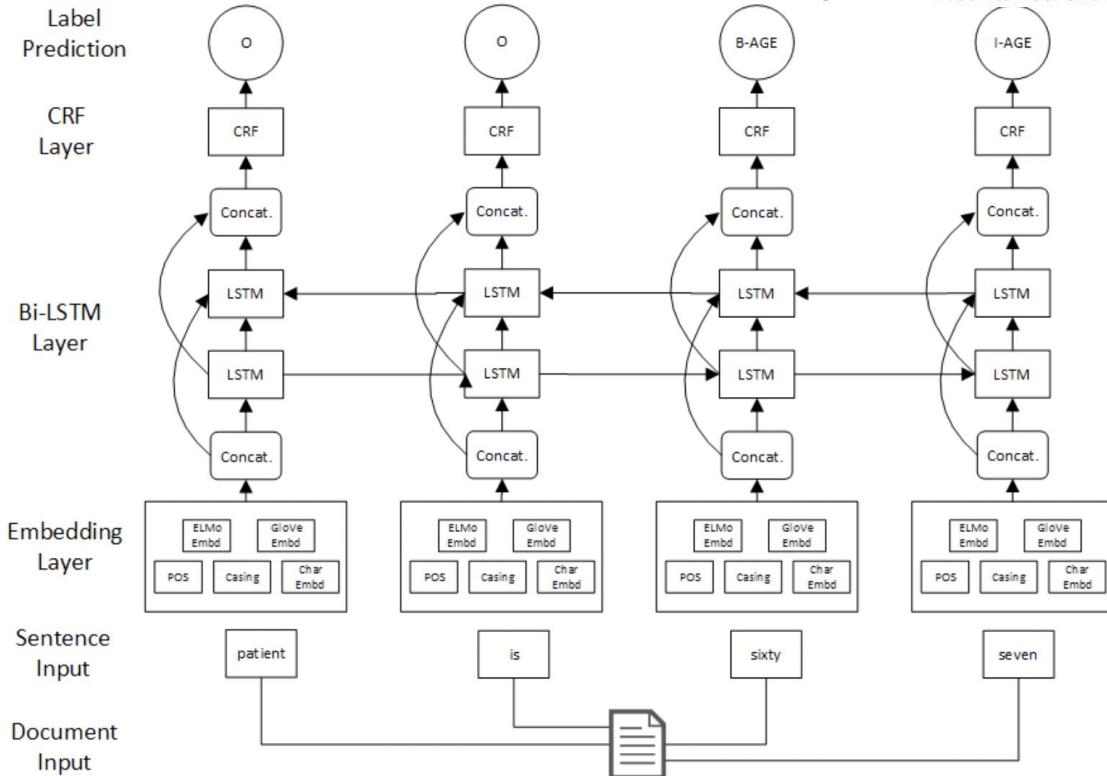
Spark NLP vs Spacy vs Stanza

Dataset	Entities	Spark - Biomedical	Spark - GloVe 6B	Stanza	SciSpacy
NBCI-Disease	Disease	89.13	87.19	87.49	81.65
BC5CDR	Chemical, Disease	89.73	88.32	88.08	83.92
BC4CHEMD	Chemical	93.72	92.32	89.65	84.55
Linnaeus	Species	86.26	85.51	88.27	81.74
Species800	Species	80.91	79.22	76.35	74.06
JNLPBA	5 types in cellular	81.29	79.78	76.09	73.21
AnatEM	Anatomy	89.13	87.74	88.18	84.14
BioNLP13-CG	16 types in Cancer Genetics	85.58	84.3	84.34	77.6

Benchmarks on BioMedical NER Datasets

NER Architecture

Char-CNN-BiLSTM



Tags	Description
B-PER	The beginning of a Person's name
I-PER	Part of a person's name
B-LOC	The beginning of a Location name
I-LOC	Part of a Location name
B-ORG	The beginning of a Organization name
I-ORG	Part of a Organization name
O	Not named-entity

BIO schema

John	B-PER
Smith	I-PER
lives	0
in	0
New	B-LOC
York	I-LOC

John Smith \Rightarrow PERSON
 New York \Rightarrow LOCATION

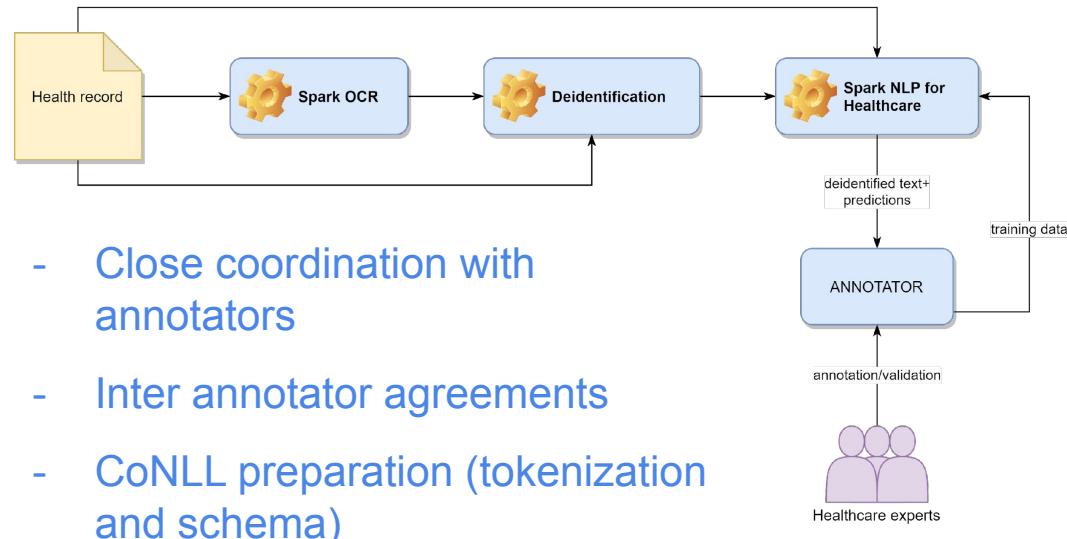
word	POS_tag	chunk_tag	NER_tag
She	PRP	O	B-person
presented	VBD	B-VP	O
with	IN	B-VP	O
left	JJ	B-NP	B-problem
upper	JJ	I-NP	I-problem
quadrant	NN	I-NP	I-problem
pain	NN	I-NP	I-problem
as	RB	O	O
well	RB	O	O
as	IN	B-VP	O
nausea	NN	B-NP	B-problem

John	B-PER
Smith	I-PER
lives	O
in	O
New	B-LOC
York	I-LOC

John Smith \Rightarrow PERSON
 New York \Rightarrow LOCATION

word	POS_tag	chunk_tag	NER_tag
She	PRP	O	B-person
presented	VBD	B-VP	O
with	IN	B-VP	O
left	JJ	B-NP	B-problem
upper	JJ	I-NP	I-problem
quadrant	NN	I-NP	I-problem
pain	NN	I-NP	I-problem
as	RB	O	O
well	RB	O	O
as	IN	B-VP	O
nausea	NN	B-NP	B-problem

NER in Healthcare



She returns today for ongoing evaluation of her EGFR mutated, stage 4 lung cancer with metastasis to her L2 vertebrae and her lungs bilaterally.

Bone negative for metastatic disease.

Patient denies any family history of cancer.

NER-DL in Spark NLP

Char-CNN-BiLSTM

	F1 : Tokens	F2 : Casing	F3 : POS	F4 : Char CNN	Labels
The					O
company					O
XYZ					Company
Private					Company
Limited					Company
works					O
in					O
the					O
health					Activity
sector					Activity
in					O
Europe					Location

Spark NLP for Healthcare

Pipeline Components

Spark NLP for Healthcare - Pipeline Components

```
from pyspark.ml import Pipeline

documentAssembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")

sentenceDetector = SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentences")

tokenizer = Tokenizer() \
    .setInputCols(["sentences"]) \
    .setOutputCol("token")

normalizer = Normalizer()\
    .setInputCols(["token"])\
    .setOutputCol("normal")

word_embeddings=WordEmbeddingsModel.pretrained()\
    .setInputCols(["document", "normal"])\ \
    .setOutputCol("embeddings")

nlpPipeline = Pipeline(stages=[\
    documentAssembler, \
    sentenceDetector, \
    tokenizer, \
    normalizer, \
    word_embeddings, \
])

nlpPipeline.fit(df).transform(df)
```

Clinical Entity Recognition	Clinical Entity Linking	Assertion Status	Relation Extraction
40 units DOSAGE of insulin glargin DRUG at night FREQUENCY	Suspect diabetes SNOMED-CT: 473127005 Lisinopril 10 MG RxNorm: 316151 Hyponatremia ICD-10: E87.1	Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY	AFTER Admitted Occurrence for Symptom due to Treatment CAUSED BY

Algorithms		Content	
Extract Knowledge	De-identify text	Medical Transformers	Linked Medical Terminologies
<ul style="list-style-type: none"> Entity Linker Entity Disambiguator Document Classifier Contextual Parser 	<ul style="list-style-type: none"> Structured Data Unstructured Text Obfuscator Generalizer 	JSL-BERT-Clinical BioBERT ClinicalBERT GloVe-Med GloVe-ICD-O BlueBERT	SNOMED-CT CPT UMLS ICD-10-CM RxNorm HPO ICD-10-PCS ICD-O LOINC
Split Text	Clean Medical Text	1000 Pretrained Models	
<ul style="list-style-type: none"> Sentence Detector Deep Sentence Detector Tokenizer nGram Generator 	<ul style="list-style-type: none"> Spell Checking Spell Correction Normalizer Stopword Cleaner 	Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	
Clinical Grammar	Find in Text	Anatomy: Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical	
<ul style="list-style-type: none"> Stemmer Lemmatizer Part of Speech Tagger Dependency Parser 	<ul style="list-style-type: none"> Text Matcher Regex Matcher Date Matcher Chunker 	Drugs: Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects	
Trainable & Tunable 		Scalable to a Cluster 	
Fast Inference 		Hardware Optimized 	
Community 			

DocumentAssembler

- ✓ Prepares data into a format that is processable by Spark NLP

"A 28-year-old female with a history
of gestational diabetes mellitus
diagnosed eight years prior to
presentation and subsequent type two
diabetes mellitus"

text

DocumentAssembler



	document	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus	0	153

document

```
documentAssembler = DocumentAssembler()\\"  
.setInputCol("text")\"  
.setOutputCol("document")
```

SentenceDetectorDLModel

- ✓ Detects sentence boundaries using a deep learning approach.
- ✓ `sentence_detector_dl_healthcare` is quite successful for clinical documents

document begin end

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m² , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dL , bicarbonate 18 mmol/L , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . Her initial serum glucose was 188 mg/dL . After a meal , her glucose was 188 mg/dL . Her serum glucose was 123 mg/dL after a meal . Serum glucose was 188 mg/dL . Her serum lipase was 52 U/L . The β -hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L – the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again . The patient was treated with an insulin drip for euDKA and HTG with a reduction in the anion gap to 13 and triglycerides to 1400 mg/dL , within 24 hours . Her euDKA was thought to be precipitated by her respiratory tract infection in the setting of SGLT2 inhibitor use . The patient was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely . She had close follow-up with endocrinology post discharge .

0 2515



		sentence	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m ² , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting .		0	433
1	Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection .		435	557
2	She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG .		559	662
3	She had been on dapagliflozin for six months at the time of presentation .		664	737
4	Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity .		739	911
5	Pertinent laboratory findings on admission were : serum glucose 111 mg/dL , bicarbonate 18 mmol/L , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 .		913	1162
6	Serum lipase was normal at 43 U/L .		1164	1198
7	Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia .		1200	1303

document

sentence

```
sentenceDetector = SentenceDetectorDLModel\  
.pretrained("sentence_detector_dl_healthcare", "en","clinical/models")\  
.setInputCols(["document"])\  
.setOutputCol("sentence")
```

Tokenizer / RegexTokenizer

- ✓ Splits words in a relevant format for NLP
- ✓ RegexTokenizer splits text by a regex pattern.
- ✓ Input for most of the DL models

	sentence	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m ² , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting .	0	433
1	Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection .	435	557
2	She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG .	559	662
3	She had been on dapagliflozin for six months at the time of presentation .	664	737
4	Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity .	739	911
5	Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 .	913	1162
6	Serum lipase was normal at 43 U/L .	1164	1198
7	Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia .	1200	1303

sentence / document

```
tokenizer = Tokenizer()\n    .setInputCols(["sentence"])\n    .setOutputCol("token")
```



	token	begin	end	sentence
0	A	0	0	0
1	28-year-old	2	12	0
2	female	14	19	0
3	with	21	24	0
4	a	26	26	0
5	history	28	34	0
6	of	36	37	0
7	gestational	39	49	0
8	diabetes	51	58	0
9	mellitus	60	67	0
73	Two	435	437	1
74	weeks	439	443	1
75	prior	445	449	1
76	to	451	452	1
77	presentation	454	465	1
78	,	467	467	1
79	she	469	471	1
80	was	473	475	1
81	treated	477	483	1
82	with	485	488	1

token

```
regexTokenizer = RegexTokenizer() \\n    .setInputCols(["document"]) \\n    .setOutputCol("regexToken") \\n    .setToLowercase(True) \\n    .setPattern("\\s+")
```

WordEmbeddingsModel

- ✓ Pretrained models can be loaded maps tokens to vectors
- ✓ `embeddings_clinical` can be preferred in clinical documents

	token	begin	end	sentence
0	A	0	0	0
1	28-year-old	2	12	0
2	female	14	19	0
3	with	21	24	0
4	a	26	26	0
5	history	28	34	0
6	of	36	37	0
7	gestational	39	49	0
8	diabetes	51	58	0
9	mellitus	60	67	0

	token	begin	end	sentence
73	Two	435	437	1
74	weeks	439	443	1
75	prior	445	449	1
76	to	451	452	1
77	presentation	454	465	1
78	,	467	467	1
79	she	469	471	1
80	was	473	475	1
81	treated	477	483	1
82	with	485	488	1



```
+-----+  
| | embeddings |  
+-----+  
| [-0.570580005645754 0.44183000922203064 0.7010200023651123 -0.417... |  
| [-0.542639970779419 0.41475999355316161 0.0321999788284302 -0.4024... |  
| [-0.2708599865436554 0.04400600120425224 -0.020260000601410866 -0... |  
| [0.6191999912261963 0.14650000631809235 -0.08592499792575836 -0.2... |  
| [-0.3397899866104126 0.20940999686717987 0.46347999572753906 -0.6... |  
+-----+
```

```
word_embeddings =  
WordEmbeddingsModel.pretrained("embeddings_clinical","en","clinical/models")\  
.setInputCols(["sentence","token"])\\  
.setOutputCol("embeddings")\
```

embeddings

MedicalNerModel/NerModel

- ✓ Assigns an NER label to each token.
- ✓ The default model is `ner_clinical`, if no name is provided.

	token	begin	end
0	A	0	0
1	28-year-old	2	12
2	female	14	19
3	with	21	24
4	a	26	26
5	history	28	34
6	of	36	37
7	gestational	39	49
8	diabetes	51	58
9	mellitus	60	67



	token	begin	end	label	confidence
0	A	0	0	O	0.9998
1	28-year-old	2	12	O	0.9991
2	female	14	19	O	0.996
3	with	21	24	O	0.9998
4	a	26	26	O	0.9989
5	history	28	34	O	0.8485
6	of	36	37	O	0.9613
7	gestational	39	49	B-PROBLEM	0.8199
8	diabetes	51	58	I-PROBLEM	0.9809
9	mellitus	60	67	I-PROBLEM	0.9607

token

named_entity

```
clinical_ner = MedicalNerModel.pretrained("ner_clinical_large","en","clinical/models")\
    .setInputCols(["sentence", "token", "embeddings"])\
    .setOutputCol("ner")\
    .setLabelCasing("upper")
```

BertForTokenClassifier

- ✓ Loads Bert Models with a token classification head for Named-Entity-Recognition (NER) tasks.

“A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus”

	begin	end	token	label	metadata
0	0	0	A	O	{'sentence': '0', 'Some(O)': '0.999487', 'Some(B-TEST)': '4.54522E-5', 'Some(I-PROBLEM...}
1	2	12	28-year-old	O	{'sentence': '0', 'Some(O)': '0.99780965', 'Some(B-TEST)': '2.268446E-4', 'Some(I-PROB...}
2	14	19	female	O	{'sentence': '0', 'Some(O)': '0.99946886', 'Some(B-TEST)': '2.7098458E-5', 'Some(I-PRO...}
3	21	24	with	O	{'sentence': '0', 'Some(O)': '0.99975836', 'Some(B-TEST)': '8.164151E-6', 'Some(I-PROB...}
4	26	26	a	O	{'sentence': '0', 'Some(O)': '0.99976975', 'Some(B-TEST)': '1.953544E-5', 'Some(I-PROB...}
5	28	34	history	O	{'sentence': '0', 'Some(O)': '0.99965894', 'Some(B-TEST)': '1.4398003E-5', 'Some(I-PRO...}
6	36	37	of	O	{'sentence': '0', 'Some(O)': '0.9997003', 'Some(B-TEST)': '6.0822117E-6', 'Some(I-PROB...}
7	39	49	gestational	B-PROBLEM	{'sentence': '0', 'Some(O)': '1.2003799E-5', 'Some(B-TEST)': '1.2508507E-5', 'Some(I-P...}
8	51	58	diabetes	I-PROBLEM	{'sentence': '0', 'Some(O)': '6.9626126E-6', 'Some(B-TEST)': '3.292549E-6', 'Some(I-PR...}
9	60	67	mellitus	I-PROBLEM	{'sentence': '0', 'Some(O)': '1.0308599E-5', 'Some(B-TEST)': '2.3517548E-6', 'Some(I-P...}
10	69	77	diagnosed	O	{'sentence': '0', 'Some(O)': '0.99924546', 'Some(B-TEST)': '1.260147E-5', 'Some(I-PROB...}

```
tokenClassifier = MedicalBertForTokenClassifier.pretrained("bert_token_classifier_ner_clinical", "en", "clinical/models")\
    .setInputCols("token", "sentence")\
    .setOutputCol("ner")\
    .setCaseSensitive(True)
```

NerConverter / NerConverterInternal

- ✓ Converts a IOB or IOB2 representation of NER to a user-friendly one, by associating the tokens of recognized entities and their label

	label	begin	end	words	confidence
0	O	27	27	a	0.9989
1	O	29	35	history	0.8485
2	O	37	38	of	0.9613
3	B-PROBLEM	40	50	gestational	0.8199
4	I-PROBLEM	52	59	diabetes	0.9809
5	I-PROBLEM	61	68	mellitus	0.9607

named_entity



	chunks	begin	end	entities	confidence
0	gestational diabetes mellitus	40	68	PROBLEM	0.9205
1	subsequent type two diabetes mellitus	118	154	PROBLEM	0.75560004
2	T2DM	158	161	PROBLEM	0.9928
3	HTG-induced pancreatitis	187	210	PROBLEM	0.97975004
4	an acute hepatitis	264	281	PROBLEM	0.9519333
5	obesity	289	295	PROBLEM	0.997

chunk

```
from sparknlp_jsl.annotator import NerConverterInternal
from sparknlp.annotator import NerConverter

ner_converter = NerConverterInternal()\
    .setInputCols(["sentence", "token", "ner"])\\
    .setOutputCol("ner_chunk")\

# Optional
    .setWhiteList(["PROBLEM"]) # List of entities to process
    .setBlackList(["TREATMENT", "TEST"]) # List of entities to be excluded.
    .setThreshold(0.99) # Confidence threshold to filter the chunk entities.
    .setReplaceLabels({"Drug_BrandName": "Drug"}) # Replace label in only NerConverterInternal
```

ChunkMerger

- ✓ Merges chunk columns coming from two or more annotators.
- ✓ NER, ContextualParser, TextMatcher, or any other annotator producing chunks are supported.

"A 63 years old man presents to the hospital with a history of recurrent infections that include cellulitis, pneumonias, and upper respiratory tract infections."

The diagram illustrates the process of merging two NER chunks into one. On the left, there are two separate tables: 'deid_ner_chunk' and 'clinical_ner_chunk'. An arrow points from these two tables to a third table on the right, labeled 'merged_ner_chunk'. The 'deid_ner_chunk' table has columns: begin, end, chunk, entity. It contains one row with values 0, 3, 4, AGE. The 'clinical_ner_chunk' table has columns: begin, end, chunk, entity. It contains four rows with values: (0, 63, 82, recurrent infections PROBLEM), (1, 97, 106, cellulitis PROBLEM), (2, 109, 118, pneumonias PROBLEM), and (3, 126, 159, upper respiratory tract infections PROBLEM). The 'merged_ner_chunk' table has columns: begin, end, chunk, entity. It contains five rows, combining the data from both tables: (0, 3, 4, AGE), (1, 63, 82, recurrent infections PROBLEM), (2, 97, 106, cellulitis PROBLEM), (3, 109, 118, pneumonias PROBLEM), and (4, 126, 159, upper respiratory tract infections PROBLEM).

deid_ner_chunk				clinical_ner_chunk				merged_ner_chunk						
begin	end	chunk	entity	begin	end	chunk	entity	begin	end	chunk	entity			
0	3	4	63	AGE	0	63	82	recurrent infections	PROBLEM	0	3	4	63	AGE
					1	97	106	cellulitis	PROBLEM	1	63	82	recurrent infections	PROBLEM
					2	109	118	pneumonias	PROBLEM	2	97	106	cellulitis	PROBLEM
					3	126	159	upper respiratory tract infections	PROBLEM	3	109	118	pneumonias	PROBLEM
								4	126	159	upper respiratory tract infections	PROBLEM		

```
chunk_merger = ChunkMergeApproach()\
    .setInputCols('clinical_ner_chunk', "deid_ner_chunk")\
    .setOutputCol('merged_ner_chunk')\
#Optional
    .setMergeOverlapping()\ # Sets whether to merge overlapping matched chunks.
    .setFalsePositivesResource()\ # Sets file with false positive pairs
    .setReplaceDictResource()\ # Sets replace dictionary pairs
    .setChunkPrecedence()\ # Sets what is the precedence when two chunks have the same start and end indices.
    .setBlackList()\ # If defined, list of entities to ignore. The rest will be processed.
```

ChunkMapper/ DocMapper

- ✓ Maps entities with their correspondings which are based on pre-defined dictionary
 - ✓ Very useful in medical coding resolution.
 - ✓ Supports multi dictionaries in a single model

"The patient was female and patient of Dr. X. and she was given Dermovate, Aspagan"

	ner_chunk	begin	end
0	Dermovate	63	71
1	Aspagan	74	80



	ner_chunk	begin	end	mapping_result	relation	all_relations
0	Dermovate	63	71	lupus	treatment	discoid lupus erythematosus:::emepeines:::psori...
1	Aspagin	74	80	ankylosing spondylitis	treatment	arthralgia:::pain:::bursitis:::headache:::migr...

token

named_entity

```
chunkerMapper_treatment= ChunkMapperModel().pretrained("drug_action_treatment_mapper", "en", "clinical/models")\
    .setInputCols(["ner_chunk"])\\
    .setOutputCol("mappings")\
    .setRels(["treatment"])# should be one of the relations from the model
```

DrugNormalizer

- ✓ Transforms text to the format used in the RxNorm and SNOMED standards

	text		normalized_text
0	Sodium Chloride/Potassium Chloride 13bag		Sodium Chloride / Potassium Chloride 13 bag
1	interferon alfa-2b 10 million unit (1 ml) injec		interferon alfa - 2b 10000000 unt (1 ml) injection
2	aspirin 10 meq/ 5 ml oral sol		aspirin 2 meq/ml oral solution

```
drugNormalizer = DrugNormalizer()\
    .setInputCols(["document"])\
    .setOutputCol("document_normalized")
#Optional
.setPolicy() # all / abbreviations / dosages
```

ContextualParser

- ✓ Allows users to extract entities from a document based on pattern matching.

A  28 year old female with a history of gestational diabetes mellitus diagnosed 8 years ago.

AGE

3 years ago, he reported an episode of HTG-induced pancreatitis .

5

months old boy with repeated concussions.

AGE

```
age = {  
    "entity": "Age",  
    "ruleScope": "sentence",  
    "matchScope": "token",  
    "regex": "\d{1,3}",  
    "prefix": ["age of", "age"],  
    "suffix": ["-years-old", "years-old", "-year-old", "-months-old",  
              "-month-old", "-months-old", "-day-old", "years old",  
              "-days-old", "month old", "years", "year", "months", "old",  
              "days old", "year old"],  
    "contextLength": 25,  
    "contextException": ["ago"],  
    "exceptionDistance": 12}
```

```
age_contextual_parser = ContextualParserApproach()\\"  
    .setInputCols(["sentence", "token"]) \  
    .setOutputCol("chunk_age") \  
    .setJsonPath("age.json") \  
    .setCaseSensitive(False) \  
    .setPrefixAndSuffixMatch(False)\ \  
    .setShortestContextMatch(True)\ \  
    .setOptionalContextRules(False)
```

Deidentification

- ✓ Allows users to mask or replace PHI (Protected Health Information) entities.

sentence

```
Patient AIQING, 25 years-old (ssid : 321-55-3699), born in Beijing,  
was transferred to the The Johns Hopkins Hospital.
```



	deid_entity_label
	Patient <NAME>, <AGE> years-old (ssid : <ID>), born in <LOCATION>, was transferred to the The <LOCATION>
	deid_same_length
	Patient [***], ** years-old (ssid : [*****]), born in [***], was transferred to the The [*****]
	deid_fixed_length
	Patient *****; ***** years-old (ssid : *****), born in *****; ***** was transferred to the The *****
	obfuscated
	Patient Olene Ronde, 41 years-old (ssid : G074713), born in China, was transferred to the Avenue Hospital

```
deidentification = DeIdentification()\  
.setInputCols(["sentence", "token", "ner_chunk"])\\  
.setOutputCol("deidentified")\\  
.setMode("mask") # "obfuscate"  
# Optional  
.setReturnEntityMappings(True)\\  
.setMaskingPolicy("entity_labels") # "same_length_chars", "deid_fixed_length", "fixed_length_chars"  
.setObfuscateDate(False)  
.setObfuscateRefSource("file") # "file", "both"
```

Spell Checker

- ✓ A flexible, configurable and "re-usable by parts" model.
- ✓ Different correction candidates for each word - **word level**.
- ✓ The surrounding text of each word, i.e. its context - **sentence level**.
- ✓ The relative cost of different correction candidates according to the edit operations at the character level it requires - **subword level**.

"Alliow me tao introdouce myhelf, I am
a man of waelth und tiaste"



"Allow me to introduce myself, I am a
man of wealth and taste"

```
spellModel = ContextSpellCheckerModel\  
.pretrained('spellcheck_dl')\  
.setInputCols("token")\  
.setOutputCol("checked")\  
.setErrorThreshold(4.0) \  
.setTradeoff(6.0)
```

TextMatcher

- ✓ Matches exact phrases (by token) provided in a file against a Document.
- ✓ A text file of predefined phrases must be provided with `setEntities`

"A 28-year-old female with a history
of gestational diabetes mellitus and
coronary artery disease"



	begin	end	entity	entity
0	39	67	gestational diabetes mellitus	DISEASE
1	73	95	coronary artery disease	DISEASE

```
entityExtractor = TextMatcher() \
    .setInputCols(["document", "token"]) \
    .setEntities("phrases.txt", ReadAs.TEXT) \
    .setOutputCol("entity") \
    .setEntityValue("DISEASE") \
    .setCaseSensitive(False)
```

Phrases.txt

gestational diabetes mellitus
coronary artery disease
Lower respiratory infections

Normalizer

- ✓ Removes all dirty characters from text following a regex pattern and transforms words based on a provided dictionary

"#publiHUCA #Emergency Emergency department observation of patients with acute heart failure prior to hospital admission: impact on short-term prognosis "



token	token	normalized
0 #publiHUCA	0 #publiHUCA	publihuca
1 #Emergency	1 #Emergency	emergency
2 Emergency	2 Emergency	emergency
3 department	3 department	department
4 observation	4 observation	observation

```
normalizer = Normalizer() \
    .setInputCols(["token"]) \
    .setOutputCol("normalized") \
    .setLowercase(True) \ # lowercase tokens
    .setCleanupPatterns(["[^\\w\\d\\s]"]) # remove punctuations (keep alphanumeric chars)
    .setSlangDictionary(path) # txt file with delimited words to be transformed into something else
```

DocumentNormalizer

- ✓ Normalize documents once that they have been processed and indexed

```
<div id="theworldsgreatest" class='my-right my-hide-small my-wide toptext'  
style="font-family:'Segoe UI',Arial,sans-serif">  
    THE WORLD'S LARGEST WEB DEVELOPER SITE  
    <h1 style="font-size:300%;">THE WORLD'S LARGEST WEB DEVELOPER SITE</h1>  
    <p style="font-size:160%;">Lorem Ipsum is simply dummy text of the printing  
and typesetting industry. Lorem Ipsum has been the industry's standard dummy  
text ever since the 1500s, when an unknown printer took a galley of type and  
scrambled it to make a type specimen book.</p>  
</div> </div> "
```

document



"the world's largest web developer site the world's
largest web developer site lorem ipsum is simply dummy
text of the printing and typesetting industry. lorem
ipsum has been the industry's standard dummy text ever
since the 1500s, when an unknown printer took a galley of
type and scrambled it to make a type specimen book. "

document

```
documentNormalizer = DocumentNormalizer() \  
.setInputCols("document") \  
.setOutputCol("normalizedDocument") \  
.setAction("clean") \  
.setPatterns(cleanUpPatterns) \  
.setReplacement(" ") \  
.setPolicy("pretty_all") \  
.setLowercase(True)
```

RegexMatcher

- ✓ Uses a reference file to match a set of regular expressions and associate them with a provided identifier.

'The patient (123-45-7890) was
discharged from the hospital on 10th
January.'



	clinical_entities	begin	end	label
0	123-45-7890	13	23	SSN

```
regex_matcher = RegexMatcher()\n    .setInputCols('document')\n    .setStrategy("MATCH_ALL")\n    .setOutputCol("regex_matches")\n    .setExternalRules(path='./regex_rules.txt', delimiter=',')
```

```
regex_rules.txt\n-----\n\d{3}.\?\d{2}.\?\d{4}, SSN
```

NerOverwriter

- ✓ Overwrites or replaces the labels of entities of specified strings.

"Mr.Brown was described warfarin 5 mg"

token	ner_label
0 Mr.Brown	O
1 was	O
2 described	O
3 warfarin	B-TREATMENT
4 5	O
5 mg	O



token	ner_label	ner_overwritten
0 Mr.Brown	O	B-People
1 was	O	O
2 described	O	O
3 warfarin	B-TREATMENT	B-DRUG
4 5	O	O
5 mg	O	O

```
nerOverwriter = NerOverwriter() \
    .setInputCols(["ner"]) \
    .setOutputCol("ner_overwritten") \
    .setNerWords(["Mr.Brown"]) \
    .setNewNerEntity("B-People") \
    .setReplaceEntities({"B-TREATMENT" : "B-DRUG"})
```

BertSentenceEmbeddings

- ✓ Calculates BERT embeddings for sequence

"Patient has a headache for the last 2 weeks, needs to get a head CT, and appears anxious when she walks fast. No alopecia noted. She denies pain"

	sentence	sentence_embeddings
0	Patient has a headache for the last 2 weeks, n...	-0.08882622 -0.09253775 -0.025744002 -0.105720...

```
Classifier = BertForSequenceClassification.pretrained("bert_sequence_classifier_covid_sentiment", "en", "clinical/models")\
.setInputCols(["document", "token"])\\
.setOutputCol("class")
```

BertForSequenceClassifier

- ✓ Loads BERT Models with sequence classification/regression for multi-class document classification tasks.

		text	result
0		British Department of Health confirms first two cases of in UK	[neutral]
1	so my trip to visit my australian exchange student just got canceled bc of coronavirus. im heartbroken :([negative]
2	I wish everyone to be safe at home and stop pandemic		[positive]

```
Classifier = BertForSequenceClassification.pretrained("bert_sequence_classifier_covid_sentiment", "en", "clinical/models")\
    .setInputCols(["document", "token"])\\
    .setOutputCol("class")
```

AssertionDLModel

- ✓ A Deep Learning based approach is used to extract Assertion Status from extracted entities and text.

"Patient has a headache for the last 2 weeks, needs to get a head CT, and appears anxious when she walks fast. No alopecia noted. She denies pain"

	entities_ner_chunk	entities_ner_chunk_class	assertion
0	a headache	PROBLEM	present
0	a head CT	TEST	present
0	anxious	PROBLEM	present
0	alopecia	PROBLEM	absent
0	pain	PROBLEM	absent

```
clinical_assertion = AssertionDLModel.pretrained("assertion_dl", "en", "clinical/models") \
    .setInputCols(["sentence", "ner_chunk", "embeddings"]) \
    .setOutputCol("assertion")
```

Relation Extraction Model

- ✓ Extracts and classifies instances of relations between named entities.
- ✓ Relation pairs need to be defined with setRelationPairs, to specify between which entities the extraction should be done.

"I experienced fatigue, muscle cramps, anxiety, aggression and sadness after taking Lipitor but no more adverse after passing Zocor."



chunk1	entity1	entity2_begin	entity2_end	chunk2	entity2	relation	score
fatigue	ADE	82	88	Lipitor	DRUG	1	
fatigue	ADE	124	128	Zocor	DRUG	0	
muscle cramps	ADE	82	88	Lipitor	DRUG	1	
muscle cramps	ADE	124	128	Zocor	DRUG	0	
anxiety	ADE	82	88	Lipitor	DRUG	1	
anxiety	ADE	124	128	Zocor	DRUG	0	
aggression	ADE	82	88	Lipitor	DRUG	1	
aggression	ADE	124	128	Zocor	DRUG	0	
sadness	ADE	82	88	Lipitor	DRUG	1	
sadness	ADE	124	128	Zocor	DRUG	0	

```
reModel = RelationExtractionModel()\
    .pretrained("re_ade_clinical", "en", 'clinical/models')\
    .setInputCols(["embeddings", "pos_tags", "ner_chunks", "dependencies"])\
    .setOutputCol("relations")\
    .setMaxSyntacticDistance(10) \
    .setRelationPairs(["drug-ade, ade-drug"])\
    .setRelationPairsCaseSensitive(False)
```

Relation Extraction DL Model

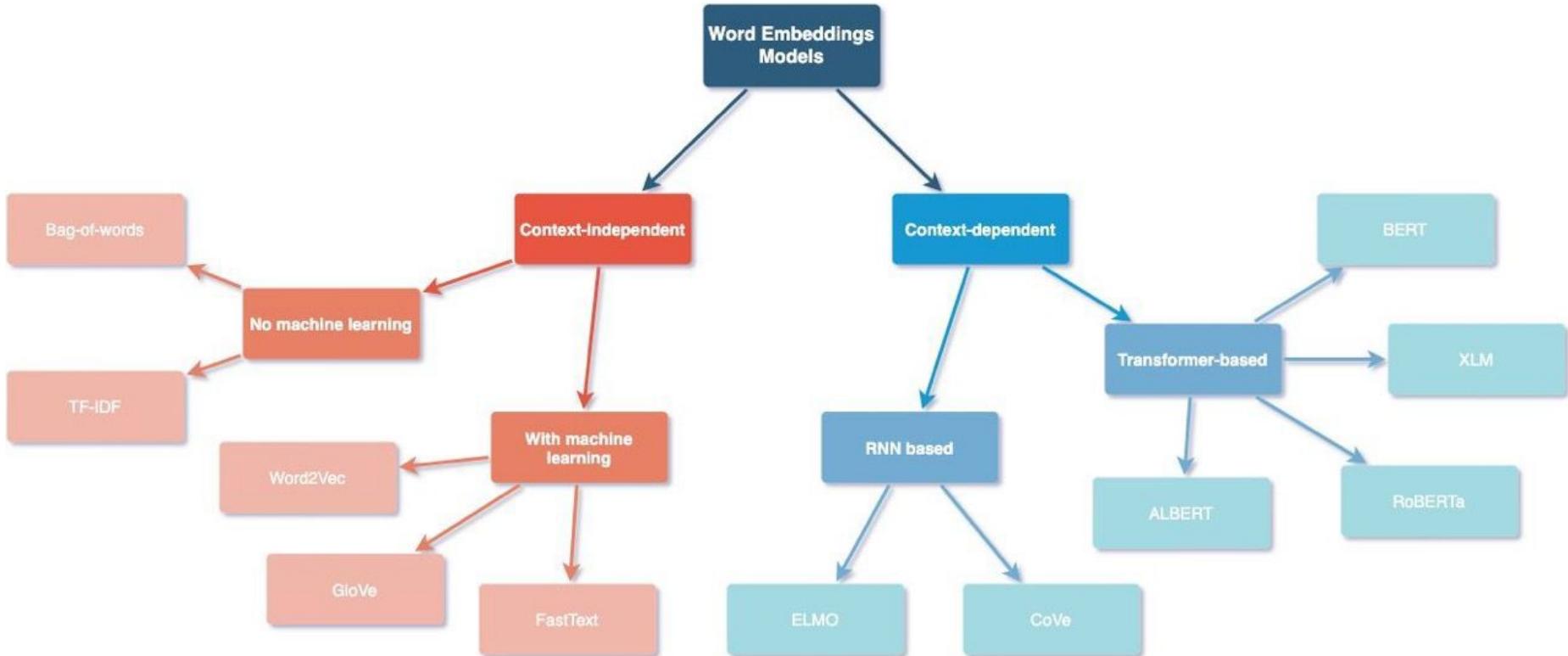
- ✓ A Deep Learning based approach is used to extract Assertion Status from extracted entities and text.

“The patient was prescribed 1 unit of Advil for 5 days after meals. The patient was also given 1 unit of Metformin daily. He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night.”

sentence	entity1_begin	entity1_end	chunk1	entity1	entity2_begin	entity2_end	chunk2	entity2	relation	
0	0	28	33	1 unit	DOSAGE	38	42	Advil	DRUG	DOSAGE-DRUG
1	0	38	42	Advil	DRUG	44	53	for 5 days	DURATION	DRUG-DURATION
2	1	96	101	1 unit	DOSAGE	106	114	Metformin	DRUG	DOSAGE-DRUG
3	1	106	114	Metformin	DRUG	116	120	daily	FREQUENCY	DRUG-FREQUENCY
4	2	190	197	40 units	DOSAGE	202	217	insulin glargine	DRUG	DOSAGE-DRUG
5	2	202	217	insulin glargine	DRUG	219	226	at night	FREQUENCY	DRUG-FREQUENCY

```
reModel = RelationExtractionModel().pretrained("posology_re")\
    .setInputCols(["embeddings", "pos_tags", "ner_chunks", "dependencies"])\
    .setOutputCol("relations")\
    .setMaxSyntacticDistance(4)
```

Clinical Word/Sentence Embeddings



Clinical Word/Sentence Embeddings

Clinical Glove
(200d)

PubMed + PMC

ICDO Glove
(200d)

PubMed + ICD10
UMLS + MIMIC III

Sent BERT

BioBert finetuned on
NLI and MedNLI

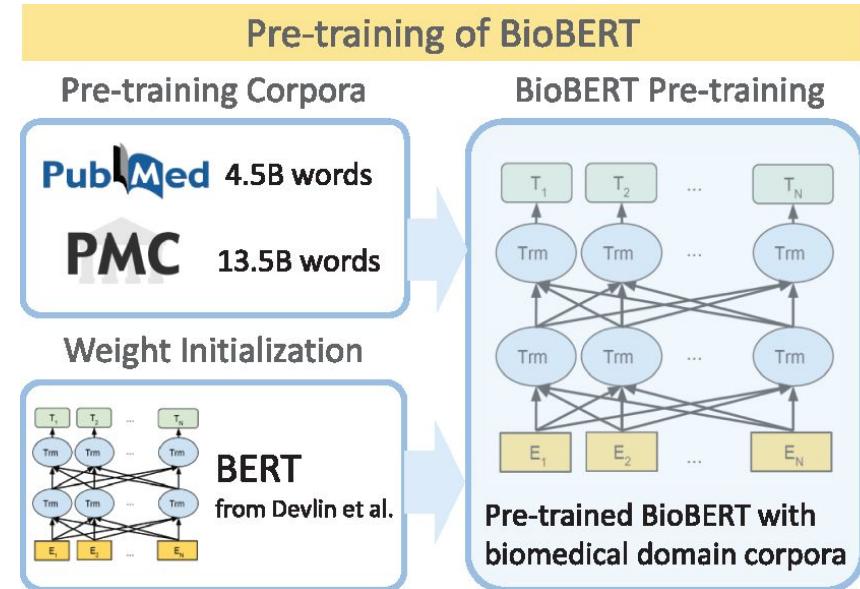
Bio/Clinical BERT

Fine tuned Pubmed + PMC + Discharge summaries



PubMed abstracts and PMC full-text articles

<https://www.nlm.nih.gov/bsd/difference.html>



Part - II

- ❖ Clinical Assertion Status Model

Assertion Status Detection

"Mother with a lung cancer, a patient is diagnosed as breast cancer in 1991 and then admitted to Mayo Clinic in Oct 2000, went under chemo for 6 months, discharged in April 2001 with a prescription of 2 mg metformin 3x per day. No sign of gynecological disorder but she suffers from acute cramps if she doesn't take her drug."

Chunk	Entity	Assertion
lung cancer	Oncological	Family
breast cancer	Oncological	Past
chemo	Treatment	Past
gynecological disorder	Disorder	Absent
acute cramps	Disorder	Conditional

```
clinical_assertion = AssertionDLModel\  
    .pretrained("assertion_dl", "en", "clinical/models")\  
    .setInputCols(["sentence", "ner_chunk", "embeddings"]) \  
    .setOutputCol("assertion")
```

Classify the assertions made on given medical concepts as being

- present,
- absent,
- possible,
- conditionally present under certain circumstances,
- hypothetically present at some future point, mentioned in the patient report but associated with someone else.

Assertion Status Detection

- The deep neural network architecture for assertion status detection in Spark NLP is based on a Bi-LSTM framework, and is a modified version of the architecture proposed by Federico Fancellu, Adam Lopez and Bonnie Webber ([Neural Networks For Negation Scope Detection](#)).
- In the proposed implementation, input units depend on the target tokens (a named entity) and the neighboring words that are explicitly encoded as a sequence using word embeddings.
- Similar to paper mentioned above, it is observed that 95% of the scope tokens (neighboring words) fall in a window of 9 tokens to the left and 15 to the right of the target tokens in the same dataset. Therefore, the same window size was implemented,
- following parameters were used: learning rate 0.0012, dropout 0.05, batch size 64 and a maximum sentence length 250.
- The model has been implemented within Spark NLP as an annotator called AssertionDLModel. After training 20 epoch and measuring accuracy on the official test set, this implementation exceeds the latest state-of-the-art accuracy benchmarks

Assertion Label	Spark NLP	Latest Best
Absent	0.944	0.937
Someone-else	0.904	0.869
Conditional	0.441	0.422
Hypothetical	0.862	0.890
Possible	0.680	0.630
Present	0.953	0.957
micro F1	0.939	0.934

Mother with a lung cancer,

AssertionDLModel

Clinical Assertion Models

model_name	Predicted Entities
1 assertion_dl	Present, Absent, Possible, conditional, hypothetical, associated_with_someone_else
2 assertion_dl_biobert	Present, Absent, Possible, conditional, hypothetical, associated_with_someone_else
3 assertion_dl_healthcare	Present, Absent, Possible, conditional, hypothetical, associated_with_someone_else
4 assertion_dl_large	Present, Absent, Possible, conditional, hypothetical, associated_with_someone_else
5 assertion_dl_radiology	Confirmed, Suspected, Negative
6 assertion_jsl	Present, Absent, Possible, Planned, Someoneelse, Past, Family, Hypothetical
7 assertion_jsl_large	present, absent, possible, planned, someoneelse, past, hypothetical
8 assertion_ml	Hypothetical, Present, Absent, Possible, Conditional, Associated_with_someone_else
9 assertion_dl_scope_L10R10	hypothetical, associated_with_someone_else, conditional, possible, absent, present
10 assertion_dl_biobert_scope_L10R10	hypothetical, associated_with_someone_else, conditional, possible, absent, present
11 assertion_jsl_augmented	Present, Absent, Possible, Planned, Past, Family, Hypothetical, SomeoneElse

Oncological Assertion Models

model_name	Predicted Entities
1 assertion_oncology_wip	Medical_History, Family_History, Possible, Hypothetical_Or_Absent
2 assertion_oncology_problem_wip	Present, Possible, Hypothetical, Absent, Family
3 assertion_oncology_treatment_wip	Present, Planned, Past, Hypothetical, Absent
3 assertion_oncology_treatment_wip	Present, Planned, Past, Hypothetical, Absent
4 assertion_oncology_response_to_treatment_wip	Present_Or_Past, Hypothetical_Or_Absent
5 assertion_oncology_test_binary_wip	Present_Or_Past, Hypothetical_Or_Absent
6 assertion_oncology_smoking_status_wip	Absent, Past, Present
7 assertion_oncology_family_history_wip	Family_History, Other
8 assertion_oncology_demographic_binary_wip	Patient, Someone_Else

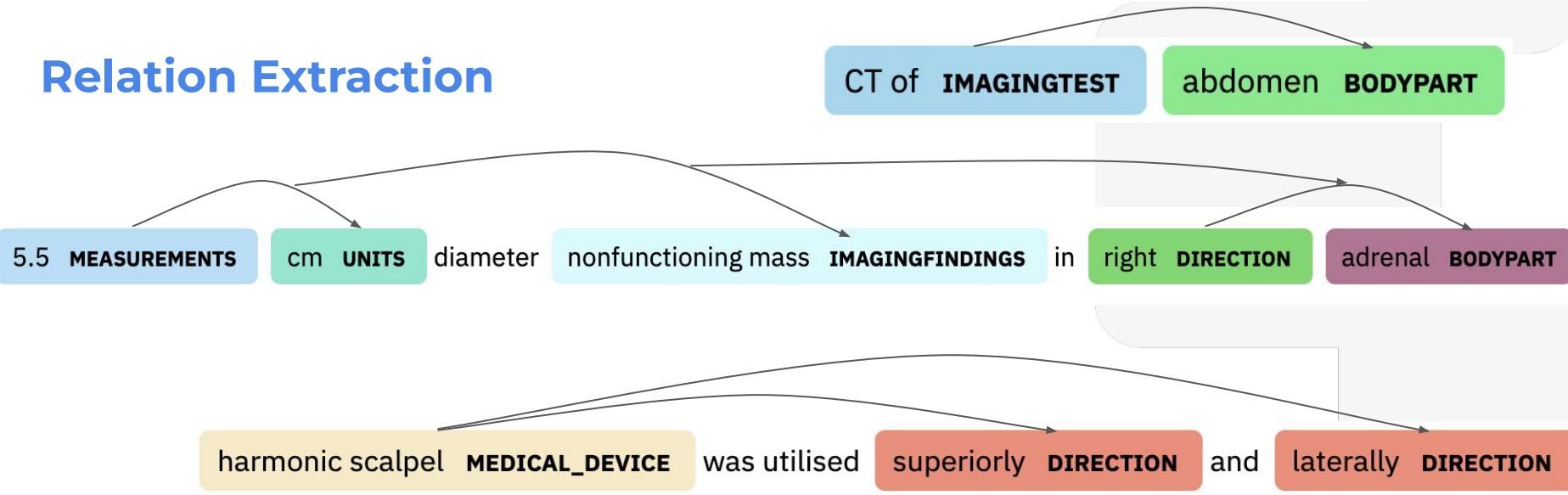
Part - III

- ❖ Clinical Relation Extraction Model
- ❖ Clinical Relation Extraction with Knowledge Graph (Neo4j)

Clinical Relation Extraction

"This is a 52-year-old inmate with a 5.5 cm diameter nonfunctioning mass in his right adrenal shown by CT of abdomen. During the umbilical hernia repair, the harmonic scalpel was utilised superiorly and laterally."

Relation Extraction



Clinical Relation Extraction

- Relation Extraction Models

index	model	index	model	index	model
1	generic_re	2	posology_re	3	re_ade_biobert
4	re_ade_clinical	5	re_ade_conversational	6	re_bodypart_directions
7	re_bodypart_problem	8	re_bodypart_proceduretest	9	re_clinical
10	re_date_clinical	11	re_drug_drug_interaction_clinical	12	re_drugprot_clinical
13	re_human_phenotype_gene_clinical	14	re_oncology_biomarker_result_wip	15	re_oncology_granular_wip
16	re_oncology_location_wip	17	re_oncology_size_wip	18	re_oncology_temporal_wip
19	re_oncology_test_result_wip	20	re_oncology_wip	21	re_temporal_events_clinical
22	re_temporal_events_enriched_clinical	23	re_test_problem_finding	24	re_test_result_date
25	re_zeroshot_biobert	26	redl_ade_biobert	27	redl_bodypart_direction_biobert
28	redl_bodypart_problem_biobert	29	redl_bodypart_procedure_test_biobert	30	redl_chemprot_biobert
31	redl_clinical_biobert	32	redl_date_clinical_biobert	33	redl_drug_drug_interaction_biobert
34	redl_drugprot_biobert	35	redl_human_phenotype_gene_biobert	36	redl_nihss_biobert
37	redl_oncology_biobert_wip	38	redl_oncology_biomarker_result_biobert_wip	39	redl_oncology_granular_biobert_wip
40	redl_oncology_location_biobert_wip	41	redl_oncology_size_biobert_wip	42	redl_oncology_temporal_biobert_wip
43	redl_oncology_test_result_biobert_wip	44	redl_temporal_events_biobert	45	

- Relation Extraction Pipelines

index	model	index	model	index	model
1	re_bodypart_directions_pipeline	4	re_human_phenotype_gene_clinical_pipeline	7	re_test_problem_finding_pipeline
2	re_bodypart_proceduretest_pipeline	5	re_temporal_events_clinical_pipeline	8	re_test_result_date_pipeline
3	re_date_clinical_pipeline	6	re_temporal_events_enriched_clinical_pipeline		

Relation	Recall	Precision	F1	SOTA
DRUG-ADE	0.66	1.00	0.80	0.76
DRUG-DOSAGE	0.89	1.00	0.94	0.91
DRUG-DURATION	0.75	1.00	0.85	0.92
DRUG-FORM	0.88	1.00	0.94	0.95*
DRUG-FREQUENCY	0.79	1.00	0.88	0.90
DRUG-REASON	0.60	1.00	0.75	0.70
DRUG-ROUTE	0.79	1.00	0.88	0.95*
DRUG-STRENGTH	0.95	1.00	0.98	0.97

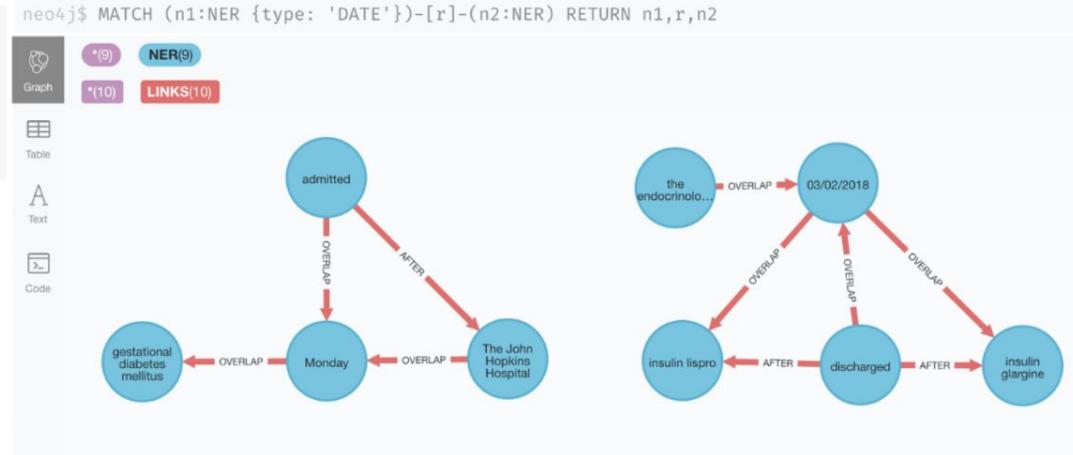
Relation	Recall	Precision	F1
OVERLAP	0.81	0.73	0.77
BEFORE	0.85	0.88	0.86
AFTER	0.38	0.46	0.43

Clinical Relation Extraction

She is admitted to The John Hopkins Hospital on Monday with a history of gestational diabetes mellitus diagnosed. She was seen by the endocrinology service and she was discharged on 03/02/2018 on 40 units of insulin glargin and 12 units of insulin lispro.

```
1 query = """
2 | MATCH (n1:NER {type: 'DATE'})-[r]-(n2:NER)
3 | RETURN n1.name AS date, r.relation AS relation, n2.name AS event
4 """
5
6 df = pd.DataFrame([dict(_) for _ in conn.query(query)])
7 df
```

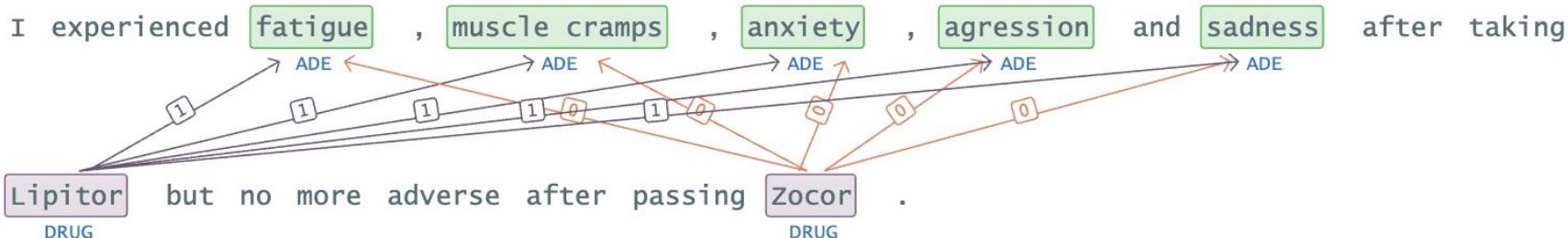
	date	relation	event
0	Monday	OVERLAP	gestational diabetes mellitus
1	Monday	OVERLAP	The John Hopkins Hospital
2	Monday	OVERLAP	admitted
3	03/02/2018	OVERLAP	insulin lispro
4	03/02/2018	OVERLAP	insulin glargin
5	03/02/2018	OVERLAP	discharged
6	03/02/2018	OVERLAP	the endocrinology service



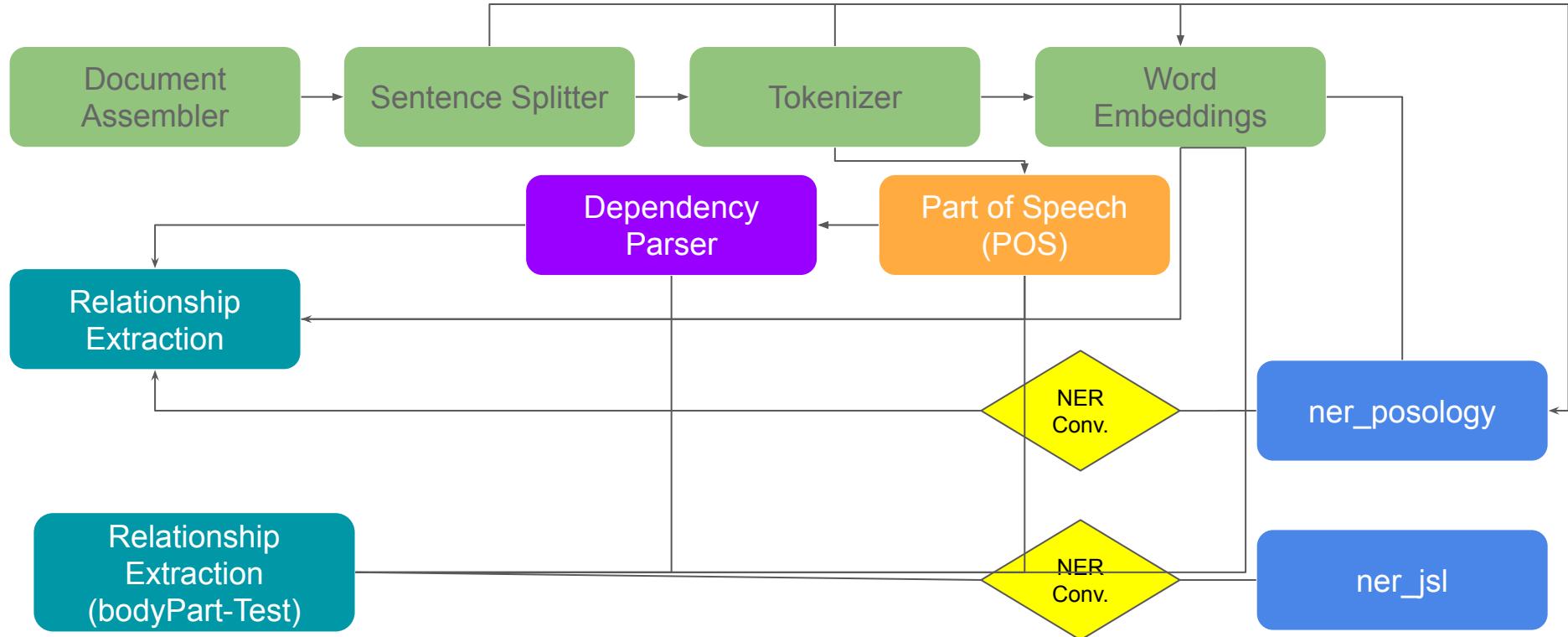
Clinical Relation Extraction

	relation	entity1	entity1_begin	entity1_end	chunk1	entity2	entity2_begin	entity2_end	chunk2	confidence
0	1	ADE	14	20	fatigue	DRUG	82	88	Lipitor	0.9996617
1	0	ADE	14	20	fatigue	DRUG	124	128	Zocor	0.9952187
2	1	ADE	23	35	muscle cramps	DRUG	82	88	Lipitor	0.9999827
3	0	ADE	23	35	muscle cramps	DRUG	124	128	Zocor	0.91462934
4	1	ADE	38	44	anxiety	DRUG	82	88	Lipitor	0.7636133
5	0	ADE	38	44	anxiety	DRUG	124	128	Zocor	0.9999691
6	1	ADE	47	55	agression	DRUG	82	88	Lipitor	0.99999833
7	0	ADE	47	55	agression	DRUG	124	128	Zocor	0.99781835
8	1	ADE	61	67	sadness	DRUG	82	88	Lipitor	1.0
9	0	ADE	61	67	sadness	DRUG	124	128	Zocor	0.9999572

I experienced fatigue, muscle cramps, anxiety, agression and sadness after taking Lipitor but no more adverse after passing Zocor.



Clinical Relation Extraction



Part - IV

- ❖ Clinical Entity Resolution
- ❖ Medical Risk Adjustment Score Calculation

Entity Resolution in Spark NLP for Healthcare

This is a 52-year-old AGE inmate with a 5.5 MEASUREMENTS cm UNITS diameter nonfunctioning mass SYMPTOM in his GENDER right DIRECTION adrenal BODYPART shown by CT of IMAGINGTEST abdomen BODYPART . During the umbilical hernia repair PROCEDURE , the harmonic scalpel MEDICAL_DEVICE was utilised superiorly DIRECTION and laterally DIRECTION .

Entity Resolution

ICD10CM, Snomed, RxNorm, CPT-4, ICD10CPS, RxCUI, ICDO

Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	76497	Unlisted computed tomography procedure
CT of abdomen	CPT-4	74150	Computed tomography, abdomen; without contrast material

weighted Sentence Chunk Embeddings (after 3.2.0)

Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	74150	Computed tomography, abdomen; without contrast material

Clinical Entity Resolution

ICD10CM

- sbiobertresolve_icd10cm_augmented
- sbiobertresolve_icd10pcs
- sbiobertresolve_icd10cm_augmented_billable_hcc
- sbiobertresolve_icd10cm
- sbiobertresolve_icd10cm_slim_normalized
- sbiobertresolve_icd10cm_slim_billable_hcc
- sbertrresolve_icd10cm_slim_billable_hcc_med
- sbiobertresolve_icd10cm_generalised

CPT

- sbiobertresolve_cpt
- sbiobertresolve_cpt_procedures_augmented
- sbiobertresolve_cpt_augmented
- sbiobertresolve_cpt_procedures_measurements_augmented

Snomed

- sbiobertresolve_snomed_auxConcepts_int
- sbiobertresolve_snomed_findings
- sbiobertresolve_snomed_findings_int
- sbiobertresolve_snomed_auxConcepts
- sbertrsolve_snomed_bodyStructure_med
- sbiobertresolve_snomed_bodyStructure
- sbiobertresolve_snomed_findings_aux_concepts
- sbertrsolve_snomed_conditions

RxNorm

- sbiobertresolve_rxnorm
- demo_sbiobertresolve_rxnorm
- sbiobertresolve_rxnorm_dispo
- sbiobertresolve_rxnorm_disposition
- sbertrsolve_rxnorm_disposition
- sbiobertresolve_rxnorm_ndc

LOINC

- sbluebertresolve_loinc
- sbiobertresolve_loinc

and more ...

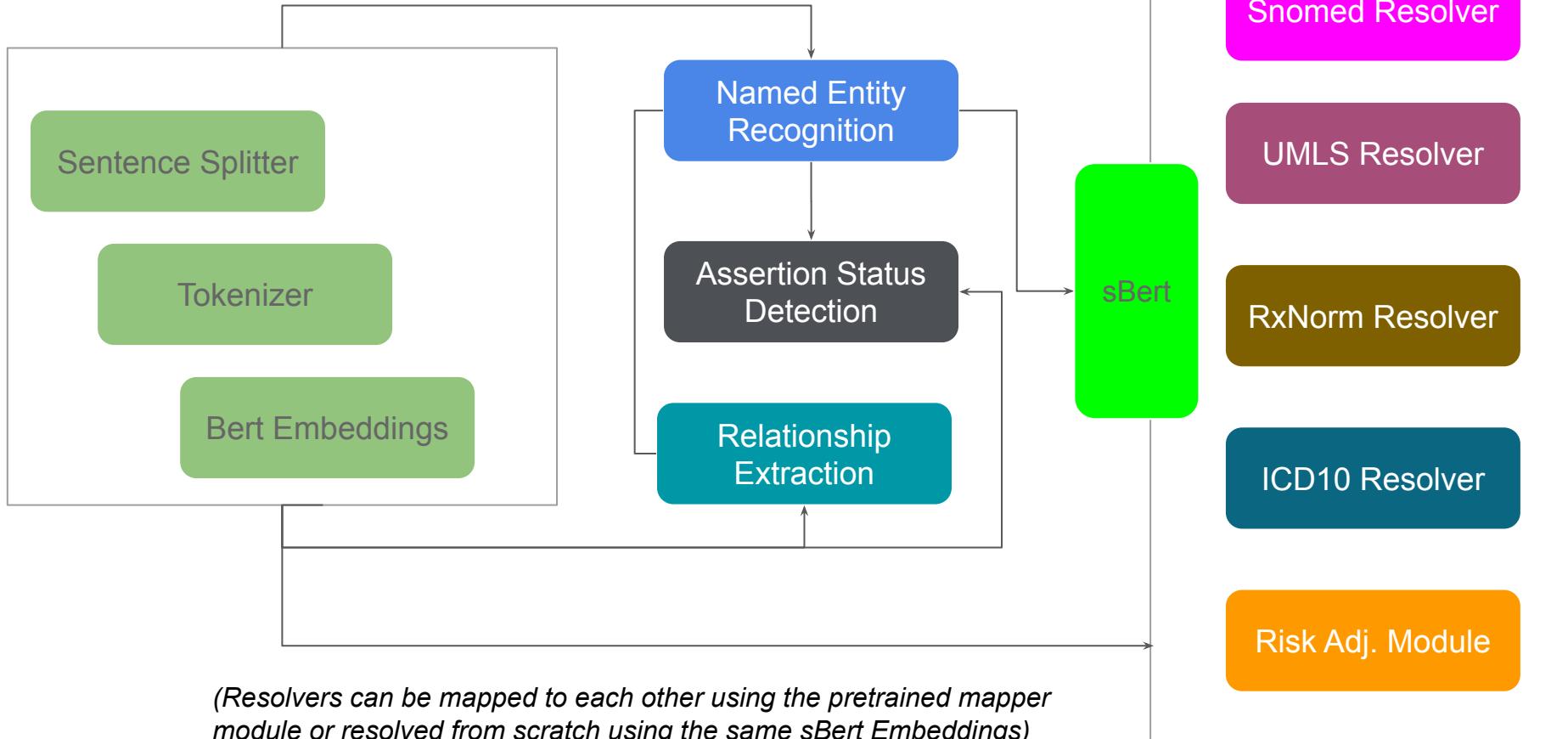
UMLS

- sbiobertresolve_umls_findings
- sbiobertresolve_umls_major_concepts
- sbiobertresolve_umls_disease_syndrome
- sbiobertresolve_umls_clinical_drugs

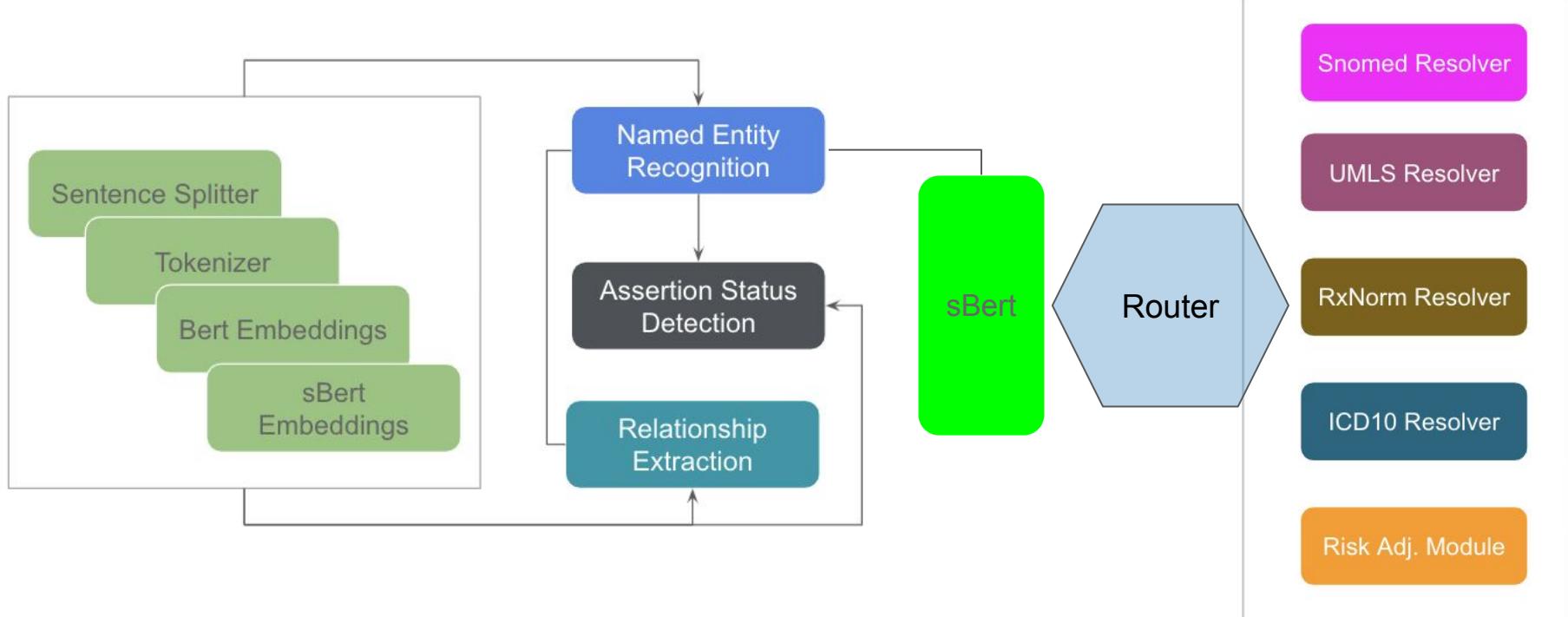
mapping

- icd10cm_snomed_mapping : ICD10 Codes to Snomed Codes
- snomed_icd10cm_mapping : Snomed Codes to ICD Codes
- icd10cm_umls_mapping : ICD Codes to UMLS Codes
- snomed_umls_mapping : Snomed Codes to UMLS Codes
- rxnorm_umls_mapping : RxNorm Codes to UMLS Codes
- mesh_umls_mapping : MeSH Codes to UMLS Codes
- rxnorm_mesh_mapping : RxNorm Codes to MeSH Codes

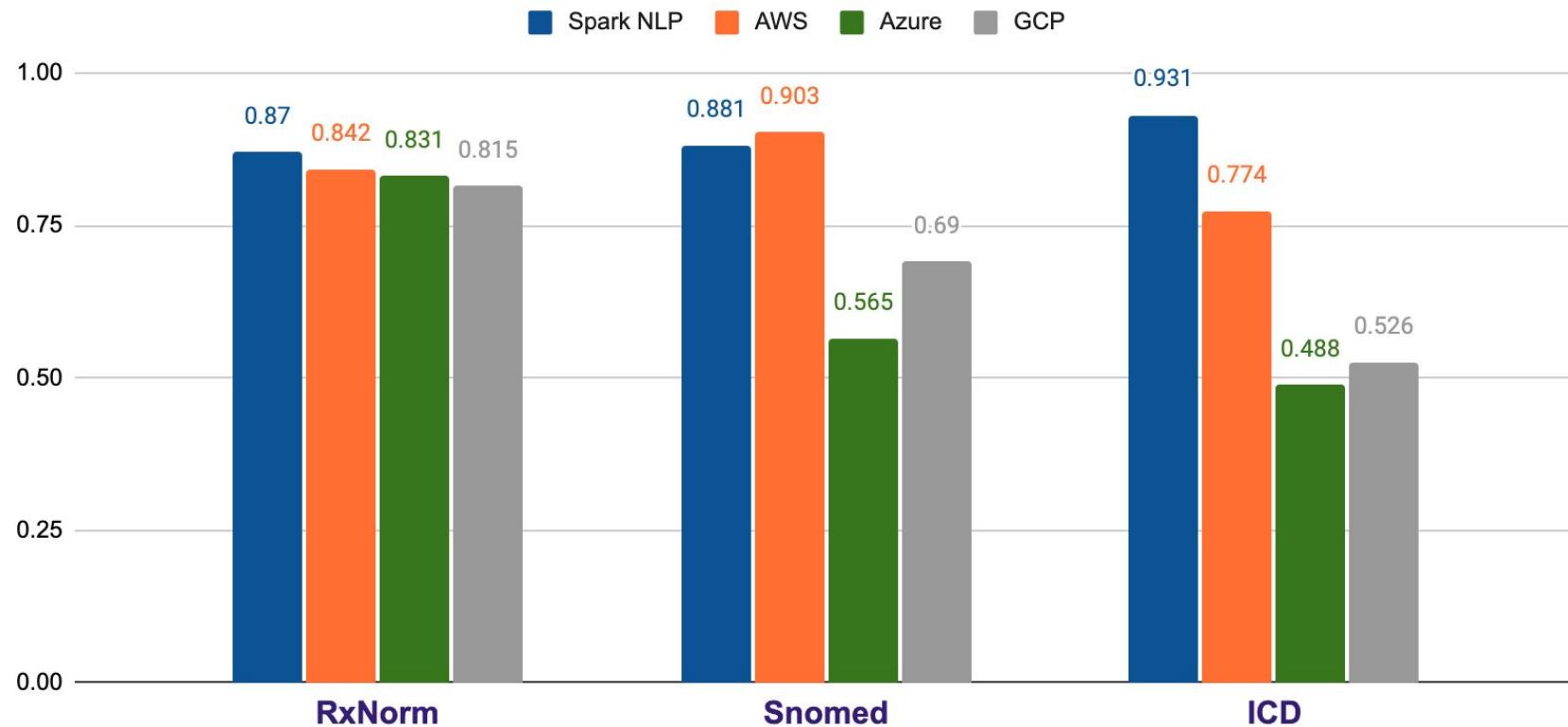
Clinical Entity Resolution



Clinical Entity Resolution



Top - 5 Results





Spark NLP for Healthcare Data Scientists

April 17-18, 2024

Veysel Kocaman
Head of Data Science
veysel@johnsnowlabs.com

Hasham UI Haq
Sr. Data Scientist

Muhammet Santas
Data Scientist

Gursev Pirge
Sr. Data Scientist

Youssef Mellah
Data Scientist

Agenda

Day	Dur.	Topic	Notebooks	Instructor
April 17 th	50 min	- Intro to John Snow Labs and Spark NLP - Healthcare NLP in Spark NLP	-	Veysel
	50 min	- Clinical Named Entity Recognition	01.0, 10.0, 01.5	Gurseyv
	50 min	- Clinical Assertion Status Model - Clinical Relation Extraction Model	02.0, 03.0	Gurseyv
	50 min	- Clinical Entity Resolution - Medicare Risk Adjustment	05.0, 05.1, 05.2, 05.3	Muhammet
April 18 th	50 min	- Retrieval Augmented Generation with Spark NLP	30.0, spark-nlp-workshop/ generative-ai	Muhammet
	50 min	- LLMs (Text Generation, Text2SQL, Summarization, QA)	23.0, 24.0, 25.1, 29.0	Hasham
	50 min	- De-Identification and Obfuscation of PHI	04.0, 04.1	Youssef
	50 min	- Text Classification (Section Headers&Classifier) - Few Shot Classifier	08.0, 08.3, 08.6, 18.1	Gurseyv

Part - V

- ❖ Retrieval Augmented Generation with Spark NLP

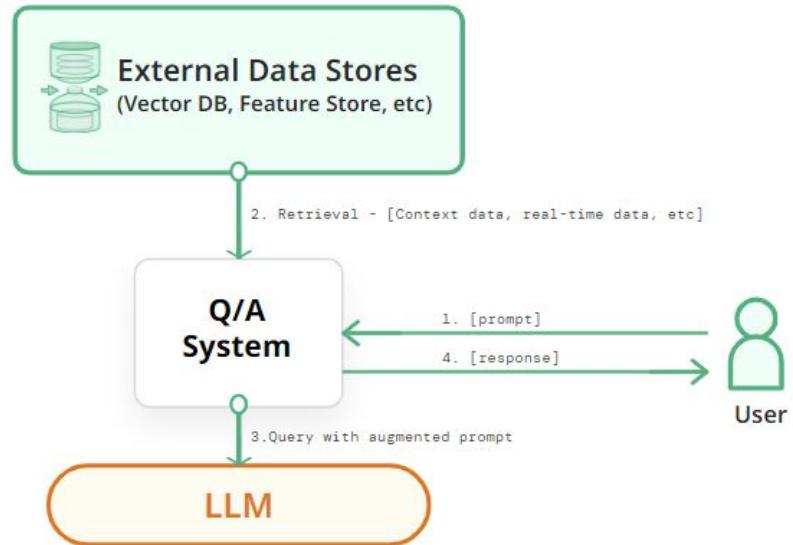
Capabilities of GPT-4 on Medical Challenge Problems

Dataset	Component	GPT-4 (5 shot)	GPT-4 (zero shot)	GPT-3.5 (5 shot)	GPT-3.5 (zero shot)	Flan-PaLM 540B* (few shot)
MedQA	Mainland China	75.31	71.07	44.89	40.31	—
	Taiwan	84.57	82.17	53.72	50.60	—
	United States (5-option)	78.63	74.71	47.05	44.62	—
	United States (4-option)	81.38	78.87	53.57	50.82	60.3**
PubMedQA	Reasoning Required	74.40	75.20	60.20	71.60	79.0
MedMCQA	Dev	72.36	69.52	51.02	50.08	56.5
MMLU	Clinical Knowledge	86.42	86.04	68.68	69.81	77.00
	Medical Genetics	92.00	91.00	68.00	70.00	70.00
	Anatomy	80.00	80.00	60.74	56.30	65.20
	Professional Medicine	93.75	93.01	69.85	70.22	83.80
	College Biology	93.75	95.14	72.92	72.22	87.50
	College Medicine	76.30	76.88	63.58	61.27	69.90

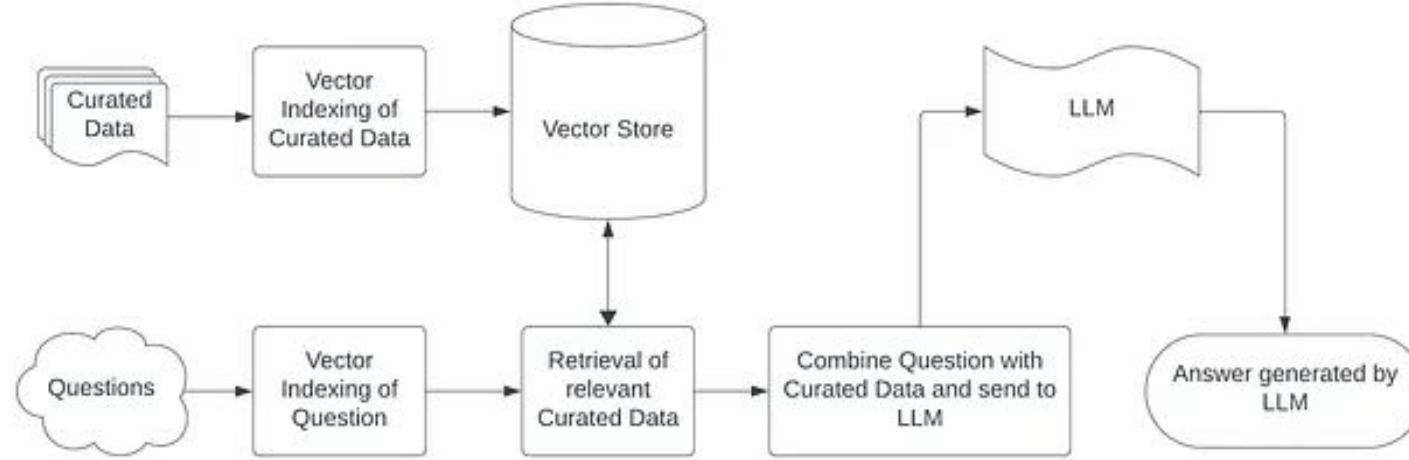
Nori et al. "Capabilities of GPT-4 on Medical Challenge Problems." *arXiv preprint 2303.13375* (2023).

Retrieval Augmented Generation (RAG)

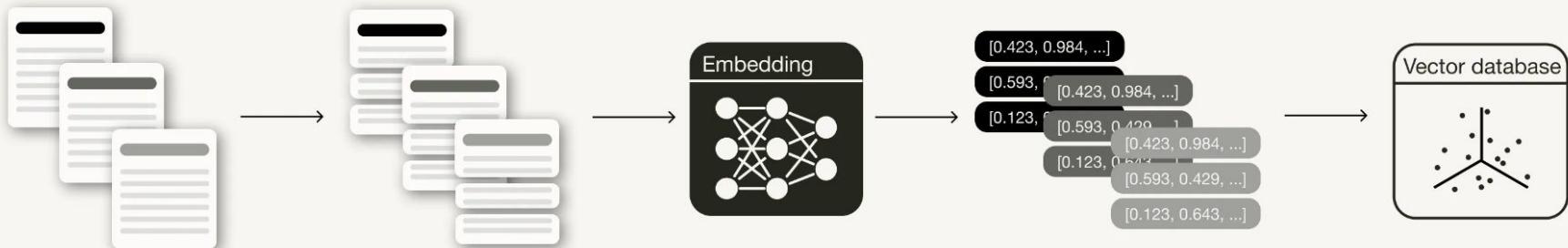
- Pre-trained Large Language Models (LLMs) like foundation models have limitations such as static learning, generating incorrect information, and potential data privacy breaches. Retrieval-Augmented Generation (RAG) address these by incorporating vector databases or feature stores, enriching prompt context and enhancing response accuracy.
- Retrieval-Augmented Generation (RAG) enhances model predictions by utilizing an external datastore at inference time. This integration enriches prompts with a mix of context, historical data, and up-to-date knowledge, leading to more accurate and relevant responses.
- RAG combines information retrieval with LLM text generation, using user prompts to fetch external context from a data store. This approach creates richer prompts with additional context, providing insights beyond the LLM's existing knowledge base.



Retrieval Augmented Generation (RAG)



Ingestion



Retrieval-Augmented Generation (RAG)

1

Source Documents

- Preprocessing (OCR, basic cleaning, formatting, ...)
- Metadata extraction (keywords, entities, author, title, ...)
- Feature engineering (table understanding, chart2text, summarization, ...)

2

Document Splitting/
Chunking

- Splitting strategy (content-aware, section-wise, task-based, char-based, tables, figures, items...)
- Max chunk size, overlap area, ...

3

Split Embeddings

- Embeddings models (e5, allmpnet, gte, bge, openai-text-ada, ... > MTEB)
- Model size and speed (384, 512, 768, 1024, ...)
- Scalability (embeddings collection at scale)

4

VectorDB

- Retrieval strategy (recursive, knn, BM25, span/ query expansion, ...)
- Postprocessing (reranking, filtering, diversity, ...)
- Speed and scalability

5

LLM

- Model performance, instruction following, guardrails, size, deployment)
- Context size (16K, 100K, ... > chat memory)
- Prompt template (*given the context splits, answer the question*)

Document Split / Chunk Augmentation In RAG

- Domain-specific (medical, finance, legal, ...)
- Context-aware (table heavy content, plain text, sections/ headers, document hierarchy, ...)
- Ideal chunk/ split size w.r.t. embeddings model and LLMs
- Chunk coherence (sentence/ paragraph awareness, overlapping size, ...)
- Metadata augmentation
 - Enrich splits with parent-doc metadata - cascading metadata into content
 - Title and section header extraction/ assignment
 - Split by sentence boundaries
 - Entity/ key phrase extraction - ADEs for drugs, drug names, symptoms etc.
 - Normalize/ group key entities - e.g. mapping diseases to ICD10
 - Summarize splits
 - Generate questions - extracts a set of questions that each split can answer
 - Augment with external KBs - knowledge graphs, reference tables, policies
 - Augment non-text content - Table-2-text, chart-2-text, numerics, formulas

Metadata Augmentation In RAG

> Enrich splits with parent-doc metadata

► `carcinogenesis_and_mutagenesis_and_impairment_of_fertility: [] 1 item`

► `overdosage: [] 1 item`

0: "10. OVERDOSAGE There have been no known experiences of overdose with diclofenac sodium topical solution. Symptoms following acute NSAID overdose are usually limited to lethargy, drowsiness, nausea, vomiting, and epigastric pain, which are generally reversible with supportive care. Gastrointestinal bleeding can occur. Hypertension, acute renal failure, respiratory depression and coma may occur, but are rare. Anaphylactoid reactions have been reported with therapeutic ingestion of NSAIDs, and may occur following an overdose. Manage patients using symptomatic and supportive care following an NSAID overdose. There are no specific antidotes. Emesis is not recommended due to a possibility of aspiration and subsequent respiratory irritation by DMSO contained in diclofenac sodium topical solution. Activated charcoal (60 to 100 g in adults, 1 to 2 g/kg in children) and/or osmotic cathartic may be indicated in patients seen within 4 hours of ingestion with symptoms or following a large overdose (5 to 10 times the usual dose). Forced diureses, alkalinization of urine, hemodialysis, or hemoperfusion may not be useful due to high protein binding. For additional information about overdose treatment, call a poison control center (1-800-222-1222)."

► `spl_medguide_table: [] 2 items`

0.76

"splits": [

```
{  
    "text": "This text is about overdosage of the drug Diclofenac Sodium.\n10. OVERDOSAGE There have been no known experiences of overdose with diclofenac sodium topical solution. Symptoms following acute NSAID overdose are usually limited to lethargy, drowsiness, nausea, vomiting, and epigastric pain, which are generally reversible with supportive care. Gastrointestinal bleeding can occur. Hypertension, acute renal failure, respiratory depression and coma may occur, but are rare. Anaphylactoid reactions have been reported with therapeutic ingestion of NSAIDs, and may occur following an overdose. Manage patients using symptomatic and supportive care following an NSAID overdose. There are no specific antidotes. Emesis is not recommended due to a possibility of aspiration and subsequent respiratory irritation by DMSO contained in diclofenac sodium topical solution. Activated charcoal (60 to 100 g in adults, 1 to 2 g/kg in children) and/or osmotic cathartic may be indicated in patients seen within 4 hours of ingestion with symptoms or following a large overdose (5 to 10 times the usual dose). Forced diureses, alkalinization of urine, hemodialysis, or hemoperfusion may not be useful due to high protein binding. For additional information about overdose treatment, call a poison control center (1-800-222-1222).",  
    "meta_info": {  
        "id": "7dc7fb31-3323-4767-e053-2991aa0abff5_abuse_overdosage_0",  
        "drug_brand_name": "Diclofenac Sodium",  
        "section": "abuse_overdosage",  
        "sub_section": "overdosage"  
    }  
},  
{  
    "text": "This text is about adverse reactions of the drug Diclofenac Sodium.\n6. ADVERSE REACTIONS The most common adverse events with diclofenac sodium topical
```

0.80

Metadata Augmentation In RAG

> Title and section header extraction/ assignment

Discharge Summary - Mesothelioma - 1

Description:

Mesothelioma, pleural effusion, atrial fibrillation, anemia, ascites, esophageal reflux, and history of deep venous thrombosis.
(Medical Transcription Sample Report)

PRINCIPAL DIAGNOSIS:

Mesothelioma.

SECONDARY DIAGNOSES:

Pleural effusion, atrial fibrillation, anemia, ascites, esophageal reflux, and history of deep venous thrombosis.

PROCEDURES

1. On August 24, 2007, decortication of the lung with pleural biopsy and transpleural fluoroscopy.
2. On August 20, 2007, thoracentesis.
3. On August 31, 2007, Port-A-Cath placement.

HISTORY AND PHYSICAL:

The patient is a 41-year-old Vietnamese female with a nonproductive cough that started last week. She has had right-sided chest pain radiating to her back with fever starting yesterday. She has a history of pericarditis and pericardectomy in May 2006 and developed cough with right-sided chest pain, and went to an urgent care center. Chest x-ray revealed right-sided pleural effusion.

PAST MEDICAL HISTORY

1. Pericardectomy.
2. Pericarditis.
2. Atrial fibrillation.
4. RNCA with intracranial thrombolytic treatment.
- 5 PTA of MCA.
6. Mesenteric venous thrombosis.
7. Pericardial window.
8. Cholecystectomy.
9. Left thoracentesis.

FAMILY HISTORY:

No family history of coronary artery disease, CVA, diabetes, CHF or MI. The patient has one family member, a sister, with history of cancer.

SOCIAL HISTORY:

She is married. Employed with the US Post Office. She is a mother of three. Denies tobacco, alcohol or illicit drug use.

default split

```
Document(page_content='Sample Type / Medical Specialty:\nHematology - Oncology\n\nSample Name:\nTrue Cut Needle Biopsy - Breast\nDescription:\nTrue cut needle biopsy of the breast. This 65-year-old female on exam was noted to have dimpling and puckering of the skin associated with nipple discharge. On exam, she has a noticeable carcinoma of the left breast with dimpling, puckering, and erosion through the skin.\n(Medical Transcription Sample Report)\n\nPREOPERATIVE DIAGNOSIS:\nCarcinoma of the left breast.\n\nPROCEDURE PERFORMED:\nTrue cut needle biopsy of the breast.\n\nGROSS FINDINGS:\nThis 65-year-old female on exam was noted to have dimpling and puckering of the skin associated with nipple discharge. On exam, she has a noticeable carcinoma of the left breast with dimpling, puckering, and erosion through the skin. At this time, a true cut needle biopsy was performed.\n\nPROCEDURE:', metadata={'document_id': 'd1987628-9184-4147-aeb6-9a2d04daabf7', 'patient_id': 'pt_11'})
```

- Default splitting doesn't observe section boundaries or headers.
- Ideally, each split should belong to one section with section title prepended.
- For unstructured/ free text documents, this requires NLP methods extract headers and split sections.

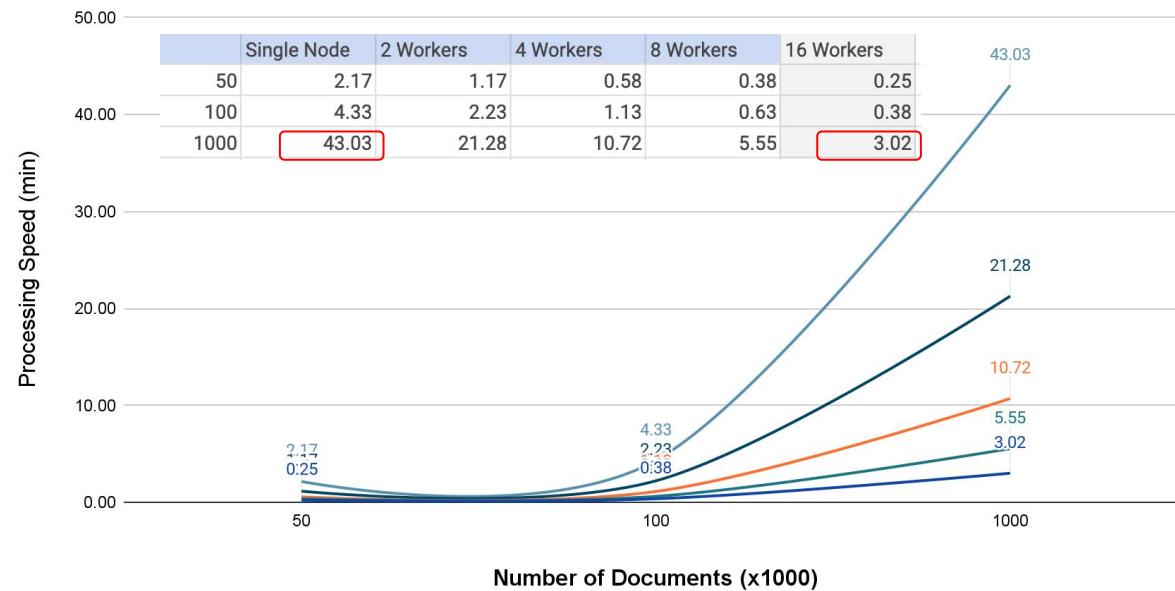
Overall MTEB English leaderboard

Rank	Model	Model Size (GB)	Embedding Dimensions
1	sionic-ai-v2		
2	sionic-ai-v1		
3	bge-large-en-v1.5	1.34	1024
4	bge-large-en	1.34	1024
5	bge-base-en-v1.5	0.44	768
6	gte-large	0.67	1024
7	gte-base	0.22	768
8	e5-large-v2	1.34	1024
9	bge-small-en-v1.5	0.13	384
10	instructor-xl	4.96	768
11	instructor-large	1.34	768
12	e5-base-v2	0.44	768
13	multilingual-e5-large	2.24	1024
14	e5-large	1.34	1024
15	gte-small	0.07	384
16	gte-small	0.07	384
17	text-embedding-ada-002	1536	
18	e5-base	0.44	768
19	e5-small-v2	0.13	384
20	instructor-base	0.44	768

Embeddings at Scale in RAG

Speed Comparison: Spark NLP vs Huggingface in Databricks for Embeddings Collection

— 16 Workers — 8 Workers — 4 Workers — 2 Workers — Single Node



By **natively scaling** on the Databricks cluster and adding more executors, **Spark NLP 5.0** achieves nearly **linear speed enhancements**.

Popular RAG Frameworks / Wrappers



haystack
by deepset



LlamaIndex



```
!pip install johnsnowlabs
```

Part - VI

- ❖ LLMs (Text Generation, Text2SQL, Summarization, Q&A)

Part - VII

- ❖ De-Identification and Obfuscation of PHI

Background & Motivation

Why de-identify?

- Organizations in possession of documents containing Protected Health Information (PHI) must follow privacy rules
- De-identification enables sharing of health data for medical research studies, policy assessments, other studies/assessments without violating patient privacy or requiring individual authorizations (HIPAA / GDPR restrictions around sharing no longer apply)

Protected Health Information (PHI)

All info created or received by an entity that:

- Relates to past/present/future health or condition of an individual; the provision of health care; or the past/present/future payment for the provision of health care
- Identifies the individual; or with respect to which there is a reasonable basis to believe the information can be used to identify the individual

PHI includes many common identifiers: names, geographic locations, dates, phone numbers, email, IDs, SSN, medical record numbers, etc

Spark NLP in Action

Spark NLP for Healthcare → De-Identification



**Deidentify structured
data**

[Live Demo](#)[Colab Netbook](#)

**Deidentify free text
documents**

[Live Demo](#)[Colab Netbook](#)

**Deidentify DICOM
documents**

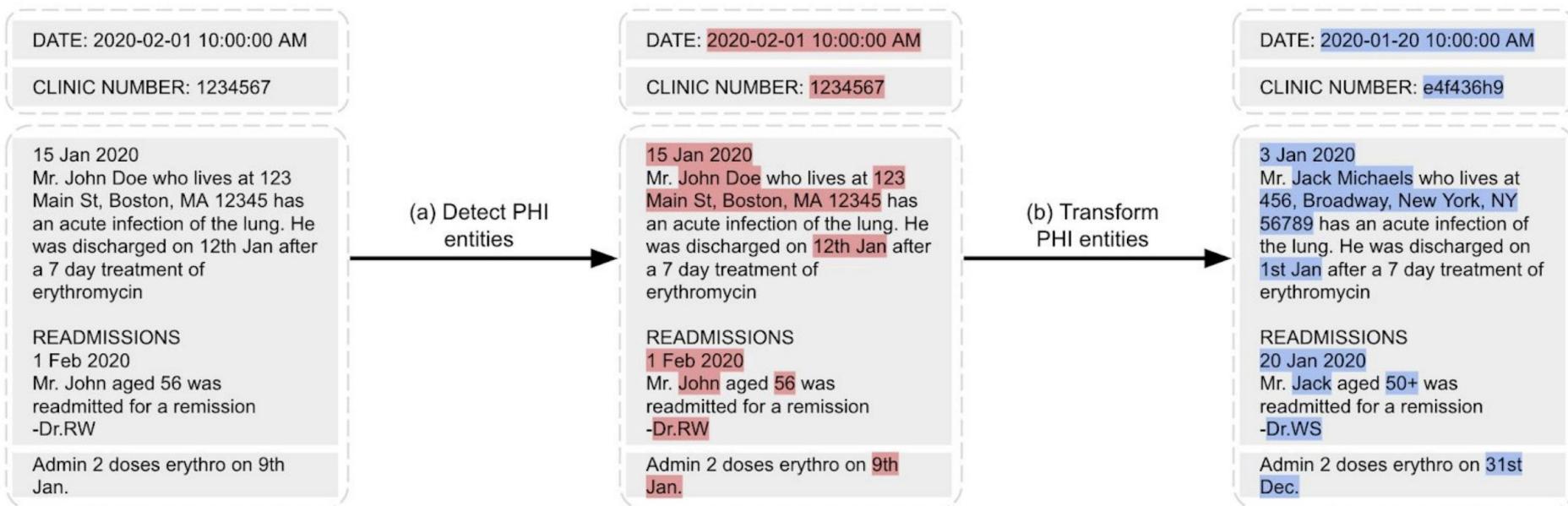
[Live Demo](#)[Colab Netbook](#)

**De-identify PDF
documents - HIPAA
Compliance**

[Live Demo](#)[Colab Netbook](#)

De-Identification

* Identifies potential pieces of content with personal information about patients and remove them by replacing with semantic tags.



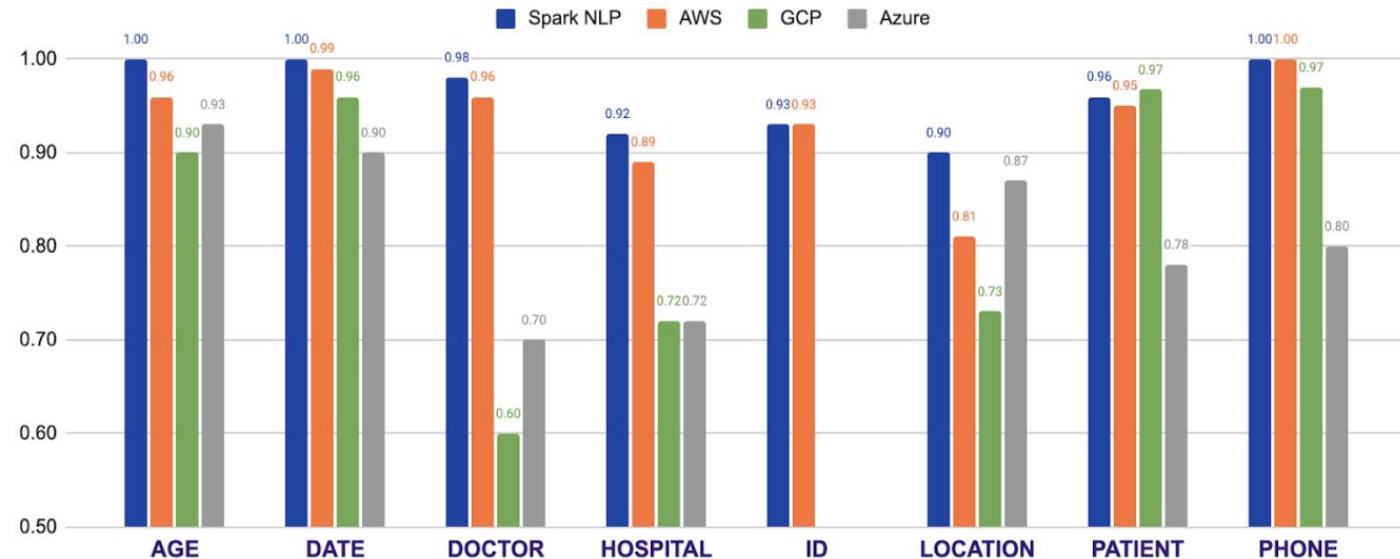
Veysel Kocaman

May 2 · 17 min read · [Listen](#)

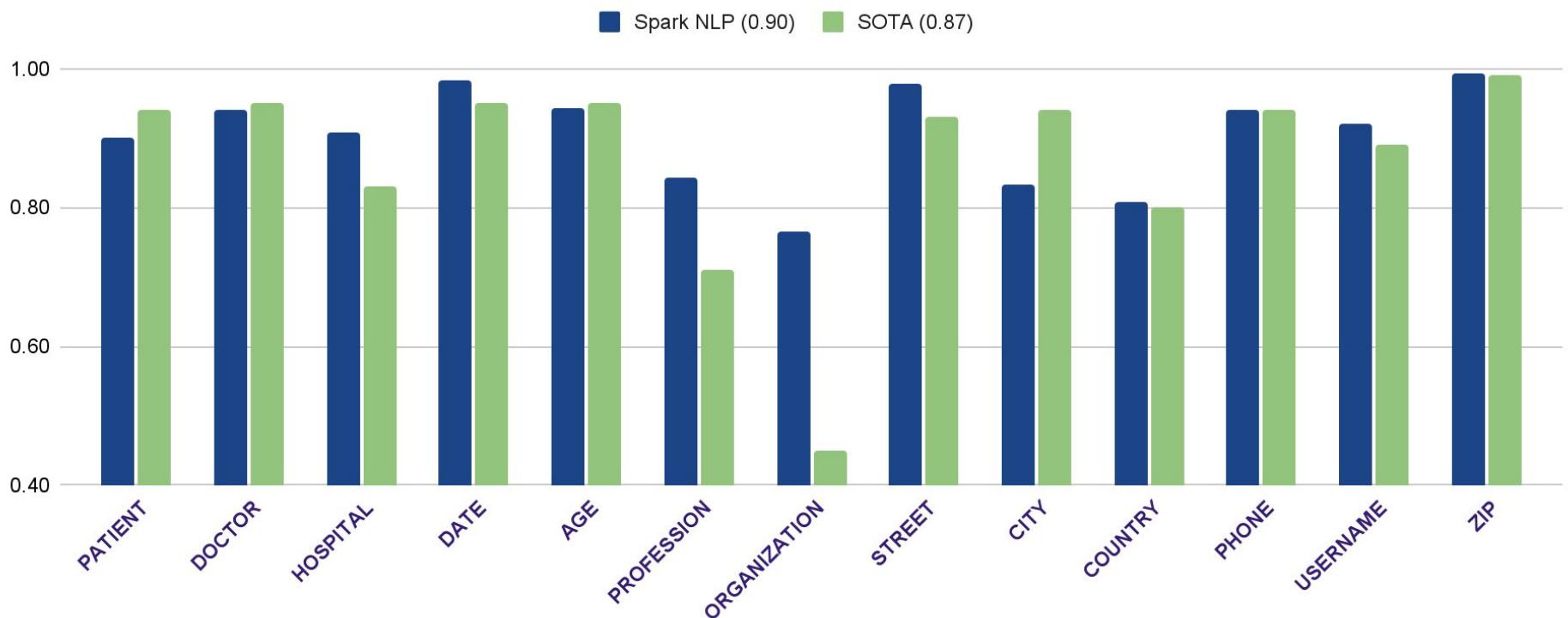


...

Comparison of Key Medical NLP Benchmarks — Spark NLP vs AWS, Google Cloud and Azure



Deidentification Benchmarks





How Providence Health De-Identified 700 Million Patient Notes with Spark NLP

Accuracy:

99.19

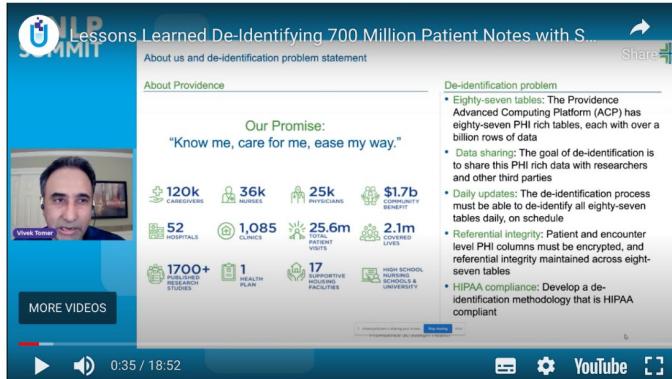
correctly de-identified sentences

Performance:

2.46 hours

to de-identify 500K patient notes.

[See how we did it](#)



De-identify



Software

We support:

- Tabular (headers, values)
- Text (NER, text matching)
- PDF: Text or Scanned
- Images (OCR & metadata)
- DICOM (OCR & metadata)

So you can:

- Replace (or delete a field)
- Mask (hash identifiers or shift dates)
- Obfuscate (name, locations, organizations)
- Generalize (disease codes, dates, addresses)

Part - VIII

- ❖ Text Classification (Section Headers & Classifier)
- ❖ Few-Shot Classifier

Spark NLP Resources

Spark NLP Official page

Spark NLP Workshop Repo

JSL Youtube channel

JSL Blogs

Introduction to Spark NLP: Foundations and Basic Components (Part-I)

Introduction to: Spark NLP: Installation and Getting Started (Part-II)

Named Entity Recognition with Bert in Spark NLP

Text Classification in Spark NLP with Bert and Universal Sentence Encoders

Spark NLP 101 : Document Assembler

Spark NLP 101: LightPipeline

<https://www.oreilly.com/radar/one-simple-chart-who-is-interested-in-spark-nlp/>

<https://blog.dominodatalab.com/comparing-the-functionality-of-open-source-natural-language-processing-libraries/>

<https://databricks.com/blog/2017/10/19/introducing-natural-language-processing-library-apache-spark.html>

<https://databricks.com/fr/session/apache-spark-nlp-extending-spark-ml-to-deliver-fast-scalable-unified-natural-language-processing>

<https://medium.com/@saif1988/spark-nlp-walkthrough-powered-by-tensorflow-9965538663fd>

<https://www.kdnuggets.com/2019/06/spark-nlp-getting-started-with-worlds-most-widely-used-nlp-library-enterprise.html>

<https://www.forbes.com/sites/forbestechcouncil/2019/09/17/why-spark-nlp-is-the-most-widely-used-nlp-library-enterprise/>

<https://medium.com/hackernoon/mueller-report-for-nerds-spark-meets-nlp-with-tensorflow-and-bert-part-1-32490a8f8f12>

<https://www.analyticsindiamag.com/5-reasons-why-spark-nlp-is-the-most-widely-used-library-enterprise/>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-training-spark-nlp-and-spacy-pipelines>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-accuracy-performance-and-scalability>

<https://www.infoworld.com/article/3031690/analytics/why-you-should-use-spark-for-machine-learning.html>



Thank you !

Doc2Chunk / Chunk2Doc

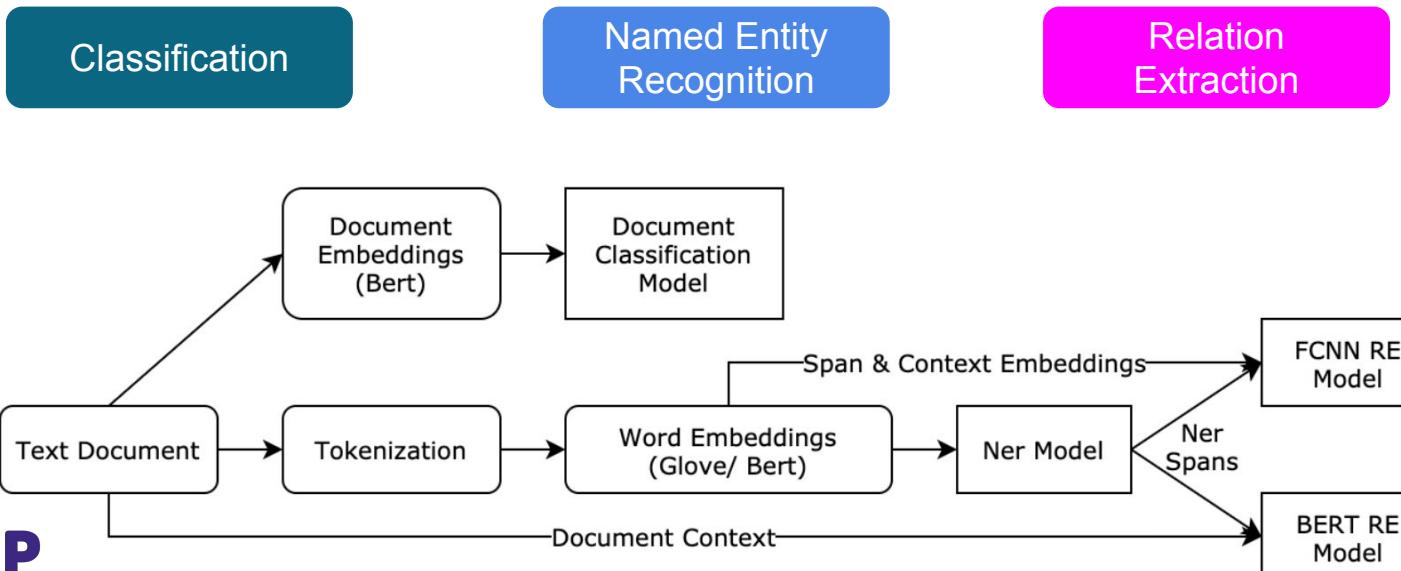
- ✓ Transform the data from one AnnotatorType to another.
- ✓ Doc2Chunk
 - Converts DOCUMENT type annotations into CHUNK type with the contents of a chunkCol.
- ✓ Chunk2Doc
 - Converts a CHUNK type column back into DOCUMENT.
 - Useful when trying to re-tokenize or do further analysis on a CHUNK result.

```
chunkAssembler = Doc2Chunk() \  
    .setInputCols("document") \  
    .setChunkCol("target") \  
    .setOutputCol("chunk")
```

```
chunkToDoc = Chunk2Doc() \  
    .setInputCols("chunk") \  
    .setOutputCol("chunkConverted")
```

Adverse Drug Reactions (ADR)

Document	Class	ADE Entity	Drug Entity	relation
I feel a bit drowsy & have a little blurred vision after taking insulin.	ADE	drowsy blurred vision	insulin insulin	Positive Positive
@yho fluvastatin gave me cramps, but lipitor suits me!	ADE	cramps cramps	fluvastatin lipitor	Positive Negative
I just took advil and haven't had any gastric problems so far.	NEG	-	-	-



Adverse Drug Reactions (ADR) Benchmark



Dataset	GLoVe Embeddings						BERT Embeddings						SOTA	
	Precision		Recall		F1		Precision		Recall		F1		F1	
	strict	relax	strict	relax	strict	relax	strict	relax	strict	relax	strict	relax		
ADE	88.32	93.77	89.26	94.80	88.78	94.27	90.0	94.47	93.56	98.22	91.75	96.31	91.3	
	87.81	93.59	88.81	94.66	88.30	94.12	89.6	94.37	93.18	98.13	91.36	96.21		
CADEC	78.14	89.04	77.14	88.01	77.62	88.50	78.53	88.63	79.03	89.32	78.76	88.95	71.9	
	71.87	86.36	71.67	86.13	71.75	86.23	72.38	86.14	73.64	87.66	72.99	86.88		
SMM4H	81.43	90.33	72.17	78.51	76.01	83.41	78.5	86.76	75.23	82.42	76.73	84.41	67.81	
	83.66	91.34	71.31	77.86	76.99	84.06	79.13	87.09	74.33	81.81	76.65	84.36		

Table 2: NER metrics on benchmark datasets. For each dataset, macro and micro averaged scores are displayed on first and second row respectively. SOTA metrics for ADE, CADEC, and SMM4H are obtained from (Yan et al. 2021), (Stanovsky, Gruhl, and Mendes 2017), and (Ge et al. 2020) respectively, and are macro-averaged.

Named Entity Recognition (NER)

Adverse Drug Reactions (ADR) Benchmark



Dataset	GLoVe (Avg.) Embeddings			BERT (Avg.) Embeddings			BERT Sentence Embeddings			SOTA
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	F1
ADE	75.96	79.53	76.86	76.91	84.96	79.37	87.41	84.72	85.96	87.0
	86.84	81.22	83.43	88.13	84.38	85.38	90.97	91.20	91.03	
CADEC	85.29	84.24	84.71	86.50	86.11	86.30	87.13	86.32	86.69	81.5
	85.99	86.10	86.0	87.38	87.43	87.40	87.78	87.86	87.79	

Table 1: Classification Metrics on benchmark datasets. For each dataset, Macro and Micro averaged scores are displayed on first and second row respectively. SOTA metrics for ADE and CADEC datasets are obtained from (Huynh et al. 2016) and (Alimova and Tutubalina 2019) respectively.

Dataset	Base (FCNN) RE			BERT RE			SOTA
	Precision	Recall	F1	Precision	Recall	F1	F1
ADE Corpus	69.11	86.22	74.70	81.31	79.03	80.10	83.74
ADE Enriched with n2c2	89.01	89.44	89.22	89.19	90.93	90.02	

Table 3: Relation Extraction performance on the ADE benchmark dataset. The test set was kept standard for a fair comparison, and all scores are macro-averaged due to high class imbalance. SOTA metrics for RE on ADE corpus as reported by (Crone 2020)

Part - V

- ❖ Contextual Parser (Rule-based NER)

❖ Key Chunk Phrase Extractor

Chunk KeyPhrase Extraction uses Bert Sentence Embeddings to determine the most relevant key phrases describing a text. The input to the model consists of chunk annotations and sentence or document annotation. The model compares the chunks against the corresponding sentences/documents and selects the chunks which are most representative of the broader text context (i.e. the document or the sentence they belong to). The key phrases candidates (i.e. the input chunks) can be generated in various ways, e.g. by NGramGenerator, TextMatcher or NerConverter. The model operates either at sentence (selecting the most descriptive chunks from the sentence they belong to) or at document level. In the latter case, the key phrases are selected to represent all the input document annotations.

Chunk Key Phrase Extractor (KCPE)

key_phrase	source	DocumentSimilarity	MMRScore	sentence
type two diabetes mellitus	NER	0.7639750686118073	0.4583850593816694	0
subsequent type diabetes	ngrams	0.7503709443591438	0.08298243928224425	0
HTG-induced pancreatitis years	ngrams	0.6817062970203589	0.11246275270031031	0
hepatitis obesity	ngrams	0.6666053470245074	0.1177052008980295	0
mellitus diagnosed years	ngrams	0.6389213391545323	0.08129479185432026	0
history gestational diabetes	ngrams	0.6219876368539883	0.0950104202982544	0
vomiting	ngrams	0.5824238088130589	0.14864183399720493	0
admitted starvation ketosis	ngrams	0.5789875069392564	0.12008073486190007	0
five-day amoxicillin respiratory	ngrams	0.5330653868257814	0.09428153526023508	0
28-year-old female history	ngrams	0.38613601247069695	0.12987678861407687	0

YAKE

keyword	score
years prior presentation	0.006335399690627251
prior presentation	0.011644010991495998
prior presentation subsequent	0.020272229518351368
weeks prior presentation	0.020272229518351368
respiratory tract infection	0.02568455658449274
anion gap	0.025965846371439553
physical examination presentation	0.02840600503736659
obtained hours presentation	0.028532992974589392
examination presentation significant	0.028532992974589392
prior	0.029673513139537065
years prior	0.03008818777992058
anion gap elevated	0.031568192739369824

CKPE

key_phrase_candidate	DocumentSimilarity
pancreatitis years prior	0.6491587146812722
diagnosed years prior	0.38594469396979897
respiratory tract infection	0.34452766290310755
patient treated insulin	0.3413457416284759
serum	0.3371024001999838
presentation revealed glucose	0.31458360368143906
examination presentation significant	0.29099950377907047
prior analysis due	0.22501711661945623
prior	0.21634008371261446
physical examination presentation	0.19165189487112474

key_phrase_candidate	source
28-year-old female history	ngram
28-year-old	NER
female history gestational	ngram
female	NER
history gestational diabetes	ngram
gestational diabetes mellitus	NER
gestational diabetes mellitus	ngram
diabetes mellitus diagnosed	ngram
mellitus diagnosed years	ngram
diagnosed years prior	ngram
eight years prior	NER
years prior presentation	ngram
prior presentation subsequent	ngram
presentation subsequent type	ngram
subsequent type diabetes	ngram
type diabetes mellitus	ngram
type two diabetes mellitus	NER
diabetes mellitus (ngram
mellitus (T2DM	ngram
(T2DM),	ngram
T2DM), prior	ngram
T2DM	NER
), prior episode	ngram
prior episode HTG-induced	ngram
episode HTG-induced pancreatitis	ngram
HTG-induced pancreatitis years	ngram
HTG-induced pancreatitis	NER
pancreatitis years prior	ngram
three years prior	NER
years prior presentation	ngram
prior presentation ,	ngram
presentation , acute	ngram
, acute hepatitis	ngram