Spark NLP for Healthcare

Modular Approach to Solve Problems at Scale in Healthcare NLP

April 20, 2022

Veysel Kocaman
Lead Data Scientist
veysel@johnsnowlabs.com

ODSC EAST 2022

Setup

RUNNING CODE:

https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Healthcare

[How to set up Google Colab]

BOOKMARK:

https://nlp.johnsnowlabs.com/models

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

spark-nlp.slack.com

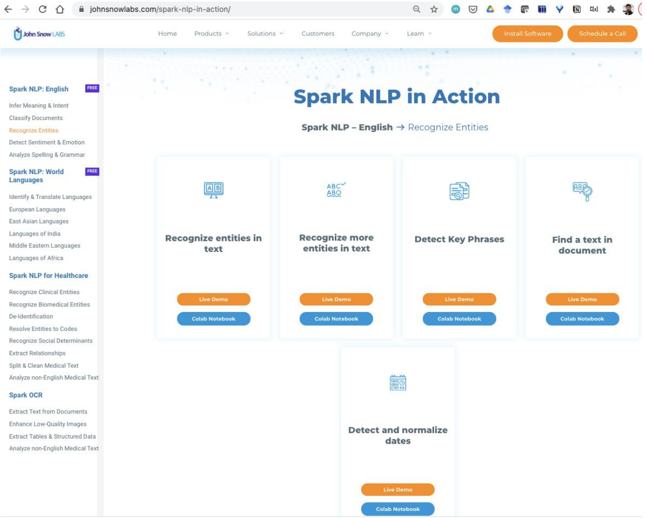
spark-nlp==3.4.2
spark-nlp-jsl==3.5.0

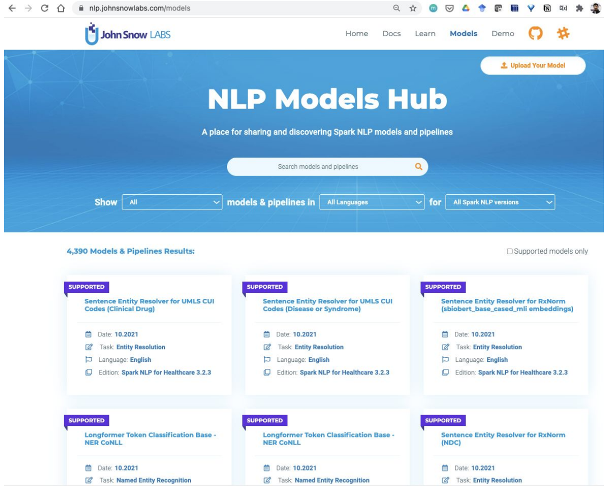
Open in Colab

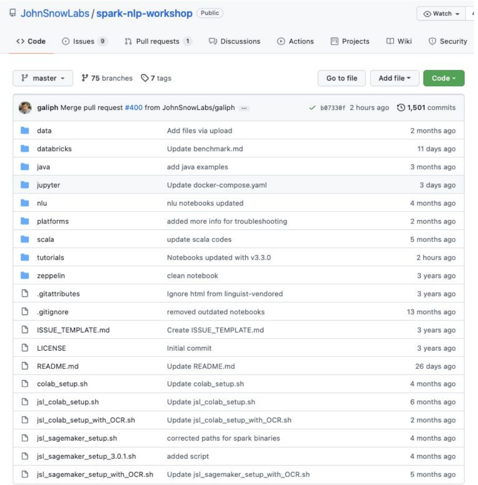
master spark-nlp-workshop / tutorials / Certification_Trainings / Healthcare /

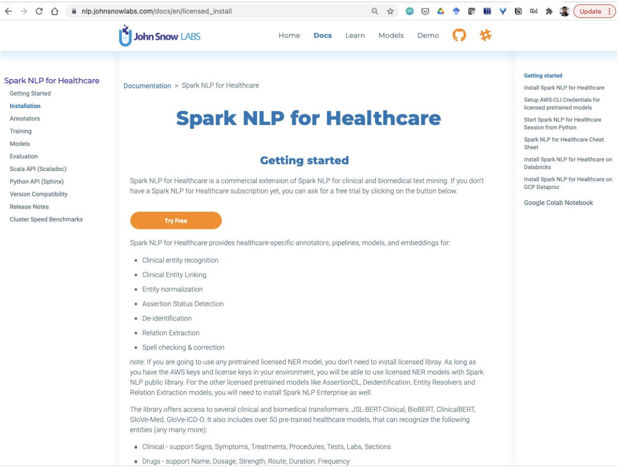
Go to file | Add file

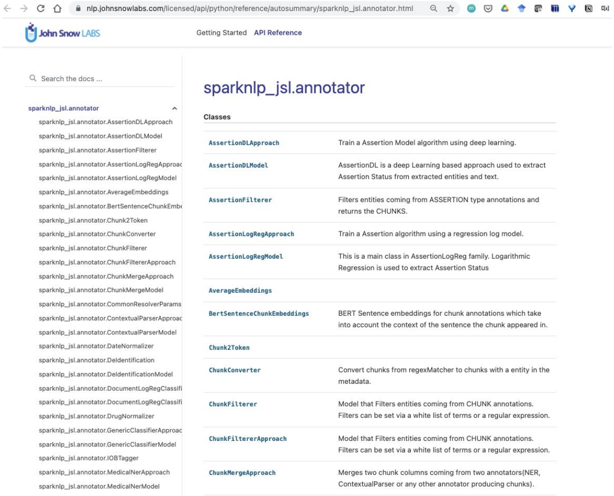
1.2.Contextual_Parser_Rule_Based_NER.ipynb	Notebooks updated with v3.3.0	2 hours ago
1.3.prepare_CoNLL_from_annotations_for_NER.ipynb	notebooks updated v3.1.1	4 months ago
1.4.Biomedical_NER_SparkNLP_paper_reproduce.ip...	Notebooks updated with v3.3.0	2 hours ago
1.5.Resume_MedicalNer_Model_Training.ipynb	Notebooks updated with v3.3.0	2 hours ago
1.6.BertForTokenClassification_NER_SparkNLP_with...	colab link updated	yesterday
1.Clinical_Named_Entity_Recognition_Model.ipynb	Notebooks updated with v3.3.0	2 hours ago
10.1.Clinical_Relation_Extraction_BodyParts_Models...	Notebooks updated with v3.3.0	2 hours ago
10.2.Clinical_RE_Knowledge_Graph_with_Neo4j.ipynb	Notebooks updated with v3.3.0	2 hours ago
10.Clinical_Relation_Extraction.ipynb	Notebooks updated with v3.3.0	2 hours ago
11.1.Healthcare_Code_Mapping.ipynb	Notebooks updated with v3.3.0	2 hours ago
11.2.Pretrained_NER_Profiling_Pipelines.ipynb	new NER profiling notebook added	20 days ago
11.Pretrained_Clinical_Pipelines.ipynb	Notebooks updated with v3.3.0	2 hours ago
12.Named_Entity_Disambiguation.ipynb	notebooks updated v3.1.2	3 months ago
13.1.Finetuning_Sentence_Entity_Resolver_Model.ip...	Notebooks updated with v3.3.0	2 hours ago
13.Snomed_Entity_Resolver_Model_Training.ipynb	Notebooks updated with v3.3.0	2 hours ago
14.German_Healthcare_Models.ipynb	notebook updated	27 days ago
15.German_Licensed_Models.ipynb	notebooks updated v3.1.2	3 months ago
16.Adverse_Drug_Event_ADE_NER_and_Classifier.ip...	notebooks updated v3.1.2	3 months ago
17.Graph_builder_for_DL_models.ipynb	notebook updated	2 months ago
19.Financial_Contract_NER.ipynb	notebooks updated v3.1.2	3 months ago
2.Clinical_Assertion_Model.ipynb	Notebooks updated with v3.3.0	2 hours ago
20.SentenceDetectorDL_Healthcare.ipynb	notebooks updated v3.1.2	3 months ago
21.Gender_Classifier.ipynb	notebooks updated v3.1.2	3 months ago
22.CPT_Entity_Resolver.ipynb	notebooks updated v3.1.2	3 months ago
23.Drug_Normalizer.ipynb	notebooks updated v3.1.2	3 months ago
24.1.Improved_Entity_Resolution_with_SentenceChu...	Add files via upload	2 months ago
24.Improved_Entity_Resolvers_in_SparkNLP_with_s...	notebook name fixed	2 months ago
25.Date_Normalizer.ipynb	colab links fixed	2 months ago
3.1.Calculate_Medicare_Risk_Adjustment_Score.ipynb	Notebooks updated with v3.3.0	2 hours ago
3.Clinical_Entity_Resolvers.ipynb	Notebooks updated with v3.3.0	2 hours ago
4.1.Pretrained_Clinical_DeIdentification.ipynb	typo fixed	5 months ago
4.Clinical_DeIdentification.ipynb	Add files via upload	3 months ago

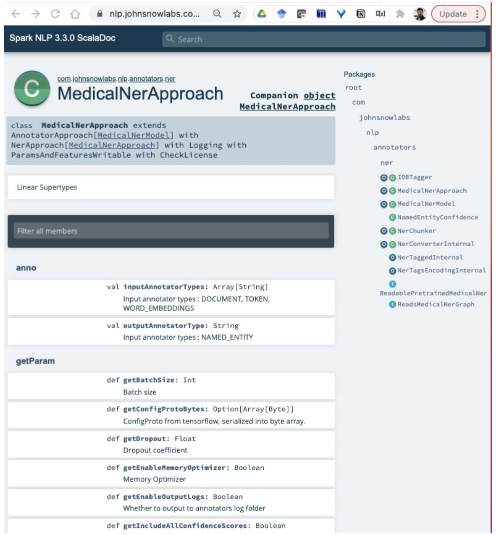

johnsnowlabs.com/spark-nlp-in-action/
John Snow LABS
Home
Products
Solutions
Customers
Company
Learn
Install Software
Schedule a Call
Spark NLP: English
FREE
Infer Meaning & Intent
Classify Documents
Recognize Entities
Detect Sentiment & Emotion
Analyze Spelling & Grammar
Spark NLP: World Languages
FREE
Identify & Translate Languages
European Languages
East Asian Languages
Languages of India
Middle Eastern Languages
Languages of Africa
Spark NLP for Healthcare
Recognize Clinical Entities
Recognize Biomedical Entities
De-Identification
Resolve Entities to Codes
Recognize Social Determinants
Extract Relationships
Split & 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
Spark NLP in Action
Spark NLP – English → Recognize Entities
Recognize entities in text
Recognize more entities in text
Detect Key Phrases
Find a text in document
Detect and normalize dates
Live Demo
Colab Notebook


nlp.johnsnowlabs.com/models
John Snow LABS
Home
Docs
Learn
Models
Demo
Upload Your Model
NLP Models Hub
A place for sharing and discovering Spark NLP models and pipelines
Search models and pipelines
Show
All
models & pipelines in
All Languages
for
All Spark NLP versions
4,390 Models & Pipelines Results:
Supported models only
SUPPORTED
Sentence Entity Resolver for UMLS CUI Codes (Clinical Drug)
Date: 10.2021
Task: Entity Resolution
Language: English
Edition: Spark NLP for Healthcare 3.2.3
SUPPORTED
Sentence Entity Resolver for UMLS CUI Codes (Disease or Syndrome)
Date: 10.2021
Task: Entity Resolution
Language: English
Edition: Spark NLP for Healthcare 3.2.3
SUPPORTED
Sentence Entity Resolver for RxNorm (sbiobert_base_cased_mli embeddings)
Date: 10.2021
Task: Entity Resolution
Language: English
Edition: Spark NLP for Healthcare 3.2.3
SUPPORTED
Longformer Token Classification Base - NER CoNLL
Date: 10.2021
Task: Named Entity Recognition
SUPPORTED
Longformer Token Classification Base - NER CoNLL
Date: 10.2021
Task: Named Entity Recognition
SUPPORTED
Sentence Entity Resolver for RxNorm (NDC)
Date: 10.2021
Task: Entity Resolution


JohnSnowLabs / spark-nlp-workshop
Public
Watch
Code
Issues 9
Pull requests 1
Discussions
Actions
Projects
Wiki
Security
master
75 branches
7 tags
Go to file
Add file
Code
galiph Merge pull request #400 from JohnSnowLabs/galiph
b07330f 2 hours ago
1,501 commits
data | Add files via upload | 2 months ago
databricks | Update benchmark.md | 11 days ago
java | add java examples | 3 months ago
jupyter | Update docker-compose.yaml | 3 days ago
nlu | nlu notebooks updated | 4 months ago
platforms | added more info for troubleshooting | 2 months ago
scala | update scala codes | 5 months ago
tutorials | Notebooks updated with v3.3.0 | 2 hours ago
zeppelin | clean notebook | 3 years ago
.gitattributes | Ignore html from linguist-vendored | 3 years ago
.gitignore | removed outdated notebooks | 13 months ago
ISSUE_TEMPLATE.md | Create ISSUE_TEMPLATE.md | 3 years ago
LICENSE | Initial commit | 3 years ago
README.md | Update README.md | 26 days ago
colab_setup.sh | Update colab_setup.sh | 4 months ago
jsl_colab_setup.sh | Update jsl_colab_setup.sh | 6 months ago
jsl_colab_setup_with_OCR.sh | Update jsl_colab_setup_with_OCR.sh | 2 months ago
jsl_sagemaker_setup.sh | corrected paths for spark binaries | 4 months ago
jsl_sagemaker_setup_3.0.1.sh | added script | 4 months ago
jsl_sagemaker_setup_with_OCR.sh | Update jsl_sagemaker_setup_with_OCR.sh | 5 months ago


nlp.johnsnowlabs.com/docs/en/licensed_install
Update
John Snow LABS
Home
Docs
Learn
Models
Demo
Spark NLP for Healthcare
Getting Started
Installation
Annotators
Training
Models
Evaluation
Scala API (Scaladoc)
Python API (Sphinx)
Version Compatibility
Release Notes
Cluster Speed Benchmarks
Documentation > Spark NLP for Healthcare
Spark NLP for Healthcare
Getting started
Spark NLP for Healthcare is a commercial extension of Spark NLP for clinical and biomedical text mining. If you don't have a Spark NLP for Healthcare subscription yet, you can ask for a free trial by clicking on the button below.
Try Free
Spark NLP for Healthcare provides healthcare-specific annotators, pipelines, models, and embeddings for:
Clinical entity recognition
Clinical Entity Linking
Entity normalization
Assertion Status Detection
De-identification
Relation Extraction
Spell checking & correction
note: If you are going to use any pretrained licensed NER model, you don't need to install licensed libray. As long as you have the AWS keys and license keys in your environment, you will be able to use licensed NER models with Spark NLP public library. For the other licensed pretrained models like AssertionDL, Deidentification, Entity Resolvers and Relation Extraction models, you will need to install Spark NLP Enterprise as well.
The library offers access to several clinical and biomedical transformers: JSL-BERT-Clinical, BioBERT, ClinicalBERT, GloVe-Med, GloVe-ICD-O. It also includes over 50 pre-trained healthcare models, that can recognize the following entities (any many more):
Clinical - support Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections
Drugs - support Name, Dosage, Strength, Route, Duration, Frequency
Getting started
Install Spark NLP for Healthcare
Setup AWS-CLI Credentials for licensed pretrained models
Start Spark NLP for Healthcare Session from Python
Spark NLP for Healthcare Cheat Sheet
Install Spark NLP for Healthcare on Databricks
Install Spark NLP for Healthcare on GCP Dataproc
Google Colab Notebook


nlp.johnsnowlabs.com/licensed/api/python/reference/autosummary/sparknlp_jsl.annotator.html
John Snow LABS
Getting Started
API Reference
Search the docs ...
sparknlp_jsl.annotator
sparknlp_jsl.annotator.AssertionDLApproach
sparknlp_jsl.annotator.AssertionDLModel
sparknlp_jsl.annotator.AssertionFilterer
sparknlp_jsl.annotator.AssertionLogRegApproach
sparknlp_jsl.annotator.AssertionLogRegModel
sparknlp_jsl.annotator.AverageEmbeddings
sparknlp_jsl.annotator.BertSentenceChunkEmbe
sparknlp_jsl.annotator.Chunk2Token
sparknlp_jsl.annotator.ChunkConverter
sparknlp_jsl.annotator.ChunkFilterer
sparknlp_jsl.annotator.ChunkFiltererApproach
sparknlp_jsl.annotator.ChunkMergeApproach
sparknlp_jsl.annotator.ChunkMergeModel
sparknlp_jsl.annotator.CommonResolverParams
sparknlp_jsl.annotator.ContextualParserApproac
sparknlp_jsl.annotator.ContextualParserModel
sparknlp_jsl.annotator.DateNormalizer
sparknlp_jsl.annotator.DeIdentification
sparknlp_jsl.annotator.DeIdentificationModel
sparknlp_jsl.annotator.DocumentLogRegClassifi
sparknlp_jsl.annotator.DocumentLogRegClassifi
sparknlp_jsl.annotator.DrugNormalizer
sparknlp_jsl.annotator.GenericClassifierApproac
sparknlp_jsl.annotator.GenericClassifierModel
sparknlp_jsl.annotator.IOBTagger
sparknlp_jsl.annotator.MedicalNerApproach
sparknlp_jsl.annotator.MedicalNerModel
sparknlp_jsl.annotator
Classes
AssertionDLApproach | Train a Assertion Model algorithm using deep learning.
AssertionDLModel | AssertionDL is a deep Learning based approach used to extract Assertion Status from extracted entities and text.
AssertionFilterer | Filters entities coming from ASSERTION type annotations and returns the CHUNKS.
AssertionLogRegApproach | Train a Assertion algorithm using a regression log model.
AssertionLogRegModel | This is a main class in AssertionLogReg family. Logarithmic Regression is used to extract Assertion Status
AverageEmbeddings
BertSentenceChunkEmbeddings | BERT Sentence embeddings for chunk annotations which take into account the context of the sentence the chunk appeared in.
Chunk2Token
ChunkConverter | Convert chunks from regexMatcher to chunks with a entity in the metadata.
ChunkFilterer | Model that Filters entities coming from CHUNK annotations. Filters can be set via a white list of terms or a regular expression.
ChunkFiltererApproach | Model that Filters entities coming from CHUNK annotations. Filters can be set via a white list of terms or a regular expression.
ChunkMergeApproach | Merges two chunk columns coming from two annotators(NER, ContextualParser or any other annotator producing chunks).


nlp.johnsnowlabs.co...
Update
Spark NLP 3.3.0 ScalaDoc
Search
com.johnsnowlabs.nlp.annotators.ner
MedicalNerApproach
Companion object MedicalNerApproach
class MedicalNerApproach extends AnnotatorApproach[MedicalNerModel] with NerApproach[MedicalNerApproach] with Logging with ParamsAndFeaturesWritable with CheckLicense
Linear Supertypes
Filter all members
anno
val inputAnnotatorTypes: Array[String]
Input annotator types : DOCUMENT, TOKEN, WORD_EMBEDDINGS
val outputAnnotatorType: String
Input annotator types : NAMED_ENTITY
getParam
def getBatchSize: Int
Batch size
def getConfigProtoBytes: Option[Array[Byte]]
ConfigProto from tensorflow, serialized into byte array.
def getDropout: Float
Dropout coefficient
def getEnableMemoryOptimizer: Boolean
Memory Optimizer
def getEnableOutputLogs: Boolean
Whether to output to annotators log folder
def getIncludeAllConfidenceScores: Boolean
Packages
root
com
johnsnowlabs
nlp
annotators
ner
IOBTagger
MedicalNerApproach
MedicalNerModel
NamedEntityConfidence
NerChunker
NerConverterInternal
NerTaggedInternal
NerTagsEncodingInternal
ReadablePretrainedMedicalNer
ReadsMedicalNerGraph

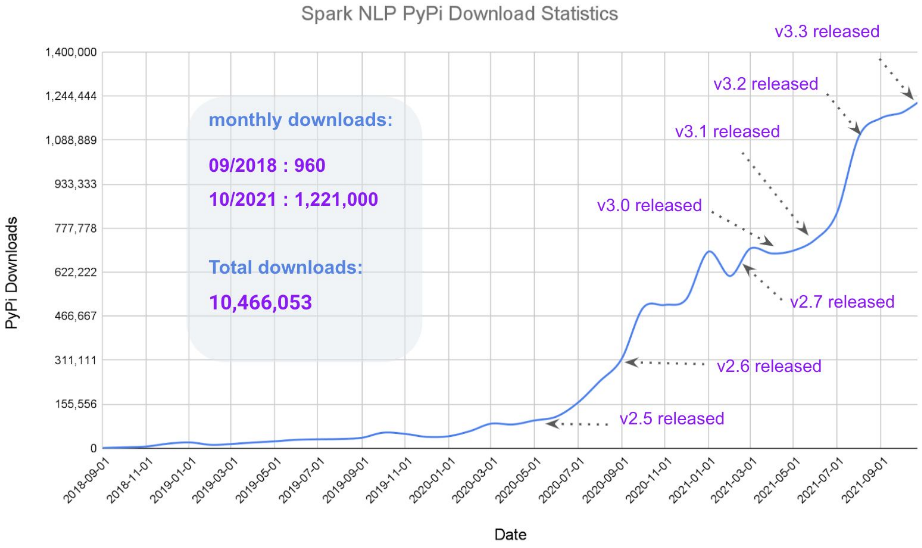
Part - I

- Overview and key concepts in Spark NLP
- NLP basics & review
- Common medical NLP use cases
- Clinical named entity recognition

Introducing Spark NLP

Daily ~ 50K
Monthly ~ 1.4M

PyPI link	https://pypi.org/project/spark-nlp
Total downloads	14,745,785
Total downloads - 30 days	1,398,927
Total downloads - 7 days	327,429



- 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)
  - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
  - Full-time development team (a new release every other week)

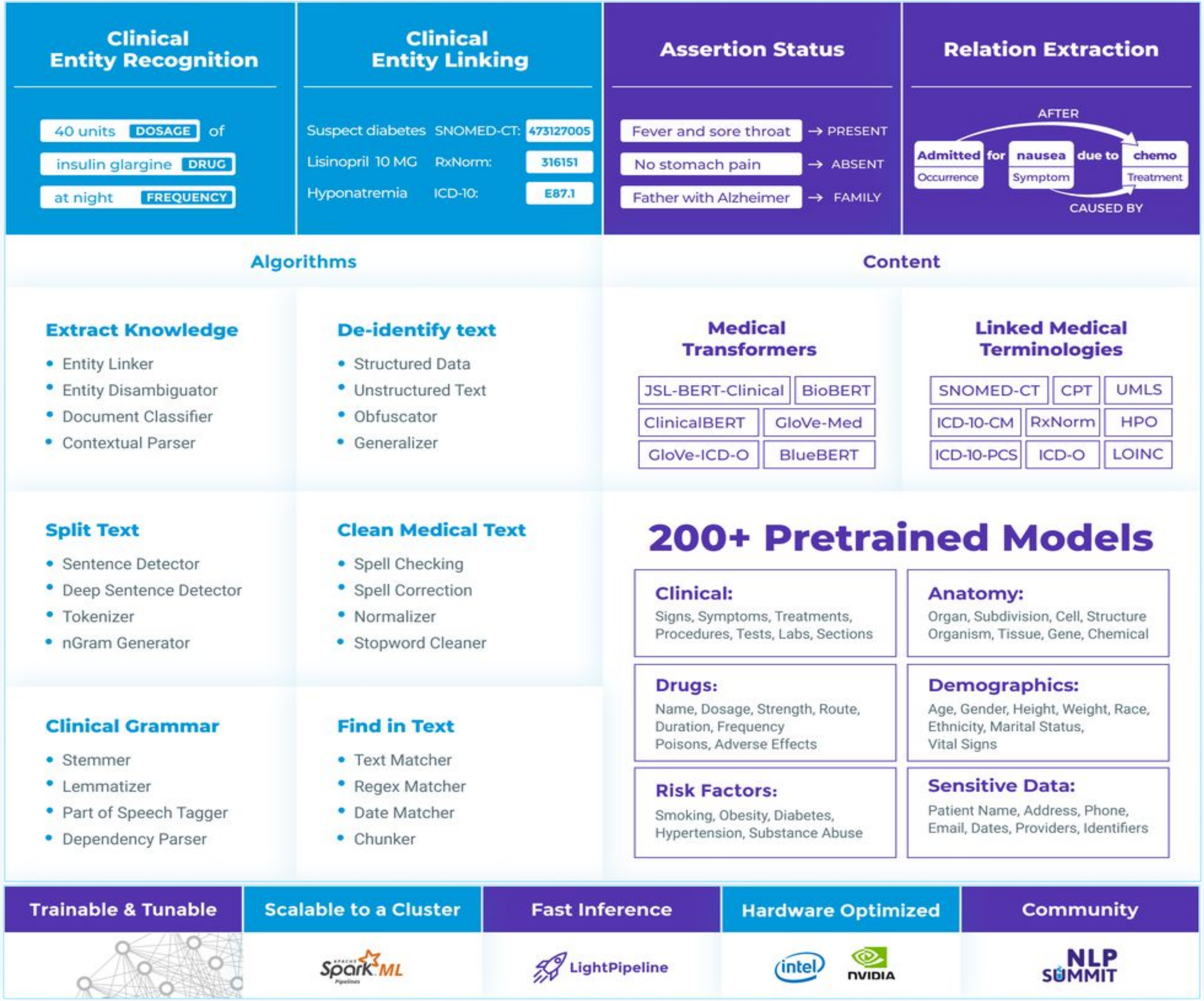
John Snow LABS

Spark NLP for Healthcare

Spark NLP for Healthcare aims to bridge this gap by providing

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

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



Academic Activities & Benchmarks

John Snow LABS

nature COMMUNICATIONS

Stanford University

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

Khushbu Agarwal[1], Sutanay Choudhury*[1], Sindhu Tipirneni[2], Pritam Mukherjee[3], Colby Ham[1], Suzanne Tamang[4], Matthew Baker[5], Siyi Tang[5], Veysel Kocaman[7], Olivier Gevaert[3,4], Robert Rallo[1], and Chandan K Reddy[2]

[1]Pacific Northwest National Laboratory, Richland, 99354, USA
[2]Department of Computer Science, Virginia Tech, Arlington, 22203, USA
[3]Stanford Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, 94305, USA
[4]Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA
[5]Department of Electrical Engineering, Stanford University, Stanford, 94305, USA
[6]Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, 94305, USA
[7]John Snow Labs, Delaware City, 19958, USA
Sutanay.Choudhury@pnnl.gov

AMIA

INFORMATICS PROFESSIONALS. LEADING THE WAY.

American Medical Informatics Association

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

Sutanay Choudhury[1], Khushbu Agarwal[1], Colby Ham[1], Pritam Mukherjee[2], Siyi Tang[2], Sindhu Tipirneni[3], Chandan Reddy[3], Suzanne Tamang[2], Robert Rallo[1], Veysel Kocaman[4]
[1]Pacific Northwest National Laboratory; [2]Stanford University; [3]Virginia Tech.; [4]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 studies[1], 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 re- ... 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-

Mining Adverse Drug Reactions from Unstructured Mediums at Scale

Health Intelligence (W3PHIAI-22) at AAAI

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 $136 billion each year in the United States alone (van Der Hooft 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,

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)

Biomedical Named Entity Recognition at Scale

Veysel Kocaman
John Snow Labs Inc.
16192 Coastal Highway
Lewes, DE , USA 19958
veysel@johnsnowlabs.com

***Abstract*—Named entity recognition (NER) is a widely applicable natural language processing task and building block of question answering, topic modeling, information retrieval, etc. In the medical domain, NER plays a crucial role by extracting meaningful chunks from clinical notes and reports, which are then fed to downstream tasks like assertion status detection, entity resolution, relation extraction, and de-identification. Reimplementing a Bi-LSTM-CNN-Char deep learning architecture on top of Apache Spark, we present a single trainable NER model that obtains new state-of-the-art results on seven public biomedical benchmarks without using heavy contextual embeddings like BERT. This includes improving BC4CHEMD to 93.72% (4.1% gain), Species800 to 80.91% (4.6% gain), and JNLPBA to 81.29% (5.2% gain). In addition, this model is freely available within a production-grade code base as part of the open-source Spark NLP library; can scale up for training and inference in any Spark cluster; has GPU support and libraries for popular programming languages such as Python, R, Scala and Java; and can be extended to support other human languages with no code changes.**

I. Introduction

Electronic health records (EHRs) are the primary source of information for clinicians tracking the care of their patients. Information fed into these systems may be found in structured fields for which values are inputted electronically (e.g. laboratory test orders or results) [1] but most of the time information in these records is unstructured making it largely inaccessible

Accurate Clinical and Biomedical Named Entity Recognition at Scale

Anonymous NAACL-HLT 2021 submission

Abstract

Named entity recognition (NER) is one of the most important building blocks of NLP tasks in the medical domain by extracting meaningful chunks from clinical notes and reports, which are then fed to downstream tasks like assertion status detection, entity resolution, relation extraction, and de-identification. Due to the growing volume of healthcare data in unstructured format, an increasingly important challenge is providing high accuracy implementations of state-of-the-art deep learning (DL) algorithms at scale. In this study, we introduce a production-grade clinical and biomedical NER algorithm based on a modified BiLSTM-CNN-Char DL architecture built on top of Apache Spark. This algorithm establishes new state-of-the-art accuracy on 7 of 8 well-known biomedical NER benchmarks and 3 clinical concept extraction challenges: 2010 i2b2/VA clinical concept extraction, 2014 n2c2 de-identification, and 2018 n2c2 medication extraction. Moreover, clinical NER models trained using this implemen-

Spark NLP: Natural Language Understanding at Scale

Veysel Kocaman, David Talby
John Snow Labs Inc.
16192 Coastal Highway
Lewes, DE , USA 19958
eysel, david}@johnsnowlabs.com

Improving Clinical Document Understanding on COVID-19 Research with Spark NLP

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

Abstract

Following the global COVID-19 pandemic, the number of scientific papers studying the virus has grown massively, leading to increased interest in automated literate review. We present a clinical text mining system that improves on previous efforts in three ways. First, it can recognize over 100 different entity types including social determinants of health, anatomy, risk factors, and adverse events in addition to other commonly used clinical and biomedical entities. Second, the text processing pipeline includes assertion status detection, to distinguish between clinical facts that are present, absent, conditional, or about someone other than the patient. Third, the deep learning models used are more accurate than previously available, leveraging an integrated pipeline of state-of-the-art pre-trained named entity recognition models, and improving on the previous best performing benchmarks for assertion status detection. We illustrate extracting trends and insights - e.g. most frequent disorders and symptoms, and most common vital signs and EKG findings – from the COVID-19 Open Research Dataset (CORD-19). The system is built using the Spark NLP library which natively supports scaling to use distributed clusters, leveraging GPU's, configurable and reusable NLP pipelines, healthcare-specific embeddings, and the ability to train models to support new entity types or human languages with no code changes.

be found in structured fields for which values are inputted electronically (e.g. laboratory test orders or results) (Liede et al. 2015) but most of the time information in these records is unstructured making it largely inaccessible for statistical analysis (Murdoch and Detsky 2013). These records include information such as the reason for administering drugs, previous disorders of the patient or the outcome of past treatments, and they are the largest source of empirical data in biomedical research, allowing for major scientific findings in highly relevant disorders such as cancer and Alzheimer's disease (Perera et al. 2014).

A primary building block in such text mining systems is named entity recognition (NER) - which is regarded as a critical precursor for question answering, topic modelling, information retrieval, etc (Yadav and Bethard 2019). In the medical domain, NER recognizes the first meaningful chunks out of a clinical note, which are then fed down the processing pipeline as an input to subsequent downstream tasks such as clinical assertion status detection (Uzuner et al. 2011), clinical entity resolution (Tzitzivacos 2007) and de-identification of sensitive data (Uzuner, Luo, and Szolovits 2007) (see Figure 1). However, segmentation of clinical and drug entities is considered to be a difficult task in biomedical NER systems because of complex orthographic structures of named entities

Peer-reviewed conference papers on Spark NLP NER

TRUSTED BY

Microsoft	Google	amazon	intel
IBM	databricks	verizon	indeed
Capital One	viacom	McKESSON	MERCK
Roche	selectdata	UiPath Reboot Work.	ASCO CANCER·LINQ DISCOVERY
cnrs	Imperial College London	Georgia Tech	STANFORD UNIVERSITY

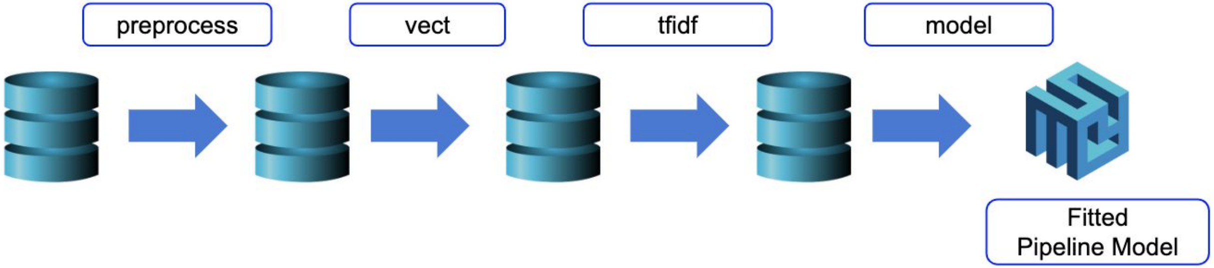
Spark NLP: Apache License 2.0

```
from pyspark.ml import Pipeline

document_assembler = 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=[
 document_assembler,
 sentenceDetector,
 tokenizer,
 normalizer,
 word_embeddings,
 ])

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

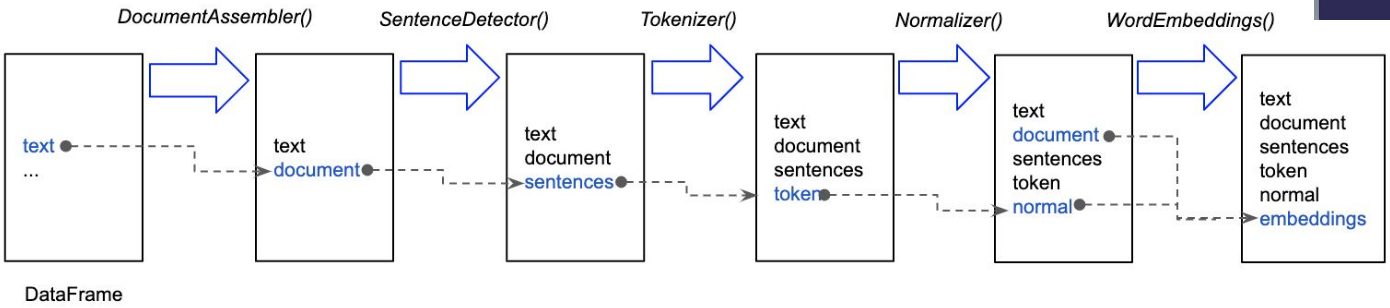
- Tokenization
- Sentence Detector
- Stop Words Removal
- Normalizer
- Stemmer
- Lemmatizer
- NGrams
- Regex Matching
- Text Matching
- Chunking
- Date Matcher
- **Part-of-speech** tagging
- **Dependency** parsing
- **Sentiment** Detection (ML models)
- **Spell Checker** (ML and DL models)
- Word **Embeddings**
- **BERT** Embeddings
- **ELMO** Embeddings
- **ALBERT** Embeddings
- **XLNet** Embeddings
- **Universal Sentence Encoder**
- **BERT** Sentence Embeddings
- Sentence Embeddings
- Chunk Embeddings
- Unsupervised **keywords extraction**
- **Language Detection** & Identification
- **Multi-class** Text Classification
- **Multi-label** Text Classification
- Multi-class **Sentiment Analysis**
- **Named entity recognition**
- Easy **TensorFlow** integration
- Full integration with Spark ML functions
- **+250 pre-trained models** in 46 languages!
- **+90 pre-trained pipelines** in 13 languages!

Introducing Spark NLP

Pipeline of annotators




```
from pyspark.ml import Pipeline

document_assembler = 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=[
 document_assembler,
 sentenceDetector,
 tokenizer,
 normalizer,
 word_embeddings,
 ])

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



Introducing Spark NLP

Spark is like a locomotive racing a bicycle. The bike will win if the load is light, it is quicker to accelerate and more agile, but with a heavy load the locomotive might take a while to get up to speed, but it's going to be faster in the end.

Faster inference

```
from sparknlp.base import LightPipeline
LightPipeline(someTrainedPipeline).annotate(someStringOrArray)
```

LightPipelines are Spark ML pipelines converted into a single machine but multithreaded task, becoming more than 10x times faster for smaller amounts of data (small is relative, but 50k sentences is roughly a good maximum).

Spark NLP in Healthcare

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

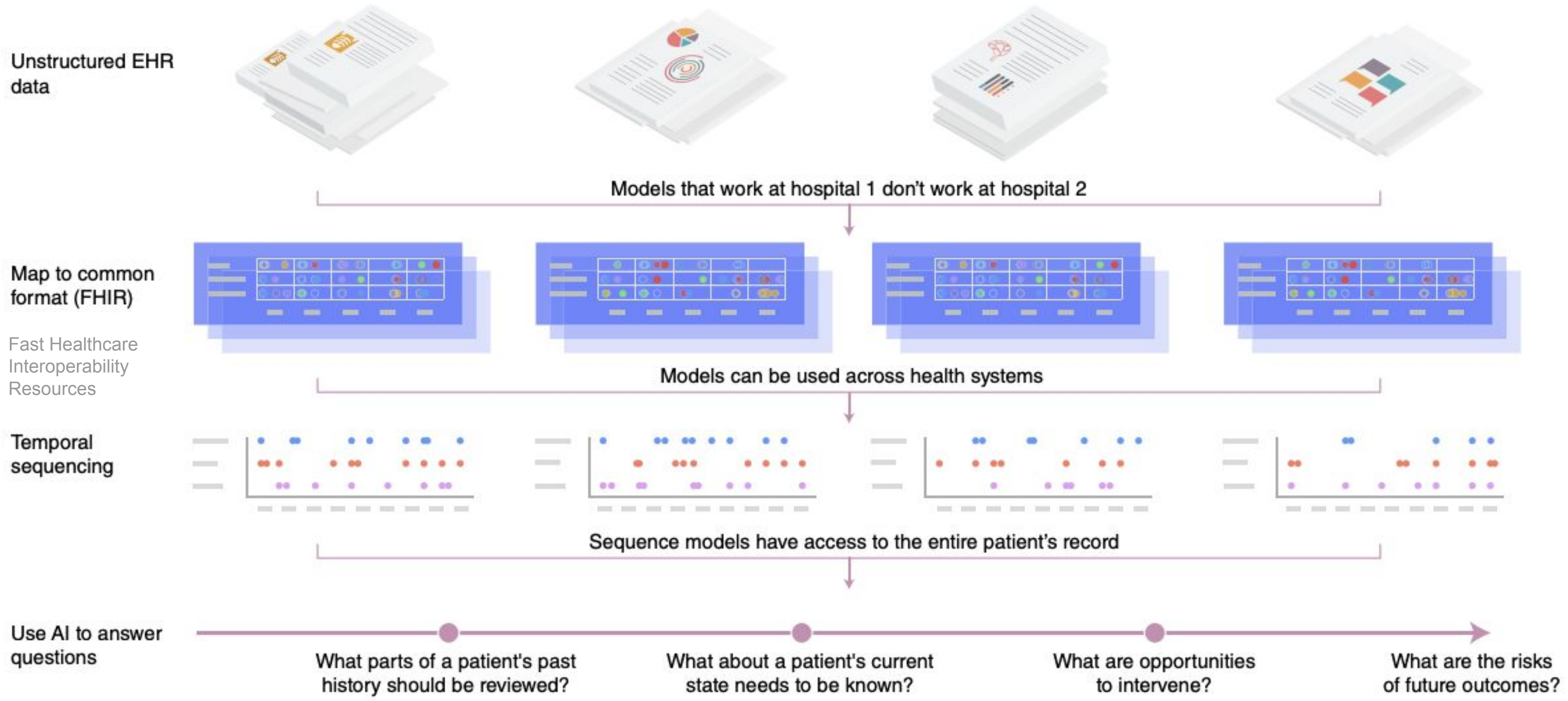
Healthcare specific needs:

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 yeatreday 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.”

↓

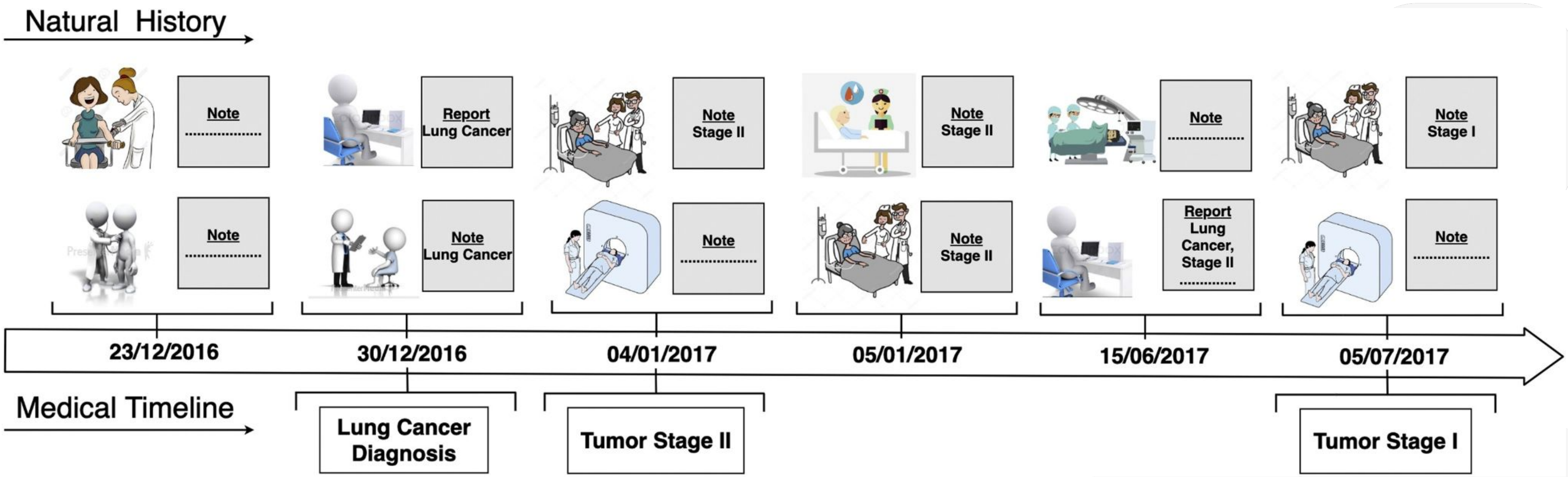
Features	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy



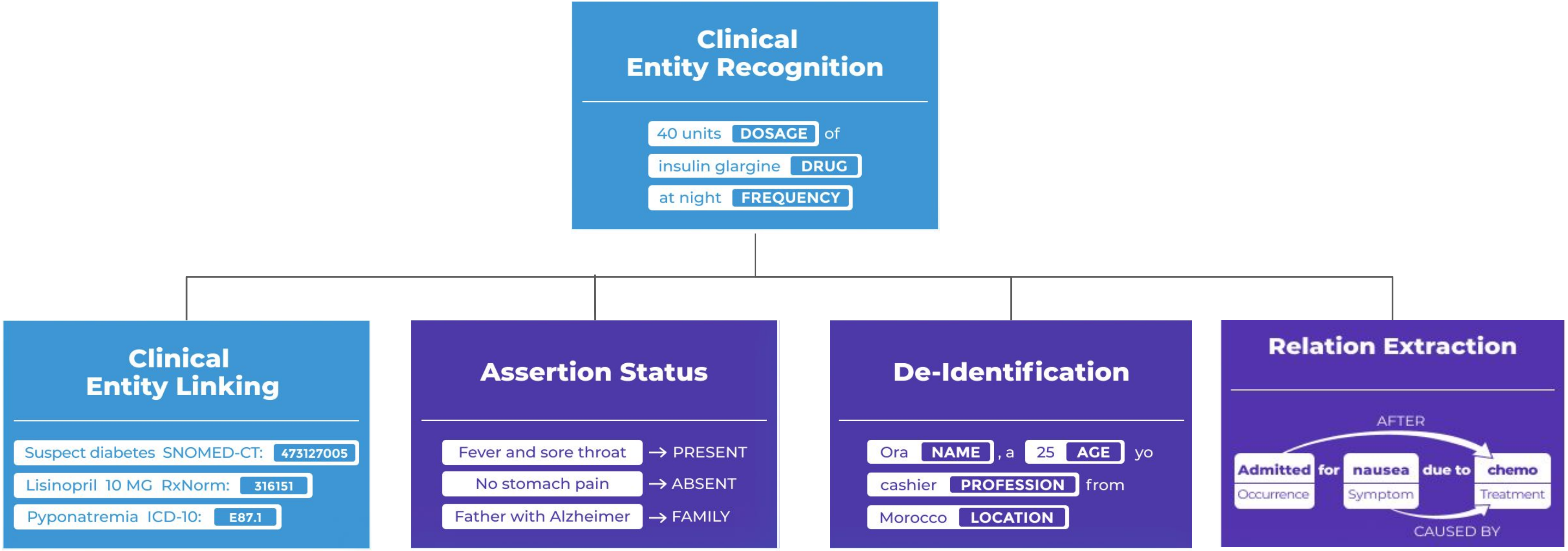
NLP SUMMIT

"Systems used to generate health data are designed for operations, not to organize data effectively for research or analytics."

Putting the clinical facts on a timeline

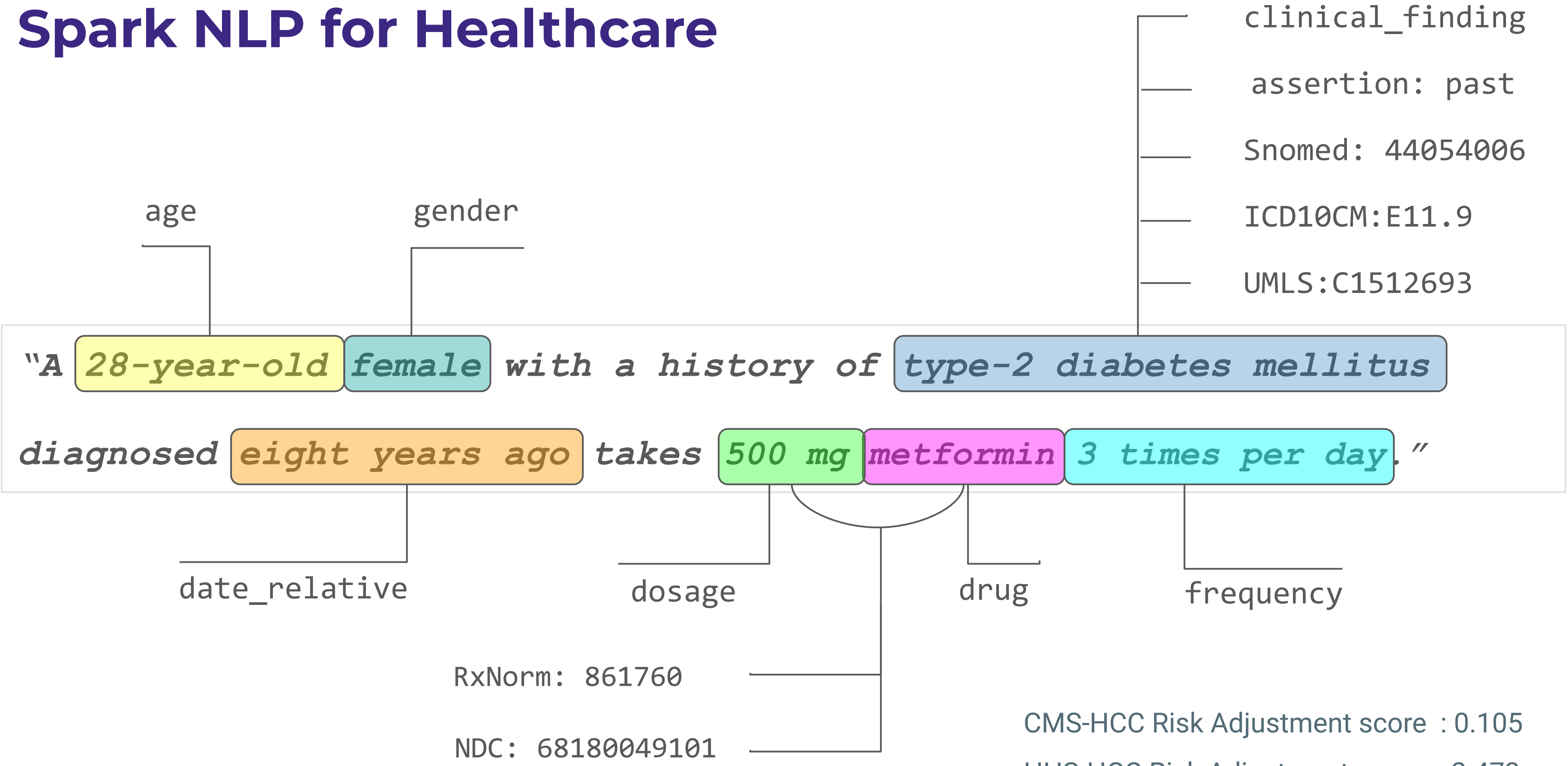


NLP in Healthcare



NLP SUMMIT

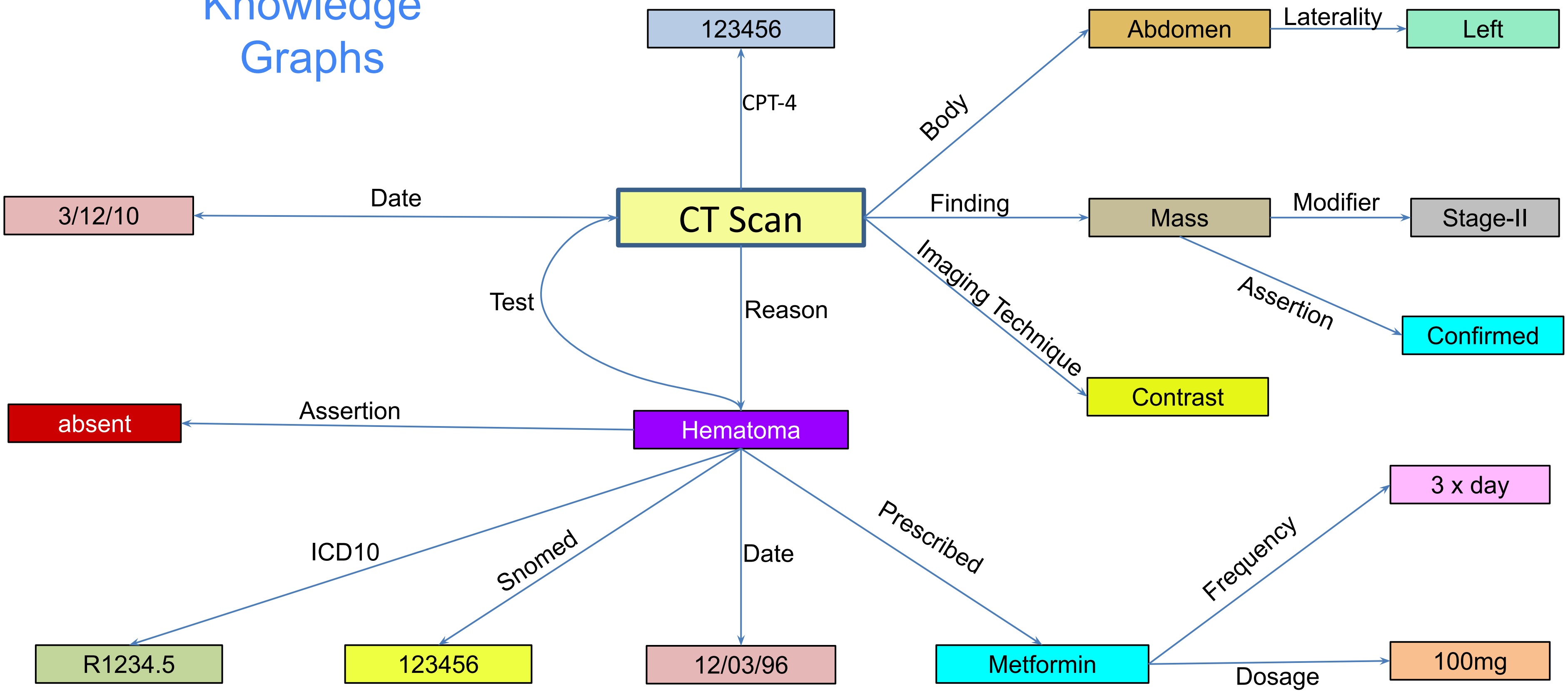
Spark NLP for Healthcare

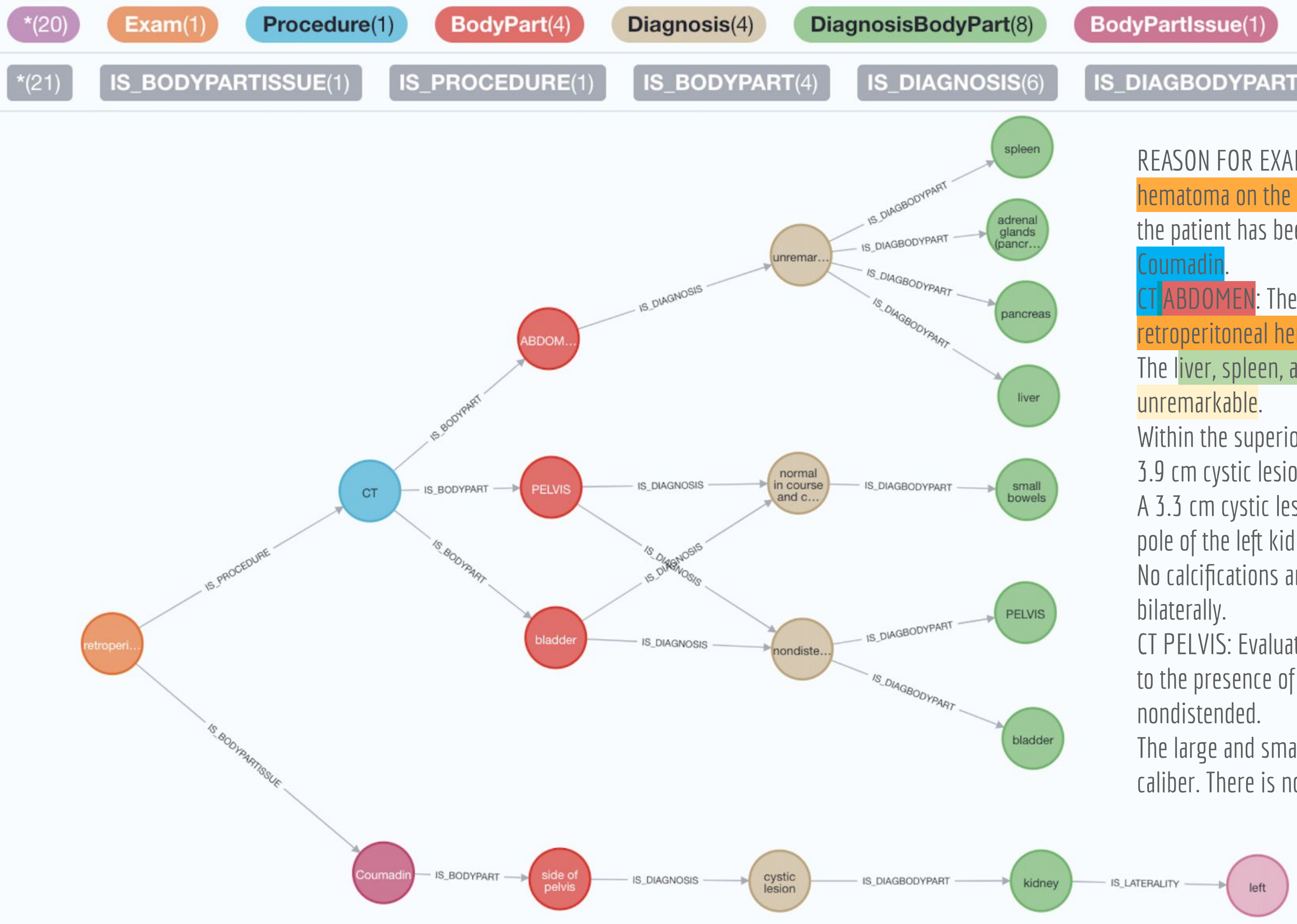


CMS-HCC Risk Adjustment score : 0.105

HHS-HCC Risk Adjustment score : 2.473

Knowledge Graphs





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.

Spark NLP for Healthcare

Named Entity Recognition

ICD10 Resolver

Snomed Resolver

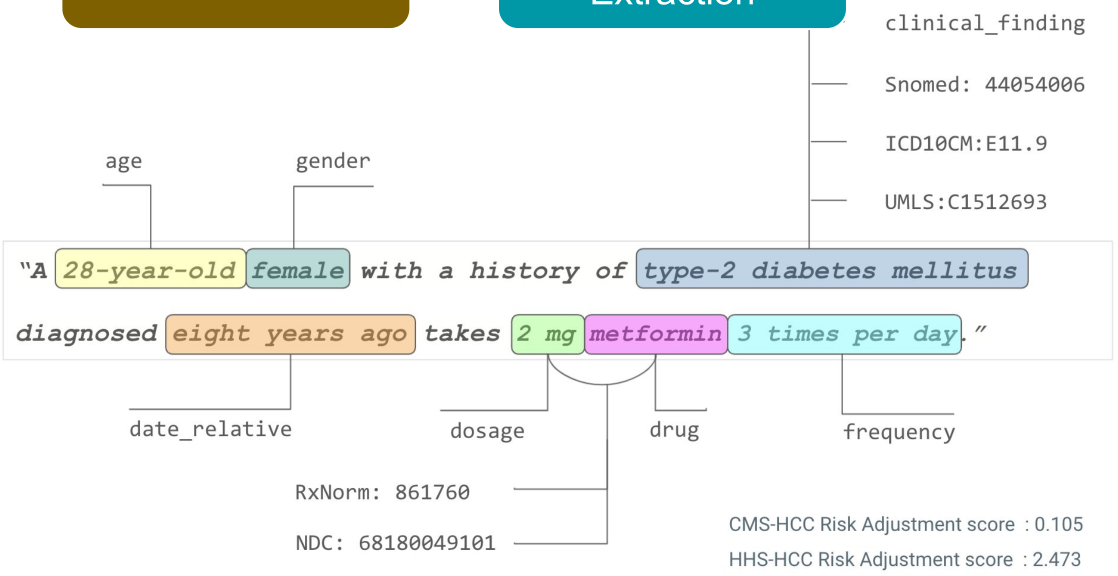
UMLS Resolver

Assertion Status Detection

Risk Adj. Module

RxNorm Resolver

Relationship Extraction



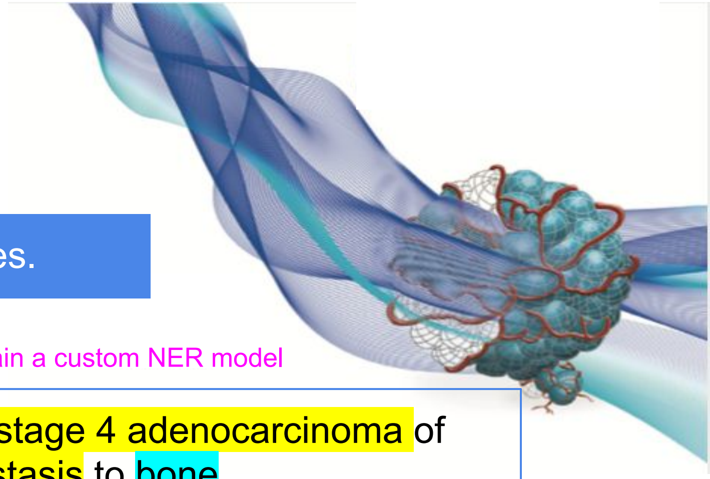
Sentence Splitter

Tokenizer

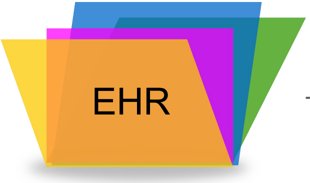
Bert Embeddings

sBert Embeddings

NLP in Healthcare



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



Extract metastasis related sentences with mets-classifier

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

- *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*

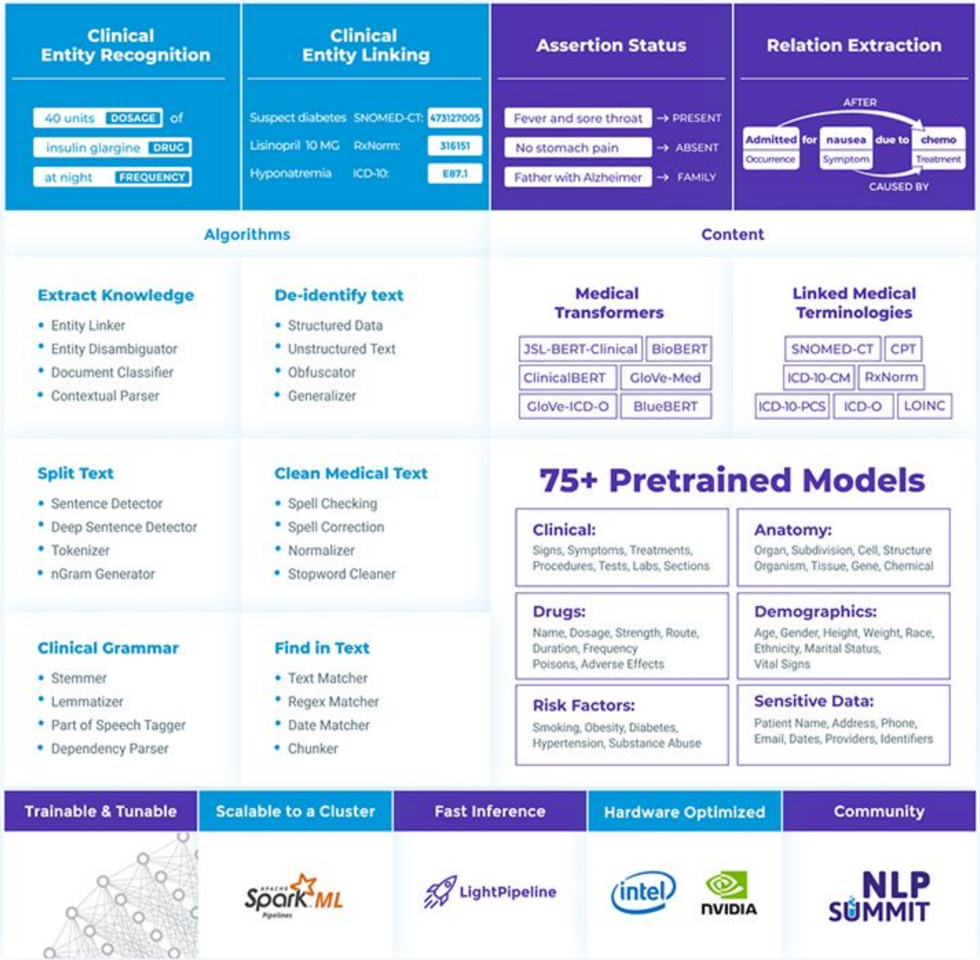
Text embeddings thru clinical word embeddings

Train a DL & ML Model

Prediction: Bone metastasis on June 2018

Clinical Named Entity Recognition (NER)

- ➢ Extract structured data from free text
- ➢ Automate record keeping & abstraction process
- ➢ Feeding downstream tasks
- ➢ Features for ML models



Pretrained NER Models

- **Clinical NER Models**

index	model_name	index	model_name	index	model_name	index	model_name
1	jsl_ner_wip_clinical	17	ner_chexpert	33	ner_deid_subentity (German)	49	ner_jsl_greedy
2	jsl_ner_wip_greedy_clinical	18	ner_clinical	34	ner_diseases_large	50	ner_jsl_slim
3	jsl_ner_wip_modifier_clinical	19	ner_clinical_icdem	35	ner_drugs	51	ner_measurements_clinical
4	jsl_rd_ner_wip_greedy_clinical	20	ner_clinical_large	36	ner_drugs_greedy	52	ner_medmentions_coarse
5	ner_ade_clinical	21	ner_clinical_large_en	37	ner_drugs_large	53	ner_posology
6	ner_ade_clinicalbert	22	ner_deid_augmented	38	ner_events_admission_clinical	54	ner_posology_experimental
7	ner_ade_healthcare	23	ner_deid_enriched	39	ner_events_clinical	55	ner_posology_greedy
8	ner_anatomy	24	ner_deid_generic_augmented	40	ner_events_healthcare	56	ner_posology_healthcare
9	ner_anatomy_coarse	25	ner_deid_generic (German)	41	ner_genetic_variants	57	ner_posology_large
10	ner_bacterial_species	26	ner_deid_large	42	ner_healthcare	58	ner_posology_small
11	ner_bionlp	27	ner_deid_sd	43	ner_human_phenotype_gene_clinical	59	ner_profiling_clinical
12	ner_cancer_genetics	28	ner_deid_sd_large	44	ner_human_phenotype_go_clinical	60	ner_radiology
13	ner_cellular	29	ner_deid_subentity_augmented	45	ner_jsl	61	ner_radiology_wip_clinical
14	ner_chemicals	30	ner_deid_synthetic	46	ner_jsl_enriched	62	ner_risk_factors
15	ner_chemprot_clinical	31	ner_deidentify_dl	47	ner_nihss	63	ner_biomarker
16	ner_abbreviation_clinical	32	ner_deid_subentity_augmented_i2b2	48	ner_diseases	64	ner_drugprot_clinical

- **BioBert NER Models**

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

- **BertForTokenClassification Clinical NER models**

	model_name
1	bert_token_classifier_ner_ade
2	bert_token_classifier_ner_clinical
3	bert_token_classifier_ner_deid
4	bert_token_classifier_ner_drugs
5	bert_token_classifier_ner_jsl
6	bert_token_classifier_ner_jsl_slim
7	bert_token_classifier_ner_bionlp
8	bert_token_classifier_ner_bacteria
9	bert_token_classifier_ner_anatomy
10	bert_token_classifier_ner_cellular
11	bert_token_classifier_ner_chemprot
12	bert_token_classifier_ner_chemicals
13	bert_token_classifier_drug_development_trials

Approach	embeddings	# of models
BiLSTM-CNN-Char	Clinical (glove)	64
BiLSTM-CNN-Char	Biobert	22
Bert for Token Cls.	Biobert	13
Total		**99**

NER JSL

Let's show an example of `ner_jsl` 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.

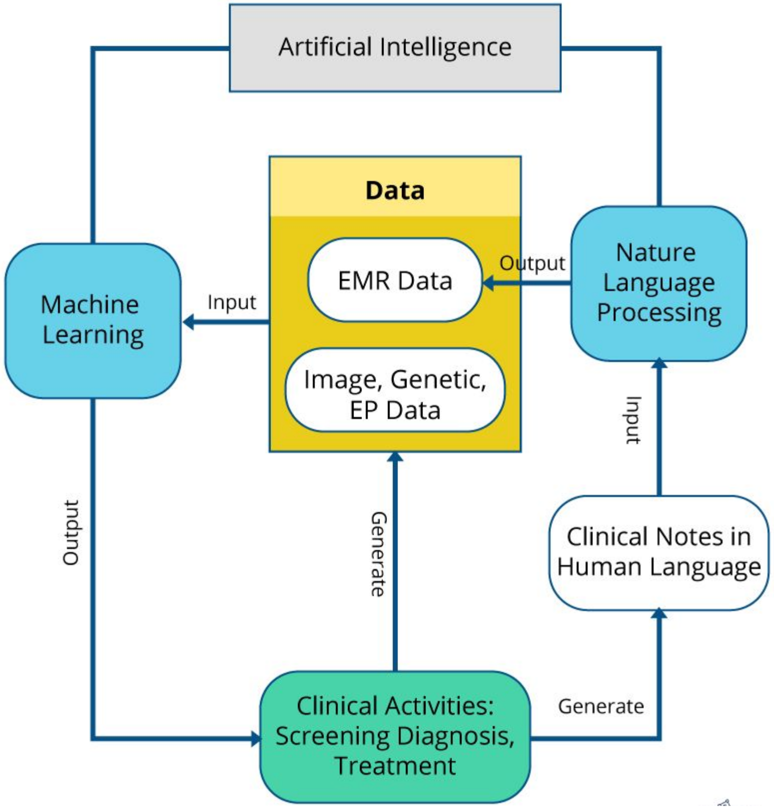
NLP in Healthcare

"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)



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, Pathological_formation, 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

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

		Spark NLP Clinical Models			AWS Medical Comprehend			GCP Healthcare API		
Entity	Sample	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

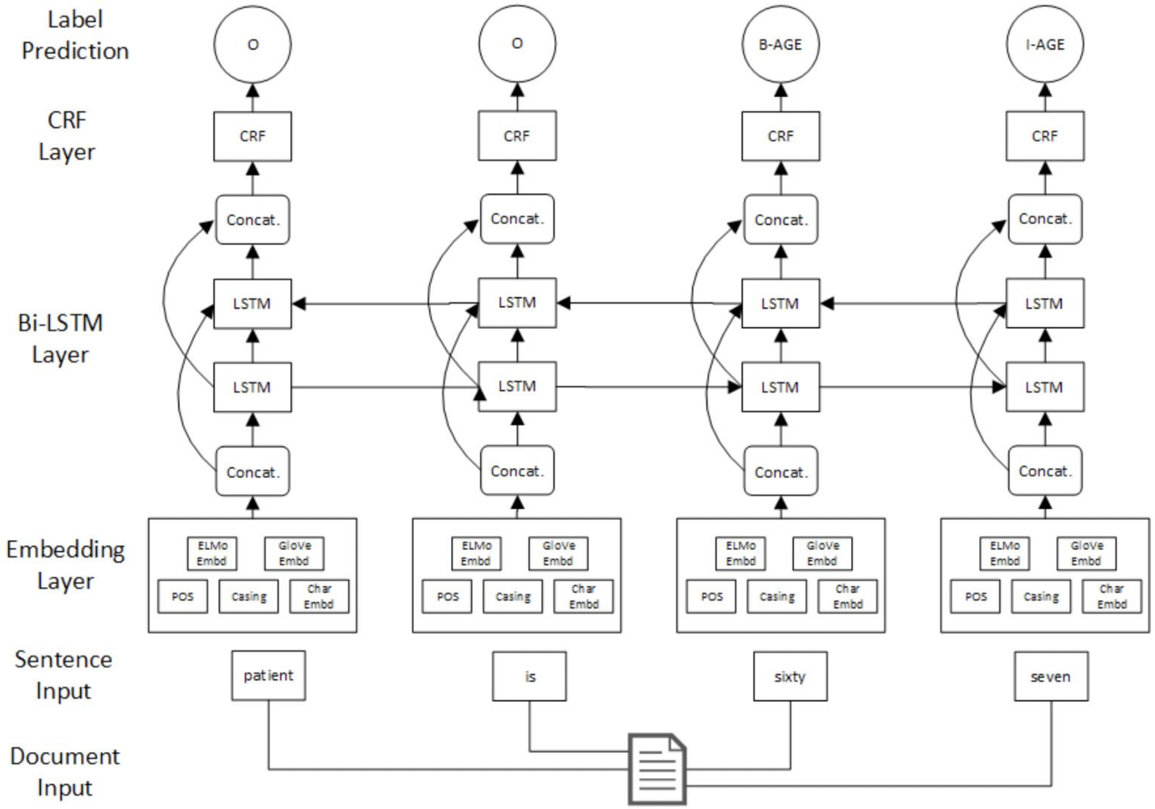
Spark NLP vs Spacy vs Stanza

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



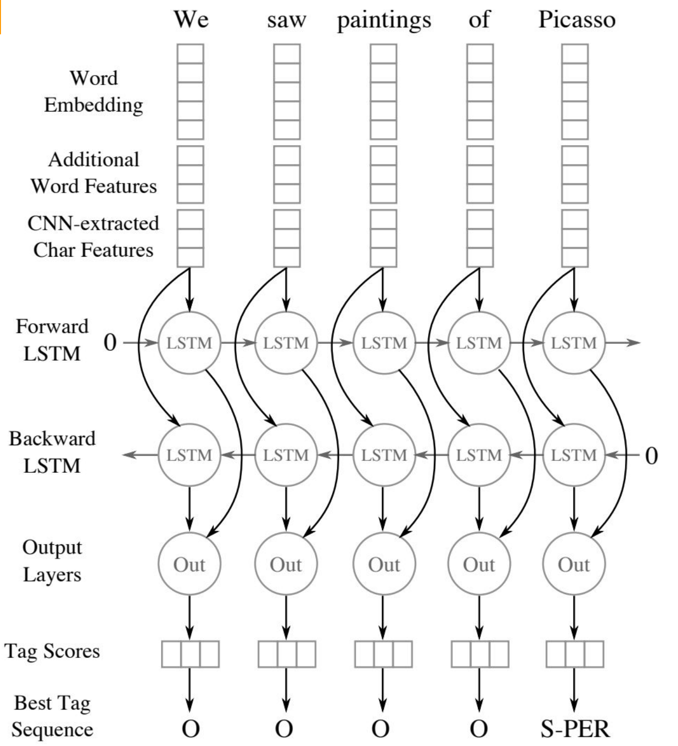


Figure 1: The (unrolled) BLSTM for tagging named entities. Multiple tables look up word-level feature vectors. The CNN (Figure 2) extracts a fixed length feature vector from character-level features. For each word, these vectors are concatenated and fed to the BLSTM network and then to the output layers (Figure).

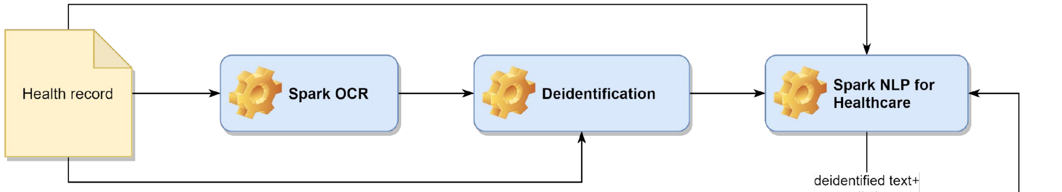
BIO schema

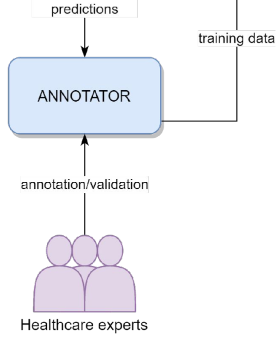
```
John     B-PER
Smith    I-PER
lives    O
in       O
New      B-LOC
York     I-LOC
```

John Smith ⇒ PERSON
New York ⇒ 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





- Close coordination with annotators
- Inter annotator agreements
- CoNLL preparation (tokenization and schema)

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.

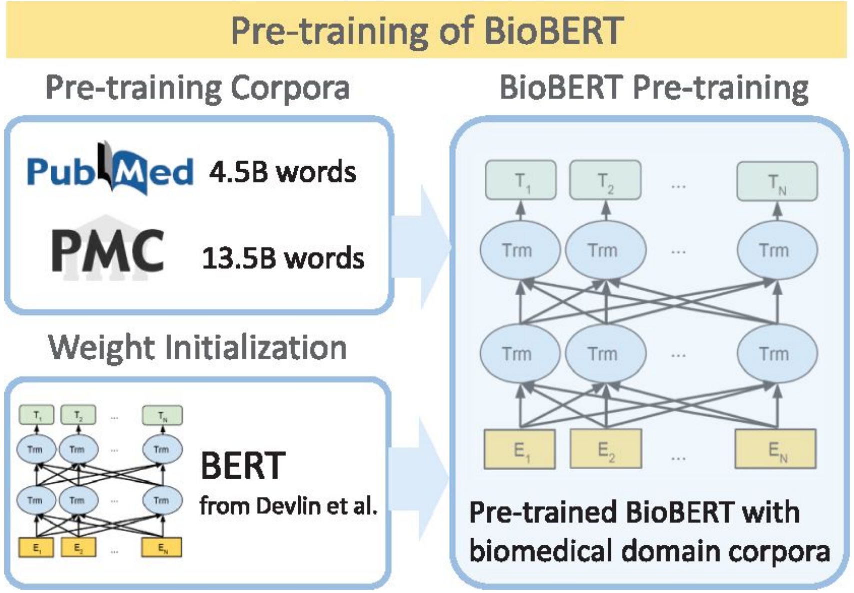
Patient denies any family history of cancer.

Clinical Word/Sentence Embeddings

Clinical Glove (200d)	ICDO Glove (200d)	Sent BERT	Bio/Clinical BERT
PubMed + PMC	PubMed + ICD10 UMLS + MIMIC III	BioBert finetuned on NLI and MedNLI	Fine tuned Pubmed + PMC + Discharge summaries

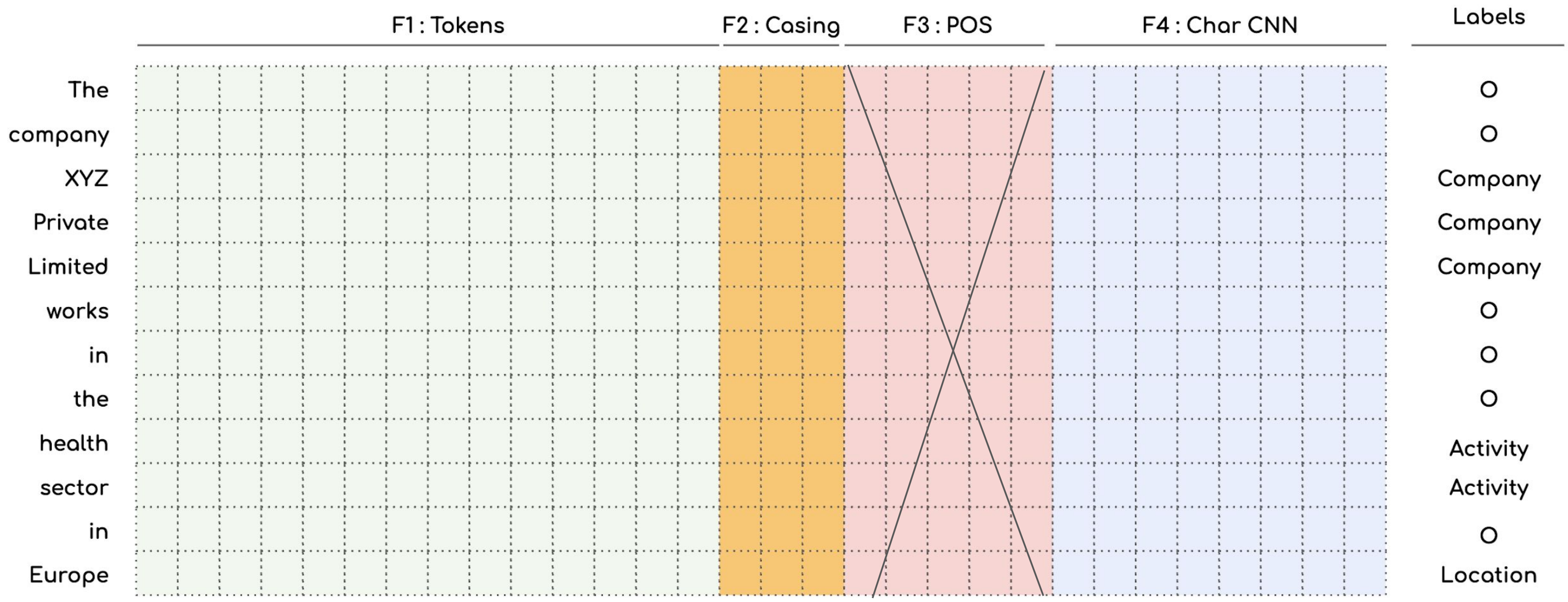
PubMed abstracts and PMC full-text articles

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



NER-DL in Spark NLP

Char-CNN-BiLSTM



Part - II

- Assertion Status detection

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 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

	model_name	Predicted Entities
1	assertion_dl	Present, Absent, Possible, Planned, Someoneelse, Past, Family, None, Hypotetical
2	assertion_dl_biobert	absent, present, conditional, associated_with_someone_else, hypothetical, possible
3	assertion_dl_healthcare	absent, present, conditional, associated_with_someone_else, hypothetical, possible
4	assertion_dl_large	hypothetical, present, absent, possible, conditional, associated_with_someone_else
5	assertion_dl_radiology	Confirmed, Suspected, Negative
6	assertion_jsl	Present, Absent, Possible, Planned, Someoneelse, Past, Family, None, Hypotetical
7	assertion_jsl_large	present, absent, possible, planned, someoneelse, past
8	assertion_ml	Hypothetical, Present, Absent, Possible, Conditional, Associated_with_someone_else

Part - III

- Entity Resolution (ICD1-, RxNorm, Snomed, etc.)

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

NLP SUMMIT

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
- sbertresolve_icd10cm_slim_billable_hcc_med
- sbiobertresolve_icd10cm_generalised

RxNorm

- sbiobertresolve_rxnorm
- demo_sbiobertresolve_rxnorm
- sbiobertresolve_rxnorm_dispo
- sbiobertresolve_rxnorm_disposition
- sbertresolve_rxnorm_disposition
- sbiobertresolve_rxnorm_ndc

LOINC

- sbluebertresolve_loinc
- sbiobertresolve_loinc

CPT

- sbiobertresolve_cpt
- sbiobertresolve_cpt_procedures_augmented
- sbiobertresolve_cpt_augmented
- sbiobertresolve_cpt_procedures_measurements_augmented

UMLS

- sbiobertresolve_umls_findings
- sbiobertresolve_umls_major_concepts
- sbiobertresolve_umls_disease_syndrome
- sbiobertresolve_umls_clinical_drugs

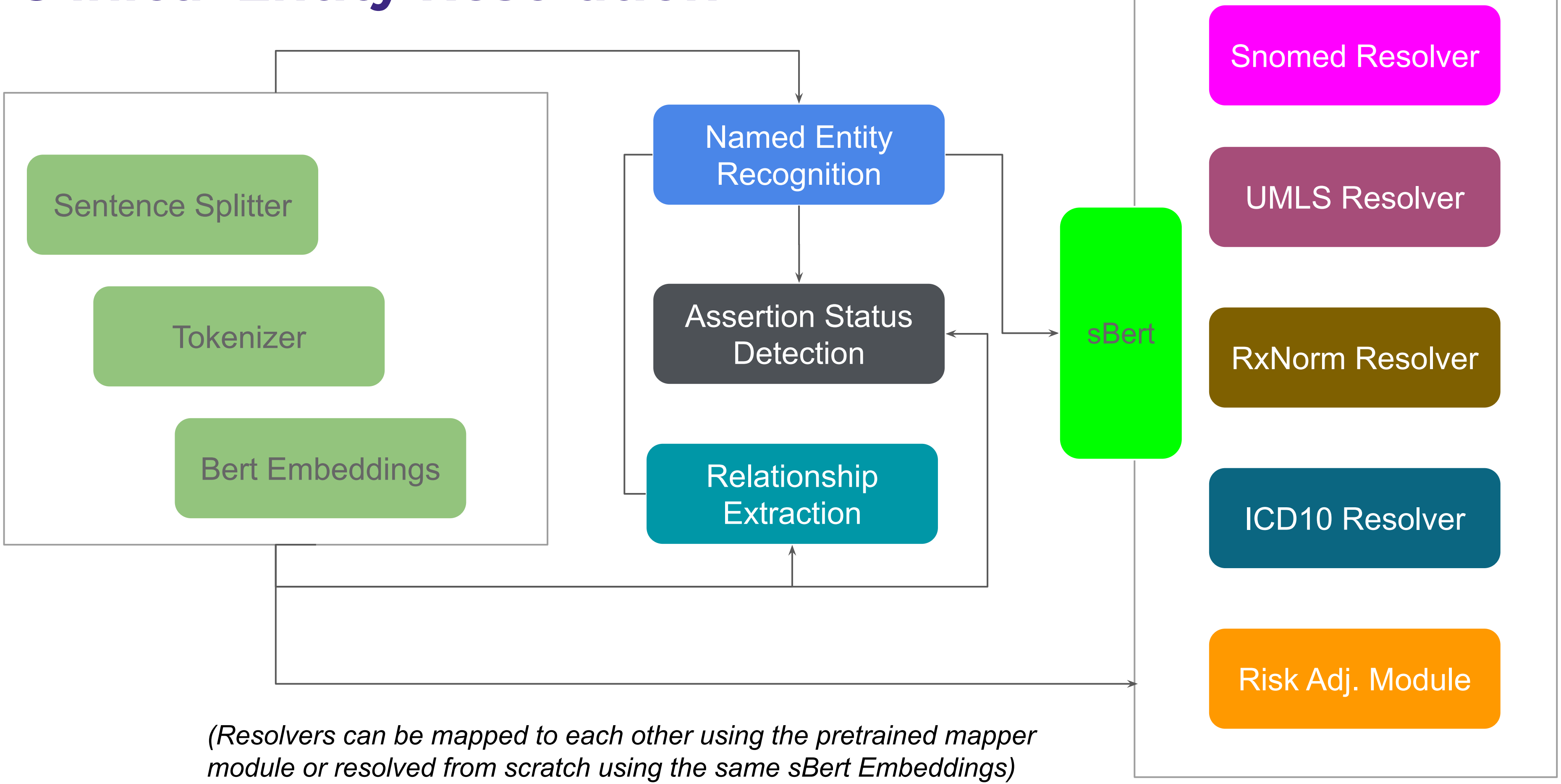
Snomed

- sbiobertresolve_snomed_auxConcepts_int
- sbiobertresolve_snomed_findings
- sbiobertresolve_snomed_findings_int
- sbiobertresolve_snomed_auxConcepts
- sbertresolve_snomed_bodyStructure_med
- sbiobertresolve_snomed_bodyStructure
- sbiobertresolve_snomed_findings_aux_concepts
- sbertresolve_snomed_conditions

mapping

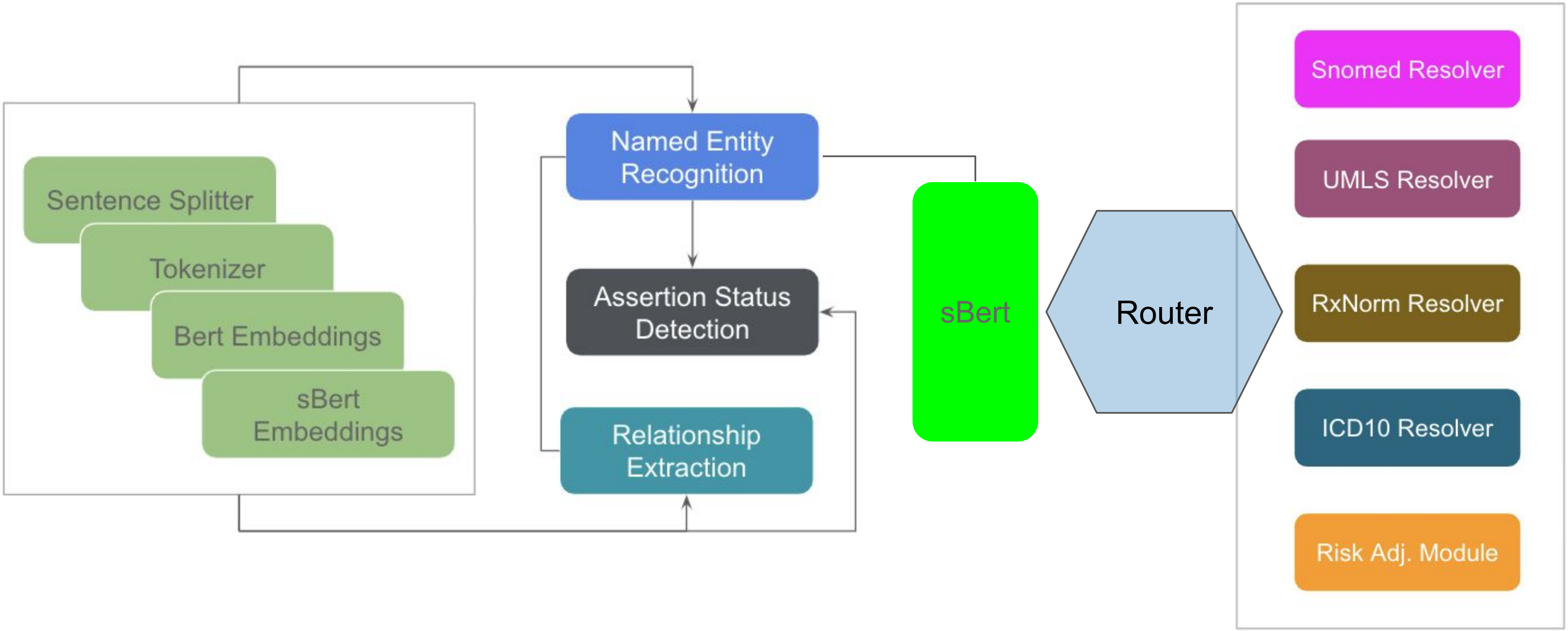
- icd10cm_snomed_mapping : ICD10 Codes to Snomed Codes
- snomed_icd10cm_mapping : Snomed Codes to ICD Codes
- icd10cm_umls_mapping : ICD Codes to UMLS Codes
- snomed_umls_mapping : Snomed Codes to UMLS Codes
- rxnorm_umls_mapping : RxNorm Codes to UMLS Codes
- mesh_umls_mapping : MeSH Codes to UMLS Codes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

Clinical Entity Resolution

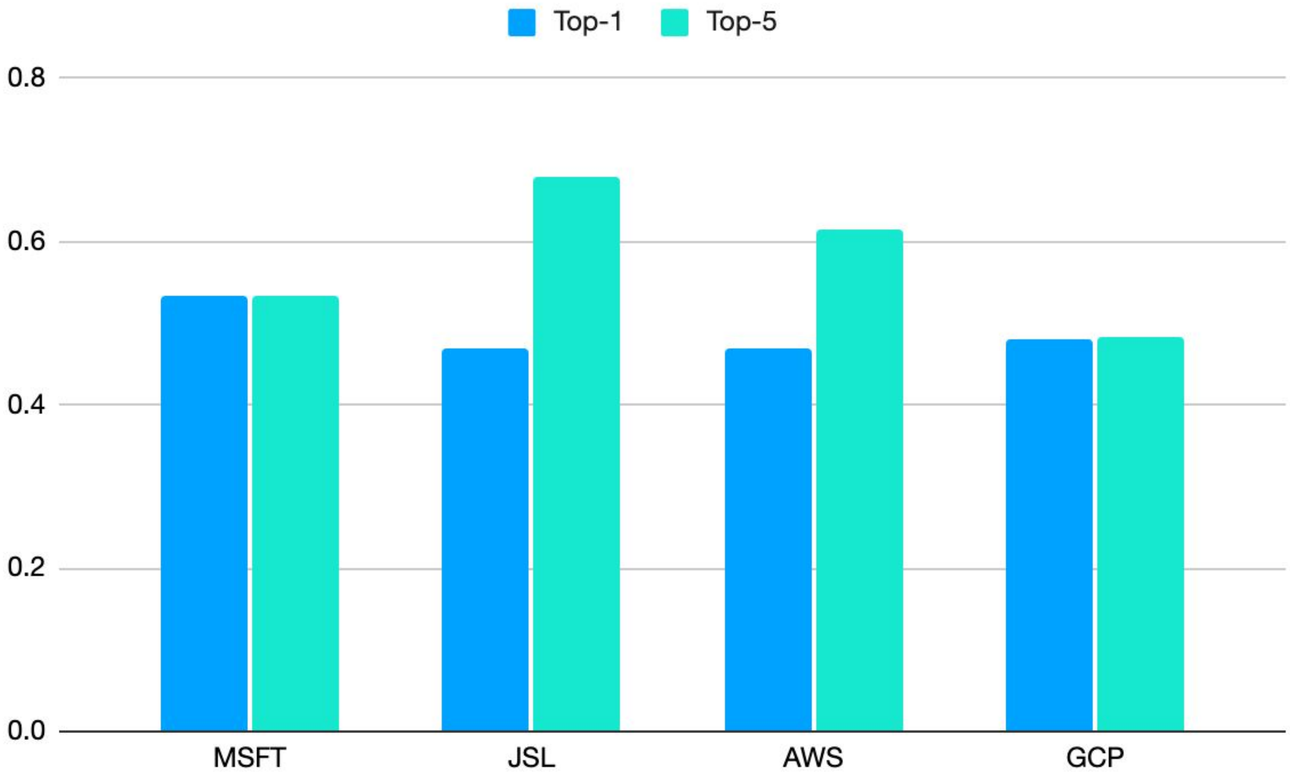


(Resolvers can be mapped to each other using the pretrained mapper module or resolved from scratch using the same sBert Embeddings)

Clinical Entity Resolution



ICD10CM Benchmark Comparison



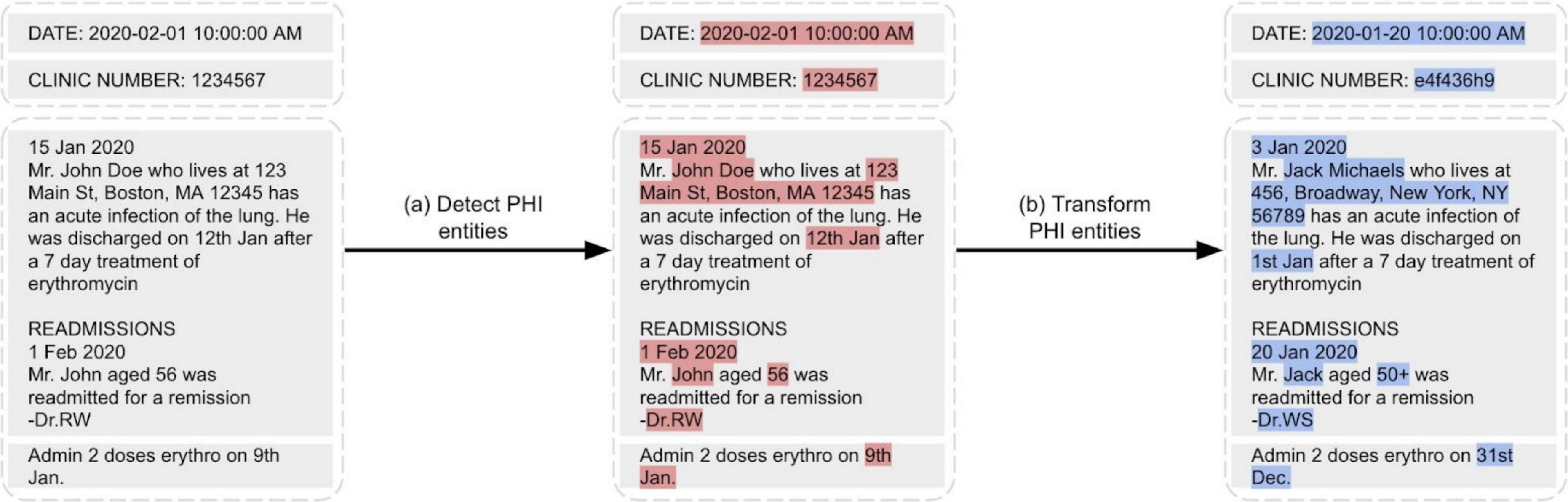
Tested on ~200 clinical concepts.

Part - IV

- De-Identification and Obfuscation of PHI data

De-Identification

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



De-Identification

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

Group Name	Included Entities
A (defined by the HIPAA Safe Harbor Implementation)	Age over 89, Phone/Fax numbers, Email addresses, Websites and URLs, IP Addresses, Dates, Social security numbers, Medical record numbers, Vehicle/Device numbers, Account/Certificate/License numbers, Health plan numbers, Biometric identifiers, Street addresses, City, Zip code, Employer names, Personal names of patients and family members
B	Group A, Doctor names, User IDs (of care providers), State
C	Group B, Hospital names, Country
D	Group C, Holidays, Days of the week, Occupations

De-Identification

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

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

Color codes:DOCTOR, HOSPITAL, DATE, STREET, MEDICALRECORD, PATIENT,

Deidentified Text

```
['A .',
 'Record date : <DATE>  , <DOCTOR> , M.D .',
 ', Name : <PATIENT> , <PATIENT> MR .',
 '# <MEDICALRECORD> Date : <DATE> PCP : <DOCTOR> , 25 month years-old , Record date : <DATE> .',
 '<HOSPITAL> .',
 '<STREET>']
```

```python
def get_deidentify_model():

    custom_ner_converter = NerConverter()\
      .setInputCols(["sentence", "token", "ner"])\
      .setOutputCol("ner_chunk")
      #.setWhiteList(entity_types)

    deidentify_pipeline = Pipeline(
            stages = [
            documentAssembler,
            sentenceDetector,
            tokenizer,
            word_embeddings,
            clinical_ner,
            custom_ner_converter,
            deidentification_rules
            ])

    empty_data = spark.createDataFrame([[""]]).toDF("text")

    model_deidentify = deidentify_pipeline.fit(empty_data)

    return model_deidentify
```

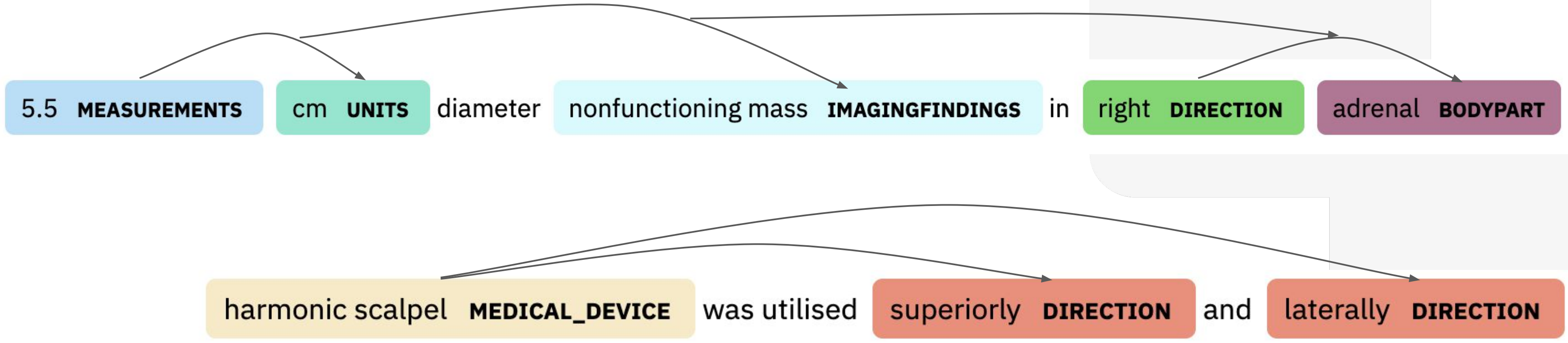
Part - V

- Relation Extraction

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

	model_name
0	re_ade_biobert
1	re_ade_clinical
2	re_bodypart_directions
3	re_bodypart_problem
4	re_bodypart_proceduretest
5	re_chemprot_clinical
6	re_clinical
7	re_date_clinical
8	re_drug_drug_interaction_clinical
9	re_human_phenotype_gene_clinical
10	re_temporal_events_clinical
11	re_temporal_events_enriched_clinical
12	re_test_problem_finding
13	re_test_result_date
14	redl_ade_biobert
15	redl_bodypart_direction_biobert
16	redl_bodypart_problem_biobert
17	redl_bodypart_procedure_test_biobert
18	redl_chemprot_biobert
19	redl_clinical_biobert
20	redl_date_clinical_biobert
21	redl_drug_drug_interaction_biobert
22	redl_human_phenotype_gene_biobert
23	redl_temporal_events_biobert

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

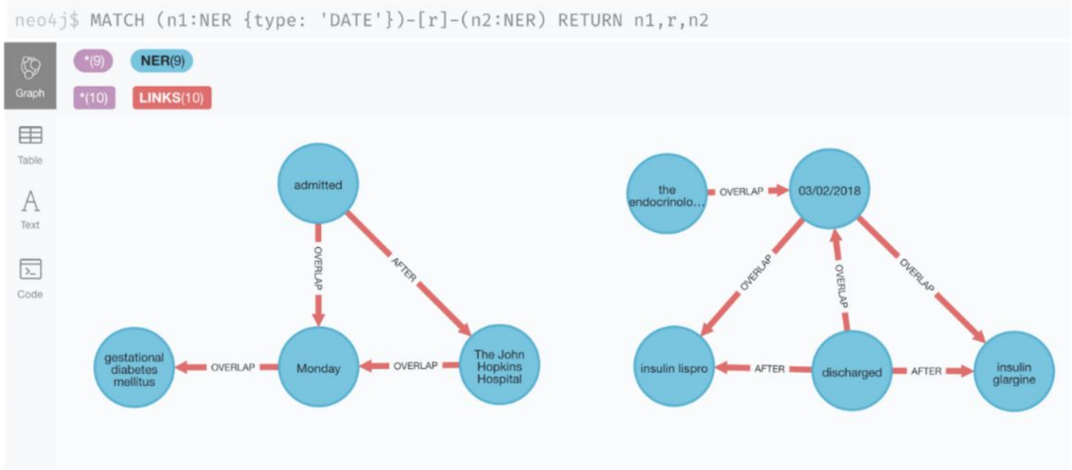
Relation	Recall	Precision	F1
OVERLAP	0.81	0.73	**0.77**
BEFORE	0.85	0.88	**0.86**
AFTER	0.38	0.46	**0.43**

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 glargine and 12 units of insulin lispro.

```
query = """
    MATCH (n1:NER {type: 'DATE'})-[r]-(n2:NER)
    RETURN n1.name AS date, r.relation AS relation, n2.name AS event
"""

df = pd.DataFrame([dict(_) for _ in conn.query(query)])
df
```

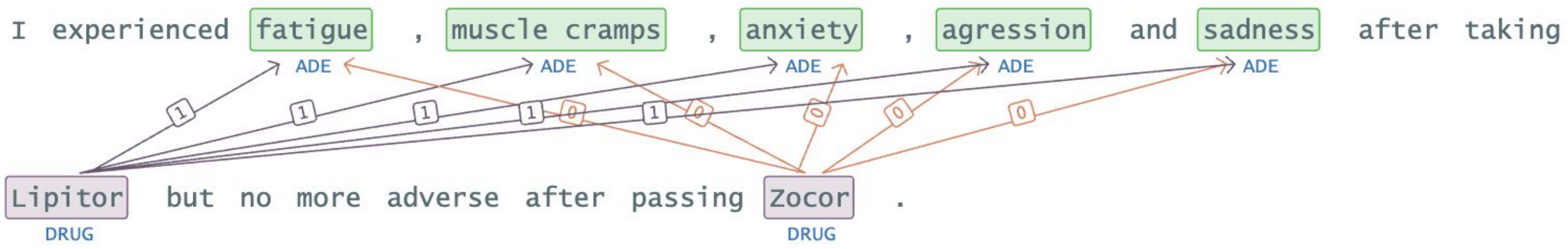
	date	relation	event
0	Monday	OVERLAP	gestational diabetes mellitus
1	Monday	OVERLAP	The John Hopkins Hospital
2	Monday	OVERLAP	admitted
3	03/02/2018	OVERLAP	insulin lispro
4	03/02/2018	OVERLAP	insulin glargine
5	03/02/2018	OVERLAP	discharged
6	03/02/2018	OVERLAP	the endocrinology service



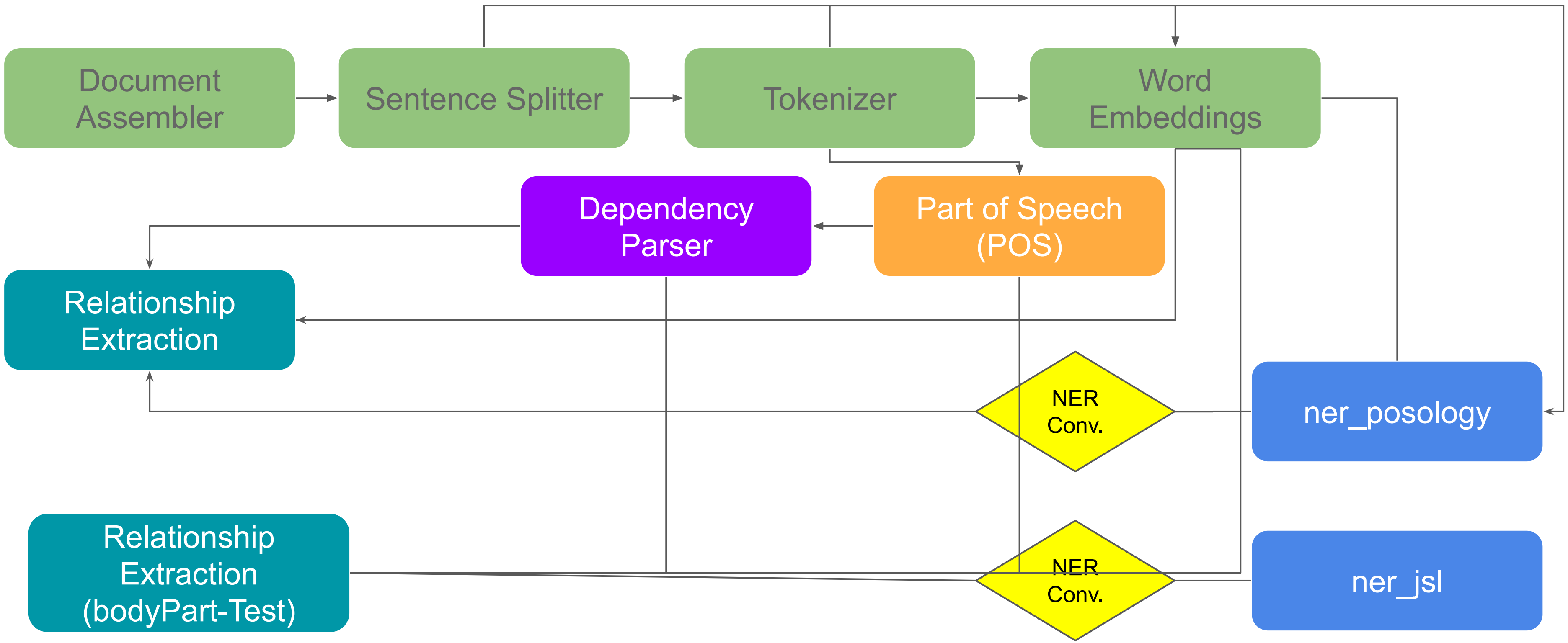
Clinical Relation Extraction

	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 - VI

- **Adverse Drug Reactions (ADR)**

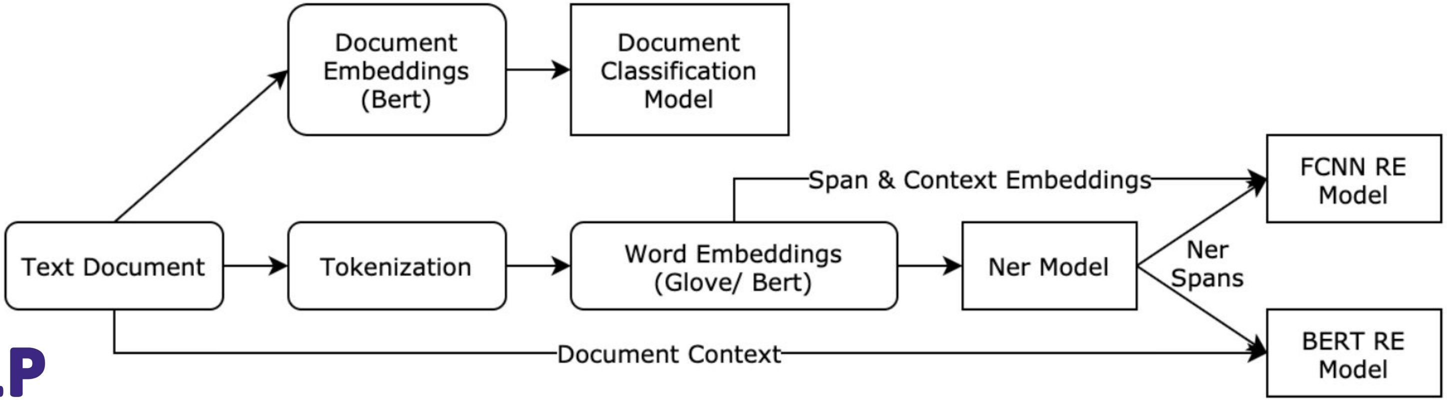
Adverse Drug Reactions (ADR)

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	-	-	-

Classification

Named Entity Recognition

Relation Extraction



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

AAAI
Association for the Advancement of Artificial Intelligence

Classification

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.

Relation Extraction

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)

NLP SUMMIT

Part - VII

❖ **Key Chunk Phrase Extractor**

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	0.095010420298254 4	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	0.029673513395379065
years prior	0.030808818777992058
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	ngram
gestational diabetes mellitus	NER
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

Thank you !

Veysel Kocaman
Principal Data Scientist
John Snow Labs

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/winning-in-health-care-ai-with-small-data/#1b2fc2555664
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-in-enterprises/
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

John Snow LABS