NLP SUMMIT HEALTHCARE

Presented by John Snow LABS

1 Line of Code to Use 600+ State-of-the-Art Clinical & Biomedical NLP Models

Christian Kasim Loan

Lead Data Scientist

John Snow Labs

Introducing Spark NLP

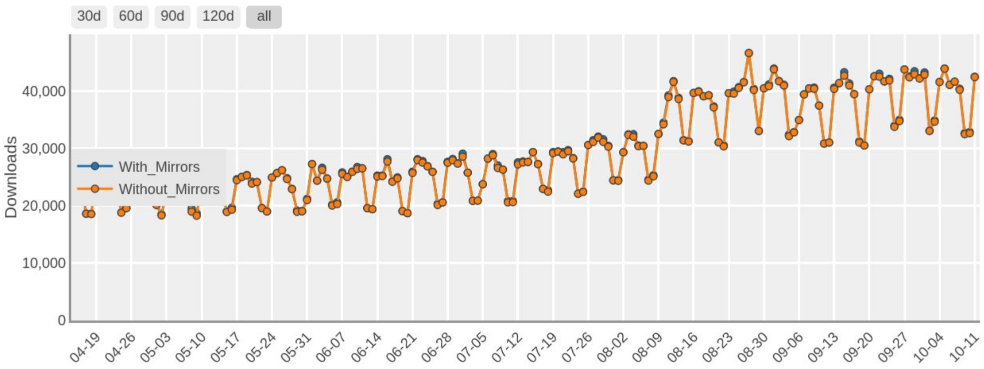
Total downloads	14,558,265
Total downloads - 30 days	1,371,781
Total downloads - 7 days	318,804

downloads 14M

downloads/month 1M

downloads/week 318k

Daily Download Quantity of spark-nlp package - Overall

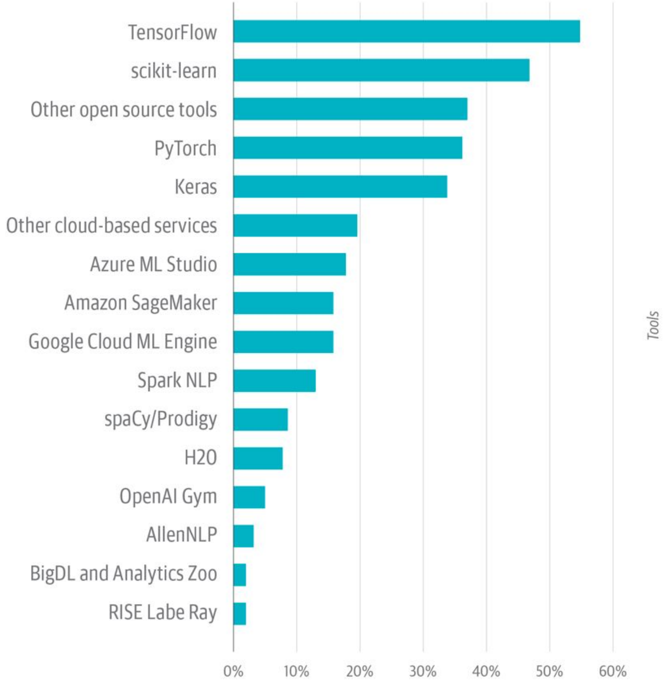


- Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. (initial release: Oct 2017)
  - A single unified solution for all your NLP needs
  - Take advantage of transfer learning and implementing the latest and greatest **SOTA** algorithms and models in NLP research
  - The most widely used NLP library in industry (5 yrs in a row)
  - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
  - Full-time development team (30 new releases in 2021, 26 new releases in 2020, 30 new releases in 2019.)

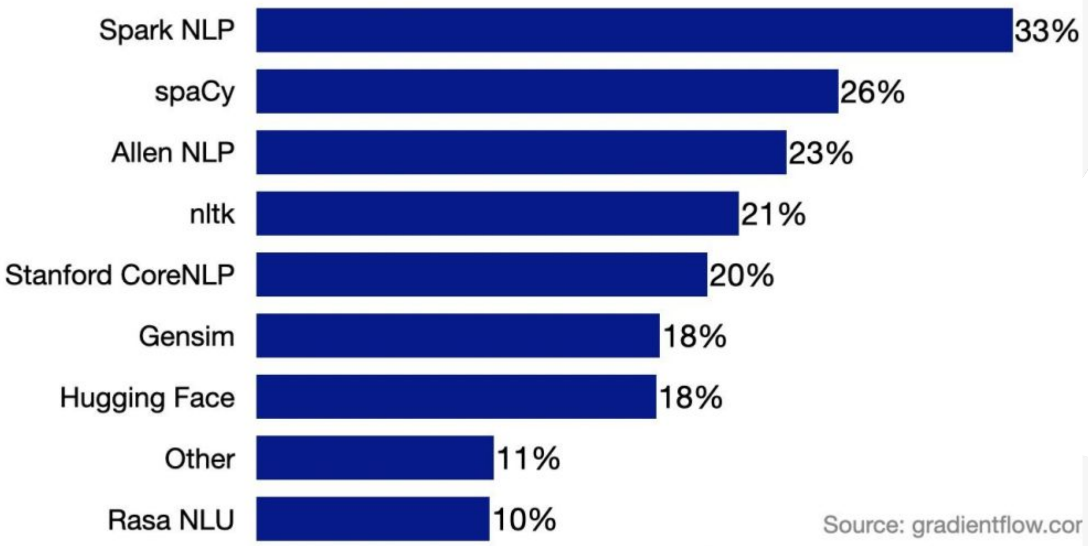
https://medium.com/spark-nlp/introduction-to-spark-nlp-foundations-and-basic-components-part-i-c83b7629ed59

Spark NLP in Industry

Which of the following AI tools do you use?



Which NLP libraries does your organization use?



NLP SUMMIT HEALTHCARE

NLP Industry Survey by Gradient Flow,
an independent data science research & insights company, September 2020

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

NLP SUMMIT HEALTHCARE

Recognized by the Technology Experts

Global 100
2021
WINNER
John Snow LABS

ghp
2021 Global
Excellence Awards

2020
corporatevision
Best Healthcare & Life Science AI Development Firm – USA

AI
ARTIFICIAL INTELLIGENCE
EXCELLENCE
AWARD
2020

International DataScience Awards
WINNER 2019

Strata Data Awards
2019
Most Significant Open Source Project
WINNER
SPARK NLP

CIO Review
ARTIFICIAL INTELLIGENCE
SOLUTION PROVIDER
OF THE YEAR - 2018

AI
PLATFORM OF THE YEAR
CIO
APPLICATIONS
2019

NLP SUMMIT HEALTHCARE

Spark NLP

- 99 total releases
- Release every two weeks for the past 4 years
- A single unified library for all your NLP/NLU need
- Active community on Slack and GitHub

NLP Feature	Spark NLP	spaCy	NLTK	CoreNLP	Hugging Face
Tokenization	Yes	Yes	Yes	Yes	Yes
Sentence segmentation	Yes	Yes	Yes	Yes	No
Steeming	Yes	Yes	Yes	Yes	No
Lemmatization	Yes	Yes	Yes	Yes	No
POS tagging	Yes	Yes	Yes	Yes	No
Entity recognition	Yes	Yes	Yes	Yes	Yes
Dep parser	Yes	Yes	Yes	Yes	No
Text matcher	Yes	Yes	No	No	No
Date matcher	Yes	No	No	No	No
Sentiment detector	Yes	No	Yes	Yes	Yes
Text classification	Yes	Yes	Yes	No	Yes
Spell checker	Yes	No	No	No	No
Language detector	Yes	No	No	No	No
Keyword extraction	Yes	No	No	No	No
Pretrained models	Yes	Yes	Yes	Yes	Yes
Trainable models	Yes	Yes	Yes	Yes	Yes
Question Answering	Yes	Yes	No	No	Yes
Text Style Transfer	Yes	No	No	No	Yes
Finance models	Yes	No	No	No	Yes
200+ Languages supported	Yes	Yes	No	No	Yes
Summarize Test	Yes	Yes	No	No	Yes
Text Generation (GPT2, T5)	Yes	Yes	No	No	Yes

NLP SUMMIT HEALTHCARE

NLP SUMMIT HEALTHCARE

OFFICIALLY SUPPORTED RUNTIMES

Java

python™

Scala

R

NVIDIA

intel XEON inside

APACHE Spark™

databricks

CLOUDERA

Azure

aws

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

NLP SUMMIT HEALTHCARE

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

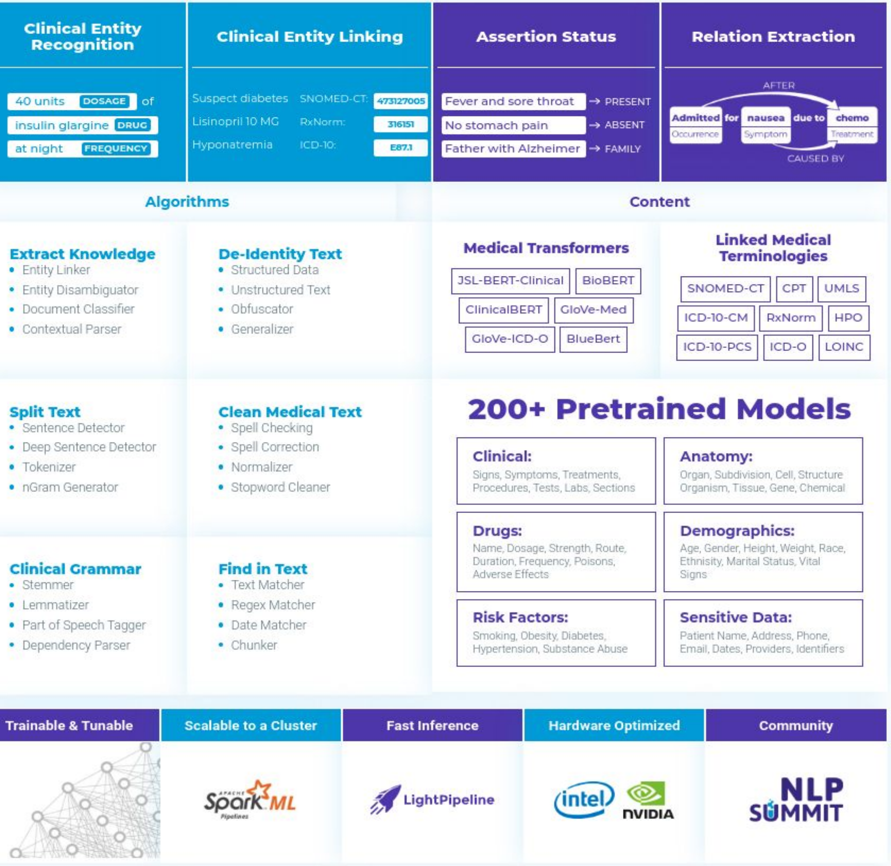
Following the global COVID-19 pandemic, the number of scientific papers studying the virus has grown massively, leading to increased interest in automated literate review. We present a clinical text mining system that improves on previous efforts in three ways. First, it can recognize over 100 different 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

Spark NLP Modules



Spark NLP Modules (Enterprise and Public)

Text Classification with Word & Sentence Embeddings & Transformers

Glove | ELMO | BERT | Electra | Longformer | Universal Sentence Encoder

Albert | XLNet | RoBerta | DistilBert | XlmRoBerta | Bert Sentence Embeddings

Word2Vec

Bert Token Classifier | DistilBert Token Classifier | Albert Token Classifier | RoBerta Token Classifier | RoBerta Token Classifier

XlmRoBerta Token Classifier | XLNet Token Classifier | LongFormer Token Classifier

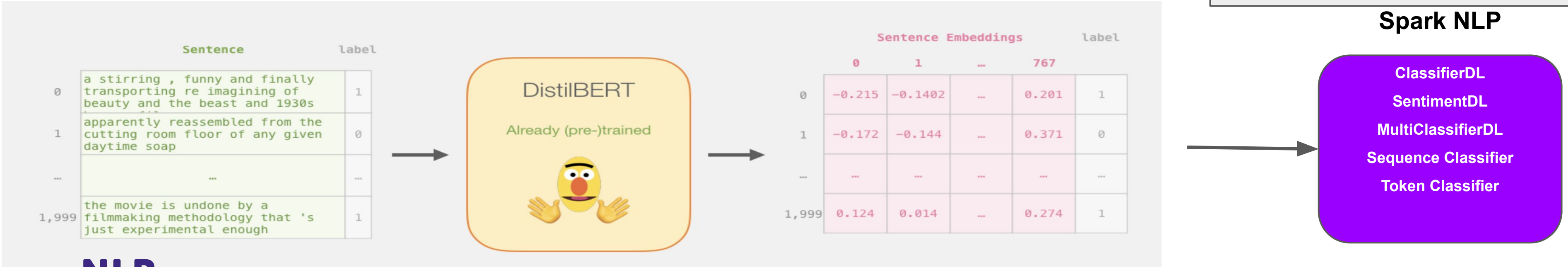
Bert Sequence Classifier | DistilBert Sequence Classifier | DistilBert Sequence Classifier

Albert Sequence Classifier | RoBerta Sequence Classifier | XlmRoBerta Sequence Classifier | XLNet Sequence Classifier | Longformer Sequence Classifier

RoBerta Sentence Embeddings | XlmRoBerta Sentence Embeddings

- Non Transformer Embedding
- Transformer Embedding
- Transformer Sequence Classifier
- Transformer Token Classifier

Spark NLP



What will we learn today

1. Learn how to leverage the 600+ healthcare models for 50 + domains
2. Quick NLU basics recap
3. All Medical Named Entity Recognizer Domains
4. All Assertion Model Domains
5. De-Identification
6. All Resolution Model Domains
7. All Relation Model Domains

How does it work?

```
model= nlu.load(model)
```

- Returns a nlu pipeline object

```
model.predict(data)
```

- Returns a pandas DF

How does it work?

```
model = nlu.load('emotion')
```

- Returns a nlu pipeline object

```
model.predict('I love NLU!')
```

- Returns a pandas DF

EMOTION DETECTION

```
nlu.load('emotion').predict('I love NLU!')
```

sentence_embeddings	category_sentence	category_surprise	category_sadness	category_joy	category_fear	sentence	category	id
[0.027570432052016258, -0.052647676318883896, ...]	0	0.012899903	0.0015578865	0.9760173	0.0095249	I love NLU!	joy	1

NLP SUMMIT HEALTHCARE

Coding Time

Go To
https://github.com/JohnSnowLabs/nlu/tree/master/examples/webinars_conferences_etc/healthcare_summit

NER Models Overview

Domain	Description	Sample NLU Spells	Sample Entities	Sample Predicted Labels	Reference Links
ADE (Adverse Drug Events)	Find adverse drug event (ADE) related entities	med_ner.ade_biobert	Aspirin , vomiting	DRUG, ADE	CADEC, Twimed
Anatomy	Find body parts, anatomical sites a nd reference related entities	med_ner.anatomy	tubules, nasopharyngeal aspirates, embryoid bodies, NK cells, Mitochondrial, tracheoesophageal fistulas, heart, colon cancer, cervical, central nervous system	Tissue structure, Organism substance, Developing anatomical structure, Cell, Cellular component, Immaterial_anatomical_entity, organ, Pathological formation, Organism_subdivision, Anatomical_system	AnEM
Cellular/ Molecular Biology	Find Genes, Molecules, Cell or general Biology related entities	med_ner.cellular.biobert	human T-cell leukemia virus type 1 Tax-responsive , primary T lymphocytes, E1A-immortalized, Spi-B mRNA, zeta-globin	DNA, Cell type, Cell_line, RNA, Protein	JNLPBA
Chemical/Genes/ Proteins	Find Chemical, Gene and Protein related entities	med_ner.chemprot.clinical	nitrogen , β-amyloid , NF-kappaB	CHEMICAL, GENE-Y, GENE-N	ChemProt
Chemical Compounds	Find general chemical compound related entities	med_ner.chemicals	resveratrol , β-polyphenol	CHEM	Dataset by John Snow Labs
Drug Chemicals	Find chemical and drug related entities	med_ner.drugs	potassium , anthracyclines, taxanes	DrugChem, DrugChem, DrugChem	i2b2 + FDA
Posology/Drugs	Find posology and drug related entities	med_ner.posology.biobert	5000 units, Aspirin, 14 days, tablets, daily, topically, 30 mg	DOSAGE, DRUG, DURATION, FORM, FREQUENCY, ROUTE, STRENGTH.	i2b2 + FDA
Risk Factors	Find risk factor of patient related entities	med_ner.risk_factors.biobert	coronary artery disease, hypertension, Smokes 2 packs of cigarettes per day, morbid obesity, Actos, Works in School, diabetic, diabetic	CAD, HYPERTENSION, SMOKER, OBESE, FAMILY HIST, MEDICATION, PHI, HYPERLIPIDEMIA, DIABETES	De-identification and Heart Disease Risk Factors Challenge da
cancer Genetics	Find cancer and genetics related entities	med_ner.cancer	human, Kir 3.3, GIRK3, potassium, GIRK, chromosome 1q21-23, pancreas, tissues, fat andskeletal muscle, KCNJ9, Type II, breast cancer, patients, anthracyclines, taxanes, vinorelbine, patients, breast, vinorelbine inpatients, anthracyclines	Amino acid, Anatomical_system, cancer, Cell, Cellular component, Developing anatomical Structure , Gene or gene product, Immaterial anatomical_entity, Multi-tissue structure, Organ, Organism, Organism_subdivision, Simple_chemical, Tissue	CG TASK of BioNLP 2013
Diseases	Find disease related entities	med_ner.diseases.biobert	the cyst, a large Prolene suture, a very small incisional hernia, the hernia cavity, omentum, the hernia, the wound lesion, The lesion, the existing scar, the cyst, the wound, this cyst down to its base, a small incisional hernia, The cyst	Disease	CG TASK of BioNLP 2013
Bacterial Species	Find bacterial species related entities	med_ner.bacterial_species	Neisseria wadsworthii, N. bacilliformis, Spirochaeta litoralis	SPECIES	Dataset by John Snow Labs
Medical Problem/Test/Treatment	Find medical problem,test and treatment related entities	med_ner.healthcare	respiratory tract infection , Ourexpression studies, atorvastatin	PROBLEM, TEST, TREATMENT	i2b2
Clinical Admission Events	Find clinical admission event related entities	med_ner.admission_events	2007, 12 AM, Headache, blood sample, presented, emergency room, daily	DATE, TIME, PROBLEM, TEST, TREATMENT, OCCURENCE, CLINICAL DEPT, EVIDENTIAL, DURATION, FREQUENCY, ADMISSION, DISCHARGE	Custom i2b2, enriched with Events
Genetic Variants	Find genetic variant related entities	en.med_ner.genetic_variants	rs1061170, p.S45P, T13046C	DNAMutation, ProteinMutation, SNP	TMVAR
PHI (Protected Healthcare Information)	Find PHI(Protected Healthcare) related entities	en.med_ner.deid	2093-01-13, David Hale, Hendrickson,</br> Ora, 7194334, 01/13/93, Oliveira, 25-year-old, 1-11-2000, Cocke County Baptist Hospital, 0295 Keats Street., (302) 786-5227, Brothers Coal-Mine	MEDICALRECORD, ORGANIZATION, DOCTOR, USERNAME, PROFESSION, HEALTHPLAN, URL, CITY, DATE, LOCATION-OTHER, STATE, PATIENT, DEVICE, COUNTRY, ZIP, PHONE, HOSPITAL, EMAIL, IDNUM, SREET, BIOID, FAX, AGE	n2c2 i2b2-PHI
Social Determinants / Demographic Data	Find Social Determinants and Demographic Data Related Entities	med_ner.jsl.enriched	21-day-old, male, congestion, mom, suctioning yellow discharge, she, problems with his breathing, perioral cyanosis, retractions, mom, Tylenol, His, his, respiratory congestion, He, tired, fussy, albuterol	Age, Diagnosis, Dosage, Drug Name, Frequency, Gender, Lab_Name, Lab_Result, Symptom_Name	Dataset by John Snow Labs

NER Models Overview - Continued

Domain	Description	Sample NLU Spells	Sample Entities	Sample Predicted Labels	Reference Links
General Clinical	Find General Clinical Entities	med_ner.jsl.wip.clinical.modifier	28-year-old, female, gestational, diabetes, mellitus, eight, years, prior, type, two, diabetes, mellitus, T2DM, HTG-induced, pancreatitis, three, years, prior, acute, hepatitis, obesity, body, mass, index, BMI, kg/m2, polyuria, polydipsia, poor, appetite, vomiting, Two, weeks, prior, she, five-day, course	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,'	Dataset by John Snow Labs
Radiology	Find Radiology related entities	med_ner.radiology.wip_clinical	Bilateral, breast, ultrasound, ovoid mass, 0.5 x 0.5 x 0.4, cm, anteromedial aspect, left, shoulder, mass, isoechoic echotexture, muscle, internal color flow, benign fibrous tissue, lipoma	ImagingTest, Imaging Technique, ImagingFindings, OtherFindings, BodyPart, Direction, Test, Symptom, Disease Syndrome Disorder, Medical_Device, Procedure, Measurements, Units	Dataset by John Snow Labs, MIMIC-CXR and MT Radiology
Radiology Clinical JSL-V1	Find radiology related entities in clinical setting	med_ner.radiology.wip_greedy_biobert	Bilateral, breast, ultrasound, ovoid mass, 0.5 x 0.5 x 0.4, cm, anteromedial aspect, left, shoulder, mass, isoechoic echotexture, muscle, internal color flow, benign fibrous tissue, lipoma	Test Result, OtherFindings, BodyPart, ImagingFindings, Disease Syndrome Disorder, ImagingTest, Measurements, Procedure, Score, Test, Medical Device, Direction, Symptom, Imaging_Technique, ManualFix, Units	Dataset by John Snow Labs
Genes and Phenotypes	Find Genes and Phenotypes (the observable physical properties of an organism) related entities	med_ner.human_phenotype.gene_biobert	APOC4 , polyhydramnios	GENE, PHENOTYPE	PGR 1, PGR 2
Normalized Genes and Phenotypes	Find Normalized Genes and Phenotypes (the observable physical properties of an organism) related entities	med_ner.human_phenotype.go_biobert	protein complex oligomerization , defective platelet aggregation	GO, HP	PGR 1, PGR 2
Radiology Clinical JSL-V2	Find radiology related entities in clinical setting	med_ner.jsl.wip.clinical.rd		Kidney Disease, HDL, Diet, Test, Imaging Technique, Triglycerides, Obesity, Duration, Weight, Social History Header, ImagingTest, Labour_Delivery, Disease Syndrome Disorder, Communicable Disease, Overweight, Units, Smoking, Score, Substance Quantity, Form, Race Ethnicity, Modifier, Hyperlipidemia, ImagingFindings, Psychological Condition, OtherFindings, Cerebrovascular Disease, Date, Test_Result, VS Finding, Employment, Death Entity, Gender, Oncological, Heart Disease, Medical Device, Total Cholesterol, ManualFix, Time, Route, Pulse, Admission Discharge, RelativeDate , O2 Saturation, Frequency, RelativeTime, Hypertension, Alcohol, Allergen, Fetus NewBorn, Birth Entity, Age, Respiration, Medical History Header, Oxygen Therapy, Section Header, LDL, Treatment, Vital Signs_Header, Direction, BMI, Pregnancy, Sexually Active or Sexual Orientation, Symptom, Clinical Dept, Measurements, Height, Family History_Header, Substance, Strength, Injury or Poisoning, Relationship Status, Blood Pressure, Drug, Temperature, , EKG Findings, Diabetes, BodyPart, Vaccine, Procedure, Dosage	Dataset by John Snow Labs
General Medical Terms	Find general medical terms and medical entities.	med_ner.medmentions		Qualitative Concept, Organization, Manufactured_Object, Amino Acid, Peptide or Protein, Pharmacologic Substance, Professional or Occupational Group, Cell Component, Neoplastic Process, Substance, Laboratory_Procedure, Nucleic Acid Nucleoside or Nucleotide, Research Activity, Gene or Genome, Indicator_Reagent_or_Diagnostic_Aid, Biologic Function, Chemical, Mammal, Molecular Function, Quantitative Concept, Prokaryote, Mental or Behavioral Dysfunction, Injury or Poisoning, Body Location or Region, Spatial Concept, Nucleotide_Sequence, Tissue, Pathologic Function, Body Substance, Fungus, Mental Process, Medical Device, Plant, Health Care Activity, Clinical Attribute, Genetic Function, Food, Therapeutic_or_Preventive_Procedure, Body Part Organ, Organ Component, Geographic Area, Virus, Biomedical or Dental Material, Diagnostic_Procedure, Eukaryote, Anatomical Structure, Organism Attribute, Molecular Biology Research Technique, Organic_Chemical, Cell, Daily or Recreational Activity,	MedMentions

Assertion Models Overview

Domain	Description	Spell	Predicted Entities	Examples	Reference Dataset
Radiology	Predict status of Radiology related entities	`assert.radiology`	`Confirmed`, `Negative`, `Suspected`	- `Confirmed` : X-Ray scan shows `cancer` in lung. - `Negative` : X-Ray scan shows no sign of `cancer` in lung. - `Suspected` :X-Ray raises `suspicion of cancer` in lung but does not confirm it.	Internal Dataset by Annotated by John Snow Lab
Healthcare/Clinical extended and Family JSL powerd	Predict status of, Healthcare/Clinical/Family related entities. Additional training with JSL Dataset	`assert.jsl`	`Present`, `Absent`, `Possible`, `Planned`, `Someoneelse`, `Past`, `Family`, `Hypotetical`	- `Present` : Patient diagnosed with `cancer` in 1999 - `Absent` : No sign of `cancer` was shown by the scans - `Possible` : Tests indicate patient might have `cancer` - `Planned` : `CT-Scan` is scheduled for 23.03.1999 - `Someoneelse` : The patient gave `Aspirin` to daugther. - `Past` : The patient has no more headaches since the operation - `Family` : The patients father has `cancer`. - `Hypotetical` : `Death` could be possible.	2010 i2b2 + Data provided by JSL
Healthcare/Clinical JSL powerd	Predict status of Healthcare/Clinical related entities. Additional training with JSL Dataset	`assert.jsl_large`	`present`, `absent`, `possible`, `planned`, `someoneelse`, `past`	- `present` : Patient diagnosed with `cancer` in 1999 - `absent` : No sign of `cancer` was shown by the scans - `possible` : Tests indicate patient might have `cancer` - `planned` : `CT-Scan` is scheduled for 23.03.1999 - `someoneelse` : The patient gave `Aspirin` to daugther - `past` : The patient has no more headaches since the operation	2010 i2b2 + Data provided by JSL
Healthcare/Clinical classic	Predict status of Healthcare/Clinical related entities	`assert.biobert`	`present` , `absent`, `possible`, `conditional`, `associated_with_someone_else` ,`hypothetical`	- `present` : Patient diagnosed with `cancer` in 1999 - `absent` : No sign of `cancer` was shown by the scans - `possible` : Tests indicate patient might have `cancer` - `conditional` If the test is positive, patient has `AIDS` - `associated with someone else` : The patients father has `cancer`. -`hypothetical` : `Death` could be possible.	2010 i2b2

Resolution Models Overview

Domain/Terminology	Description	Sample NLU Spells	Sample Entities	Sample Predicted Codes	Reference Links
ICD-10 / ICD-10-CM (International Classification of Diseases - Clinical Modification)	Get ICD-10-CM codes of Medical and Clinical Entities. The ICD-10 Clinical Modification (ICD-10-CM) is a modification of the ICD-10, authorized by the World Health Organization, used as a source for diagnosis codes in the U.S. Be aware, ICD10-CM is often referred to as ICD10	resolve.icd10cm.augmented	hypertension, gastritis	I10, K2970	ICD-10-CM, WHO ICD-10-CM
ICD-10-PCS (International Classification of Diseases - Procedure Coding System)	Get ICD-10-PCS codes of Medical and Clinical Entities. The International Classification of Diseases, Procedure Coding System (ICD-10-PCS), is a U.S. cataloging system for procedural code It is maintaining by Centers for Medicare & Medicaid Services	resolve.icd10pcs	hypertension, gastritis	DWY18ZZ, 04723Z6	ICD10-PCS, CMS ICD-10-PCS
ICD-O (International Classification of Diseases, Oncollogy) Topography & Morphology codes	Get ICD-0 codes of Medical and Clinical Entities. The International Classification of Diseases for Oncology (ICD-O), is a domain-specific extension of the International Statistical Classification of Diseases and Related Health Problems for tumor diseases.	resolve.icdo.base	metastatic lung cancer	9050/3 + C38.3, 8001/3 + C39.8	ICD-O Histology Behaviour dataset
HCC (Hierachical Conditional Categories)	Get HCC codes of Medical and Clinical Entities. Hierarchical condition category (HCC) relies on ICD-10 coding to assign risk scores to patients. Along with demographic factors (such as age and gender), insurance companies use HCC coding to assign patients a risk adjustment factor (RAF) score.	resolve.hcc	hypertension, gastritis	139, 188	HCC
ICD-10-CM + HCC Billable	Get ICD-10-CM and HCC codes of Medical and Clinical Entities.	resolve.icd10cm.augmented_billable	metastatic lung cancer	C7800 + ['1', '1', '8']	ICD10-CM HCC
CPT (Current Procedural Terminology)	Get CPT codes of Medical and Clinical Entities. The Current Procedural Terminology(CPT) is developed by the American Medical Association (AMA) and used to assign codes to medical procedures/services/diagonstics. The codes are used to derive the amount of payment a healthcare provider may receives from insurance companies for the provided service.receives	resolve.cpt.procedures_measurements	calcium score, heart surgery	82310, 33257	CPT
LOINC (Logical Observation Identifiers Names and Codes)	Get LOINC codes of Medical and Clinical Entities. Logical Observation Identifiers Names and Codes (LOINC) developed by the U.S. organization Regenstrief Institute	resolve.loinc	acute hepatitis, obesity	28083-4, 50227-8	LOINC
HPO (Human Phenotype Ontology)	Get HPO codes of Medical and Clinical Entities.	resolve.HPO	cancer, bipolar disorder	0002664, 0007302, 0100753	HPO
UMLS (Unified Medical Language System) CUI	Get UMLS codes of Medical and Clinical Entities.	resolve.umls.findings	vomiting, polydipsia, hepatitis	C1963281, C3278316, C1963279	UMLS
SNOMED International (Systematized Nomenclature of Medicine)	Get SNOMED (INT) codes of Medical and Clinical Entities.	resolve.snomed.findings_int	hypertension	148439002	SNOMED
SNOMED CT (Clinical Terms)	Get SNOMED (CT) codes of Medical and Clinical Entities.	resolve.snomed.findings	hypertension	73578008	SNOMED
SNOMED Conditions	Get SNOMED Conditions codes of Medical and Clinical Entities.	resolve.snomed_conditions	schizophrenia	58214004	SNOMED
RxNorm and RxCUI (Concept Uingue Indentifier)	Get Normalized RxNorm and RxCUI codes of Medical, Clinical and Drug Entities.	resolve.rxnorm	50 mg of eltrombopag oral	825427	[RxNorm Overview] [November 2020 RxNorm Clinical Drugs ontology gr

- Relation Domains

Domain	Description	Sample NLU Spells	Predictable Relationships and Explanation
Dates and Clinical Entities	Predict binary temporal relationship between Date Entities and Clinical Entities	relation.date	- 1 for Date Entity and Clinical Entity are related. - 0 for Date Entity and Clinical Entity are not related
Body Parts and Directions	Predict binary direction relationship between Bodypart Entities and Direction Entities	relation.bodypart.direction	- 1 for Body Part and Direction are related - 0 for Body Part and Direction are not related
Body Parts and Problems	Predict binary location relationship between Bodypart Entities and Problem Entities	relation.bodypart.problem	- 1 for Body Part and Problem are related - 0 for Body Part and Problem are not related
Body Parts and Procedures	Predict binary application relationship between Bodypart Entities and Procedure Entities	relation.bodypart.procedure	- 1 for Body Part and Test/Procedure are related - 0 for Body Part and Test/Procedure are not related
Adverse Effects between drugs (ADE)	Predict binary effect relationship between Drugs Entities and Adverse Effects/Problem Entities	relation.ade	- 1 for Adverse Event Entity and Drug are related - 0 for Adverse Event Entity and Drug are not related
Phenotype abnormalities,Genes and Diseases	Predict binary caused by relationship between Phenotype Abnormality Entities, Gene Entities and Disease Entities	relation.humen_phenotype_gene	- 1 for Gene Entity and Phenotype Entity are related - 0 for Gene Entity and Phenotype Entity are not related
Temporal events	Predict multi-class temporal relationship between Time Entities and Event Entities	relation.temporal_events	- AFTER if Any Entity occured after Another Entity - BEFORE if Any Entity occured before Another Entity - OVERLAP if Any Entity during Another Entity
Dates and Tests/Results	Predict multi-class temporal cause,reasoning and conclusion relationship between Date Entities, Test Entities and Result Entities	relation.test_result_date	- relation.test_result_date - is_finding_of for Medical_Entity is found because of Test_Entity - is_result_of for Medical_Entity reason for doing Test_Entity - is_date_of for Date_Entity relates to time of Test/Result - 0 : No relationship
Clinical Problem, Treatment and Tests	Predict multi-class cause,reasoning and effect relationship between Treatment Entities, Problem Entities and Test Entities	relation.clinical	- TrIP : A certain treatment has improved/cured a medical problem - TrWP : A patient's medical problem has deteriorated or worsened because of treatment - TrCP : A treatment caused a medical problem - TrAP : A treatment administered for a medical problem - TrNAP : The administration of a treatment was avoided because of a medical problem - TeRP : A test has revealed some medical problem - TeCP : A test was performed to investigate a medical problem - PIP : Two problems are related to each other
DDI Effects of using Multiple Drugs (Drug Drug Interaction)	Predict multi-class effects, mechanisms and reasoning for DDI effects(Drug Drug Interaction) relationships between Drug Entities	relation.drug_drug_interaction	- DDI-advise when an advice/recommendation regarding a Drug_Entity and Drug_Entity is given - DDI-effect when Drug_Entity and Drug_Entity have an effect on the human body (pharmacodynamic mechanism). Including a clinical finding, signs or symptoms, an increased toxicity or therapeutic failure. - DDI-int when effect between Drug_Entity and Drug_Entity is already known and thus provides no additional information. - DDI-mechanism when Drug_Entity and Drug_Entity are affected by an organism (pharmacokinetic). Such as the changes in levels or concentration in a drug. Used for DDIs that are described by their PK mecha - DDI-false when a Drug_Entity and Drug_Entity have no interaction mentioned in the text.
Posology (Drugs, Dosage, Duration, Frequency, Strength)	Predict multi-class posology relationships between Drug Entities, Dosage Entities, Strength Entities, Route Entities, Form Entities, Duration Entities and Frequency Entities	relation.posology	- DRUG-ADE if Problem_Entity Adverse effect of Drug_Entity - DRUG-DOSAGE if Dosage_Entity refers to a Drug_Entity - DRUG-DURATION if Duration_Entity refers to a Drug_Entity - DRUG-FORM if Mode/Form_Entity refers to intake form of Drug_Entity - DRUG-FREQUENCY if Frequency_Entity refers to usage of Drug_Entity - DRUG-REASON if Problem_Entity is reason for taking Drug_Entity - DRUG-ROUTE if Route_Entity refer to administration method of Drug_Entity - DRUG-STRENGTH if Strength_Entity refers to Drug_Entity
Chemicals and Proteins	Predict Regulator, Upregulator, Downregulator, Agonist, Antagonist, Modulator, Cofactor, Substrate relationships between Chemical Entities and Protein Entities	relation.chemprot	- CPR:1 if One ChemProt Entity is Part of of Another ChemProt Entity - CPR:2 if One ChemProt Entity is Regulator (Direct or Indirect) of Another ChemProt Entity - CPR:3 if One ChemProt Entity is Upregulator/Activator/Indirect Upregulator of Another ChemProt Entity - CPR:4 if One ChemProt Entity is Downregulator/Inhibitor/Indirect Downregulator of Another ChemProt Entity - CPR:5 if One ChemProt Entity is Agonist of Another ChemProt Entity - CPR:6 if One ChemProt Entity is Antagonist of Another ChemProt Entity - CPR:7 if One ChemProt Entity is Modulator (Activator/Inhibitor) of Another ChemProt Entity - CPR:8 if One ChemProt Entity is Cofactor of Another ChemProt Entity - CPR:9 if One ChemProt Entity is Substrate and product of of Another ChemProt Entity - CPR:10 if One ChemProt Entity is Not Related to Another ChemProt Entity

Relation Domain Examples

Domain	Sentence With Relationships	Predicted Relationships for Sample Sentence	Reference Links
Dates and Clinical Entities	This 73 y/o patient had CT on 1/12/95 , with cognitive decline since 8/11/94 .	- 1 for CT and 1/12/95 - 0 for cognitive decline and 1/12/95 - 1 for cognitive decline and 8/11/94	Internal Dataset by Annotated by John Snow Labs
Body Parts and Directions	MRI demonstrated infarction in the upper - brain stem , left cerebellum and right basil ganglia	- 1 for uppper and brain stem - 0 for upper and cerebellum - 1 for left and cerebellum	Internal Dataset by Annotated by John Snow Labs
Body Parts and Problems	Patient reported numbness in his left hand and bleeding from ear .	- 1 for numbness and hand - 0 for numbness and ear - 1 for bleeding and ear	Internal Dataset by Annotated by John Snow Labs
Body Parts and Procedures	The chest was scanned with portable ultrasound and amputation was performed on foot	- 1 for chest and portable ultrasound - 0 for chest and amputation - 1 for foot and amputation	Internal Dataset by Annotated by John Snow Labs
Adverse Effects between drugs (ADE)	Taking Lipitor for 15 years, experienced much sever fatigue ! Doctor moved me to voltaren 2 months ago , so far only experienced cramps	- 1 for sever fatigue and Liptor - 0 for sever fatigue and voltaren - 0 for cramps and Liptor - 1 for cramps and voltaren	Internal Dataset by Annotated by John Snow Labs
Phenotype abnormalities,Genes and Diseases	She has a retinal degeneration , hearing loss and renal failure , short stature, Mutations in the SH3PXD2B gene coding for the Tks4 protein are responsible for the autosomal recessive.	- 1 for hearing loss and SH3PXD2B - 0 for retinal degeneration and hearing loss - 1 for retinal degeneration and autosomal recessive	PGR aclAntology
Temporal events	She is diagnosed with cancer in 1991 . Then she was admitted to Mayo Clinic in May 2000 and discharged in October 2001	- OVERLAP for cancer and 1991 - AFTER for additted and Mayo Clinic - BEFORE for admitted and discharged	Temporal JSL Dataset and n2c2
Dates and Tests/Results	On 23 March 1995 a X-Ray applied to patient because of headache , found tumor in brain	- is finding of for tumor and X-Ray - is result of for headache and X-Ray - is_date_of for 23 March 1995 and X-Ray	Internal Dataset by Annotated by John Snow Labs
Clinical Problem, Treatment and Tests	- TrIP : infection resolved with antibiotic course - TrWP : the tumor was growing despite the drain - TrCP : penicillin causes a rash - TrAP : Dexamphetamine for narcolepsy - TrNAP : Ralafen was not given because of ulcers - TeRP : an echocardiogram revealed a pericardial effusion - TeCP : chest x-ray for pneumonia - PIP : Azotemia presumed secondary to sepsis	- TrIP for infection and antibiotic course - TrWP for tumor and drain - TrCP for penicillin and rash - TrAP for Dexamphetamine and narcolepsy - TrNAP for Ralafen and ulcers - TeRP for echocardiogram and pericardial effusion - TeCP for chest x-ray and pneumonia - PIP for Azotemia and sepsis	2010 i2b2 relation challenge
DDI Effects of using Multiple Drugs (Drug Drug Interaction)	- DDI-advise : UROXATRAL should not be used in combination with other alpha-blockers - DDI-effect : Chlorthalidone may potentiate the action of other antihypertensive drugs - DDI-int : The interaction of omeprazole and ketoconazole has been established - DDI-mechanism : Grepafloxacin may inhibit the metabolism of theobromine - DDI-false : Aspirin does not interact with Chlorthalidone	- DDI-advise for UROXATRAL and alpha-blockers - DDI-effect for Chlorthalidone and antihypertensive drugs - DDI-int for omeprazole and ketoconazole - DDI-mechanism for Grepafloxacin and theobromine - DDI-false for Aspirin and Chlorthalidone	DDI Extraction corpus
Posology (Drugs, Dosage, Duration, Frequency, Strength)	- DRUG-ADE : had a headache after taking Paracetamol - DRUG-DOSAGE : took 0.5ML of Celstone - DRUG-DURATION : took Aspirin daily for two weeks - DRUG-FORM : took Aspirin as tablets - DRUG-FREQUENCY : Aspirin usage is weekly - DRUG-REASON : Took Aspirin because of headache - DRUG-ROUTE : Aspirin taken orally - DRUG-STRENGTH : 2mg of Aspirin	- DRUG-ADE for headache and Paracetamol - DRUG-DOSAGE for 0.5ML and Celstone - DRUG-DURATION for Aspirin and for two weeks - DRUG-FORM for Aspirin and tablets - DRUG-FREQUENCY for Aspirin and weekly - DRUG-REASON for Aspirin and headache - DRUG-ROUTE for Aspirin and orally - DRUG-STRENGTH for 2mg and Aspirin	Magge, Scotch, Gonzalez-Hernandez (2018)
Chemicals and Proteins	- CPR:1 (Part of) : The amino acid sequence of the rabbit alpha(2A)-adrenoceptor has many interesting properties. - CPR:2 (Regulator) : Triacsin inhibited ACS activity - CPR:3 (Upregulator) : Ibandronate increases the expression of the FAS gene - CPR:4 (Downregulator) : Vitamin C treatment resulted in reduced C-Rel nuclear translocation - CPR:5 (Agonist) : Reports show tricyclic antidepressants act as agnonists at distinct opioid receptors - CPR:6 (Antagonist) : GDC-0152 is a drug triggers tumor cell apoptosis by selectively antagonizing LAPs - CPR:7 (Modulator) : Hydrogen sulfide is a allosteric modulator of ATP-sensitive potassium channels - CPR:8 (Cofactor) : polyinosinic:polycytidylic acid and the IFNα/β demonstrate capability of endogenous IFN. - CPR:9 (Substrate) : ZIP9 plays an important role in the transport and toxicity of Cd(2+) cells - CPR:10 (Not Related) : Studies indicate that GSK-3β inhibition by palinurin cannot be competed out by ATP	- CPR:1 (Part of) for amino acid and rabbit alpha(2A)-adrenoceptor - CPR:2 (Regulator) for Triacsin and ACS - CPR:3 (Upregulator) for Ibandronate and FAS gene - CPR:4 (Downregulator) for Vitamin C and C-Rel - CPR:5 (Agonist) for tricyclic antidepressants and opioid receptors - CPR:6 (Antagonist) (Antagonist) for GDC-0152 and LAPs - CPR:7 (Modulator) for Hydrogen sulfide and ATP-sensitive potassium channels - CPR:8 (Cofactor) for polyinosinic:polycytidylic acid and IFNα/β - CPR:9 (Substrate) for ZIP9 and Cd(2+) cells - CPR:10 (Not Related) for GSK-3β and ATP	ChemProt Paper

NLU : Apache License 2.0

```python
# Multiple binary sentiment classifiers trained on various datasets
nlu.load('classify.sentiment').predict('I love NLU and Python WebDev Conf 2021!')
nlu.load('classify.sentiment.imdb').predict('The Matrix was a pretty good movie')
nlu.load('classify.sentiment.twitter').predict('@elonmusk Tesla stock price is too high imo')

# Translate between 200 languages
nlu.load('en.translate_to.zh').predict('NLU can translate between 200 languages!')

# Spellchecking
nlu.load('spell').predict('I liek to live dangertus!')

# Extract Named Entities
nlu.load('ner').predict('Donald Trump and John Biden dont share many oppinions')

# Unsupervised Keyword Extraction
nlu.load('yake').predict('Weights extract keywords withouth requiring weights!')

# Over 50+ classifiers on various problems
nlu.load('classify.emotion').predict('He was suprised by the diversity of NLU')
nlu.load('classify.spam').predict('Hello you are the heir to a 100 Million fortune!')
nlu.load('classify.fakenews').predict('Unicorns landed on mars!')
nlu.load('classify.sarcasm').predict('love the teachers who give exams the day after halloween')
nlu.load('en.classify.question').predict('How expensive is the Watch?')
nlu.load('en.classify.toxic').predict('You are to stupid')
nlu.load('classify.cyberbullying').predict('Women belong in the kitchen!') #sorry

# Get BERTology and Transformer Embeddings for Sentences and Words
nlu.load('bert').predict('BERTolgy Word embeddings!')
nlu.load('bert elmo albert glove').predict('Multiple BERTolgy Word embeddings!')
nlu.load('embed_sentence.bert ').predict('BERTolgy Sentence embeddings!')

# Text cleaning and Pre-Processing
nlu.load('lemmatize').predict('Get me the lemmatized version of a string')
nlu.load('normalize').predict('Get me the lemmatized version of a string')
nlu.load('clean').predict('Get me the lemmatized version of a string')

# Grammatical Parts of Speech
nlu.load('pos').predict('Extract Parts of Speech')
```

- 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!

NLP SUMMIT HEALTHCARE

nlu_viz_cheatsheet.py

```python
data = 'I want some pretty visualizations please!'

# Viz detected entities
nlu.load('ner').viz(data)

# Viz labeled dependency tree and POS tags
nlu.load('dep.typed').viz(data)

# Viz asserted statuses
nlu.load('med_ner.clinical assert').viz(data)

# Viz resolved sentences
nlu.load('med_ner.jsl.wip.clinical resolve.icd10cm').viz(data)

# Viz resolved entities
nlu.load('med_ner.jsl.wip.clinical resolve_chunk.rxnorm.in').viz(data)

# Viz extracted relationships between entities
nlu.load('med_ner.jsl.wip.clinical relation.temporal_events').viz(data)
```

nlu_streamlit_cheatsheet.py

```python
# Input text data for visualizations
data = 'Billy from Berlin wants some pretty visualizations please!'

# Full UI with self generating Python code snippets for every feature
nlu.load('ner').viz_streamlit(data)

# Classification
nlu.load('sentiment').viz_streamlit_classes(data)

# Named Entity Recognition (NER)
nlu.load('ner').viz_streamlit_ner(data)

# Dependency Tree and Part of Speech (POS) Tags
nlu.load('dep.typed').viz_streamlit_dep_tree(data)

# Token features
nlu.load('stemm pos spell').viz_streamlit_token(data)

# Calculate similarity between two texts based on Word Embeddings
nlu.load('bert').viz_streamlit_word_similarity(['I love  NLU! <3','I also Streamlit! <3'])

# Raw visualizations in Streamlit
nlu.load('<Model>').viz(data, write_to_streamlit=True)

# Predict on any datatype and returns Pandas df
nlu.load('<Model>').predict(data)

# Enable caching
nlu.enable_streamlit_caching()
```

NLP SUMMIT HEALTHCARE

More NLU Ressources

- Join our Slack
- NLU Website
- NLU Github
- Many more NLU example tutorials
- Overview of every powerful nlu 1-liner
- Checkout the Modelshub for an overview of all models
- Checkout the NLU Namespace where you can find every model as a tabel
- Intro to NLU article
- In Depth and Easy Sentence Similarity Tutorial, with StackOverflow Questions using BERTology embeddings
- 1 line of Python code for BERT, ALBERT, ELMO, ELECTRA, XLNET, GLOVE, Part of Speech with NLU and t-SNE

Webinars and Videos

- NLU & Streamlit Tutorial
- Multi Lingual NLU Webinar - Tutorial on Chinese News dataset
- John Snow Labs NLU: Become a Data Science Superhero with One Line of Python code
- Python Web Def Conf - Python's NLU library: 4,000+ Models, 200+ Languages, State of the Art Accuracy, 1 Line of Code
- NYC/DC NLP Meetup with NLU

Thank you.

christian@JohnSnowLabs.com

@ckl_it

In/Christian-Kasim-Loan

Medium.com/@Christian.Kasim.Loan

NLP SUMMIT HEALTHCARE