Biomedical Named Entity Recognition at Scale

CADL@ICPR 2020
Jan 11, 2021

John Snow LABS

Veysel Kocaman
Lead Data Scientist

David Talby, PhD
CTO

John Snow LABS

Spark NLP Modules

Clinical Entity Recognition

40 units DOSAGE of
insulin glargine DRUG
at night FREQUENCY

Clinical Entity Linking

Suspect diabetes SNOMED-CT: 473127005
Lisinopril 10 MG RxNorm: 316151
Hyponatremia ICD-10: E87.1

Assertion Status

Fever and sore throat → PRESENT
No stomach pain → ABSENT
Father with Alzheimer → FAMILY

Relation Extraction

AFTER
Admitted (Occurrence) for nausea (Symptom) due to chemo (Treatment)
CAUSED BY

Algorithms

Extract Knowledge
- Entity Linker
- Entity Disambiguator
- Document Classifier
- Contextual Parser

De-identify text
- Structured Data
- Unstructured Text
- Obfuscator
- Generalizer

Split Text
- Sentence Detector
- Deep Sentence Detector
- Tokenizer
- nGram Generator

Clean Medical Text
- Spell Checking
- Spell Correction
- Normalizer
- Stopword Cleaner

Clinical Grammar
- Stemmer
- Lemmatizer
- Part of Speech Tagger
- Dependency Parser

Find in Text
- Text Matcher
- Regex Matcher
- Date Matcher
- Chunker

Content

Medical Transformers

JSL-BERT-Clinical, BioBERT, ClinicalBERT, GloVe-Med, GloVe-ICD-O

Linked Medical Terminologies

SNOMED-CT, CPT, ICD-10-CM, ICD-O, ICD-10-PCS, RxNorm, LOINC

50+ Pretrained Models

Clinical: Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections

Anatomy: Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical

Drugs: Name, Dosage, Strength, Route, Duration, Frequency

Demographics: Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs

Risk Factors: Smoking, Obesity, Diabetes, Hypertension, Substance Abuse

Sensitive Data: Patient Name, Address, Phone, Email, Dates, Providers, Identifiers

Trainable & Tunable	Scalable to a Cluster	Fast Inference	Hardware Optimized	Community
	Apache Spark ML Pipelines	LightPipeline	intel, NVIDIA	NLP SUMMIT

Entity Recognition

I love Lucy PERSON

Information Extraction

They met Last week DATE --> 29-04-2020

Sentiment Analysis

Document Classification

Algorithms

Split Text
- Sentence Detector
- Deep Sentence Detector
- Tokenizer
- nGram Generator

Clean Text
- Spell Checking
- Spell Correction
- Normalizer
- Stopword Cleaner

Understand Grammar
- Stemmer
- Lemmatizer
- Part of Speech Tagger
- Dependency Parser

Find in Text
- Text Matcher
- Regex Matcher
- Date Matcher
- Chunker

Content

Transformers

GloVe, ELMO, BERT, ALBERT, XLNet

Languages

Bulgarian, Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Finnish, Norwegian, Polish, Portuguese, Spanish, Romanian, Russian, Slovak, Swedish, Turkish, Ukrainian

Models

90+ Pretrained

Pipelines

70+ Pretrained

Trainable & Tunable	Scalable to a Cluster	Fast Inference	Hardware Optimized	Community
	Apache Spark ML Pipelines	LightPipeline	intel, NVIDIA	NLP SUMMIT

spark-nlp

Summary

PyPI link	https://pypi.org/project/spark-nlp
Total downloads	2,674,517
Total downloads - 30 days	376,927
Total downloads - 7 days	81,617

Daily ~ 10K
Monthly ~ 350K

The most widely used NLP library in industry (3 yrs in a row)

Biomedical Named Entity Recognition at Scale

- 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.
- Delivering the first production-grade scalable NER model implementation.
- 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).
- 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.

Biomedical Named Entity Recognition at Scale

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m2 , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The β-hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again .

Clinical NER

Color codes:PROBLEM, TREATMENT, TEST,

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,

Deid NER

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.

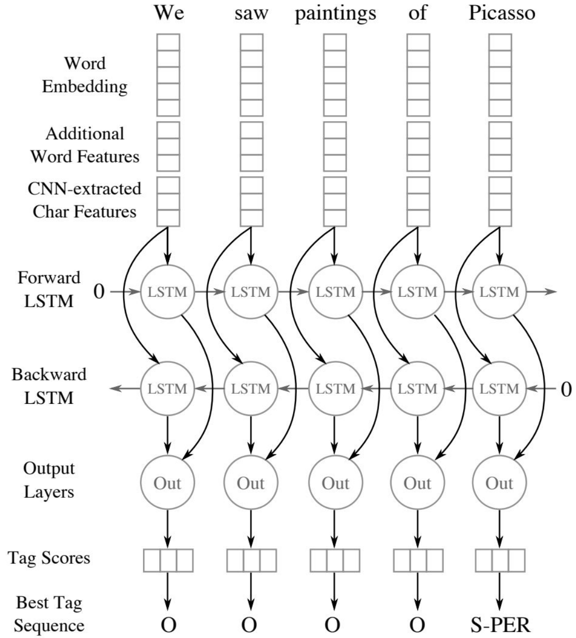
Faster inference

```
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).

NER Model Implementation in Spark NLP

Feature-engineered machine learning systems	Dict	SP	DU	EN	GE
Carreras et al. (2002) binary AdaBoost classifiers	Yes	81.39	77.05	-	-
Malouf (2002) - Maximum Entropy (ME) + features	Yes	73.66	68.08	-	-
Li et al. (2005) SVM with class weights	Yes	-	-	88.3	-
Passos et al. (2014) CRF	Yes	-	-	90.90	-
Ando and Zhang (2005a) Semi-supervised state of the art	No	-	-	89.31	75.27
Agerri and Rigau (2016)	Yes	**84.16**	**85.04**	**91.36**	**76.42**
Feature-inferring neural network word models					
Collobert et al. (2011) Vanilla NN +SLL / Conv-CRF	No	-	-	81.47	-
Huang et al. (2015) Bi-LSTM+CRF	No	-	-	84.26	-
Yan et al. (2016) Win-BiLSTM (English), FF (German) (Many fets)	Yes	-	-	88.91	**76.12**
Collobert et al. (2011) Conv-CRF (SENNA+Gazetteer)	Yes	-	-	89.59	-
Huang et al. (2015) Bi-LSTM+CRF+ (SENNA+Gazetteer)	Yes	-	-	**90.10**	-
Feature-inferring neural network character models					
Gillick et al. (2015) – BTS	No	**82.95**	**82.84**	**86.50**	**76.22**
Kuru et al. (2016) CharNER	No	82.18	79.36	84.52	70.12
Feature-inferring neural network word + character models					
Yang et al. (2017)	Yes	85.77	**85.19**	91.26	-
Luo (2015)	Yes	-	-	91.20	-
Chiu and Nichols (2015)	Yes	-	-	**91.62**	-
Ma and Hovy (2016)	No	-	-	91.21	-
Santos and Guimaraes (2015)	No	82.21	-	-	-
Lample et al. (2016)	No	85.75	81.74	90.94	**78.76**
Bharadwaj et al. (2016)	Yes	**85.81**	-	-	-
Dernoncourt et al. (2017)	No	-	-	90.5	-
Feature-inferring neural network word + character + affix models					
Re-implementation of Lample et al. (2016) (100 Epochs)	No	85.34	85.27	90.24	78.44
Yadav et al. (2018)(100 Epochs)	No	86.92	87.50	90.69	78.56
Yadav et al. (2018) (150 Epochs)	No	**87.26**	**87.54**	90.86	**79.01**

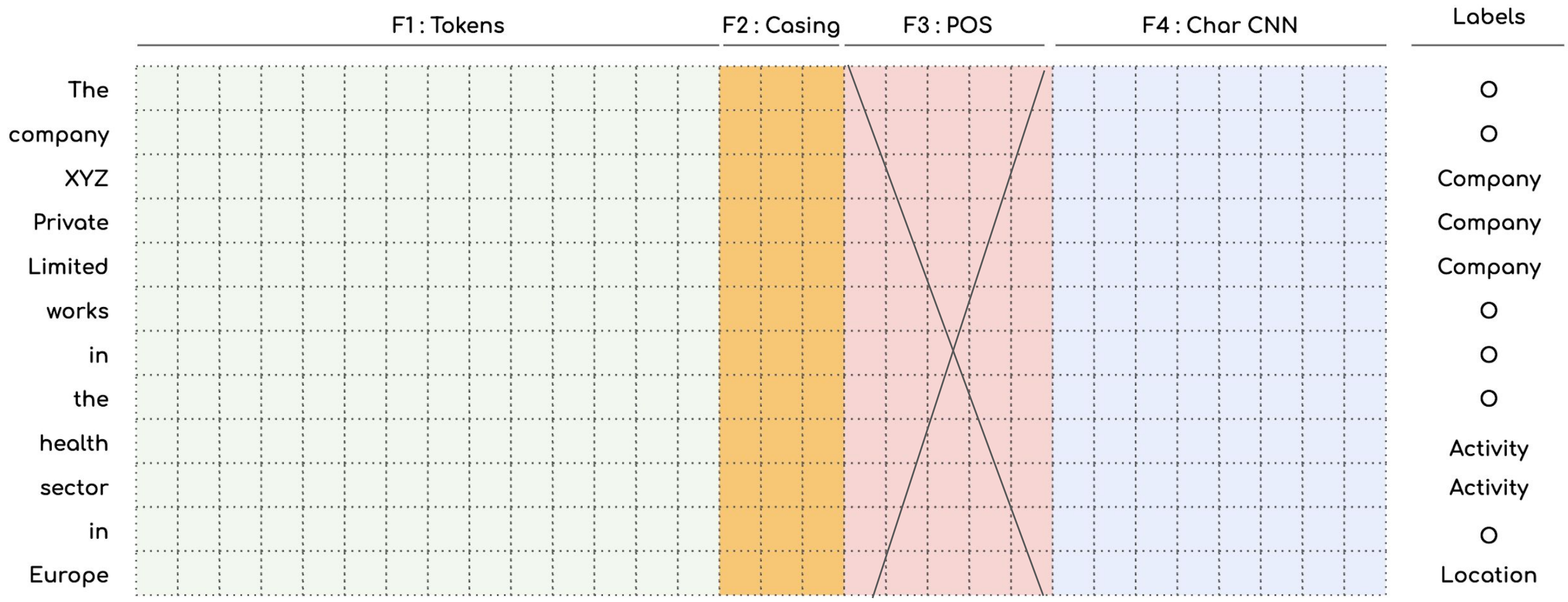


Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

John Snow LABS

NER Model Implementation in Spark NLP

Char-CNN-BiLSTM

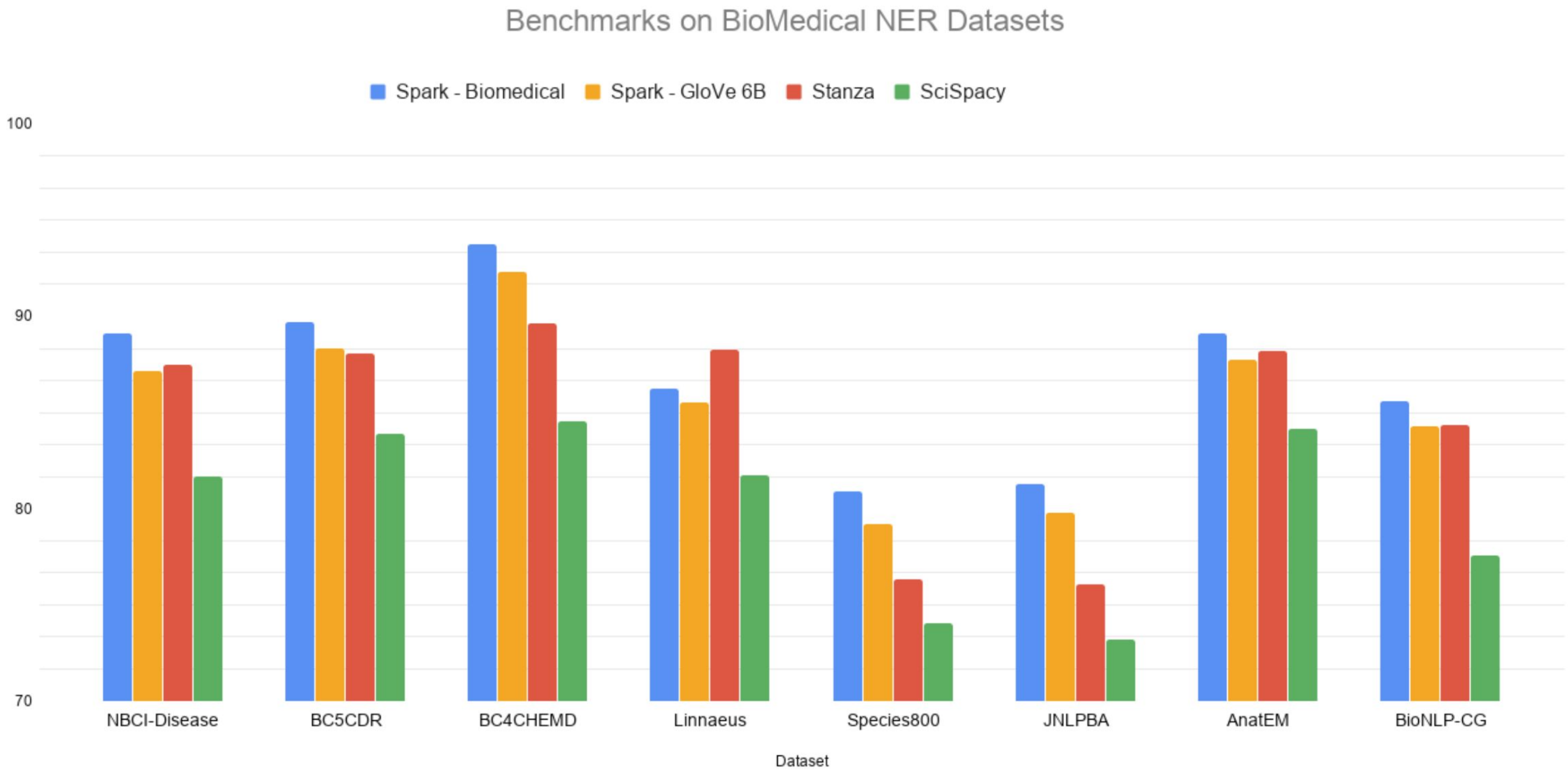


Biomedical Named Entity Recognition at Scale

TABLE II: NER performance across different datasets in the biomedical domain. All scores reported are micro-averaged test F1 excluding O's. Stanza results are from the paper reported in Zhang et al. [2020], SciSpaCy results are from the scispacy-medium models reported in Neumann et al. [2019]. The official training and validation sets are merged and used for training and then the models are evaluated on the original test sets. For reproducibility purposes, we use the preprocessed versions of these datasets provided by Wang et al. [2019] and also used by Stanza. Spark-x prefix in the table indicates our implementation. Bold scores represent the best scores in the respective row.

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

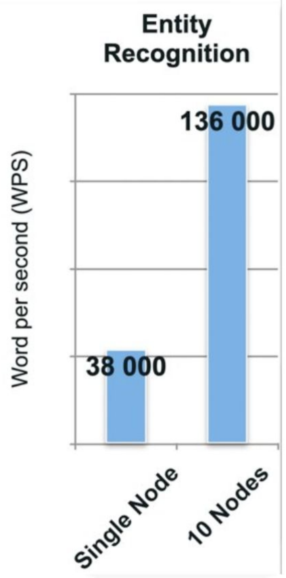
Biomedical Named Entity Recognition at Scale



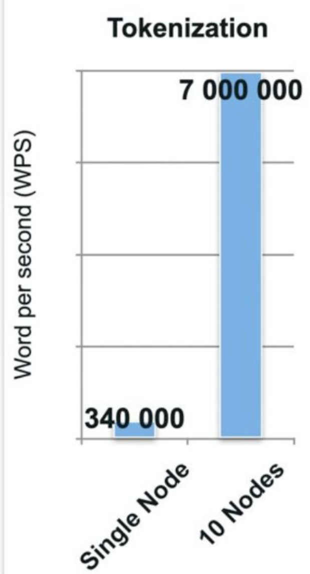
Biomedical Named Entity Recognition at Scale

TF in Keras vs TF in Apache Spark

Table 3: Performance evaluation on biomedical NER datasets using the same BiLSTM-CNN-Char architecture in TensorFlow and Spark NLP under the same settings for each dataset. The Spark NLP implementation beats the same architecture 7 out of 8 times in terms of macro F1 score and is faster to train in half of the datasets *(macro average F1 score, embeddings glove6B_300d, lr 0.001, dropout 0.5, LSTM state size 200, epoch 10, batch size 128, optimizer Adam)*. Bold letters represent best results.



	Tensorflow 1.15 (Keras)		Spark NLP	
Dataset	time (sec)	macro-F1	time (sec)	macro-F1
BC5CDR-disease	409	0.840	**336**	**0.858**
BC5CDR-chem	438	0.848	**367**	**0.894**
BC4CHEMD	2954	0.890	**2719**	**0.936**
NCBI-Disease	312	0.882	**269**	**0.883**
JNLPBA	**495**	0.705	743	**0.758**
Species800	**215**	0.813	232	**0.820**
Linnaeus	**709**	**0.787**	730	0.759



Biomedical Named Entity Recognition at Scale

Word Embeddings Coverage Ratio

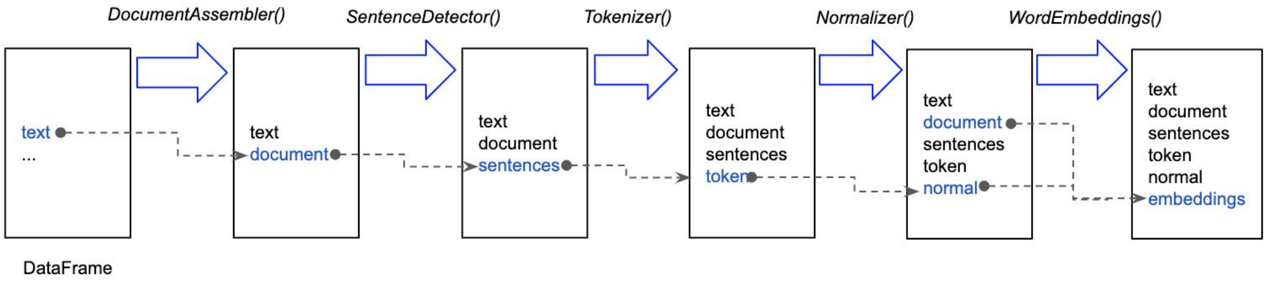
TABLE I: Word embeddings coverage ratios on biomedical datasets. Our domain specific embeddings have near-perfect word coverages. The average word coverage of our implementation of domain specific word embeddings (we call it Spark-Biomedical Embeddings in this study) is 99.5% and the average word coverage of Glove6B embeddings is 96.1% on the biomedical datasets used in this study)

Dataset	Spark-Biomedical Embeddings		Spark-Glove6B Embeddings	
	Training set	Test set	Training set	Test set
NBCI-Disease	99.700	99.695	96.703	96.710
BC5CDR	99.171	99.106	96.059	95.795
BC4CHEMD	99.571	99.551	96.409	96.434
Linnaeus	99.162	99.181	96.801	96.867
Species800	99.350	99.345	95.909	96.258
JNLPBA	99.530	99.496	92.566	92.690
AnatEM	99.580	99.623	96.992	96.945
BioNLP-CG	99.859	99.814	97.750	96.663

Biomedical Named Entity Recognition at Scale

Appendix

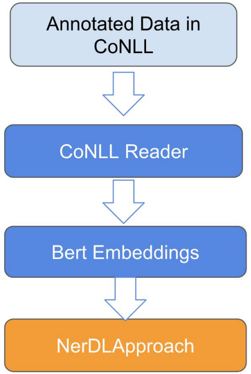
```
from pyspark.ml import Pipeline
import sparknlp
from sparknlp.training import CoNLL
from sparknlp.annotator import *

spark = sparknlp.start()

training_data = CoNLL().readDataset(spark, ’
    BC5CDR_train.conll’)

word_embedder = WordEmbeddings.pretrained(’
    wikiner_6B_300’, ’xx’) \
  .setInputCols(["sentence",’token’])\
  .setOutputCol("embeddings")

nerTagger = NerDLApproach()\
   .setInputCols(["sentence", "token", "embeddings"])
     \
   .setLabelColumn("label")\
   .setOutputCol("ner")\
   .setMaxEpochs(10)\
   .setDropout(0.5)\
   .setLr(0.001)\
   .setPo(0.005)\
   .setBatchSize(8)\
   .setValidationSplit(0.2)\

pipeline = Pipeline(
    stages = [
    word_embedder,
    nerTagger
  ])

ner_model = pipeline.fit(training_data)
```



BIO schema

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

John Smith ⇒ PERSON
New York ⇒ LOCATION



John Snow LABS

Biomedical Named Entity Recognition at Scale

CADL@ICPR 2020

Jan 11, 2021

John Snow LABS

Veysel Kocaman
Lead Data Scientist

David Talby, PhD
CTO