TraiNER
Active Entity Recognition

Enterprise Data Management WS 17/18

Beuth Hochschule für Technik

Team

Frontend	Backend	Machine Learning
Luise Napieralski Philipp Behrendt	Benjamin Rühl Stephan Hausdörfer Philipp Hannasky	Tom Oberhauser Robin Mehlitz Marlene Brüggemann Christopher Kümmel

Masterprojekt:
Simon Lischka, Wadim Lewin, Vladimir Schmidt, Robin Mehlitz, Tom Oberhauser

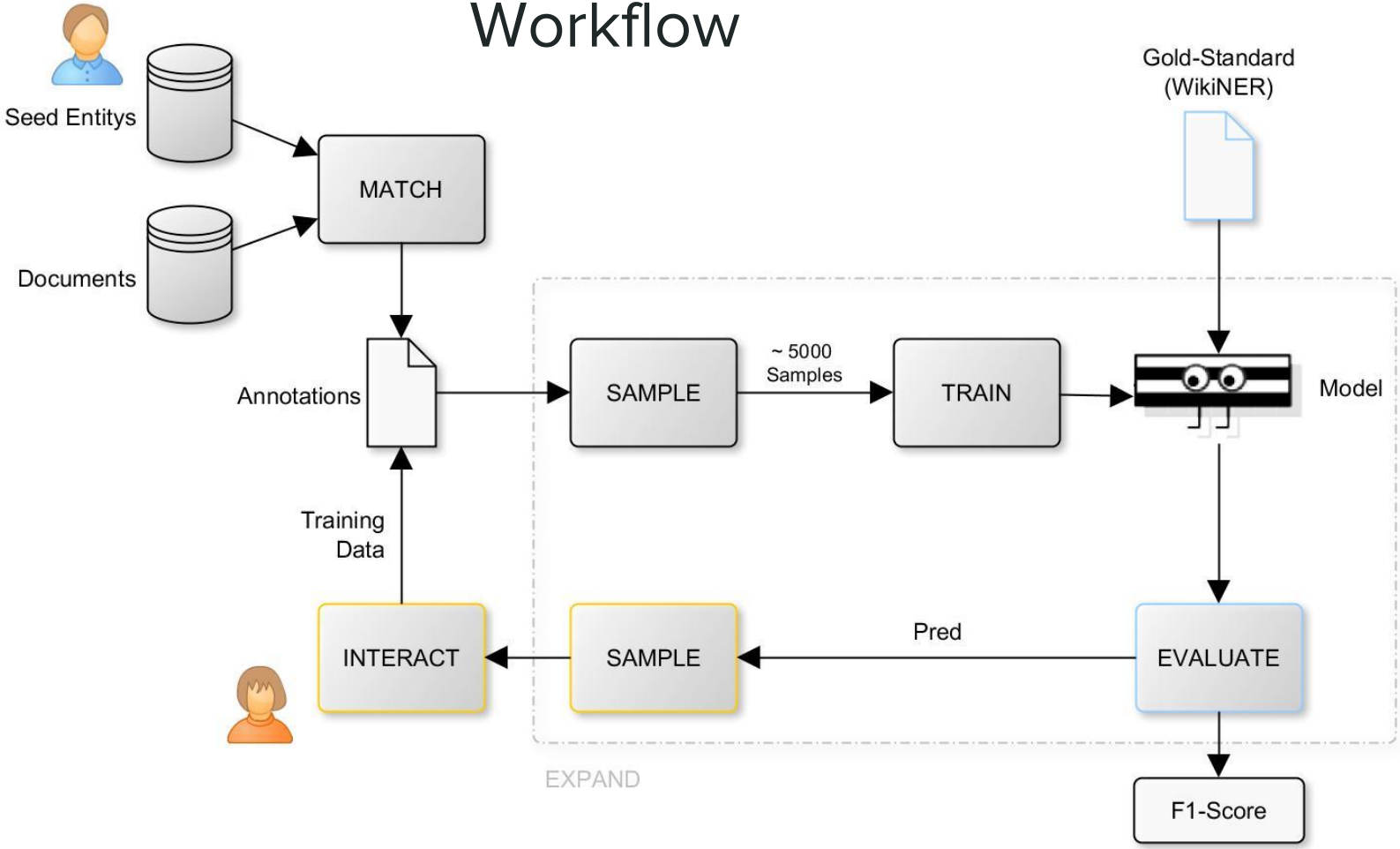
Project Description

- Problem
  - Not enough domain specific training data for NER models creation
  - Example
    - Finding fashion brands / disease names in texts
    - A lot of documents given
    - No gold standard to train / evaluate the model
- Solution approaches
  - Manual creation with personnel / manpower ➜ high effort in time and costs
  - Entity list and string matches ➜ too less / too many matches ➜ too less profit for too much time / costs effort
- Our approach - Active Learning
  - User trains the model iterative ➜ lower effort ➜ good training data / model

Demo

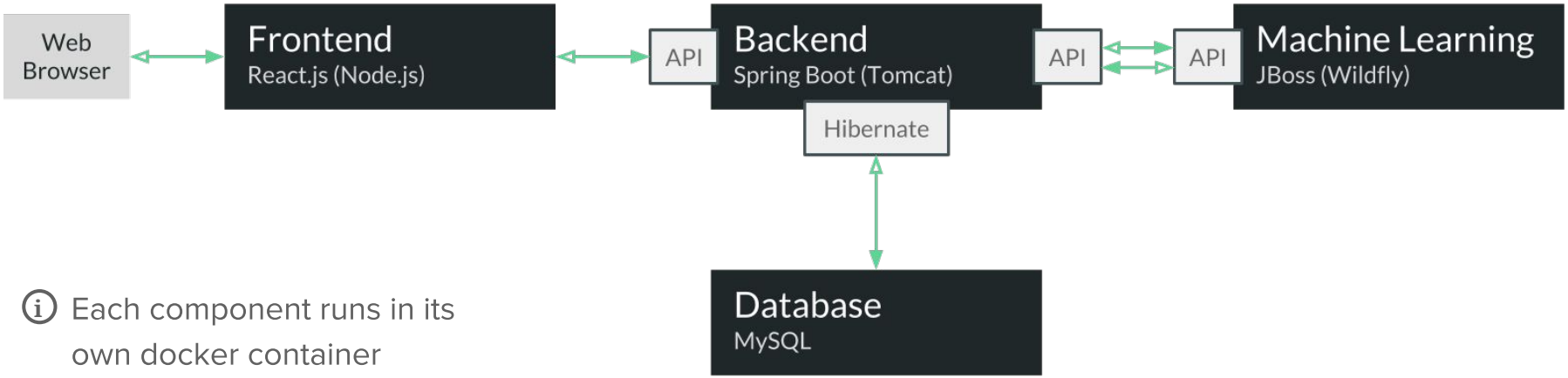
Workflow

Workflow


Seed Entitys
Documents
MATCH
Gold-Standard
(WikiNER)
Annotations
SAMPLE
~ 5000
Samples
TRAIN
Model
Training
Data
INTERACT
SAMPLE
Pred
EVALUATE
EXPAND
F1-Score

Architecture

Architecture

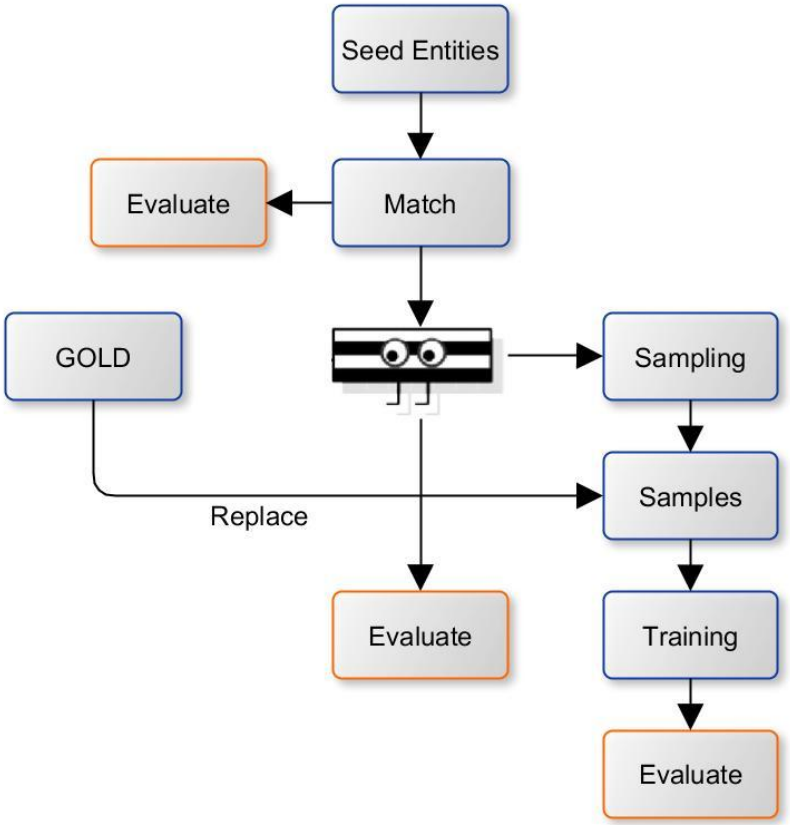


ⓘ Each component runs in its own docker container

Demo

Machine Learning Component Evaluation

Evaluation Architecture for “Gold User”



WikiNER Subset

- Dataset
  - 250 random documents
  - 7611 sentences (out of approx 142k)
  - with Gold Standard
- Seed-Entities
  - Wikidata - EN - Names
  - approx 5.6 millions

- Match

Docs	Tokns	Anns	Pred	TP	FP	TN	FN	TAcc	Prec	Rec	F1
250	193029	16832	44058	13016	31042	0	3816	100.00	29.54	77.33	42.75

Sampling Strategy Evaluation - F1 - I

#	Re-Train w/o	Re-Train w/ random	New-Train w/ random	Re-Train w/ uncertainty	New-Train w/ uncertainty
1	42,51	42,89	43,10	43,44	43,10
2	42,92	42,74	42,79	42,48	43,48
3	43,32	43,72	42,52	43,64	43,35
4	42,43	43,01	42,00	42,94	43,32
5	42,81	43,92	43,37	43,10	43,24
6	43,19	43,24	42,24	43,25	48,16

batch size = 32 sentences; epochs = 1;

Sampling Strategy Evaluation - F1 - II

#	Re-Train w/ uncertainty (5 epochs)	Re-Train w/ uncertainty (200 sentences - 1 epoch)	Re-Train w/ uncertainty (200 sentences - 10 epochs)
1	42,95	42,83	42,74
2	43,16	44,52	44,35
3	43,00	44,52	43,32
4	43,05	43,84	44,08
5	43,11	43,26	44,03
6	43,16	43,35	43,69

batch size = 32 sentences;

Match vs. Model - I

- Medicine Dataset
  - 9229 Sentences
  - no Gold Standard
- Seed-Entities
  - 14235 medical terms

Training on full dataset

Docs	Tokns	Anns	Pred	TP	FP	TN	FN	TAcc	Prec	Rec	F1
1078	185056	5146	5144	5143	1	0	3	96.40	99.98	99.94	99.96

Training on random 5000 sentences

Docs	Tokns	Anns	Pred	TP	FP	TN	FN	TAcc	Prec	Rec	F1
1078	185056	5146	4800	4214	586	0	932	96.80	87.79	81.89	84.74

batch size = 32 sentences; epochs = 20;

Match vs. Model - II

Match

Autism is a developmental disorder characterized by troubles with social interaction and communication. Often there is also restricted and repetitive behavior. Parents usually notice signs in the first two or three years of their child's life. These signs often develop gradually, though some children with autism reach their developmental milestones at a normal pace and then worsen. Autism is caused by a combination of genetic and environmental factors. Risk factors include certain infections during

Trained Model (5000 sentences)

Autism is a developmental disorder characterized by troubles with social interaction and communication. Often there is also restricted and repetitive behavior. Parents usually notice signs in the first two or three years of their child's life. These signs often develop gradually, though some children with autism reach their developmental milestones at a normal pace and then worsen. Autism is caused by a combination of genetic and environmental factors. Risk factors include certain infections during

Match vs. Model - III

Match

arthritis. The most common forms are osteoarthritis (degenerative joint disease) and rheumatoid arthritis. Osteoarthritis usually occurs with age and affects the fingers, knees, and hips. Rheumatoid arthritis is an autoimmune disorder that often affects the hands and feet. Other types include gout, lupus, fibromyalgia, and septic arthritis. They are all types of rheumatic disease. Treatment may include resting the joint and alternating between applying ice and heat. Weight loss and exercise may also be

Trained Model (5000 sentences)

arthritis. The most common forms are osteoarthritis (degenerative joint disease) and rheumatoid arthritis. Osteoarthritis usually occurs with age and affects the fingers, knees, and hips. Rheumatoid arthritis is an autoimmune disorder that often affects the hands and feet. Other types include gout, lupus, fibromyalgia, and septic arthritis. They are all types of rheumatic disease. Treatment may include resting the joint and alternating between applying ice and heat. Weight loss and exercise may also be

Sum up

Milestone Results

- Milestone 1
  - First UI design
  - First match on WikiNER but match was broken
  - Whole process designed
- Milestone 2
  - UI implemented with first functionality (annotation interaction)
  - Frontend, Backend, Machine Learning Server standalones, not integrated
  - Fixed match operator
  - “Gold User” architecture implemented
- Milestone 3
  - Frontend - Backend - Machine Learning Server integrated
  - “Gold User” for Machine Learning evaluation used
  - For Machine Learning sample strategies implemented

Lessons learned

- Faster integration of the components
- Training took long - outsourced to cluster
- Convenient domain specific dataset

Future Work

Future Work

- Evaluation of model when no gold standard is given
- Export and retrain of the model from the UI
- Apply annotations over all documents
- Optical feedback during training

- Learning how to Active Learn: A Deep Reinforcement Learning Approach

  Meng Fang, Yuan Li, Trevor Cohn - *https://arxiv.org/abs/1708.02383*
- Medicine dataset with GOLD annotations
  - take articles from WikiNER which point to those from diseases dataset